



**Manual**  
**NATIONAL WORKSHOP CUM TRAINING PROGRAMME**  
**ON**  
**STATISTICAL TOOLS FOR RESEARCH DATA ANALYSIS (Series II)**

**DURATION OF THE PROGRAMME**  
**TWO WEEKS**  
(Starting from 29TH May, 2017)

by  
**Society For Application of Statistics  
in Agriculture and Allied Sciences (SASAA)**  
and  
**Department of Agricultural Statistics  
Bidhan Chandra Krishi Viswavidyalaya**

**Venue**  
**Department of Agricultural Statistics, Faculty of Agriculture**  
**Bidhan Chandra Krishi Viswavidyalaya**  
P.O.- Krishi Viswavidyalaya, Nadia, West Bengal, 741252





**Manual**  
**NATIONAL WORKSHOP CUM TRAINING**  
**PROGRAMME**  
**ON**  
**STATISTICAL TOOLS FOR RESEARCH DATA ANALYSIS**  
**(Series II)**

**DURATION OF THE PROGRAMME**  
**TWO WEEKS**  
**(Starting from 29<sup>TH</sup> May, 2017)**

**by**  
**Society For Application of Statistics**  
**in Agriculture and Allied Sciences (SASAA)**  
**and**  
**Department of Agricultural Statistics**  
**Bidhan Chandra Krishi Viswavidyalaya**

**Venue**  
**Department of Agricultural Statistics, Faculty of Agriculture**  
**Bidhan Chandra Krishi Viswavidyalaya**  
**P.O.- Krishi Viswavidyalaya, Nadia, West Bengal, 741252**

Organiser  
Society for Application of Statistics in Agriculture and Allied Sciences(SASAA)  
And  
Department of Agricultural Statistics  
Bidhan Chandra Krishi Viswavidyalaya  
Mohanpur-741252  
Nadia, West Bengal, India

National Workshop cum Training Programme  
on  
Statistical Tools for Research Data Analysis (Series-II)

**Compiled by**

Prof. A Majumder  
Prof. P. K. Sahu  
Prof. D Mazumdar  
Dr Mrs.B Bhattacharyya

***Edited by:***

Prof . A Majumder and Prof. P K Sahu  
Secretary, SASAA      Head, Dept. Ag. Statistics, BCKV

**Printed at**

Unimage  
10, Roy Bagan Street, Kolkata-700012  
Ph. : 2533 2956

**Published by**

AkiNik Publications  
C-11/169, Sector 3, Rohini, Delhi-110085  
ISBN: 978-93-85895-89-0

Content of the articles are the absolute opinion of the authors, neither the Society nor the Department of Agricultural Statistics is responsible for the same.



**Bidhan Chandra Krishi Viswavidyalaya**  
**P.O. Krishi Viswavidyalaya, Mohanpur 741252**  
**District: Nadia, West Bengal, India**  
**Website: www.bckv.edu.in**

(+91) 03473-222666  
(+91) 033-25879772  
Fax : (+91) (03473)-222275  
Email : ddpatra@rediffmail.com  
bckvvc@gmail.com  
Cell : +91-9830071278

**Dr. DD Patra,** Ph D (IARI, New Delhi), FNA Sc., FISSS, FNAAS  
Vice-Chancellor

No. VC/BCKV/114/57  
Date: 22.5.2017

### MESSAGE

It is a great pleasure to know that The Society for Application of Statistics in Agriculture and Allied Sciences (SASAA) in collaboration with The Department of Agricultural Statistics, Faculty of Agriculture, Bidhan Chandra Krishi Viswavidyalaya is going to organize a two weeks long National Workshop cum Training Programme on “Statistical Tools for Research Data Analysis (Series II)”, in the Department of Agricultural Statistics, Faculty of Agriculture, Bidhan Chandra Krishi Viswavidyalaya starting from 29<sup>th</sup> May, 2017.

Statistics provides scientific tools for planning of experiments or projects, collection of data from the experiment or for the project, analysis of data, and drawing valid inferences for recommendation or implementation. Therefore, the theme of the workshop chosen by the organizers is very much relevant to the researchers in agriculture and allied fields.

I hope this programme will benefit the scientific community, enrich and encourage it to use statistical tools in proper manner for the betterment of social welfare activities.

(D. D. Patra)



# WEST BENGAL UNIVERSITY OF ANIMAL AND FISHERY SCIENCES

68, KSHUDIRAM BOSE SARANI, BELGACHIA, KOLKATA-700 037, PHONE : 2556 3450, RESI. : (033) 2550 2229, FAX : 91-33-2557 1986  
web : [www.wbuafsci.ac.in](http://www.wbuafsci.ac.in), e-mail : [drpbiswas56@yahoo.com](mailto:drpbiswas56@yahoo.com)

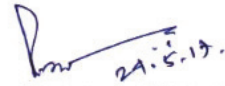
Prof. Purnendu Biswas, Ph.D.  
Vice-Chancellor

No. : VCSMWBUAFS/ M-5/274  
Date : 24.5.2017

## MESSAGE

I am indeed glad to know that the Society for Application of Statistics and Allied Sciences (SASAA) in collaboration with the Department of Agricultural Statistics, Bidhan Chandra Krishi Viswavidyalaya has taken initiative to organize a two weeks long National Workshop cum Training Programme on "Statistical Tools for Research Data Analysis (Series II)", starting from 29<sup>th</sup> May, 2017. Dependence of Indian Economy on Agriculture is well reflected by its contribution in the national GDP. Statisticians and the Agricultural Statistics being complementary to each other always play a significant role in this regard. Such a programme to make the Scientists, Teachers and other users familiar about the befitting use of statistics in their respective field is well conceived and useful one.

I wish the workshop cum training programme a grand success.

  
(Purnendu Biswas)

Prof. A Majumder  
General Secretary  
Society for Application of Statistics in  
Agriculture & Allied Sciences (SASAA)  
Deptt. of Agricultural Statistics  
Bidhan Chandra Krishi Viswavidyalaya  
Mohanpur, Nadia, West Bengal-741252  
E-mail: [secretarysasaa@gmail.com](mailto:secretarysasaa@gmail.com)



Bidhan Chandra Krishi Viswavidyalaya  
Faculty of Agriculture  
P. O. Krishi Viswavidyalaya, Dt. – Nadia  
West Bengal India; Pin-741252

**Prof. M. Pramanik**  
Dean

**E-Mail:** [deanofag.bckv@gmail.com](mailto:deanofag.bckv@gmail.com)  
**Website:** [www.bckv.edu.in](http://www.bckv.edu.in)  
**Ph. No. 033-25878338,**  
**03473-222656**  
**Fax- 03473-222273**

Ref. No. \_\_\_\_\_

Date: 24.05.2017

**Message**

I am happy to note that the Society for Application of Statistics in Agriculture and Allied Sciences (SASAA) in collaboration with the Department of Agricultural Statistics, Bidhan Chandra Krishi Viswavidyalaya is going to organize a two weeks long National Workshop cum Training Programme on “Statistical Tools for Research Data Analysis (Series II)”, at the Department of Agricultural Statistics, BCKV during 29<sup>th</sup> May to 9<sup>th</sup> June 2017.

I am really delighted to know that the society, SASAA, has taken many positive steps towards better use and appropriate applications statistical tools in greater scientific arena. I have been informed that the programme will be addressed by various luminaries in the field of statistics followed by hands on training to the participants.

I hope the challenging task taken up by a small group of statisticians from SASAA as well as from the Department of Agricultural Statistics will remain an example to the scientific community of the country.

I wish the workshop cum training programme a great success.

DEAN  
FACULTY OF AGRICULTURE

PROF. (DR) SATYABRATA PAL

Honorary Visiting Professor  
International Statistical Education Centre  
Indian Statistical Institute, Kolkata

Formerly, Senior Post-Doctoral Fellow  
International Rice Research Institute  
Los Banos, Manila, Pholippines

Formerly, Chairman and Member  
Finance Committee (Statutory)  
International Biometric Society, USA

Formerly  
Dean, Post-Graduate Studies & Professor of Statistics  
Bidhan Chandra Krishi Viswavidyalaya  
Nadia, West Bengal, Pin-741252

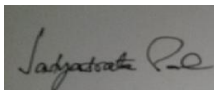
Principal  
Swami Vivekananda Institute of Management and  
Computer Science, Kolkata – 700104 &  
NSHM College of Management and Technology  
Durgapore, West Bengal  
Dean, Computer Applications, Regent Educational  
Research Foundation, Barrackpore, Kolkata

MESSAGE FROM THE DESK OF PRESIDENT

The Society for Applications of Statistics in Agriculture and Allied Sciences (SASAA) has stepped into its fourth year of existence holding with it a rich heritage of accomplishment of scientific activities in the national academic scenario in the discipline of Statistics, which has bequeathed on us a feeling of pride and at the same time a burden of responsibility in order to retaining the momentum in the future years with ardent zeal along with a mission to reach at a coveted place being certified by the academia. Needless to mention that the horizon of coverage of the modern scientific material in the published Papers in our Journal RASHI is ever-increasing with the successive issues and also the array of the topics being covered by the Resource Persons in this Workshop is, indeed, assuming more modernity in terms of research-development.

At the moment our prime objective staring us is to acquire the E-ISSN No. and Print No. for RASHI. In our latest endeavor, minor technical fault crept in our form submission, at the moment we have to proceed with all efforts instantly (as soon) to submit our revised application form. The days (in respect of jobs) before the Statisticians have assumed more competitive, now a days they have to face challenges from the Software professionals, Economists, Computer scientists and primarily, Mathematicians who have been in the foray much before the Statisticians have laid their foot-prints.

On this occasion of organisation of the National Workshop cum Training programme on “Statistical Tools for Research Data Analysis (Series – II)” of fifteen days duration (commencing 29<sup>th</sup> May, 2017), I greet the efforts of the members of the Society and most importantly, the Faculty and Staff of the Department of Agricultural Statistics and wish that the collective endeavour we have rendered will yield fruits in the sense of enriching the knowledge-base of the participants and, indeed, after having undergone the training they (the participants) will find themselves more proficient in analysing data in their respective fields. I convey my heart-felt thanks to the Resource persons for taking pains to come over to BCKV from their respective places to deliver the lectures addressed to the knowledge-enhancement of the participants and welcome all of them on this august occasion.



(SATYABRATA PAL)

President, SASAA

Elected Member, International Statistical Institute, Hague, Netherlands  
Fellow, Royal Statistical society, London, UK  
Fellow, Indian Association of Hydrologists, NIH, Roorkey, U.P., India  
Fellow, West Bengal Academy of Science and Technology, W.B., India  
Fellow, Inland Fisheries Society of India, CICFRI, Barrackpore, W.B., India



# **Society for Application of Statistics in Agriculture and Allied Sciences(SASAA)**



Registration No. S/2L No. 33489 of 2014-15

**Secretariat: DEPARTMENT OF AGRICULTURAL STATISTICS**

**Bidhan Chandra Krishi Viswavidyalaya**

**Mohanpur, Nadia, West Bengal, India – 741252**

**Ph. 03473 223256-9 Extn. 279, [www.sasaa.org](http://www.sasaa.org)**



The main objective of the Society for Application of Statistics in Agriculture and Allied Sciences (SASAA) is promotion of statistics to researchers of Agriculture and Allied Sciences. For fulfilling the above, the society is trying to encompass its activities to organize national level workshop cum training programmes, national level seminars, symposia, publication of journals etc.. The society had already organized one workshop cum training programme in 2015 as first major activity by organizing a national level workshop cum training programme on Statistical Tools for Research Data Analysis. After getting over helming responses from various parts of scientific community across the nation, we have decided to organize the similar type workshop cum training programme on Statistical Tools for Research Data Analysis at least once in two years interval. As a result, we are organizing the second workshop cum training programme on Statistical Tools for Research Data Analysis as series II, in collaboration with the Department of Agricultural Statistics, BCKV. Resource persons are different reputed academicians and statisticians from different nationally and internationally reputed institutes and universities have come forward to teach the trainees or participants of the present workshop cum training programme. We are really grateful to them. Participants from different universities and institutes across the nation have shown their interest in the programme. Again we are thankful to them. Unfortunately, due to some limitations, we cannot circulate the brochure of the workshop cum training programme well ahead of its commencement. We are thankful to Kribhco, Indofil industries and Nova Seeds for extending financial support for this programme otherwise it would have been very difficult to organize the programme.

The society must be thankful to Prof. D. D. patra, Hon'ble Vice Chancellor, Bidhan Chandra Krishi Viswavidyalaya, for keen interest, guidance and encouragement towards this society and also its present programme. We thankfully acknowledge the active cooperation and help of Prof. M. Pramanick, Dean, Faculty of Agriculture, Bidhan Chandra Krishi Viswavidyalaya. We also thankfully acknowledge the cooperation and help of Mr. G. Pal, Registrar, BCKV, Prof. J. K. Hore, Dean, Faculty of Horticulture, Prof. S Mukherjee, Dean, Faculty of Ag. Engineering, Prof. R. K. Biswas, Dean P.G. Studies, BCKV, Prof. S. Pal, Director of Research, Prof. K. Brahmachari, Director of Extension Education, Dr. S. Mitra, Director of Farms and Heads of all the departments of BCKV.

The society thankfully acknowledges the cooperation and help received from the teachers, staff members, research scholars and students of the Department of Agricultural Statistics, BCKV towards organization of this event.

A. Majumder

Secretary, SASAA

Date: 29.05.2017

---

**Contact: Email: [secretarysasaa@gmail.com](mailto:secretarysasaa@gmail.com); Ph: 91 9433841687**

# Bidhan Chandra Krishi Viswavidyalaya

## DEPARTMENT OF AGRICULTURAL STATISTICS

Mohanpur, Nadia, West Bengal, India – 741252

Ph. 03473 223256-9 Extn. 279, FAX 03473 -222273

*DR. P. K. SAHU*

*Professor & Head*



Residence : B2/258, Kalyani, Nadia

West Bengal, India-741235

Ph: 91-033- 25823552(R)

+91 9433841687(m)

Email: pksbckv@gmail.com

### National Workshop cum Training Programme on Statistical Tools for Research Data Analysis (Series II)

Now – a – days one can hardly find a discipline of Science in which Statistics has not been used. Data and, or information are the ingredient of statistical sciences. To make these data speak for itself, one needs to process and put these under various statistical treatments. Statistical tools have the unique characteristic of unearthing the otherwise hidden information from a given set of data. Statistical sciences also have the power of generalizing the information or conclusion resulted from a particular structured/non structured/survey/experimental data pertaining to life science, social science, behavioral science, medical science and other fields of modern sciences.

To make the scientists/teachers and practitioners in different fields of modern sciences aware, the **Society for Application of Statistics in Agriculture and Allied Sciences (SASAA)** in collaboration with **Department of Agricultural Statistics**, Bidhan Chandra Krishi Viswavidyalaya, Mohanpur is organizing a two week long Workshop cum Training Programme, 2<sup>nd</sup> time with in a span of two years, at the Department of Agricultural Statistics, Bidhan Chandra Krishi Viswavidyalaya, Mohanpur. Participants from different states, different state and central institutes have come to join the programme. Eminent statisticians from University of Calcutta, Indian Statistical Institute (ISI), Kolkata; Indian Institute of Science and Education Research (IISER), Kolkata; University of Kalyani; Indian Council of Agricultural Research (ICAR) and Bidhan Chandra Krishi Viswavidyalaya are expected to take part in the programme. Main characteristic feature of the programme would be hands on training with and, or without use of statistical softwares preceded by the theoretical aspects in diverse fields of study. Data management and extraction of information could be discussed in the programme followed by hands on training to the participants of different modern scientific disciplines.

Helps, co-operation and guidance received from different parts of the University Administration under the leadership of our Hon'ble Vice Chancellor, Prof D D Patra, remains to be specially mentioned. Active financial support from ICAR and generosity of the University administration could make it possible to develop such a beautiful laboratory where the training programme could take place. Though a very few in number, the help and support rendered by the teachers, supporting staff and students of the department, once again remind us that "Where there is will there is way!".

Sincere and dedicated efforts are there from all corners to make this program a successful one; even then we solicit suggestions and guidance from each and every one for its further improvement.

Hope with active support from every corner the department of Agricultural Statistics and SASAA will be able to organize, even better than this type of programme in future.

Mohanpur

P K Sahu

Date: 25-5-2017

## CONTENT

Title	Author	Page No.
Technical Programme		1
Modelling in Agriculture and Allied Sciences	S. P. Mukherjee	2
Clustering and Classification: some computational issues	Asis Kumar Chattopadhyay	4
Some Advances in Survey Sampling Theory & Practice	Arijit Chaudhuri	9
Design and analysis of Factorial Experiments	G. M. Saha	11
Statistical methods for Spatital Data	Satyabrata Pal	23
Fuzzy Liner Regression using SAS software	G Sathish, Minakshi Mishra and Prof. D. Mazumdar	24
Business Intelligent in Cloud Computing	Manas Kumar Sanyal	36
Abstract of presentation of Some Work on Non-Sampling Errors	Prof. Pulakesh Maiti	37
Misuses of Statistics	Prof. Shantiranjana Pal	39
Cluster Analysis	Prof. Pradip Kumar Sahu	45
Association between the set of Macro and Micro climatic parameters with the set of Crop growth parameters of wheat crop following “Monte Carlo simulation” Technique: Redundancy Analysis (RDA)	G Sathish, Minakshi Mishra and Prof. D. Mazumdar	64
Some Problem-aspects associated with Regression Analysis	Kiranmoy Das	65
Linear regression using SAS	Dr. (Mrs) B.Bhattacharyya	70
Analysis of variance and Basic Experimental Designs	Anurup Majumder	83
Artificial Neural Network and its applications	G Sathish, Pavana Kumar S.T. and Prof. D. Mazumdar	104
Non-parametric Test	Ajit Kumar Das	109
Use of Auxiliary Information in Sample surveys	Ajit Kumar Das	117
Regression Analysis and its Application	Prof. P K Sahu	121

Missing Plot Technique	Prof. Anurup Majumder	156
Basics of Regression and PCA	Asok K. Nanda	162
Probability: How to Model?	Asok K. Nanda	170
Statistical Assessment of Agreement in Treatment Effects Comparisons	Bikas K Sinha	207
Soil Heterogeneity and its treatment — Optimum Plot Size Determination	Bikas K Sinha	222
Time Series Modeling- an overview	Dr K K Goswami	243
Analysis of Covariance	Premadhis Das	262
Principal Component Analysis (PCA)	A K Nanda	268
Concepts & Methodological aspects in Farm Business Analysis	Dr. A K Nandi	291
Split-Plot and Strip Plot Designs	R N Panda	323
Multivariate Analysis	Asish Kr Chottapadhyay	330
Sample Size Determination	Bikas K Sinha	352
Application of Multiple Criteria Decision Making Approach in Agriculture	Prof. Anurup Majumder	364

## National Workshop cum Training programme on “Statistical Tool for Research Data Analysis (Series II)”

Starting from 29<sup>th</sup> May, 2017, Duration two weeks

Organised by Society for Application of Statistics and Allied Sciences (SASAA), in collaboration with

Department of Agricultural Statistics, F/ Agriculture, BCKV

## Technical Programme

29.05.2017	11:00 - 12:15pm	12:15 - 12:30pm	12:30 - 1:30 pm	1:30 - 2:30 pm	2:30 - 3:30 pm	3:30 - 4:30 pm
Inaugural Day	Inauguration	Tea Break	Lecture by Chief Guest, Prof SP Mukhopadhyay	Lunch	Lecture on ‘Some Advances in Survey Sampling Theory & Practice’ by Prof. A. Choudhuri, ISI, Kol.	Lecture on ‘Statistical Assessment of Agreement in Treatment Effects Comparisons’ by Prof. BK Sinha, Former Advisor, Prime Minister, GOI, ISI, Kol.

## Technical Session:

Time/Day	10:45 - 11:45am	11:45- 12:00noon	12:00noon-1:00pm	1:00 - 2:00 pm	2:00 - 3:00 pm	3:00 - 4:45 pm
30.05.2017	Lec. 1 (AKD, BCKV)	Tea	Lec. 2(KKG, ICAR)	Lunch	Lec.3(AKD, BCKV)	Practical
31.05.2017	Lec. 4(SRP, BCKV)	Tea	Lec. 5(PKS, BCKV)	Lunch	Lec.6(AM, BCKV)	Practical
01.06.2017	Lec. 7(PM, ISI)	Tea	Lec. 8(RMP, BCKV)	Lunch	Lec.9(BB, BCKV)	Practical
02.06.2017	Lec. 10(AN, IISER)	Tea	Lec. 11(AM, BCKV)	Lunch	Lec.12(BB, BCKV)	Practical
05.06.2017	Lec. 13(SP, ISI)	Tea	Lec. 14(KD, ISI)	Lunch	Lec.15(DM, BCKV)	Practical
06.06.2017	Lec. 16(GMS, ISI)	Tea	Lec. 17(BKS, ISI)	Lunch	Lec.18(PD, KU)	Practical
07.06.2017	Lec. 19(AN, BCKV)	Tea	Lec. 20(RNP, KU)	Lunch	Lec.21(MS, KU)	Practical
08.06.2017	Lec. 22(AC, CU)	Tea	Lec. 23(DM, BCKV)	Lunch	Lec.24(DM, BCKV)	Practical
09.06.2017	Valedictory Session	Lunch				

03.06. 2017 (Saturday) and 04.06. 2017 (Sunday) are scheduled for visit to Places of Academic interest.

## List of Resource persons:

PD - Prof. P. Das, KU	RNP – Prof. R. N. Panda, KU	AKD – Prof. A. K. Das, BCKV	DM – Prof. D. Mazumdar, BCKV
KD – Prof. K. Das, ISI, Kol	PM – Prof. P. Maity, ISI, Kol	SP – Prof. S. Pal, ISI, Kol	KKG – Dr. K. K. Goswami, ICAR
AN – Prof. A. Nanda, IISER, Kol	AM – Prof. A. Majumder, BCKV	SRP – Prof. SR. Pal, BCKV	BB – Dr. B. Bhattacharyya, BCKV
GMS – Prof. G. M. Saha, ISI, Kol	BKS – Prof. B. K. Sinha, ISI, Kol	NM – Prof. N. Mondal, CU	AN – Prof. A. Nandi, BCKV
PKS – Prof. P. K. Sahu, BCKV	AC – Prof. A. Chattopadhyay, CU	MS – Prof. M. Sanyal, KU	RMP – Prof. RM Panda, BCKV

N.B.- Programmes are subjected to last minute change.

## Modelling in Agriculture and Allied Sciences

S.P.Mukherjee

Models are simplified (occasionally simplistic) representations of Reality. Models are used to facilitate studies on real-life systems, processes and phenomena. Some models are meant to provide a useful description of some such observable phenomenon, some others are meant to relate some phenomena with certain other phenomena and provide explanations for the phenomena of interest, while some models are meant to control a phenomenon in terms of its outcome kept at some desired level or within some tolerable limits, and still others which attempt to optimize the outcome of some phenomenon. Descriptive and explanatory models are also used for prediction purposes.

Modelling is an exercise that involves model development (or selection), model testing for relevance and adequacy, and model-solving. And this exercise is an essential component of any scientific study in any branch of human enquiry.

The phenomenon universally observed is “variation” across individuals, time, and space. And in many cases, such variations are caused by some uncontrollable chance factors over and above some controllable assignable factors affecting the output of a system or the outcome of a process, rendering the latter ‘unpredictable’. Thus, variations in weather conditions across days in a crop season are random, as are variations in yield from one plant to another of the same variety receiving the same treatment.

To represent such random variations in some variable(s) as revealed in a sample of observations, we make use of probability models through which we pass from the particular sample to reach some conclusion relating to the phenomenon (population or ensemble) itself. From the existing huge stock of probability models, we can select one that fits our observations best and facilitates inductive inferences. A choice among alternatives can be based on some criterion—usually a penalized likelihood of the sample observations, assuming a particular model as true—which is to be obtained from the sample data. Otherwise, we have to develop a model to incorporate special features of the observed data and of the underlying phenomenon. In fact, many new models have been added to the current kit of models, e.g. exponentiated probability models or skewed normal distribution or generalized exponential or gamma distribution and a host of similar others.

Growth (including decay and decline) is the most common phenomenon we come across in many areas of investigation. Growth of a population of living beings including micro-organisms or of inanimate entities like land areas put under different crops requires quantification and explanation. Thus the growth of a human population along with growth in urbanization and industrialization can provide an explanation of growth in the proportion of land area put under non-agricultural purposes. And the consequence of the latter could be linked to changes in agricultural practices meant to increase productivity of land. Changes in cropping pattern adopted by the farming community can be linked to increase in Gross Value Added per unit area mandated by a change in living conditions of farmers.

Whenever we speak of changes—essentially over time—we deal with dynamic models, as opposed to static ones which are focused on an existing situation only. Most often differential or differential-difference equations—a single equation or a set—are used as models with parameters that may be constants or time-dependent or even random. Thus the logistic law of growth, which has undergone many modifications and extensions, was derived from a simple differential equation

$$\frac{1}{P(t)} \frac{dP(t)}{dt} = [1 - kP(t)]$$
 the L.H.S corresponds to the relative rate of growth at time  $t$  with  $P(t)$  denoting population size at time  $t$  and  $k$  is a constant that is related to the limiting population size. The solution that can be used to predict the population at a future time  $t$  is given by

$$P(t) = L / [1 + \exp \{r(\beta - t)\}]$$

Some models are deterministic, while others are probabilistic depending on whether some of the input or outcome variables or variables that constrain the process or the phenomenon are random (affected by some

uncontrollable chance factors besides controllable assignable factors and hence not exactly predictable ) or not... Stochastic process models are being used currently to track growth under uncertainty situations.

Recognising the reality that population at any future point of time  $t$  depends on many factors including some which cannot be exactly predicted or even cannot be explicitly identified, we try to develop probabilistic models for this purpose. The output in such a model is a probability distribution for a point of time  $t$  given by  $P_n(t) = \Pr \{ N(t)=n \}$   $n= 0,1,2,3,\dots$  and  $t \geq 0$ . Assuming that population growth is determined by the birth and death rates ( which can be interpreted as the probability that a birth takes place in a small time interval say  $(t, t+ \Delta t)$  and similarly a death occurs in such a small interval), we can write out an infinite system of probabilistic differential- difference equations as

In the large category of explanatory models, we have the regression models and structural relations models. Regression models depict dependence relations connecting different phenomena and the associated variables. Essentially meant to relate phenomena taking place concurrently or to relate one current phenomenon with another that happened earlier, regression models are widely used in all scientific studies. We have a large variety of regression models depending on the nature of the variables involved and the nature of relations connecting them. Somewhat different are structural equations models where a variable that plays the role of an independent regressor or explanatory variable in one equation can become the dependent variable in another equation. Usually, these are linear relations that connect some exogenous with some endogenous variables. Some of the variables could be latent, besides the usual manifest variables. Models representing socio-economic phenomena which are not uncommon in Agriculture belong to this category. A lot remains to be done in estimating regression parameters subject to some constraints imposed on them like in a linear regression involving two independent variables  $X_1$  and  $X_2$  affecting the dependent variable  $Y$  in the form

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + e \quad \text{with } e \text{ as the random error}$$

it may be natural to require that  $\beta_1 > \beta_2$  or the other way round or, simply that  $\beta_1 > 0$ .

In such cases, we have to take recourse to constrained optimization of the likelihood or the sum of squared residuals.

Models to control and, subsequently, to optimize a system output or the response / outcome of a process are generally branded as ' optimum-seeking models' and are well illustrated by response surface methodologies in the context of multi-factor agricultural experiments to indicate the optimum treatment combination. It is possible that we eventually end up with a near-optimal or a satisfactory solution to our search for an optimum. In fact, a response surface model is just a regression model or a set of regression models one each for a single response variable. The solution is rather complicated and does not admit of a unique approach.

Models representing observed phenomena linked to farming and the farmers that go beyond the traditional models for physical aspects of farming to account for human behaviour aspects that are influenced by many external human interventions are assuming greater and greater significance. Appropriate models have to be developed—may be through necessary modifications—have to be developed and solved towards this. Models to explain changes in land use for different purposes—agricultural, residential, industrial, recreational, commercial etc.—as also in use of cultivable land to be put under different crops or crop rotations have been sometimes more qualitative but provide deep insights into social, cultural, political and economic systems and their operations.

## Clustering and Classification: some computational issues

Asis Kumar Chattopadhyay  
Department of Statistics, Calcutta University

### Abstract

In the present work attempts have been made to highlight different computational problems related to clustering, classification and dimension reduction techniques depending on input data type and underlying model assumptions of the different statistical methods. The effect of directional data has been considered on the basis of a real environmental data set.

### 1. Introduction

Cluster Analysis, also called data segmentation, has a variety of goals. All relate to grouping or segmenting a collection of objects (also called observations, individuals, cases, or data rows) into subsets or "clusters", such that those within each cluster are more closely related to one another than objects assigned to different clusters. Central to all of the goals of cluster analysis is the notion of degree of similarity (or dissimilarity) between the individual objects being clustered. There are two major methods of clustering - hierarchical clustering and k-means clustering.

Statistical techniques for classification are essentially of two types. Members of the first type are used to construct a sensible and informative classification of an initially unclassified set of data; these are known as cluster analysis methods. The information on which the derived classification is based is generally a set of variable values recorded for each object or individual in the investigation, and clusters are constructed so that individuals within clusters are similar with respect to their variable values and different from individuals in other clusters. The second set of statistical techniques concerned with classification is known as discriminant or assignment methods. Here the classification scheme is known a priori and the problem is how to devise rules for allocating unclassified individuals to one or other of the known classes.

Different Statistical techniques are available for clustering and classification (Fraix Burnet D et al (2015), De T et al (2013) and references there in). But depending on the nature of the different types of data the following problems often arise and in some cases a proper solution is still not available.

1. Sometimes the data set under consideration has a distributional form (usually normal) and sometimes it is of non normal nature. Based on the above point, there is a justification needed about which clustering or classification technique should be used so that it reflects the proper nature of the data set provided. This problem is more relevant for classification as most of the classification methods are model based. For clustering most of the methods are non parametric in nature and as such the above problem is not very serious. But here also basic assumption is that the nature of the variables under study are continuous where as under practical situations these may be categorical like binary, nominal, ordinal and even directional (particularly for environmental and Astronomical data). Under such situations standard similarity/dissimilarity measures will not work.
2. The clustering techniques which require an inherent model assumption are known as Model Based Methods, whereas the clustering technique where no modelling assumption or distributional form is needed may be termed as Non-Model based Methods. Hence based on the nature of data set, one has to decide about proper application of the two types of techniques.
3. Even if one decides about the proper methods for the data set at hand, there are several techniques available under both the categories and no pre defined criteria can be set to judge which technique is the best for the situation under consideration.



4. The above point arises the need of a comparative study among various available techniques and a computational analysis of all the methods. Once all the methods are implemented, it requires a criterion to decide upon the best technique based on a post classifier. So an appropriate post classification approach is also needed in this regard. For a post classification approach, a pre-classifier or training sample is required. Since in this type of techniques a prior knowledge of classification is provided, these are called Supervised Learning. All other techniques where no prior classification is provided are known as Unsupervised Learning.
5. A comparative validity algorithm may be helpful for predicting the superiority of different techniques.
6. At present big data issues related to data size is quite common. In statistical terms this problems may be tackled in terms of both the number of observations and the variables considered. Many standard clustering techniques fails to deal with such big data sets. Thus some dimension reduction methods may be applied at first and then clustering may be performed on the reduced data set. Some data mining techniques are very helpful under such situations.
9. The above criteria also needs to be validated depending on whether the data is Gaussian or non-Gaussian. That means the dimension reduction techniques may vary according as the data set has a distributional form or not.
10. Finally and most importantly after all these considerations, the similarity of grouping of objects obtained from different methods should be checked in terms of some physical properties.

## 2. Hierarchical Clustering Technique

Central to all of the goals of cluster analysis is the notion of degree of similarity (or dissimilarity) between the individual objects being clustered. There are two major methods of clustering -hierarchical clustering and k-means clustering. In hierarchical clustering the data are not partitioned into a particular cluster in a single step. Instead, a series of partitions takes place, which may run from a single cluster containing all objects to  $n$  clusters each containing a single object. Hierarchical Clustering is subdivided into agglomerative methods, which proceed by series of fusions of the  $n$  objects into groups, and divisive methods, which separate  $n$  objects successively into finer groups. Agglomerative techniques are more commonly used. Hierarchical clustering may be represented by a two dimensional diagram known as dendrogram which illustrates the fusions or divisions made at each successive stage of analysis.

### 2.1. Agglomerative methods:

An agglomerative hierarchical clustering procedure produces a series of partitions of the data,  $C_n, C_{n-1}, \dots, C_1$ . The first  $C_n$  consists of  $n$  single object 'clusters', the last  $C_1$ , consists of single group containing all  $n$  cases.

At each particular stage the method joins together the two clusters which are closest together (most similar). (At the first stage, of course, this amounts to joining together the two objects that are closest together, since at the initial stage each cluster has one object.) Differences between methods arise because of the different ways of defining distance (or similarity) between clusters.

A key step in a hierarchical clustering is to select a distance measure. A simple measure is Manhattan distance, equal to the sum of absolute distances for each variable. The name comes from the fact that in a two-variable case, the variables can be plotted on a grid that can be compared to city streets, and the distance between two points is the number of blocks a person would walk.

A more common measure is Euclidean distance, computed by finding the square of the distance between each variable, summing the squares, and finding the square root of that sum. In the two-variable case, the distance is analogous to finding the length of the hypotenuse in a triangle; that is, it is the distance as the crow flies. A

review of cluster analysis in health psychology research found that the most common distance measure in published studies in that research area is the Euclidean distance or the squared Euclidean distance.

To calculate distance between two clusters it is required to define two representative points from the two clusters. Different linkage measures like "single linkage", "complete linkage", "average linkage" etc have been proposed for this purpose.

## 2.2. Similarity for any type of data

The above mentioned dissimilarity/similarity measures are applicable to continuous type data only. But generally we work with mixed type data sets which includes different types like continuous, discrete, binary, nominal, ordinal etc. Gower J.C.(1971) has proposed a general measure as follows:

The Gower's Coefficient of Similarity:

Two individuals  $i$  and  $j$  may be compared on a character  $k$  and assigned a score  $s_{ijk}$ . There are many ways of calculating  $s_{ijk}$ , some of which are described below.

Gower's similarity index  $S_{ij}$  is defined as

$$S_{ij} = \frac{\sum_1^k s_{ijk}}{\sum_1^k \delta_{ijk}}$$

where  $\delta_{ijk} = 1$  if when character  $k$  can be compared for observations  $i$  and  $j$   
 $= 0$  otherwise

For continuous (quantitative) variables with values  $x_{1k}, x_{2k}, \dots, x_{nk}$  for the  $k$ th variable

$$s_{ijk} = 1 - |x_{ik} - x_{jk}| / R_k$$

where  $R_k$  is the range of the variable  $k$  and may be the total range in population or the range in the sample..

For a categorical (qualitative) character with  $m$  categories ( $m=2$  for binary variable)

$s_{ijk} = 0$  if  $i$  and  $j$  are totally different  
 $= p$  (positive fraction) if there is some degree of agreement  
 $= 1$  when  $i$  and  $j$  are same

## 3. Dimension Reduction

When the data set is large ( in terms of number of variables ) one may first apply some appropriate dimension reduction technique and then perform clustering on the reduced data set. One must keep in mind that the discriminant usefulness of distances is lost in high dimension parameter spaces since distances tend to become similar (one of the aspects of the "curse of dimensionality").

### 3.1 Principal Component Analysis (PCA)

In this technique, given a data set of observations on correlated variables, an orthogonal transformation is performed to convert it into a set of uncorrelated variables called the principal components. The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance. One rule of thumb is to consider those components whose variances are greater than one in the reduced space. Principal components are guaranteed to be independent only if the variables are jointly normally distributed.

### 3.2 Independent Component Analysis (ICA)

One of the most recent powerful statistical techniques for analyzing large data sets is independent component analysis (ICA), see Comon (1994) for the original description of ICA. Such data sets are generally multivariate in nature. The common problem is to find a suitable representation of the multivariate data. For the sake of computational and conceptual simplicity such representation is sought as a linear transformation of the original data. Principal component analysis, factor analysis, projection pursuit are some popular methods for linear transformation. But ICA is different from other methods, because it looks for the components in the representation that are both statistically independent and non Gaussian. In essence, ICA separates statistically independent component data, which is the original source data, from an observed set of data mixtures. All information in the multivariate data sets are not equally important. We need to extract the most useful information. Independent component analysis extracts and reveals useful hidden factors from the whole datasets. ICA defines a generative model for the observed multivariate data, which is typically given as a large database of samples. See Hyvarinen et al. (2001), Comon and Jutten (2010) and Lee (1998) for booklength discussions on ICA. ICA can be applied in various fields like neural network (Fiori 2003), studying EEG data (Bartlett et al. 1995), speech processing (Kumaran et al. 2005), brain imaging (McKeown et al. 1997), signal separation (Adali et al. 2009), telecommunications (Hyvarinen et al. 2002), econometrics (Bonhomme and Robin 2009), etc. Chattopadhyay et al(2015) has applied ICA for environmental pollution data yet.

## 4. Environmental Data

We consider a data set containing measurements collected from flights conducted in June and July 2002 over the Walnut Creek watershed in central Iowa, USA. The study was part of the Soil Moisture Experiment 2002 (SMEX02) and the Soil Moisture Atmosphere Coupling Experiment (SMACEX), run by Canada's National Research Council (NRC). See SMEX02 Soil Moisture Atmosphere Coupling Experiment, Iowa, <http://nsidc.org/data/nsidc-0232.html>, for a detailed description of the methodology and the data.

The aircraft carried numerous sensors and flew several flights per day along six tracks in the watershed area. These data were collected as part of a validation study for the Advanced Microwave Scanning Radiometer—Earth Observing System (AMSRE). AMSRE is a mission instrument launched aboard NASA's Aqua Satellite on 04 May 2002. AMSR-E validation studies linked to SMEX are designed to evaluate the accuracy of AMSR-E soil moisture data. Specific validation objectives include assessing and refining soil moisture algorithm performance; verifying soil moisture estimation accuracy; investigating the effects of vegetation, surface temperature, topography, and soil texture on soil moisture accuracy; and determining the regions that are useful for AMSR-E soil moisture measurements.

Variables for this data set include air temperature (TEMP, in °C), dew point temperature (DEWP, in °C), radiometric surface temperature (KT19, in °C), greenness index (GRN, ratio of 730/660 nm reflected radiation), net radiation from wingtip sensor (NETRD, in W/m<sup>2</sup>), CO<sub>2</sub> concentration (in ppm), wind direction (in ° true), wind speed (in m/s), sensible heat flux (H, in W/m<sup>2</sup>), latent heat flux (LE, in W/m<sup>2</sup>), CO<sub>2</sub> flux (WC, in mg/m<sup>2</sup>/s), friction velocity computed from momentum flux ( $U^*$ , in m/s) and ozone flux, corrected (in mg/m<sup>2</sup>/s). Here wind direction is a directional variable and the remaining set is continuous.

### 4.1. Conversion of directional data to linear

Note that PCA or ICA has been developed for linear continuous data but here one variable, viz. wind direction, is circular in nature. This is an important covariate as it is believed that quite often wind brings pollutants from neighbouring places to any particular place under consideration, especially if there are some industrial regions nearby. But it is not immediate how to include this type of data for classification directly or through PCA or ICA. A density plot of this data clearly shows the wind direction has a bimodal distribution, the two modes are near 0° and 200°. Thus we are motivated to consider two main directions, east and west (approximately), which correspond to 0° and 180°.

Chattopadhyay et al (2015) proposed a method of conversion from circular to linear where they considered standard cosine angular distance of an angle  $\theta$  from a fixed angle  $\phi$ , defined by  $d\phi = 1 - \cos(\theta - \phi)$ , which is in the linear scale, and  $d \in [0, 2]$ . Thus, for a wind direction of  $\theta$ , we may consider two distances  $d_0 = 1 - \cos(\theta - 0^\circ)$  and  $d_{180} = 1 - \cos(\theta - 180^\circ)$ , both of which are linear. So, instead of taking  $\theta$  in our analysis, we consider the pair  $(d_{\max}, d_{\text{sign}})$ , where  $d_{\max} = \max(d_0, d_{180})$  and  $d_{\text{sign}} = +1$  if  $d_{\max} = d_0$  and  $d_{\text{sign}} = -1$  if  $d_{\max} = d_{180}$ . Alternately, if we want to ignore the sign we can work with  $\theta^* = 2 \times \theta$ , which is approximately unimodal with mode near  $45^\circ$ . We may work with  $d^* = 1 - \cos(\theta^* - 45^\circ)$ .

## 5. Conclusion

From the above discussions it is very clear that although clustering and classification problems are widely used under different disciplines by scientists from several areas, one should always take care of the nature of data in order to apply the methods successfully. In the introduction we have listed several such problems and only a few are discussed in latter sections. It is quite expected that one may identify many other computation based problems which are not listed here.

## References:

- Adali T, Jutten C, Romano JMT, Barros AK (2009) Independent component analysis and signal separation. In: Proceedings of 8th international conference, ICA 2009, Paraty, Brazil, March 15–18, 2009. Springer, Berlin
- Bonhomme S, Robin J-M (2009) Consistent noisy independent component analysis. *J Econ* 149:12–25
- Chattopadhyay A.K. Mondal S and Biswas A (2015), Independent component analysis and clustering for pollution data, *Environmental and Ecological Statistics*, 22, 33-43
- Comon P (1994) Independent component analysis, a new concept? *Signal Process* 36:287–314
- Comon P, Jutten C (2010) *Handbook of blind source separation, independent component analysis and applications*. Academic Press, Oxford, UK
- De, T., Chattopadhyay, T., and Chattopadhyay, A. K. (2013). Comparison among clustering and classification techniques on the basis of galaxy data. *Calcutta Stat. Assoc. Bull.* 65, 257–260.
- Fiori S (2003) Overview of independent component analysis technique with an application to synthetic aperture radar (sar) imagery processing. *Neural Networks* 16(3–4):453–467
- Fraix-Burnet D., Thuillard M and Chattopadhyay A.K.(2015), Multivariate approaches to classification in extragalactic astronomy, *Frontiers in Astronomy and Space Science*, 2, 1-17
- Gower J.C. (1971) A General Coefficient of Similarity and Some of Its Properties. *Biometrics*, Vol. 27, No. 4. (Dec., 1971), pp. 857-871
- Hyvärinen A, Karhunen J, Oja E (2001) *Independent component analysis*. Wiley, New York
- Hyvärinen A, Karhunen J, Oja E (2002) *Telecommunications*. In: *Independent component analysis*, Ch 2. Wiley, New York. doi:10.1002/0471221317
- Kumaran RS, Narayanan K, Gowdy JN (2005) Myoelectric signals for multimodal speech recognition. *INTERSPEECH* 2005:1189–1192
- Lee T-W (1998) *Independent component analysis: theory and applications*. Kluwer, Boston, MA
- McKeown MJ, Makeig S, Jung T-P, Brown GG, Kindermann SS, Sejnowski TJ (1997) Analysis of fmri data by decomposition into independent components. *Am Acad Neurol Abstr* 48:A417

## Some Advances in Survey Sampling Theory & Practice

Arijit Chaudhuri

D.B. Lahiri, the famous Survey Statistician and Sampling expert, a close collaborator of Professor P.C. Mahalanobis once wrote that to an intelligent person but not expert in Survey Sampling techniques it is not easy to communicate an explanation as to why the same value observed for the sample mean based on a (1) Simple random sample taken without replacement, (2) Simple random sample taken with replacement a same number of draws as in (1) and (3) Simple random sample taken with replacement a number of times with an average equal to the sample-size same as in (1) should each be unbiased for the population mean but with different levels in accuracy.

The same D.B. Lahiri gave us a method of sampling a unit from a population with a probability of selection proportional to size-measure of the unit (called PPS) in a much simpler way than a prevailing selection procedure. He also gave a method of sample-selection with a probability proportional to the sample-sum of the size-measures showing that the ratio estimator based on the sample is exactly unbiased for the population total. An elegant formula for the exact variance of this Lahiri ratio estimator along with an unbiased estimator for this is available on adopting Hajek's theory reformulated by JNK Rao. Lahiri also introduced the systematic sampling procedure with selection probabilities proportional to size-measures. A systematic sample with a single draw does not yield an unbiased variance estimator for a linear estimator for a population total. The reason is that every such sample has a fixed number separating the starting point in the draws and as a consequence every pair of units in the population thereby cannot have a positive inclusion-probability. Chaudhuri and Pal have introduced a modification allowing a random variable separating the respective starting draws leading to positive-valued inclusion-probabilities of every paired unit and hence exactly unbiased variance estimator from a single modified systematic sample. Chaudhuri and Pal have derived a generalization of Yates & Grundy formula for the variance of Horvitz-Thompson estimator for a population total and its unbiased estimator pointing out a condition for the uniform non-negativity of the latter.

Chaudhuri has modified numerous existing randomized response techniques to show their applicability to general methods of sample selection liberating from their exclusive dependence on simple random sample selection with replacement. He developed optional randomization response theory covering qualitative as well as quantitative variables vitiated by social stigma. He permits (1) a respondent to respond either directly or following a randomized device or (2) respond directly or randomly without divulging which alternative is actually followed. He has further given general procedures to examine how far protection is afforded to a respondent agreeing to give out truthful randomized responses. He has further examined Bayesian methods in the contexts of randomized response techniques (RRT). Usually RRT's involve Bernoullian trials while implementing RR devices but if inverse Bernoullian trials are tried, there is possibility for improved accuracy. To examine this with certain classical RR devices problems are encountered and so necessary amendments in the original RR devices also Chaudhuri has introduced.

Chaudhuri (2012) also examined small area estimation procedures showing the role of modelling in survey sampling.

Chaudhuri has given techniques of Adaptive Sampling and Network Sampling as devices to increase information contents of survey data when a sample contains inadequate substance of relevance when the sampling design is quite general.

In finite population sampling Taylor series expansion is a convenient method to study efficacies of estimated correlation coefficients. But this approach fails in utilizing Spearman's rank correlation coefficient. However Chaudhuri has demonstrated how Kendall's correlation coefficient TAU may be studied in respect of its efficacies by Taylor's expansion method in case of general sample selection methods.

Chaudhuri has further demonstrated how multi-stage sampling is quite efficacious in auditing large-scale sample survey data.

*References :*

- *Chaudhuri, Arijit (2010). Essentials of survey sampling. Prentice Hall of India, Delhi*
- *Chaudhuri, Arijit (2011). Randomized response and indirect questioning techniques in surveys. Taylor & Francis, CRC Press. Chapman & Hall.*
- *Chaudhuri, Arijit (2012). Developing small do0mainstatisticsLAP, Germany.*
- *Chaudhuri, Arijit & Christofides, T.C. (2013). Indirect questioning in sample surveys. Springer Verlag Heidel burg*
- *Chaudhuri, Arijit (2014). Modern survey Sampling . Taylor & Francis, CRC Press. Chapman & Hall.*
- *Chaudhuri, Arijit (2015). Network and adaptive Sampling. Taylor & Francis. CRC Press, Chapman & Hall*

# DESIGN AND ANALYSIS OF FACTORIAL EXPERIMENTS

G.M.Saha  
Indian Statistical Institute  
Kolkata 700108

## 1. INTRODUCTION

An experiment in which two or more factors, each at two or more levels are studied simultaneously with the primary objective of estimating and testing various main effects and interactions, is called a **factorial experiment**, or a **multi-factor experiment**. When all the factors have the same number of levels, it is called a **symmetrical** factorial experiment. Otherwise, it is an **asymmetrical** factorial experiment. If there are  $n$  factors, say,  $F_1, F_2, \dots, F_n$  at  $s_1, s_2, \dots, s_n$  levels respectively, then we call it a  $s_1 \times s_2 \times \dots \times s_n$  factorial. In particular, when  $s_1 = s_2 = \dots = s_n = s$ , it is an  $s^n$  factorial experiment. We shall confine ourselves in these lectures to only  $2^n$  ( $n=2, 3, 4, 5$ ) and  $3^n$  ( $n=2, 3$ ) experiments.

Apart from the objective of examining interactions, the advantage of economising on experimental resources also suggest that one should prefer factorial experiments to several single factor experiments. For example, if there are 2 factors each at 2 levels, and if we want to compare treatment effects using a randomized block design with error based on 12 d.f., we need 5 replications of  $2^2$  treatment combinations requiring  $5 \times 4 = 20$  experimental units, while in two single factor experiments with similar precision, we shall need  $13 \times 2 + 13 \times 2 = 52$  experimental units. Thus the factorial experiment saves a lot of resources in comparison to the corresponding two separate single factor experiments.

Depending on the heterogeneity of the experimental units involved, we decide upon an appropriate design of experiment, like CRD, or RBD, or LSD, or even split-plot or strip-plot design, and then try to estimate and test various treatment contrasts of interest, like main effects and interactions from the resulting data.

## 2. MAIN EFFECTS & INTERACTIONS ( $2^2$ Experiment)

Let A and B be two factors, each at two levels, say  $a_0, a_1$  and  $b_0, b_1$  respectively. Thus this  $2^2$  experiment has four level combinations or treatment combinations given by  $a_0b_0, a_1b_0, a_0b_1, a_1b_1$ . Using notations due to Yates we shall denote them by (1) or simply 1, a, b, ab respectively. Clearly the rule of notation is that if a factor is at level 0, the corresponding small letter is absent, and if it is at level 1, the corresponding small letter is present. We shall use the same notations to denote

the treatment combinations as well as their effects, and the meaning will usually be clear from the context.

We define simple effect (s.e.) of A at  $b_0$  as  $(a - 1)$ , while the s.e. of A at  $b_1$  as  $(ab - b)$ . The main effect of A, denoted by simply A, is defined as the average of these two simple effects. Thus  $A = (a - 1 + ab - b)/2$ ; or A is symbolically equal to  $(a - 1)(b + 1)/2$ . Similarly the main effect of B is defined as the average of  $(b - 1)$  and  $(ab - a)$ ; and so B is symbolically equal to  $(a + 1)(b - 1)/2$ . Now, if A and B do not interact, we would expect s.e. of A at  $b_1$  and the s.e. of A at  $b_0$  to be of the same order. Thus a difference of these two simple effects would measure the interaction between the two factors. And so, we define the interaction  $AB = (ab - b - a + 1)/2$ , or AB is symbolically equal to  $(a - 1)(b - 1)/2$ . The three effects A, B, and AB that we have just defined can be presented in the following tabular form.

Table of signs of the treatment combinations.

Effects	(1)	a	b	ab	Divisor
A	-	+	-	+	2
B	-	-	+	+	2
AB	+	-	-	+	2

Note that mathematically A, B and AB are simply a set of three mutually orthogonal treatment contrasts that can be formed with four treatment combinations, but statistically they represent as many main effects and interactions of the factors under study. Consequently, the best estimates of these effects are obtainable from the same expressions by replacing the treatment combinations by the respective treatment totals and multiplying the divisors by r, where r is the replication number. If  $[x]$  denotes the total of the r observations under the treatment combination x, then, estimated  $A = (-[1] + [a] - [b] + [ab])/2r = [A]/2r$ , say, where  $[A]$  is called the factorial effect total for main effect A. And therefore the sum of squares due to A, i.e., SSA is equal to  $[A]^2/4r$ , carrying 1 d.f. Similarly, estimated  $B = [B]/2r$ , where  $[B] = (-[1] - [a] + [b] + [ab])$ , and  $SSB = [B]^2/4r$  carrying 1 df. And estimated  $AB = [AB]/2r$ , where  $[AB] = ([1] - [a] - [b] + [ab])$ , and  $SSAB = [AB]^2/4r$  carrying 1 df. Depending on the design used the ANOVA (analysis of variance) table can now be set up. If an RBD with r replications are used, then the ANOVA would be as follows.

Source of Variation	d.f.	S.S.	M.S. = S.S./df	F
Replications	$r - 1$	SSR	MSR	-
A	1	SSA	MSA	MSA/MSE
B	1	SSB	MSB	MSB/MSE
AB	1	SSAB	MSAB	MSAB/MSE
Error	$3(r - 1)$	SSE (by subtraction)	MSE	-
Total	$4r - 1$	SST	-	-

Here  $SSR = (R_1^2 + R_2^2 + \dots + R_r^2)/4 - CF$ ,  $CF = G^2/4r$ , and  $SST = (\text{Sum of squares of all the } 4r \text{ observations}) - CF$ , G being the grand total and  $R_i$  being the total of the  $i$ th replication.



### 3. A 2<sup>3</sup> EXPERIMENT

We can now easily extend the above developments of a 2<sup>2</sup> experiment to those of a 2<sup>3</sup> experiment. Let the factors be denoted by A, B, C. Then using Yates' notations the 8 treatment combinations are (1), a, b, ab, c, ac, bc, abc. There are 2<sup>3</sup> - 1 = 7 main effects and interactions, and these are defined symbolically as follows.

$$A=(a-1)(b+1)(c+1)/2^2, B=(a+1)(b-1)(c+1)/2^2, C=(a+1)(b+1)(c-1)/2^2,$$

$$AB=(a-1)(b-1)(c+1)/2^2, AC=(a-1)(b+1)(c-1)/2^2, BC=(a+1)(b-1)(c-1)/2^2,$$

$$ABC=(a-1)(b-1)(c-1)/2^2.$$

Expanding these expressions we can easily prepare the following table of signs.

Table of signs of the treatment combinations

Effects	(1)	a	b	ab	c	ac	bc	abc	Divisor
A	-	+	-	+	-	+	-	+	4
B	-	-	+	+	-	-	+	+	4
C	-	-	-	-	+	+	+	+	4
AB	+	-	-	+	+	-	-	+	4
AC	+	-	+	-	-	+	-	+	4
BC	+	+	-	-	-	-	+	+	4
ABC	-	+	+	-	+	-	-	+	4

An inspection of the signs above reveals a simple rule : (i) if an effect X has an even number of factors in it, then the sign of a treatment combination x in the expression of X is + when x has an even number of letters common with X, and - when it has an odd number of letters common with X; while (ii) if X has an odd number of factors in it, then the sign of x in the expression of X is + when x has an odd number of letters common with X, and - when x has an even number of letters common with X, 0 in this rule being considered an even number. We also note that we can obtain the signs of say AB by simply multiplying the corresponding signs of A and B, and so on. Thus without actually expanding the expressions for the various main effects and interactions we can construct the table of signs by simply using the rule just described.

The 7 effects of the above 2<sup>3</sup> experiment can be estimated and their sums of squares obtained using the above table of signs. Note that the sum of squares due to an effect is equal to the square of the factorial effect total divided by r<sup>2</sup><sup>3</sup>, where r is the replication number. However, it is clear that the process can be very tedious if the number of factors is relatively large. In such cases Yates' algorithm of sums and differences is employed to compute the various factorial effect totals, i.e., the quantities [A], [B], ..., [ABC]. We shall illustrate this algorithm via an actual analysis. But before that here is a little description of the procedure. If a column has entries (x<sub>1</sub>, x<sub>2</sub>, x<sub>3</sub>, x<sub>4</sub>, x<sub>5</sub>, x<sub>6</sub>, x<sub>7</sub>, x<sub>8</sub>) in this order, on which we perform the operation of sum and difference (OSD), then we get as a result a new column of entries (y<sub>1</sub>, y<sub>2</sub>, y<sub>3</sub>, y<sub>4</sub>, y<sub>5</sub>, y<sub>6</sub>, y<sub>7</sub>, y<sub>8</sub>) in this order, where, y<sub>1</sub>=x<sub>1</sub>+x<sub>2</sub>, y<sub>2</sub>=x<sub>3</sub>+x<sub>4</sub>, y<sub>3</sub>=x<sub>5</sub>+x<sub>6</sub>, y<sub>4</sub>=x<sub>7</sub>+x<sub>8</sub>, and y<sub>5</sub>=x<sub>2</sub>-x<sub>1</sub>, y<sub>6</sub>=x<sub>4</sub>-x<sub>3</sub>, y<sub>7</sub>=x<sub>6</sub>-x<sub>5</sub>, y<sub>8</sub>=x<sub>8</sub>-x<sub>7</sub>. We perform this OSD recursively 3 times (3 being the number of factors in the experiment) on the column of the treatment totals written against the column of the treatment combinations in Yates' standard order,

such as (1), a, b, ab, c, ac, bc, abc. The final column obtained in this process will be that of the factorial effect totals in the order G, [A], [B], [AB], [C], [AC], [BC], [ABC], where G is the grand total of all the  $r^2^3$  observations.

Example 1. Let us analyse the data from a  $2^3$  factorial experiment conducted in 3 randomised complete blocks. The three factors were the fertilizers, viz., nitrogen (N), phosphorus (P), and potassium (K). The yields along with the totals are as follows.

Table of treatment combinations and their yields

Blocks	(1)	n	p	np	k	nk	pk	npk	Total
B-1	101	106	312	373	265	291	391	450	2289 (B <sub>1</sub> )
B-2	106	89	324	338	272	306	407	449	2291 (B <sub>2</sub> )
B-3	87	128	323	324	279	334	423	471	2369 (B <sub>3</sub> )
Total	294 (T <sub>1</sub> )	323 (T <sub>2</sub> )	959 (T <sub>3</sub> )	1035 (T <sub>4</sub> )	816 (T <sub>5</sub> )	931 (T <sub>6</sub> )	1221 (T <sub>7</sub> )	1370 (T <sub>8</sub> )	6949 (G)

Grand Total G = 6949 Number of Observations  $n = r^2^3 = 3.8 = 24$

Correction Factor  $CF = G^2/n = 2012025.042$

Total S.S. (SST) =  $(101^2 + 106^2 + 312^2 + \dots + 423^2 + 471^2) - CF = 352843.958$

Block (Replication) S.S. (SSR) =  $(2289^2 + 2291^2 + 2369^2)/8 - CF = 520.333$

Treatment S.S. (SSTr.) =  $(294^2 + 323^2 + \dots + 1370^2)/3 - CF = 348651.291$

Therefore, Error S.S. (SSE) =  $SST - SSR - SSTr. = 3672.334$

Yates' table of sums and differences yielding Factorial effect totals

Treatment combinations in Yates' standard order	Treatment Totals	OSD on Column 2	OSD on Column 3	OSD on Column 4	Factorial Effect Totals
Column 1	Column 2	Column 3	Column 4	Column 5	Column 6
(1)	294	617	2611	6949	G
n	323	1994	4338	369	[N]
p	959	1747	105	2221	[P]
np	1035	2591	264	81	[NP]
k	816	29	1377	1727	[K]
nk	931	76	844	159	[NK]
pk	1221	115	47	- 533	[PK]
npk	1370	149	34	- 13	[NPK]

Therefore S.S. due to N (SSN) =  $[N]^2/r^2^3 = 369^2/24 = 5673.375$  etc. This leads us to the ANOVA table as follows.

Source of Variation	d.f.	S.S.	M.S.	F
Replications	3-1=2	520,333	260.167	0.9918
Treatments	$2^3-1=7$	348651.291	49807.328	189.880*
N	1	5673.375	5673.375	21.629*
P	1	205535.042	205535.042	783.558*
K	1	124272.042	124272.042	473.761*
NP	1	273.375	273.375	1.0422
NK	1	1053.375	1053.375	4.016
PK	1	11837.041	11837.041	45.126*
NPK	1	7.041	7.041	0.027
Error	$2 \times 7 = 14$	3672.337	262.310	
Total	$24-1=23$	352843.958		

(\* indicates significance at 5% level)

Before we end this section, we would like to draw attention to the recursive rule that we are following to write down the treatment combinations in Yates' standard order. We begin with (1) or 1. Next we introduce the small letter corresponding to the first factor, say a. This gives us  $2^1$  treatment combinations. Then we bring in the small letter, say b for the second factor, and 'multiply' all the previous combinations by b, giving us 1, a, b, ab as the  $2^2$  treatment combinations. Next comes c for the third factor, and we 'multiply' all the previous combinations by c, yielding us 1, a, b, ab, c, ac, bc, abc. And we continue the 'multiplication' until the last factor letter comes.

Thus, for the  $2^4$  and  $2^5$  experiments the treatment combinations in standard order are respectively :

(1, a, b, ab, c, ac, bc, abc, d, ad, bd, abd, cd, acd, bcd, abcd), and  
 (1, a, b, ab, c, ac, bc, abc, d, ad, bd, abd, cd, acd, bcd, abcd, e, ae, be, abe, ce, ace, bce, abce, de, ade, bde, abde, cde, acde, bcde, abcde). Expressions for the various main effects and interactions for these experiments can now be easily written down by simply extending those for the  $2^3$  experiments that we had developed earlier in this section. ANOVA's can also be similarly set up for for the  $2^4$  and  $2^5$  factorial experiments by simple extensions of the previous results.

We would close this section with the remark that if a latin square design were used instead of a randomized block design, then in the ANOVA table of the factorial experiment we should have Row S.S. and Column S.S. instead of the Block or Replication S.S., before the Treatments S.S. and its split up into the single df components representing the main effects and interactions. This modification in the analysis of variance is applicable to any kind of orthogonal design used in the experiment.

## 4. CONFOUNDING IN 2<sup>n</sup> EXPERIMENTS

If the number of factors in a 2<sup>n</sup> factorial experiment is large, then the number of treatment combinations is also very large, and consequently in such situations it may not be feasible or advisable to use randomized complete blocks, because then the basic assumption of the homogeneity of the experimental units within the large sized blocks itself is questionable. Use of special kinds of incomplete block designs is recommended in such situations. Confounding is a design technique which helps us to form these incomplete blocks without complicating the analysis much, and without compromising on the precision of the estimates of the more important effects like, say the main effects and the two factor interactions.

Consider a 2<sup>3</sup> experiment with A, B, C as factors, and suppose we want to split the set of 8 treatment combinations into 2 incomplete blocks of size 4 each. Notice that in the definition of the interaction ABC the 4 treatment combinations (a, b, c, abc) have + sign while the remaining 4 treatment combinations (1, ab, ac, bc) has – sign. Let us assign the first 4 treatments into one block, say B-I, and the second 4 treatments into another block, say B-II of size 4 each. So, B-I and B-II make one replication. Observe that if we use the 8 observations from this replication of two blocks of 4 plots each, the best estimates of any effect other than ABC are all free from block effects, and hence these effects are estimable with full precision as in an orthogonal design like say RBD. However interaction effect ABC got mixed up or entangled with the block effects, and is rendered non-estimable in this replication. We say ABC is confounded with the blocks of this replication.

If an effect is confounded in all the replications of the design, we say it is totally confounded in the design, and its df and S.S. do not appear in the ANOVA table, reducing thereby the treatment df by one. If, on the other hand, an effect is confounded in some but not all the replications, then we say it is partially confounded, and its S.S. is computed from Yates' table of sums and differences with an adjustment at the end. Loss of information on an effect X is defined as the ratio of the number of replications in which X is confounded to the total number of replications in the design. Thus the loss of information of a totally confounded effect is 1, while that of a partially confounded effect is less than 1.

Consider a 2<sup>4</sup> experiment with A, B, C, D as factors. Suppose we want to use blocks of size 8 only. Then we have 2 blocks per replication. The following confounding scheme is called a balanced confounding scheme, because interactions of the same order are confounded in the same number of replications. The order of an interaction is defined to be the number of factors in it minus 1.

Replication	Effect to be confounded	Replication	Effect to be confounded
I	ABC	III	ABD
II	ACD	IV	BCD

Interactions of order 2 are confounded in one replication, while effects of all other orders are confounded in zero replications. So this design is a balanced confounded design. The block contents of this design are as follows.

Replication	Effect confounded	Block	Treatment combinations
I	ABC	1	(1), ab, ac, bc, d, abd, acd, bcd
		2	a, b, c, abc, ad, bd, cd, abcd
II	ACD	3	(1), ac, ad, cd, b, abc, abd, bcd
		4	a, c, d, acd, ab, bc, bd, abcd
III	ABD	5	(1), ab, ad, bd, c, abc, acd, bcd
		6	a, b, d, abd, ac, bc, cd, abcd
IV	BCD	7	(1), bc, bd, cd, a, abc, abd, acd
		8	b, c, d, bcd, ab, ac, ad, abcd

The method of construction that we have followed here is the following. The block with the treatment combination (1) is called the key block. Other blocks in a replication are called the non-key blocks of the replication. Since size of the key block is  $2^3$ , we have 3 (the index) independent treatment combinations in it and the rest are obtainable from them by ‘multiplication’ of the independent treatment combinations, taking two at a time, three at a time, and so on. In this rule of multiplication squares of letters are taken to be 1, i.e.,  $a^2=b^2=c^2= \dots=1$ . Thus,  $a.ab=b$ ,  $ab.bd=ad$  etc. The independent treatment combinations of the key block of a replication are those that have an even number of letters common with the effect(s) to be confounded in the replication, 0 being considered an even number for the purpose. Once the key block is obtained, a non-key block is obtained by first writing down any treatment combination that has not appeared in any of the blocks already obtained, say  $x$ , and then multiplying all the treatment combinations of the key block by  $x$ .

In general a  $(2^n, 2^k)$  confounded design refers to a  $2^n$  factorial experiment in  $2^{n-k}$  blocks of  $2^k$  plots each per replication. For such a design we need to confound  $(2^{n-k} - 1)$  interactions in each replication, of which only  $(n-k)$  are independent while the rest are generalised interactions which are obtainable by ‘multiplying’ the independent interactions taking two at a time, three at a time, and so on, where the multiplication follows the rule as explained earlier.

With this background we can now discuss a  $(2^5, 2^3)$  balanced confounded design and its analysis. Let the factors be denoted by A, B, C, D, E. We must confound 3 interactions in each replication of which 2 are independent and one is generalised. We notice that there are  ${}^5C_3=10$  three factor interactions, and  ${}^5C_4=5$  four factor interactions. If we can confound all these 15 interactions each exactly in one replication, we shall need five replications for the balanced confounded design. By trial and error we get the following balanced scheme. The key blocks of all the replications are also obtained by following the method of construction as described earlier.

Replication	Effects to be confounded	Independent treatment combinations of the key block
I	ABD, ACE, BCDE	(1), ade, bd, ce
II	ACD, ACE, ABDE	(1), ad, be, cde
III	ADE, BCD, ABCE	(1), ae, bde, cde
IV	ABE, CDE, ABCD	(1), ade, bde, cd
V	ABC, BDE, ACDE	(1), ac, bce, de

For illustration purposes we shall construct below all the four blocks of Replication I only, by following the method as described earlier.

#### Replication I

Block 1	Block 2	Block 3	Block 4
(1)	a	b	c
ade	de	abde	acde
bd	abd	d	bcd
ce	ace	bce	e
abe	be	ae	abce
acd	cd	abcd	ad
bcde	abcde	cde	bde
abc	bc	ac	ab

As an exercise you are advised to construct the 16 other blocks of this balanced confounded design. To develop the ANOVA table for this design, we now prepare Yates' table of sums and differences and perform OSD five times to obtain the factorial effect totals  $[A]$ ,  $[B]$ ,  $[AB]$ , ...,  $[ABCDE]$ . S.S. due to an effect  $[X]$  is  $[X]^2/r2^n$ , if  $[X]$  is unconfounded in the design. If however  $[X]$  is partially confounded in the design, then its S.S. is  $[X]^2/r'2^n$ , where  $r'$  is the number of replications in which  $[X]$  is not confounded and  $[X]^* = [X] - [X]^a$ , where again  $[X]^a = \{(\text{sum of totals of those blocks which belong to the replications in which } X \text{ is confounded, and which contain treatment combinations having an odd number of letters common with } X) - (\text{sum of totals of those blocks which belong to the replications in which } X \text{ is confounded, and which contain treatment combinations having an even number of letters common with } X)\}$ , provided  $X$  has an odd number of factors in it. The words 'odd' and 'even' inside  $\{ \}$  exchange places when  $X$  has an even number of factors in it. Thus  $[ABD]^a = (B_2+B_3) - (B_1+B_4)$ , where  $B_i$ 's are the block totals of the Replication I.

Before we close this section, we suggest, you construct and develop analysis for, a balanced confounded design for a  $(2^3, 2^2)$  factorial experiment in (i) 3 replications, and (ii) 4 replications. Also try to find out the interactions confounded in a replication of a  $(2^5, 2^3)$  experiment in which one of the blocks has the treatment combinations ( a, bd, ce, cd, be, abc, ade, abcde) by following the reverse process of the method of construction that we had outlined earlier.

## 5. 3<sup>rd</sup> EXPERIMENTS

Suppose A and B are two factors, each at three (equi-spaced) levels denoted by  $a_0, a_1, a_2$  and  $b_0, b_1, b_2$  respectively. The  $3^2=9$  treatment combinations are  $a_0b_0, a_1b_0, a_2b_0, a_0b_1, a_1b_1, a_2b_1, a_0b_2, a_1b_2, a_2b_2$ . In Yates' notations these are (1), a,  $a^2, b, ab, a^2b, b^2, ab^2, a^2b^2$ . It will, however, be more convenient for us to use the vector notations for the treatment combinations, by which the set of 9 treatment combinations is  $S = (00, 10, 20, 01, 11, 21, 02, 12, 22)$ . Clearly here  $x_1x_2$  represents a treatment combination with the first factor at level  $x_1$  and the second factor at level  $x_2$ , where  $x_1=0, 1, 2$  and  $x_2=0, 1, 2$ . With 9 treatments we can think of  $9-1=8$  mutually orthogonal treatment contrasts carrying 8 df. We shall now try to form these contrasts in such a manner that they represent the main effects A, B and interaction AxB, carrying respectively 2, 2 and 4 df. To define the main effect A, we first divide S into three subsets  $A_0, A_1$  and  $A_2$  of three treatments each, where  $A_i$  consists of the treatment combinations that satisfy  $x_1=i, i=0, 1, 2$ . Thus  $A_0=(00, 01, 02)$ . Similarly,  $A_1=(10, 11, 12)$  and  $A_2=(20, 21, 22)$ . Using same notation for a treatment combination as well as its effect, we define  $(A)_i$  as the sum of the effects of the treatments in  $A_i$ . So,  $(A)_i=(i0 + i1 + i2), i=0, 1, 2$ . We now define the main effect of A by a set of two mutually orthogonal contrasts among  $(A)_i$ 's,  $i=0, 1, 2$ . Thus if  $(c_0, c_1, c_2)$  and  $(d_0, d_1, d_2)$  are such that  $c_0+c_1+c_2=d_0+d_1+d_2=0$ , and  $c_0d_0+c_1d_1+c_2d_2=0$ , then the two contrasts defining the main effect A are  $C_A^1 = c_0(A)_0+c_1(A)_1+c_2(A)_2$ , and  $C_A^2 = d_0(A)_0+d_1(A)_1+d_2(A)_2$ . On similar lines, if  $B_i$ =the subset of treatments in S which satisfy  $x_2=i, i=0, 1, 2$ , then  $C_B^1 = c_0(B)_0+c_1(B)_1+c_2(B)_2$  and  $C_B^2 = d_0(B)_0+d_1(B)_1+d_2(B)_2$  define the main effect of B. Interaction AxB has two components, called AB and  $AB^2$ , each carrying 2 df, where the defining equations are  $x_1+x_2=0, 1, 2 \pmod{3}$ , and  $x_1+2x_2=0, 1, 2 \pmod{3}$  respectively. Here additions and multiplications of the levels are done modulo 3. By this we mean that whenever a result exceeds 3, we replace it by the remainder left when divided by 3. Thus, by  $a=b \pmod{3}$ , we mean  $a=3q+b$ , i.e., when a is divided by 3, q is the quotient and b is the remainder. So, the subsets of treatment combinations that define AB and  $AB^2$  are as follows.

Subset	Treatment combinations	Subset	Treatment combinations
$AB_0$	00, 12, 21	$AB^2_0$	00, 11, 22
$AB_1$	10, 22, 01	$AB^2_1$	10, 21, 02
$AB_2$	20, 02, 11	$AB^2_2$	20, 01, 12

So,  $C_{AB}^1 = c_0(00+12+21)+c_1(10+22+01)+c_2(20+02+11)$ , and  $C_{AB}^2 = d_0(00+12+21)+d_1(10+22+01)+d_2(20+02+11)$ , define the interaction AB with 2 df. Similarly we can define interaction  $AB^2$  with 2 df. Note that these  $4 \times 2=8$  contrasts are really as many mutually orthogonal treatment contrasts which represent our main effects and interactions.

Having defined the four effects each with 2 df, we would now like to get their best estimates and sum of squares due to them. Let for an effect X,  $[X]_i$  denote the sum of the totals of the treatment combinations in the subset of treatments  $X_i$ . Thus  $[A]_i = T(i0) + T(i1) + T(i2), i=0, 1, 2$ , where  $T(x_1x_2)$  stands for the total of r observations under the treatment combination  $x_1x_2$ , r being the replication number. Then S.S. due to main effect A =  $([A]_0^2+[A]_1^2+[A]_2^2)/3r - G^2/9r$ , where  $G = [A]_0+[A]_1+[A]_2$ , the grand total of all the 9r observations. In general S.S. due an

effect X carrying 2 df =  $([X]_0^2 + [X]_1^2 + [X]_2^2) / 3r - G^2 / 9r$ . The following will, therefore, be the ANOVA of a  $3^2$  experiment conducted in r randomized complete blocks.

Sources of Variation	d.f.	S.S.	M.S. = S.S./d.f.	F=MSX/MSE
Replications	r-1	SSR	MSR	-
Treatments	9-1=8	SSTr	-	-
A	2	SSA	MSA	$F_A$
B	2	SSB	MSB	$F_B$
AB	2	SSAB	MSAB	$F_{AB}$
$AB^2$	2	$SSAB^2$	$MSAB^2$	$F_{AB^2}$
Error	8(r-1)	SSE (by subtraction)	MSE	-
Total	9r-1	SST	-	-

Here  $SSR = (R_1^2 + R_2^2 + \dots + R_r^2) / 9 - CF$ ,  $SST = (\text{Sum of squares of all the } 9r \text{ observations}) - CF$ ,  $CF = G^2 / 9r$ ,  $R_i$  being the total of observations in the  $i$ th replication.

Consider now a  $3^3$  experiment with A, B, C as factors. There are  $3^3 - 1 = 26$  df due to treatments which are split into 13 main effects and interactions, each carrying 2 df. The set of 27 treatment combinations is given by  $S = (x_1 x_2 x_3)$ , with  $x_i = 0, 1, 2$  (mod 3). The effects and their defining equations are as follows.

Table of effects and defining equations of a  $3^3$  experiment

Effect	Def. Equation	Effect	Def. Equation
A	$x_1 = 0, 1, 2 \pmod{3}$	BC	$x_2 + x_3 = 0, 1, 2 \pmod{3}$
B	$x_2 = 0, 1, 2 \pmod{3}$	$BC^2$	$x_2 + 2x_3 = 0, 1, 2 \pmod{3}$
C	$x_3 = 0, 1, 2 \pmod{3}$	ABC	$x_1 + x_2 + x_3 = 0, 1, 2 \pmod{3}$
AB	$x_1 + x_2 = 0, 1, 2 \pmod{3}$	$AB^2C$	$x_1 + 2x_2 + x_3 = 0, 1, 2 \pmod{3}$
$AB^2$	$x_1 + 2x_2 = 0, 1, 2 \pmod{3}$	$ABC^2$	$x_1 + x_2 + 2x_3 = 0, 1, 2 \pmod{3}$
AC	$x_1 + x_3 = 0, 1, 2 \pmod{3}$	$AB^2C^2$	$x_1 + 2x_2 + 2x_3 = 0, 1, 2 \pmod{3}$
$AC^2$	$x_1 + 2x_3 = 0, 1, 2 \pmod{3}$	-	-

For example, the partition of S into  $ABC_0$ ,  $ABC_1$ , and  $ABC_2$  would be like :

$ABC_0 = (000, 102, 012, 111, 210, 201, 021, 222, 120)$  ;

$ABC_1 = (100, 202, 112, 211, 010, 001, 121, 022, 220)$  ;

$ABC_2 = (200, 002, 212, 011, 110, 101, 221, 122, 020)$  .

You can practice such partitioning with respect to all the other 12 effects, and check that in the process you have constructed 26 mutually orthogonal treatment contrasts defining 13 effects each carrying 2 df. To computer sum of squares due to an



effect X carrying 2 df we use the formula :  $SSX = ([X]_0^2 + [X]_1^2 + [X]_2^2) / (r.3^2) - CF$ , where  $CF = G^2 / (r.3^3)$ , G being the grand total, i.e.,  $G = [X]_0 + [X]_1 + [X]_2$ . Now depending on the design used we can set up the ANOVA table for the  $3^3$  experiment.

The ideas explained above can be easily extended to those of a  $3^4$  experiment. If A, B, C, D denote the four factors, then there are four main effects, A,B,C,D;  ${}^4C_{2,2} = 12$  two factor interactions like  $AB^u, \dots, CD^u$ , where  $u=1, 2$ ;  ${}^4C_{3,2} = 16$  three factor interactions like  $AB^u C^v, \dots, BC^u D^v$ , where  $u,v=1,2$ ; and  ${}^4C_{4,2} = 8$  four factor interactions like  $AB^u C^v D^w$ , where  $u,v,w=1,2$ . These  $(4+12+16+8)=40$  effects each carrying 2 df represent  $40 \times 2 = 80$  mutually orthogonal treatment contrasts among 81 treatment combinations of a  $3^4$  factorial experiment. The defining equations of an effect  $X = A^\alpha B^\beta C^\gamma D^\delta$  are clearly :  $\alpha x_1 + \beta x_2 + \gamma x_3 + \delta x_4 = 0, 1, 2 \pmod{3}$ , where  $\alpha, \beta, \gamma, \delta$  take values 0, 1, or 2. To obtain the partition  $X_0$ , we have to solve the system of linear equation :  $\alpha x_1 + \beta x_2 + \gamma x_3 + \delta x_4 = 0 \pmod{3}$  for the unknowns  $x_1, x_2, x_3, x_4$ , which take values 0, 1, or 2 only. There are 3 independent solutions of this system, which, say, are  $t_1, t_2$  and  $t_3$ . These 3 generate all the  $3^3$  solutions in  $X_0$ . In fact they are :  $c_1 t_1 + c_2 t_2 + c_3 t_3 \pmod{3}$ , where  $c_1, c_2, c_3 = 0, 1, 2$ . For example, to obtain the partition  $AB^2 CD^2_0$ , we have to solve the equation  $x_1 + 2x_2 + x_3 + 2x_4 = 0 \pmod{3}$ . From inspection we see that 1001, 0102 and 0011 are three independent solutions of this equation. Combining these three vectors linearly with coefficient 0, 1, 2, we get all the 27 treatment combinations of  $AB^2 CD^2_0$  : (0000, 1001, 0102, 0011, 1100, 1012, 0110, 1111, 1202, 1020, 0121, 2101, 2010, 0212, 2002, 0201, 0022, 2200, 2021, 0220, 2222, 2101, 2010, 0212, 1202, 1020, 0121). To get  $AB^2 CD^2_1$  we start with one treatment combination that has not appeared before, say, x, and then add x to all the 27 treatment combinations of  $AB^2 CD^2_0$ . Similarly we write some treatment combination, say y, which did not appear before, and then add y to all the treatment combinations of  $AB^2 CD^2_0$ , to get  $AB^2 CD^2_2$ . To compute sum of squares due to an effect X carrying 2 df, we first write down  $X_0, X_1, X_2$ , and then get the totals  $[X]_0, [X]_1, [X]_2$ , and use the formula :  $S.S. = ([X]_0^2 + [X]_1^2 + [X]_2^2) / (r.3^{4-1}) - CF$ , where  $CF = G^2 / (r.3^4)$ , G being the grand total, i.e.,  $G = [X]_0 + [X]_1 + [X]_2$ . Then we compute SSR, SST and SSE by subtraction, and set up the ANOVA table as shown earlier.

## 6. CONFOUNDING IN $3^n$ EXPERIMENTS

As you may notice for large n,  $3^n$  is very large, and then the use of complete blocks of  $3^n$  homogeneous plots may either be not feasible or advisable. And then we go for smaller (incomplete) blocks of size  $3^k$ , where  $k < n$ , by using the technique of confounding as in the case of  $2^n$  experiments for large n. The construction follows the same principle as before. That is, when  $k=n-1$ , we use the defining partitions of an effect as our blocks, thereby confounding the effect with the blocks of the replication. When  $k < n-1$ , we confound two or more independent and their generalised interactions with the blocks of the replication. And, as earlier, if  $X_1$  and  $X_2$  are two independent interactions, then the generalized interactions are obtained by multiplying them, like  $X_1 X_2$  and  $X_1 X_2^2$ , the indices being reduced modulo 3. While multiplying if the index of the first factor becomes 2, then we square the interaction to reduce the index of the first factor to one, a convention we follow while writing down the distinct interaction effects among the factors.

Consider now a  $(3^3, 3^2)$  experiment with A, B, C as factors. There are  ${}^3C_3 \cdot 2^{3-1} = 4$  three factor interactions : ABC,  $AB^2C$ ,  $ABC^2$  and  $AB^2C^2$  in this experiment. So, confounding each of them in one replication, we can get a balanced confounded design in four replications. To construct the two independent treatment combinations of the key block of the replication in which we would confound ABC, we have to solve the equation  $x_1 + x_2 + x_3 = 0 \pmod{3}$ , i.e.,  $x_1 + x_2 = 2x_3 \pmod{3}$ . Obviously, these are ( 102, 012). Note that we have used the two rows of the unit matrix of order two under  $x_1$  and  $x_2$  and then determined  $x_3$  satisfying the given equation, to find out the two independent treatment combinations of the key block. All the 9 treatment combinations of the key block are, therefore, ( 000, 102, 012, 111, 120, 201, 021, 222, 210 ). Next, we find 100 had not appeared earlier. So, we add 100 to all the treatments of the key block to get one non-key block : ( 100, 202, 112, 211, 220, 001, 121, 022, 010 ). Since 200 had not appeared so far, we can add this treatment to all the treatments of the key block to get the other non-key block. So, it is : ( 200, 002, 212, 011, 020, 101, 221, 122, 110 ). Proceeding on similar lines we can construct all the 12 blocks of this balanced confounded design.

Let us now consider a  $(3^4, 3^2)$  factorial design with A, B, C, D as factors. By trial and error we can construct a balanced confounded design in four replications in this case also. The confounding scheme is as follows.

Replication	Interactions to confound
I	$AB^2C$ , $AB^2D^2$ , $ACD$ , $BCD^2$
II	$AB^2C^2$ , $ABD$ , $ACD^2$ , $BC^2D^2$
III	$ABC$ , $AB^2D$ , $AC^2D^2$ , $BC^2D$
IV	$AB^2C$ , $ABD^2$ , $AC^2D$ , $BCD$

Note that all the  ${}^4C_3 \cdot 2^2 = 16$  three factor interactions are confounded in this design, each in exactly one replication. Hence the design is balanced. We shall now obtain the contents of the key block of Replication I only, and others can be obtained similarly. For the first two interactions of Replication I, the defining equations are :  $x_1 + 2x_2 + x_3 = 0 \pmod{3}$ , and  $x_1 + 2x_2 + 2x_4 = 0 \pmod{3}$ , i.e.,  $x_3 = 2x_1 + x_2$  and  $x_4 = x_1 + 2x_2$ . Hence taking 10 and 01 under  $x_1$  and  $x_2$ , we have the two independent treatment combinations of the key block of Replication I as : 1021 and 0112. Generating others from these two, we have the key block : (0000, 1021, 0112, 1100, 1212, 2012, 0221, 2200, 2121).

## 7. BOOKS FOR REFERENCE

1. Das, M. N. and Giri, N. C. (1986). **Design and Analysis of Experiments**. Wiley Eastern Limited (Second edition).
2. Montgomery, Douglas C. (1976) **Design and Analysis of Experiments**. John Wiley & Sons, New York.
3. Cochran, W. G. and Cox, G. M. (1967) **Experimental Designs**. John Wiley & Sons, New York.

## STATISTICAL METHODS FOR SPATIAL DATA

Satyabrata Pal

Honorary Visiting Professor

International Statistical Education Centre, Indian Statistical Institute, Kolkata

### ABSTRACT OF THE LECTURE TO BE PRESENTED IN THE WORKSHOP

Statistics connotes the science of uncertainty, its goal is to attempt to model order in presence of disorder. It is an extremely powerful research tool. It is a recognisable phenomenon that even in a big scale if disorder is the law, there exist a smaller scale when the data do not fit the theory exactly which necessitates to investigate the residual uncertainty. The level of disorder is measured by a quantity called entropy. If there are  $k$  possible states of nature that occur at random with probabilities,  $p_1, p_2, \dots, p_k$  (arising from a probability distribution with  $p_i \geq 0$  and  $\sum_1^k p_i = 1$ ), the entropy (Shannon) is defined as  $E = - \sum_1^k p_i \cdot \log p_i$ , where, in general,  $x \cdot \log x = 0$  for  $x = 0$ . Mean ( $\mu = \sum_1^k i \cdot p_i$ ) and variance  $\sigma^2 = \sum_1^k (i - \mu)^2 \cdot p_i$ ; Statisticians use  $\sigma^2$  to quantify disorder, but it can be seen that  $E$  and  $\sigma^2$  are closely related.

Modelling data requires some conditions – the data are random and are identically distributed. Lack of homogeneity in data is accounted for by assuming non-constant mean, often the mean is assumed to be a linear combination of several explanatory variables. Apart from such large scale variation there may exist inhomogeneous small scale variation. Such small scale variations call for treatment with respect to heteroscedastic behaviour. Also dependent data model is also a reality in some situations. To tackle dependence, intraclass-correlation structures and serial-correlation structures are used. However, such tools offer little scope in case of spatial data where dependence is present in all directions. Spatial data can be observed over unidirectional flow of time. The natural consequence is that data which are close together, in time or space, are correlated. Spatial models are used in soil science, epidemiology, crop science, ecology, forestry, astronomy, atmospheric science, geology, etc, and, specially, in cases when data are observed at different spatial locations, and dependence becomes a reality, here again the degree of dependence is also a criterion to be looked into. Usually, models need to be flexible than their temporal counterparts (as past, present and future have no analogy in space). Yet, to deal with data where space-time dependencies are likely, two approaches can be pursued - departure from independence paradigm could be modelled or statistical procedures are developed that would be robust to those departures.

Coming to the discipline of agricultural science, it is a truth that R. A. Fisher was clearly aware of spatial dependencies a hundred years ago (while he was at Rothamsted Experimental Station) and he suggested the fundamental principles of Design of Experiments so as to take care of it. But it was later realised that even after application of fundamental principles, data coming from actual field experiments remain correlated. Fairfield Smith observed that by choosing plot dimension error variance can be reduced. Papadakis, Bartlett, Wilkinson, Besag and Kempton developed nearest neighbour models to take care of the correlation.

Data types in the above context are: Geostatistical data, Lattice data and Point Patterns. Description of such data sets will be discussed.

For Geostatistical data, different procedures (and concepts), viz., square-root difference, pocket plot, large and small scale variation, variogram, median polish estimation technique, estimator (also robust estimator) of variogram, valid co-variogram and variogram model fitting, cross-validation of fitted variogram, will be discussed. Also the Kriging (co-kriging) procedures (for spatial prediction) will be discussed.

Treatment of Lattice data will be covered under the purview of the lecture.

All of the above procedures will be discussed with examples, mostly live,

## FUZZY LINEAR REGRESSION USING SAS SOFTWARE

G Sathish, Minakshi Mishra and Prof D. Mazumdar

Department of Agricultural Statistics, BCKV, Nadia- 741252 (W.B)

Multiple linear regression modelling is a very powerful technique and is extensively used in agricultural research. This technique estimates linear relationship between dependent (response) and independent (explanatory) variables. If  $X_i, i=1,2,\dots,n$  are explanatory variables and  $Y$  is response variable, the model is expressed as :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + e \quad (1)$$

where  $\beta$ 's are parameters and  $e$  is the error term assumed to be following a normal distribution. The parameters are generally estimated using method of least squares. A good description of various aspects of multiple linear regression methodology is given in Draper and Smith (1998).

One drawback of the above methodology is that the underlying relationship is assumed to be crisp or precise, as it gives a precise value of response for a set of values of explanatory variables. However, in a realistic situation, the underlying relationship is not a crisp function of a given form; it contains some vagueness or impreciseness. So, by assuming a crisp relationship, some vital information may be lost (Slowinski 1998). A very promising technique of fuzzy regression has been developed. This technique can be applied to solve agricultural research problems.

A fuzzy regression model corresponding to equation (1) can be written as:

$$Y = A_0 + A_1 X_1 + \dots + A_n X_n \quad (2)$$

Here explanatory variables  $X_i$ 's, as before, are assumed to be precise. However as mentioned above, response variable  $Y$  is not crisp but is instead fuzzy in nature. This implies that the parameters are also fuzzy in nature. Our aim is to estimate these parameters, it is assumed that  $A_i$ 's are symmetric fuzzy numbers (i.e. vagueness is expressible as equidistant from the center) and so can be represented by intervals. For example,  $A_i$  can be expressed as fuzzy set given by:

$$A_i = \langle a_{1c}, a_{1w} \rangle \quad (3)$$

where  $a_{1c}$  is centre and  $a_{1w}$  is radius or vagueness associated. The above fuzzy set describes belief of regression coefficient around in terms of symmetric triangular membership function. This methodology is applied when the underlying phenomenon is fuzzy which means that the response variable is fuzzy and the relationship is also considered to be fuzzy.

Method of estimation of parameters of equation (2) is different from that of equation (1). In fuzzy regression methodology, parameters are estimated by minimizing total vagueness in the model, i.e. sum of radii of predicted intervals, from equation (2):

$$Y_j = A_0 + A_1 X_{1j} + \dots + A_n X_{nj}$$

Using equation (3),

$$y_j = \langle a_{0c}, a_{0w} \rangle + \langle a_{1c}, a_{1w} \rangle x_{1j} + \dots + \langle a_{nc}, a_{nw} \rangle x_{nj} = \langle y_{jc}, y_{jw} \rangle, \text{ say}$$

Thus,

$$y_{jc} = a_{0c} + a_{1c} x_{1j} + \dots + a_{nc} x_{nj} \tag{4a}$$

$$y_{jw} = a_{0w} + a_{1w} |x_{1j}| + \dots + a_{nw} |x_{nj}| \tag{4b}$$

As  $y_{jw}$  represents radius and so cannot be negative, therefore on the right hand side of equation 4(b), absolute values of  $x_{ij}$  are taken. Then parameters  $A_i$  are estimated by minimizing the quantity, which is total vagueness of the model-data set combination, subject to the constraints that each data point must fall within estimated value of response variable. This can be visualized as the following linear programming problem (Tanaka, 1987):

$$\text{Minimize } \sum_{j=1}^m (a_{0w} + a_{1w} |x_{1j}| + \dots + a_{nw} |x_{nj}|) \tag{5}$$

$$\text{Subject to } \left\{ \left( a_{0c} + \sum_{i=1}^n a_{ic} x_{ij} \right) - \left( a_{0w} + \sum_{i=1}^n a_{iw} x_{ij} \right) \right\} \leq Y_j$$

$$\left\{ \left( a_{0c} + \sum_{i=1}^n a_{ic} x_{ij} \right) + \left( a_{0w} + \sum_{i=1}^n a_{iw} x_{ij} \right) \right\} \geq Y_j \text{ and } a_{iw} \geq 0$$

To solve the above linear programming problem, Simplex procedure (Taha, 1997) is generally employed.

**ILLUSTRATION 1:** Suppose that the response variable is dry-matter accumulation (Y) and the explanatory variable is plant height (X1). Only the data pertaining to maturity level, i.e. 60 days after sowing (DAS), are considered for data analysis and the same are presented in Table 1 for ready reference.

**Table 1: Data of dry matter accumulation and plant height for effect of sulphur on growth of greengram crop**

Dry-matter accumulation (g/m <sup>2</sup> )	Plant Height (cm)	
247.32	60.41	
324.52	61.08	
364.56	64.98	
328.44	64.16	
349.48	62.99	
339.92	65.20	
320.48	63.24	
357.16	67.19	

The linear regression model and fuzzy regression model are fitted to the above data using SAS, version 9.3 software package and following are the SAS codes and results obtained:

**Method of linear regression (LR)**

Title 'Method of least square';

**data** plant;

input y x1;

247.32 60.41

324.52 61.08

364.56 64.98

328.44 64.16

349.48 62.99

339.92 65.20

320.48 63.24

357.16 67.19

;

**proc reg;**

model y=x1;

output out=all;

proc print data=all;

run;

quit;

**Method of Fuzzy regression (FR) (OPTMODEL)**

Title 'Linear programming';

**data** plant;

input y x1;

datalines;

247.32 60.41

324.52 61.08

364.56 64.98

328.44 64.16

349.48 62.99

339.92 65.20

320.48 63.24

357.16 67.19

```

;
run;
proc optmodel;
set j= 1..8;
number y{j}, x1{j};
read data plant into [_n_] y x1;
/*Print y x1*/
print y x1;
number n init 8; /* Total number of Observations*/
/* Decision Variables*/
var aw{1..2}>=0; /*Theses three variables are bounded*/
var ac{1..2}; /* These three variables are not bounded*/
/* Objective function*/
min z1= aw[1] * n + sum{i in j} x1[i] * aw[2];
/*Linear Constraints*/
con c{i in 1..n}: ac[1]+x1[i]*ac[2]-aw[1]-x1[i]*aw[2]<= y[i];
con c1{i in 1..n}: ac[1]+x1[i]*ac[2]+aw[1]+x1[i]*aw[2]>= y[i];
expand; /* This provides all equations */
solve;
print ac aw;
quit;

```

**RESULTS:****Partial SAS output:****Method of linear regression (LR)**

Parameter Estimates				
Variable	Parameter Estimate	Standard Error	t-value	Sig.
Intercept	-460.09564	280.45936	-1.64	0.1520
X1	12.39596	4.40349	2.82	0.0306

**Method of Fuzzy regression (FR) (OPTMODEL)**

Parameter Estimates		
Variable	Parameter Estimate(ac)	Standard Error (aw)
Constant	-698.185	33.173
X1	16.201	0.000

The fitted model for LR is

$$Y = -460.10 + 12.40 X1 \quad (6)$$

Standard Errors (280.46) (4.40)

The fitted model for FR is

$$Y = \langle -698.19, 33.17 \rangle + \langle 16.20, 0 \rangle X1 \quad (7)$$

In order to compare performance of above 2 approaches, viz. linear regression methodology and fuzzy regression methodology, width of prediction intervals corresponding to each observed value of response variable is computed.

For the former, upper limits of prediction interval are computed from the prediction equation (6) by taking the coefficient as their corresponding estimated values plus standard error, i.e.

$$Y = (-460.10 + 280.46) + (12.40 + 4.40) X1$$

Similarly, lower limits of prediction interval for linear regression model is computed using the equation (6),

$$Y = (-460.10 - 280.46) + (12.40 - 4.40) X1$$

Further, for fuzzy regression model, the prediction equations for computing upper and lower limits, obtained from equation (7), are respectively

$$Y = (-698.19 + 33.17) + (16.20 + 0) X1$$

And

$$Y = (-698.19 - 33.17) + (16.20 - 0) X1$$

The width of prediction intervals in respect of linear regression model and fuzzy regression model corresponding to observed explanatory variable is computed in MS Excel and the results are reported in the following Table 2.



**Table 2: Fitting of LR and FLR**

Observed Dry matter accumulation	Predicted_LR			Predicted_FLR		
	upper limit	lower limit	width	upper limit	lower limit	width
247.32	835.22	-257.73	1092.95	313.69	247.34	66.35
324.52	846.47	-252.37	1098.85	324.55	258.20	66.35
364.56	911.99	-221.20	1133.20	387.73	321.38	66.35
328.44	898.22	-227.76	1125.97	374.44	308.10	66.35
349.48	878.56	-237.11	1115.67	355.49	289.14	66.35
339.92	915.69	-219.45	1135.13	391.29	324.95	66.35
320.48	882.76	-235.11	1117.87	359.54	293.19	66.35
357.16	949.12	-203.54	1152.66	423.53	357.19	66.35
Average width			1121.54	Average width		<b>66.35</b>

From the Table 2, average width for former was found to be 1121.54, while that for latter was only 66.35, indicating thereby the superiority of fuzzy regression methodology. In reality, underlying phenomenon is fuzzy; therefore, as emphasized above, correct methodology to obtain relationship between response and explanatory variable is to apply fuzzy regression methodology rather than linear regression methodology.

**ILLUSTRATION 2:** For more than one explanatory variables i.e. Multiple linear regression (MLR). Data (2009-10) given in the PhD Thesis of Minakshi Mishra are considered. She studied the relationship of macro climatic parameters with the growth processes parameters using FLR techniques. The response variable is wheat yield (Y) and the explanatory variables are Tmax (X1), Tmin (X2), RH-I (X3), RH-II (X4), Rainfall (X5), LAI (X6) and Plant height (X7) are considered for data analysis and the same are presented in Table 3 for ready reference.

Table 3: Yield, macro climatic parameters (Tmax, Tmin, RH-I, RH-II, Rainfall) and growth processes (LAI and Plant height) of wheat crop

Yield	Tmax	Tmin	RH-I	RH-II	Rainfall	LAI	Plntht
2530.92	32.19	21.29	93.71	59.86	0.00	3.04	46.31
2617.34	32.67	20.29	92.14	53.14	0.00	3.32	53.92
2532.20	29.24	20.06	93.71	60.43	0.00	3.19	50.23
3495.24	27.74	13.11	93.29	46.43	0.00	3.91	48.60
3342.85	28.36	13.94	94.29	53.00	0.00	3.69	57.13
3192.43	26.99	12.29	95.14	52.00	0.00	3.50	52.92
3968.47	28.27	14.73	94.00	52.00	0.00	4.25	52.46
3669.77	24.79	7.87	92.71	42.71	0.00	4.03	61.83

3484.49	22.04	8.40	97.71	59.00	0.00	3.85	57.13
2164.08	23.54	9.70	94.43	49.57	0.00	2.72	49.23
2375.86	21.46	10.81	97.71	63.14	0.00	3.14	57.38
2305.49	26.11	8.09	93.14	40.71	0.00	2.94	53.88
2902.78	27.71	10.16	92.43	36.00	0.00	3.39	50.81
3026.27	29.26	15.17	91.86	46.71	0.00	3.58	57.67
2697.51	29.11	16.10	91.86	45.00	0.20	3.22	54.96
3262.30	32.29	17.26	92.29	43.00	0.83	3.69	54.77
3342.01	35.66	21.20	91.29	31.57	0.00	3.95	61.58
3073.50	34.30	19.40	88.71	31.71	0.00	3.50	57.65

The multiple linear regression model and fuzzy regression model are fitted to the above data using SAS, version 9.3 software package and following are the SAS codes and results given below:

Method of Multiple linear regression (MLR)

Title 'Method of least square';

data wheatMa;

input y x1 x2 x3 x4 x5 x6 x7 ;

cards;

```

2530.92      32.19      21.29      93.71      59.86      0.00      3.04      46.31
2617.34      32.67      20.29      92.14      53.14      0.00      3.32      53.92
2532.20      29.24      20.06      93.71      60.43      0.00      3.19      50.23
3495.24      27.74      13.11      93.29      46.43      0.00      3.91      48.60
3342.85      28.36      13.94      94.29      53.00      0.00      3.69      57.13
3192.43      26.99      12.29      95.14      52.00      0.00      3.50      52.92
3968.47      28.27      14.73      94.00      52.00      0.00      4.25      52.46
3669.77      24.79      7.87      92.71      42.71      0.00      4.03      61.83
3484.49      22.04      8.40      97.71      59.00      0.00      3.85      57.13
2164.08      23.54      9.70      94.43      49.57      0.00      2.72      49.23
2375.86      21.46      10.81      97.71      63.14      0.00      3.14      57.38
2305.49      26.11      8.09      93.14      40.71      0.00      2.94      53.88
2902.78      27.71      10.16      92.43      36.00      0.00      3.39      50.81
3026.27      29.26      15.17      91.86      46.71      0.00      3.58      57.67
2697.51      29.11      16.10      91.86      45.00      0.20      3.22      54.96
3262.30      32.29      17.26      92.29      43.00      0.83      3.69      54.77
3342.01      35.66      21.20      91.29      31.57      0.00      3.95      61.58
3073.50      34.30      19.40      88.71      31.71      0.00      3.50      57.65

```

```
;
proc reg;
model Y= X1 X2 X3 X4 X5 X6 X7;
output out=all;
proc print data=all;
run;
quit;
```

Method of Fuzzy regression (FR) (OPTMODEL)

Title 'Linear programming';

data wheatMa;

infile datalines;

input y x1 x2 x3 x4 x5 x6 x7;

datalines;

2530.92	32.19	21.29	93.71	59.86	0.00	3.04	46.31
2617.34	32.67	20.29	92.14	53.14	0.00	3.32	53.92
2532.20	29.24	20.06	93.71	60.43	0.00	3.19	50.23
3495.24	27.74	13.11	93.29	46.43	0.00	3.91	48.60
3342.85	28.36	13.94	94.29	53.00	0.00	3.69	57.13
3192.43	26.99	12.29	95.14	52.00	0.00	3.50	52.92
3968.47	28.27	14.73	94.00	52.00	0.00	4.25	52.46
3669.77	24.79	7.87	92.71	42.71	0.00	4.03	61.83
3484.49	22.04	8.40	97.71	59.00	0.00	3.85	57.13
2164.08	23.54	9.70	94.43	49.57	0.00	2.72	49.23
2375.86	21.46	10.81	97.71	63.14	0.00	3.14	57.38
2305.49	26.11	8.09	93.14	40.71	0.00	2.94	53.88
2902.78	27.71	10.16	92.43	36.00	0.00	3.39	50.81
3026.27	29.26	15.17	91.86	46.71	0.00	3.58	57.67
2697.51	29.11	16.10	91.86	45.00	0.20	3.22	54.96
3262.30	32.29	17.26	92.29	43.00	0.83	3.69	54.77
3342.01	35.66	21.20	91.29	31.57	0.00	3.95	61.58
3073.50	34.30	19.40	88.71	31.71	0.00	3.50	57.65

;

run;

proc optmodel;

set j= 1..18;

number y{j}, x1 {j}, x2 {j}, x3 {j}, x4 {j}, x5 {j}, x6 {j}, x7 {j};

```

read data wheatMa into [_n_] y x1 x2 x3 x4 x5 x6 x7; /*Print y x1 x2*/
print y x1 x2 x3 x4 x5 x6 x7;
number n init 18;
var aw{1..8}>=0;
var ac{1..8};
min z1= aw[1]*n + sum{i in j} x1[i] * aw[2] + sum{i in j} x2[i] * aw[3] + sum{i in j} x3[i] * aw[4] + sum{i in
j} x4[i] * aw[5]+ sum{i in j} x5[i] * aw[6] + sum{i in j} x6[i] * aw[7] + sum{i in j} x7[i] * aw[8];
con c{i in 1..n}: ac[1] + x1[i]*ac[2] + x2[i]*ac[3] + x3[i]*ac[4] + x4[i]*ac[5] + x5[i]*ac[6] + x6[i]*ac[7] +
x7[i]*ac[8] - aw[1] - x1[i]*aw[2] - x2[i]*aw[3]-x3[i]*aw[4] - x4[i]*aw[5] - x5[i]*aw[6] - x6[i]*aw[7] -
x7[i]*aw[8]<= y[i];
con c1 {i in 1..n}: ac[1] + x1[i]*ac[2] + x2[i]*ac[3] + x3[i]*ac[4] + x4[i]*ac[5] + x5[i]*ac[6] + x6[i]*ac[7] +
x7[i]*ac[8] + aw[1] + x1[i]*aw[2] + x2[i]*aw[3] + x3[i]*aw[4] + x4[i]*aw[5] + x5[i]*aw[6] + x6[i]*aw[7] +
x7[i]*aw[8]>= y[i];
expand; /* This provides all equations */
solve;
print ac aw;
quit;

```

Note: y= yield, x1= Tmax, x2= Tmin, x3= RH-I, x4= RH-II, x5= Rainfall, x6= LAI and x7= Plntht

## RESULTS (SAS output)

**Table 4. Method of Multiple linear regression (MLR)**

Parameter Estimates				
Variable	Parameter Estimate	Standard Error	t-value	Sig.
(Constant)	16.973	3221.18	0.005	0.996
Tmax	29.816	34.923	0.854	0.413
Tmin	-33.831	25.385	-1.333	0.212
RH-I	-16.007	33.629	-0.476	0.644
RH-II	6.872	8.566	0.802	0.441
Rainfall	52.162	145.188	0.359	0.727
LAI	1249.798	81.125	15.406	0.000
Plntht	-10.751	7.881	-1.364	0.202

**Table 5. Method of Fuzzy regression (FR) (OPTMODEL)**

Parameter Estimates		
Variable	Parameter Estimate(ac)	Standard Error (aw)
Constant	-3017.6165	0.0000
Tmax	41.1475	4.1317
Tmin	-34.9422	0.0000
RH-I	15.3737	0.0000
RH-II	2.4009	0.0000
Rainfall	0.0000	0.0000
LAI	1266.5993	0.0000
Plntht	-11.3984	0.0000

The fitted MLR model for showing the relationship of macro climatic parameters with growth processes parameters to predicted wheat yield during 2009-10 is

$$Y = 16.97 + 29.82 \text{ Tmax} - 33.83 \text{ Tmin} - 16.01 \text{ RH-I} + 6.87 \text{ RH-II} + 52.16 \text{ Rainfall} + 1249.798 \text{ LAI} - 10.75 \text{ Plntht}$$

Upper and lower limits of prediction interval for MLR model are computed from the above prediction equation by taking the coefficient as their corresponding estimated values plus or minus standard error (Table 4), i.e.

$$Y = (16.97 + 3221.18) + (29.82 + 34.92) \text{ Tmax} + (-33.83 + 25.39) \text{ Tmin} + (-16.01 + 33.63) \text{ RH-I} + (6.87 + 8.57) \text{ RH-II} + (52.16 + 145.19) \text{ Rainfall} + (1249.80 + 81.13) \text{ LAI} + (-10.75 + 7.88) \text{ Plntht}$$

And

$$Y = (16.97 - 3221.18) + (29.82 - 34.92) \text{ Tmax} + (-33.83 - 25.39) \text{ Tmin} + (-16.01 - 33.63) \text{ RH-I} + (6.87 - 8.57) \text{ RH-II} + (52.16 - 145.19) \text{ Rainfall} + (1249.80 - 81.13) \text{ LAI} + (-10.75 - 7.88) \text{ Plntht}$$

The fitted FLR model for showing the relationship of macro climatic parameters with growth processes parameters to predicted wheat yield during 2009-10 is

$$Y = \langle -3017.62, 0 \rangle + \langle 41.15, 4.13 \rangle \text{ Tmax} + \langle -34.94, 0 \rangle \text{ Tmin} + \langle 15.37, 0 \rangle \text{ RH-I} + \langle 2.40, 0 \rangle \text{ RH-II} + \langle 0, 0 \rangle \text{ Rainfall} + \langle 1266.60, 0 \rangle \text{ LAI} + \langle -11.40, 0 \rangle \text{ Plntht}$$

Upper and lower limits of prediction interval for FLR model are computed by taking values of aic and aiw where  $i = 1$  to 8 (Table 5) respectively as,

$$Y = (-3017.62 + 0) + (41.15 + 4.13) \text{ Tmax} + (-34.94 + 0) \text{ Tmin} + (15.37 + 0) \text{ RH-I} + (2.40 + 0) \text{ RH-II} + (0 + 0) \text{ Rainfall} + (1266.60 + 0) \text{ LAI} + (-11.40 + 0) \text{ Plntht}$$

And

$$Y = (-3017.62 - 0) + (41.15 - 4.13) \text{ Tmax} + (-34.94 - 0) \text{ Tmin} + (15.37 - 0) \text{ RH-I} + (2.40 - 0) \text{ RH-II} + (0 - 0) \text{ Rainfall} + (1266.60 - 0) \text{ LAI} + (-11.40 - 0) \text{ Plntht}$$

**Table 6: Fitting of MLR and FLR**

Observed Yield	Predicted_MLR			Predicted_FLR			
	upper limit	lower limit	width	upper limit	lower limit	width	
2530.92	11625.73	-6696.37	18322.10	2598.34	2332.38	265.96	
2617.34	11895.48	-6355.61	18251.09	2892.84	2622.86	269.98	
2532.20	11644.22	-6505.99	18150.21	2656.09	2414.44	241.65	
3495.24	12344.42	-5171.01	17515.43	3720.48	3491.23	229.25	
3342.85	12186.98	-5692.84	17879.82	3382.31	3147.98	234.33	
3192.43	11867.05	-5775.72	17642.77	3192.39	2969.40	222.99	
3968.47	12912.90	-4981.76	17894.66	4106.55	3872.93	233.62	
3669.77	12257.15	-4912.16	17169.31	3758.56	3553.75	204.81	
3484.49	12192.15	-5324.72	17516.87	3560.94	3378.80	182.15	
2164.08	10590.12	-6406.80	16996.92	2165.77	1971.23	194.54	
2375.86	11245.37	-6312.06	17557.43	2551.45	2374.14	177.31	
2305.49	10892.10	-6073.39	16965.49	2524.98	2309.19	215.79	
2902.78	11492.57	-5584.81	17077.38	3100.05	2871.04	229.01	
3026.27	11941.05	-5782.94	17723.99	3176.46	2934.70	241.76	
2697.51	11464.40	-6224.22	17688.61	2707.18	2466.60	240.58	
3262.30	12396.69	-5823.78	18220.47	3419.05	3152.26	266.79	
3342.01	12540.29	-5760.50	18300.79	3633.00	3338.35	294.65	
3073.50	11835.54	-5973.33	17808.87	3068.86	2785.42	283.43	
Average width			17704.57	Average width			<b>234.92</b>

From the Table 6, average width for former was found to be 17704.57, while that for latter was only 234.92, indicating thereby the superiority of fuzzy regression methodology. In reality, underlying phenomenon is fuzzy; therefore, as emphasized above, correct methodology to obtain relationship between response and explanatory variables is to apply fuzzy regression methodology rather than multiple linear regression methodology.

#### References

Draper N R and Smith H. 1998. Applied Regression Analysis, 3rd Edn. John Wiley and Sons, New York, USA.

- Mishra M and Mazumdar D. 2017. Application of soft-computing techniques on macro and micro climatic variables to forecast yields of major crops. PhD Thesis, BCKV. 1-141.
- Slowinski R. 1998. Fuzzy Sets in Decision Analysis. Operations Research and Statistics, Kluwer Academic Publishers, Boston.
- Taha H A. 1997. Operations research: An introduction, 6th Ed., Prentice Hall, New Jersey.
- Tanaka H. 1987. Fuzzy data analysis by possibilistic linear models. Fuzzy Sets and Systems 24: 363-375.

## **Business Intelligent in Cloud Computing**

**By**

**Manas Kumar Sanyal**

With the advancement of the technologies, organizations are highly dependent on the Information System (IS). Information System empowers the Organizations to gain competitive advantageous over their competitors. Business intelligence (BI) is combination of applications, technologies, processes and architectures which can be used to identify the historical, current and predictive views of business operations. BI Technologies can be used to support the collection, analysis, presentation and dissemination of business information. Business intelligence can be used to support a wide range of business decisions ranging from operational to strategic. Basic operating decisions include product positioning or pricing. Strategic business decisions include priorities, goals and directions at the broadest level. In all cases, BI is most effective when it combines data derived from the market in which a company operates (external data) with data from company sources internal to the business such as financial and operations data (internal data). When combined, external and internal data can provide a more complete picture which, in effect, creates an "intelligence" that cannot be derived by any singular set of data.

Business Intelligence (BI) helps organizations in processing internal as well as external data and converting them into required business information. Cloud computing has its own set of advantages that brings competitive solutions for the organizations. By combining these two, the results have been even more advantageous. Cloud based business intelligence and analytics is one of the fastest growing business management solutions even when the economy is taking a downturn. Organizations keep spending on BI as it helps them provide business information and market trends that are highly beneficial.

The use of Business Intelligence (BI) in the cloud is a game-changer, as it makes BI affordable and easily available as compared to traditional BI. It is expected that customers will slowly but surely migrate from in-house BI to BI in the cloud. In this lecture, we want to highlight the current need of the organizations to adapt BI in Cloud Computing.



## Abstract of presentation of Some Work On Non-Sampling Errors

**Prof.Pulakesh Maiti ISI Kolkata**

Usually, there are four stages in the operation of a sample survey namely.

- (1)Frame structure/Frame mechanism;
- (2) Sampling Mechanism;
- (3)Response Mechanism:
- (4)Measurement and/or imputation Mechanism.

All stages contribute to errors; The error occurred at the stage 2 is known as Sampling error, whereas the others arising from remaining three stages are known as Non-sampling Errors. Non-sampling errors may also occur during development of a schedule and or at the tabulation stage. Therefore assessment of errors associated with an estimator or a class of estimators must be made taking care of all the factors/errors above called total error comprising of sampling error as well as non sampling error .Sampling error has been considered mostly in survey literature, whereas much attention has not been paid in controlling non sampling errors, although there are enough evidences of existence of errors arising out of frame structure, incomplete data and/or measurement problem. Just as Sampling error is measured through standard error /sampling variance, there should come up something like non-sampling variance to deal with non sampling error. Hence, it is necessary that efficiency of an estimator should be judged with respect to total variance and not just by sampling error alone. Total Error can be decomposed in to the following:

$$\text{Total MSE/Variance}=\text{Sampling variance} + \text{Non-sampling Variance};$$

Non sampling variance can further be decomposed as

$$\begin{aligned} \text{Non-Sampling Variance} &= \text{Response Variance}+ \text{Measurement Variance} \\ &+ \text{Imputation variance.} \end{aligned}$$

The present series of lectures would focus mainly on determining ways and means to control and estimate non-sampling variance along with the sampling variance associated with an estimator for a parameter-linear and/or non-linear. The errors may be classified and quantified as follows:

(1)A frame ----list frame based on population units or a frame created through indirect identifiability of the inferential population units may have such problems as incompleteness causing coverage problem, multiplicity problem etc. In such situations how necessary measures as adopting multiple frames and defining appropriate estimators etc. may be defined would be discussed.

(2)Incomplete data cause non response problem. Analysis under this situation can be made by defining the estimator through either revising the original weights due to sampling design by response probabilities, without making any effort to make the data complete or On the other hand, incomplete data can be made complete by imputations by type-single or multiple and by category- Mean imputation or random imputation or regression type imputation etc..How this type of nonresponse problem could be tackled both from the points of deterministic as well as stochastic partotioning of the inferential population in to respondent and Non-respondent Groups would be discussed.

(3)The data collected need not always be error free, and with these erroneous data how analysis due to measurement error and/or imputation error can be made would be discussed, assuming that measurement taken from a responding unit at an instant has a distribution having finite mean and variance, and the measurement variance associated with an estimator can be discussed.

Estimation procedures developed by taking care of the above factors would be discussed with the Total Variance/Mean Square Error associated with the estimator(s) in judging their efficiencies under some sampling schemes.

Finally, we can discuss some methods of repeat Measurement and/or of Repeatability of samples as a whole with randomisation of the observations within the a sample for estimating total variance/mean square error as a whole or for estimating different components of the total variance, and we can close our discussions indicating some further research areas.

Investigation of the properties of an estimator in the light of above considerations in general and in particular in estimating the population total. We shall also discuss the problem of estimating Mean Square.

## Misuses of Statistics Prof. Shantaranjan Pal

"Lies, damned lies, and statistics" is a phrase describing the persuasive power of numbers, particularly the use of Statistics to bolster weak arguments. It is also sometimes colloquially used to doubt statistics used to prove an opponent's point. Mark Twain popularized the saying in *Chapters from My Autobiography*, published in the *North American Review* in 1906. "Figures often beguile me," he wrote, "particularly when I have the arranging of them myself; in which case the remark attributed to Disraeli would often apply with justice and force: 'There are three kinds of lies: lies, damned lies, and statistics.'"

Alternative attributions include, among many others (for example Walter Bagehot and Arthur James Balfour) the radical English journalist and politician Henry Du Pré Labouchère (1831–1912), Jervoise Athelstane Baines,<sup>[3]</sup> and British politician and man of letters Leonard H. Courtney, who used the phrase in 1895 and two years later became president of the Royal Statistical Society. Courtney is quoted by Baines (1896) as attributing the phrase to a "wise statesman" but he may have been referring to a future statesman rather than a past one. The phrase has also been attributed to Arthur Wellesley, 1st Duke of Wellington.

The earliest instance of the phrase found in print dates to a letter written in the British newspaper *National Observer* on June 8, 1891, published June 13, 1891, p. 93(-94): NATIONAL PENSIONS [To the Editor of The National Observer] London, 8 June 1891 "Sir, —It has been wittily remarked that there are three kinds of falsehood: the first is a 'fib,' the second is a downright lie, and the third and most aggravated is statistics. It is on statistics and on the absence of statistics that the advocate of national pensions relies...". Later, in October 1891, as a query in *Notes and Queries*, the pseudonymous questioner, signing as "St Swithin", asked for the originator of the phrase, indicating common usage even at that date The pseudonym has been attributed to Eliza Gutch.

That phrase can be found in *Nature*, page 74 November 26, 1885: "A well-known lawyer, now a judge, once grouped witnesses into three classes: simple liars, damned liars, and experts. He did not mean that the expert uttered things which he knew to be untrue, but that by the emphasis which he laid on certain statements, and by what has been defined as a highly cultivated faculty of evasion, the effect was actually worse than if he had."

Statistics may be a principled means of debate with opportunities for agreement, but this is true only if the parties agree to a set of rules. Misuses of statistics violate the rules.

Or to put it another way:

False facts are highly injurious to the progress of science, for they often long endure; but false views, if supported by some evidence, do little harm, as every one takes a salutary pleasure in proving their falseness; and when this is done, one path towards error is closed and the road to truth is often at the same time opened.

— Charles Darwin, *The Descent of Man* (1871), Vol. 2, 385.

Many misuses of statistics occur because

- ~ The source is a subject matter expert, not a statistics expert.<sup>[8]</sup> The source may incorrectly use a method or interpret a result.
- ~ The source is a statistician, not a subject matter expert.<sup>[9]</sup> An expert should know when the numbers being compared describe different things. Numbers change, as reality does not, when legal definitions or political boundaries change.
- ~ The subject being studied is not well defined.<sup>[10]</sup> While IQ tests are available and numeric it is difficult to define what they measure; Intelligence is an elusive concept. Publishing "impact" has the same problem.<sup>[11]</sup> A seemingly simple question about the number of words in the English language immediately encounters questions about archaic forms, accounting for prefixes and suffixes, multiple definitions of a word, variant spellings, dialects, fanciful creations (like ectoplastistics from ectoplasm and statistics),<sup>[12]</sup> technical vocabulary...
- ~ Data quality is poor.<sup>[13]</sup> Apparel provides an example. People have a wide range of sizes and body shapes. It is obvious that apparel sizing must be multidimensional. Instead it is complex in unexpected ways. Some apparel is sold by size only (with no explicit consideration of body shape), sizes vary by country and

manufacturer and [some sizes](#) are deliberately misleading. While sizes are numeric, only the crudest of statistical analyses is possible using the size numbers with care.

- ~ The popular press has limited expertise and mixed motives.<sup>[14]</sup> If the facts are not "newsworthy" (which may require exaggeration) they may not be published. The motives of advertisers are even more mixed.
- ~ "Politicians use statistics in the same way that a drunk uses lamp-posts—for support rather than illumination" - Andrew Lang (WikiQuote) "What do we learn from these two ways of looking at the same numbers? We learn that that a clever propagandist, right or left, can almost always find a way to present the data on economic growth that seems to support her case. And we therefore also learn to take any statistical analysis from a strongly political source with handfuls of salt."<sup>[15]</sup> The term statistics originates from numbers generated for and utilized by the state. Good government may require accurate numbers, but popular government may require supportive numbers (not necessarily the same). "The use and misuse of statistics by governments is an ancient art."

### Misuse of statistics may occur due to

1. Statistics usually produces probabilities; conclusions are provisional
2. The provisional conclusions have errors and error rates. Commonly 5% of the provisional conclusions of significance testing are wrong
3. Statisticians are not in complete agreement on ideal methods
4. Statistical methods are based on assumptions which are seldom fully met
5. Data gathering is usually limited by ethical, practical and constraints.

### Many misuses of statistics occur because

- ~ The source is a subject matter expert, not a statistics expert. The source may incorrectly use a method or interpret a result.
- ~ The source is a statistician, not a subject matter expert. An expert should know when the numbers being compared describe different things. Numbers change, as reality does not, when legal definitions or political boundaries change.
- ~ The subject being studied is not well defined. While [IQ tests](#) are available and numeric it is difficult to define what they measure; Intelligence is an elusive concept. Publishing "impact" has the same problem. A seemingly simple question about the number of words in the English language immediately encounters questions about archaic forms, accounting for prefixes and suffixes, multiple definitions of a word, variant spellings, dialects, fanciful creations (like ectoplastistics from ectoplasm and statistics), technical vocabulary...
- ~ Data quality is poor. Apparel provides an example. People have a wide range of sizes and body shapes. It is obvious that apparel sizing must be multidimensional. Instead it is complex in unexpected ways. Some [apparel](#) is sold by size only (with no explicit consideration of body shape), sizes vary by country and manufacturer and [some sizes](#) are deliberately misleading. While sizes are numeric, only the crudest of statistical analyses is possible using the size numbers with care.
- ~ The popular press has limited expertise and mixed motives. If the facts are not "newsworthy" (which may require exaggeration) they may not be published. The motives of advertisers are even more mixed. "Politicians use statistics in the same way that a drunk uses lamp-posts—for support rather than illumination" - Andrew Lang (WikiQuote) "What do we learn from these two ways of looking at the same numbers? We learn that that a clever propagandist, right or left, can almost always find a way to present the data on economic growth that seems to support her case. And we therefore also learn to take any statistical analysis from a strongly political source with handfuls of salt."<sup>[15]</sup> The term statistics originates from numbers generated for and utilized by the state. Good government may require accurate numbers, but popular government may require supportive numbers (not necessarily the same). "The use and misuse of statistics by governments is an ancient art." If a research team wants to know how 300 million people feel about a certain topic, it would be impractical to ask all of them. However, if the team picks a random sample of about 1000 people, they can be fairly certain that the results given by this group are representative of what the larger group would have said if they had all been asked.

This confidence can actually be quantified by the [central limit theorem](#) and other mathematical results. Confidence is expressed as a probability of the true result (for the larger group) being within a certain range of

the estimate (the figure for the smaller group). This is the "plus or minus" figure often quoted for statistical surveys. The probability part of the confidence level is usually not mentioned; if so, it is assumed to be a standard number like 95%.

The two numbers are related. If a survey has an estimated error of  $\pm 5\%$  at 95% confidence, it also has an estimated error of  $\pm 6.6\%$  at 99% confidence.  $\pm x\%$  at 95% confidence is always  $\pm 1.32x\%$  at 99% confidence for a normally distributed population.

The smaller the estimated error, the larger the required sample, at a given confidence level.

at 95.4% confidence:

$\pm 1\%$	would	require	10,000	people.
$\pm 2\%$	would	require	2,500	people.
$\pm 3\%$	would	require	1,111	people.
$\pm 4\%$	would	require	625	people.
$\pm 5\%$	would	require	400	people.
$\pm 10\%$	would	require	100	people.
$\pm 20\%$	would	require	25	people.
$\pm 25\%$	would	require	16	people.

$\pm 50\%$  would require 4 people.

People may assume, because the confidence figure is omitted, that there is a 100% certainty that the true result is within the estimated error. This is not mathematically correct. "...the null hypothesis is never proved or established, but it is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis." (Fisher in *The Design of Experiments*)

Many reasons for confusion exist including the use of double negative logic and terminology resulting from the merger of Fisher's "significance testing" (where the null hypothesis is never accepted) with "hypothesis testing". (where some hypothesis is always accepted).

This can—using the judicial analogue above—be compared with the truly guilty defendant who is released just because the proof is not enough for a guilty verdict. This does not prove the defendant's innocence, but only that there is not proof enough for a guilty verdict.

Many people may not realize that the randomness of the sample is very important. In practice, many opinion polls are conducted by phone, which distorts the sample in several ways, including exclusion of people who do not have phones, favoring the inclusion of people who have more than one phone, favoring the inclusion of people who are willing to participate in a phone survey over those who refuse, etc. Non-random sampling makes the estimated error unreliable.

On the other hand, people may consider that statistics are inherently unreliable because not everybody is called, or because they themselves are never polled. People may think that it is impossible to get data on the opinion of dozens of millions of people by just polling a few thousands. This is also inaccurate.<sup>[a]</sup> A poll with perfect unbiased sampling and truthful answers has a mathematically determined margin of error, which only depends on the number of people polled.

However, often only one margin of error is reported for a survey. When results are reported for population subgroups, a larger margin of error will apply, but this may not be made clear. For example, a survey of 1000 people may contain 100 people from a certain ethnic or economic group. The results focusing on that group will be much less reliable than results for the full population. If the margin of error for the full sample was 4%, say, then the margin of error for such a subgroup could be around 13%.

There are also many other measurement problems in population surveys.

The problems mentioned above apply to all statistical experiments, not just population surveys. When a statistical test shows a correlation between A and B, there are usually six possibilities:

1. A causes B.
2. B causes A.
3. A and B both partly cause each other.
4. A and B are both caused by a third factor, C.

5. B is caused by C which is correlated to A.
6. The observed correlation was due purely to chance.

The sixth possibility can be quantified by statistical tests that can calculate the probability that the correlation observed would be as large as it is just by chance if, in fact, there is no relationship between the variables. However, even if that possibility has a small probability, there are still the five others.

If the number of people buying ice cream at the beach is statistically related to the number of people who drown at the beach, then nobody would claim ice cream causes drowning because it's obvious that it isn't so. (In this case, both drowning and ice cream buying are clearly related by a third factor: the number of people at the beach).

This fallacy can be used, for example, to prove that exposure to a chemical causes cancer. Replace "number of people buying ice cream" with "number of people exposed to chemical X", and "number of people who drown" with "number of people who get cancer", and many people will believe you. In such a situation, there may be a statistical correlation even if there is no real effect. For example, if there is a perception that a chemical site is "dangerous" (even if it really isn't) property values in the area will decrease, which will entice more low-income families to move to that area. If low-income families are more likely to get cancer than high-income families (this can happen for many reasons, such as a poorer diet or less access to medical care) then rates of cancer will go up, even though the chemical itself is not dangerous. It is believed<sup>[24]</sup> that this is exactly what happened with some of the early studies showing a link between EMF ([electromagnetic fields](#)) from power lines and [cancer](#).

In well-designed studies, the effect of false causality can be eliminated by assigning some people into a "treatment group" and some people into a "control group" at random, and giving the treatment group the treatment and not giving the control group the treatment. In the above example, a researcher might expose one group of people to chemical X and leave a second group unexposed. If the first group had higher cancer rates, the researcher knows that there is no third factor that affected whether a person was exposed because he controlled who was exposed or not, and he assigned people to the exposed and non-exposed groups at random. However, in many applications, actually doing an experiment in this way is either prohibitively expensive, infeasible, unethical, illegal, or downright impossible. For example, it is highly unlikely that an [IRB](#) would accept an experiment that involved intentionally exposing people to a dangerous substance in order to test its toxicity. The obvious ethical implications of such types of experiments limit researchers' ability to empirically test causation.

"...the null hypothesis is never proved or established, but it is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis." (Fisher in [The Design of Experiments](#)) Many reasons for confusion exist including the use of double negative logic and terminology resulting from the merger of Fisher's "significance testing" (where the null hypothesis is never accepted) with "hypothesis testing" (where some hypothesis is always accepted).

#### **Confusing statistical significance with practical significance**

Statistical significance is a measure of probability; practical significance is a measure of effect. A baldness cure is statistically significant if a sparse peach-fuzz usually covers the previously naked scalp. The cure is practically significant when a hat is no longer required in cold weather and the barber asks how much to take off the top. The bald want a cure that is both statistically and practically significant; It will probably work and if it does, it will have a big hairy effect. Scientific publication often requires only statistical significance. This has led to complaints (for the last 50 years) that statistical significance testing is a misuse of statistics-

#### **Data dredging**

[Data dredging](#) is an abuse of [data mining](#). In data dredging, large compilations of data are examined in order to find a correlation, without any pre-defined choice of a [hypothesis](#) to be tested. Since the required [confidence interval](#) to establish a relationship between two parameters is usually chosen to be 95% (meaning that there is a 95% chance that the relationship observed is not due to random chance), there is thus a 5% chance of finding a correlation between any two sets of completely random variables. Given that data dredging efforts typically examine large datasets with many variables, and hence even larger numbers of pairs of variables, spurious but apparently statistically significant results are almost certain to be found by any such study.

Note that data dredging is a valid way of *finding* a possible hypothesis but that hypothesis *must* then be tested with data not used in the original dredging. The misuse comes in when that hypothesis is stated as fact without further validation.

"You cannot legitimately test a hypothesis on the same data that first suggested that hypothesis. The remedy is clear. Once you have a hypothesis, design a study to search specifically for the effect you now think is there. If the result of this test is statistically significant, you have real evidence at last."

### **Data manipulation**

Informally called "fudging the data," this practice includes selective reporting and even simply making up false data.

Examples of selective reporting abound. The easiest and most common examples involve choosing a group of results that follow a pattern consistent with the preferred hypothesis while ignoring other results or "data runs" that contradict the hypothesis.

Psychic researchers have long disputed studies showing people with ESP ability. Critics accuse ESP proponents of only publishing experiments with positive results and shelving those that show negative results. A "positive result" is a test run (or data run) in which the subject guesses a hidden card, etc., at a much higher frequency than random chance

Scientists, in general, question the validity of study results that cannot be reproduced by other investigators. However, some scientists refuse to publish their data and methods<sup>L</sup>

Data manipulation is a serious issue/consideration in the most honest of statistical analyses. Outliers, missing data and non-normality can all adversely affect the validity of statistical analysis. It is appropriate to study the data and repair real problems before analysis begins. "[I]n any scatter diagram there will be some points more or less detached from the main part of the cloud: these points should be rejected only for cause."

### **Other fallacies**

Pseudoreplication is a technical error associated with analysis of variance. Complexity hides the fact that statistical analysis is being attempted on a single sample (N=1). For this degenerate case the variance cannot be calculated (division by zero).

The gambler's fallacy assumes that an event for which a future likelihood can be measured had the same likelihood of happening once it has already occurred. Thus, if someone had already tossed 9 coins and each has come up heads, people tend to assume that the likelihood of a tenth toss also being heads is 1023 to 1 against (which it was before the first coin was tossed) when in fact the chance of the tenth head is 50% (assuming the coin is unbiased).

The prosecutor's fallacy<sup>[31]</sup> has led, in the UK, to the false imprisonment of women for murder when the courts were given the prior statistical likelihood of a woman's 3 children dying from Sudden Infant Death Syndrome as being the chances that their already dead children died from the syndrome. This led to statements from Roy Meadow that the chance they had died of Sudden Infant Death Syndrome were extremely small (one in millions). The courts then handed down convictions in spite of the statistical inevitability that a few women would suffer this tragedy. The convictions were eventually overturned (and Meadow was subsequently struck off the U.K. Medical Register for giving "erroneous" and "misleading" evidence, although this was later reversed by the courts).<sup>[32]</sup> Meadow's calculations were irrelevant to these cases, but even if they were, using the same methods of calculation would have shown that the odds against two cases of infanticide were even smaller (one in billions).<sup>[32]</sup>

**The lucid fallacy. Probabilities are based on simple models that ignore real (if remote) possibilities. Poker players do not consider that an opponent may draw a gun rather than a card. The insured (and governments) assume that insurers will remain solvent, but see AIG and systemic risk.**

### **Discarding unfavorable data**

All a company has to do to promote a neutral (useless) product is to find or conduct, for example, 40 studies with a confidence level of 95%. If the product is really useless, this would on average produce one study showing the product was beneficial, one study showing it was harmful and thirty-eight inconclusive studies (38 is 95% of 40). This tactic becomes more effective the more studies there are available. Organizations that do not publish every study they carry out, such as tobacco companies denying a link between smoking and cancer, anti-smoking advocacy groups and media outlets trying to prove a link between smokings and various ailments, or miracle pill vendors, are likely to use this tactic.

Ronald Fisher considered this issue in his famous lady tasting tea example experiment (from his 1935 book, *The Design of Experiments*). Regarding repeated experiments he said, "It would clearly be illegitimate, and would rob our calculation of its basis, if unsuccessful results were not all brought into the account."

Another term related to this concept is cherry picking.

### Loaded question

The answers to surveys can often be manipulated by wording the question in such a way as to induce a prevalence towards a certain answer from the respondent. For example, in polling support for a war, the questions:

~ Do you support the attempt by the USA to bring freedom and democracy to other places in the world?

~ Do you support the unprovoked military action by the USA?

will likely result in data skewed in different directions, although they are both polling about the support for the war. A better way of wording the question could be "Do you support the current US military action abroad?" A still more nearly neutral way to put that question is "What is your view about the current US military action abroad?" The point should be that the person being asked has no way of guessing from the wording what the questioner might want to hear.

Another way to do this is to precede the question by information that supports the "desired" answer. For example, more people will likely answer "yes" to the question "Given the increasing burden of taxes on middle-class families, do you support cuts in income tax?" than to the question "Considering the rising federal budget deficit and the desperate need for more revenue, do you support cuts in income tax?"

The proper formulation of questions can be very subtle. The responses to two questions can vary dramatically depending on the order in which they are asked.<sup>[17]</sup> "A survey that asked about 'ownership of stock' found that most Texas ranchers owned stock, though probably not the kind traded on the New York Stock Exchange."

### Overgeneralization

Overgeneralization is a fallacy occurring when a statistic about a particular population is asserted to hold among members of a group for which the original population is not a representative sample.

For example, suppose 100% of apples are observed to be red in summer. The assertion "All apples are red" would be an instance of overgeneralization because the original statistic was true only of a specific subset of apples (those in summer), which is not expected to be representative of the population of apples as a whole.

A real-world example of the overgeneralization fallacy can be observed as an artifact of modern polling techniques, which prohibit calling cell phones for over-the-phone political polls. As young people are more likely than other demographic groups to lack a conventional "landline" phone, a telephone poll that exclusively surveys responders of calls landline phones, may cause the poll results to undersample the views of young people, if no other measures are taken to account for this skewing of the sampling. Thus, a poll examining the voting preferences of young people using this technique may not be a perfectly accurate representation of young peoples' true voting preferences as a whole without overgeneralizing, because the sample used excludes young people that carry only cell phones, who may or may not have voting preferences that differ from the rest of the population.

Overgeneralization often occurs when information is passed through nontechnical sources, in particular mass media.

### References:

1. *The Design of Experiments*: Ronald Fisher 1935 Nonparametric statistics for the behavioral sciences. Siegel, Sidney New York, NY, US: McGraw-Hill Nonparametric statistics for the behavioral sciences. (1956). xvii 312 pp
2. Weatherburn, Don (November 2011), "Uses and abuses of crime statistics" *Crime and Justice Bulletin: Contemporary Issues in Crime and Justice*, NSW Bureau of Crime Statistics and Research, **153**, ISBN 9781921824357, ISSN 1030-1046, Archived from the original on June 21, 2014
3. Strasak, Alexander M.; Qamruz Zaman; Karl P. Pfeiffer; Georg Göbel; Hanno Ulmer (2007). "Statistical errors in medical research—a review of common pitfalls". *Swiss Medical Weekly*. **137**: 44–49. PMID 17299669
4. Indrayan, Abhaya (2007). "Statistical fallacies in orthopedic research". *Indian Journal of Orthopaedics*. **41** (1): 37. doi:10.4103/0019-5413.30524



## **Cluster Analysis**

**Prof. Pradip Kumar Sahu**

Head, Department of Agricultural Statistics,

Bidhan Chandra Krishi Vishwavidyalaya

Mohanpur, Nadia-741252

Contact Number: +91-8478912537

E mail I D: [pksbckv@gmail.com](mailto:pksbckv@gmail.com)

Multivariate Analysis is a comprehensive and perhaps the most realistic approach for describing a system involving multiple characters.

It is an extension of the commonly applied single variable statistical procedures to multiple variables of a system, duly considering the relationship among the variables.

Procedures under Multivariate Analysis have wide scope for application in agriculture and allied sciences.

# Multivariate Analysis

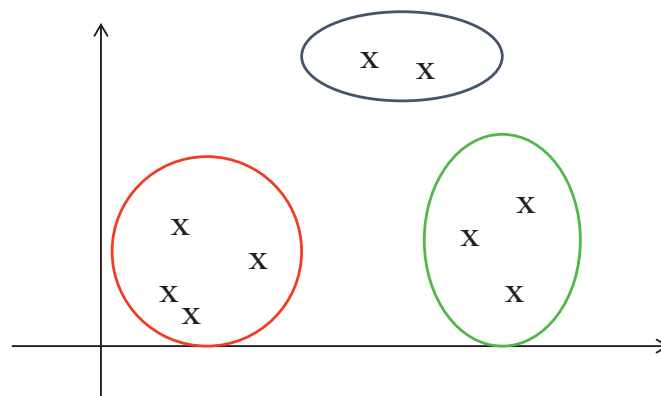
- Multiple linear regression
- Discriminant analysis
- Principal component analysis
- Canonical correlation
- Cluster analysis
- .....

## Cluster Analysis

Clustering is essentially a multivariate technique but can also be used in univariate case.

- The goal of clustering is to
  - Group individuals that are close (or **similar**) to each other

- Example



## Cluster Analysis

- Cluster Analysis deals with procedures for classifying the objects on the basis of their observational vectors into homogeneous groups, referred as clusters
- Procedures for formation of clusters were basically developed in Taxonomy for classification of Operating Taxonomic Units (OTU) of insects
- A separate branch of Numerical Taxonomy was also developed for this purpose, due to the efforts of Sneath and Sokal (1973).
- Objects are classified on the basis of "similarities" between them.
- Similarity is measured through the inter-object distances

### *Approaches to Clustering*

Hierarchical  
Optimization  
Ordination  
Clumping and  
Density Search

### Hierarchical Clustering Methods

1. Single Linkage Method
2. Complete Linkage Method
3. Average Linkage Method
4. Centroid Method and
5. Ward's Minimum Variance Method

By and large the whole process of **Hierarchical Clustering** accomplished in two phases

### Phase-1: Working out the distance

- Working out the distances between any two individuals considering all the characters under study together.

### Phase-2: Linking

- Joining the similar elements in to one groups and relatively dissimilar elements in to different groups.

### Distance Function:

If there are two objects (P, Q) with their observations X and Y, then

$d(P, Q)$  is a distance function if it has the following properties:

- **Symmetry:**  $d(P, Q) = d(Q, P)$
- **Non-Negativity:**  $d(P, Q) \geq 0$
- **Definiteness:**  $d(P, Q) = 0$  if & only if  $P = Q$
- **Triangle Inequality:**  $d(P, Q) \leq d(P, R) + d(R, Q)$

Some of the commonly applied distance measures are:

Euclidean Distance, Minkowski's Metric, Karl Pearson's Distance, Mahalanobis Standardized  $D^2$  Distance and Correlation.

# Various Distance Measures

## Euclidean distance

$$D(x, y) = \sqrt{\sum_{i=1}^m (x_i - y_i)^2}$$

where X and Y are the units measured over  $i = 1, 2, \dots, m$  characters.

## Squared Euclidean distance

$$D(x, y) = \sum_{i=1}^m (x_i - y_i)^2$$

The distance between any two objects is not affected by the addition of new objects to the analysis.

These are greatly affected by the differences in scale among the dimensions.

## Example

### Example data set

Subject ID	A	B
S1	5	5
S2	6	6
S3	15	14
S4	16	15
S5	25	20
S6	30	19

### Euclidean Distance (ED) between objects S1 (5, 5) and S2 (6, 6).

$$\begin{aligned}
 ED &= \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \\
 &= \sqrt{(6 - 5)^2 + (6 - 5)^2} \\
 &= 1.41
 \end{aligned}$$

Similarly, ED between S5 and S6 =

$$= 5.0990$$

	S1	S2	S3	S4	S5	S6
S1	0	1.4142	13.4536	14.8660	25	28.6530
S2	1.4142	0	12.0415	13.4536	23.6008	27.2940
S3	13.4536	12.0415	0	1.4142	11.6619	15.8114
S4	14.8660	13.4536	1.4142	0	10.2956	14.560
S5	25	23.6008	11.6619	10.2956	0	5.0990
S6	28.6530	27.2940	15.8114	14.560	5.0990	0



### City-block (Manhattan) distance

- ✓ Simply the sum of the absolute differences across dimensions.
- ✓ Yields result similar to the Euclidean distance.
- ✓ The effect of outlier is minimized as the differences are not squared.

$$D(x, y) = \sum_{i=1}^m |x_i - y_i|, i = 1, 2, \dots, m$$

### Chebychev distance

- ✓ Most simplest distance measure.
- ✓ Take care of only the distance of the character which has maximum absolute differences between the two units of the population.

$$D(x, y) = \text{Maximum} |x_i - y_i|, i = 1, 2, \dots, m$$

### City-block (Manhattan) distance

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_10	C_11	C_12	C_13	C_14	C_15	C_16	C_17	C_18	C_19	C_20
C_1	0	63	45	91	114	106	109	137	211	62	25	99	69	75	126	132	119	199	68	57
C_2	63	0	75	50	59	56	141	190	170	26	77	67	29	48	158	167	92	157	67	73
C_3	45	75	0	68	85	102	106	167	190	60	50	75	57	49	94	118	96	175	45	32
C_4	91	50	68	0	46	65	141	213	140	44	105	36	27	34	127	161	74	127	36	47
C_5	114	59	85	46	0	27	176	246	116	54	128	73	47	58	170	198	107	104	70	78
C_6	106	56	102	65	27	0	174	238	121	56	121	90	60	75	190	200	124	108	82	90
C_7	109	141	106	141	176	174	0	120	280	123	88	159	138	145	54	92	177	264	133	129
C_8	137	190	167	213	246	238	120	0	344	192	122	225	200	204	125	65	239	331	198	187
C_9	211	170	190	140	116	121	280	344	0	169	228	136	147	145	256	295	170	19	154	161
C_10	62	26	60	44	54	56	123	192	169	0	77	57	31	38	144	157	82	157	46	55
C_11	25	77	50	105	128	121	88	122	228	77	0	113	83	88	105	108	131	214	83	72
C_12	99	67	75	36	73	90	159	225	136	57	113	0	39	34	136	176	40	123	41	52
C_13	69	29	57	27	47	60	138	200	147	31	83	39	0	23	135	151	69	134	40	46
C_14	75	48	49	34	58	75	145	204	145	38	88	34	23	0	119	155	54	132	26	30
C_15	126	158	94	127	170	190	54	125	256	144	105	136	135	119	0	74	153	241	112	107
C_16	132	167	118	161	198	200	92	65	295	157	108	176	151	155	74	0	190	282	149	138
C_17	119	92	96	74	107	124	177	239	170	82	131	40	69	54	153	190	0	158	66	76
C_18	199	157	175	127	104	108	264	331	19	157	214	123	134	132	241	282	158	0	137	150
C_19	68	67	45	36	70	82	133	198	154	46	83	41	40	26	112	149	66	137	0	19
C_20	57	73	32	47	78	90	129	187	161	55	72	52	46	30	107	138	76	150	19	0

**Chebychev distance.**

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_10	C_11	C_12	C_13	C_14	C_15	C_16	C_17	C_18	C_19	C_20
C_1	0	38	30	40	65	68	25	55	144	29	16	54	35	37	37	58	79	132	36	35
C_2	38	0	41	24	28	31	63	92	106	8	53	45	17	27	73	95	69	94	27	27
C_3	30	41	0	44	69	72	22	57	147	33	27	37	39	32	32	54	48	135	24	14
C_4	40	24	44	0	25	28	65	95	104	18	56	21	7	12	75	98	46	91	20	29
C_5	65	28	69	25	0	12	91	120	78	36	81	32	30	37	101	123	50	66	45	55
C_6	68	31	72	28	12	0	94	123	76	39	84	37	33	40	103	126	62	63	48	58
C_7	25	63	22	65	91	94	0	41	169	55	18	58	61	54	21	32	63	157	45	36
C_8	55	92	57	95	120	123	41	0	199	84	39	88	90	83	63	49	105	186	75	66
C_9	144	106	147	104	78	76	169	199	0	114	159	111	108	115	179	201	116	12	124	133
C_10	29	8	33	18	36	39	55	84	114	0	45	39	11	21	65	87	63	102	20	20
C_11	16	53	27	56	81	84	18	39	159	45	0	51	51	44	33	42	75	147	36	32
C_12	54	45	37	21	32	37	58	88	111	39	51	0	28	18	68	91	24	99	18	22
C_13	35	17	39	7	30	33	61	90	108	11	51	28	0	10	71	93	53	96	15	25
C_14	37	27	32	12	37	40	54	83	115	21	44	18	10	0	64	86	42	103	9	18
C_15	37	73	32	75	101	103	21	63	179	65	33	68	71	64	0	22	63	167	55	46
C_16	58	95	54	98	123	126	32	49	201	87	42	91	93	86	22	0	85	189	78	68
C_17	79	69	48	46	50	62	63	105	116	63	75	24	53	42	63	85	0	104	43	43
C_18	132	94	135	91	66	63	157	186	12	102	147	99	96	103	167	189	104	0	112	121
C_19	36	27	24	20	45	48	45	75	124	20	36	18	15	9	55	78	43	112	0	9
C_20	35	27	14	29	55	58	36	66	133	20	32	22	25	18	46	68	43	121	9	0

**Power distance**

- ❖ More general form of Euclidean distance.
- ❖ Takes care of such situation where we may want to change the progressive weight that is placed on character differences on which the objects are very different

$$D(x, y) = \left( \sum_{i=1}^m |x_i - y_i|^p \right)^{\frac{1}{q}}, i = 1, 2, \dots, m$$

where  $p$  and  $q$  are the parameters to be defined by the user

**Percent disagreement**

Most useful for categorical data

$$D(x, y) = \frac{i' (= \text{number of characters where } x_i \neq y_i)}{i}, i = 1, 2, \dots, m$$

Clearly  $i' \leq i$



**Power distance**

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_10	C_11	C_12	C_13	C_14	C_15	C_16	C_17	C_18	C_19	C_20
C_1	0	40	31	53	72	71	42	66	152	34	16	64	44	47	56	65	84	140	41	37
C_2	40	0	47	28	34	32	70	102	113	11	54	46	18	29	82	99	71	102	34	38
C_3	31	47	0	45	69	74	39	80	149	37	30	44	40	33	41	59	58	137	25	16
C_4	53	28	45	0	29	35	73	115	105	24	64	23	11	15	78	100	48	93	21	30
C_5	72	34	69	29	0	13	97	135	81	39	86	42	31	38	105	126	63	69	47	56
C_6	71	32	74	35	13	0	99	133	82	39	85	52	35	45	109	128	75	70	52	61
C_7	42	70	39	73	97	99	0	54	175	62	32	75	68	65	25	39	87	163	58	51
C_8	66	102	80	115	135	133	54	0	213	98	55	122	107	107	68	49	135	202	101	93
C_9	152	113	149	105	81	82	175	213	0	119	166	112	111	116	182	204	121	12	125	134
C_10	34	11	37	24	39	39	62	98	119	0	47	40	14	22	73	91	64	108	24	28
C_11	16	54	30	64	86	85	32	55	166	47	0	71	56	55	45	50	87	154	49	42
C_12	64	46	44	23	42	52	75	122	112	40	71	0	29	19	74	98	25	99	23	30
C_13	44	18	40	11	31	35	68	107	111	14	56	29	0	13	75	95	53	99	19	27
C_14	47	29	33	15	38	45	65	107	116	22	55	19	13	0	68	90	42	104	11	18
C_15	56	82	41	78	105	109	25	68	182	73	45	74	75	68	0	31	78	170	60	52
C_16	65	99	59	100	126	128	39	49	204	91	50	98	95	90	31	0	104	192	82	73
C_17	84	71	58	48	63	75	87	135	121	64	87	25	53	42	78	104	0	109	44	47
C_18	140	102	137	93	69	70	163	202	12	108	154	99	99	104	170	192	109	0	113	122
C_19	41	34	25	21	47	52	58	101	125	24	49	23	19	11	60	82	44	113	0	10
C_20	37	38	16	30	56	61	51	93	134	28	42	30	27	18	52	73	47	122	10	0

**Percent disagreement**

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_10	C_11	C_12	C_13	C_14	C_15	C_16	C_17	C_18	C_19	C_20
C_1	0.00	1.00	1.00	1.00	1.00	1.00	0.92	1.00	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
C_2	1.00	0.00	1.00	1.00	0.92	1.00	1.00	1.00	1.00	0.92	1.00	1.00	1.00	1.00	1.00	1.00	0.92	1.00	1.00	1.00
C_3	1.00	1.00	0.00	1.00	1.00	0.92	1.00	1.00	1.00	1.00	1.00	0.85	1.00	1.00	1.00	1.00	1.00	0.92	1.00	0.92
C_4	1.00	1.00	1.00	0.00	0.92	0.85	1.00	1.00	1.00	0.92	0.92	1.00	1.00	1.00	1.00	1.00	1.00	0.92	1.00	1.00
C_5	1.00	0.92	1.00	0.92	0.00	0.92	1.00	1.00	1.00	0.85	0.85	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
C_6	1.00	1.00	0.92	0.85	0.92	0.00	1.00	1.00	1.00	1.00	0.92	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.92
C_7	0.92	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
C_8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
C_9	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	0.92	1.00	1.00	1.00	1.00	1.00	0.85	0.92
C_10	1.00	0.92	1.00	0.92	0.85	1.00	1.00	1.00	1.00	0.00	1.00	0.85	0.92	0.92	1.00	1.00	1.00	1.00	1.00	1.00
C_11	1.00	1.00	1.00	0.92	0.85	0.92	1.00	1.00	1.00	1.00	0.00	0.92	0.92	1.00	1.00	1.00	1.00	1.00	1.00	0.92
C_12	1.00	1.00	0.85	1.00	0.92	0.92	1.00	1.00	1.00	0.85	0.92	0.00	0.92	0.92	1.00	1.00	0.92	0.92	1.00	0.92
C_13	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.92	0.92	0.92	0.92	0.00	0.77	1.00	1.00	0.92	0.92	0.92	1.00
C_14	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.92	1.00	0.92	0.77	0.00	0.92	1.00	0.92	0.92	1.00	1.00
C_15	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.92	0.00	1.00	1.00	1.00	1.00	1.00
C_16	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00
C_17	1.00	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.92	0.92	0.92	1.00	1.00	0.00	0.92	1.00	1.00
C_18	1.00	1.00	0.92	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.92	0.92	0.92	1.00	1.00	0.92	0.00	1.00	1.00
C_19	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.85	1.00	1.00	1.00	0.92	1.00	1.00	1.00	1.00	1.00	0.00	0.92
C_20	1.00	1.00	0.92	1.00	1.00	0.92	1.00	1.00	0.92	1.00	0.92	0.92	1.00	1.00	1.00	1.00	1.00	1.00	0.92	0.00

## Mahalanobis distance

- Introduced by P. C. Mahalanobis in 1936.

The Mahalanobis distance of an observation  $x = (x_1, x_2, x_3, \dots, x_N)^T$  from a set of observations with mean  $\mu = (\mu_1, \mu_2, \mu_3, \dots, \mu_N)^T$  Covariance matrix S is defined as

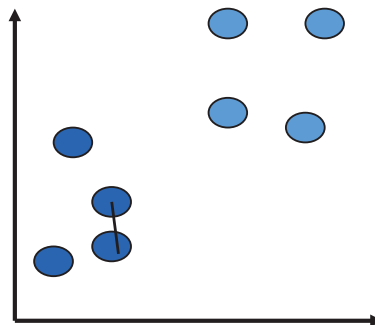
$$D_M(x) = \sqrt{(x - \mu)^T S^{-1} (x - \mu)}$$

### Amalgamation technique or Linkage Rules

- At first, each unit is taken as one cluster.
- Based on various principles the basic clusters are linked together i.e. more number of units are added to the basic clusters.

# Single Linkage

- The minimum of all pairwise distances between points in the two clusters
- Tends to produce long, “loose” clusters



## How Single linkage works?

### Example data set

Subject ID	A	B
S1	5	5
S2	6	6
S3	15	14
S4	16	15
S5	25	20
S6	30	19

**Euclidean Distance (ED) between objects S1 (5, 5) and S2 (6, 6).**

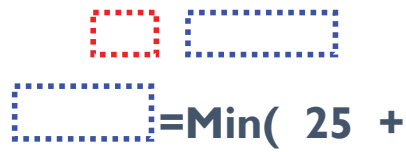
$$ED = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

$$= \sqrt{(6 - 5)^2 + (6 - 5)^2}$$

$$= 1.41$$

	S1	S2	S3	S4	S5	S6
S1	0	1.4142	13.4536	14.8660	25	28.6530
S2	1.4142	0	12.0415	13.4536	23.6008	27.2940
S3	13.4536	12.0415	0	1.4142	11.6619	15.8114
S4	14.8660	13.4536	1.4142	0	10.2956	14.560
S5	25	23.6008	11.6619	10.2956	0	5.0990
S6	28.6530	27.2940	15.8114	14.560	5.0990	0

Distance between  $S_5$  and  $\{S1, S2\} = \text{Min}(d_{(5,1)} + d_{(5,2)})$



23.6008 )

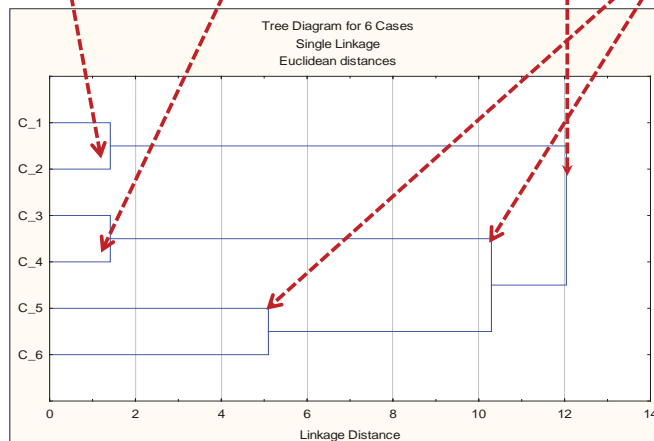
	S1	S2	S3	S4	S5	S6
S1	0	1.4142	13.4536	14.8660	25	28.6530
S2	1.4142	0	12.0415	13.4536	23.6008	27.2940
S3	13.4536	12.0415	0	1.4142	11.6619	15.8114
S4	14.8660	13.4536	1.4142	0	10.2956	14.560
S5	25	23.6008	11.6619	10.2956	0	5.0990
S6	28.6530	27.2940	15.8114	14.560	5.0990	0

After merging objects S1 and S2 then S3 and S4      After merging objects S5 and S6

	{S1, S2}	{S3, S4}	S5	S6		(S1,S2)	(S3,S4)	(S5,S6)	
{S1,S2}	0	13.4536	24.3.608	27.294	Min	(S1,S2)	0	13.4536	26.1369
{S3,S4}	13.4536	0	10.2956	14.560		(S3,S4)	13.4536	0	10.2956
S5	23.608	10.2956	0	5.0990		(S5,S6)	23.608	10.2956	0
S6	27.294	14.560	5.0990	0					

Contd

	(S1,S2)	(S3,S4) and (S5,S6)
(S1,S2)	0	13.4536
(S3,S4) and (S5,S6)	13.4536	0



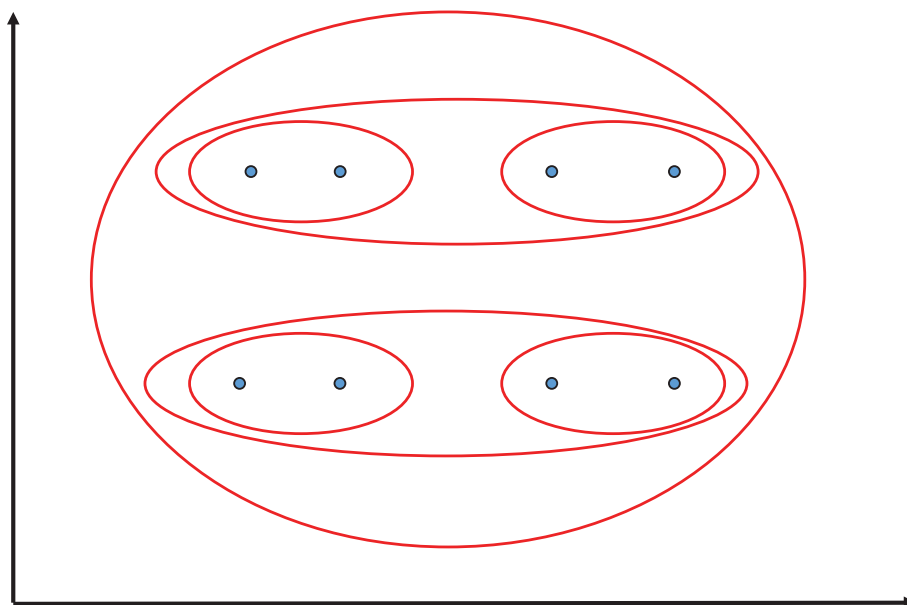
Dendrogram produced by the hierarchical clustering algorithm

- Clustering methods differ only with regard to the computation of the inter-cluster distances with the remaining objects during every step of the merger.
- If  $C_1$  is the initial Cluster with objects (1, 2), then the distance of  $C_1$  with the remaining objects "j" ( $j \neq 1, 2: 3, \dots, N$ ) is computed as follows:

Single Linkage	Min ( $d_{1j}, d_{2j}$ )
Complete Linkage	Max ( $d_{1j}, d_{2j}$ )
Average Linkage	Average ( $d_{1j}, d_{2j}$ )

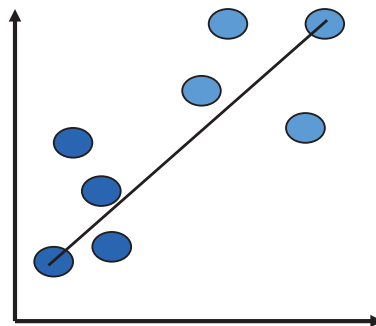
- In Ward's Minimum Variance Method, the merger during each step is based on relatively **Minimum Variance** of the distances among the objects.

## Single Linkage Example

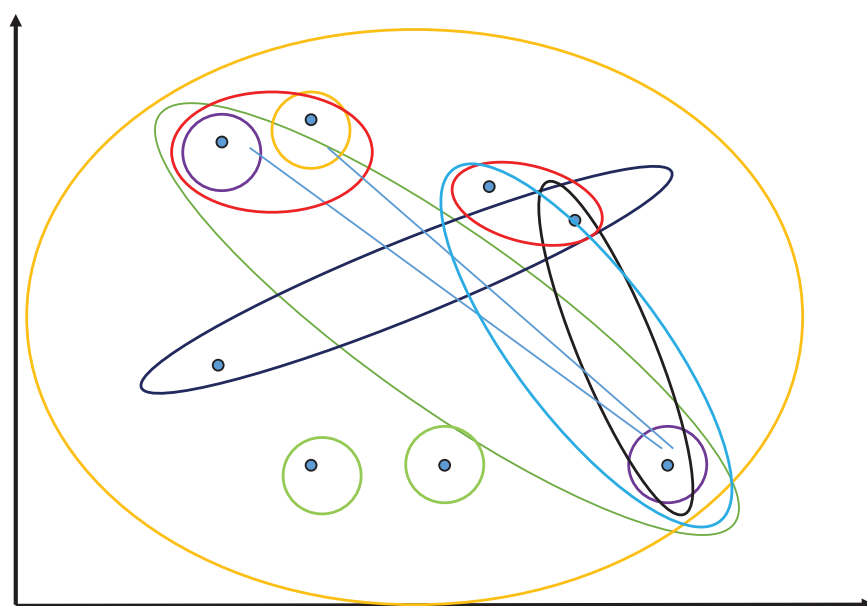


## Complete Linkage

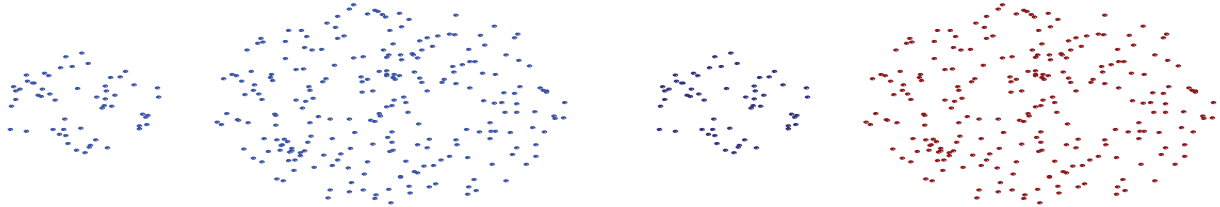
- The maximum of all pairwise distances between points in the two clusters
- Tends to produce very tight clusters



## Complete Linkage Example



## single-linkage clustering

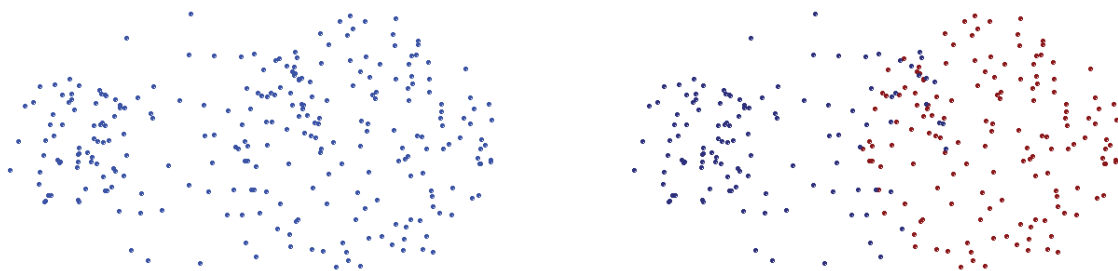


**Original Points**

**Two Clusters**

- Can handle non-elliptical shapes

## single-linkage clustering

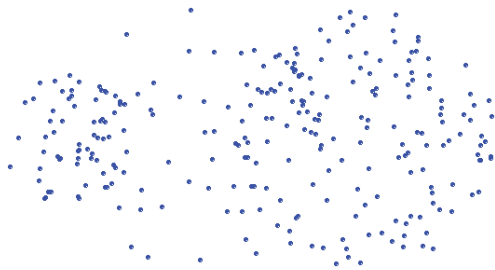


**Original Points**

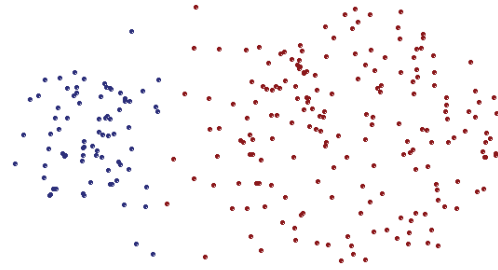
**Two Clusters**

- Sensitive to noise and outliers
- It produces long, elongated clusters

## complete-linkage clustering



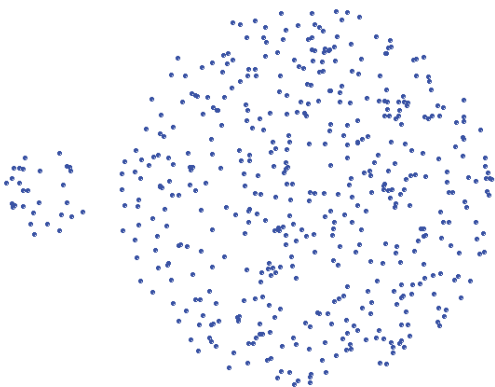
**Original Points**



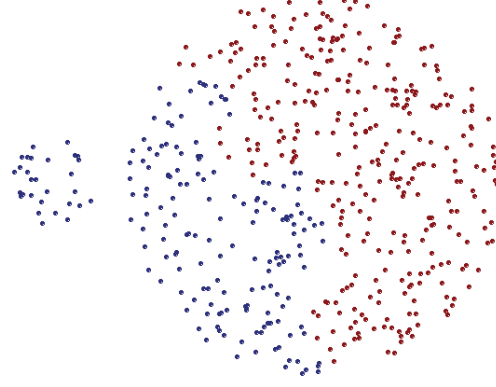
**Two Clusters**

- More balanced clusters (with equal diameter)
- Less susceptible to noise

## Complete-linkage clustering



**Original Points**



**Two Clusters**

- Tends to break large clusters
- All clusters tend to have the same diameter – small clusters are merged with larger ones



### Unweighted pair-group average (UPGMA)

- ❖ The average distance between all pairs of objects in the two different clusters is the distance between two clusters.
- ❖ Very efficient when the objects form natural distinct "clumps,"

### Weighted pair-group average(WPGMA).

- ❖ This method is one step ahead of the UPGMA method through the incorporation of size of the respective cluster as the weight.
- ❖ As such this method is very useful when we are expecting variable cluster sizes.

### Unweighted pair-group centroid

The distance between two clusters is determined as the difference between centroids

### Weighted pair-group centroid

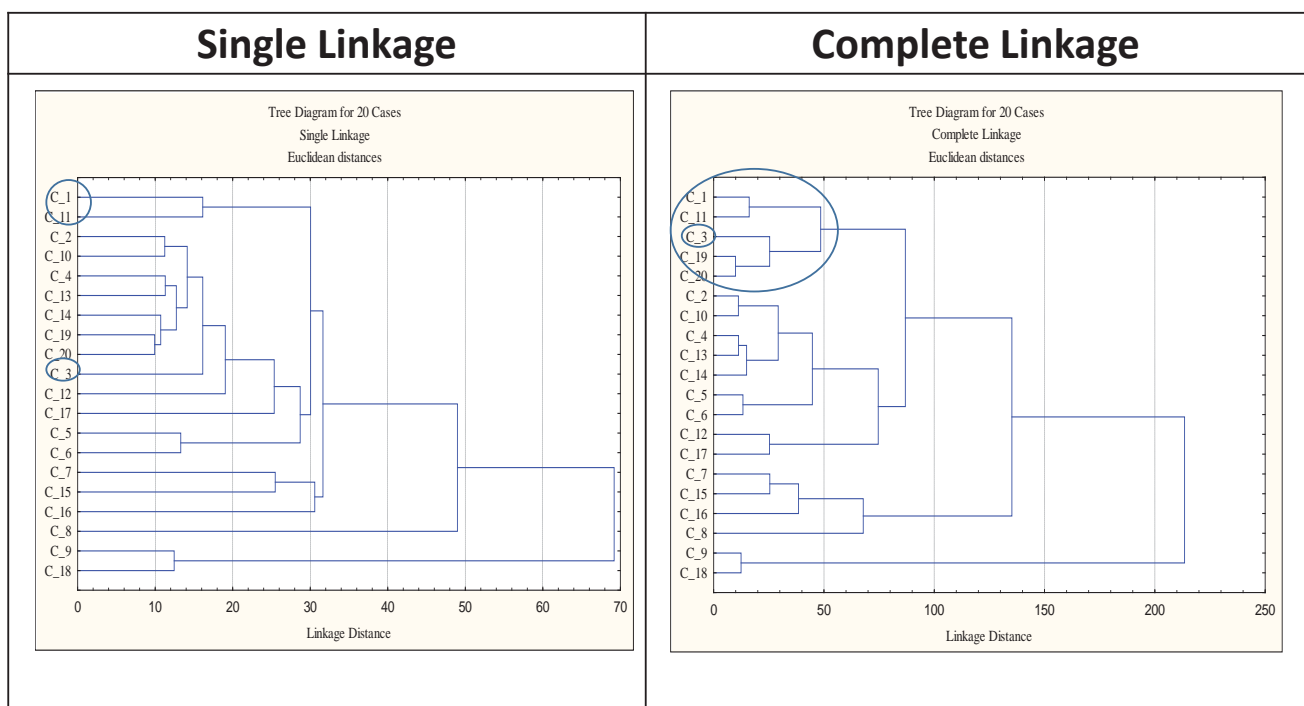
This method is identical to the previous one, except that weighting is introduced into the computations to take into consideration differences in cluster sizes (i.e., the number of objects contained in them).

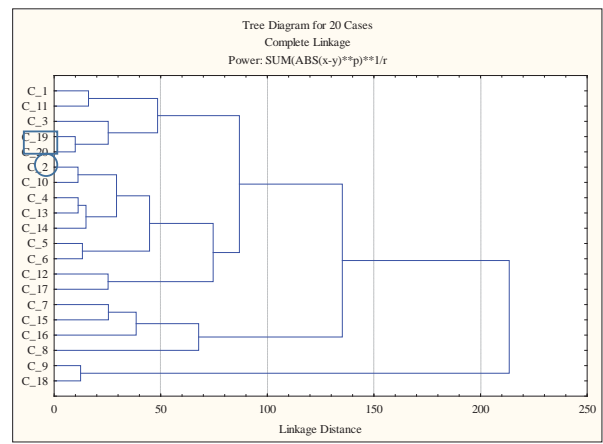
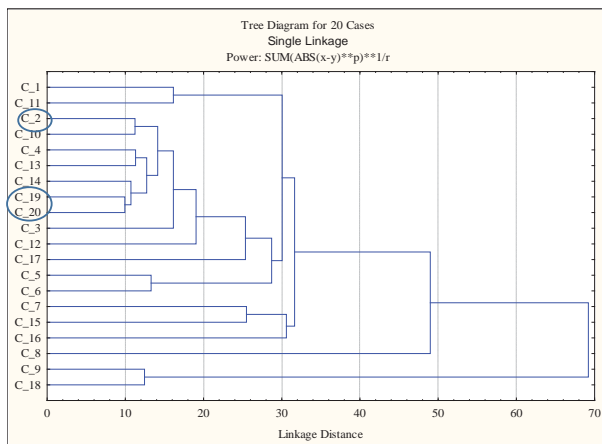
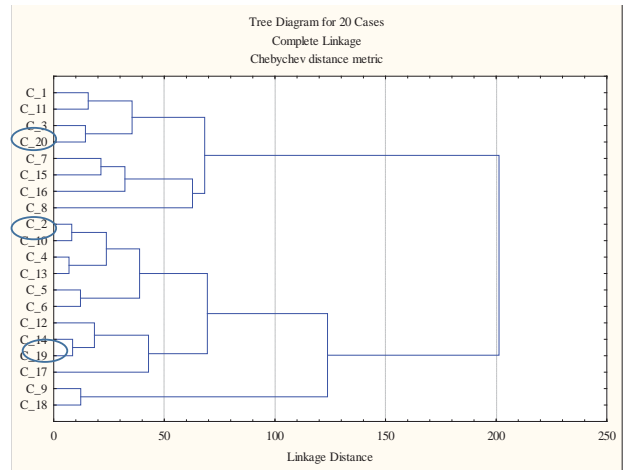
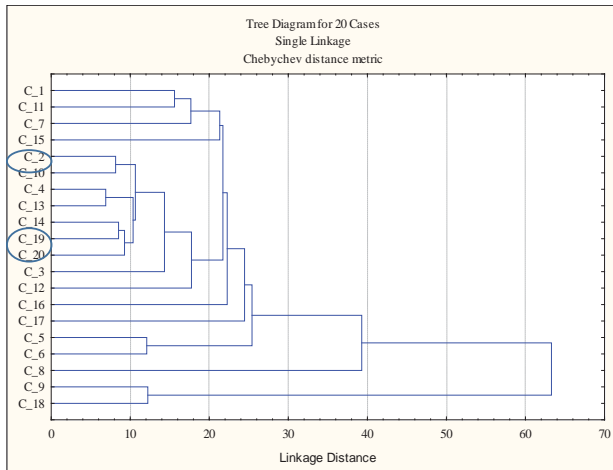
Thus, when there are (or one suspects there to be) considerable differences in cluster sizes, this method is preferable to the previous one.

## Ward's method

- ✓ It uses Analysis of Variance approach to evaluate the distances between clusters.
- ✓ This method attempts to minimize the Sum of Squares (SS) of any two (hypothetical) clusters that can be formed at each step.
- ✓ Though this method is very efficient, it has tendency to produce small clusters.

## Comparison of Linkage method





## **Association between the set of Macro and Micro climatic parameters with the set of Crop growth parameters of wheat crop following “Monte Carlo simulation” technique: redundancy analysis (RDA)**

**G Sathish, Minakshi Mishra and Prof D. Mazumdar**  
**Department of Agricultural Statistics, BCKV, Nadia- 741252 (W.B)**

RDA is a multivariate analysis technique for two sets of variables. The most practical situation is the one in which these are sets of explanatory (x) and response (y) variables. RDA of the y-set on x-set can be considered a set of simultaneous regression equations of the variables in the y-set on those in the x-set, thereby using only a small number of linear combinations of the x-variables. Therefore, the technique is a form of reduced rank regression. It can also be considered a principal component analysis of the projections of the y variables on the space spanned by the x variables.

De Leeuw (1987), who described the history of RDA, mentioned Kelley (1940) as the one who introduced RDA. RDA is a method to extract and summarise the variation in a set of response variables that can be explained by a set of explanatory variables. More accurately, RDA is a direct gradient analysis technique which summarises linear relationships between components of response variables that are "redundant" with (i.e. "explained" by) a set of explanatory variables. All multivariate analyses were performed by using the software CANOCO, Version 4.5 for Windows (TerBraak, 1989).

The results of RDA were visualized in the form of ordination diagrams in the Canodraw for Windows program. Variables are represented as symbol such as lines with arrows pointing in the direction of maximal variation. Variables with lines close to each other and headed in the same (opposite) direction are highly positively (negatively) correlated. Two lines at a 90-degree angle indicate that the corresponding variables are uncorrelated. The inclusive forward selection procedure was employed for sorting out the factors explaining the most variance in the Y data and then, Monte Carlo test with 499 permutations was carried out for significance testing of the selected X factors.

**Macro-climatic parameters:** Maximum temperature (max. temp/ Tmax), Minimum temperature (min. temp/ Tmin), Relative humidity morning (RH-I), Relative humidity evening (RH-II) and Rainfall (Rf).

**Micro-climatic parameters:** Absorbed photosynthetic active radiation (APAR), Reflected photosynthetic active radiation (RPAR) and Transmitted photosynthetic active radiation (TPAR).

**Growth processes parameters:** Leaf area index (LAI), plant height (Plntht) and Yield (Y).

## Some Problem-aspects Associated with Regression Analysis

**Kiranmoy Das**

**Assistant Professor**

**Interdisciplinary Statistical Research Unit,  
Indian Statistical Institute (ISI), Kolkata, India.**

**[email:kiranmoy.das@gmail.com](mailto:kiranmoy.das@gmail.com)**

### ABSTRACT OF THE TALK

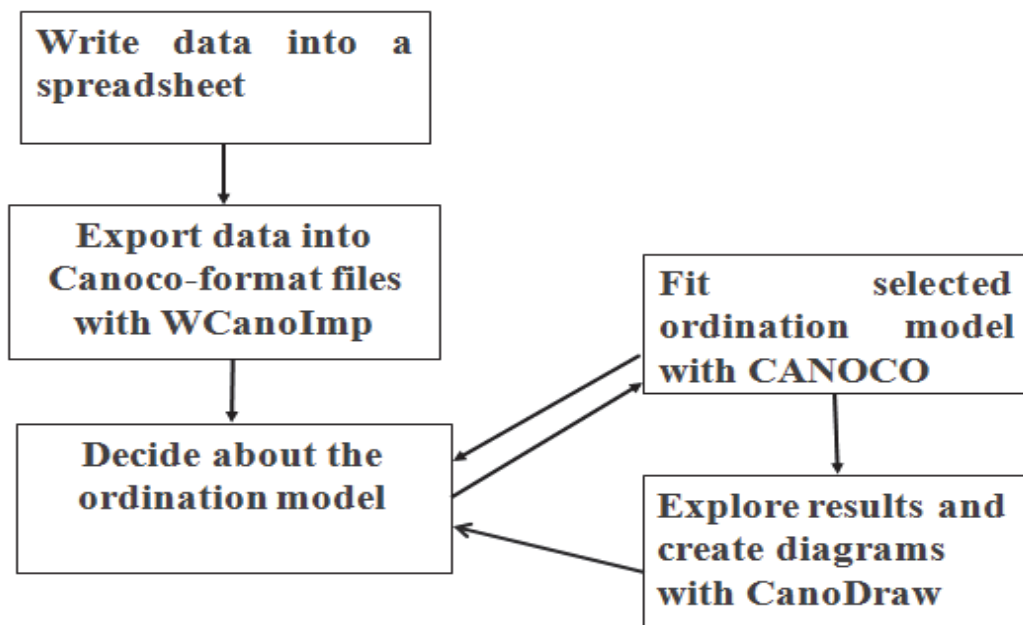
I will focus on two important aspects of the regression analysis which are not addressed typically in the usual class lectures. The first issue is the endogeneity which occurs when the errors are correlated with one or more predictors. This is quite common in the simultaneous equations models. The two-stage regression technique based on two-stage least squares method is the treatment for this issue. We will briefly go through two-stage regression techniques with examples in agriculture, biology, and many other related disciplines. In the presence of endogeneity, the ordinary least squares result in the inconsistent parameter estimates and hence inconsistent scientific inference.

Second issue is the incomplete data problem which might be caused by truncation and/or censorship through some selection mechanism. This also includes the zero-inflated data. In many real applications, we observe a weighty proportion of zeros in the response as well as some of the predictor variables. Since there is a hidden mechanism of producing zero values, it is extremely important to model this with proper adjustment. We will go through some standard methods for handling the zero inflation, e.g. Tobit model, Two-part model, latent variable model etc. The zero inflation is a special case of truncation and/or censorship and under such situations again the ordinary least squares result in biased estimates and hence inconsistent inference. We will demonstrate the two-stage least squares under such settings and the estimation method.

We will also go through some standard methods for imputing the missing data under a regression setting. Incomplete data results in biased estimates and hence under some model assumptions, the missing values can be imputed for a better and powerful inference. Under different missingness pattern, e.g. missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR), we will go through the imputation methods based on regression. If time permits, the more advanced multiple imputation (MI) techniques will also be discussed. MI essentially imputed the missing set of values multiple times and then for each set of the imputed values, estimates the underlying model parameters. Finally, the estimates are averaged over all imputations and thus the final estimates of the model parameters are obtained with the estimates of the corresponding standard errors. The usefulness of the multiple imputations compared to the single imputation methods will also be discussed with appropriate examples.

All our theoretical methods and techniques will be illustrated by real examples from various disciplines, majorly from the agricultural sciences as per the background of the participants. Some statistical packages will be shown such that the participants can implement the computations as per their requirements.

A simplified flow chart for RDA using CANOCO 4.5 software



An algorithm for RDA using CANOCO 4.5 software

1. Data variable analysis: Dependent & Independent variables
2. Extract patterns from the explained variation only (Direct gradient analysis)
3. Define both data file for dependent and independent sets of variables
4. Choose RDA as linear direct gradient analysis
5. Focus scaling on: inter sample distance
6. Dependent variables : all divide by standard deviation
7. No transformation
8. Samples all standardized by norms
9. Dependent variables actually standardized by error variance
10. Manual selection as forward selection of independent variables
11. Best K= Total number of independent variables
12. Use Monte Carlo permutation tests under full model.

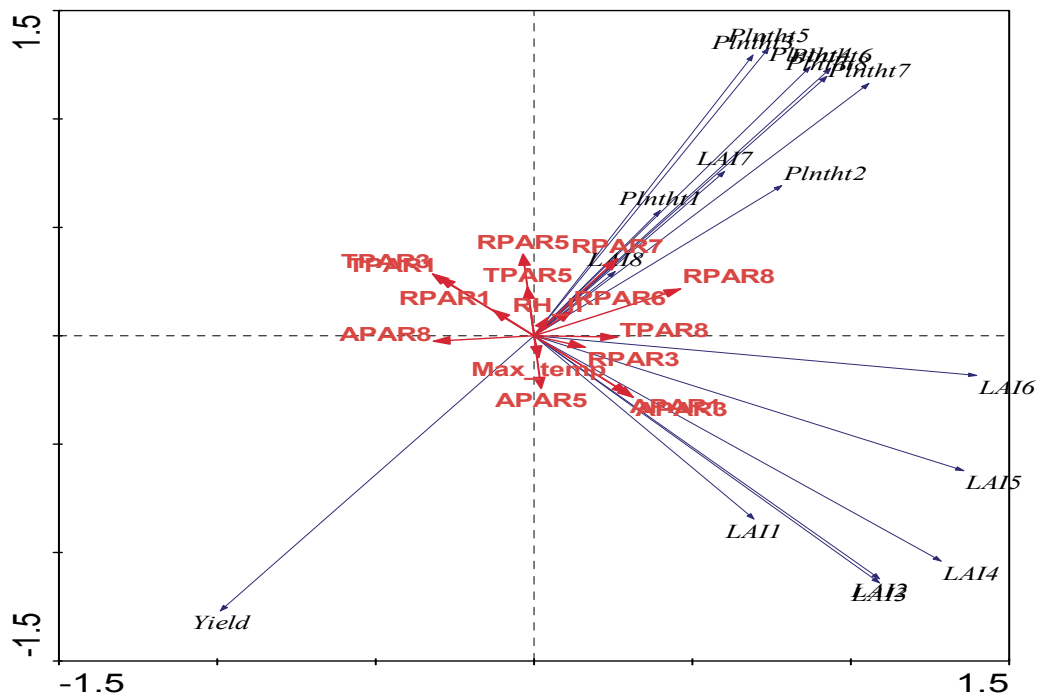
No. of permutations: 499, Set seeds: 23239, Randomize: 945, Unrestricted permutations.

**RESULT**

The inflation factor of selected meteorological parameters (macro and micro climatic), X-set was below 5 for wheat crop (Table 1). Association of meteorological parameters and growth parameters including yield of wheat crop were shown in biplot figure (Figure 1).

**Table 1 Forward selection of X set (micro and macro climatic) variables for wheat crop**

Selected variables name	Mean	Standard Deviation	Inflation factor
APAR of Week 1	51.81	19.95	4.06
APAR of Week 3	63.99	15.91	4.77
APAR of Week 5	83.52	9.81	3.71
APAR of Week 8	88.68	3.91	1.59
RPAR of Week 1	5.07	1.16	2.56
RPAR of Week 3	4.63	0.61	1.48
RPAR of Week 5	4.19	0.68	2.94
RPAR of Week 6	3.57	0.45	1.66
RPAR of Week 7	3.66	0.46	4.12
RPAR of Week 8	3.96	0.57	2.55
TPAR of Week 1	43.11	19.22	0.00
TPAR of Week 3	31.39	15.87	0.00
TPAR of Week 5	12.29	9.49	0.00
TPAR of Week 8	7.36	3.71	0.00
Max. temp	28.06	3.71	1.71
RH-II	49.60	10.63	2.44



**Figure 1** Bi-plot showing RDA results for association between the set of meteorological parameters (macro and micro climatic) with the growth parameters including yield of wheat crop (APAR1 to 8: APAR of week 1 to week 8, RPAR1 to 8: RPAR of week 1 to 8, TPAR1 to 8: TPAR of week 1 to 8, LAI1 to 8: LAI of week 1 to week 8, Plntht1 to 8: Plntht of week 1 to 8)

The first axis revealed that the LAI and Plant height of week 1 to week 8 along with APAR of week 1, week 3, and week 5, RPAR of week 3, week 6, week 7 and week 8, TPAR of week 8, maximum temperature and RH-II don't contribute to yield but on the second axis, it was revealed that the LAI of week 1 to week 6 along with APAR of week 1, week 3, week 5 and week 8, RPAR of week 3, TPAR of week 8 and maximum temperature were important variables to promote yield (Figure 1).

**Table 2** The RDA results for wheat crop of X set (micro and macro parameters) and Y set (growth parameters) on axis 1-4

Axes	1	2	3	4
Eigen values	0.372	0.249	0.102	0.037
X - Y correlations	0.895	0.91	0.925	0.882
Cumulative percentage variance				
of Y data	37.2	62.1	72.3	76.0
of X - Y relation	46.8	78.1	90.9	95.6
Sum of all eigen values	1			
Sum of all canonical eigen values	0.795			



Table 2 indicated that 62.1% of the total variance of growth parameters and 78.1% of the macro-micro meteorological parameters along with growth parameters were explained by the first two canonical associations represented by first two axes.

### References

De Leeuw J. 1987. On the history of redundancy analysis. Internal publication, Department of Data theory, University of Leiden.

Kelley T L. 1940. Talents and tasks: their conjunction in a democracy of wholesome living and national defense. Harvard Education Papers no.1, Graduate School of Education, Harvard University.

TerBraak C J F 1989. CANOCO- An extension of DECORANA to analyze species- environment relationships. Hydrobiologia 184 (3): 169-170.

## Linear regression using SAS

By

Dr(Mrs) B.Bhattacharyya, BCKV

Regression analysis is the art of fitting straight lines to patterns of data. In a linear regression model, the variable of interest (the so-called “dependent” variable) is predicted from  $k$  other variables (the so-called “independent” variables) using a linear equation.

Regression analysis discovers the relationship between dependent and explanatory variables. More specifically, this statistical relationship rather speaks of some significant association in the data. Linear regression attempts to draw a line that comes closest to the data by finding the slope and intercept that define the line and minimize regression errors. However, many relationships in data do not follow a straight line, in which cases statisticians use nonlinear regression instead.

Quite often it happens that a dependent variable is not explained by single variable. In these cases, multiple regression analysis is done, which attempts to explain dependent variable as a function of more than one independent variable. Multiple regressions can also be linear and nonlinear. Multiple linear regression is one of the statistical tools used for discovering relationships between variables. It is used to find the linear model that best predicts the dependent variable from the independent variables. A data set with  $p$  independent variables has  $2^p$  possible subset models to consider since each of the  $p$  variables is either included or excluded from the model, not counting interaction terms.

The model can be expressed as

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + \epsilon_i$$

where  $k$  is the number of independent variables, the betas are constants and the epsilons are independent and identically distributed (i.i.d.) normal random variables with mean zero.  $Y_i$ , the dependent variable for case  $i$ ,  $X_{1i}$ 's the independent variables for case  $i$ ,  $\beta_0$ , the intercept (if all  $x$ 's are zero, the expected value of  $y$  is  $\beta_0$ ) and  $\beta_j$  being the slope of a linear regression.

### Estimates of the Model Parameters

- The estimates of the  $\beta$  coefficients are the values that minimize the sum of squared errors for the sample.
- $b$  represents a sample estimate of a  $\beta$  coefficient.

$$\frac{SSE}{n - p}$$

- $MSE = \frac{SSE}{n - p}$  estimates  $\sigma^2$ , the variance of the errors. In the formula,  $n$  = sample size,  $p$  = number of  $\beta$  coefficients in the model (including the intercept) and  $SSE$  = sum of squared

errors. Notice that for simple linear regression  $p = 2$ .

- $S = \sqrt{\text{MSE}}$  estimates  $\sigma$  and is known as the *regression standard error* or the *residual standard error*.
- In the case of two predictors, the estimated regression equation yields a plane (as opposed to a line in the simple linear regression setting). For more than two predictors, the estimated regression equation yields a hyperplane.

### Interpretation of the Model Parameters

- Each  $\beta$  coefficient represents the change in the mean response,  $E(y)$ , per unit increase in the associated predictor variable when all the other predictors are held constant.
- For example,  $\beta_1$  represents the change in the mean response,  $E(y)$ , per unit increase in  $x_1$  when  $x_2, x_3, \dots, x_{p-1}$  are constants.
- The intercept term,  $\beta_0$ , represents the mean response,  $E(y)$ , when all the predictors  $x_1, x_2, x_3, \dots, x_{p-1}$ , are all zero (which may or may not have any practical meaning).

### Predicted Values and Residuals

A **predicted value** is calculated as  $\hat{y}_i = b_0 + b_1x_{i,1} + b_2x_{i,2} + \dots + b_{p-1}x_{i,p-1}$ , where the  $b$  values come from statistical software and the  $x$ -values are specified by us.

A **residual (error)** term is calculated as  $e_i = y_i - \hat{y}_i$  the difference between an actual and a predicted value of  $y$ .

A **plot of residuals versus predicted values** ideally should resemble a horizontal random band. Departures from this form indicates difficulties with the model and/or data.

Other residual analyses can be done exactly as we did in simple regression. For instance, we might wish to examine a normal probability plot (NPP) of the residuals. Additional plots to consider are plots of residuals versus each  $x$ -variable separately. This might help us identify sources of curvature or nonconstant variance.

### ANOVA Table

Source	df	SS	MS	F
Regression	$p - 1$	SSR	$\text{MSR} = \text{SSR} / (p - 1)$	$\text{MSR} / \text{MSE}$
Error	$n - p$	SSE	$\text{MSE} = \text{SSE} / (n - p)$	
Total	$n - 1$	SSTO		

### Coefficient of Determination, R-squared, and Adjusted R-squared

As in simple linear regression,  $R^2 = SSR/SSTO = 1 - SSE/SSTO$ , and represents the proportion of variation in  $y$  (about its mean) "explained" by the multiple linear regression model with predictors,  $x_1, x_2, \dots$ .

If we start with a simple linear regression model with one predictor variable,  $x_1$ , then add a second predictor variable,  $x_2$ , SSE will decrease (or stay the same) while SSTO remains constant, and so  $R^2$  will increase (or stay the same). In other words,  $R^2$  always increases (or stays the same) as more predictors are added to a multiple linear regression model, *even if the predictors added are unrelated to the response variable*. Thus, by itself,  $R^2$  may not be useful to identify which predictors should be included in a model and which should be excluded.

An alternative measure, adjusted  $R^2$ , does not necessarily increase as more predictors are added, and can be used to identify which predictors should be included in a model and which should be excluded. Adjusted  $R^2 =$

$$1 - \left( \frac{n-1}{n-p} \right) (1 - R^2),$$

and while it has no practical interpretation, is useful for such model building purposes.

Simply stated, when comparing two models used to predict the same response variable, we generally prefer the model with the higher value of adjusted  $R^2$ .

### Significance Testing of Each Variable

Within a multiple regression model, it is very important to know whether a particular  $x$ -variable is making a useful contribution to the model. That is, given the presence of the other  $x$ -variables in the model, does a particular  $x$ -variable help us predict or explain the  $y$ -variable? For instance, suppose that we have three  $x$ -variables in the model. The general structure of the model could be

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon.$$

As an example, to determine whether variable  $x_1$  is a useful predictor variable in this model, we could test

$$H_0: \beta_1 = 0$$

$$H_A: \beta_1 \neq 0.$$

If the null hypothesis above were the case, then a change in the value of  $x_1$  would not change  $y$ , so  $y$  and  $x_1$  are not linearly related. Also, we would still be left with variables  $x_2$  and  $x_3$  being present in the model. When we cannot reject the null hypothesis above, we should say that we do not need variable  $x_1$  in the model given that variables  $x_2$  and  $x_3$  will remain in the model. In general, the interpretation of a slope in multiple regression can be tricky. Correlations among the predictors can change the slope values dramatically from what they would be in separate simple regressions.

To carry out the test, statistical software will report  $p$ -values for all coefficients in the model. Each  $p$ -value will be based on a  $t$ -statistic calculated as

$$t^* = (\text{sample coefficient} - \text{hypothesized value}) / \text{standard error of coefficient}.$$

For our example above, the  $t$ -statistic is:

$$t^* = \frac{b_1 - 0}{\text{se}(b_1)} = \frac{b_1}{\text{se}(b_1)}.$$

Note that the hypothesized value is usually just 0, so this portion of the formula is often omitted.

### Regression Diagnostics

- Linearity - the relationships between the predictors and the outcome variable should be linear
- Normality - the errors should be normally distributed - technically normality is necessary only for the  $t$ -tests to be valid, estimation of the coefficients only requires that the errors be identically and independently distributed
- Homogeneity of variance (homoscedasticity) - the error variance should be constant
- Independence - the errors associated with one observation are not correlated with the errors of any other observation
- Model specification - the model should be properly specified (including all relevant variables, and excluding irrelevant variables)

Additionally, there are issues that can arise during the analysis that, while strictly speaking, are not assumptions of regression, are none the less, of great concern to regression analysts.

- Influence - individual observations that exert undue influence on the coefficients
- Collinearity - predictors that are highly collinear, i.e. linearly related, can cause problems in estimating the regression coefficients.

Many graphical methods and numerical tests have been developed over the years for regression diagnostics. We will explore these methods and show how to verify regression assumptions and detect potential problems using SAS.

### Testing of Data series

Linear regression model may not be directly applicable to all data sets. It is recommended that data should be plotted first of all. By examining these initial plots, it can be assessed quickly whether the data have linear relationships or interactions are present. The Nonlinearity may be detected from scatter plots or otherwise. Transformations on either the predictor variable,  $X$ , or the response variable,  $Y$ , may often be sufficient to make the linear regression model appropriate for the transformed data.

### **SAS CODES for testing of Raw Data:**

```
/*RAW DATA EXAMINATION*/ proc  
plot data=dataplot; plot Y*X1 Y*X2  
Y*X3 ; run;
```

### **Modelling**

The REG procedure can be used to build and test the assumptions of the data we propose to model.

### **Root MSE**

The RMSE needs to be small compared to other models. The value of Root MSE will be dependent on the values of the Y variable are used for modeling. As a guideline, you want the value for each of the variables in your model to have a Type III SS p-value of 0.05 or less. Other approaches to finding good models are having a small PRESS statistic (found in REG as Predicted Resid SS (Press)) or having a CP statistic of  $p-1$  where  $p$  is the number of parameters in your model. CP can also be found using PROC REG.

### **Test of Assumptions**

Plots of residuals against the predictor variable or against the fitted values are not only helpful to study whether a linear regression function is appropriate but also to examine whether the variance of the error terms are constant.

### **Tests for Normality of residuals**

The normal probability plot of the errors look linear and fall in line diagonally then the errors are normally distributed.

### **Kolmogorov-Smirnov Test:**

The **Kolmogorov–Smirnov test (K–S test or KS test)** is a nonparametric test of the equality of continuous, one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution (one-sample K–S test), or to compare two samples (two-sample K–S test).

### **Shapiro-Wilks Test**

The Shapiro-Wilks test for normality is one of three general normality tests designed to detect all departures from normality. It is comparable in power to the other two tests. The test rejects the hypothesis of normality when the p-value is less than or equal to 0.05. Failing the normality test allows you to state with 95% confidence the data does not fit the normal distribution. Passing the normality test only allows you to state no significant departure from normality was found.

### **Testing for Autocorrelation**

Autocorrelation is when an error term is related to a previous error term. This situation can happen with time series data such as monthly sales. The Durbin-Watson statistic can be used to check if autocorrelation exist. The Durbin-Watson statistic is calculated by using the DW option in REG. The Durbin-Watson statistic test for first order correlation of error terms. The Durbin Watson statistic ranges from 0 to 4.0. Generally a D-W statistic of

2.0 indicates the data are independent.

Procreg;

model Y = X1 X2 X3;

Output out=resdat r=resid p=pred;

Data check;

set resdat;

Procunivariate normal plot; varresid;

Title 'Test of Normality of Residuals';

Run;

### **The PROC REG provides nine methods of model selection**

NONE stands for no selection. This method is the default and uses the full model given in the MODEL statement to fit the linear regression.

FORWARD stands for forward selection. This method starts with no variables in the model and adds variables one by one to the model. At each step, the variable added is the one that maximizes the fit of the model. You can also specify groups of variables to treat as a unit during the selection process. An option enables you to specify the criterion for inclusion.

BACKWARD stands for backward elimination. This method starts with a full model and eliminates variables one by one from the model. At each step, the variable with the smallest contribution to the model is deleted. You can also specify groups of variables to treat as a unit during the selection process. An option enables you to specify the criterion for exclusion.

STEPWISE stands for stepwise regression, forward and backward. This method is a modification of the forward-selection method in that variables already in the model do not necessarily stay there. You can also specify groups of variables to treat as a unit during the selection process. Again, options enable you to specify criteria for entry into the model and for remaining in the model.

MAXR stands for maximum  $R^2$  improvement. This method tries to find the best one-variable model, the best two-variable model, and so on. The MAXR method differs from the

STEPWISE method in that many more models are evaluated with MAXR, which considers all switches before making any switch. The STEPWISE method may remove the "worst" variable without considering what the "best" remaining variable might accomplish, whereas

MAXR would consider what the "best" remaining variable might accomplish. Consequently,

MAXR typically takes much longer to run than STEPWISE.

MINR stands for minimum  $R^2$  improvement. This method closely resembles MAXR, but the switch chosen is the one that produces the smallest increase in  $R^2$ .

RSQUARE finds a specified number of models having the highest  $R^2$  in each of a range of model sizes.

CP finds a specified number of models with the lowest  $C_p$  within a range of model sizes.

ADJRSQ finds a specified number of models having the highest adjusted  $R^2$  within a range of model sizes.

**Some Keywords of PROC REG**

<b>Keyword</b>	<b>Statistic</b>
COOKD.	Cook's D influence statistics
COVRATIO.	standard influence of observation on covariance of betas
DFFITS.	standard influence of observation on predicted value
H.	leverage
LCL.	lower bound of 100(1-a)% confidence interval for individual prediction
LCLM.	lower bound of 100(1-a)% confidence interval for the mean of the dependent variable
PREDICTED.   PRED.   P.	predicted values
PRESS.	residuals from refitting the model with current observation deleted
RESIDUAL.   R.	residuals

STDI.	standard error of the individual predicted value
STDP.	standard error of the mean predicted value
STDR.	standard error of the residual
STUDENT.	residuals divided by their standard errors
UCL.	upper bound of 100(1-a)% confidence interval for individual prediction
UCLM.	upper bound of 100(1-a)% confidence interval for the mean of the dependent variables



### **Collinearity**

First, look at multicollinearity from a conventional viewpoint. The absence of multi-collinearity is essential to a multiple regression model. In regression when several predictors (regressors) are highly correlated, this problem is called multi-collinearity or collinearity. When things are related, we say they are linearly dependent on each other because you can nicely fit a straight regression line to pass through many data points of those variables. Collinearity simply means co-dependence. Collinearity is problematic when one's purpose is explanation rather than mere prediction. Collinearity makes it more difficult to achieve significance of the collinear parameters. But if such estimates are statistically significant, they are as reliable as any other variables in a model. And even if they are not significant, the sum of the coefficient is likely to be reliable. In this case, increasing the sample size is a viable remedy for collinearity when prediction instead of explanation is the goal. However, if the goal is explanation, measures other than increasing the sample size are needed.

### **Unusual and influential data**

A single observation that is substantially different from all other observations can make a large difference in the results of your regression analysis. If a single observation (or small group of observations) substantially changes your results, you would want to know about this and investigate further. There are three ways that an observation can be unusual.

**Outliers:** In linear regression, an outlier is an observation with large residual. In other words, it is an observation whose dependent-variable value is unusual given its values on the predictor variables. An outlier may indicate a sample peculiarity or may indicate a data entry error or other problem.

**Leverage:** An observation with an extreme value on a predictor variable is called a point with high leverage. Leverage is a measure of how far an observation deviates from the mean of that variable. These leverage points can have an effect on the estimate of regression coefficients.

**Influence:** An observation is said to be influential if removing the observation substantially changes the estimate of coefficients. Influence can be thought of as the product of leverage and outlierness.

### **Illustrative example :**

**Example:** The following data was collected through a pilot sample survey on Hybrid Jowar crop on yield and biometrical characters. The biometrical characters were average Plant Population (PP), average Plant Height (PH), average Number of Green Leaves (NGL) and Yield (kg/plot).

S.No.	PP	PH	NGL	Yield	S.No.	PP	PH	NGL	Yield
1	142.00	0.525	8.2	2.470	24	55.55	0.265	5.0	0.430
2	143.00	0.640	9.5	4.760	25	88.44	0.980	5.0	4.080
3	107.00	0.660	9.3	3.310	26	99.55	0.645	9.6	2.830
4	78.00	0.660	7.5	1.970	27	63.99	0.635	5.6	2.570
5	100.00	0.460	5.9	1.340	28	101.77	0.290	8.2	7.420
6	86.50	0.345	6.4	1.140	29	138.66	0.720	9.9	2.620
7	103.50	0.860	6.4	1.500	30	90.22	0.630	8.4	2.000
8	155.99	0.330	7.5	2.030	31	76.92	1.250	7.3	1.990
9	80.88	0.285	8.4	2.540	32	126.22	0.580	6.9	1.360
10	109.77	0.590	10.6	4.900	33	80.36	0.605	6.8	0.680
11	61.77	0.265	8.3	2.910	34	150.23	1.190	8.8	5.360
12	79.11	0.660	11.6	2.760	35	56.50	0.355	9.7	2.120
13	155.99	0.420	8.1	0.590	36	136.00	0.590	10.2	4.160
14	61.81	0.340	9.4	0.840	37	144.50	0.610	9.8	3.120
15	74.50	0.630	8.4	3.870	38	157.33	0.605	8.8	2.070
16	97.00	0.705	7.2	4.470	39	91.99	0.380	7.7	1.170
17	93.14	0.680	6.4	3.310	40	121.50	0.550	7.7	3.620
18	37.43	0.665	8.4	1.570	41	64.50	0.320	5.7	0.670
19	36.44	0.275	7.4	0.530	42	116.00	0.455	6.8	3.050
20	51.00	0.280	7.4	1.150	43	77.50	0.720	11.8	1.700
21	104.00	0.280	9.8	1.080	44	70.43	0.625	10.0	1.550
22	49.00	0.490	4.8	1.830	45	133.77	0.535	9.3	3.280
23	54.66	0.385	5.5	0.760	46	89.99	0.490	9.8	2.690

1. Obtain correlation coefficient between each pair of the variables PP, PH, NGL and yield.
2. Obtain partial correlation between NGL and yield after removing the linear effect of PP and PH.
3. Give a scatter plot of the variable PP.
4. Fit a multiple linear regression equation by taking yield as dependent variable and biometrical characters as explanatory variables. Print the matrices used in the regression computations.

5. Test the significance of the regression coefficients and also equality of regression coefficients of (a) PP and PH (b) PH and NGL
6. Obtain the predicted values corresponding to each observation in the data set.
7. Check for the linear relationship among the biometrical characters, i.e., multi-collinearity in the data.
8. Fit the multiple linear regression model without intercept.

**Data Input:**

For performing analysis, input the data in the following format.

{Here serial number is termed as SN, plant population as PP, average plant height as PH, average number of green leaves as (NGL) and yield as YLD. It may, however, be noted that one can retain the same name or can code in any other fashion}.

Prepare a SAS data file using

**Data** corr; /\*one can enter any other name for data\*/

**input** sn pp ph ngl yld;

**cards;**

1	142.00	0.525	8.2	2.47
2	143.00	0.64	9.5	4.76
3	107.00	0.66	9.3	3.31
4	78.00	0.66	7.5	1.97
5	100.00	0.46	5.9	1.34
6	86.50	0.345	6.4	1.14
7	103.50	0.86	6.4	1.5
8	155.99	0.33	7.5	2.03
9	80.88	0.285	8.4	2.54
10	109.77	0.59	10.6	4.9
11	61.77	0.265	8.3	2.91
12	79.11	0.66	11.6	2.76
13	155.99	0.42	8.1	0.59
14	61.81	0.34	9.4	0.84
15	74.50	0.63	8.4	3.87
16	97.00	0.705	7.2	4.47
17	93.14	0.68	6.4	3.31
18	37.43	0.665	8.4	1.57
19	36.44	0.275	7.4	0.53

20	51.00	0.28	7.4	1.15
21	104.00	0.28	9.8	1.08
22	49.00	0.49	4.8	1.83
23	54.66	0.385	5.5	0.76
24	55.55	0.265	5.0	0.43
25	88.44	0.98	5.0	4.08
26	99.55	0.645	9.6	2.83
27	63.99	0.635	5.6	2.57
28	101.77	0.29	8.2	7.42
29	138.66	0.72	9.9	2.62
30	90.22	0.63	8.4	2.00
31	76.92	1.25	7.3	1.99
32	126.22	0.58	6.9	1.36
33	80.36	0.605	6.8	0.68
34	150.23	1.19	8.8	5.36
35	56.50	0.355	9.7	2.12
36	136.00	0.59	10.2	4.16
37	144.50	0.61	9.8	3.12
38	157.33	0.605	8.8	2.07
39	91.99	0.38	7.7	1.17
40	121.50	0.55	7.7	3.62
41	64.50	0.32	5.7	0.67
42	116.00	0.455	6.8	3.05
43	77.50	0.72	11.8	1.70
44	70.43	0.625	10.0	1.55
45	133.77	0.535	9.3	3.28
46	89.99	0.49	9.8	2.69

;

/\* Obtain correlation coefficient between each pair of the variables PP, PH, NGL and yield using the following SAS statements\*/

```
proc corr;
var pp ph ngl yld;
run;
```

/\* Obtain partial correlation between NGL and yield after removing the linear effect of PP and PH by using the following SAS statements\*/

```
proc corr;  
var ngl yld;  
partial pp ph;  
run;
```

/\* Obtain the scatter plot using the following SAS statements \*/

```
proc plot;  
plot pp*yld = '*';  
/*pp=VERTICAL AXIS yld = HORIZONTAL AXIS.*/  
run;
```

/\* Fit a multiple linear regression equation by taking yield as dependent variable and biometrical characters as explanatory variables. Print the matrices used in the regression computations using the following SAS statements\*/

```
proc reg;  
model yld= pp ph ngl/p r influence vif collin xpx i;  
/* testing the significance of regression coefficients. This is also done by default in regression fitting*/  
test1: test pp=0;  
test2: test ph=0;  
test3: test ngl=0;  
*testing the equality of two regression coefficients;  
test4: test pp-ph=0;  
test4a: test pp=ph=0;  
/*test 4 tests the equality of regression coefficients of pp and ph, whereas test4a test whether regression  
coefficients of pp and ph simultaneously are significantly different from zero*/  
test5: test ph-ngl=0;  
test5a: test ph=ngl=0;  
run;  
/*
```

**p:** It calculates predicted values from the input data and the estimated model. The display includes the observation number, the ID variable (if one is specified), the actual and predicted values, and the residual. If the CLI, CLM, or R option is specified, the P option is unnecessary

**r:** Requests an analysis of the residuals. The results include everything requested by the p option plus the standard errors of the mean predicted and residual values, the studentized residual, and Cook's *D* statistic to measure the influence of each observation on the parameter estimates.

**influence:** Computes influence statistics

**vif:** Produces variance inflation factors with the parameter estimates. Variance inflation is the reciprocal of tolerance.

**collin:** produces collinearity analysis. It requests a detailed analysis of collinearity among the regressors. This includes eigenvalues, condition indices, and decomposition of the variances of the estimates with respect to each eigenvalue.

**xpx:** Displays the **X'X** crossproducts matrix for the model. The crossproducts matrix is bordered by the **X'Y** and **Y'Y** matrices.

**i:** displays sums-of-squares and crossproducts matrix. It displays the **(X'X)<sup>-1</sup>** matrix. The inverse of the crossproducts matrix is bordered by the parameter estimates and SSE matrices. \*/

/\* A regression model without intercept can be fitted by any of the following two procedures\*/

**proc reg;**

**model** yld=pp ph ngl;

**restrict** intercept=0; /\* A RESTRICT statement is used to place restrictions on the parameter estimates in the MODEL preceding it. \*/

**run;**

**proc reg;**

**model** yld=pp ph ngl/noint; /\* Use the NOINT option to fit a model without an intercept term \*/

**run;**

## ANALYSIS OF VARIANCE AND BASIC EXPERIMENTAL DESIGNS

Anurup Majumder, BCKV

### Analysis of variance:

Analysis of variance (ANOVA) is a technique for investigating how much of the variability in a set of observations could be ascribed due to different causes of variation. Or, simply, it is basically a technique of partitioning the total or overall variation into different assignable sources of variation. Further, it helps in testing whether the variation due to any particular component is significant as compared to residual variation that can occur among the observational units, using F test.

It is one of the most powerful techniques of statistical analysis which has been developed to test the hypothesis of equality of the sample mean values whether any significant difference is present or not assuming that the samples belong to a same population. The difference of two samples mean value can be done by using t test. Analysis of variance is the generalization of the comparison of several samples mean values. The idea of Analysis of variance (ANOVA) was introduced by Sir R. A. Fisher, the noted British Statistician. He used the technique in agricultural research, which was basically, the partitioning the total variance of a set of data into different components associated with different recognized sources of variation in agricultural field experiments. Later, it is found useful in experimental situations of every branch of science even in social study.

### Application:

We may take several experimental situations as examples where this technique can be applied and valid decision can be taken. Suppose an agronomist may like to see the performances of different available varieties of a crop in  $r$  number of identical experimental plots. Or, a poultry farm owner may like to see the quality and quantity of meat of chickens of different breeds under identical conditions of his farm. Or, one sociologist may want to examine the level of intelligence (IQ score) of children of different races in identical conditions of schooling. In all the abovementioned examples, the differences of the sample mean values will be tested by the technique of analysis of variance.

**Basic assumptions for analysis of variance:** In order to perform the analysis, certain basic assumptions are made about the observations and effects. Point wise assumptions are given below:

1. All effects for different sources of variation are additive in nature.
2. Experimental errors are independent.
3. Experimental errors ( $e_{ij}$ ) are independently and identically distributed as normal with mean zero and constant variance ( $\sigma^2$ );  $e_{ij} \sim \text{i.i.d. } N(0, \sigma^2)$ .

The interpretation of the analysis of variance is valid only when the above assumptions are met although slight deviations from these assumptions do not cause much harm.

### Analysis of variance of one way classified data:

If there are  $n$  number of observations grouped into  $k$  number of classes and in each group there are  $n_i$  ( $i = 1, 2, \dots, k$ ) number of observations. Let  $y_{ij}$  be the response of the  $i$ th member of the  $j$ th group. Let the response  $y_{ij}$  be the additional effect of the sources due to  $j$ th group and error effect of the study. Then we can write it in the form of a mathematical linear model as:

$$y_{ij} = \mu + \alpha_i + e_{ij}, \quad i=1, 2, \dots, k; j = 1, 2, \dots, n_i,$$

where  $y_{ij}$  is the response of the  $j$ th individual unit belonging to the  $i$ th category or group,  $\mu$  is overall population mean,  $\alpha_i$  is the effect of being in the  $i$ th group and  $e_{ij}$  is a random error which follows normal distribution, attached to the  $(ij)$ th observation. This constitutes a one-way analysis of variance model which can be expanded further by adding more and more effects as applicable to a particular situation. When more than one known source of variation is involved, the model is referred as multi-way analysis of variance model.

The next stage is to estimate the effect of main source of variation i.e.,  $\alpha_i$ .

For estimating  $\alpha_i$ , the normal equations are obtained by method of least squares.

Let  $E = \sum_{ij} e_{ij}^2 = \sum_{ij} (y_{ij} - \mu - \alpha_i)^2$ , then the normal equations are obtained by taking the partial

differentiation of  $E$  with respect to unknown parameters  $\mu$  and  $\alpha_i$ , then equating to zero, we get the following  $k+1$  normal equations. This  $E$  is known as Error sum of squares.

$$\frac{\partial E}{\partial \mu} = -2 \sum_{ij} (y_{ij} - \mu - \alpha_i) = 0, \quad (1)$$

$$\frac{\partial E}{\partial \alpha_i} = -2 \sum_j (y_{ij} - \mu - \alpha_i) = 0, \quad (i = 1, 2, \dots, k). \quad (2)$$

From (1) we get,

$$\sum_{ij} y_{ij} = \sum_i n_i (\mu + \alpha_i) = G, \quad \text{where } G \text{ is the grand total of all observations.}$$

From (2) we get,

$$\sum_j y_{ij} = n_i (\mu + \alpha_i) = T_i, \quad \text{where } T_i \text{ is the total of } i\text{th Group.}$$

$$\mu + \alpha_i = \frac{T_i}{n_i} = \bar{T}, \quad \text{where } \bar{T} \text{ is the mean of } i\text{th Group.}$$

Here all the equations are not independent. The summation of all  $k$  equations of (2) will give the equation in (1).

Thus, individual  $\alpha_i$ 's are not estimable. We can only estimate  $\mu + \alpha_i$ . Now,

$$\begin{aligned} E &= \sum_{ij} (y_{ij} - \mu - \alpha_i)^2 \\ &= \sum_{ij} y_{ij} (y_{ij} - \mu - \alpha_i), \quad \text{the other term vanishes by virtue of normal equations.} \end{aligned}$$



$$= \sum_{ij} y_{ij}^2 - \sum_i T_i(\mu + \alpha_i) = \sum_{ij} y_{ij}^2 - \sum_i \frac{T_i^2}{n_i} \quad (3)$$

Thus the error sum of squares is obtained.

Here, we are going to test the differences among the effects of k different groups. For the purpose, we make the null hypothesis as  $H_0: \alpha_1 = \alpha_2 = \dots = \alpha_k$ , i.e., all the group effects are equal. Under the above hypothesis, the model becomes,

$y_{ij} = \mu + e_{ij}$ , with only one parameter  $\mu$ , because the group itself is not a source of variation in  $y_{ij}$  according to our null hypothesis,  $H_0$ .

The changed normal equation will be:

$$E_1 = \sum_{ij} e_{ij}^2 = \sum_{ij} (y_{ij} - \mu)^2, \text{ where } E_1 \text{ is the changed error sum of squares. Then after the partial}$$

differentiation of  $E_1$  with respect to  $\mu$  and equating to zero, we will get:

$$-2 \sum_{ij} (y_{ij} - \mu) = 0,$$

$$\text{Or, } \hat{\mu} = \frac{\sum_{ij} y_{ij}}{n} = \bar{y}. \text{ Substituting the value of } \hat{\mu} \text{ in } E_1 \text{ we get,}$$

$$E_1 = \sum_{ij} (y_{ij} - \bar{y})^2 = \sum_{ij} y_{ij}^2 - \frac{G^2}{n}.$$

Hence, the sum of squares due to the (k-1) group contrasts will be

$$E_1 - E = \sum_{ij} y_{ij}^2 - \frac{G^2}{n} - \left( \sum_{ij} y_{ij}^2 - \sum_i \frac{T_i^2}{n_i} \right) \\ = \sum_i \frac{T_i^2}{n_i} - \frac{G^2}{n} \text{ with } k-1 \text{ degrees of freedom. The factor } G^2/n \text{ is known as correction factor. Using the}$$

correction factor, the error sum of square (E) in equation (3) can be rewritten as:

$$E = \left( \sum_{ij} y_{ij}^2 - \frac{G^2}{n} \right) - \left( \sum_i \frac{T_i^2}{n_i} - \frac{G^2}{n} \right) = \text{Total sum of squares} - \text{Group sum of squares.}$$

The ratio of two mean squares, viz., Group mean square and error mean square under the null hypothesis stated

above will follow F distribution with (k-1) and (n-k) degrees of freedom. We can write  $\frac{E_1 - E}{E} \cdot \frac{n - k}{k - 1}$  will

follow F distribution with (k-1) and (n-k) degrees of freedom. The value of calculated F will be compared with the Table value of F distribution at 5% or 1% level of significance with (k-1) and (n-k) degrees of freedom. Next we will go for the Analysis of variance (ANOVA) table for testing the null hypothesis.

**ANOVA Table:**

Sources of variation	Degrees of freedom (df)	Sum of squares (SS)	Mean squares $\left( MS = \frac{SS}{df} \right)$	Computed F-ratio $\frac{MSS}{MSE}$	Tabulated F with group df and Error df
Between groups	k-1	GSS	GMS = GSS/(k-1)	GMS/EMS	
Within groups (error)	n-k	ESS	EMS = ESS/(n-k)		
Total	n-1	TSS			

GSS = Group Sum of Squares, ESS = Error Sum of Squares, TSS = Total Sum of Squares, GMS = Group Mean Squares and EMS = Error Mean Squares.

The estimate of group effects is  $\hat{\mu} + \hat{\alpha}_i = \frac{T_i}{n_i} = \bar{y}_i$ .

Variance of any two group contrast will be

$V(\alpha_i - \alpha_j) = \frac{2\sigma^2}{n_i} = \frac{2 \text{ Error Mean Square}}{n_i}$ , the estimate of Error variance ( $\sigma^2$ ) is the error mean square

which is Error sum of square/ error degrees of freedom.

**Example:** The data presented in the following table represents the weight (in gm) of randomly collected eggs from five different breeds of hens.

The analysis of variance for the sample data is conducted as follows.

Model of analysis:

$$y_{ij} = \mu + \alpha_i + e_{ij}, i=1, 2, \dots, k; j = 1, 2, \dots, n_i,$$

where  $y_{ij}$  is the response of the  $j$ th individual unit belonging to the  $i$ th category or group,  $\mu$  is overall population mean,  $\alpha_i$  is the effect of being in the  $i$ th group and  $e_{ij}$  is a random error which follows normal distribution, attached to the  $(ij)$ th observation.

$H_0: \alpha_1 = \alpha_2 = \dots = \alpha_k$ , against  $H_1: \alpha_1 \neq \alpha_2 \neq \dots \neq \alpha_k$ ,  $H_1: \alpha_1 \neq \alpha_2 \neq \dots \neq \alpha_k$ .

Step 1. Compute the group totals, group means, grand total and grand mean. Here, the number of groups or breeds =  $k = 5$  and number of eggs collected from different breeds is not equal. The table presents the number of eggs from each breed with their weight (gm). Here  $n = \text{Total number of observation} = 17 (= 4+3+4+3+3)$ .

**Table: The weight of eggs for 5 selected breeds**

	Egg weight of Breeds (gm)					Overall
	1	2	3	4	5	
1	158	167	149	153	157	
2	154	163	155	161	164	
3	151	168	158	153	163	
4	142		156			
Total	605	498	618	467	484	2672
Mean	151.25	166	154.5	155.667	161.333	157.75

Step 2. Compute the correction factor  $C.F = \frac{G^2}{n} = 419975.5$

Step 3. Compute the total sum of squares  $TSS = \sum_{ij} y_{ij}^2 - \frac{G^2}{n}$

$$TSS = (158^2 + 167^2 + \dots + 163^2) - 419975.5 = 730.4706$$

Step 4. Compute the Group sum of squares  $GSS = \sum_i \frac{T_i^2}{n_i} - \frac{G^2}{n}$

$$GSS = (605^2 / 4 + 498^2 / 3 + 618^2 / 4 + 467^2 / 3 + 484^2 / 3) - 419975.5 = 461.3873$$

Step 5. Compute the error sum of squares as  $ESS = TSS - GSS = 269.0833$

Step 6. Compute the mean squares for group and error. These are obtained using Equations  $GMS = GSS/(k-1)$  and  $EMS = ESS/(n-k)$ .

Step 8. Summarize the results as shown below:

Table for ANOVA :

Sources of variation	Degree of freedom (df)	Sum of squares (SS)	Mean squares $\left( MS = \frac{SS}{df} \right)$	Computed F-ratio	Tabular F
Between Groups	5-1 = 4	461.3873	115.3468	5.144	3.26 at 5% level of sig.
Within groups or Error	16 - 4 = 12	269.0833	22.4236		
Total	17-1 = 16	730.4706			

Compare the computed value of F with tabular value of F at 4 and 12 degrees of freedom. In this example, the computed value of F (5.144) is greater than the tabular value (3.26) at 5% level of significance. Thus the null hypothesis is rejected. It may thus be concluded that there are significant differences among the mean of egg weights of different groups.

#### Analysis of variance of two way classified data:

If there are two factors A and B, acting simultaneously and p levels of factor A and q levels of factor B, are involved in the action. Then,  $y_{ij}$  be the response of the joint action of the  $i$ th ( $= 1, 2, \dots, p$ ) level of factor A and  $j$ th ( $= 1, 2, \dots, q$ ) level of factor B. The total number of observation  $n = pq$ . The model of the two way data will be:

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}, \quad i = 1, 2, \dots, p; j = 1, 2, \dots, q,$$

where  $y_{ij}$  is the response of the  $i$ th level of A at  $j$ th level of B,  $\mu$  is overall observation mean,  $\alpha_i$  is the effect of  $i$ th level of factor A,  $\beta_j$  is the effect of  $j$ th level of factor B and  $e_{ij}$  is a random error which follows normal distribution, attached to the  $(ij)$ th observation.

The basic assumptions are same as in one way data, which are (i) The different source effects are additive (ii) The errors  $e_{ij}$  are independently and identically distributed as normal with mean zero and constant variance ( $\sigma^2$ );  $e_{ij} \sim \text{i.i.d. } N(0, \sigma^2)$ .

For estimating  $\alpha_i$ , the normal equations are obtained by method of least squares.

Let  $E = \sum_{ij} e_{ij}^2 = \sum_{ij} (y_{ij} - \mu - \alpha_i - \beta_j)^2$ , then the normal equations are obtained by taking the partial differentiation of E with respect to unknown parameters  $\mu$ ,  $\alpha_i$  and  $\beta_j$  then equating to zero, we get the following  $(p + q + 1)$  normal equations. This E is known as Error sum of squares.

$$\frac{\partial E}{\partial \mu} = -2 \sum_{ij} (y_{ij} - \mu - \alpha_i - \beta_j) = 0, \quad (1)$$

$$\frac{\partial E}{\partial \alpha_i} = -2 \sum_j (y_{ij} - \mu - \alpha_i - \beta_j) = 0, \quad (i = 1, 2, \dots, p). \quad (2)$$

$$\frac{\partial E}{\partial \beta_j} = -2 \sum_i (y_{ij} - \mu - \alpha_i - \beta_j) = 0, \quad (j = 1, 2, \dots, q). \quad (3)$$

By simplification of the above equations we get,

$$pq\mu + q \sum_i \alpha_i + p \sum_j \beta_j = G \quad (4)$$

$$q\mu + q\alpha_i + \sum_j \beta_j = T_i, \quad (i = 1, 2, \dots, p). \quad (5)$$

$$p\mu + \sum_i \alpha_i + p\beta_j = B_j, \quad (j = 1, 2, \dots, q). \quad (6)$$

where  $\sum_{ij} y_{ij} = G$ ,  $\sum_j y_{ij} = T_i$  and  $\sum_i y_{ij} = B_j$ .

These equations are not independent as the sum of p equations in (2) or sum of q equations of (3) will give the equation at (1). From (5) we get

$$\hat{\mu} + \hat{\alpha}_i = \frac{T_i}{q} - \sum_j \frac{\hat{\beta}_j}{q} = \frac{T_i}{q} - \bar{\beta}.$$

From (6) we get  $\hat{\mu} + \hat{\beta}_j = \frac{B_j}{p} - \sum_i \frac{\hat{\alpha}_i}{p} = \frac{B_j}{p} - \bar{\alpha}$  and from (4) we get

$$\hat{\mu} + \bar{\alpha} + \bar{\beta} = \bar{y}, \quad \text{where } \sum_i \frac{\hat{\alpha}_i}{p} = \bar{\alpha}, \quad \sum_j \frac{\hat{\beta}_j}{q} = \bar{\beta} \quad \text{and} \quad \sum_{ij} \frac{y_{ij}}{pq} = \bar{y}.$$

Substituting these in equation

$$E = \sum_{ij} (y_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j)^2 = \sum_{ij} y_{ij} (y_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j), \quad \text{other term vanishes by virtue of normal equations.}$$

$$\begin{aligned} E &= \sum_{ij} y_{ij}^2 - \hat{\mu}G - \sum_i \hat{\alpha}_i T_i - \sum_j \hat{\beta}_j B_j \\ &= \sum_{ij} y_{ij}^2 - \frac{\sum_i T_i^2}{q} - \frac{\sum_j B_j^2}{p} + \frac{G^2}{pq}. \end{aligned}$$

Let us make the null hypothesis,  $H_0: \alpha_1 = \alpha_2 = \dots = \alpha_p$ , then the changed error sum of square will be  $E_1 = \sum_{ij} (y_{ij} - \mu - \beta_j)^2$ . The normal equations will also be changed accordingly and ultimately  $E_1 =$

$$\sum_{ij} y_{ij}^2 - \frac{\sum_j B_j^2}{p}.$$

Hence the sum of square due to factor A will be

$$\text{SS due to A} = E_1 - E = \frac{\sum_i T_i^2}{q} - \frac{G^2}{pq} \quad \text{with } p-1 \text{ degrees of freedom. Here } \frac{G^2}{pq} \text{ is the value of correction factor}$$

(CF). Hence,

$\frac{E_1 - E}{E} \times \frac{pq - p - q - 1}{p - 1}$  is distributed as F with (p-1) and (pq-p-q-1) degrees of freedom. This value of calculated F will be compared with the tabulated F with (p-1) and (pq-p-q-1) degrees of freedom. Ultimately we can conclude that the null hypothesis will be rejected or not. If we are interested to see the differences of the factor B, the null hypothesis will be  $H_0: \beta_1 = \beta_2 = \dots = \beta_q$ , then the changed error sum of square will be  $E_2 = \sum_{ij} e_{ij}^2 = \sum_{ij} (y_{ij} - \mu - \alpha_i)^2$ . The other steps are similar to the process for factor A.

Next we will go for the Analysis of variance (ANOVA) table for presentation.

**ANOVA Table:**

Sources of variation	Degrees of freedom (df)	Sum of squares (SS)	Mean squares $\left( MS = \frac{SS}{df} \right)$	Computed F-ratio $\frac{MSS}{MSE}$	Tabulated F with group df and Error df
Between Factor A	p-1	SS due to A	AMS = SS due to A / (p-1)	AMS/EMS	
Between Factor B	q-1	SS due to B	BMS = SS due to B / (q-1)	BMS/EMS	
Within groups (error)	(p-1)(q-1)	ESS	EMS = ESS / (p-1)(q-1)		
Total	pq-1	TSS			

SS due to A = Sum of Squares for A, SS due to B = Sum of Squares for B, ESS = Error Sum of Squares, TSS = Total Sum of Squares, AMS = Mean Squares for A, BMS = Mean Squares for B and EMS = Error Mean Squares.

The estimate of Factor A effects is  $\hat{\mu} + \hat{\alpha}_i = \frac{T_i}{q} = \bar{y}_i$ .

The estimate of Factor A effects is  $\hat{\mu} + \hat{\beta}_j = \frac{B_j}{p} = \bar{y}_j$ .

Variance of contrast of any two levels of factor A will be

$V(\alpha_i - \alpha_{i'}) = \frac{2\sigma^2}{q} = \frac{2 \text{ Error Mean Square for A}}{q}$ , ( $i \neq i'$ ), the estimate of Error variance ( $\sigma^2$ ) is the

error mean square which is Error sum of square/ error degrees of freedom.

In the similar way  $V(\beta_j - \beta_{j'})$  can be calculated.

**Example:** The data presented in the following table represents the two way classification of rate of increase of weight of 5 breeds of chicken (kg) for three different batches in a poultry farm. We are interested to see, whether the variations of breeds in the rate of increase of weight of chicks are significantly different or not.

The analysis of variance for the sample data is conducted as follows.

Model of analysis:

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}, i = 1, 2, \dots, 5; j = 1, 2, \dots, 3,$$

where  $y_{ij}$  is the response of the  $i$ th breed at  $j$ th season,  $\mu$  is overall observation mean,  $\alpha_i$  is the effect of  $i$ th breed,  $\beta_j$  is the effect of  $j$ th batch and  $e_{ij}$  is a random error which follows normal distribution, attached to the  $(ij)$ th observation.

$H_0: \alpha_1 = \alpha_2 = \dots = \alpha_5$ , against  $H_1: \alpha_1 \neq \alpha_2 \neq \dots \neq \alpha_5$ .

**Table:** The rate of change of weight of chickens in different batches

	Breeds weight change (kg)					Total	Mean
	1	2	3	4	5		
Batch1	0.58	0.67	0.49	0.53	0.57	2.84	0.57
Batch2	0.54	0.63	0.55	0.61	0.64	2.97	0.59
Batch3	0.51	0.68	0.58	0.53	0.63	2.93	0.59
Total	1.63	1.98	1.62	1.67	1.84	8.74	
Mean	0.54	0.66	0.54	0.56	0.61		

Step 1. Compute Grand total (G), Total of  $i$ th Breed ( $T_i$ ) and Total of  $j$ th Batch ( $B_j$ ) as following:

$$\sum_{ij} y_{ij} = G, \sum_j y_{ij} = T_i \text{ and } \sum_i y_{ij} = B_j, i = 1, 2, 3, 4, 5; j = 1, 2, 3.$$

Compute Grand mean =  $\bar{y} = \frac{G}{pq}$ , compute  $i$ th breed mean =  $\bar{T}_i = \frac{T_i}{q}$  and compute  $j$ th Batch Mean =

$$\bar{B}_j = \frac{B_j}{p}.$$

Step 2. Compute the correction factor C.F =  $\frac{G^2}{pq} = 5.09$

Step 3. Compute the total sum of squares  $TSS = \sum_{ij} y_{ij}^2 - \frac{G^2}{pq}$ ,

$$TSS = (0.58^2 + 0.67^2 + \dots + 0.63^2) - 5.09 = 0.05$$

Step 4. Compute the Sum of squares Due to Breeds  $= \sum_i \frac{T_i^2}{q} - \frac{G^2}{pq} = Br.SS$

$$Br.SS = (1.63^2 + 1.98^2 + 1.62^2 + 1.67^2 + 1.84^2)/3 - 5.09 = 0.03$$

Step 5. Compute the Sum of squares Due to Batches  $= \sum_j \frac{B_j^2}{p} - \frac{G^2}{pq} = Ba.SS$

$$Ba.SS = (2.84^2 + 2.97^2 + 2.93^2)/5 - 5.09 = 0.0018$$

Step 5. Compute the error sum of squares as  $ESS = TSS - Br.SS - Ba.SS = 0.01$

Step 6. Compute the mean squares for Breed, Batch and error. These are obtained using Equations  $BrMS = Br.SS/(p-1)$ ,  $BaMS = Ba.SS/(q-1)$  and  $EMS = ESS/(p-1)(q-1)$ .

Step 8. Summarize the results as shown below:

**ANOVA Table:**

Sources of variation	Degrees of freedom (df)	Sum of squares (SS)	Mean squares	Computed F-ratio	Tabulated F with group df and Error df
Between Factor Breed	5-1 = 4	Br.SS	Br.MS = Br.SS/4	Br.MS/EMS = 4.90	3.84
Between Factor Batch	3-1 = 2	Ba.SS	Ba.MS = Ba.SS/2	Ba.MS/EMS = 0.53	4.46
Within groups (error)	(5-1)(3-1) = 8	ESS	EMS = ESS/8		
Total	15-1 = 14	TSS			

Compare the computed value of F with tabular value of F at 4 and 8 degrees of freedom for comparing the mean values of different breeds. In this example, the computed value of F (4.90) is greater than the F tabular value (3.84) at 5% level of significance. Thus the null hypothesis is rejected. It may thus be concluded that there are significant differences among the means of different breeds. But the mean values of three different batches are not significantly different.



### Transformation of data:

One of the most important assumptions made in analysis of variance is that the observations should follow normal distribution. In general the agricultural field experiments parameters like yield measurements, height and girth measurements, etc. follow normal distribution. However, certain types of measurements in biological fields, like percentage infestation, germination percentage, insect population, etc., do not follow normal distribution as such. It is, therefore, necessary to transform the data into a different scale, so that the transformed data may follow normal distribution. Quite often, the data are subjected to certain scale transformations such that in the transformed scale, the constant variance assumption is realized. Some of such transformations can also correct for departures of observations from normality because unequal variance is many times related to the distribution of the variable also. Certain methods are available for identifying the transformation needed for any particular data set (Montgomery and Peck, 1982) but one may also resort to certain standard forms of transformations depending on the nature of the data. The most common of such transformations are *logarithmic transformation*, *square root transformation* and *angular transformation*.

### Some guidelines are worked out which indicate the appropriate transformation for a given set of data.

Firstly, mean and variance of the observed set of data of any experiment may be undergone through a graph. The graph would be drawn mean versus variance. In case we get a straight line, we may conclude that variance is proportionately changing with mean and in that case **square root transformation** should be done.

Here,  $y$ , the observed response should be changed to  $\sqrt{y}$ . Thus, this type of transformation is recommended when linear relationship is commonly observed between mean and variance. When the data consist of a small portion of the whole numbers (*e.g.*, number of bacterial column per plate count, weeds per plot, earthworms per square metre of soil, insects caught in traps, etc.) this square root transformation is recommended when the observed values fall within the range of 1 to 10. But when many zeros are present in the data set, the transformation should be,  $\sqrt{y + 0.5}$  instead of only  $\sqrt{y}$ .

Secondly, when the mean is proportional to standard deviation of the observed data, **logarithmic transformation** should be used for analysis of variance. A good example is data from an experiment involving various types of insecticides. To study the effectiveness of insecticides, this type of transformation is required. Insect counts on the treated experimental unit may be small while for the ineffective ones or in the untreated experimental units, the counts may range from 100 to several thousands. If zeros are present in the data, it is advisable to add 1 to each observation before making the transformation. The log transformation is particularly effective when the data show positively skewed distributions. It is also used to achieve additive property of effects in certain cases. When data relate to proportions, one can use the transformation  $y = \log \frac{x}{1-x}$ , where  $x$  represents to observed proportions. For example, if  $A'$  denotes the area of plant leaf affected by a disease and  $A$  denotes the total area of the leaf, then  $x = A'/A$  which will lie between 0 to 1.

Again all kinds of proportions do not require logarithmic transformation. When the observed data relate a binomial population an **angular transformation** (also termed as **sine inverse transformation**) should be used.

Here, the transformed equation is  $\theta = \sin^{-1}\sqrt{p}$ , where  $\theta$  is the transformed data for analysis and  $p$  is the observed percentage converted to proportion. However, when all observed percentages lie between 30 to 70, the data need not to undergo transformation. Because such observations follow approximately normal distribution. Hence no transformation is required. When the data contain 0 or 1 values for  $p$ , before taking angular values, replace 0 with  $(1/4n)$  and 1 with  $[1-(1/4n)]$ , where  $n$  is the number of observations based on which  $p$  is estimated for each group.

Once the transformation has been made the analysis of variance is carried out in usual way with the transformed data and the conclusions are drawn on the basis of transformed variates. The arithmetic mean and their confidence intervals are again retransformed into original form for final presentation.

### **Experimental Designs or Design of Experiments:**

Experimental design actually deals with methods of constructing and analyzing comparative options under study. It may be considered for two main processes. The first one is the construction and planning of experiments and secondly, the analysis process through the technique analysis of variance. Thus, experimental designs or design of experiments comprises the process of planning of experiments, analyzing the results through analysis of variance techniques and drawing inferences from the experiments. **The technique mainly used to draw the inferences is 'analysis of variance'**. The designing of experiments and analysis of experiments strongly influence one another. The faulty choice in designing will lead to wrong analysis and ultimately, wrong judgments. The process of experimental design starts from basic lay-out of the experiment which are capable of facilitating the analysis of the experiment with a better precision. For the purpose of a good designing, some basic principles of designs should be followed accurately. There are three basic principles of experimental designs, viz., **Replication, Randomization and Local control**. These principles increase the precision of the experiments and these will also help to draw valid inferences from the experiments.

### **Basic principles of experimental designs:**

**Replication:** In design of experiments replication refers to repetition. The repetition of treatments by applying them to more than one experimental unit is known as replication. In any statistical experiment replication is necessary to get an estimate of the experimental error variation caused due to environmental or any other uncontrollable factors. To estimate the error variation is necessary for any experiment. Replication also increases the precision of the experiments. If we repeat a single treatment  $r$  number of times, the mean of the treatment will be subjected to a standard error  $\frac{\sigma}{\sqrt{r}}$ , where  $\sigma$  is the standard deviation of individual experimental plots or units and this standard error has to be estimated from the experiment. Thus, if  $r$  increases the standard error will decrease. Thus, replication of treatments in an experiment helps in reducing the standard error of the experiment in addition to provide an estimate of error variation.

**Randomization:** For any kind of objective comparison, it is necessary that the treatments should be allotted randomly to different experimental units. To make any experiment bias free randomization is necessary. Statistical procedures employed in making inferences about treatments provide a good and valid result only

when the treatments are allotted randomly to various experimental plots. By random allocation of treatments in the long run will be subjected to equalize or neutralize the environmental or uncontrollable effects.

**Local Control:** Every experiment with replication provides an estimate of error variation, but it is not desirable to have a large experimental error. The experimental error can be minimized by making use of the fact that the adjacent areas in the field of experiment should be relatively more homogenous than the areas which are widely separated. One usual procedure in this regard is that the entire experimental field or material should be divided into different groups by taking homogeneous plots or units together and the treatments may be allotted randomly to different units in each group. This procedure is commonly known as local control. The main objective of the local control is to reduce the error by suitably modifying the allocation of treatments to the experimental units.

The readers may follow that the term ‘treatment’ is used in explaining the basic principles of design of experiments. It is better to give an idea about it by a simple definition.

**Treatment:** In experimental designs, treatment refers to any stimulus which is directly related to response or observation of an experiment and which is applied so as to observe the change of effect (positive or negative) of the response in an experimental situation or to compare its effect with other stimuli used in the same experiment. In practice, treatment may refer to any physical substance or a procedure or any thing, which has been capable of controlled application according to the requirement of the experimenter.

**Example:** Suppose an experimenter wants to compare the yield of available varieties of rice in West Bengal. Here the varieties are the treatments. Suppose one research scholar of Agronomy, wants to establish that three time irrigations to rabi crops is better than four or more irrigations. Here, the different numbers of irrigations are the treatments. Again, line sowing of jute crop is better than other methods of sowing jute crop. Here the sowing methods are the treatments.

#### **Basic Designs:**

**Completely Randomised Design:** The simplest design using only two essential or basic principles of experimental designs, viz., replication and randomization. It is commonly known as CRD. In this design of experiment, it is assumed that the whole experimental material is homogeneous. Then the experimental material will be divided into a number of small unit or plots depending upon the number of treatments under study. The treatments are allotted randomly to those experimental plots or units. The process of allocation of treatments can be done by regular process of randomization. Firstly, the total number of units or plots ( $n$ ) required has to be finalized. Suppose there are  $v$  number of treatments and each treatment is replicated  $r_i$  number of times ( $i = 1, 2,$

...,  $v$ ). Then  $n (= \sum_{i=1}^v r_i)$  is the total number of plots required for the experiment. Now, the entire experimental material should be divided into  $n$  number of equal experimental plots or units. In the next step, experimenter should give serial numbers (from 1 to  $n$ ) to each of the  $v$  treatments with their replications. Suppose the first treatment is replicated  $r_1$  number of times, then these treatments will be framed with 1 to  $r_1$  number. The second treatment, suppose, replicated for  $r_2$  number of times then these treatments will be framed with serial members from  $(r_1 + 1)$  to  $r_2$ . The process will continue up to the last  $v$ th treatment which is replicated for  $r_v$  number of times. The experimental units are also marked with the serial numbers from 1 to  $n$ . Now total  $n$  treatments will

be placed in n number of experimental units by following the rule of randomization. If the first number selected from the random table is m, then the treatment which corresponds to m should be placed to first plot. Repetition of any number is not allowed here. Then we proceed for the selection of treatment for the second plot. Likewise all n treatments will be placed on n experimental units or plots.

**Analysis of CRD:** The analysis of CRD is following the technique of analysis of variance of one way classified data. Let there be v number of treatments and ith treatment be replicated for  $r_i$  number of times. Here the additive fixed effect model is:

$$y_{ij} = \mu + \alpha_i + e_{ij}, \quad i=1, 2, \dots, v; j = 1, 2, \dots, r_i, \text{ where}$$

$y_{ij}$  is the yield or response of jth replication of ith treatment;

$\mu$  is the general mean effect of all observations;

$\alpha_i$  is the ith treatment effect and

$e_{ij}$  is the error effect of jth replication of ith treatment, which follows normal distribution with zero mean and  $\sigma^2$  variance.

Following computations are done according to the process of analysis of variance of one way classified data.

$$\text{Grand Total} = G = \sum_{i=1}^v \sum_{j=1}^{r_i} y_{ij}, \quad \text{Total number of observations} = n = \sum_{i=1}^v r_i$$

$$\text{Correction factor} = CF = \frac{G^2}{n}, \quad \text{Total of ith treatment} = T_i = \sum_{j=1}^{r_i} y_{ij},$$

$$\text{Mean of ith treatment} = \bar{T}_i = \frac{T_i}{r_i}.$$

Then the sum of squares are calculated:

$$\text{Total Sum of Squares} = TSS = \sum_{i=1}^v \sum_{j=1}^{r_i} y_{ij}^2 - CF$$

$$\text{Treatment Sum of Squares} = Tr.SS = \sum_{i=1}^v \frac{T_i^2}{r_i} - CF$$

$$\text{Error Sum of Squares} = ESS = TSS - Tr.SS.$$

The null hypothesis of the experiment will be:

$$H_0: \alpha_1 = \alpha_2 = \dots = \alpha_v, \text{ against } H_1: \alpha_1 \neq \alpha_2 \neq \dots \neq \alpha_v.$$

Then we go for analysis of variance table:

## ANOVA Table:

Sources of Variation or SOV	Degrees of Freedom or df	Sum of squares or SS	Mean squares or MS	F Ratio	
				Calculated	Tabulated
Treatment	v -1	Tr.SS	Tr.MS = Tr.SS/(v-1)	Tr.MS/ EMS	
Error	n-v	ESS	EMS = ESS/(n-v)		
Total	n-1	TSS			

**Conclusion about the experiment:** If the calculated value of F is less than the tabulated value of F for (v-1) and (n-v) degrees of freedom at 5% level of significance, then the null hypothesis is not rejected. Then we can conclude that the test is insignificant and the treatments under study are equally effective. Otherwise, if the calculated value of F is greater than or equals to the tabulated value of F for (v-1) and (n-v) degrees of freedom at 5% level of significance, then the null hypothesis is rejected. Then we can conclude that the test is **significant** and the treatments under study are not equally effective. It is to be noted, that if the level of significance is 1%, then the comparison should be done with the Tabulated value of F at 1% level of significance with (v-1) and (n-v) degrees of freedom. If the calculated value of F is greater than or equals to tabulated F, then we can conclude that the test is **highly significant**. Generally, the conclusions are drawn on the basis of 5% level of significance. If the null hypothesis is rejected, it is clear that the treatments are not equally effective. Then the most vital question should come. Which one is the best or worst treatment? Here we should go for calculating the standard error of difference of treatment mean values (SEd) for calculating the critical difference value or CD value.

The estimated standard error of difference between  $i$ th and  $i'$ th ( $i \neq i'$ ) = 1, 2, ..., v) treatment mean values =

$$SEd = \sqrt{EMS \left( \frac{1}{r_i} + \frac{1}{r_{i'}} \right)}, \text{ where } r_i \text{ and } r_{i'} \text{ are the replications of } i\text{th and } i'\text{th treatments } (i \neq i') = 1, 2, \dots, v).$$

Then the CD value will be calculated for comparing  $i$ th and  $i'$ th ( $i \neq i'$ ) = 1, 2, ..., v) treatment mean values will be = CD (for  $i$ th and  $i'$ th treatments) = SEd x  $t_{\alpha}$ , where  $t_{\alpha}$  is the tabulated value of t- distribution at  $\alpha\%$  level of significance ( $\alpha$  may be 5% or 1%) with error degrees of freedom i.e., (n-v).

For comparing the treatment mean values, firstly we have to calculate the mean of v treatments. Then make the difference of the pair of treatments  $i$ th and  $i'$ th. If the difference is less than the CD value, we can conclude that the treatments are equally effective. Otherwise, if the difference is greater than or equals to the CD value the treatments are not equally effective.

**Example:** Following data gives increase in weight of chicks (gm) after certain period when they fed four kinds of feed. Analyze the data and draw your conclusion.

Increase in weight of chicks	Feed 1	Feed 2	Feed 3	Feed 4
	150	145	140	152
	158	148	142	158
	162	155	146	154
	160	146	148	155
	162	150	142	160

**Randomised Block Design (RBD) or Randomised Complete Block Design (RCBD):** In order to control the heterogeneity in one direction in the experimental material, it is desirable to divide the entire experimental into homogeneous groups of units, known as blocks. The treatments are randomly allotted to these blocks. Separate random allocation is done for each block. This procedure gives rise to a design known as Randomised Block Design (RBD) or Randomised Complete Block Design (RCBD). This design is an application of analysis of variance of two way classified data. This design is an arrangement of  $v$  treatments in each of the  $r$  blocks such that each of the treatments occurs once and only once in each block.

The design lay out for RBD is very simple. Firstly, the experimental material should be divided into more or less homogeneous blocks of equal sizes. The number of such blocks will be the number of replications,  $r$ . Now, each block will be again divided into  $v$  number of small experimental units of equal sizes. The  $v$  treatments are then allotted randomly to the units of each block. Thus every treatment is placed  $r$  number of times in the whole set up. In agriculture, the experimental field can be divided into homogeneous blocks. In a Dairy farm, if the cows of same age group are kept under one shed, then it should be divided into  $r$  number of rows with the provision of  $v$  number of cows in each row. Other management practices and environmental conditions like air, light, humidity etc. should be maintained with equal care. Then the  $v$  treatments, may be medicines or vitamins should be given to each cow in each row of the shed randomly to  $v$  number of cows.

#### Analysis of RBD:

The analysis of the RBD is more or less similar to analysis of variance of two way classified data. Here the linear additive model is:

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}, i = 1, 2, \dots, v; j = 1, 2, \dots, r,$$

where  $y_{ij}$  is the response of the  $i$ th treatment at  $j$ th block,  $\mu$  is overall observation mean,  $\alpha_i$  is the effect of  $i$ th treatment,  $\beta_j$  is the effect of  $j$ th block and  $e_{ij}$  is a random error which follows normal distribution, attached to the  $i$ th treatment in  $j$ th block. Thus,  $e_{ij} \sim N(0, \sigma^2)$ .

Thus, it is clear that in the design  $v$  treatments are replicated  $r$  times.

Following computations are done according to the process of analysis of variance of two way classified data.

$$\text{Grand Total} = G = \sum_{i=1}^v \sum_{j=1}^r y_{ij}, \text{ Total number of observations} = n = vr$$

$$\text{Correction factor} = CF = \frac{G^2}{n}, \text{ Total of } i\text{th treatment} = T_i = \sum_{j=1}^r y_{ij},$$

$$\text{Total of } j\text{th Block} = B_j = \sum_{i=1}^v y_{ij}; \text{ Mean of } i\text{th treatment} = \bar{T}_i = \frac{T_i}{r}.$$

$$\text{Then, Total Sum of Squares} = TSS = \sum_{i=1}^v \sum_{j=1}^r y_{ij}^2 - CF;$$

$$\text{Treatment Sum of Squares} = Tr.SS = \frac{1}{r} \sum_{i=1}^v T_i^2 - CF$$

$$\text{Block Sum of Squares} = Bl.SS = \frac{1}{v} \sum_{j=1}^r B_j^2 - CF$$

$$\text{Error Sum of Squares} = ESS = TSS - Tr.SS - Bl.SS$$

The null hypothesis for comparing the treatment effects of the experiment will be:

$$H_0: \alpha_1 = \alpha_2 = \dots = \alpha_v, \text{ against } H_1: \alpha_1 \neq \alpha_2 \neq \dots \neq \alpha_v.$$

One may also compare the block effects of the experiment by taking the null hypothesis as equality of block effects.

For testing the null hypothesis we proceed for the analysis of variance table:

**ANOVA Table:**

Sources of Variation or SOV	Degrees of Freedom or df	Sum of squares or SS	Mean squares or MS	F Ratio	
				Calculated	Tabulated
Treatment	v-1	Tr.SS	$Tr.MS = \frac{Tr.SS}{v-1}$	$\frac{Tr.MS}{EMS}$	
Block	r-1	Bl.SS	$Bl.MS = \frac{Bl.SS}{r-1}$		
Error	(v-1)(r-1)	ESS	$EMS = \frac{ESS}{(v-1)(r-1)}$		
Total	vr-1	TSS			

**Conclusion about the experiment:** Similar to CRD.

**Example:** The following table gives the lay out and yields (kg.) of five varieties of rice in an experiment in four randomised blocks. Analyse the data and make the conclusion about the varieties of rice.

Blocks	I	130.6 (2)*	127.7 (1)	124.0 (5)	127.8 (3)	116.2 (4)
	II	129.3 (3)	115.0 (4)	122.5 (1)	128.8 (2)	117.0 (5)
	III	122.7 (5)	131.0 (2)	114.1 (4)	134.9 (1)	128.5 (3)
	IV	114.1 (4)	122.1 (3)	117.1 (5)	139.5 (2)	136.0 (1)

\* = Figures in parenthesis denotes the treatment or variety number.

### Latin Square Design:

To control the heterogeneity in two directions in the experimental material, we commonly use the design known as Latin Square Design (LSD). In such designs, two restrictions are imposed by forming groups in two directions, row wise and column wise. Treatments are allotted in such a way that every treatment occurs either in each row or in each column just once and only once. We know that in a latin square of order  $s$  has  $s^2$  positions in a two way arrangement of  $s$  rows and  $s$  columns. Every cell of any row or any column of the square has  $s$  latin letters precisely once. The structure of the design is also looked like a latin square, thus the design is named as latin square design. Here, the treatment effects can be estimated with the elimination of row effects and column effects consequently the error variation is reduced considerably.

Lay out of the design should be followed by the guidelines of Fisher and Yate's table or by Statistical tables written by Rao, Mitra and Mathai. Commonly it has been done by cyclical rotation of  $v$  number of letters as given below:

Let  $v$  be the number of treatments and  $v = 5$ .

A	B	C	D	E
E	A	B	C	D
D	E	A	B	C
C	D	E	A	B
B	C	D	E	A

The letters represents the treatments.

### Analysis of Latin Square design:

The additive model of the design is

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + e_{ijk}, \quad i = 1, 2, \dots, v; j = 1, 2, \dots, v, \text{ \& } k = 1, 2, \dots, v.$$

where  $y_{ijk}$  is the response of the  $i$ th treatment at  $j$ th row and  $k$ th column,  $\mu$  is overall observation mean,  $\alpha_i$  is the effect of  $i$ th treatment,  $\beta_j$  is the effect of  $j$ th row,  $\gamma_k$  is the  $k$ th column effect and  $e_{ijk}$  is a random error which follows normal distribution, attached to the  $i$ th treatment in  $j$ th row and  $k$ th column. Thus,  $e_{ijk} \sim N(0, \sigma^2)$ .

Thus, it is clear that in the design  $v$  treatments are replicated  $v$  times.

Following computations are done according to the process of analysis of variance of a latin square design.

$$\text{Grand Total} = G = \sum_{i=1}^v \sum_{j=1}^v y_{ijk} = \sum_{i=1}^v \sum_{k=1}^v y_{ijk} = \sum_{j=1}^v \sum_{k=1}^v y_{ijk},$$

Total number of observations =  $n = v^2$ .



Correction factor =  $CF = \frac{G^2}{n}$ , Total of  $i$ th treatment =  $T_i = \sum_{j=1}^v y_{ijk}$ , ignoring the column position =  $\sum_{k=1}^v y_{ijk}$ , ignoring row position.

Total of  $j$ th row =  $R_j = \sum_{i=1}^v y_{ijk}$  ignoring the column position =  $\sum_{k=1}^v y_{ijk}$  ignoring the treatment number.

Total of  $k$ th column =  $C_k = \sum_{i=1}^v y_{ijk}$  ignoring the row position =  $\sum_{j=1}^v y_{ijk}$  ignoring the treatment number.

Mean of  $i$ th treatment =  $\bar{T}_i = \frac{T_i}{v}$ .

Then, the sum of squares are calculated:

Total Sum of Squares =  $TSS = \sum_{j=1}^v \sum_{k=1}^v y_{ij}^2 - CF$

Treatment Sum of Squares =  $Tr.SS = \frac{1}{v} \sum_{i=1}^v T_i^2 - CF$

Row Sum of Squares =  $RSS = \frac{1}{v} \sum_{j=1}^v R_j^2 - CF$

Column Sum of Squares =  $CSS = \frac{1}{v} \sum_{k=1}^v C_k^2 - CF$

Error Sum of Squares =  $ESS = TSS - RSS - CSS - Tr.SS$

The null hypothesis for comparing the treatment effects of the experiment will be:

$H_0: \alpha_1 = \alpha_2 = \dots = \alpha_v$ , against  $H_1: \alpha_1 \neq \alpha_2 \neq \dots \neq \alpha_v$ .

One may also compare the block effects of the experiment by taking the null hypothesis as equality of block effects. For testing the null hypothesis we proceed for the analysis of variance table:

**ANOVA Table:**

Sources of Variation or SOV	Degrees of Freedom or df	Sum of squares or SS	Mean squares or MS	F Ratio	
				Calculated	Tabulated
Treatment	v-1	Tr.SS	$Tr.MS = \frac{Tr.SS}{v-1}$	$\frac{Tr.MS}{EMS}$	
Row	v-1	RSS	$RMS = \frac{RSS}{v-1}$		
Column	v-1	CSS	$CMS = \frac{CSS}{v-1}$		
Error	(v-1)(v-2)	ESS	$EMS = \frac{ESS}{(v-1)(v-2)}$		
Total	v <sup>2</sup> - 1	TSS			

**Conclusion about the experiment:** Similar to CRD

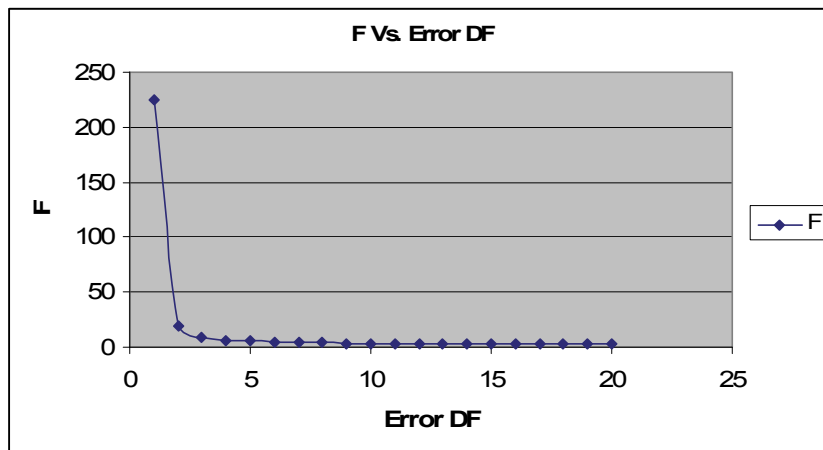
**Example:** The results of an experiment in a latin square design in a field of five varieties of rice are given below in suitable units. Analyse the the data and draw your conclusions.

A 141.5	B 143.5	C 135.0	D 147.0	E 137.0
E 139.0	A 143.5	B 141.0	C 133.2	D 148.4
D 142.9	E 139.0	A 142.0	B 139.0	C 129.0
C 130.0	D 147.8	E 140.0	A 144.0	B 145.6
B 139.7	C 128.5	D 142.6	E 134.3	A 142.5

The letter represents the variety and the corresponding yield values are given.

**Minimum number of replications:**

It has been observed that the error degrees of freedom is a function of the replication number r. The probability values of a F- distribution is linked with treatment degrees of freedom and error degrees of freedom. The treatment number in an experiment is fixed by the experimenter and it can not be changed so for the treatment degrees of freedom. The rate of change of F probability (table) values is very much unstable for error degrees of freedom less than 10 and it becomes more or less stable at more than 10 error degrees of freedom for any particular treatment degrees of freedom. The situation will be clear by the figure1 as below:



**Fig. 1: The curve of F values for different error df for fixed treatment df 4**

To avoid the unstable zone of F distribution, an experiment should be planned in such a way that F distribution will come into a stable zone. Consequently, the error degrees of freedom should be greater than 10. In RBD or in CRD (equal replication) the error degrees of freedom should be at least 10 as it is a product function of replication. This guide line should be followed for determination of number of replications required for any experiment with  $v$  number of treatments. However, there are some biological and medical experiments, where increase a single replication is very costly or very difficult because the specimens are very scarce. In those cases we can not be strict on the above minimum number of replications. Under the above guide line, we can conclude that we should avoid a latin square design with 3 treatments in a single latin square.

## ARTIFICIAL NEURAL NETWORKS AND ITS APPLICATIONS

**G Sathish, Pavana Kumar S.T. and Prof D. Mazumdar**  
**Department of Agricultural Statistics, BCKV, Nadia- 741252 (W.B)**

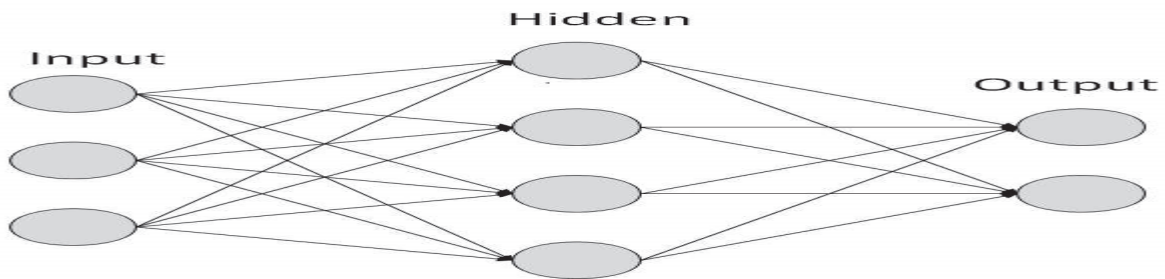
**Artificial neural networks (ANNs)** or **connectionist systems** are a computational model used in machine learning (Warren McCulloch and Walter Pitts, 1943), computer science and other research disciplines, which is based on a large collection of connected simple units called artificial neurons, loosely analogous to axons in a biological brain. Connections between neurons carry an activation signal of varying strength. If the combined incoming signals are strong enough, the neuron becomes activated and the signal travels to other neurons connected to it. Such systems can be trained from examples, rather than explicitly programmed, and excel in areas where the solution or feature detection is difficult to express in a traditional computer program. Like other machine learning methods, neural networks have been used to solve a wide variety of tasks, like computer vision and speech recognition, that are difficult to solve using ordinary rule-based programming.

Typically, neurons are connected in layers, and signals travel from the first (input), to the last (output) layer. Modern neural network projects typically have a few thousand to a few million neural units and millions of connections; their computing power is similar to a worm brain, several orders of magnitude simpler than a human brain. The signals and state of artificial neurons are real numbers, typically between 0 and 1. There may be a threshold function or limiting function on each connection and on the unit itself, such that the signal must surpass the limit before propagating. Back propagation is the use of forward stimulation to modify connection weights, and is sometimes done to train the network using known correct outputs. However, the success is unpredictable: after training, some systems are good at solving problems while others are not. Training typically requires several thousand cycles of interaction.

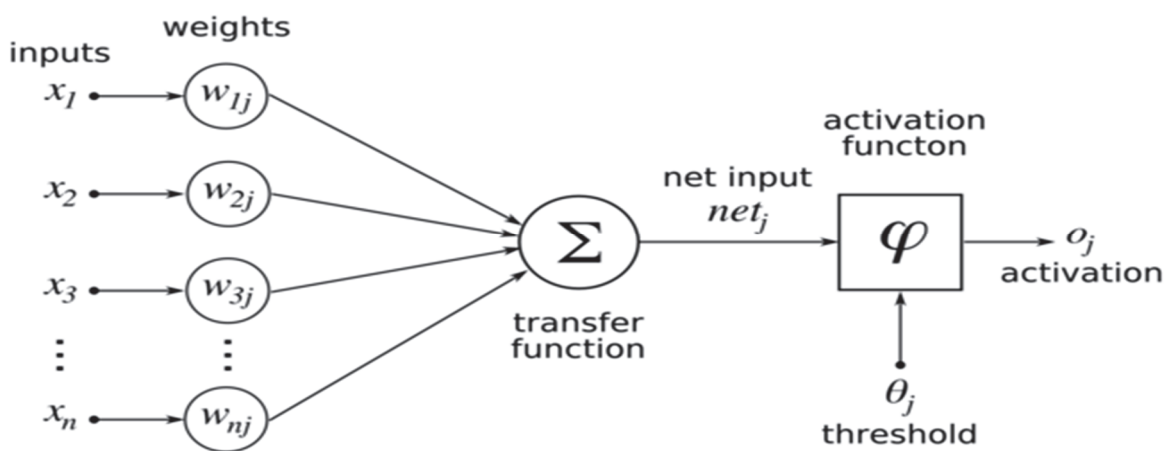
The goal of the neural network is to solve problems in the same way that a human would, although several neural network categories are more abstract. New brain research often stimulates new patterns in neural networks. One innovative approach is use of connections which span further to connect processing layers rather than adjacent neurons. Other research being explored with the different types of signal over time that axons propagate, such as deep learning, interpolates greater complexity than a set of boolean variables being simply on or off. Newer types of network are more free flowing in terms of stimulation and inhibition, with connections interacting in more chaotic and complex ways. Dynamic neural networks are the most advanced, in that they dynamically can, based on rules, form new connections and even new neural units while disabling others.

### **Levenberg-Marquardt Algorithm**

The Levenberg–Marquardt algorithm, which was independently developed by Kenneth Levenberg and Donald Marquardt (Hao Yu, Bogdan M. Wilamowski, 2010), provides a numerical solution to the problem of minimizing a nonlinear function. It is fast and has stable convergence. In the artificial neural-networks field, this algorithm is suitable for training small- and medium-sized problems.



Input and Output connection in ANN



Mathematical representation of a Neural Network

The basic idea of the Levenberg–Marquardt algorithm is that it performs a combined training process: around the area with complex curvature, the Levenberg–Marquardt algorithm switches to the steepest descent algorithm, until the local curvature is proper to make a quadratic approximation; then it approximately becomes the Gauss–Newton algorithm, which can speed up the convergence significantly. The derivation of the Levenberg–Marquardt algorithm will be presented in four parts:

- A. Steepest descent algorithm
- B. Newton’s method
- C. Gauss–Newton’s algorithm and
- D. Levenberg– Marquardt algorithm

Neural networks and other AI machine learning models are prone to “over fitting” .When the numbers of parameters are increase, significant over fitting can be observed and resulted in poor generalization. The over fitting is a frequent problem in multilayer perceptron (MLP) and there several theories developed to determine the optimum neural network architecture and which maximizes the generalization. The performance of ANN could be

analysed using model performance indicators such as model efficiency, correlation coefficient and root mean square error are expressed by:

$$\text{Model Efficiency (MEF)} = 1 - \frac{\sum_{i=1}^N (Q_i - Q_i^*)^2}{\sum_{i=1}^N (Q_i - \bar{Q})^2}$$

$$\text{Correlation Coefficient (CC)} = \frac{\sum_{i=1}^N (Q_i - \bar{Q})(Q_i^* - \bar{Q}^*)}{\sqrt{\left(\sum_{i=1}^N (Q_i - \bar{Q})^2\right)\left(\sum_{i=1}^N (Q_i^* - \bar{Q}^*)^2\right)}}$$

$$\text{Root Mean Square Error (RMSE)} = \left( (1/n) \sum_{i=1}^N (Q_i - Q_i^*)^2 \right)^{0.5}$$

Where, Q-Observed, Q\*-Predicted,  $\bar{Q}$  -Average observed,  $\bar{Q}^*$  -Average predicted

Machine Learning in ANNs

ANNs are capable of learning and they need to be trained. There are several learning strategies –

- **Supervised Learning** – It involves a teacher that is scholar than the ANN itself. For example, the teacher feeds some example data about which the teacher already knows the answers.

For example, pattern recognizing. The ANN comes up with guesses while recognizing. Then the teacher provides the ANN with the answers. The network then compares it guesses with the teacher’s “correct” answers and makes adjustments according to errors.

- **Unsupervised Learning** – It is required when there is no example data set with known answers. For example, searching for a hidden pattern. In this case, clustering i.e. dividing a set of elements into groups according to some unknown pattern is carried out based on the existing data sets present.
- **Reinforcement Learning** – This strategy built on observation. The ANN makes a decision by observing its environment. If the observation is negative, the network adjusts its weights to be able to make a different required decision the next time.

**Example:**

**Artificial neural network for Gall midge incidence in rice**

Lavenberg-Marquardt algorithm was employed for training the data set which is two-layer feed-forward network having input as weather parameters to predict and compare with the target (Pest/Disease). The Artificial neural network for Gall midge with one hidden unit in the hidden layer with sigmoid function as an activation function and which was of the form  $\{g(\text{netweather})=1/(1+e^{-\text{netweather}})\}$  while the linear function  $\{g(\text{netweather})=\text{netweather}\}$  as an activation function for output neurons. Therefore, 3 weights for input to hidden neurons, 1 weight for hidden to output neurons and two bias weights were chosen. For training network 75 % of data, each for validation and testing 12.5 % data were used by random division process and the weights of 3 input value in input layer were denoted by  $I_i$  ( $i=1, 2, 3$ ), Weights of one hidden unit were denoted as  $H_j$  ( $j=1$ ) and weight of output node was denoted by  $O$ , here in this study researcher used only one output node. The bias weights were denoted as (One hidden and one Output bias)  $B_1H$ ,  $B_2O$  and the performance of the L-M trained neural network regression model was accessed by their Mean square error (MSE) in the validation stage and coefficient of determination ( R-square ) between input and the target. Weights for input, hidden output and bias are obtained as,

$I_1H_1$	-3.43	$I_2H_1$	-1.509	$I_3H_1$	-0.164
		$H_2O_1$	1.522		
		$B_1H$	4.432		
		$B_2O$	-0.482		

Fig: Neural Network Architecture for Predicting Gall Midge Incidence

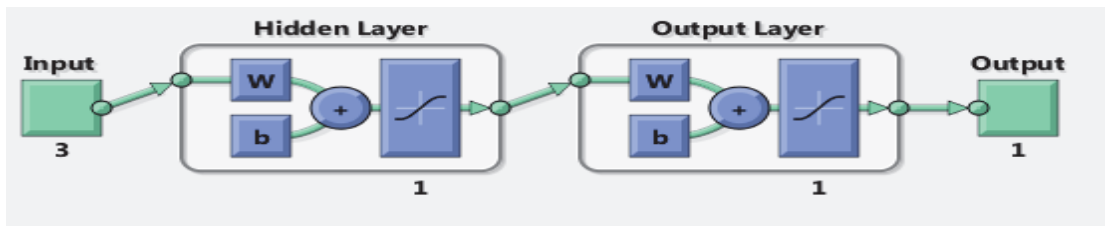
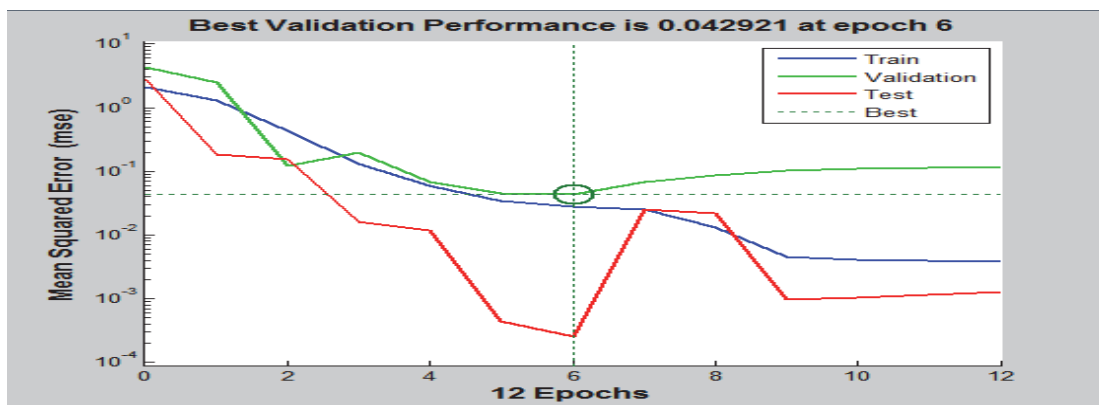
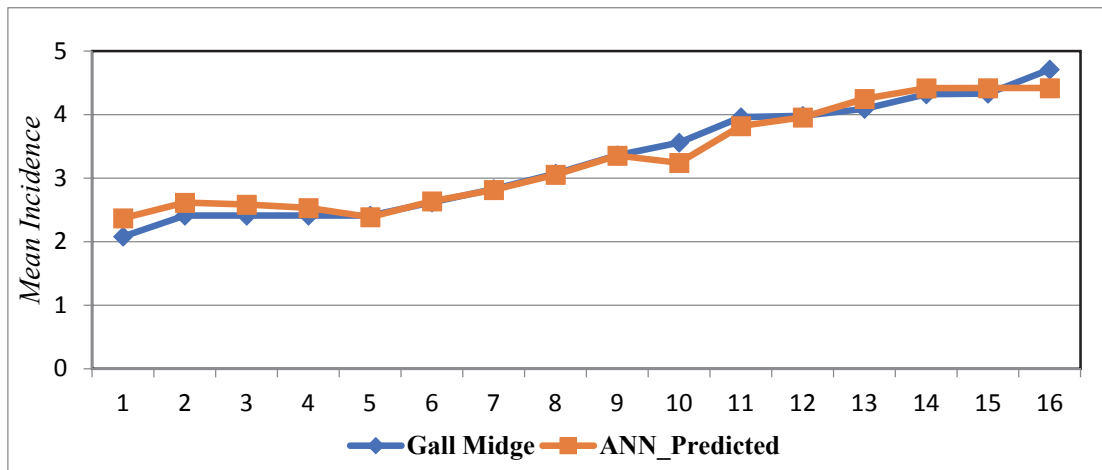


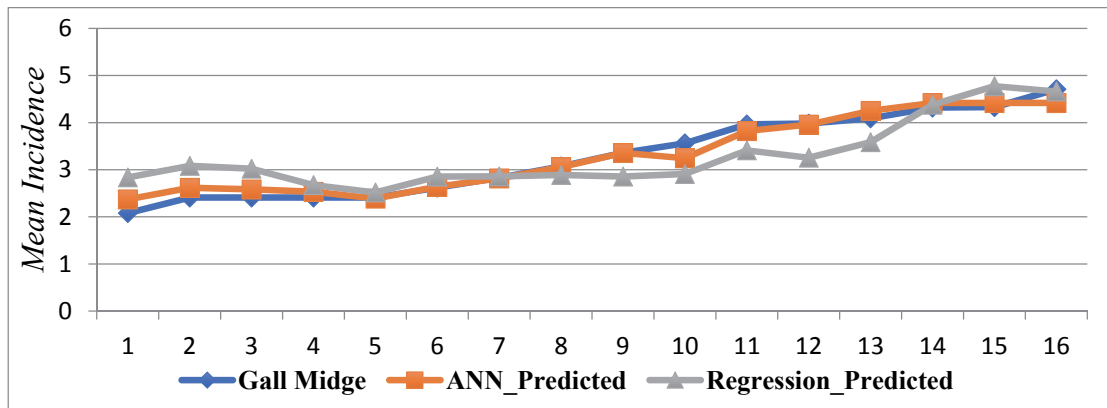
Fig: Best Validation



Observed and Predicted Using ANN (Closely Predicted By ANN)



Superiority of ANN over Regression Method



**References:**

Bogdan M. Wilamowski, Hao Yu: Neural Network Learning Without Backpropagation. IEEE Trans. Neural Networks 21(11): 1793-1803 (2010).

McCulloch, Warren; Walter Pitts (1943). "A Logical Calculus of Ideas Immanent in Nervous Activity". Bulletin of Mathematical Biophysics. 5 (4)



## Non-parametric Test

Ajit Kumar Das  
BCKV, Mohanpur, Nadia-741252

### 1. Introduction

A non-parametric test is a test that does not depend on the particular form of the basic frequency function from which the samples are drawn. In other words, non-parametric test does not make any assumption regarding the form of the population. Below we shall mention briefly the comparative study of parametric test and non-parametric test.

1. If a sample is drawn from a population and the form of the population distribution (say, normal distribution) is known and the observations are random and independent, then one can use parametric test. In non-parametric test, no assumption is made about the form of frequency function of the parent population from which the sample is drawn.

2. No parametric technique will be applicable to the data which are mere classification (i.e., which are measured in nominal scale) but non-parametric method exists to deal with such data. Since the socio-economic data are not, in general, normally distributed, non-parametric tests have found applications in psychometry, sociology and educational statistics.

3. Non-parametric tests are available to deal with data which are given in ranks or whose seemingly numerical scores have the strength of the ranks. For example, no parametric test can be applied if the scores are given in grades such as A, B, C, D, etc. However, in non-parametric tests, we make an assumption that the sample is drawn from a population having continuous distribution function  $F(x)$ . Many non-parametric tests are available in the literature. We discuss some of them.

#### One sample non-parametric tests:

- (i) Ordinary sign test
- (ii) Wilcoxon signed- rank test
- (iii) Run test
- (iv) Kolmogorov-Smirnov test

#### Paired sample non-parametric tests:

- (i) Sign test (ii) Wilcoxon signed- rank test

#### Two or more samples non-parametric tests:

- (i) ) Kolmogorov-Smirnov two sample test
- (ii) Mann-Whitney U-test
- (iii) Kruskal-Wallis one way analysis

**Measures of association for bivariate samples:**

- (i) Spearman’s rank correlation coefficient
- (ii) Kendall’s rank correlation coefficient

**2. One sample non- parametric tests**

**(i) Ordinary Sign test**

The simplest possible non-parametric test about the location parameter only of a population is a sign test. Thus to test the null hypothesis  $H_0 : \theta = \theta_0$  where  $\theta$  is the median of a continuous distribution, we replace each sample observation ( in the order they are obtained) either by ‘+’ sign or by ‘-’ sign according as it is greater than or smaller then  $\theta_0$  . However any observation equal to  $\theta_0$  is simply discarded. Then the total number of ‘+’ and ‘-’ signs will be  $m$  which should be  $\leq$  sample size  $n$ . The number of ‘+’ sign will then follow a binomial distribution with parameter  $m$  and  $p = \frac{1}{2}$  . Right (or left) tail end of a symmetrical binomial distribution will serve as a critical region for testing  $H_0 : \theta = \theta_0$  against  $H_1 : \theta > \theta_0$  (or  $\theta < \theta_0$ ). For testing  $H_0 : \theta = \theta_0$  against  $H_1 : \theta \neq \theta_0$  we should consider both tail end of the symmetrical binomial distribution as the critical region.

**(ii) Wilcoxon signed- rank test**

This test is used to test the hypothesis that the observations have come from symmetrical population with a common specified median , say ,  $\mu_0$  . Thus the problem is to test

$H_0 : \mu = \mu_0$  . The signed-rank statistic  $T^+$  is computed as follows : 1. Subtract  $\mu_0$  from each observation.

- 2. Rank the resulting differences in order of size , discarding sign.
- 3. Restore the sign of the original difference to the corresponding rank .
- 4. Obtain  $T^+$  , the sum of the positive ranks.

Similarly,  $T^-$  is the sum of the negative ranks. Then under  $H_0$ , we expect  $T^+$  and  $T^-$  to be the same. We also note that

$$T^+ + T^- = \sum_{i=1}^n i = \frac{n(n+1)}{2} .$$

The statistic  $T^+$  (or  $T^-$ ) is known as the Wilcoxon statistic. A large value of  $T^+$  ( or equivalently, a small value of  $T^-$  ) means that most of the large deviation from  $\mu_0$  are positive and therefore we reject  $H_0$  in favour of the alternative  $H_1 : \mu > \mu_0$  .

Thus the test rejects  $H_0$  at the level  $\alpha$

- if  $T^+ < C_1$  when  $H_1 : \mu < \mu_0$
- if  $T^+ > C_2$  when  $H_1 : \mu > \mu_0$
- if  $T^+ < C_3$  or  $T^+ > C_4$  when  $H_1 : \mu \neq \mu_0$

where  $C_1, C_2, C_3$  and  $C_4$  are such that

$$P[T^+ < C_1] = \alpha$$

$$P[T^+ > C_2] = \alpha$$

$$P[T^+ < C_3] + P[T^+ > C_4] = \alpha$$

### (iii) Run test

Suppose we have  $n$  observations. We like to test  $H_0$ : The set of observations are random against  $H_1$ : They are not random.

At first we find the median of the sample observations. Then we replace each original observation either by '+' or '-' sign according as it is larger or smaller than the median. Any observation equal to median is simply discarded. A run is defined to be a sequence of values of the same kind bounded by the values of other kind. We compute the total number of runs 'r'. Too many values of 'r' as well as too small values of 'r' give an indication of non-randomness. Thus the test rejects  $H_0$  at the level  $\alpha$  if  $r < r_1$  or  $r > r_2$  where  $r_1$  and  $r_2$  are such that

$$P[r < r_1] = \alpha/2, P[r > r_2] = \alpha/2.$$

The one-sample run test is based on the order or sequence in which the individual scores or observations originally were obtained.

### (iv) Kolmogorov-Smirnov test

Let  $X_1, X_2, \dots, X_n$  be a sample from continuous distribution function  $F(x)$ . We are to test  $H_0: F(x) = F_0(x) \forall x$  against  $H_1: F(x) \neq F_0(x)$  for some  $x$ . Suppose  $F_n(x)$  is the sample (empirical) distribution function corresponding to any given  $x$ ; that is, if the number of observation  $\leq x$  is  $k$ , then

$$F_n(x) = \frac{k}{n}.$$

Test statistic under  $H_0$  is given by

$$D_n = \text{Sup} |F_n(x) - F_0(x)|$$

which is known as Kolmogorov-Smirnov statistic. The distribution of  $D_n$  does not depend on  $F_0$  as long as  $F_0$  is continuous.  $H_0$  is rejected if  $D_n > D_{n,\alpha}$ . Similarly,

the one-sided KS statistics for one sided alternatives are the following: (i) for the alternative  $H^+ : F(x) \geq F_0(x) \forall x$  the appropriate statistic is

$$D_n^+ = \text{Sup}[F_n(x) - F_0(x)]$$

(ii) for the alternative  $H^- : F(x) \leq F_0(x) \forall x$  the appropriate statistic is

$$D_n^- = \text{Sup}[F_0(x) - F_n(x)].$$

The statistics  $D_n^+$  and  $D_n^-$  have the same distribution because of symmetry. The test rejects  $H_0$  if  $D_n^+ > D_{n,\alpha}^+$  when alternative is  $F(x) \geq F_0(x) \forall$  and rejects  $H_0$  if  $D_n^- > D_{n,\alpha}^-$  when

alternative is  $F(x) \leq F_o(x) \forall x$  at the level  $\alpha$ .

### 3. Paired sample non-parametric tests

#### (i) sign test

The single-sample sign test can be easily modified to apply to sampling from a bivariate population. Let  $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$  be a random sample from a bivariate population. Let  $d_i = X_i - Y_i, i=1, 2, \dots, n$  and assume that  $d_i$  has an absolutely continuous distribution. Our null hypothesis is  $H_0: \text{Median}(d)=0$ . So in paired sample case instead of dealing with the sample values, one has to deal with the differences  $d$  and perform the test in the same way as we do for one sample case. This test is the non-parametric version of paired 't' test.

#### (ii) Wilcoxon signed-rank test

The sign test utilizes information simply about the direction of the differences within pairs. If the relative magnitude as well as the direction of the differences is considered, a most powerful test can be made. The Wilcoxon matched-pairs signed-ranks test gives more weight to a pair which shows a large difference between the two conditions than to a pair which shows a small difference. If  $M_d$  is the median of the population of differences, then the null hypothesis is that  $M_d=0$  and the alternative hypothesis is one of  $M_d > 0, M_d < 0$  or  $M_d \neq 0$ . The observed differences  $d_i$  are ranked in increasing order of absolute magnitude and the sum of ranks is computed for all the differences of like sign. The test statistic  $T$  is the smaller of these two rank-sums. Pairs with  $d_i=0$  are not counted. Now conclusion of the test procedure remains same as in case of one sample test.

### 4. Two or more samples non-parametric tests

#### (i) Mann-Whitney U-test

Suppose we have two independent samples  $X_1, X_2, \dots, X_m$  and  $Y_1, Y_2, \dots, Y_n$  from two populations with continuous distribution functions  $F$  and  $G$  respectively.

Test  $H_0: F(x) = G(x) \forall x$  against  $H_1: F(x) \neq G(x)$  for some  $x$ . Thus we want to test whether two samples have come from the same population. The only assumption is that  $F(x)$  and  $G(x)$  are both continuous. Under the Mann-Whitney procedure, combine the data of two samples and arrange them in ascending order. Keep track which observation belongs to which sample. Statistic  $U$  is defined as the number of times  $Y$ 's precede the  $X$ 's in the combined sequence of  $(m+n)$  variate values. The value of  $U$  can be obtained by considering the sum of the ranks  $R_2$  of  $Y$ 's (the case when  $Y$  precedes  $X$ ) in the ordered combined sequence. The formula for  $U$  is  $U = mn + \frac{n(n+1)}{2} - R_2$ . Similarly, if the ranks of  $X$ 's are counted (the case when  $X$  precedes  $Y$ ), the value of  $U'$  can be obtained by the formula  $U' = mn + \frac{m(m+1)}{2} - R_1$  where  $R_1$  is the sum of ranks of  $X$ 's in the combined sequence of  $X$ 's and  $Y$ 's. It is noted that  $U' = mn - U$ . Clearly  $U=0$  if all the  $Y$ 's are larger than  $X$ 's and  $U=mn$  if all the  $Y$ 's are smaller than  $X$ 's. Thus  $0 \leq U \leq mn$ . Calculated value of  $U$  is compared with

table of U with different values of m, n and  $\alpha$  and conclusion can be drawn accordingly.

**(ii) Kolmogorov-Smirnov two sample test**

Let  $X_1, X_2, \dots, X_m$  and  $Y_1, Y_2, \dots, Y_n$  be independent random samples from continuous distribution functions F and G respectively. Let the order statistics be  $X_{(1)}, X_{(2)}, \dots, X_{(m)}$  and  $Y_{(1)}, Y_{(2)}, \dots, Y_{(n)}$ . Let us write

for the empirical distribution functions

$$F_m(x) = \begin{cases} 0 & , \quad x < X_{(1)} \\ \frac{k}{m} & , \quad X_{(k)} \leq x < X_{(k+1)}, \quad k = 1, 2, \dots, m-1 \\ 1 & , \quad x \geq X_{(m)} \end{cases}$$

and

$$G_n(x) = \begin{cases} 0 & , \quad x < Y_{(1)} \\ \frac{k}{n} & , \quad Y_{(k)} \leq x < Y_{(k+1)}, \quad k = 1, 2, \dots, n-1 \\ 1 & , \quad x \geq Y_{(n)} \end{cases}$$

In a combined ordered arrangement of m X's and n Y's,  $F_m$  and  $G_n$  represent the respective proportions of X and Y values that do not exceed x. Under  $H_0 : F(x) = G(x) \forall x$ , we expect a reasonable agreement between the two sample df's. We define

$$D_{m,n} = \text{Sup} |F_m(x) - G_n(x)|.$$

Then  $D_{m,n}$  may be used to test  $H_0$  against  $H_1 : F(x) \neq G(x)$  for some x. The test rejects  $H_0$  at level  $\alpha$  if  $D_{m,n} \geq D_{m,n;\alpha}$  where  $P_{H_0} [D_{m,n} \geq D_{m,n;\alpha}] \leq \alpha$ .

**(iii) Kruskal-Wallis test:**

Kruskal-Wallis test is one of the most frequently used method in non-parametric statistics for analysing data in one way classification. It is equivalent to one way analysis of variance in parametric methods.

Kruskal-Wallis test is based on the following assumptions:

1. The observations are independent within and between samples
2. The variable under study is continuous.
3. The populations are identical except possibly in respect of median.

We test  $H_0$ : All the k populations are identical against  $H_1$ : At least one pair of populations do not have the same median.

Let there be k independent samples from k populations of sizes  $n_1, n_2, \dots, n_k$ . The observations in k samples can be presented in the tabular form as given below:

Sample Numbers

1	2	..... i .....	k
X11	X21	Xi1	Xk1
X12	X22	Xi2	Xk2
.	.	.	.
.	.	.	.
.	.	.	.
$x_{1n_1}$	$x_{2n_2}$	$x_{in_i}$	$x_{kn_k}$

Now we assign rank to each observation from 1 to  $n = \sum_{i=1}^k n_i$  by pooling all the sample

observations and write them in ascending order. The sum of ranks is equal to  $\frac{n(n+1)}{2}$ .

Suppose  $R_i$  is the actual sum of ranks of observations in sample i .

To test  $H_o$ , Kruskal-Wallis test statistic is a weighted sum of squares of deviations of the sum of ranks of the treatments from the expected sum of ranks, using reciprocals of sample size as the weights.

Thus Kruskal-Wallis test statistic is

$$H = \frac{12}{n(n+1)} \sum_{i=1}^k \frac{1}{n_i} \left[ R_i - \frac{n_i(n+1)}{2} \right]^2 = \frac{12}{n(n+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} - 3(n+1).$$

The statistic H is approximately distributed as  $\chi^2$  with (k-1)d.f. subject to the condition that  $n_i$  should be large i.e. each  $n_i$  should not be less than 5.  $H_o$  is rejected if calculated  $\chi^2 >$  table value of  $\chi^2$  at  $\alpha$  level of significance and (k-1) d.f. Otherwise it is accepted.

**Note:** If there are a number of ties, one must adjust H for ties. The adjustment factor

$$C = \frac{\sum T}{n(n^2 - 1)}$$

where  $T = (t^3 - t)$  for t, the number of tied observations in a group and  $\sum$  is

over all such groups. The corrected test statistic is  $H_c = \frac{H}{C}$ .

**5. Measures of association for bivariate sample**

**(i) Spearman’s rank correlation coefficient:**

In many situations, the individuals are ranked by two judges or the measurements taken for two variables are assigned ranks within the samples independently. Now it is desired to know the extent of association between the ranks. The method of calculating the association between ranks was given by Charles Edward Spearman in 1906 and is known as Spearman’s rank correlation.

Let  $(X_i, Y_i)$  ( $i=1,2,\dots,n$ ) be a sample from a bivariate population. If the sample values  $X_i$  and  $Y_i$  are each ranked from 1 to  $n$  in increasing order of magnitude separately and if the  $X$ 's and  $Y$ 's have continuous distribution functions, we get a unique set of rankings. If for  $n$  individuals,  $d_i$  is the difference between ranks of the  $i^{\text{th}}$  individual for  $i=1, 2,\dots,n$ ; the formula for Spearman's rank correlation is

$$r_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}.$$

The value of  $r_s$  lies between -1 and +1. If  $X$  and  $Y$  are independent then  $E(r_s)=0$ . Also population Spearman's rank correlation coefficient i.e.  $\rho_s = 0 \Rightarrow E(r_s)=0$ . Here we test

$H_0 : \rho_s = 0$  against  $H_1 : \rho_s \neq 0$ . The test statistic  $t = \frac{r_s \sqrt{n-2}}{\sqrt{1-r_s^2}}$  has  $(n-2)$  d.f. The decision about

$H_0$  is taken in the usual way. For large samples under  $H_0$ , the random variable  $Z = r_s \sqrt{n-1}$  has approximately a standard normal distribution. The approximation is good for  $n \geq 10$ .

### (ii) Kendall's rank correlation coefficient

Kendall's rank correlation coefficient  $\tau$  is suitable for the paired ranks as in case of Spearman's rank correlation. We now need to find an estimate of  $\tau$  from the sample. Using sample observations, Kendall's measure of association becomes

$$T = \frac{\text{Actual value}}{\text{Maximum possible value}} = \frac{2S}{n(n-1)}.$$

The procedure for calculating  $T$  consists of the following steps:

**Step 1:** Arrange the rank of the first set ( $X$ ) in ascending order and rearrange the ranks of the second set ( $Y$ ) in such a way that  $n$  pairs of rank remain the same.

**Step 2:** After operating Step 1, the ranks of  $X$  are in natural order. Now we are left to determine how many pairs of ranks on the set  $Y$  are in their natural order and how many are not. A number is said to be in natural order if it is smaller than the succeeding number and is coded as +1 and also if it is greater than its succeeding number then it will not be taken in natural order and will be coded as -1. In this way all pairs of the set ( $Y$ ) will be considered and assigned the values +1 and -1.

**Step 3:** Find the sum 'S' of all the coded values. Here we test  $H_0 : \tau = 0$  against  $H_1 : \tau \neq 0$ .

Thus we reject  $H_0$  if the observed value of  $|T| > t_{\alpha/2}$  where  $P[|T| > t_{\alpha/2} / H_0] = \alpha$ . The values of

$t_\alpha$  are given in the table for selected values of  $n$  and  $\alpha$ . If  $n$  is large under  $H_0 : \tau = 0$ ,  $\frac{3\sqrt{n}}{2}T \sim N(0,1)$  and we can test the independence of  $x$  and  $y$ .

### References

1. Rohatgi, V.K. (1976): An Introduction to Probability Theory and Mathematical Statistics, Wiley Eastern Limited, New Delhi
2. Sahu, P.K., Pal, S.R. and Das, A.K. (2015): Estimation and Inferential Statistics, Springer, India
3. Siegel, Sidney (1956): Non-parametric Statistics for the behavioural sciences, International Student Edition, Tokyo



## Use of Auxiliary Information in Sample surveys

Ajit Kumar Das

BCKV, Mohanpur, Nadia-741252

### 1. Introduction

The purpose of a sample survey is the collection of information by observing only a part of the population for drawing an inference about the characteristic of a population or universe.

A **finite population** is a collection of  $N$  (given) well defined, identifiable and observable elements (units). The number of elements in a finite population is called population size. Usually it is denoted by  $N$ . A collection of agricultural fields in a village, a collection of households in an area, a collection of industrial units in an urban area are some examples of survey populations. Let  $y$  be a study variable having values  $Y_1, Y_2, \dots, Y_N$ . These are the population values with respect to the characteristic  $y$  under study. Any real valued function of the population values is called **parameter**. For example, the population mean, population variance, population range etc. are parameters. Usually parameters are not known but we need to estimate them.

Now we define a sample. A **sample** is a small part or a representative part of the population. The number of elements in a sample is denoted by  $n$  and it is referred to as sample size. After the sample is selected, data are collected from the sample units. We shall denote by  $y_i$  the value of  $y$  on the unit selected at the  $i^{\text{th}}$  draw ( $i=1, 2, \dots, n$ ). Any real valued function of the sample values is called **statistic**. For example, the sample mean, sample variance, sample proportion etc. are statistics. A statistic is called an estimator when it is used to estimate a parameter. Statistics are random variables due to sampling fluctuation but parameters are constants. We mention that a sample is a small part of the population. Now the question is: how do we choose? Choosing a small part of the population according to a probability sampling scheme is called **probability sampling**. One of the simplest method of drawing a sample from a population is simple random sampling.

**Simple random sampling(SRS)** is a method of drawing a sample of size  $n$  from a population of  $N$  units such that each and every unit in the population have an equal probability of being included in the sample. Under this sampling procedure with replacement, sample mean is an unbiased estimator for the population mean and the sampling variance of the sample mean is  $\sigma^2/n$  where  $\sigma^2$  is the population variance.

Since  $\sigma^2$  is unknown, it is estimated unbiased by the sample mean square  $s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$ .

Thus unbiased estimate of  $V(\bar{y}) = s^2/n$  and estimate of standard error of  $\bar{y}$  is  $s/\sqrt{n}$ . The precision of an unbiased estimator  $e_1$  with respect to another unbiased estimator  $e_2$  is measured as  $\frac{V(e_2)}{V(e_1)} \times 100\%$ . If the precision exceeds unity then  $e_1$  is an estimator superior to  $e_2$ .

If the population is heterogeneous (with respect to the character under study) then the whole population is divided into several subpopulations or strata each of which is more homogeneous than the entire population and samples are drawn independently from each stratum. This is known as **stratified sampling**. It is noted that it is not necessary to use the same sampling scheme in all strata. Depending on the nature of the strata, different sampling schemes can be used in different strata. It is noted that, under stratified sampling, the variance of the estimator can be reduced from the one based on a random sample drawn from the whole population.

So far we have considered only character  $y$  under study. In some situations it may happen that other information about the population units are available or may be made available. For example, in the above mentioned populations, the area of agricultural fields, household size, number of workers in industrial units etc. are the auxiliary variables( $x$ ). The auxiliary information may be utilized in three basic ways:

- (i) At the pre-selection stage or the designing stage. For example, the information may be used in stratifying the population.
- (ii) At the selection stage. For example, in selecting the units with unequal probabilities; the probabilities of selection being based on the values of an auxiliary variate  $x$ .
- (iii) At the post selection stage or at the estimation stage. For example, in constructing the estimators, for a parameter, using the available information.

Obviously the auxiliary information may be used in mixed ways also.

## 2. Auxiliary information used in stratified sampling

One way to estimate the population mean with greater precision is to divide the population into several groups each of which is more homogeneous in respect of the  $y$ -values than the entire population. Since  $y$ - values are not known, units which are believed to be homogeneous from some prior knowledge or from the knowledge about the distribution of some closely related variable  $x$  are grouped together. Let us take one example:

In a village there are 50 agricultural fields and the area of agricultural fields ( $x$ ) is known. The character  $y$  under study is the yield and the auxiliary character  $x$  is the area. The values of  $y$  in the population are not known. In the absence of knowledge about the distribution of  $y$ , the frequency distribution of a closely related auxiliary variable  $x$  may be used to form homogeneous groups. On the basis of area of agricultural fields we have the following table:

Stratum boundaries(area in suitable unit)	Number of agricultural fields ( $N_h$ )	Area under wheat crop $X_h$ (in suitable unit)	$n_h = \frac{n}{N} N_h$	$n_h = \frac{n}{X} X_h$
<50	10	$X_1$	3	$\frac{n}{X} X_1$
50-100	25	$X_2$	7.5	$\frac{n}{X} X_2$
>100	15	$X_3$	4.5	$\frac{n}{X} X_3$
Total	50	$X$	15	15

For construction of strata the auxiliary information on the area of agricultural fields is used and another information on area under wheat crop for each stratum is obtained.

Suppose a sample of 15 agricultural fields is to be drawn. Once the sample size  $n$  is fixed next arises the question of deciding the sample size  $n_h$  meant for the stratum  $h$ ,  $h=1,2,\dots,L$ . One allocation is proportional allocation. Under this allocation,

$$n_h = \frac{n}{N} N_h, h=1,2,\dots,L.$$

For allocating the sample size in different stratum, we may use auxiliary character on area ( $X_h$ ) under wheat crop i.e.  $n_h$  is proportional to the size  $X_h$  of stratum. Under this allocation,

$$n_h = \frac{n}{X} X_h, \text{ where } X = \sum_h X_h.$$

In this example the auxiliary information on area of agricultural fields is used for construction of strata and the other auxiliary information on area under wheat crop for each stratum is used for allocation of sample units for different strata. From the above table, out of  $n=15$ , 3 fields are to be selected from the first stratum, 8 fields from second stratum and 4 fields from third stratum using proportional allocation.

### 3. Auxiliary information used at selection stage

In sampling from a finite population, often the value of some auxiliary character  $x$  closely related to the study variable  $y$  are available for all the units of the population. In such a situation, instead of sampling the units with equal probability with replacement or without replacement one may sample the units with probability proportional to size-measure  $x$  with replacement or without replacement. Let  $X_i$  and  $Y_i$  be the values of the size variable and the study variable for  $i^{\text{th}}$  population unit,  $i=1,2,\dots,N$ . The purpose is to select units with probability proportional to size measure  $X_i$  and with replacement (**PPSWR**). Here

$$P_i \propto X_i \Rightarrow P_i = kX_i. \text{ Summing both sides we get } k = \frac{1}{X} \text{ where } X = \sum_{i=1}^N X_i = \text{population total of the}$$

auxiliary variable  $x$ . Hence  $P_i = \frac{X_i}{X}$  for  $i=1,2,\dots,N$ . In order to select units with these probabilities, one can use either ‘‘Cumulative Total Method’’ or ‘‘Lahiri’s Method’’.

#### Cumulative Total Method

A table of cumulative total of sizes of the units is made. Let  $T_1=X_1$ ,  $T_2=X_1+X_2$ ,  $T_3=X_1+X_2+X_3, \dots, T_N=X_1+X_2+\dots+X_N$ . A random number, say  $R$ , is drawn between 1 and  $T_N (=X)$ . The unit  $i$  is selected if  $T_{i-1} < R \leq T_i$ . The process is repeated  $n$  times to get a sample of size  $n$ . It can be seen that in this method,

$$\text{the probability of selecting the } i^{\text{th}} \text{ unit in a given draw is } \frac{T_i - T_{i-1}}{X} = \frac{X_i}{X}.$$

#### Lahiri’s Method

If the number of units in the population is very large, cumulation of sizes may be tedious. This is avoided in this method. Let  $M$  be an integer greater than or equal to the maximum of the sizes  $X_1, X_2, \dots, X_N$ .

**Step 1** Select a random number  $i$  from 1 to  $N$ .

**Step 2** Select a random number  $R$  from 1 to  $M$ .

If  $R \leq X_i$ , then  $i^{\text{th}}$  unit is selected. Otherwise reject the unit  $i$  and repeat the trial (i.e. Steps 1 and 2) till a unit is selected. The whole process is repeated  $n$  times. It can be shown that in this method also the probability of selecting of  $i^{\text{th}}$  unit remains  $P_i$  at each draw.

### 3. Auxiliary information used at estimation stage

In sections 2 and 3 we have used auxiliary information in making stratification and in drawing a sample respectively. Here we use auxiliary variable  $x$  for estimating the population mean or total of a study variable  $y$  in order to get improved estimators. Suppose a sample of size  $n$  is drawn from a population of  $N$  units with equal probability (SRS). Two such methods are (i) ratio method and (ii) regression method. The ratio estimator and the regression estimator of the population mean  $\bar{Y}$  of a study variable  $y$  is given by

$$\hat{Y}_R = \frac{\bar{y}}{\bar{x}} \bar{X} \quad \text{and} \quad \hat{Y}_{Rg} = \bar{y} - b(\bar{x} - \bar{X})$$

respectively where  $\bar{y}$  and  $\bar{x}$  are the sample means of the characters  $y$  and  $x$ ;  $\bar{X}$ , the population mean of the character  $x$ , must be known and  $b$  is the sample regression coefficient of  $y$  on  $x$ . It is shown that both the estimators are biased. The ratio estimator is more efficient than the usual mean estimator  $\bar{y}$  if

$$\rho > \frac{1}{2} \cdot \frac{C_x}{C_y}$$

where  $C_y$ ,  $C_x$  are the population coefficient of variations of  $y$  and  $x$ ;  $\rho$  is the population correlation coefficient between  $x$  and  $y$ . The regression estimator is always efficient than  $\bar{y}$  unless  $\rho = 0$  and the regression estimator is always efficient than the ratio estimator unless  $\beta = R$  where  $\beta$  is the population regression coefficient of  $y$  on  $x$  and  $R = \frac{\bar{Y}}{\bar{X}}$  is the ratio of two population means.

If  $\rho = 0$  i.e. correlation coefficient is 0 then there is no use of auxiliary character  $x$  and we simply use the usual mean estimator  $\bar{y}$ . It is also noted that when  $\beta = R$  i.e.  $\bar{Y} = \beta \bar{X}$  then ratio estimator and regression estimator are equally efficient but bias of ratio estimator is zero. We can conclude that if the regression line of  $y$  on  $x$  is perfectly linear but passes through the origin then one should prefer ratio estimator than regression estimator.

The ratio and regression estimators assume the knowledge of the population mean  $\bar{X}$  of the auxiliary variable  $x$ . However there are some situations where the population mean of the auxiliary variable will not be known in advance. In such cases, two-phase sampling can be used for getting estimators. Sometimes it may happen that  $p(\geq 2)$  auxiliary variables are available. In this situation one may define different estimators for the population mean of a study variable  $y$ .

## References

Mukhopadhyay, Parimal (2000): Theory and Methods of Survey Sampling, Prentice-Hall of India, New Delhi

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

## Regression Analysis and Its Application

Prof. P K Sahu

Correlation coefficient measures the degree of linear association between any two given variables. But the actual relationship between or among the variables is given by the line of regression. The technique by which we can analyze the relationship among the correlated variables is known as regression analysis in the theory of statistics. Francis Galton was to coin the term 'Regression' in his famous paper "Family likeness in stature" in the proceedings of Royal Society, London in 1886. **Regression analysis is the study of dependence of one variable (the dependent variable) on one or more independent (explanatory variables) variables.** In agricultural and other experiment mainly three types of variables are recorded : a) the treatments or factors such as variety, insecticide, doses or type of fertilizers, different chemical treatments, different management practices etc., b) Environmental parameters like rainfall, temperature, humidity, sunshine hours, wind speed etc. and c) various responses in the form of different growth and yield parameters, qualitative changes etc. Now the task of a statistician to work out the actual relationship between or among the variables under study. And this is being accomplished through regression analysis.

In different socio-economic studies different demographic, social, economical, educational etc. parameters are studied to find out the dependence of the ultimate variables, say adoption index, awareness, empowerment status etc on these parameters. Regression analysis is a technique by virtue of which one can study the relationship.

### **Objective:**

**Thus the main objective of regression analysis is to estimate and/or predict the average value of the dependent variable given the values for independent /explanatory variables.**

Thus in regression analysis the dependent variables is the function of one or more independent variables, and can be represented in the form of

$$Y = f(X_i)$$

This type of regression analysis is known as function/deterministic dependency. But in statistics one deals with random/stochastic variables *i.e.*, the variables having probability distributions. So in statistical regression the dependent variable is the function of one or more independent variables and the error term, which can be represented in the form of  $Y = f(X_i, u_i)$ .

### **Cause-Effect relationship ?**

As such statistical regression analysis does not imply cause and effect relationship between the explanatory variables and the dependent variable. The idea of causation never comes from statistical theories it comes from outside the area of statistics. To know more about causation one should consider Granger test of causality.

### **Prediction equation:**

The regression equation can also be used as prediction equation, under the assumption that the trend of change in Y (the dependent variables) corresponding to change in X (or the  $X_i$ s) (the independent variables) remains the same. Once the constants are estimated from a given set of observations, the value of the dependent variable corresponding to any value of (X) (or set of values of  $X_i$ s) within the range of X (or  $X_i$ s) can be worked out. To some extent the prediction can be made for Y for the value(s) of X ( $X_i$ s) beyond the range but not too far beyond the values taken for calculation.

**Type of Regression :**

Regression analysis may be of two main types

- (i) **Linear regression and**
- (ii) **Non-linear regression.**

Again both the above mentioned two types of regressions can be categorized in to

- a) **simple regression and**
- b) **multiple regression depending up on the number of variables present in the regression equations.**

What do we mean by linearity?

a) **Linearity in variables:** A regression equation is called linear regression equation in variables if none of the variables in the equation has got power other than unity;

b) **Linearity in parameter :** A regression equation will be called as linear regression equation in parameter if none of the parameters in the equation has got power other than unity.

$Y = \alpha + \beta_1 X_1 + \beta_1^2 X_2 + \dots + \varepsilon$  (1) is a linear regression equation in variable but

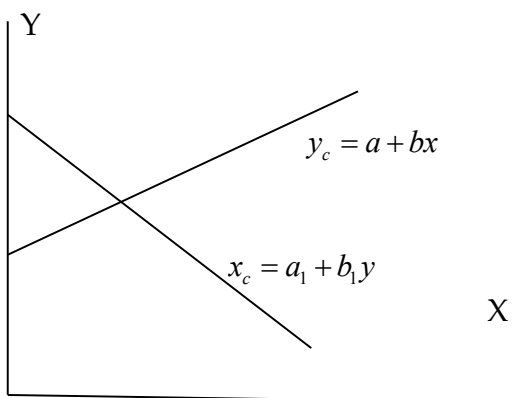
$Y = \alpha + \beta_1 X_1 + \beta_2 \sqrt{X_2} + \beta_3 X_3 + \dots + \varepsilon$  (2) is not a linear regression equation in variable.

Equation (1) is the example of regression equation non-linear in parameter where as equation (2) is the example of linear regression equation in parameter.

Generally, the regression equations, which are linear in parameter(s), are termed as linear regression equation.

**Simple linear regression analysis:**

The two regression lines i.e. Regression line of Y on X and X on Y can be represented as follows:



**a) Simple linear regression equation:**

Let we have a linear regression equation  $X_1 = 15 + 1.2X_2$ . We shall now examine what are the information we can have from the above relationship.

- i) The relationship between  $X_1$  and  $X_2$  is linear and it is an example of simple linear regression equation with two variables in the equation
- ii)  $X_1$  and  $X_2$  are the dependent and independent variable respectively in the relationship.
- iii) 15 is known as the intercept constant and it is the mean value of  $X_1$  under the given condition; the line of regression starts at 15 scale of the  $X_1$  axis.
- iv) 1.2 is known as the regression coefficient; it indicates that there would 1.2 unit change in the value of the dependent variable  $X_1$  with a unit change in the independent variable  $X_2$ . It is also the slope of the regression line.

**b) Multiple linear regression equation:**

Suppose we have a multiple linear regression equation  $X_1 = 10 + 1.2X_2 - 0.8X_3 + 1.7X_4 + 0.096X_5$ . From this relationship one can have the following information:

- i) The relationship between the variable  $X_1$  and the variables  $X_2, X_3, X_4$  and  $X_5$  is linear.
- ii) In the relationship  $X_1$  is the dependent and  $X_2, X_3, X_4, X_5$  are the independent variables
- iii) For the given set of values for the variables, the line of regression touches the Y axis at 10.
- iv) 1.2 (the coefficient of  $X_2$ ), 0.8 (the coefficient of  $X_3$ ), 1.7 (the coefficient of  $X_4$ ), 0.096 (the coefficient of  $X_5$ ) along with the intercept 10 are the parameters of the regression equation and are also known as the partial regression coefficients of the variable  $X_2, X_3, X_4, X_5$  respectively.
- v) From the partial regression coefficient one can infer that excepting  $X_3$  all other independent variables are positively correlated with the dependent variable  $X_1$ .
- vi) One unit change in  $X_2$  variable keeping other variable at constant level will result in 1.2 unit change (in the same direction) in the dependent variable  $X_1$ . Whereas one unit change in the variable  $X_3$  will result in 0.8 unit change in the dependent variable  $X_1$  in opposite direction; one unit increase in  $X_3$  will result in 0.8 unit decrease in  $X_1$ . Other regression coefficients can also be interpreted similar way.

**Assumption of linear regression model:**

The linear regression equation is based on certain assumptions, some of which are quite obvious but some are needed for better statistical treatment during further analysis towards drawing meaningful interpretation about the population under investigation.

- (1) The regression equation is linear in parameter i.e.  $X_1 = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \dots + \beta_k X_k + u$
- (2) Independent variables are non-stochastic i.e., values taken by  $X_2, X_3, X_4, X_5, \dots, X_k$  are fixed in repeated samples.
- (3) For a given set of values of  $X_2, X_3, X_4, X_5, \dots, X_k$ , the expected value of random variable  $u$  is zero i.e.,  $E(u_i) = 0$
- (4)  $Var(u_i) = E(u_i^2) - (E(u))^2 = E(u_i^2) = \sigma^2$  [ $\because E(u) = 0$  by assumption (i)] =  $\sigma^2$   
 $\Rightarrow Var(u_i/X_1) = Var(u_i/X_2) = Var(u_i/X_3) \dots = \sigma^2$
- (5) There should not be any auto-correlation between the disturbances  
 $Cov(u_i, u_j/X_i, X_j) = 0 \quad i \neq j$

- (6) Non existence of correlation between disturbances and independent variables  
 i.e.  $r_{u_i, X_i} = 0 \Rightarrow \text{Cov}(u_i, X_i)$   
 $= E[u_i - E(u_i)] [X_i - E(X_i)]$   
 $= E[u_i (X_i - E(X_i))]; [E[E(u_i)[X_i - E(X_i)]]$  vanishes, because  $E(u_i) = 0$   
 $= E(u_i \cdot X_i) - E(u_i) E(X_i)$   
 $= E(u_i X_i)$  must be equal to zero  
 i.e.,  $E(u_i X_i) = 0$
- (7) Multicollinearity should not exist among the independent variables, i.e.  $r_{X_i X_j} = 0$ ; otherwise there will be a problem of estimation of the regression parameters.
- (8) The number of observations ( $n$ ) must be greater than the number of parameters ( $j$ ) (number of variables in the regression equation) to be estimated i.e.,  $n > j$  ( $= 1, \dots, k$ )
- (9) In a given sample the independent variables ( $X$ 's) should be variable in true to the sense i.e. the values of  $X$ 's must not be constant i.e.,  $\text{Var}(X) > 0$
- (10) Correct specification of the regression model is an essential condition i.e., the model should clearly spell out (1) functional form of the model (linear in this case) (2) the variables and the number of variables to be included (3) probabilistic assumption about the variables.

**Example: Find out the correlation between weight of eggs and number of eggs laid per cycle in certain poultry bird and find regression of weight of eggs on hatching.**

Weight of egg(s)(g)	45	48	49	50	51	52	53	54	55	56	57	58	59	60	61
Hatching	80	80	85	88	92	92	90	91	92	92	89	86	84	82	80

**Solution:** From the given problem, we are to calculate the correlation coefficient between weights of eggs ( $Y$ ) and the hatching ( $X$ ); also we are to find out the regression equation of  $Y$  on  $X$ . Now from the above information, let us construct the following table:

Observations	Weight of eggs( $Y$ )	Hatching ( $X$ )	$(Y - \bar{Y})$	$(X - \bar{X})$	$(Y - \bar{Y})^2$	$(X - \bar{X})^2$	$(X - \bar{X})(Y - \bar{Y})$
1	45	80	-8.87	-6.87	78.62	47.15	60.88
2	48	80	-5.87	-6.87	34.42	47.15	40.28
3	49	85	-4.87	-1.87	23.68	3.48	9.08
4	50	88	-3.87	1.13	14.95	1.28	-4.38
5	51	92	-2.87	5.13	8.22	26.35	-14.72
6	52	92	-1.87	5.13	3.48	26.35	-9.58
7	53	90	-0.87	3.13	0.75	9.82	-2.72
8	54	91	0.13	4.13	0.02	17.08	0.55
9	55	92	1.13	5.13	1.28	26.35	5.82
10	56	92	2.13	5.13	4.55	26.35	10.95
11	57	89	3.13	2.13	9.82	4.55	6.68
12	58	86	4.13	-0.87	17.08	0.75	-3.58
13	59	84	5.13	-2.87	26.35	8.22	-14.72
14	60	82	6.13	-4.87	37.62	23.68	-29.85
15	61	80	7.13	-6.87	50.88	47.15	-48.98
<b>Sum</b>	<b>808.00</b>	<b>1303.00</b>	<b>0.00</b>	<b>0.00</b>	<b>311.73</b>	<b>315.73</b>	<b>5.73</b>
<b>Average</b>	<b>53.87</b>	<b>86.87</b>					



$$\text{Cov}(X, Y) = S_{xy} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{n} = \frac{5.73}{15} = 0.38$$

$$S^2_x = \frac{\sum (X - \bar{X})^2}{n} = \frac{315.73}{15} = 21.05$$

$$S^2_y = \frac{\sum (Y - \bar{Y})^2}{n} = \frac{311.73}{15} = 20.78$$

$$\therefore r_{xy} = \frac{\text{Cov}(X, Y)}{S_x S_y} = \frac{0.38}{\sqrt{(21.05 \times 20.78)}} = 0.0183$$

So there is very small correlation between the two variables.

**Regression analysis:**

Parameter estimation:  $b_{yx} = \frac{S_{xy}}{S^2_x} = \frac{0.38}{21.05} = 0.0182$

Intercept( $b_0$ ) =  $\bar{Y} - b\bar{X} = 53.87 - 0.0185 \times 86.87 = 52.29$

Hence regression equation of weight of eggs on hatching is  $Y = 52.29 + 0.0182X$

[Regression Exercise.xls](#)

**Properties of regression coefficient:**

**1. Regression coefficient measures** the amount of change expected in the dependent variable due to a unit change in the independent variable.

**2. The sign of regression coefficient is the sign of the correlation coefficient:**

We know that  $r_{x_1x_2} \frac{S_{x_1}}{S_{x_2}} = b_{x_1x_2}$

**3. Correlation coefficient is the geometric mean between the two regression coefficients**

**4. We have**

$$b_1 b_2 = r^2_{x_1x_2} \leq 1$$

$$\Rightarrow b_2 \leq \frac{1}{b_1} \text{ or, } b_1 \leq \frac{1}{b_2}$$

**5. We know,**

$$(\sqrt{b_1} - \sqrt{b_2})^2 \geq 0$$

$$\frac{b_1 + b_2}{2} \geq \sqrt{b_1 b_2} \geq r$$

**6. Regression coefficient does not have any range**

Regression coefficient of  $X_1$  on  $X_2$  is given as  $r_{x_1x_2} \frac{S_{x_1}}{S_{x_2}}$

We know  $-1 \leq r_{xy} \leq 1, \infty \geq S_x \geq 0$  and  $\infty \geq S_y \geq 0$

So the range of  $b_{yx}$  is

We know  $1 \geq r_{x_1x_2} \geq -1$ ,  $-\infty \leq S_{x_1} \leq \infty$  and  $-\infty \leq S_{x_2} \leq \infty$

So the range of byx is

$$b_{x_1x_2} = r_{x_1x_2} \frac{S_{x_1}}{S_{x_2}} = \pm 1 \frac{\infty \geq S_{x_1} \geq 0}{\infty \geq S_{x_2} \geq 0}$$

$$\therefore \infty \geq b_{x_1x_2} \geq -\infty$$

**7. Regression coefficients does not depend on change of origin but depend on scales**

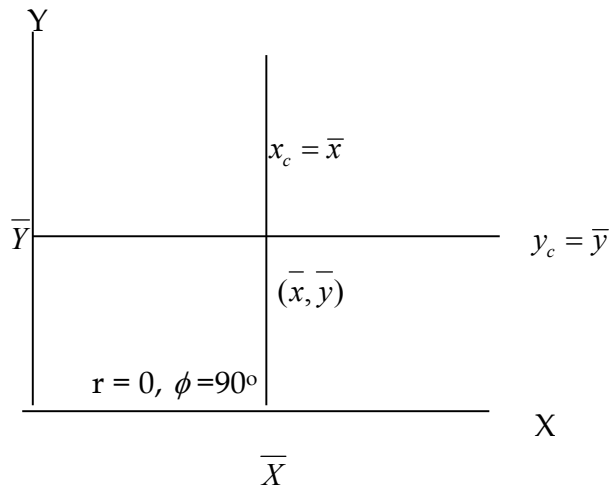
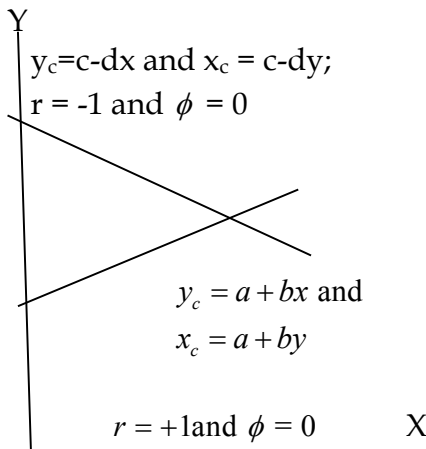
**8. Angle between two lines of regression**

$$\tan \theta = \frac{b_2 \sim b_1}{1 + b_1 b_2} = \frac{\frac{S_{x_2}}{r_{x_1x_2} S_{x_1}} - r_{x_1x_2} \frac{S_{x_2}}{S_{x_1}}}{1 + \frac{S_{x_2}}{S_{x_1}} \frac{S_{x_2}}{S_{x_1}}} = \frac{\frac{S_{x_2} - r_{x_1x_2}^2 S_{x_2}}{r_{x_1x_2} S_{x_1}}}{\frac{S_{x_1}^2 + S_{x_2}^2}{S_{x_1}^2}} = \frac{\frac{S_{x_2} (1 - r_{x_1x_2}^2)}{r_{x_1x_2} S_{x_1}}}{\frac{S_{x_1}^2 + S_{x_2}^2}{S_{x_1}^2}} = \frac{S_{x_2} S_{x_1} (1 - r_{x_1x_2}^2)}{(S_{x_1}^2 + S_{x_2}^2) r_{x_1x_2}}$$

$$\therefore \theta = \tan^{-1} \left[ \frac{(1 - r_{x_1x_2}^2)}{r_{x_1x_2}} \frac{S_{x_2} S_{x_1}}{(S_{x_1}^2 + S_{x_2}^2)} \right]$$

Putting  $r = \pm 1$  we have  $\theta = 0$  i.e. the two regression lines coincide each other and when  $r = 0$ ,  $\theta = 90^\circ$ , i.e. the regression lines are perpendicular to each other.

**9. If the variables are measured from their respective means then the regression equation passes through the origin.** [Regression Exercise.xls](#)



**Expectations and variances of the regression parameters**

The expectations of  $b_0$  and  $b_1$  are given as  $E(b_0) = \beta_0$  and  $E(b_1) = \beta_1$ . The corresponding variances of the estimators are given as

$$Var(b_0) = \sigma_{b_0}^2 = \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}_2^2}{SS_{xx}} \right), \text{ and } Var(b_1) = \sigma_{b_1}^2 = \frac{\sigma^2}{SS_{xx}}; \text{ where } \sigma^2 \text{ is the error variance of}$$

the regression model.

Under the normality assumption of the dependent variable and as both the regression estimators are linear functions of the dependent variable, so these are also assumed to behave

like normal variate. As estimator of  $\sigma^2$  is  $s^2$  thus, by replacing  $\sigma^2$  in the above variance estimates we have

$$Var(b_0) = \sigma_{b_0}^2 = \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}_2}{SS_{xx}} \right) = s^2 \left( \frac{1}{n} + \frac{\bar{x}_2}{SS_{xx}} \right), \text{ and } Var(b_1) = \sigma_{b_1}^2 = \frac{\sigma^2}{SS_{xx}} = \frac{s^2}{SS_{xx}} \text{ and the}$$

corresponding standard errors are the square roots of the variances. Here  $S^2$  is the residual mean sum of squares and is given as  $\frac{\sum_{i=1}^n (x_{1i} - \hat{x}_1)^2}{d.f.}$  and  $x_1$  is the dependent variable; d.f. is the

number of observation -no. of parameters estimated in the model, for simple regression equation model d.f would be  $n-2$ .

It should be noted that the variance of residuals is not equal to the error variance,  $Var(e_i) \neq \sigma^2$ . The residual variance depends on independent variable  $x_2$ . But when sample size is large,  $Var(e_i) \approx \sigma^2$ , which is estimated by  $s^2$ , that is,  $E(s^2) = \sigma^2$ .

**Test of significance for regression coefficient**

In order to test the same we have the following null and alternative hypotheses respectively:

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

Assuming that the dependent variable follow normal distribution, we have the following test statistic under the given hypotheses:

$$t = \frac{b_1 - 0}{SE(b_1)} = \frac{b_1}{\sqrt{\frac{s^2}{SS_{xx}}}} = \frac{b_1}{s} \sqrt{SS_{xx}} \text{ with } n-2 \text{ degrees of freedom. At } \alpha \text{ level of significance the null}$$

hypothesis  $H_0$  is rejected if the computed value of  $|t| \geq t_{\alpha/2, (n-2)}$ , where  $t_{\alpha/2, (n-2)}$  is a critical value of  $t$  distribution with  $(n - 2)$  degrees of freedom under  $H_0$ .

**Multiple linear regression analysis:**

A more general case is the multiple regression equation in which more than two variables are involved. Simple linear regression equation may be considered as special case of multiple linear regression equation in which only two variables are involved.

$X_1 = \beta_1 + \beta_2 X_2 + \dots + \beta_k X_k$

But instead of the population, we are provided with a sample, so from sample observations we are to work out a linear regression equation of the form  $X_1 = b_1 + b_2 X_2 + \dots + b_k X_k$  using the sets of values for  $X_1, X_2, \dots, X_k$ .

This equation is also true for all sets of observation where variables are measured from their respective means, that means.

Observation	$x_1$	$x_2$	$x_3$	$x_k$
1	$x_{11}$	$x_{21}$	$x_{31}$	$x_{k1}$
2	$x_{12}$	$x_{22}$	$x_{32}$	$x_{k2}$
3	$x_{13}$	$x_{23}$	$x_{33}$	$x_{k3}$
.....	.....	.....	.....	.....
$i$	$x_{1i}$	$x_{2i}$	$x_{3i}$	$x_{ki}$
.....	.....	.....	.....	.....
$n$	$x_{1n}$	$x_{2n}$	$x_{3n}$	$x_{kn}$

Now multiplying both sides of the above equation  $x_1 = b_2x_2 + \dots + b_kx_k$  by  $x_2, x_3, \dots, x_k$  respectively and taking sum we get following k equations; known as normal equations.

$$\left. \begin{aligned} \sum x_1x_2 &= b_2 \sum x_2^2 + b_3 \sum x_2x_3 + b_4 \sum x_2x_4 + \dots + b_k \sum x_2x_k \\ \sum x_1x_3 &= b_2 \sum x_2x_3 + b_3 \sum x_3^2 + b_4 \sum x_3x_4 + \dots + b_k \sum x_3x_k \\ \sum x_1x_4 &= b_2 \sum x_2x_4 + b_3 \sum x_3x_4 + b_4 \sum x_4^2 + \dots + b_k \sum x_4x_k \\ &\dots\dots\dots \\ \sum x_1x_k &= b_2 \sum x_2x_k + b_3 \sum x_3x_k + b_4 \sum x_4x_k + \dots + b_k \sum x_k^2 \end{aligned} \right\}$$

Now these  $b_2, b_3, \dots, b_k$  are the estimates of the  $\beta_2, \beta_3, \dots, \beta_k$ .

Solving the above k-1 equations, k-1 regression coefficients can be obtained.

**Multiple linear regression equation taking three variables:**

In multiple linear regression analysis with three variables at a time, one is dependent variable (say  $X_1$ ) and two independent variables (say  $X_2$  and  $X_3$ ). From the above normal equations for particular case of three variables, the normal equations would be

$$\left. \begin{aligned} \sum x_1x_2 &= b_2 \sum x_2^2 + b_3 \sum x_3x_2 \\ \sum x_1x_3 &= b_2 \sum x_2x_3 + b_3 \sum x_3^2 \end{aligned} \right\}$$

Solving the above two equations in (8) we shall get

$$\frac{\sum x_3^2 \sum x_1x_2 - \sum x_2x_3 \sum x_1x_3}{\sum x_2^2 \sum x_3^2 - \sum x_2x_3} = b_2 \text{ and}$$

$$\frac{\sum x_2^2 \sum x_1x_3 - \sum x_2x_3 \sum x_1x_2}{\sum x_2^2 \sum x_3^2 - \sum x_2x_3} = b_3$$

**Example:** Following table gives data pertaining to 20 units for three variables  $X_1, X_2$  and  $X_3$ . Find out the linear regression equation of  $X_1$  on  $X_2$  and  $X_3$ .

Observation	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
$X_1$	1.83	1.56	1.85	1.9	1.7	1.8	1.85	1.73	1.95	1.67	1.82	1.84	1.74	1.68	1.62	1.82	1.91	1.61	1.64	1.85
$X_2$	13	10	12	14	12	13	12	10	14	13	16	14	11	12	11	15	15	12	13	15
$X_3$	12	10	11	11	12	11	11	10	11	12	14	14	9	8	9	13	13	9	10	13

**Solution:**

Observation	$X_1$	$X_2$	$X_3$	$x_1 = X_{1i} - \bar{X}_1$	$x_2 = X_{2i} - \bar{X}_2$	$x_3 = X_{3i} - \bar{X}_3$	$x_1^2$	$x_2^2$	$x_3^2$	$x_1x_2$	$x_1x_3$	$x_2x_3$
1	1.83	13.00	12.00	0.061	0.150	0.850	0.004	0.023	0.722	0.009	0.052	0.128
2	1.56	10.00	10.00	-0.209	-2.850	-1.150	0.044	8.123	1.323	0.596	0.240	3.278
3	1.85	12.00	11.00	0.081	-0.850	-0.150	0.007	0.722	0.023	-0.069	-0.012	0.128
4	1.90	14.00	11.00	0.131	1.150	-0.150	0.017	1.323	0.023	0.151	-0.020	-0.173

5	1.70	12.00	12.00	-0.069	-0.850	0.850	0.005	0.722	0.722	0.059	-0.059	-0.722
6	1.80	13.00	11.00	0.031	0.150	-0.150	0.001	0.023	0.023	0.005	-0.005	-0.023
7	1.85	12.00	11.00	0.081	-0.850	-0.150	0.007	0.722	0.023	-0.069	-0.012	0.128
8	1.73	10.00	10.00	-0.039	-2.850	-1.150	0.002	8.123	1.323	0.111	0.045	3.278
9	1.95	14.00	11.00	0.181	1.150	-0.150	0.033	1.323	0.023	0.208	-0.027	-0.173
10	1.67	13.00	12.00	-0.099	0.150	0.850	0.010	0.023	0.722	-0.015	-0.084	0.128
11	1.82	16.00	14.00	0.051	3.150	2.850	0.003	9.923	8.123	0.161	0.145	8.978
12	1.84	14.00	14.00	0.071	1.150	2.850	0.005	1.323	8.123	0.082	0.202	3.278
13	1.74	11.00	9.00	-0.029	-1.850	-2.150	0.001	3.423	4.623	0.054	0.062	3.978
14	1.68	12.00	8.00	-0.089	-0.850	-3.150	0.008	0.722	9.923	0.076	0.280	2.678
15	1.62	11.00	9.00	-0.149	-1.850	-2.150	0.022	3.423	4.623	0.276	0.320	3.978
16	1.82	15.00	13.00	0.051	2.150	1.850	0.003	4.623	3.423	0.110	0.094	3.978
17	1.91	15.00	13.00	0.141	2.150	1.850	0.020	4.623	3.423	0.303	0.261	3.978
18	1.61	12.00	9.00	-0.159	-0.850	-2.150	0.025	0.722	4.623	0.135	0.342	1.828
19	1.64	13.00	10.00	-0.129	0.150	-1.150	0.017	0.023	1.323	-0.019	0.148	-0.173
20	1.85	15.00	13.00	0.081	2.150	1.850	0.007	4.623	3.423	0.174	0.150	3.978
<b>Total</b>	<b>35.37</b>	<b>257</b>	<b>223</b>				<b>0.237</b>	<b>54.550</b>	<b>56.550</b>	<b>2.336</b>	<b>2.125</b>	<b>42.450</b>
<b>Average</b>	<b>1.769</b>	<b>12.850</b>	<b>11.150</b>									

From above table we have,

$$\sum x_1^2 = 0.237; \sum x_2^2 = 54.550; \sum x_3^2 = 56.550; \sum x_1x_2 = 2.336; \sum x_1x_3 = 2.125;$$

$$\sum x_2x_3 = 42.450; (\sum x_2x_3)^2 = 1802.003$$

We Know that,

$$b_1 = \frac{\sum x_3^2 \sum x_2x_1 - \sum x_2x_3 \sum x_3x_1}{\sum x_2^2 \sum x_3^2 - (\sum x_2x_3)^2} = \frac{56.550 \times 2.336 - 42.450 \times 2.125}{54.550 \times 56.550 - 1802.003}$$

$$= \frac{132.100 - 90.206}{3084.803 - 1802.003} = \frac{41.894}{1282.799} = 0.032$$

$$b_2 = \frac{\sum x_2^2 \sum x_3x_1 - \sum x_2x_3 \sum x_2x_1}{\sum x_2^2 \sum x_3^2 - (\sum x_2x_3)^2} = \frac{54.550 \times 2.125 - 42.450 \times 2.336}{54.550 \times 56.550 - 1802.003}$$

$$= \frac{115.918 - 99.163}{3084.803 - 1802.003} = \frac{16.755}{1282.799} = 0.013$$

Hence the linear regression of  $x_1$  on  $x_2$  and  $x_3$  will be

$x_1 = 0.032x_2 + 0.013x_3$  transforming back to original variables we have

$(X_1 - \bar{X}_1) = (X_2 - \bar{X}_2)0.032 + (X_3 - \bar{X}_3)0.013$ , putting value of  $\bar{X}_1$ ,  $\bar{X}_2$  and  $\bar{X}_3$  we have

$$(X_1 - 1.769) = (X_2 - 12.850)0.032 + (X_3 - 11.150)0.013$$

$$X_1 = 1.769 + 0.032X_2 - 0.411 + 0.013X_3 - 0.144$$

$$X_1 = 1.214 + 0.032X_2 + 0.013X_3$$

[Regression Exercise.xls](#)

Differences are to be noted

### Multiple Correlation:

Under multiple variables situation the correlation coefficient between the dependent variable and the joint effect of the independent variables is known as the multiple correlation coefficient.

Let we have k variables  $X_1, X_2, \dots, X_k$  of which  $X_1$  is the dependent variable and others are independent variables. The joint effects of the independent variables  $X_2, X_3, X_4, \dots, X_k$  on  $X_1$  is the estimated value of the dependent variable  $X_1$  i.e.  $\hat{X}_1$  from the regression equation

$$R_{X_1.X_2.X_3.X_4\dots X_k} = \frac{\text{Cov}(X_1, \hat{X}_1)}{\sqrt{V(X_1)}\sqrt{V(\hat{X}_1)}}$$

We know that  $\text{Cov}(X_1, \hat{X}_1) = \text{Cov}(\hat{X}_1 + u, \hat{X}_1)$  [ $\because X_1 = \hat{X}_1 + u$  and u is the error]

$$= V(\hat{X}_1) + \text{Cov}(\hat{X}_1, e) \quad (\because \text{Cov}(\hat{X}_1, e), \text{ by assumption})$$

$$= V(\hat{X}_1)$$

$$\therefore R_{X_1.X_2.X_3.X_4\dots X_k} = \frac{\text{Cov}(X_1, \hat{X}_1)}{\sqrt{V(X_1)}\sqrt{V(\hat{X}_1)}} = \frac{V(\hat{X}_1)}{\sqrt{V(X_1)}\sqrt{V(\hat{X}_1)}} = \sqrt{\frac{V(\hat{X}_1)}{V(X_1)}}$$

Squaring both the sides we get  $R^2_{X_1.X_2.X_3.X_4\dots X_k} = \frac{V(\hat{X}_1)}{V(X_1)}$ , which is known as the **coefficient of**

**determination. Thus, coefficient determination is a measure of the proportion of variance of the dependent variable explained by the joint effects of the independent variables**

So,

$$\text{TSS} = \sum_{i=1}^n (X_{1i} - \bar{X}_1)^2$$

$$= \sum_{i=1}^n (X_{1i} - \hat{X}_{1i} + \hat{X}_{1i} - \bar{X}_1)^2$$

$$= \sum_{i=1}^n ((X_{1i} - \hat{X}_{1i}) + (\hat{X}_{1i} - \bar{X}_1))^2$$

$$= \sum_{i=1}^n (X_{1i} - \hat{X}_{1i})^2 + \sum_{i=1}^n (\hat{X}_{1i} - \bar{X}_1)^2 + 2 \sum_{i=1}^n (X_{1i} - \hat{X}_{1i})(\hat{X}_{1i} - \bar{X}_1)$$

$$= \text{RSS} + \text{RgSS} + 2 \sum_{i=1}^n u_i (\hat{X}_{1i} - \bar{X}_1) \quad \text{where, RgSS and RSS are sum of squares due to regression and residual respectively}$$

$$= \text{RSS} + \text{RgSS} + 2 \sum_{i=1}^n u_i \hat{X}_{1i} - 2\bar{X}_1 \sum_{i=1}^n u_i$$

$$= \text{RSS} + \text{RgSS} + 0 + 0 \quad (\text{ by assumptions})$$

$$= \text{RSS} + \text{RgSS}$$

$$\therefore \text{TSS} = \text{RgSS} + \text{RSS}$$

$$\text{or } \frac{\text{TSS}}{\text{TSS}} = \frac{\text{RgSS}}{\text{TSS}} + \frac{\text{RSS}}{\text{TSS}}$$

$$\text{or, } \frac{\sum_{i=1}^n (X_{1i} - \bar{X}_1)^2}{\sum_{i=1}^n (X_{1i} - \bar{X}_1)^2} = \frac{\sum_{i=1}^n (\hat{X}_{1i} - \bar{X}_1)^2}{\sum_{i=1}^n (X_{1i} - \bar{X}_1)^2} + \frac{\sum_{i=1}^n (X_{1i} - \hat{X}_{1i})^2}{\sum_{i=1}^n (X_{1i} - \bar{X}_1)^2}$$

$$\text{or, } 1 = \frac{V(\hat{X}_1)}{V(X_1)} + \frac{V(u)}{V(X_1)}$$

$$= R^2 + \frac{V(u)}{V(X_1)}$$

$$\therefore R^2 = 1 - \frac{V(u)}{V(X_1)} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

The RSS being the sum of squares, it can take only the positive value including zero i.e.  $\text{RSS} \geq 0$ . Moreover as  $\text{TSS} = \text{RgSS} + \text{RSS}$ , so RSS can have maximum value equal to TSS. Thus,  $\text{TSS} \geq \text{RSS} \geq 0$ .

$\therefore$  When  $\text{RSS} = 0$  then  $R^2 = 1$ ; again when  $\text{RSS} = \text{TSS}$  then  $R^2 = 0$

$$\therefore 1 \geq R^2 \geq 0$$

When  $R^2 = 1$ , then it implies perfect fittings, that total variation of the dependent variable has been explained by its linear relationship with the independent variables. Again, when  $R^2 = 0$ , then it implies no fittings, that means zero per cent of the total variation of the dependent variable has been explained by its linear relationship with the independent variables.

### Interpretation of $R^2$

Suppose, after framing a regression equation of  $X_1$  on  $X_2, X_3, X_4, \dots, X_k$ , we have calculated  $R^2 = 0.9$ ; this means 90% of the variations in the dependent variable has been explained by its linear relationship with the independent variables, leaving 10% unexplained.

**Thus the value of  $R^2$  measures the explaining power of the linear regression equation.**

### Game of maximisation of $R^2$

An experimenter is always in search of a relationship which can explain the dependent variable to the greatest possible extent. As such the experimenter tries to include more and more number of variables in the linear regression equation with an aim to get as much  $R^2$  as possible, so that the relationship could explain more and more variation in the dependent variable. Sometimes, with the aim of maximising  $R^2$  the experimenter includes such variables which might not have significant contribution towards the objective of the study. Thus, the process of maximising  $R^2$  by including more and more number of variables in the regression model is known as "Game of maximisation of  $R^2$ ". As we have already pointed out that the number of variables and the variables to be included in the regression equation is not guided by the statistical theory but by the subject on hand and relevance of the variables under given conditions; it does not matter how much is the value of the  $R^2$ , in the process!

**Adjusted  $R^2$**  is such a measure developed, which is **not** a non-decreasing function of number of variables in the regression equation, like  $R^2$ . Adjusted  $R^2$  is defined as:

$$\begin{aligned} \bar{R}^2_{X_1, X_2, X_3, X_4, \dots, X_k} &= 1 - \frac{RMS}{TMS} \\ &= 1 - \frac{TSS - RgSS / (n - k)}{TSS / (n - 1)} = 1 - \frac{(n - 1)}{(n - k)} + \frac{(n - 1)}{(n - k)} \frac{RgSS}{TSS} \\ &= 1 - \frac{(n - 1)}{(n - k)} (1 - R^2) \end{aligned}$$

Thus adjusted  $R^2$  is taking in to consideration the associated degrees of freedom. In any regression model we have  $K \geq 2$  thereby indicating that  $\bar{R}^2 < R^2$  that means as the number of independent variables increases  $\bar{R}^2$  increases lesser than  $R^2$

Again when  $R^2 = 1$ ,  $\bar{R}^2 = 1 - \frac{n - 1}{n - k} (1 - R^2) = 1$

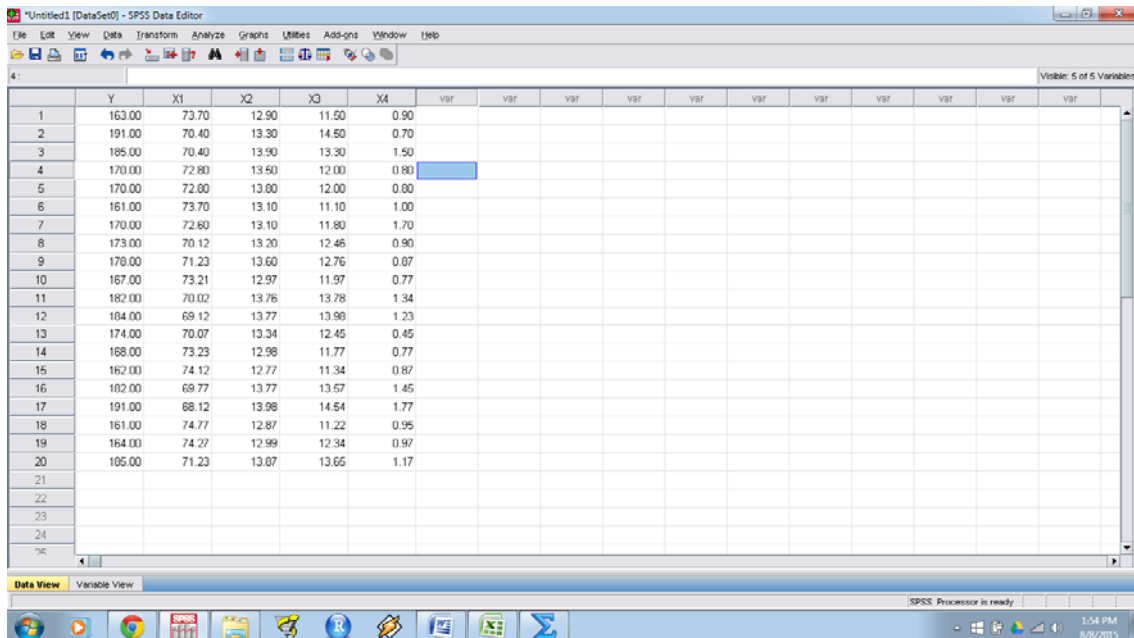
When  $R^2 = 0$  then  $\bar{R}^2 = 1 - \frac{n - 1}{n - k} (1 - 0) = \frac{n - k - n + 1}{n - k} = \frac{1 - k}{n - k}$

now,  $k \geq 2$  so  $\bar{R}^2$  is negative.

**Thus though  $R^2 \geq 0$ ,  $\bar{R}^2$  can be less than zero.**

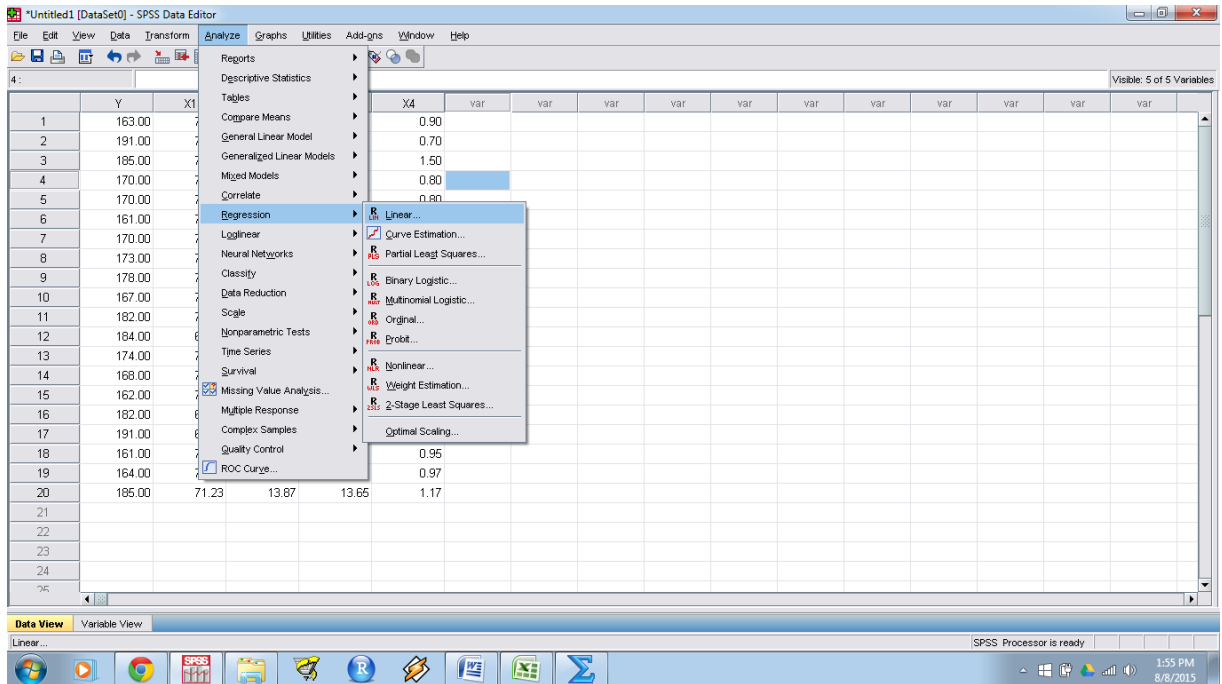
**Estimation of multiple linear regression equation using SPSS:**

**Step1:** Showing data structure to perform the regression analysis in data editor menu of SPSS.

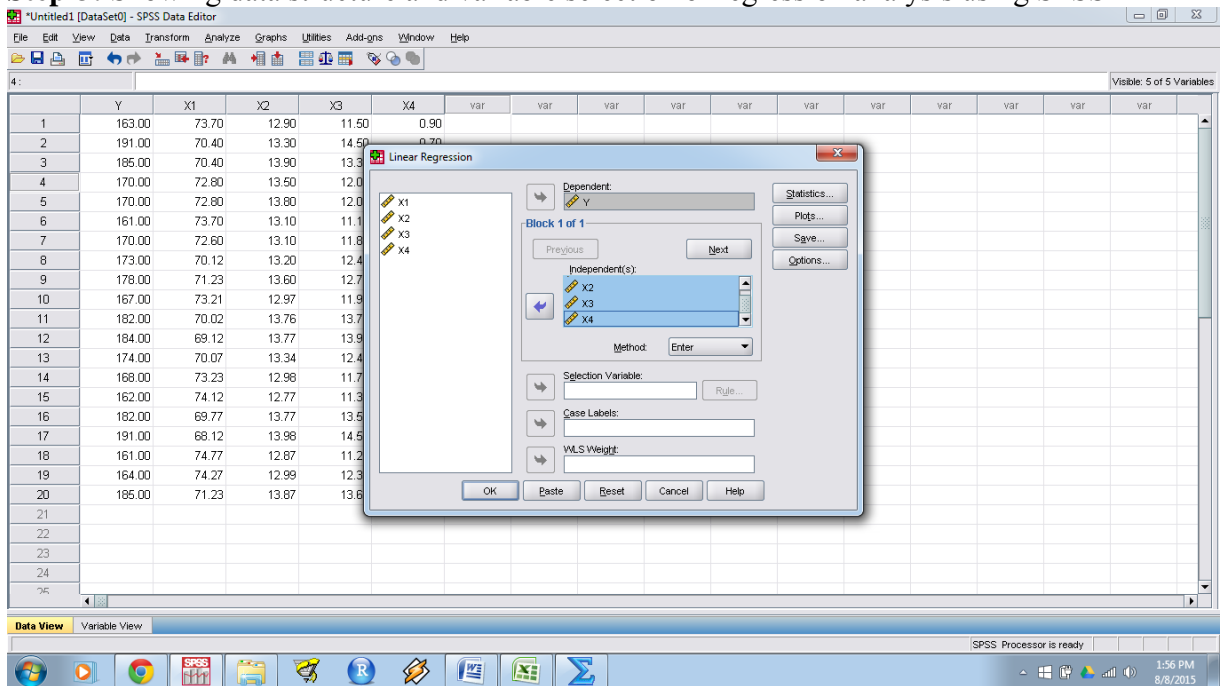


**Step2:** Showing data structure and selection of appropriate options in analysis menu of SPSS





**Step-3: Showing data structure and variable selection of regression analysis using SPSS**



**Step4:** Click on Ok button, out of Regression analysis will be displayed in output window of SPSS.

Variables Entered/Removed <sup>b</sup>			
Model	Variables Entered	Variables Removed	Method
1	X4, X1, X2, X3 <sup>a</sup>		Enter
a. All requested variables entered.			
b. Dependent Variable: Y			

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.977 <sup>a</sup>	.954	.942	2.40175
a. Predictors: (Constant), X4, X1, X2, X3				

ANOVA <sup>b</sup>						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1814.424	4	453.606	78.636	.000 <sup>a</sup>
	Residual	86.526	15	5.768		
	Total	1900.950	19			
a. Predictors: (Constant), X4, X1, X2, X3						
b. Dependent Variable: Y						

Coefficients <sup>a</sup>						
Model		Unstandardized		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	102.200	66.966		1.526	.148
	X1	-.815	.606	-.159	-1.343	.199
	X2	3.533	2.299	.143	1.537	.145
	X3	6.596	1.056	.721	6.244	.000
	X4	-.041	1.803	-.001	-.023	.982
a. Dependent Variable: Y						

### Test for regression parameters:

Suppose we like to test the hypothesis  $H_{01} : \beta_1 = \beta_2 = \dots = \beta_k = 0$ . This is equivalent to testing  $H_{01} : \rho_{y.12\dots k} = 0$  where  $\rho_{y.12\dots k}$  is the population multiple correlation coefficient of y with  $x_1, x_2, \dots, x_k$ .

Then the test statistic under  $H_0$  is given by

$$F = \frac{\sum_{i=1}^k b_i S_{y_i}}{S_{yy} - \sum_{i=1}^k b_i S_{y_i}} \cdot \frac{n-k-1}{k}$$

$$= \frac{R_{y.12\dots k}^2}{1 - R_{y.12\dots k}^2} \cdot \frac{n-k-1}{k}$$

In case F is significant we reject  $H_0$  and infer that there is usefulness of variables  $x_1, x_2, \dots, x_k$  in prediction of y.

**General Linear Regression Model:**

General linear regression model is closely linked with polynomial regression. Consider a bivariate normal distribution (y,  $x_2$ ). The conditional expectation of y given  $x_2$  is the linear regression of y on  $x_2$ . The true linear regression is of the form  $\eta_{y.2} = \beta_1 x_1 + \beta_2 x_2$  where  $x_1$  is dummy variable having value 1 and the model is  $y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$ .

Sometimes it may happen that the regression of y on  $x_2$  is far from linear. A non-linear regression of the form  $\eta_{y.2} = \beta_1 x_1 + \beta_2 x_2 + \beta_1 x_2^2 + \dots + \beta_k x_2^{k-1}$ .....(1) will fit the data better. The regression equation (1) known as polynomial regression equation. The model is, therefore,  $y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_1 x_{2i}^2 + \dots + \beta_k x_{2i}^{k-1} + \varepsilon_i$ . In this model  $y_i$  is a linear function of  $\beta$ 's. Hence the model is also a general linear regression model. If we redesignate  $x_2^2$  by  $x_3, x_2^3$  by  $x_4, \dots, x_2^{k-1}$  by  $x_k$  then we precisely get the model  $y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_1 x_{3i} + \dots + \beta_k x_{ki} + \varepsilon_i$ . Thus polynomial regression is a special case of general linear regression model.

For n number of observations the regression equation can be written as:

- 1  $Y_1 = \beta_1 X_{11} + \beta_2 X_{21} + \beta_3 X_{31} + \beta_4 X_{41} + \dots + \beta_k X_{k1} + u_1$
- 2  $Y_2 = \beta_1 X_{12} + \beta_2 X_{22} + \beta_3 X_{32} + \beta_4 X_{42} + \dots + \beta_k X_{k2} + u_2$
- 3  $Y_3 = \beta_1 X_{13} + \beta_2 X_{23} + \beta_3 X_{33} + \beta_4 X_{43} + \dots + \beta_k X_{k3} + u_3$
- ⋮
- ⋮
- ⋮
- n  $Y_n = \beta_1 X_{1n} + \beta_2 X_{2n} + \beta_3 X_{3n} + \beta_4 X_{4n} + \dots + \beta_k X_{kn} + u_n$

$$\Rightarrow \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}_{n \times 1} = \begin{pmatrix} X_{11} & X_{21} & X_{31} & X_{41} & \dots & X_{k1} \\ X_{12} & X_{22} & X_{32} & X_{42} & \dots & X_{k2} \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots \\ X_{1n} & X_{2n} & X_{3n} & X_{4n} & \dots & X_{kn} \end{pmatrix}_{n \times k} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}_{k \times 1} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}_{n \times 1}$$

In matrix notation the above equations can be written as

$$\underline{Y} = \underline{X}\underline{\beta} + \underline{u} \quad \text{with } E(\underline{Y}) = \underline{X}\underline{\beta}, E(\underline{u}) = 0 \text{ and } E(\underline{u}\underline{u}') = \sigma^2 \underline{I}, \text{ where } \underline{I} \text{ is an } n \times n \text{ unit matrix.}$$

Our objective will be to minimize  $L = \underline{u}'\underline{u}$

$$\begin{aligned} \text{Let } L = \underline{u}'\underline{u} &= (\underline{Y} - \underline{X}\underline{\beta})'(\underline{Y} - \underline{X}\underline{\beta}) \\ &= \underline{Y}'\underline{Y} - 2\underline{\beta}'\underline{X}'\underline{Y} + \underline{\beta}'\underline{X}'\underline{X}\underline{\beta} \end{aligned}$$

Let  $\underline{b}$  be the least square estimators of  $\underline{\beta}$ . Setting

$$\frac{\partial L}{\partial \underline{\beta}} = 0 \text{ and writing } \underline{b} \text{ for } \underline{\beta} \text{ we have}$$

$$2\underline{X}'\underline{Y} - 2\underline{X}'\underline{X}\underline{b} = 0$$

$$\text{or, } \underline{X}'\underline{X}\underline{b} = \underline{X}'\underline{Y}$$

$$\text{or, } \underline{b} = (\underline{X}'\underline{X})^{-1}\underline{X}'\underline{Y}, \text{ provided } (\underline{X}'\underline{X})^{-1} \text{ exists,}$$

$$\text{where } (\underline{X}'\underline{X}) = \underline{S} = \begin{pmatrix} \sum X_{1i}^2 & \sum X_{1i}X_{2i} & \dots & \sum X_{1i}X_{ki} \\ \sum X_{1i}X_{2i} & \sum X_{2i}^2 & \dots & \sum X_{2i}X_{ki} \\ \dots & \dots & \dots & \dots \\ \sum X_{1i}X_{ki} & \sum X_{2i}X_{ki} & \dots & \sum X_{ki}^2 \end{pmatrix}$$

$$\underline{X}'\underline{Y} = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1n} \\ X_{21} & X_{22} & \dots & X_{2n} \\ \dots & \dots & \dots & \dots \\ X_{k1} & X_{k2} & \dots & X_{kn} \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ \cdot \\ Y_n \end{pmatrix} = \begin{pmatrix} \sum X_{1i}Y_i \\ \sum X_{2i}Y_i \\ \cdot \\ \sum X_{ki}Y_i \end{pmatrix}$$

$$\underline{b} = \begin{pmatrix} b_1 \\ b_2 \\ \cdot \\ b_k \end{pmatrix}$$

We demonstrate below both the stepwise forward and backward regression  
**Stepwise forward regression**

**Table 1: Models summary**

Model		Coefficients		R <sup>2</sup>	t	Sig.
		B	Std. Error			
1	(Constant)	145.339	6.182	.846	23.512	.000
	(X <sub>1</sub> )	4.540	.457			
2	(Constant)	134.110	9.922	.862	13.517	.000
	(X <sub>1</sub> )	4.077	.551			

	X <sub>2</sub>	1.253	.880		1.423	.173
3	(Constant)	132.549	10.475	.865	12.654	.000
	( X <sub>1</sub> )	4.232	.623		6.795	.000
	X <sub>2</sub>	2.418	2.205		1.097	.289
	X <sub>3</sub>	-1.079	1.863		-.579	.571

a Dependent Variable: (Y)

**Table 2 : ANOVA tables for different models**

Model		Sum of Squares	df	Mean Square	F	Sig.
	Regression	1990.168	1	1990.168	98.745	.000
	Residual	362.782	18	20.155		
	Total	2352.950	19			
	Regression	2028.774	2	1014.387	53.195	.000
	Residual	324.176	17	19.069		
	Total	2352.950	19			
	Regression	2035.423	3	678.474	34.188	.000
	Residual	317.527	16	19.845		
	Total	2352.950	19			

1 Predictors: (Constant), ( X<sub>1</sub>)2 Predictors: (Constant), ( X<sub>1</sub>), X<sub>2</sub>3 Predictors: (Constant), ( X<sub>1</sub>), X<sub>2</sub>, X<sub>3</sub>

Dependent Variable: (Y)

From the correlation table it is clear that the dependent variable yield (Y) has highest correlation with X<sub>1</sub> followed by X<sub>2</sub> and X<sub>3</sub>. So we first include X<sub>1</sub> in the model and find that it explains 84.6% variation in yield through this relationship. In subsequent steps we introduce the variables X<sub>2</sub> and X<sub>3</sub> respectively and find that though the overall relationships are significant but neither the individual coefficients are significant at P = 0.05 (the desired level of significance) nor they have increased the explaining power of the model to a great extent. So these variables are redundant under the given context.

**Stepwise backward regression**

Model		Coefficients		R <sup>2</sup>	t	Sig.
		B	Std. Error			
3	(Constant)	132.549	10.475	.865	12.654	.000
	( X <sub>1</sub> )	4.232	.623		6.795	.000
	X <sub>2</sub>	2.418	2.205		1.097	.289
	X <sub>3</sub>	-1.079	1.863		-.579	.571
	(Constant)	134.110	9.922	.862	13.517	.000
	( X <sub>1</sub> )	4.077	.551		7.401	.000
	X <sub>2</sub>	1.253	.880		1.423	.173
	(Constant)	145.339	6.182	.846	23.512	.000
	( X <sub>1</sub> )	4.540	.457		9.937	.000

a Dependent Variable: (Y)

## ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
	Regression	2035.423	3	678.474	34.188	.000
	Residual	317.527	16	19.845		
	Total	2352.950	19			
	Regression	2028.774	2	1014.387	53.195	.000
	Residual	324.176	17	19.069		
	Total	2352.950	19			
	Regression	1990.168	1	1990.168	98.745	.000
	Residual	362.782	18	20.155		
	Total	2352.950	19			

1 Predictors: (Constant), X<sub>3</sub>, ( X<sub>1</sub>), X<sub>2</sub>

2 Predictors: (Constant), ( X<sub>1</sub>), X<sub>2</sub>

3 Predictors: (Constant), ( X<sub>1</sub>)

Dependent Variable: (Y)

In the stepwise backward method of regression we first introduce all the variables in the model and the model summary along with the analysis of variance is given in table above respectively. It is found in the first step that except the variable X<sub>1</sub>, other two variables X<sub>2</sub> and X<sub>3</sub> have the non-significant coefficients (at pre-assigned significance level of 0.05) and the most non-significant coefficient being found in case of variable X<sub>3</sub>. Hence in the next step we drop the variable X<sub>3</sub> and again fit the regression model with the variable X<sub>1</sub> and X<sub>2</sub>. In this case we find not much of reduction in explaining power of the model (only 0.3%) compared to the previous model, having all the variables, moreover the coefficient of X<sub>2</sub> still remains non-significant. So in next step we drop the variable X<sub>2</sub> and find that the model with only X<sub>1</sub>, as explanatory variable, is sufficient to explain as much as 84.6% variations in the dependent variable yield (Y) coupled with significant coefficient for the variable X<sub>1</sub>. Thus we conclude that, under the given context, the variable, number of hill per square meter (X<sub>1</sub>) is useful in explaining the variations in yield(Y) of paddy and the other variables considered are redundant.

## ANOTHER EXAMPLE

VARIABLES FOR REGRESSION (RESPONSE VAR. IS LAST)

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

DET. of correlation matrix of Predictors = 0.18131702E-02

MULTIPLE R-SQ= 0.6442 MULTIPLE R = 0.8026

F-VALUE FOR R = 4.00 WITH 19AND 42 DFS

R-BAR SQ(ADJUSTED R-SQ)= 0.4832

CHAR	BETA	BETAxR	STRU-R	REG COEF-B	SE OF B	T-VAL OF B	VIF
1	-0.018	-1.062	0.479	-0.015	0.128	0.118	2.672
2	-0.055	0.896	-0.130	-0.109	0.245	0.444	1.826
3	-0.040	1.905	-0.383	-0.025	0.085	0.300	2.101
4	0.181	4.616	0.205	0.893	0.616	1.451	1.829
5	-0.263	11.498	-0.351	-0.903	0.470	1.920	2.213
6	0.592	62.612	0.849	0.391	0.099	3.969	2.625
7	0.028	0.972	0.278	0.081	0.334	0.243	1.567
8	-0.055	1.347	-0.197	-0.217	0.456	0.475	1.579
9	0.097	1.972	0.163	0.478	0.585	0.816	1.681
10	0.017	-0.169	-0.082	0.017	0.139	0.123	2.156
11	-0.111	5.013	-0.363	-0.557	0.664	0.839	2.063
12	-0.208	5.611	-0.216	-0.945	0.476	1.985	1.298
13	0.079	0.409	0.041	0.122	0.231	0.530	2.630
14	0.062	0.562	0.072	0.065	0.116	0.561	1.455
15	-0.104	-1.035	0.080	-0.952	1.059	0.899	1.585
16	0.019	0.046	0.019	0.000	0.000	0.109	3.661
17	-0.104	2.793	-0.216	0.000	0.000	0.831	1.847
18	0.105	2.810	0.215	0.000	0.000	0.851	1.789
19	-0.069	-0.798	0.093	0.000	0.000	0.372	3.995

INTERCEPT CONSTANT= 9.49 S.E. OF INTERCEPT= 7.50  
 ESTIMATE OF SIGMA S= 3.71  
 INDEX OF MULTICOLLINEARITY RL= 0.21354600E+01

step-down equation-- one predictor dropped

VARIABLES FOR REGRESSION (RESPONSE VAR. IS LAST)

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 17 18 19 20

DET. of correlation matrix of Predictors = 0.66385912E-02

MULTIPLE R-SQ= 0.6441 MULTIPLE R = 0.8025  
 F-VALUE FOR R = 4.32 WITH 18AND 43 DFS

R-BAR SQ(ADJUSTED R-SQ)= 0.4951

CHAR	BETA	BETAxR	STRU-R	REG COEF-B	SE OF B	T-VAL OF B	VIF
1	-0.023	-1.380	0.479	-0.020	0.120	0.164	2.387
2	-0.054	0.879	-0.130	-0.107	0.242	0.442	1.815
3	-0.041	1.950	-0.383	-0.026	0.084	0.311	2.092
4	0.175	4.477	0.205	0.867	0.557	1.555	1.532
5	-0.266	11.647	-0.351	-0.914	0.453	2.019	2.101
6	0.594	62.869	0.849	0.393	0.096	4.074	2.571
7	0.025	0.855	0.278	0.072	0.318	0.225	1.454
8	-0.056	1.383	-0.197	-0.223	0.448	0.497	1.558
9	0.092	1.860	0.163	0.451	0.524	0.860	1.378
10	0.019	-0.191	-0.082	0.019	0.135	0.143	2.106
11	-0.108	4.898	-0.363	-0.544	0.645	0.843	1.997
12	-0.210	5.676	-0.216	-0.955	0.460	2.076	1.241
13	0.072	0.370	0.041	0.111	0.203	0.547	2.079
14	0.062	0.561	0.072	0.065	0.114	0.566	1.455
15	-0.101	-1.008	0.080	-0.928	1.023	0.907	1.511
17	-0.108	2.908	-0.216	0.000	0.000	0.921	1.666
18	0.109	2.924	0.215	0.000	0.000	0.944	1.610

19            -0.058        -0.679        0.093        0.000        0.000        0.373        2.951

INTERCEPT CONSTANT=            9.89    S.E. OF INTERCEPT=            6.50  
 ESTIMATE OF SIGMA S=            3.67  
 INDEX OF MULTICOLLINEARITY RL= 0.18613850E+01

step-down equation-- one predictor dropped

VARIABLES FOR REGRESSION (RESPONSE VAR. IS LAST)  
 1 2 3 4 5 6 7 8 9 11 12 13 14 15 17 18 19 20

DET. of correlation matrix of Predictors = 0.13983872E-01

MULTIPLE R-SQ= 0.6439    MULTIPLE R = 0.8024  
 F-VALUE FOR R =            4.68    WITH 17AND 44 DFS

R-BAR SQ(ADJUSTED R-SQ)= 0.5063

CHAR	BETA	BETAxR	STRU-R	REG COEF-B	SE OF B	T-VAL OF B	VIF
1	-0.021	-1.281	0.479	-0.018	0.118	0.155	2.370
2	-0.050	0.810	-0.130	-0.098	0.232	0.424	1.705
3	-0.040	1.921	-0.383	-0.026	0.083	0.310	2.090
4	0.174	4.456	0.205	0.862	0.550	1.567	1.528
5	-0.258	11.302	-0.351	-0.887	0.406	2.185	1.726
6	0.593	62.699	0.849	0.392	0.095	4.123	2.552
7	0.027	0.952	0.278	0.080	0.309	0.257	1.409
8	-0.052	1.275	-0.197	-0.205	0.426	0.482	1.442
9	0.096	1.950	0.163	0.472	0.496	0.952	1.264
11	-0.110	4.987	-0.363	-0.554	0.635	0.873	1.975
12	-0.210	5.655	-0.216	-0.951	0.454	2.095	1.237
13	0.076	0.391	0.041	0.117	0.196	0.597	1.986
14	0.065	0.583	0.072	0.067	0.112	0.603	1.417
15	-0.096	-0.950	0.080	-0.874	0.942	0.929	1.311
17	-0.108	2.905	-0.216	0.000	0.000	0.931	1.666
18	0.108	2.898	0.215	0.000	0.000	0.948	1.605
19	-0.047	-0.553	0.093	0.000	0.000	0.352	2.256

INTERCEPT CONSTANT=            9.60    S.E. OF INTERCEPT=            6.11  
 ESTIMATE OF SIGMA S=            3.63  
 INDEX OF MULTICOLLINEARITY RL= 0.17374640E+01

step-down equation-- one predictor dropped

VARIABLES FOR REGRESSION (RESPONSE VAR. IS LAST)  
 2 3 4 5 6 7 8 9 11 12 13 14 15 17 18 19 20

DET. of correlation matrix of Predictors = 0.33145368E-01

MULTIPLE R-SQ= 0.6437    MULTIPLE R = 0.8023  
 F-VALUE FOR R =            5.08    WITH 16AND 45 DFS

R-BAR SQ(ADJUSTED R-SQ)= 0.5170

CHAR	BETA	BETAxR	STRU-R	REG COEF-B	SE OF B	T-VAL OF B	VIF
2	-0.051	0.823	-0.130	-0.100	0.229	0.436	1.701
3	-0.042	2.021	-0.383	-0.027	0.081	0.331	2.068
4	0.178	4.556	0.205	0.881	0.530	1.662	1.451



5	-0.258	11.286	-0.351	-0.886	0.401	2.206	1.725
6	0.578	61.151	0.849	0.382	0.070	5.440	1.424
7	0.027	0.925	0.278	0.077	0.305	0.253	1.406
8	-0.055	1.349	-0.197	-0.217	0.415	0.523	1.396
9	0.099	1.997	0.163	0.484	0.485	0.996	1.236
11	-0.113	5.117	-0.363	-0.568	0.621	0.914	1.934
12	-0.206	5.563	-0.216	-0.936	0.438	2.137	1.175
13	0.084	0.434	0.041	0.130	0.175	0.740	1.630
14	0.070	0.628	0.072	0.072	0.105	0.687	1.294
15	-0.094	-0.932	0.080	-0.857	0.925	0.927	1.292
17	-0.105	2.831	-0.216	0.000	0.000	0.928	1.625
18	0.103	2.775	0.215	0.000	0.000	0.950	1.495
19	-0.045	-0.525	0.093	0.000	0.000	0.339	2.225

INTERCEPT CONSTANT= 8.99 S.E. OF INTERCEPT= 4.61  
 ESTIMATE OF SIGMA S= 3.59  
 INDEX OF MULTICOLLINEARITY RL= 0.15673170E+01

step-down equation-- one predictor dropped

VARIABLES FOR REGRESSION (RESPONSE VAR. IS LAST)  
 2 3 4 5 6 8 9 11 12 13 14 15 17 18 19 20

DET. of correlation matrix of Predictors = 0.46585953E-01

MULTIPLE R-SQ= 0.6432 MULTIPLE R = 0.8020  
 F-VALUE FOR R = 5.53 WITH 15AND 46 DFS

R-BAR SQ(ADJUSTED R-SQ)= 0.5268

CHAR	BETA	BETAxR	STRU-R	REG COEF-B	SE OF B	T-VAL OF B	VIF
2	-0.046	0.750	-0.130	-0.091	0.224	0.406	1.660
3	-0.049	2.326	-0.383	-0.031	0.079	0.392	1.988
4	0.182	4.657	0.205	0.900	0.520	1.732	1.423
5	-0.260	11.372	-0.351	-0.892	0.397	2.248	1.719
6	0.582	61.680	0.850	0.385	0.068	5.620	1.384
8	-0.061	1.499	-0.197	-0.241	0.400	0.603	1.323
9	0.097	1.964	0.163	0.475	0.479	0.991	1.230
11	-0.108	4.882	-0.363	-0.541	0.606	0.894	1.879
12	-0.207	5.586	-0.217	-0.939	0.433	2.168	1.174
13	0.086	0.446	0.041	0.133	0.173	0.770	1.620
14	0.075	0.678	0.072	0.078	0.102	0.767	1.234
15	-0.094	-0.939	0.080	-0.863	0.915	0.943	1.291
17	-0.102	2.759	-0.216	0.000	0.000	0.917	1.610
18	0.107	2.881	0.215	0.000	0.000	1.006	1.466
19	-0.046	-0.541	0.094	0.000	0.000	0.353	2.222

INTERCEPT CONSTANT= 8.96 S.E. OF INTERCEPT= 4.57  
 ESTIMATE OF SIGMA S= 3.55  
 INDEX OF MULTICOLLINEARITY RL= 0.15481830E+01

step-down equation-- one predictor dropped

VARIABLES FOR REGRESSION (RESPONSE VAR. IS LAST)  
 2 3 4 5 6 8 9 11 12 13 14 15 17 18 20

DET. of correlation matrix of Predictors = 0.10349380E+00

MULTIPLE R-SQ= 0.6422 MULTIPLE R = 0.8014  
 F-VALUE FOR R = 6.03 WITH 14AND 47 DFS

R-BAR SQ(ADJUSTED R-SQ)= 0.5356

CHAR	BETA	BETAxR	STRU-R	REG COEF-B	SE OF B	T-VAL OF B	VIF
2	-0.045	0.741	-0.131	-0.090	0.222	0.404	1.660
3	-0.046	2.175	-0.383	-0.029	0.078	0.371	1.978
4	0.174	4.459	0.205	0.860	0.503	1.712	1.356
5	-0.242	10.626	-0.352	-0.832	0.355	2.341	1.406
6	0.581	61.588	0.850	0.384	0.068	5.663	1.381
8	-0.051	1.260	-0.197	-0.202	0.381	0.531	1.223
9	0.098	1.996	0.163	0.482	0.474	1.016	1.228
11	-0.123	5.591	-0.363	-0.619	0.559	1.107	1.631
12	-0.207	5.612	-0.217	-0.942	0.429	2.195	1.174
13	0.086	0.444	0.041	0.133	0.171	0.773	1.619
14	0.074	0.671	0.073	0.077	0.101	0.766	1.234
15	-0.095	-0.947	0.080	-0.869	0.906	0.959	1.291
17	-0.120	3.246	-0.216	0.000	0.000	1.221	1.278
18	0.094	2.538	0.216	0.000	0.000	0.951	1.294

INTERCEPT CONSTANT= 8.57 S.E. OF INTERCEPT= 4.39  
 ESTIMATE OF SIGMA S= 3.52  
 INDEX OF MULTICOLLINEARITY RL= 0.14107780E+01

step-down equation-- one predictor dropped

VARIABLES FOR REGRESSION (RESPONSE VAR. IS LAST)  
 2 4 5 6 8 9 11 12 13 14 15 17 18 20

DET. of correlation matrix of Predictors = 0.20467445E+00

MULTIPLE R-SQ= 0.6412 MULTIPLE R = 0.8007  
 F-VALUE FOR R = 6.60 WITH 13AND 48 DFS

R-BAR SQ(ADJUSTED R-SQ)= 0.5440

CHAR	BETA	BETAxR	STRU-R	REG COEF-B	SE OF B	T-VAL OF B	VIF
2	-0.066	1.085	-0.131	-0.131	0.190	0.691	1.238
4	0.185	4.745	0.206	0.914	0.477	1.917	1.243
5	-0.252	11.082	-0.352	-0.866	0.340	2.547	1.311
6	0.581	61.732	0.851	0.384	0.067	5.719	1.380
8	-0.050	1.239	-0.197	-0.199	0.377	0.527	1.222
9	0.102	2.071	0.163	0.499	0.468	1.067	1.216
11	-0.130	5.915	-0.364	-0.654	0.546	1.197	1.585
12	-0.208	5.624	-0.217	-0.942	0.425	2.217	1.173
13	0.094	0.485	0.042	0.145	0.167	0.866	1.561
14	0.076	0.689	0.073	0.079	0.100	0.792	1.231
15	-0.083	-0.833	0.080	-0.763	0.852	0.896	1.162
17	-0.126	3.405	-0.216	0.000	0.000	1.306	1.247
18	0.102	2.761	0.216	0.000	0.000	1.068	1.231

INTERCEPT CONSTANT= 8.10 S.E. OF INTERCEPT= 4.16  
 ESTIMATE OF SIGMA S= 3.48  
 INDEX OF MULTICOLLINEARITY RL= 0.12923860E+01

step-down equation-- one predictor dropped

VARIABLES FOR REGRESSION (RESPONSE VAR. IS LAST)  
2 4 5 6 9 11 12 13 14 15 17 18 20

DET. of correlation matrix of Predictors = 0.25008515E+00

MULTIPLE R-SQ= 0.6391 MULTIPLE R = 0.7994  
F-VALUE FOR R = 7.23 WITH 12AND 49 DFS

R-BAR SQ(ADJUSTED R-SQ)= 0.5507

CHAR	BETA	BETAxR	STRU-R	REG COEF-B	SE OF B	T-VAL OF B	VIF
2	-0.063	1.030	-0.131	-0.124	0.188	0.661	1.231
4	0.185	4.769	0.206	0.916	0.473	1.935	1.243
5	-0.241	10.623	-0.352	-0.828	0.330	2.511	1.250
6	0.596	63.511	0.852	0.394	0.064	6.148	1.275
9	0.103	2.094	0.163	0.503	0.464	1.084	1.216
11	-0.131	5.975	-0.364	-0.659	0.542	1.214	1.585
12	-0.205	5.564	-0.217	-0.929	0.421	2.206	1.169
13	0.086	0.449	0.042	0.133	0.164	0.812	1.536
14	0.068	0.615	0.073	0.070	0.098	0.720	1.197
15	-0.080	-0.804	0.080	-0.734	0.844	0.870	1.157
17	-0.119	3.222	-0.217	0.000	0.000	1.254	1.222
18	0.109	2.952	0.216	0.000	0.000	1.157	1.209

INTERCEPT CONSTANT= 7.29 S.E. OF INTERCEPT= 3.84

ESTIMATE OF SIGMA S= 3.46

INDEX OF MULTICOLLINEARITY RL= 0.12742680E+01

step-down equation-- one predictor dropped

VARIABLES FOR REGRESSION (RESPONSE VAR. IS LAST)  
4 5 6 9 11 12 13 14 15 17 18 20

DET. of correlation matrix of Predictors = 0.30796091E+00

MULTIPLE R-SQ= 0.6359 MULTIPLE R = 0.7974  
F-VALUE FOR R = 7.94 WITH 11AND 50 DFS

R-BAR SQ(ADJUSTED R-SQ)= 0.5558

CHAR	BETA	BETAxR	STRU-R	REG COEF-B	SE OF B	T-VAL OF B	VIF
4	0.180	4.653	0.206	0.889	0.469	1.896	1.234
5	-0.237	10.487	-0.353	-0.813	0.327	2.486	1.245
6	0.605	64.792	0.854	0.400	0.063	6.339	1.250
9	0.094	1.930	0.164	0.462	0.457	1.009	1.193
11	-0.134	6.142	-0.365	-0.673	0.539	1.250	1.582
12	-0.202	5.520	-0.218	-0.917	0.418	2.192	1.167
13	0.067	0.350	0.042	0.103	0.157	0.659	1.419
14	0.072	0.659	0.073	0.075	0.097	0.774	1.191
15	-0.078	-0.781	0.080	-0.710	0.838	0.846	1.155
17	-0.108	2.932	-0.217	0.000	0.000	1.160	1.183
18	0.122	3.316	0.217	0.000	0.000	1.329	1.158

INTERCEPT CONSTANT= 6.15 S.E. OF INTERCEPT= 3.41

ESTIMATE OF SIGMA S= 3.44

INDEX OF MULTICOLLINEARITY RL= 0.12523900E+01

step-down equation-- one predictor dropped

VARIABLES FOR REGRESSION (RESPONSE VAR. IS LAST)

4 5 6 9 11 12 14 15 17 18 20

DET. of correlation matrix of Predictors = 0.43704206E+00

MULTIPLE R-SQ= 0.6327 MULTIPLE R = 0.7954  
F-VALUE FOR R = 8.79 WITH 10AND 51 DFS

R-BAR SQ(ADJUSTED R-SQ)= 0.5607

CHAR	BETA	BETAxR	STRU-R	REG COEF-B	SE OF B	T-VAL OF B	VIF
4	0.173	4.495	0.207	0.855	0.463	1.844	1.219
5	-0.227	10.126	-0.354	-0.781	0.322	2.429	1.217
6	0.621	66.838	0.857	0.410	0.061	6.766	1.169
9	0.090	1.848	0.164	0.440	0.454	0.970	1.187
11	-0.103	4.738	-0.366	-0.517	0.481	1.075	1.274
12	-0.210	5.767	-0.218	-0.953	0.413	2.311	1.147
14	0.075	0.690	0.073	0.078	0.096	0.811	1.188
15	-0.079	-0.794	0.080	-0.718	0.834	0.861	1.155
17	-0.104	2.837	-0.218	0.000	0.000	1.126	1.177
18	0.127	3.455	0.217	0.000	0.000	1.390	1.151

INTERCEPT CONSTANT= 6.54 S.E. OF INTERCEPT= 3.34

ESTIMATE OF SIGMA S= 3.42

INDEX OF MULTICOLLINEARITY RL= 0.11885260E+01

step-down equation-- one predictor dropped

VARIABLES FOR REGRESSION (RESPONSE VAR. IS LAST)

4 5 6 9 11 12 15 17 18 20

DET. of correlation matrix of Predictors = 0.51934447E+00

MULTIPLE R-SQ= 0.6280 MULTIPLE R = 0.7925  
F-VALUE FOR R = 9.75 WITH 9AND 52 DFS

R-BAR SQ(ADJUSTED R-SQ)= 0.5636

CHAR	BETA	BETAxR	STRU-R	REG COEF-B	SE OF B	T-VAL OF B	VIF
4	0.158	4.149	0.208	0.783	0.453	1.727	1.174
5	-0.227	10.187	-0.356	-0.780	0.321	2.433	1.217
6	0.623	67.619	0.860	0.412	0.060	6.820	1.167
9	0.072	1.486	0.165	0.351	0.439	0.800	1.118
11	-0.095	4.408	-0.367	-0.477	0.477	1.001	1.261
12	-0.225	6.236	-0.219	-1.023	0.402	2.545	1.097
15	-0.071	-0.728	0.081	-0.654	0.827	0.790	1.144
17	-0.100	2.751	-0.218	0.000	0.000	1.089	1.174
18	0.141	3.892	0.218	0.000	0.000	1.592	1.104

INTERCEPT CONSTANT= 7.77 S.E. OF INTERCEPT= 2.96

ESTIMATE OF SIGMA S= 3.41

INDEX OF MULTICOLLINEARITY RL= 0.11620250E+01

step-down equation-- one predictor dropped

VARIABLES FOR REGRESSION (RESPONSE VAR. IS LAST)  
4 5 6 9 11 12 17 18 20

DET. of correlation matrix of Predictors = 0.59425858E+00

MULTIPLE R-SQ= 0.6235 MULTIPLE R = 0.7896  
F-VALUE FOR R = 10.97 WITH 8AND 53 DFS

R-BAR SQ(ADJUSTED R-SQ)= 0.5667

CHAR	BETA	BETAxR	STRU-R	REG COEF-B	SE OF B	T-VAL OF B	VIF
4	0.160	4.223	0.208	0.791	0.452	1.752	1.174
5	-0.206	9.309	-0.357	-0.708	0.306	2.312	1.118
6	0.615	67.244	0.863	0.407	0.060	6.798	1.153
9	0.069	1.437	0.165	0.337	0.437	0.771	1.116
11	-0.094	4.399	-0.369	-0.473	0.475	0.996	1.261
12	-0.233	6.502	-0.220	-1.059	0.398	2.661	1.083
17	-0.100	2.774	-0.219	0.000	0.000	1.094	1.174
18	0.148	4.112	0.219	0.000	0.000	1.683	1.093

INTERCEPT CONSTANT= 6.87 S.E. OF INTERCEPT= 2.73  
ESTIMATE OF SIGMA S= 3.40  
INDEX OF MULTICOLLINEARITY RL= 0.11466810E+01

step-down equation-- one predictor dropped

VARIABLES FOR REGRESSION (RESPONSE VAR. IS LAST)  
4 5 6 11 12 17 18 20

DET. of correlation matrix of Predictors = 0.66335306E+00

MULTIPLE R-SQ= 0.6193 MULTIPLE R = 0.7870  
F-VALUE FOR R = 12.55 WITH 7AND 54 DFS

R-BAR SQ(ADJUSTED R-SQ)= 0.5699

CHAR	BETA	BETAxR	STRU-R	REG COEF-B	SE OF B	T-VAL OF B	VIF
4	0.141	3.752	0.209	0.698	0.434	1.610	1.090
5	-0.202	9.187	-0.358	-0.693	0.304	2.278	1.114
6	0.622	68.404	0.866	0.411	0.059	6.924	1.144
11	-0.100	4.713	-0.370	-0.503	0.472	1.067	1.252
12	-0.235	6.584	-0.221	-1.065	0.397	2.687	1.083
17	-0.109	3.040	-0.220	0.000	0.000	1.205	1.156
18	0.155	4.320	0.220	0.000	0.000	1.771	1.084

INTERCEPT CONSTANT= 7.88 S.E. OF INTERCEPT= 2.38  
ESTIMATE OF SIGMA S= 3.38  
INDEX OF MULTICOLLINEARITY RL= 0.11317680E+01

step-down equation-- one predictor dropped

VARIABLES FOR REGRESSION (RESPONSE VAR. IS LAST)  
4 5 6 12 17 18 20

DET. of correlation matrix of Predictors = 0.83076311E+00

MULTIPLE R-SQ= 0.6113 MULTIPLE R = 0.7818  
 F-VALUE FOR R = 14.41 WITH 6AND 55 DFS

R-BAR SQ(ADJUSTED R-SQ)= 0.5689

CHAR	BETA	BETAxR	STRU-R	REG COEF-B	SE OF B	T-VAL OF B	VIF
4	0.162	4.372	0.211	0.803	0.423	1.899	1.034
5	-0.215	9.914	-0.360	-0.739	0.302	2.448	1.093
6	0.648	72.222	0.871	0.428	0.057	7.492	1.058
12	-0.249	7.065	-0.222	-1.128	0.393	2.874	1.059
17	-0.091	2.590	-0.221	0.000	0.000	1.028	1.119
18	0.136	3.836	0.221	0.000	0.000	1.584	1.038

INTERCEPT CONSTANT= 6.73 S.E. OF INTERCEPT= 2.12  
 ESTIMATE OF SIGMA S= 3.39  
 INDEX OF MULTICOLLINEARITY RL= 0.10666770E+01

step-down equation-- one predictor dropped

VARIABLES FOR REGRESSION (RESPONSE VAR. IS LAST)  
 4 5 6 12 18 20

DET. of correlation matrix of Predictors = 0.92930320E+00

MULTIPLE R-SQ= 0.6038 MULTIPLE R = 0.7770  
 F-VALUE FOR R = 17.07 WITH 5AND 56 DFS

R-BAR SQ(ADJUSTED R-SQ)= 0.5684

CHAR	BETA	BETAxR	STRU-R	REG COEF-B	SE OF B	T-VAL OF B	VIF
4	0.171	4.673	0.212	0.848	0.421	2.014	1.023
5	-0.192	8.951	-0.363	-0.659	0.292	2.258	1.020
6	0.663	74.818	0.877	0.438	0.056	7.774	1.028
12	-0.259	7.447	-0.224	-1.175	0.390	3.011	1.044
18	0.144	4.111	0.222	0.000	0.000	1.683	1.030

INTERCEPT CONSTANT= 5.85 S.E. OF INTERCEPT= 1.95  
 ESTIMATE OF SIGMA S= 3.39  
 INDEX OF MULTICOLLINEARITY RL= 0.10290140E+01

step-down equation-- one predictor dropped

VARIABLES FOR REGRESSION (RESPONSE VAR. IS LAST)  
 4 5 6 12 20

DET. of correlation matrix of Predictors = 0.95671790E+00

MULTIPLE R-SQ= 0.5837 MULTIPLE R = 0.7640  
 F-VALUE FOR R = 19.98 WITH 4AND 57 DFS

R-BAR SQ(ADJUSTED R-SQ)= 0.5545

CHAR	BETA	BETAxR	STRU-R	REG COEF-B	SE OF B	T-VAL OF B	VIF
4	0.181	5.109	0.215	0.896	0.427	2.101	1.018
5	-0.191	9.213	-0.369	-0.656	0.296	2.212	1.020
6	0.673	78.510	0.892	0.445	0.057	7.781	1.023

12            -0.241            7.169            -0.227            -1.093            0.393            2.780            1.028

INTERCEPT CONSTANT=            6.39    S.E. OF INTERCEPT=            1.95

ESTIMATE OF SIGMA S=            3.44

INDEX OF MULTICOLLINEARITY    RL= 0.10225140E+01

### Problem to be worked out

Yield (gm) (Y)	Hill/m <sup>2</sup> ( X1)	Tiller/ Hill (X2)	Effective tiller/ Hill (X3)	Panicle Length (cm)(X4)	Grain/ Panicle(X5)
y	X1	X2	X3	X4	X5
200	12	22	15	16	40
210	13	24	16	18	45
205	13	23	16	18	42
195	11	20	14	15	35
185	10	19	13	14	30
220	16	24	15	16	50
210	15	22	14	17	45
212	14	22	14	17	45
200	13	20	13	15	40
205	12	21	13	14	42
202	12	20	12	13	41
208	13	21	13	14	42
207	13	21	13	14	42
215	14	23	15	17	47
210	16	22	15	17	45
190	10	18	12	13	35
185	10	17	11	12	32
220	16	23	14	16	47
225	17	25	16	18	49
215	17	22	14	16	47

### Contribution of explanatory variable:

Starting with a set of variables, logical and theoretical aspects will lead the researchers initially to include the core variables in the model. But how to judge the next variable to be included in the model? We have given a rough idea about this in example 10.6 (Note). However, this can be done by judging the change in regression sum of squares due to inclusion of new variable into the model and can be measured in terms of significant improvement in  $R^2$  value. Significance of inclusion of a new variable can be tested with the help of F test.

$$F = \frac{(R^2_{\text{new}} - R^2_{\text{old}})/k' (= \text{No. of new var.})}{(1 - R^2_{\text{new}})/n - k'(\text{new})}$$

with  $H_0(R^2_{\text{New}} - R^2_{\text{old}} = 0)$  at  $(k', n - k)$  d.f.

where  $k'$  is the number of new variables included in the model and  $k$  is the total no. of variables in the new model.

Generally when  $\bar{R}^2$  is increased due to inclusion of a variable in the model then the variable is retained in the model. This happens (approximately) when the t value of the partial regression co-efficient of the newly introduced variable is greater than unity in absolute value (against the test  $H_0: \beta = 0$ ).

Similarly, by introducing a group of variables if the F value increases by unity the  $R^2$  also increases (as thumb rule) and one can retain the new variables because these increase the explanatory power of the new model.

**Structural stability of regression models:**

Suppose the regression models for paddy yield with amount of nitrogen applied before and after the green revolution i.e. during the period 1960-70 (say) and 1971-1980 (say) respectively in India are as follows:

$$Y_t = \alpha_1 + \alpha_2 N_t + u_{1t} \text{ and}$$

$$Y_t = \beta_1 + \beta_2 N_t + u_{2t}$$

Our objective is to verify whether these two relationships are structurally different or not; in other words, to verify whether the parameters of the above equations are different or not. If these are not different then there exists structural stability of the model, otherwise not.

To answer this problem, one can use Chow test as given by Gregory Chow (1960) in “Test of equality between sets of co-efficient in two linear regressions” in *Econometrica*, vol. - 28, No. (3) pp. 591-605. We present below a summary of the Chow test.

**Chow test:**

It is essentially a F-test , the steps are as follows -

Given that

$$\hat{Y}_t = \hat{\alpha}_1 + \hat{\alpha}_2 N_t + \hat{u}_{1t} - \text{pre GR period with } n_1 \text{ no. of observation.}$$

$$\hat{Y}_t = \hat{\beta}_1 + \hat{\beta}_2 N_t + \hat{u}_{2t} - \text{post GR period with } n_2 \text{ no. of observation.}$$

- Assumptions:** (1)  $u_{1t}$  &  $u_{2t} \sim N(0, \sigma^2)$   
 (2)  $u_{1t}$  &  $u_{2t}$  are independently distributed.

**Steps :**

- (1) Construct a regression equation taking all the  $n_1 + n_2$  no. of observations together and let it be  $\hat{Y}_p = \hat{\alpha}_0 + \hat{\beta}_0 N_t + \hat{u}_p$  with  $n_1 + n_2 - 2df$ .
- (2) Add together the RSSs obtained in two different regression equations for two different periods i.e.,  $\sum \hat{u}_{1t}^2 + \sum \hat{u}_{2t}^2$ . The d.f. for it will be  $n_1 - k + n_2 - k = n_1 + n_2 - 2k$ . ( because of assumption - 2).
- (3) Subtract the above sum of squares of the two residuals from the sum of square of residual for pooled sample i.e.,

$$\sum \hat{u}_p^2 - (\sum \hat{u}_{1t}^2 + \sum \hat{u}_{2t}^2) \text{ degree of freedom for the above will be } (n_1 + n_2 - k) -$$



$$(n_1 + n_2 - 2k) = k$$

$$(4) \text{ Perform } F = \frac{\{\sum \hat{u}^2 p - (\sum u^2_{1t} + \sum \hat{u}^2_{2t})\}/k}{(\sum \hat{u}^2_{1t} + \sum \hat{u}^2_{2t})/n_1 + n_2 - 2k}$$

With  $k, n_1 + n_2 - 2k$  d.f. to test  $H_0 : (\alpha_i = \beta_i) i = 1, 2, \dots, k$ .

### Multicollinearity :

Multicollinearity is defined as the presence of linear or near linear relationship among the explanatory variables in a regression analysis/exercise.

Actually while estimating the parameters of regression equations through OLS, one of the important assumption is the assumption of  $\text{Cov}(X_i, X_j) = 0$  where  $i \neq j$

Clearly, the violation of the above assumption (*i.e.* due to the existence of multicollinearity), may have certain implication on the regression analysis.

### Consequences of Multicollinearity:

When multicollinearity is in the form of perfect linear relationship ( $r_{x_i x_j} = 1$ ) among the explanatory variable then a) estimates of the co-efficient are indeterminate and b) standard errors of these estimates become infinitely large.

Let  $\gamma = b_0 + b_1 X_1 + b_2 X_2 + u$  and the explanatory variable  $X_1$  &  $X_2$  are perfectly linearly related or  $X_2 = K X_1$  (say)

$$\Rightarrow \hat{b}_1 = \frac{(\sum x_1 y)(\sum x_2^2) - (\sum x_2 y)(\sum x_1 x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2} \quad \text{and}$$

$$\Rightarrow \hat{b}_2 = \frac{(\sum x_2 y)(\sum x_1^2) - (\sum x_1 y)(\sum x_1 x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2} \quad [\text{x and y are taken in their standardised form}]$$

Let us substitute  $x_2 = k x_1$  in  $b_1$  &  $b_2$  we have

$$\hat{b}_1 = \frac{k^2(\sum x_1 y) - (\sum x_1^2) - k^2(\sum x_1 y)(\sum x_1^2)}{k^2(\sum x_1^2)^2 - k^2(\sum x_1^2)^2} = 0/0$$

$$\begin{aligned} \hat{b}_2 &= \frac{(\sum x_2 y)(\sum x_1^2) - (\sum x_1 y)(\sum x_1 x_2)}{k^2(\sum x_1^2) - (k\sum x_1^2)^2} \\ &= \frac{k(\sum x_1 y)(\sum x_1^2) - k(\sum x_1 y)(\sum x_1^2)}{k^2(\sum x_1^2) - k^2(\sum x_1^2)} = \frac{0}{0} \end{aligned}$$

$$b) \text{ Var } (\hat{b}_1) = \sigma_u^2 \frac{\sum x_2^2}{\sum x_1^2 \sum x_2^2 - (\sum x_1 x_2)^2} = \sigma_u^2 \frac{k^2 \sum x_1^2}{k^2 \sum x_1^2 \sum x_1^2 - k^2 (\sum x_1^2)^2} = \frac{\sigma_u^2 \sum x_1^2}{0} = \infty \quad \text{similarly}$$

$$\text{Var } (\hat{b}_2) = \infty$$

On the other hand if the explanatory variables are not perfectly collinear but are correlated to certain degree *i.e.*  $0 < r_{x_i x_j} < 1$ ; the effects of collinearity are uncertain. Different opinions have been put forwarded by different authors to tackle this type of

multicollinearity : 1) one thing is clear that estimates of the co-efficient do not become biased due to the existence of multicollinearity, because OLS estimates do not require explanatory variables to be un-correlated for unbiasedness, 2) on the contrary sample with multicollinear Xs may give rise to imprecise and unstable estimates.

### Heteroscedasticity

One of the important assumptions of the regression technique is that variance of the disturbance term (*i.e.* conditional on the chosen values of the explanatory variable, X)  $u_i$  should be constant *i.e.*  $E(u_i^2) = \sigma^2$ . But in many cases it is found that  $E(u_i^2) \neq \text{constant}$  rather  $E(u_i^2) = \sigma_i^2$ . Thus the assumption of equal spread (homoscedastic) is violated, giving rise to heteroscedasticity.

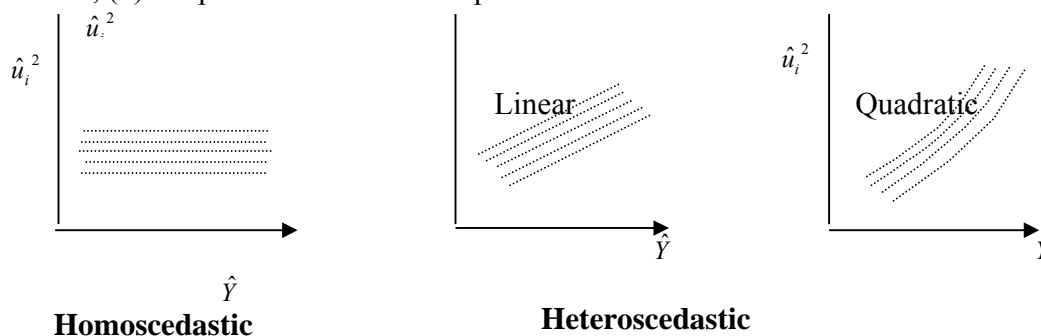
### Reasons for heteroscedasticity

- (1) With the development of more and more modern techniques of model formation, errors are presumed to be reduced.
- (2) As the agricultural income increases farmers' discretionary power increases for disposition of resources, and thereby increasing the possibility of increase in  $\sigma_i^2$ .
- (3) Refinement of data collection technique is another reason. As the data collection technique improves  $\sigma_i^2$  is likely to decrease.
- (4) Presence of outlier increases the  $\sigma_i^2$ .
- (5) Misspecification of CLRM.

**Consequences:** Under heteroscedastic condition the OLS estimates though remain unbiased but loses BLUE property. Moreover the problem of establishing confidence interval and testing of coefficients with 't' and 'F' will give inaccurate result, and mostly provide non-significant coefficient  $\hat{\beta}_i$  because of its large SE.

### Detection of heteroscedasticity:

Detection of heteroscedasticity is not at all an easy task. Several informal and formal procedures are available. Among the informal methods (a) Analysis of the nature of the problem, (b) Graphical methods are important.



### Remedial measure:

If  $\sigma_i^2$  are known then one can go for weighted least square technique (WLS) to obtain BLUE estimators. Weighted least square technique is a special case of Generalized least square

method. In this method OLS is applied on transformed (weighted) variables in such a way to satisfy the standard least square assumptions.

### Regression and the Dummy Variable:

Variables are generally categorized into quantitative and qualitative variables. So far in regression analysis we have discussed about quantitative variables. In this section we shall discuss regression with qualitative variables (*viz.* sex, race, nationality, colour, complexion, religion, response etc.). In regression analysis these qualitative variables are often assigned certain values 0/1, (0 indicating absence of certain attributes and 1, otherwise) are introduced into the regression model as dummy variable. Again, qualitative variables can also be grouped/classified into more than two categories e.g., education status, educated/not educated, primary/other than primary, secondary/other than secondary and so on. In this case each class can be designed as one dummy variable with value 0 (absence) or 1 (presence) and like that. Now, the question is to identify the number of dummy variables to be introduced in the regression model. Simple answer is, introduce number of dummy variables one less than the number of classes of the attribute. This is to avoid the problem of multicollinearity among the explanatory variables. It is clear that so far we have considered dummy variables as explanatory variable. But, it is not necessary that the dummy variable should only be considered as explanatory variable. We will discuss all of them while discussing Regression on dummy dependent variable. For the time being we will discuss only those cases where dummy variables are used as independent variable.

The simplest example of dummy variable is given as follows:

$$Y_i = \alpha + \beta_1 D_i + \beta_2 X_i + u_i$$

where  $Y_i$  is the agricultural income of  $i^{\text{th}}$  farmer

$D_i = 1$ , if the person is educated

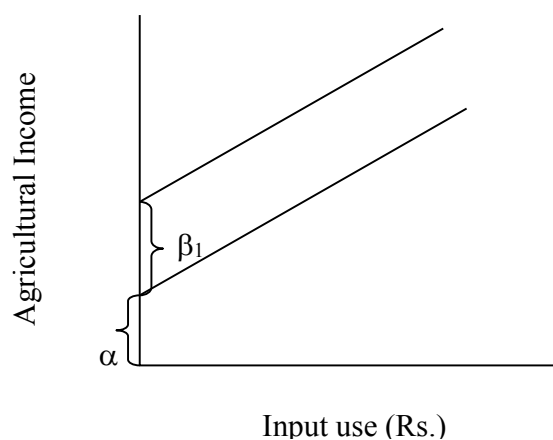
= 0, otherwise.

$X_i$  = input use in rupee

Thus the average agricultural earning of an educated person is given by

$$E(Y_i | D_i = 0) = \alpha + \beta_2 X_i. \text{ Thus this } \alpha \text{ gives the average agricultural earning of uneducated}$$

agricultural farmers and  $\beta_1$  represents the difference in earning between educated and uneducated farmer.



Testing for stability of the regression can be made usual following the procedure laid down earlier.

**Interaction Effects:** It may so happen in many cases there exists interaction between two qualitative variables for example if we include resource availability to a farmer (i.e., resource rich farmer-1 and resource poor farmer-0) in addition to education, we may come across with the situation of interaction because an educated farmer can either be resource rich or resource poor and so on. The regression model in this can be represented as follows:

$$Y_i = \alpha + \alpha_1 D_{1i} + \alpha_2 D_{2i} + \alpha_3 (D_{1i} D_{2i}) + \beta X_i + u_i$$

where,

$\alpha_1$  = Differential effect of being an educated farmer.

$\alpha_2$  = Differential effect of being Resource rich farmer

$\alpha_3$  = Differential effect of being an educated resource rich farmer.

$D_{1i}$  = 1, if educated  
= 0, if uneducated

$D_{2i}$  = 1, if resource rich farmer  
= 0, if resource poor farmer.

**Regression with Dummy Dependent Variable:**

In many of the social studies responses came in the form of yes/no, has/has not, and so on. If one try to predict these answers (dependent variable) having certain other independent information (independent variables) one has to go for regression with dummy dependent variable. Typical characteristics of these types of regression are that the dependent variable is of dichotomous in nature. There are four different approaches to this type of regression analysis.

- 1) The Linear Probability Model (LPM).
- 2) The Logit Model.
- 3) The Probit Model.
- 4) The Tobit Model.

1. **LPM:** We know that possession of heavy agricultural implement depends upon agricultural income of the farmer concerned, along with other factors. Let us take an example:

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

where X = Agricultural income

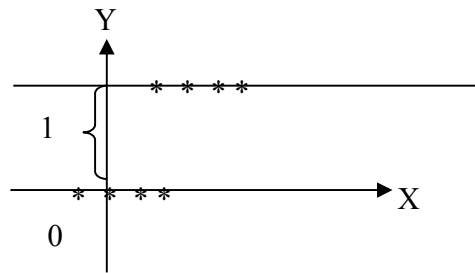
Y = 1, if the farmer has own capital  
= 0, if the farmer does not have own capital.

In this model

$E(Y_i = 1|X_i)$  is the conditional expectation of having capital given a particular agricultural income. In other words it is nothing but the probability of having own capital given agricultural income  $X_i$ . As such this model is known as Linear Probability Model.

Demerit:

- 1) Whatever may be the value of X, the value of Y will either be zero or one, thereby giving rise to a truncated form of scattered diagram.



- 2) Some of the estimated values of Y may be negative or may be greater than one.
- 3) Normality assumption is violated and

Solution: With the help of weighted least square but some of the observation will not be accounted in the method.

2. **The Logit Model:** From the discussion of LPM it is clear that we should have such a probability model in which (1) as  $X_i$  increases  $P = E(Y = 1|X_i)$  but never goes beyond 0 – 1 interval and (2) the relationship should be non-linear between  $E(Y_i)$  and  $X_i$  for the reason that  $E(Y_i)$  tends toward zero slowly as  $X_i$  tends towards two extremities i.e., vary high or very small.

Unlike LPM in Logit  $E(Y = 1|X_i) = P_i$  is given as

$$P_i = \frac{1}{1 + e^{-(\beta_1 + \beta_2 X_i)}}$$

$$= \frac{1}{1 + e^{-z_i}} \text{ [putting } \beta_1 + \beta_2 X_i \text{]}$$

where  $P_i$  is the probability of having own capital so the  $P_i' =$  probability of not having own capital is given by

$$1 - P_i = 1 - \frac{1}{1 + e^{-z_i}} = \frac{1}{1 + e^{z_i}}$$

$\frac{P_i}{1 - P_i}$  is simply odds ratio in favour of having own capital.

$L_i = \ln\left(\frac{P_i}{1 - P_i}\right) = z_i = \beta_1 + \beta_2 X_i$  and this L is called logit; it is both linear in parameter

as well as in variable, where  $\beta_1$  is log odds in favour of having own capital with '0' agricultural income and  $\beta_2$  measures the change in L i.e., log odds in favour of having own capital with change in X.

- ◆ As  $P \rightarrow 1$  from 0,  $L \Rightarrow -\infty \rightarrow +\infty$  i.e., although the probabilities lie between 0 and 1 the logits are not so.
- ◆ Logit is linear in X but L is not linear.

**Steps in formation of Logit Regression:**

1. Calculate  $\hat{p}_i$  for each agricultural income level by  $p_i = \frac{n_i}{N_i}$ , where  $n_i$  = number of persons having a tractor out of total  $N$  number of person having Agricultural Income  $i$ .

Agricultural Income	$N_i$ (Total)	$N_i$ (no. of farmers having own capital)
$x_1$	$N_1$	$n_1$
$x_2$	$N_2$	$n_2$
.	.	.
.	.	.
$x_k$	$N_k$	$n_k$

2. Calculate  $\hat{\alpha}_i = l_n \left[ \frac{\hat{P}_i}{(1 - \hat{P}_i)} \right]$

3. Regress  $\sqrt{w_i} L_i$  on  $\sqrt{w_i} X_i$  with

OLS where  $w_i = N_i \hat{p}_i (1 - \hat{p}_i)$ .

This transformation is done to avoid heteroscedasticity

- 3) **The Probit Model:** Instead of taking logistic CDF (Cumulative Density Function) sometimes it is found suitable to use normal CDF. Thus the model comes out as a result of estimation of normal CDF if popularly known as Probit Model or the Normit Model. A detail discussion on probit analysis has been made in volume two of this book.

**Regression vs. causality:**

Let us consider situations in which two or more variables affects each other with distributed lags. Can any one say the variables causes each other? Can we assign any direction of causality? **Granger’s test for causality** may provide an answer to the above problems. In economics like Cobweb phenomenon there are certain parameters in which the present value of the variable not only depend on its preceding values but also preceding values of the other related variables. For example: price of jute can depend not only on its past values but also on the lag values of production area under jute and lag values of area under competing crops grown during the jute season etc. For the time being let us restrict to two variables only i.e. price of jute ( $P_r$ ) and production of jute ( $P$ ).

$$P_{rt} = \alpha_1 P_{1t} + \alpha_2 P_{2t} + \alpha_3 P_{3t} + \dots + \beta_1 P_{1,t-1} + \beta_2 P_{2,t-2} + \beta_3 P_{3,t-3} \dots + u_{1t}$$

$$= \sum_{i=1}^n \alpha_i P_{r(t-i)} + \sum_{j=1}^n \beta_j P_{r(t-j)} + u_{1t} \tag{1}$$

and,  $P_t = \lambda_1 P_{1t} + \lambda_2 P_{2t} + \lambda_3 P_{3t} + \dots + \gamma_1 P_{1,t-1} + \gamma_2 P_{2,t-2} + \dots + u_{2t}$

$$= \sum_{i=1}^m \lambda_i P_{r(t-i)} + \sum_{j=1}^m \gamma_j P_{j(t-j)} \tag{2}$$

**Assumption:** Production information are given completely in the time series data

- Problem:** (1) Price of jute is caused by production of jute i.e.,  
 $\sum \alpha_i \neq 0$  and  $\sum \gamma_j = 0$   
 (2) Production is caused by price of jute i.e.,  
 $\sum \alpha_i = 0$  and  $\sum \lambda_j \neq 0$

- (3) Two way causality i.e., both price causes production and production causes price i.e. coefficients are statistically significantly different from zero in both the regressions.
- (4) Existence of no causality i.e. coefficients are not statistically significant from zero.

**Procedure:**

- (1) Regress current price on all lagged price only, get residual sum of square (RSS<sub>1</sub>).
- (2) Regress current price on all lagged price and include lagged productions also. Get RSS<sub>2</sub>.
- (3) Calculate
$$F = \frac{(RSS_1 - RSS_2)/m}{RSS_2/(n-k)}$$
 with m, n-k df.  
Where m is the number of lagged production and k is the number of parameter estimated in step-2.
- (4) If Cal F > Tab F then  
H<sub>0</sub>:  $\sum \alpha_i = 0$  is rejected i.e., production causes price.
- (5) Repeat the steps 1-4 with the mode-2 i.e. to check whether P<sub>r</sub>→P.
- (6) Conclude accordingly.

**Demerit:** How to fix the number of lagged terms to be included in the two Regression equations.

**FURTHER READING**

1. Draper, N. R., and Smith, H. (1998). Applied Regression Analysis, 3rd edition. **Wiley**, New York
2. Sahu P K (2013). Agriculture and Applied Statistics - I. 2nd Reprint. **Kalyani Publishers**, New Delhi, India
3. Sahu Pradip Kumar (2013) : Research methodology : A guide for researchers in Agricultural Science, Social Science and Other Related Fields. **Springer**.
4. Sahu Pradip Kumar (2016): Applied Statistics for Agriculture, Veterinary, Fishery, Dairy and Allied Fields. **Springer**.

## **Missing Plot Technique**

**Prof. Anurup Majumder**

Department of Agricultural Statistics,  
Bidhan Chandra Krishi Vishwavidyalaya  
Mohanpur, Nadia-741252

Contact Number: +91-8478912537

E mail I D: [anurupbckv@gmail.com](mailto:anurupbckv@gmail.com)

## **Introduction**

It happens many time in conducting the experiments that some observation are missed. This may happen due to several reasons. For example, in a clinical trial, suppose the readings of blood pressure are to be recorded after three days of giving the medicine to the patients. Suppose the medicine is given to 20 patients and one of the patient doesn't turn up for providing the blood pressure reading. Similarly, in an agricultural experiment, the seeds are sown and yields are to be recorded after few months. Suppose some cattle destroys the crop of any plot or the crop of any plot is destroyed due to storm, insects etc.



In such cases, one option is to

- somehow estimate the missing value on the basis of available data,
- replace it back in the data and make the data set complete.

Now conduct the statistical analysis on the basis of completed data set as if no value was missing by making necessary adjustments in the statistical tools to be applied.

We discuss here the classical missing plot technique proposed by Yates which involve the following steps:

- Estimate the missing observations by the values which makes the error sum of squares to be minimum.
- Substitute the unknown values by the missing observations.
- Express the error sum of squares as a function of these unknown values.
- Minimize the error sum of squares using principle of maxima/minima, i.e., differentiating it with respect to the missing value and put it to zero and form a linear equation.

- Form as many linear equation as the number of unknown values (i.e., differentiate error sum of squares with respect to each unknown value).
- Solve all the linear equations simultaneously and solutions will provide the missing values.
- Impute the missing values with the estimated values and complete the data.
- Apply analysis of variance tools.
- The error sum of squares thus obtained is corrected but treatment sum of squares are not corrected.
- The number of degrees of freedom associated with the total sum of squares are subtracted by the number of missing values and adjusted in the error sum of squares. No change in the degrees of freedom of sum of squares due to treatment is needed.

### One missing observations in RBD

Suppose one observation in (i, j)<sup>th</sup> cell is missing and let this be x. The arrangement of observations in RBD then will look like as follows:

	Blocks							Block Total
		1	2	...	i	...	b	
Treatments	1	$y_{11}$	$y_{21}$	...	$y_{i1}$	...	$y_{b1}$	$B_1 = y_{o1}$
	2	$y_{12}$	$y_{22}$	...	$y_{i2}$	...	$y_{b2}$	$B_2 = y_{o2}$
	.	.	.		.		.	.
	.	.	.		.		.	.
	.	.	.		.		.	.
	j	$y_{1j}$	$y_{2j}$	...	$y_{ij} = x$	...	$y_{bj}$	$B_j = y'_{oj} + x$
	.	.	.		.		.	.
v	$y_{1v}$	$y_{2v}$	...	$y_{iv}$	...	$y_{bv}$	$B_b = y_{ob}$	
Treatment Totals	$T_1 = y_{1o}$	$T_2 = y_{2o}$	=	$T_i = y'_{io} + x$		$y_{vo}$	Grand Total $G = y'_{oo} + x$	

where  $y'_{oo}$  : total of known observations

$y'_{oj}$  : total of known observations in  $j^{\text{th}}$  block

$y'_{oi}$  : total of known observations in  $i^{\text{th}}$  treatment

$$\text{Correction factor (CF)} = \frac{(G')^2}{n} = \frac{(y'_{oo} + x)^2}{bv}$$

$$TSS = \sum_{i=1}^b \sum_{j=1}^v y_{ij}^2 - CF$$

$$= (x^2 + \text{terms which are constant with respect to } x) - CF$$

$$SSBl = \frac{1}{b} [(y'_{io} + x)^2 + \text{terms which are constant with respect to } x] - CF$$

$$SSTr = \frac{1}{v} [(y'_{oj} + x)^2 + \text{terms which are constant with respect to } x] - CF$$

$$SSE = TSS - SSBl - SSTr$$

$$= x^2 - \frac{1}{b}(y'_{io} + x)^2 - \frac{1}{v}(y'_{oj} + x)^2 + \frac{(y'_{oo} + x)^2}{bv} + (\text{terms which are constant with respect to } x) - CF.$$

Find  $x$  such that  $SSE$  is minimum

$$\frac{\partial(SSE)}{\partial x} = 0 \Rightarrow 2x - \frac{2(y'_{io} + x)}{b} - \frac{2(y'_{oj} + x)}{v} + \frac{2(y'_{oo} + x)}{bv} = 0$$

$$\text{or } x = \frac{vy'_{io} + by'_{oj} - y'_{oo}}{(b-1)(v-1)}$$

## Two missing observations in RBD

If there are two missing observations, then let them be  $x$  and  $y$ .

- Let the corresponding row sums (block totals) are  $(R_1 + x)$  and  $(R_2 + y)$ .
- Column sums (treatment totals) are  $(C_1 + x)$  and  $(C_2 + y)$ .
- Total of known observations is  $S$ .

Then

$$SSE = x^2 + y^2 - \frac{1}{b}[(R_1 + x)^2 + (R_2 + y)^2] - \frac{1}{v}[(C_1 + x)^2 + (C_2 + y)^2] + \frac{1}{bv}(S + x + y)^2$$

+ terms independent of  $x$  and  $y$ .

Now differentiate  $SSE$  with respect to  $x$  and  $y$ , as

$$\frac{\partial(SSE)}{\partial x} = 0 \Rightarrow x - \frac{R_1 + x}{b} - \frac{C_1 + x}{b} + \frac{S + x + y}{bv} = 0$$

$$\frac{\partial(SSE)}{\partial y} = 0 \Rightarrow y - \frac{R_2 + y}{v} - \frac{C_2 + y}{v} + \frac{S + x + y}{bv} = 0.$$

Thus solving the following two linear equations in  $x$  and  $y$ , we obtain the estimated missing values

$$(b-1)(v-1)x = bR_1 + vC_1 - S - y$$

$$(b-1)(v-1)y = bR_2 + vC_2 - S - x.$$

### **Adjustments to be done in analysis of variance**

- (i) Obtain the within block sum of squares from incomplete data.
- (ii) Subtract correct error sum of squares from (i) . This given the correct treatment sum of squares.
- (iii) Reduce the degrees of freedom of error sum of squares by the number of missing observations.
- (iv) No adjustments in other sum of squares are required.

#### **Example:-**

To find out the best source of nitrogen at 60 kg./ha. , an experiment was conducted on Paddy with five sources of nitrogen in four randomized blocks at paddy breeding centre, Coimbatore, Tamil Nadu . The yield (Kg./Plot) data for different treatment are given in the following table, Analyze the Data.

<b>Blocks</b>	<b>Ammonium Sulphate</b>	<b>Ammonium chloride</b>	<b>Urea</b>	<b>Chilean nitrate</b>	<b>Ammonium Sulphate nitrate</b>
<b>1</b>	<b>25.4</b>	<b>32.5</b>	<b>37.5</b>	<b>22.5</b>	<b>20.5</b>
<b>2</b>	<b>17.3</b>	<b>_</b>	<b>25.4</b>	<b>14.7</b>	<b>21.5</b>
<b>3</b>	<b>22.4</b>	<b>28.4</b>	<b>30.1</b>	<b>23.5</b>	<b>23.5</b>
<b>4</b>	<b>30.5</b>	<b>33.4</b>	<b>34.5</b>	<b>22.4</b>	<b>28.5</b>

## Basics of Regression and PCA

Asok K. Nanda

Department of Mathematics & Statistics  
IISER Kolkata

June 2, 2017



Asok K. Nanda, IISER Kolkata

Principal Component Analysis

## Model and Assumptions

The model is

$$\mathbf{y}^{n \times 1} = \mathbf{X}^{n \times k} \boldsymbol{\beta}^{k \times 1} + \mathbf{u}$$

### Assumptions:

- (i) The relation between response and regressors is linear in parameters.
- (ii)  $E(\mathbf{u}) = 0$
- (iii)  $V(\mathbf{u}) = \sigma^2 I$
- (iv)  $Cov(u_i, u_j) = 0$ . [This follows from (iii)]
- (v)  $\mathbf{u} \sim \text{Normal}$ . [This and (iv) give that  $u_i$  are independent]
- (vi)  $X$  is non-stochastic
- (vii)  $R(X) = k$

### Questions:

- What will happen if one or more assumption(s) is/are violated?
- What remedial measures can be taken in case of violation of assumption(s)?



Asok K. Nanda, IISER Kolkata

Principal Component Analysis

## Effect of Violation of Assumptions

- If (i) is violated, **non-linear regression** technique will be adopted.
- If (ii) is violated, then  $E(u_i) = w (\neq 0) \forall i$ . In this case the model is

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \dots + \beta_k x_{ki} + u_i$$

which gives

$$E(y_i | x_i, i = 1, 2, \dots, k) = (\beta_1 + w) = \beta_2 x_{2i} + \beta_3 x_{3i} + \dots + \beta_k x_{ki}$$

Here  $\beta_1$  cannot be estimated unbiasedly. But  $\beta_1$  is of less importance and the slope terms are more important (and they remain unaffected). Again, if  $E(u_i) = w_i$ , then the slope coefficient may be biased and this case may be crucial.



## Effect of Violation of Assumptions (Contd.)

- Violation of (iii) leads to **heteroscedasticity**.

**Example** For a person learning typing, the variance of the mistakes done in typing does not remain constant day by day.

If the non-constant error variance is not eliminated, the LSEs will still be unbiased but the minimum variance property may be lost, and as a result, the regression coefficients will have larger variance.

The OLS does not work here. We need to give less weightage to the observations coming from populations with higher variability. This is done in **weighted least square method**.

We also use **delta method** to stabilize the variance.



## Effect of Violation of Assumptions (Contd.)

- If (iv) is violated, we have a auto-regressive process

$$y_t = \rho y_{t-1} + u_t$$

**Example:** Suppose one family consumes 20kg of rice per month. If in a month some guests come and 30 kg of rice is bought, and some amount of rice becomes extra, this will affect the amount of rice bought in the next month. So, the rice bought in two months are not uncorrelated. If the presence of guests in a house is very common, then this could be included in the model. This correlation is called autocorrelation.



## Effect of Violation of Assumptions (Contd.)

If the errors are positively autocorrelated, the variance will be seriously under-estimated, as a result, the testing of hypotheses of regression coefficients may show that one particular regression coefficient contribute significantly to the model when actually it is not. Confidence interval will be shorter than usual. Generally, under-estimation of variance gives the analysts a false impression of accuracy.

If the autocorrelation is due to omission of regressor and that is identified, then include that in the model to avoid auto-correlation.

Autocorrelation may be detected by Durbin-Watson Test and Cochrane-Orcutt Method may be applied to get rid of autocorrelation.





## Effect of Violation of Assumptions (Contd.)

- Assumption (v) is not essential if our objective is estimation only. LSEs are BLUE even if normality is not there. However, for testing purpose we need this assumption.
- If we are working with secondary data, then we do not have any control on Assumption (vi). However, if  $x_i$  are stochastic, the data can be analysed with bit more calculations.
- Assumption (vii) is violated, we encounter the problem of **multicollinearity**. If  $|r_{12}| \rightarrow 1$ , then  $V(\hat{\beta}_1) \rightarrow \infty$  and  $Cov(\hat{\beta}_1, \hat{\beta}_2) \rightarrow \pm\infty$ . This means that the different samples may lead to different estimators that could be widely apart. The CI will be wider. However, the LSEs are still BLUE.

Multicollinearity can be handled by **collecting additional data** or **model specification**.



## Why PCA

The main use of PCA is to reduce the dimension of the data under consideration. To be specific:

- Suppose a response variable  $Y$  is to be regressed against a large number of covariates.
- Retaining all covariates may lead to severe multicollinearity or non-identifiability of regression coefficients.
- Standard errors will be unacceptably large, and predictions may be very inaccurate, if no remedy is undertaken.



## What does PCA do?

$\mathbf{X}$  : A  $p$ -dimensional vector

$\Sigma$  :  $V(\mathbf{X})$

$\lambda_1 > \lambda_2 > \dots > \lambda_p$  : Eigenvalues of  $\Sigma$

**Assumption:**  $p$  is large.

- Finds a small set of linear combinations of the covariates which are uncorrelated with each other. This will avoid the multicollinearity problem.
- Ensures that the linear combinations chosen have maximal variance. A good regression design chooses values of the covariates which are spread out.
- This fewer number of linear combinations account for most of the variation present in  $\mathbf{X}$ , measured by trace of  $\Sigma$ .



## How many Principal Components should one use?

- The objective is to use only the first few components.
- The usual technique is to look for where there is a sharp drop in the component variance. Remember that a good regression design will have spread out covariates, so the components with small variance (i.e. small eigenvalues) will be omitted.



## Mathematical Formulation of PCA

- Let  $\mathbf{X}$  be replaced by  $l_1' \mathbf{X}$  where  $l_1$  is such that  $V(l_1' \mathbf{X})$  is maximum with  $l_1' l_1 = 1$ .
- This restriction is required, otherwise variance can be infinitely large.
- The desired linear combination is then  $u_1$  with  $V(u_1) = \lambda_1$
- If  $\frac{\lambda_1}{\sum_{i=1}^p \lambda_i}$ , the proportion of variation explained by  $u_1$ , is large (95% or so), then  $\mathbf{X}$  can be well replaced by  $u_1$
- Otherwise, we consider another linear combination  $l_2' \mathbf{X} \ni V(l_2' \mathbf{X})$  is maximum with  $l_2' l_2 = 1$  and  $Cov(l_2' \mathbf{X}, u_1) = 0$
- The desired linear combination is then  $u_2$  with  $V(u_2) = \lambda_2$
- If  $\frac{\lambda_1 + \lambda_2}{\sum_{i=1}^p \lambda_i}$ , the proportion of variation explained by  $u_1$  and  $u_2$ , is large then  $\mathbf{X}$  can be well replaced by  $u_1$  and  $u_2$ .
- Continue



## Calculation of PCA

To find PCs and their variances  $\equiv$  To find eigenvalues  $\lambda_i$  of  $\Sigma$  and the corresponding eigenvectors  $\beta_i \ni \beta_i' \beta_i = 1, \beta_i' \beta_j = 0$  for all  $i \neq j$

### Calculation of $\lambda_1$ :

- Start with any arbitrary vector  $\mathbf{x}_0$  and compute successively

$$\mathbf{x}_i = \frac{\Sigma \cdot \mathbf{x}_{i-1}}{\sqrt{\mathbf{x}_{i-1}' \Sigma \mathbf{x}_{i-1}}}, ; i = 1, 2, \dots$$

- If  $\mathbf{x}_{i-1} \approx \mathbf{x}_i$ , then  $\beta_1 = \frac{\mathbf{x}_i}{\sqrt{\mathbf{x}_i' \mathbf{x}_i}}$  and  $\lambda_1 = \sqrt{\mathbf{x}_i' \Sigma \mathbf{x}_i}$

### Calculation of $\lambda_2 (> \lambda_1)$ :

- Define  $\Sigma_2 = \Sigma - \lambda_1 \beta_1 \beta_1'$
- Apply above procedure with  $\Sigma$  replaced by  $\Sigma_2$
- Continue



## Example

**Note:** If the variables do not have the same unit of measurement, we replace  $X_i$  by  $Z_i = \frac{X_i - \mu_i}{\sigma_i}$  to standardize.

**Example:** In order to study the yield of a certain crop, five characteristics *Irrigation, Fertilizer, Temperature, Pesticide, Soil Quality* were measured in 14 consecutive years. The data obtained were standardized. The variance-covariance matrix is calculated as under.

$$\Sigma = \begin{pmatrix} 4.308 & 1.683 & 1.803 & 2.155 & -0.253 \\ & 1.768 & 0.588 & 0.177 & 0.176 \\ & & 0.801 & 1.065 & -0.158 \\ & & & 1.970 & -0.357 \\ & & & & 0.504 \end{pmatrix}$$



## Example (Contd.)

The eigenvalues are obtained as (6.931, 1.786, 0.390, 0.230, 0.014). The corresponding eigenvectors are

$$\begin{aligned} & (0.781, 0.306, 0.334, 0.426, -0.054), \\ & (-0.071, -0.764, 0.083, 0.579, -0.262), \\ & (0.004, -0.162, 0.015, 0.220, 0.962), \\ & (0.542, -0.545, 0.050, -0.636, -0.051), \\ & (-0.302, -0.010, 0.937, -0.173, 0.024) \end{aligned}$$

It is observed that 74.1% of the total variation is explained by the first principal component

$$u_1 = 0.781X_1 + 0.306X_2 + 0.334X_3 + 0.426X_4 - 0.054X_5$$



## Example (Contd.)

93.2% of the total variation is explained jointly by  $u_1$ . The second principal component is

$$u_2 = -0.071X_1 - 0.764X_2 + 0.083X_3 + 0.579X_4 - 0.262X_5.$$

97.4% of the total variation is explained jointly by  $u_1$ ,  $u_2$  and the third principal component is

$$u_3 = 0.004X_1 - 0.162X_2 + 0.015X_3 + 0.220X_4 + 0.962X_5.$$

**Remark:** Although PCA reduces the data size, it is not advisable to use for regression analysis.



# THANK YOU



## Probability : How to Model?

**Asok K. Nanda**

Department of Mathematics & Statistics  
IISER Kolkata

June 2, 2015



Asok K. Nanda, IISER Kolkata

Probability : How to Model?

## Modeling

**Aim:** To learn how to model a random phenomenon.

**Random Experiment:** An experiment is said to be random if

- 1 the outcomes of the experiment be known in advance
- 2 the outcome of a particular performance cannot be predicted with certainty
- 3 the experiment may be repeated under identical conditions

**Sample Space:** Collection of all possible outcomes of a random experiment. It is generally denoted by  $\Omega$ .



Asok K. Nanda, IISER Kolkata

Probability : How to Model?

## Modeling

**Aim:** To learn how to model a random phenomenon.

**Random Experiment:** An experiment is said to be random if

- 1 the outcomes of the experiment be known in advance
- 2 the outcome of a particular performance cannot be predicted with certainty
- 3 the experiment may be repeated under identical conditions

**Sample Space:** Collection of all possible outcomes of a random experiment. It is generally denoted by  $\Omega$ .



## Modeling

**Aim:** To learn how to model a random phenomenon.

**Random Experiment:** An experiment is said to be random if

- 1 the outcomes of the experiment be known in advance
- 2 the outcome of a particular performance cannot be predicted with certainty
- 3 the experiment may be repeated under identical conditions

**Sample Space:** Collection of all possible outcomes of a random experiment. It is generally denoted by  $\Omega$ .



## Modeling

**Aim:** To learn how to model a random phenomenon.

**Random Experiment:** An experiment is said to be random if

- 1 the outcomes of the experiment be known in advance
- 2 the outcome of a particular performance cannot be predicted with certainty
- 3 the experiment may be repeated under identical conditions

**Sample Space:** Collection of all possible outcomes of a random experiment. It is generally denoted by  $\Omega$ .



## Modeling

**Aim:** To learn how to model a random phenomenon.

**Random Experiment:** An experiment is said to be random if

- 1 the outcomes of the experiment be known in advance
- 2 the outcome of a particular performance cannot be predicted with certainty
- 3 the experiment may be repeated under identical conditions

**Sample Space:** Collection of all possible outcomes of a random experiment. It is generally denoted by  $\Omega$ .





## Modeling (contd.)

**Experiment 1:** Tossing of a coin.

$$\Omega = \{Head, Tail\}$$

To model this experiment, we associate some probability with each of the elements in  $\Omega$ . Let the probability of the outcome 'Head' be  $p$  and the probability of the outcome 'Tail' be  $q$ , w h e r e  $p, q \geq 0$  a n d  $p + q = 1$ .

If the coin is known to be unbiased, we have  $p = q = 0.5$ .



## Modeling (contd.)

**Experiment 1:** Tossing of a coin.

$$\Omega = \{Head, Tail\}$$

To model this experiment, we associate some probability with each of the elements in  $\Omega$ . Let the probability of the outcome 'Head' be  $p$  and the probability of the outcome 'Tail' be  $q$ , w h e r e  $p, q \geq 0$  a n d  $p + q = 1$ .

If the coin is known to be unbiased, we have  $p = q = 0.5$ .



## Modeling (contd.)

**Experiment 1:** Tossing of a coin.

$$\Omega = \{Head, Tail\}$$

To model this experiment, we associate some probability with each of the elements in  $\Omega$ . Let the probability of the outcome 'Head' be  $p$  and the probability of the outcome 'Tail' be  $q$ , where  $p, q \geq 0$  and  $p + q = 1$ .

If the coin is known to be unbiased, we have  $p = q = 0.5$ .



## Modeling (contd.)

**Experiment 1:** Tossing of a coin.

$$\Omega = \{Head, Tail\}$$

To model this experiment, we associate some probability with each of the elements in  $\Omega$ . Let the probability of the outcome 'Head' be  $p$  and the probability of the outcome 'Tail' be  $q$ , where  $p, q \geq 0$  and  $p + q = 1$ .

If the coin is known to be unbiased, we have  $p = q = 0.5$ .



## Modeling (contd.)

**Experiment 1:** Tossing of a coin.

$$\Omega = \{Head, Tail\}$$

To model this experiment, we associate some probability with each of the elements in  $\Omega$ . Let the probability of the outcome 'Head' be  $p$  and the probability of the outcome 'Tail' be  $q$ , where  $p, q \geq 0$  and  $p + q = 1$ .

If the coin is known to be unbiased, we have  $p = q = 0.5$ .



## Modeling (contd.)

**Experiment 2:** Tossing of one usual die.

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

To model this experiment, let us take  $p_i$  as the probability of getting number  $i$ ,  $i = 1, 2, \dots, 6$ . Clearly,  $p_i \geq 0$  for all  $i$  and  $\sum_{i=1}^6 p_i = 1$ .



## Modeling (contd.)

**Experiment 2:** Tossing of one usual die.

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

To model this experiment, let us take  $p_i$  as the probability of getting number  $i$ ,  $i = 1, 2, \dots, 6$ . Clearly,  $p_i \geq 0$  for all  $i$  and  $\sum_{i=1}^6 p_i = 1$ .



## Modeling (contd.)

**Experiment 2:** Tossing of one usual die.

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

To model this experiment, let us take  $p_i$  as the probability of getting number  $i$ ,  $i = 1, 2, \dots, 6$ . Clearly,  $p_i \geq 0$  for all  $i$  and  $\sum_{i=1}^6 p_i = 1$ .



## Modeling (contd.)

**Experiment 2:** Tossing of one usual die.

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

To model this experiment, let us take  $p_i$  as the probability of getting number  $i$ ,  $i = 1, 2, \dots, 6$ . Clearly,  $p_i \geq 0$  for all  $i$  and  $\sum_{i=1}^6 p_i = 1$ .



## Modeling (contd.)

Now we consider a different kind of experiment.

**Experiment 3:** Choose a number at random from  $(0, 1]$ .

$$\Omega = \{x : 0 < x \leq 1\},$$

an uncountable sample space.

Let us model this experiment in the same way as we have done in the previous experiment. For doing this, we associate, to every number  $x \in (0, 1]$ , a number  $p(x) \geq 0$ , called the probability of  $x$ , such that

$$\sum_{x \in (0, 1]} p(x) = 1$$



## Modeling (contd.)

Now we consider a different kind of experiment.

**Experiment 3:** Choose a number at random from  $(0, 1]$ .

$$\Omega = \{x : 0 < x \leq 1\},$$

an uncountable sample space.

Let us model this experiment in the same way as we have done in the previous experiment. For doing this, we associate, to every number  $x \in (0, 1]$ , a number  $p(x) \geq 0$ , called the probability of  $x$ , such that

$$\sum_{x \in (0, 1]} p(x) = 1$$



## Modeling (contd.)

Now we consider a different kind of experiment.

**Experiment 3:** Choose a number at random from  $(0, 1]$ .

$$\Omega = \{x : 0 < x \leq 1\},$$

an uncountable sample space.

Let us model this experiment in the same way as we have done in the previous experiment. For doing this, we associate, to every number  $x \in (0, 1]$ , a number  $p(x) \geq 0$ , called the probability of  $x$ , such that

$$\sum_{x \in (0, 1]} p(x) = 1$$



## Modeling (contd.)

Now we consider a different kind of experiment.

**Experiment 3:** Choose a number at random from  $(0, 1]$ .

$$\Omega = \{x : 0 < x \leq 1\},$$

an uncountable sample space.

Let us model this experiment in the same way as we have done in the previous experiment. For doing this, we associate, to every number  $x \in (0, 1]$ , a number  $p(x) \geq 0$ , called the probability of  $x$ , such that

$$\sum_{x \in (0, 1]} p(x) = 1$$



## Modeling (contd.)

In order to have some idea about the value of  $p(x)$ , let us first study the chance of the outcome belonging to the different intervals of  $(0, 1]$ .

First the interval is divided into two equal halves as  $(0, 1/2]$  and  $(1/2, 1]$ . Since the number  $X$  is chosen at random from  $(0, 1]$ , we have

$$P(X \leq 1/2) = P(X > 1/2) = 1/2.$$

Now, let us divide the interval into 4 equal parts as

$(0, 1/4]$ ,  $(1/4, 1/2]$ ,  $(1/2, 3/4]$ ,  $(3/4, 1]$ . Clearly,

$$\begin{aligned} P(X \leq 1/4) &= P(1/4 < X \leq 1/2) \\ &= P(1/2 < X \leq 3/4) = P(3/4 < X \leq 1) = 1/4 \end{aligned}$$



## Modeling (contd.)

In order to have some idea about the value of  $p(x)$ , let us first study the chance of the outcome belonging to the different intervals of  $(0, 1]$ . First the interval is divided into two equal halves as  $(0, 1/2]$  and  $(1/2, 1]$ . Since the number  $X$  is chosen at random from  $(0, 1]$ , we have

$$P(X \leq 1/2) = P(X > 1/2) = 1/2.$$

Now, let us divide the interval into 4 equal parts as

$(0, 1/4]$ ,  $(1/4, 1/2]$ ,  $(1/2, 3/4]$ ,  $(3/4, 1]$ . Clearly,

$$\begin{aligned} P(X \leq 1/4) &= P(1/4 < X \leq 1/2) \\ &= P(1/2 < X \leq 3/4) = P(3/4 < X \leq 1) = 1/4 \end{aligned}$$



## Modeling (contd.)

In order to have some idea about the value of  $p(x)$ , let us first study the chance of the outcome belonging to the different intervals of  $(0, 1]$ . First the interval is divided into two equal halves as  $(0, 1/2]$  and  $(1/2, 1]$ . Since the number  $X$  is chosen at random from  $(0, 1]$ , we have

$$P(X \leq 1/2) = P(X > 1/2) = 1/2.$$

Now, let us divide the interval into 4 equal parts as

$(0, 1/4]$ ,  $(1/4, 1/2]$ ,  $(1/2, 3/4]$ ,  $(3/4, 1]$ . Clearly,

$$\begin{aligned} P(X \leq 1/4) &= P(1/4 < X \leq 1/2) \\ &= P(1/2 < X \leq 3/4) = P(3/4 < X \leq 1) = 1/4 \end{aligned}$$





## Modeling (contd.)

In order to have some idea about the value of  $p(x)$ , let us first study the chance of the outcome belonging to the different intervals of  $(0, 1]$ . First the interval is divided into two equal halves as  $(0, 1/2]$  and  $(1/2, 1]$ . Since the number  $X$  is chosen at random from  $(0, 1]$ , we have

$$P(X \leq 1/2) = P(X > 1/2) = 1/2.$$

Now, let us divide the interval into 4 equal parts as

$(0, 1/4]$ ,  $(1/4, 1/2]$ ,  $(1/2, 3/4]$ ,  $(3/4, 1]$ . Clearly,

$$\begin{aligned} P(X \leq 1/4) &= P(1/4 < X \leq 1/2) \\ &= P(1/2 < X \leq 3/4) = P(3/4 < X \leq 1) = 1/4 \end{aligned}$$



## Modeling (contd.)

In general, if we partition  $(0, 1]$  into  $2^n$  intervals as

$$\left( \frac{k}{2^n}, \frac{k+1}{2^n} \right], \quad k = 0, 1, \dots, (2^n - 1), \quad n \geq 1,$$

then we have

$$P\left( \frac{k}{2^n} < X \leq \frac{k+1}{2^n} \right) = \frac{1}{2^n},$$

for  $k = 0, 1, 2, \dots, (2^n - 1), \quad n \geq 1$ .



## Modeling (contd.)

**Note:** As the partition becomes finer (*i.e.*,  $n \rightarrow \infty$ ), the chance of the outcome belonging to a specific interval decreases, *i.e.*,  $1/2^n \downarrow 0$ , as  $n \uparrow \infty$ .

**Question:** As  $n \rightarrow \infty$ , can we get  $p(x) > 0$ , for any particular  $x \in (0, 1]$ ?



## Modeling (contd.)

**Note:** As the partition becomes finer (*i.e.*,  $n \rightarrow \infty$ ), the chance of the outcome belonging to a specific interval decreases, *i.e.*,  $1/2^n \downarrow 0$ , as  $n \uparrow \infty$ .

**Question:** As  $n \rightarrow \infty$ , can we get  $p(x) > 0$ , for any particular  $x \in (0, 1]$ ?



## Modeling (contd.)

If possible, let  $p(x) = \epsilon > 0$ , for some  $x \in (0, 1]$  ( $\epsilon$  is small enough).

Now,  $\exists$  some  $n \geq 1 \Rightarrow \frac{1}{2^n} < \epsilon$

Since  $x \in (0, 1]$ ,

$$\exists k \in \{0, 1, 2, \dots, 2^n - 1\} \Rightarrow x \in \left( \frac{k}{2^n}, \frac{k+1}{2^n} \right]$$

Thus,

$$p(x) \leq P_n \left[ \frac{k}{2} X \leq \leq < \frac{k+1}{2^n} \right] = \frac{1}{2^n}$$

*i.e.*,  $\epsilon \leq \frac{1}{2^n} < \epsilon$ . Hence,  $P(X = x) = p(x) = 0 \forall x \in (0, 1]$ .



## Modeling (contd.)

If possible, let  $p(x) = \epsilon > 0$ , for some  $x \in (0, 1]$  ( $\epsilon$  is small enough).

Now,  $\exists$  some  $n \geq 1 \Rightarrow \frac{1}{2^n} < \epsilon$

Since  $x \in (0, 1]$ ,

$$\exists k \in \{0, 1, 2, \dots, 2^n - 1\} \Rightarrow x \in \left( \frac{k}{2^n}, \frac{k+1}{2^n} \right]$$

Thus,

$$p(x) \leq P_n \left[ \frac{k}{2} X \leq \leq < \frac{k+1}{2^n} \right] = \frac{1}{2^n}$$

*i.e.*,  $\epsilon \leq \frac{1}{2^n} < \epsilon$ . Hence,  $P(X = x) = p(x) = 0 \forall x \in (0, 1]$ .



## Modeling (contd.)

If possible, let  $p(x) = \epsilon > 0$ , for some  $x \in (0, 1]$  ( $\epsilon$  is small enough).

Now,  $\exists$  some  $n \geq 1 \Rightarrow \frac{1}{2^n} < \epsilon$

Since  $x \in (0, 1]$ ,

$$\exists k \in \{0, 1, 2, \dots, 2^n - 1\} \Rightarrow x \in \left( \frac{k}{2^n}, \frac{k+1}{2^n} \right]$$

Thus,

$$p(x) \leq P \left[ \frac{k}{2^n} < X \leq \frac{k+1}{2^n} \right] = \frac{1}{2^n} < \epsilon$$

*i.e.*,  $\epsilon \leq \frac{1}{2^n} < \epsilon$ . Hence,  $P(X = x) = p(x) = 0 \forall x \in (0, 1]$ .



## Modeling (contd.)

If possible, let  $p(x) = \epsilon > 0$ , for some  $x \in (0, 1]$  ( $\epsilon$  is small enough).

Now,  $\exists$  some  $n \geq 1 \Rightarrow \frac{1}{2^n} < \epsilon$

Since  $x \in (0, 1]$ ,

$$\exists k \in \{0, 1, 2, \dots, 2^n - 1\} \Rightarrow x \in \left( \frac{k}{2^n}, \frac{k+1}{2^n} \right]$$

Thus,

$$p(x) \leq P \left[ \frac{k}{2^n} < X \leq \frac{k+1}{2^n} \right] = \frac{1}{2^n} < \epsilon$$

*i.e.*,  $\epsilon \leq \frac{1}{2^n} < \epsilon$ . Hence,  $P(X = x) = p(x) = 0 \forall x \in (0, 1]$ .



## Modeling (contd.)

If possible, let  $p(x) = \epsilon > 0$ , for some  $x \in (0, 1]$  ( $\epsilon$  is small enough).

Now,  $\exists$  some  $n \geq 1 \Rightarrow \frac{1}{2^n} < \epsilon$

Since  $x \in (0, 1]$ ,

$$\exists k \in \{0, 1, 2, \dots, 2^n - 1\} \Rightarrow x \in \left( \frac{k}{2^n}, \frac{k+1}{2^n} \right]$$

Thus,

$$p(x) \leq P \left[ \frac{k}{2^n} < X \leq \frac{k+1}{2^n} \right] = \frac{1}{2^n} < \epsilon$$

*i.e.*,  $\epsilon \leq \frac{1}{2^n} < \epsilon$ . Hence,  $P(X = x) = p(x) = 0 \forall x \in (0, 1]$ .



## Modeling (contd.)

Thus, the probability associated with each outcome of the experiment is zero. This is surprising to note that each point from  $(0, 1]$  has probability zero although  $P(X \in (0, 1]) = 1$ . This reminds us of the fact that the mass of a point on a paper is zero, although the mass of a portion of a paper, which consists of uncountably many points, is non-zero.



## Modeling (contd.)

Thus, the probability associated with each outcome of the experiment is zero. This is surprising to note that each point from  $(0, 1]$  has probability zero although  $P(X \in (0, 1]) = 1$ . This reminds us of the fact that the mass of a point on a paper is zero, although the mass of a portion of a paper, which consists of uncountably many points, is non-zero.



## Modeling (contd.)

Thus, the probability associated with each outcome of the experiment is zero. This is surprising to note that each point from  $(0, 1]$  has probability zero although  $P(X \in (0, 1]) = 1$ . This reminds us of the fact that the mass of a point on a paper is zero, although the mass of a portion of a paper, which consists of uncountably many points, is non-zero.



## Modeling (contd.)

**Note:** Clearly, this kind of modeling (which is obtained by following the one for Experiment 1 and Experiment 2) is of no use. This shows that we cannot imitate the model of discrete sample space exactly.

**Question:** How to model then?



## Modeling (contd.)

**Note:** Clearly, this kind of modeling (which is obtained by following the one for Experiment 1 and Experiment 2) is of no use. This shows that we cannot imitate the model of discrete sample space exactly.

**Question:** How to model then?



## Modeling (contd.)

Note that our aim is to find the probability of an event  $A \subseteq \Omega$ , by associating probabilities to the outcomes of an experiment, such that  $P(A) = \sum_{x \in A} p(x)$ . But we have seen from the above discussion that, for an uncountably many outcomes, we cannot talk about the probability of individual outcome. So, we do not want to define probability of outcome and then probability of events. Rather, we try to define probability of an event directly.



## Modeling (contd.)

Note that our aim is to find the probability of an event  $A \subseteq \Omega$ , by associating probabilities to the outcomes of an experiment, such that  $P(A) = \sum_{x \in A} p(x)$ . But we have seen from the above discussion that, for an uncountably many outcomes, we cannot talk about the probability of individual outcome. So, we do not want to define probability of outcome and then probability of events. Rather, we try to define probability of an event directly.





## Modeling (contd.)

Note that our aim is to find the probability of an event  $A \subseteq \Omega$ , by associating probabilities to the outcomes of an experiment, such that  $P(A) = \sum_{x \in A} p(x)$ . But we have seen from the above discussion that, for an uncountably many outcomes, we cannot talk about the probability of individual outcome. So, we do not want to define probability of outcome and then probability of events. Rather, we try to define probability of an event directly.



## Modeling (contd.)

Before we answer this, let us talk of events.

Question: What is an event?

Answer: Any subset of  $\Omega$  is an event(?)

Let  $\mathbb{F}$  be the class of all subsets of  $\Omega = (0, 1]$ . Now, for every event  $A \in \mathbb{F}$  we want to associate a number  $P(A)$  to denote the probability that selected outcome is in  $A$ . So,  $P$  must satisfy the following properties.



## Modeling (contd.)

Before we answer this, let us talk of events.

**Question:** What is an event?

**Answer:** Any subset of  $\Omega$  is an event(?)

Let  $\mathbb{F}$  be the class of all subsets of  $\Omega = (0, 1]$ . Now, for every event  $A \in \mathbb{F}$  we want to associate a number  $P(A)$  to denote the probability that selected outcome is in  $A$ . So,  $P$  must satisfy the following properties.



## Modeling (contd.)

Before we answer this, let us talk of events.

**Question:** What is an event?

**Answer:** Any subset of  $\Omega$  is an event(?)

Let  $\mathbb{F}$  be the class of all subsets of  $\Omega = (0, 1]$ . Now, for every event  $A \in \mathbb{F}$  we want to associate a number  $P(A)$  to denote the probability that selected outcome is in  $A$ . So,  $P$  must satisfy the following properties.



## Modeling (contd.)

Before we answer this, let us talk of events.

**Question:** What is an event?

**Answer:** Any subset of  $\Omega$  is an event(?)

Let  $\mathbb{F}$  be the class of all subsets of  $\Omega = (0, 1]$ . Now, for every event  $A \in \mathbb{F}$  we want to associate a number  $P(A)$  to denote the probability that selected outcome is in  $A$ . So,  $P$  must satisfy the following properties.



## Modeling (contd.)

Before we answer this, let us talk of events.

**Question:** What is an event?

**Answer:** Any subset of  $\Omega$  is an event(?)

Let  $\mathbb{F}$  be the class of all subsets of  $\Omega = (0, 1]$ . Now, for every event  $A \in \mathbb{F}$  we want to associate a number  $P(A)$  to denote the probability that selected outcome is in  $A$ . So,  $P$  must satisfy the following properties.



## Modeling (contd.)

Before we answer this, let us talk of events.

**Question:** What is an event?

**Answer:** Any subset of  $\Omega$  is an event(?)

Let  $\mathbb{F}$  be the class of all subsets of  $\Omega = (0, 1]$ . Now, for every event  $A \in \mathbb{F}$  we want to associate a number  $P(A)$  to denote the probability that selected outcome is in  $A$ . So,  $P$  must satisfy the following properties.



## Modeling (contd.)

For every  $A \subseteq (0, 1]$ ,

$$(i) 0 \leq P(A) \leq 1$$

$$(ii) P(\emptyset) = 0, P(\Omega) = 1$$

(iii) If  $A_1, A_2, \dots \in \mathbb{F}$ , with  $A_i \cap A_j = \emptyset$ , for  $i \neq j$ , then

$$P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

[If  $A = \{x_1, x_2, \dots, x_k\} \subseteq \Omega$  where  $x_i$ 's are the outcomes of the experiment, then  $P(A) = \sum_{i=1}^k p(x_i) = \sum_{i=1}^k P(X = x_i)$ , where  $X$  is the number chosen from  $(0, 1]$ .]

$$(iv) P[X \in \left(\frac{k-1}{2^n}, \frac{k}{2^n}\right]] = \frac{1}{2^n}, \text{ for } k = 0, 1, 2, \dots, (2^n - 1), n \geq 1.$$



## Modeling (contd.)

For every  $A \subseteq (0, 1]$ ,

- (i)  $0 \leq P(A) \leq 1$
- (ii)  $P(\emptyset) = 0, P(\Omega) = 1$
- (iii) If  $A_1, A_2, \dots \in \mathbb{F}$ , with  $A_i \cap A_j = \emptyset$ , for  $i \neq j$ , then

$$P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

[If  $A = \{x_1, x_2, \dots, x_k\} \subseteq \Omega$  where  $x_i$ 's are the outcomes of the experiment, then  $P(A) = \sum_{i=1}^k p(x_i) = \sum_{i=1}^k P(X = x_i)$ , where  $X$  is the number chosen from  $(0, 1]$ .]

- (iv)  $P[X \in [\frac{k}{2^n}, \frac{k+1}{2^n}]] = \frac{1}{2^n}$ , for  $k = 0, 1, 2, \dots, (2^n - 1), n \geq 1$ .



## Modeling (contd.)

For every  $A \subseteq (0, 1]$ ,

- (i)  $0 \leq P(A) \leq 1$
- (ii)  $P(\emptyset) = 0, P(\Omega) = 1$
- (iii) If  $A_1, A_2, \dots \in \mathbb{F}$ , with  $A_i \cap A_j = \emptyset$ , for  $i \neq j$ , then

$$P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

[If  $A = \{x_1, x_2, \dots, x_k\} \subseteq \Omega$  where  $x_i$ 's are the outcomes of the experiment, then  $P(A) = \sum_{i=1}^k p(x_i) = \sum_{i=1}^k P(X = x_i)$ , where  $X$  is the number chosen from  $(0, 1]$ .]

- (iv)  $P[X \in [\frac{k}{2^n}, \frac{k+1}{2^n}]] = \frac{1}{2^n}$ , for  $k = 0, 1, 2, \dots, (2^n - 1), n \geq 1$ .



## Modeling (contd.)

For every  $A \subseteq (0, 1]$ ,

- (i)  $0 \leq P(A) \leq 1$
- (ii)  $P(\emptyset) = 0, P(\Omega) = 1$
- (iii) If  $A_1, A_2, \dots \in \mathbb{F}$ , with  $A_i \cap A_j = \emptyset$ , for  $i \neq j$ , then

$$P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

[If  $A = \{x_1, x_2, \dots, x_k\} \subseteq \Omega$  where  $x_i$ 's are some outcomes of the experiment, then  $P(A) = \sum_{i=1}^k p(x_i) = \sum_{i=1}^k P(X = x_i)$ , where  $X$  is the number chosen from  $(0, 1]$ .]

- (iv)  $P[X \in (\frac{k}{2^n}, \frac{k+1}{2^n}]] = \frac{1}{2^n}$ , for  $k = 0, 1, 2, \dots, (2^n - 1), n \geq 1$ .



## Modeling (contd.)

**Target:** To get a function  $P : \mathbb{F} \rightarrow (0, 1]$  such that the above four conditions are satisfied. So, an important question now remains is whether such a  $P$  exists.

**Theorem:** There exists no such  $P$  satisfying all the above four conditions.



## Modeling (contd.)

**Target:** To get a function  $P : \mathbb{F} \rightarrow (0, 1]$  such that the above four conditions are satisfied. So, an important question now remains is whether such a  $P$  exists.

**Theorem:** There exists no such  $P$  satisfying all the above four conditions.



## Modeling (contd.)

**Target:** To get a function  $P : \mathbb{F} \rightarrow (0, 1]$  such that the above four conditions are satisfied. So, an important question now remains is whether such a  $P$  exists.

**Theorem:** There exists no such  $P$  satisfying all the above four conditions.



## Modeling (contd.)

Looking into the above theorem, we have to relax some requirement(s) on  $P$  in order that the existence of  $P$  is guaranteed.

But a careful observation shows that all the four conditions are equally important.

So, only one place where we can be a little less ambitious is the domain of the definition of  $P$ , i.e., in place of the whole

$\mathbb{F}$ , we may take a subset of  $\mathbb{F}$ , i.e., the class of those subsets of  $\Omega$  whose probabilities are of some practical interest.

In order to do that we define probability for events but do not allow every subset to be an event. Then a natural question arises-

Question: Which subsets to allow?



## Modeling (contd.)

Looking into the above theorem, we have to relax some requirement(s) on  $P$  in order that the existence of  $P$  is guaranteed.

But a careful observation shows that all the four conditions are equally important.

So, only one place where we can be a little less ambitious is the domain of the definition of  $P$ , i.e., in place of the whole

$\mathbb{F}$ , we may take a subset of  $\mathbb{F}$ , i.e., the class of those subsets of  $\Omega$  whose probabilities are of some practical interest.

In order to do that we define probability for events but do not allow every subset to be an event. Then a natural question arises-

Question: Which subsets to allow?





## Modeling (contd.)

Looking into the above theorem, we have to relax some requirement(s) on  $P$  in order that the existence of  $P$  is guaranteed.

But a careful observation shows that all the four conditions are equally important.

So, only one place where we can be a little less ambitious is the domain of the definition of  $P$ , i.e., in place of the whole  $\mathbb{F}$ , we may take a subset of  $\mathbb{F}$ , i.e., the class of those subsets of  $\Omega$  whose probabilities are of some practical interest.

In order to do that we define probability for events but do not allow every subset to be an event. Then a natural question arises-

Question: Which subsets to allow?



## Modeling (contd.)

Looking into the above theorem, we have to relax some requirement(s) on  $P$  in order that the existence of  $P$  is guaranteed.

But a careful observation shows that all the four conditions are equally important.

So, only one place where we can be a little less ambitious is the domain of the definition of  $P$ , i.e., in place of the whole  $\mathbb{F}$ , we may take a subset of  $\mathbb{F}$ , i.e., the class of those subsets of  $\Omega$  whose probabilities are of some practical interest.

In order to do that we define probability for events but do not allow every subset to be an event. Then a natural question arises-

Question: Which subsets to allow?



## Modeling (contd.)

Looking into the above theorem, we have to relax some requirement(s) on  $P$  in order that the existence of  $P$  is guaranteed.

But a careful observation shows that all the four conditions are equally important.

So, only one place where we can be a little less ambitious is the domain of the definition of  $P$ , i.e., in place of the whole  $\mathbb{F}$ , we may take a subset of  $\mathbb{F}$ , i.e., the class of those subsets of  $\Omega$  whose probabilities are of some practical interest.

In order to do that we define probability for events but do not allow every subset to be an event. Then a natural question arises-

Question: Which subsets to allow?



## Modeling (contd.)

Looking into the above theorem, we have to relax some requirement(s) on  $P$  in order that the existence of  $P$  is guaranteed.

But a careful observation shows that all the four conditions are equally important.

So, only one place where we can be a little less ambitious is the domain of the definition of  $P$ , i.e., in place of the whole  $\mathbb{F}$ , we may take a subset of  $\mathbb{F}$ , i.e., the class of those subsets of  $\Omega$  whose probabilities are of some practical interest.

In order to do that we define probability for events but do not allow every subset to be an event. Then a natural question arises-

Question: Which subsets to allow?



## Modeling (contd.)

- (a)  $\emptyset$  and  $\Omega$  should be allowed.
- (b) In this present experiment, the probability of the outcome lying in a specific interval is important. So, all intervals are to be allowed.
- (c) If any  $A \subseteq \Omega$  is allowed, then  $A^c$  should be allowed. This is because if the occurrence of something is important, then the non-occurrence of the same is also equally important.
- (d) If the individual occurrence of  $A_1, A_2, \dots$  is important, then the occurrence of at least one of  $A_1, A_2, \dots$  is also important. So, if  $A_1, A_2, \dots$  are allowed then  $\cup_{i=1}^{\infty} A_i$  should also be allowed.



## Modeling (contd.)

- (a)  $\emptyset$  and  $\Omega$  should be allowed.
- (b) In this present experiment, the probability of the outcome lying in a specific interval is important. So, all intervals are to be allowed.
- (c) If any  $A \subseteq \Omega$  is allowed, then  $A^c$  should be allowed. This is because if the occurrence of something is important, then the non-occurrence of the same is also equally important.
- (d) If the individual occurrence of  $A_1, A_2, \dots$  is important, then the occurrence of at least one of  $A_1, A_2, \dots$  is also important. So, if  $A_1, A_2, \dots$  are allowed then  $\cup_{i=1}^{\infty} A_i$  should also be allowed.



## Modeling (contd.)

- (a)  $\emptyset$  and  $\Omega$  should be allowed.
- (b) In this present experiment, the probability of the outcome lying in a specific interval is important. So, all intervals are to be allowed.
- (c) If any  $A \subseteq \Omega$  is allowed, then  $A^c$  should be allowed. This is because if the occurrence of something is important, then the non-occurrence of the same is also equally important.
- (d) If the individual occurrence of  $A_1, A_2, \dots$  is important, then the occurrence of at least one of  $A_1, A_2, \dots$  is also important. So, if  $A_1, A_2, \dots$  are allowed then  $\cup_{i=1}^{\infty} A_i$  should also be allowed.



## Modeling (contd.)

- (a)  $\emptyset$  and  $\Omega$  should be allowed.
- (b) In this present experiment, the probability of the outcome lying in a specific interval is important. So, all intervals are to be allowed.
- (c) If any  $A \subseteq \Omega$  is allowed, then  $A^c$  should be allowed. This is because if the occurrence of something is important, then the non-occurrence of the same is also equally important.
- (d) If the individual occurrence of  $A_1, A_2, \dots$  is important, then the occurrence of at least one of  $A_1, A_2, \dots$  is also important. So, if  $A_1, A_2, \dots$  are allowed then  $\cup_{i=1}^{\infty} A_i$  should also be allowed.



## Modeling (contd.)

Let  $\mathcal{B}$  be the smallest class of subsets of  $\Omega$  satisfying all the above conditions. Then  $\mathcal{B}$  is formally defined as the smallest class of subsets of  $\Omega$  such that

$$(i) \emptyset, \Omega \in \mathcal{B}$$

$$(ii) A \in \mathcal{B} \Rightarrow A^c \in \mathcal{B}$$

$$(iii) A_1, A_2, \dots \in \mathcal{B} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{B}$$

$$(iv) \binom{k}{2^n}, \binom{k+1}{2^n} \in \mathcal{B} \quad \forall k = 0, 1, 2, \dots, (2^n - 1), n \geq 1$$

[The last inequality is required because otherwise,  $\{\emptyset, \Omega\}$  will be the smallest class of subsets of  $\Omega$  satisfying (i)-(iii). But the class  $\{\emptyset, \Omega\}$  is of no practical use.]



## Modeling (contd.)

Let  $\mathcal{B}$  be the smallest class of subsets of  $\Omega$  satisfying all the above conditions. Then  $\mathcal{B}$  is formally defined as the smallest class of subsets of  $\Omega$  such that

$$(i) \emptyset, \Omega \in \mathcal{B}$$

$$(ii) A \in \mathcal{B} \Rightarrow A^c \in \mathcal{B}$$

$$(iii) A_1, A_2, \dots \in \mathcal{B} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{B}$$

$$(iv) \binom{k}{2^n}, \binom{k+1}{2^n} \in \mathcal{B} \quad \forall k = 0, 1, 2, \dots, (2^n - 1), n \geq 1$$

[The last inequality is required because otherwise,  $\{\emptyset, \Omega\}$  will be the smallest class of subsets of  $\Omega$  satisfying (i)-(iii). But the class  $\{\emptyset, \Omega\}$  is of no practical use.]



## Modeling (contd.)

Let  $\mathcal{B}$  be the smallest class of subsets of  $\Omega$  satisfying all the above conditions. Then  $\mathcal{B}$  is formally defined as the smallest class of subsets of  $\Omega$  such that

- (i)  $\emptyset, \Omega \in \mathcal{B}$
- (ii)  $A \in \mathcal{B} \Rightarrow A^c \in \mathcal{B}$
- (iii)  $A_1, A_2, \dots \in \mathcal{B} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{B}$
- (iv)  $\binom{k}{2^n}, \binom{k+1}{2^n} \in \mathcal{B} \forall k = 0, 1, 2, \dots, (2^n - 1), n \geq 1$

[The last inequality is required because otherwise,  $\{\emptyset, \Omega\}$  will be the smallest class of subsets of  $\Omega$  satisfying (i)-(iii). But the class  $\{\emptyset, \Omega\}$  is of no practical use.]



## Modeling (contd.)

Let  $\mathcal{B}$  be the smallest class of subsets of  $\Omega$  satisfying all the above conditions. Then  $\mathcal{B}$  is formally defined as the smallest class of subsets of  $\Omega$  such that

- (i)  $\emptyset, \Omega \in \mathcal{B}$
- (ii)  $A \in \mathcal{B} \Rightarrow A^c \in \mathcal{B}$
- (iii)  $A_1, A_2, \dots \in \mathcal{B} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{B}$
- (iv)  $\binom{k}{2^n}, \binom{k+1}{2^n} \in \mathcal{B} \forall k = 0, 1, 2, \dots, (2^n - 1), n \geq 1$

[The last inequality is required because otherwise,  $\{\emptyset, \Omega\}$  will be the smallest class of subsets of  $\Omega$  satisfying (i)-(iii). But the class  $\{\emptyset, \Omega\}$  is of no practical use.]



## Modeling (contd.)

Let  $\mathcal{B}$  be the smallest class of subsets of  $\Omega$  satisfying all the above conditions. Then  $\mathcal{B}$  is formally defined as the smallest class of subsets of  $\Omega$  such that

- (i)  $\emptyset, \Omega \in \mathcal{B}$
- (ii)  $A \in \mathcal{B} \Rightarrow A^c \in \mathcal{B}$
- (iii)  $A_1, A_2, \dots \in \mathcal{B} \Rightarrow \cup_{i=1}^{\infty} A_i \in \mathcal{B}$
- (iv)  $(\frac{k}{2^n}, \frac{k+1}{2^n}] \in \mathcal{B} \forall k = 0, 1, 2, \dots, (2^n - 1), n \geq 1$

[The last inequality is required because otherwise,  $\{\emptyset, \Omega\}$  will be the smallest class of subsets of  $\Omega$  satisfying (i)-(iii). But the class  $\{\emptyset, \Omega\}$  is of no practical use.]



## Borel Field

**Remark:** It can be shown that  $\mathcal{B}$  is a proper subset of  $\mathbb{F}$  and  $\mathcal{B}$  meets all the requirements, *i.e.*, we can always define a  $P$  on elements of  $\mathcal{B}$  such that all the above four conditions are satisfied.  $\mathcal{B}$  is known as Borel field.



## One Interesting Problem

A Sultan thought of increasing the number of women in his country, as compared to the number of men, so that the men could have larger harems. To accomplish this, he proposed the following law:

*“As soon as a mother gave birth to her first son, she would be forbidden to have any more child”*

Sultan’s argument was that some families would have several girls and only one boy, but no family could have more than one boy. It should not be long until the females would greatly outnumber the males.

Question: What is your view on this?



## One Interesting Problem

A Sultan thought of increasing the number of women in his country, as compared to the number of men, so that the men could have larger harems. To accomplish this, he proposed the following law:

*“As soon as a mother gave birth to her first son, she would be forbidden to have any more child”*

Sultan’s argument was that some families would have several girls and only one boy, but no family could have more than one boy. It should not be long until the females would greatly outnumber the males.

Question: What is your view on this?







Question: What is your view on this?



## One Interesting Problem

A Sultan thought of increasing the number of women in his country, as compared to the number of men, so that the men could have larger harems. To accomplish this, he proposed the following law:

*“As soon as a mother gave birth to her first son, she would be forbidden to have any more child”*

Sultan’s argument was that some families would have several girls and only one boy, but no family could have more than one boy. It should not be long until the females would greatly outnumber the males.

Question: What is your view on this?



Imagine that on the same day each of 8 mothers gives birth to their first child. We would expect half to be boys and half to be girls. Because of the law, the 4 mothers who are having girls would be allowed to have second child. It is expected that the 2 mothers will have sons and 2 will have daughters. The two who have daughters would have another child and it is expected to have one boy and one girl. If we add up we see that the number of boys and girls are same (seven each).



THANK YOU



**BCKV Workshop**  
**Bikas K Sinha**  
**[bikassinha1946@gmail.com]**  
**Retired Faculty, ISI, Kolkata**

**"Statistical Assessment of Agreement in  
Treatment Effects Comparisons"**

**National Workshop-cum-Training Programme  
On Statistical Tools for Research Data Analysis  
[Series II]**

**BCKV, May 29 – June 9, 2017**

## *Quotes of the Day.....*

- *There is more to see (on the ground)*
- *than*
- *what we see (from the sky).*
- **\*\*\*\*\***
- *A Man's Feet Should Be Planted*
- *In His Home Country .....*
- **BUT**
- *His Eyes Should Survey The World !!!*

## Unplanned Balanced Experiment....

- There are 'k' pairs of experimental plots ....same size and shape within each pair.....homogeneous in fertility gradient of the 2 plots in each pair...
- In each pair of plots...randomly apply Tr A to one plot and Tr B to the other plot....
- Generate 'yield' data  $y(A)$  and  $y(B)$  in due time...
- k pairs..... $\{y_i(A), y_i(B)\}; i=1, 2, \dots, k$ .
- Treatments A and B are '*in agreement*' provided  $|y_i(A) - y_i(B)| < \text{a tolerable qty} = \eta$ , say for *substantial* proportion of the k pairs  $i=1, 2, \dots, k$

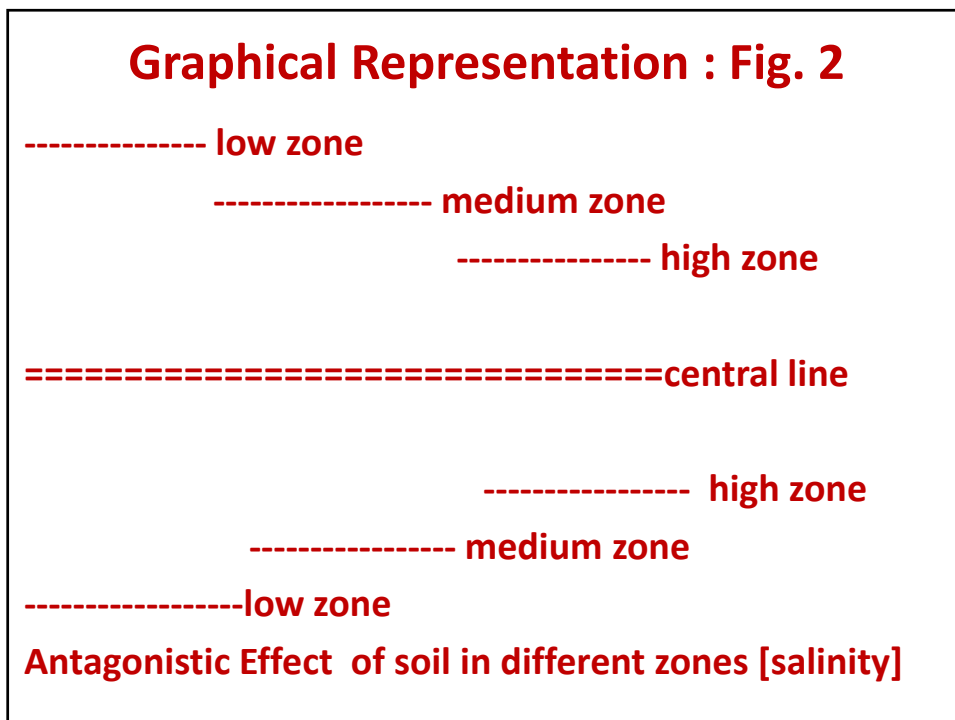
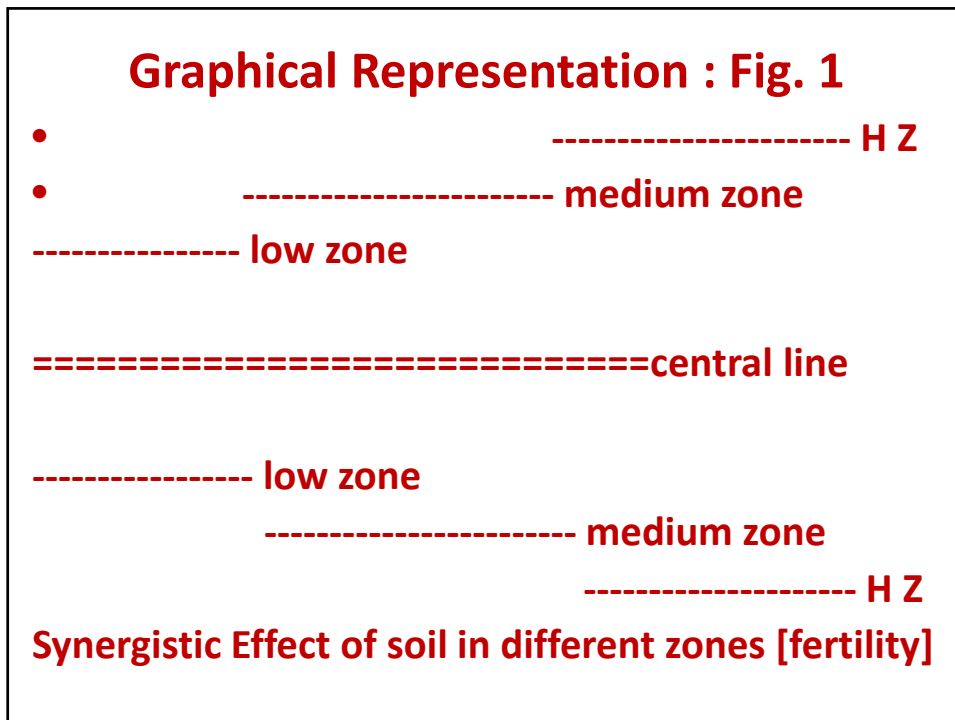
## *Flexibility. . . .*

Shapes / Sizes / Land Type.....similarity within each pair...that's it.....could be very much different from one place to another.....zonal effects....negligible or serious.....

- Zonal classification.....the experimental pairs of plots may be grouped into several zones....  
wrt fertility : low / medium / high zones  
wrt salinity : low / medium / high zones....  
or, any other plausible criterion .....

For low fertility [high salinity] zone :  $Y_i(A) - Y_i(B)$  is likely to be relatively smaller/higher than that for medium or high fertility zones.....

So....agreement has to be understood in the proper context.....



## Assessment of Agreement based on Individual Performance of Tr. pairs

Recall : Treatments A and B are 'in agreement' provided  $Z_i = |y_i(A) - y_i(B)| < \eta$ , a specified qty for substantial proportion of the k pairs,  $i=1, 2, \dots, k$ .

*This is non-parametric or distribution-free approach.*

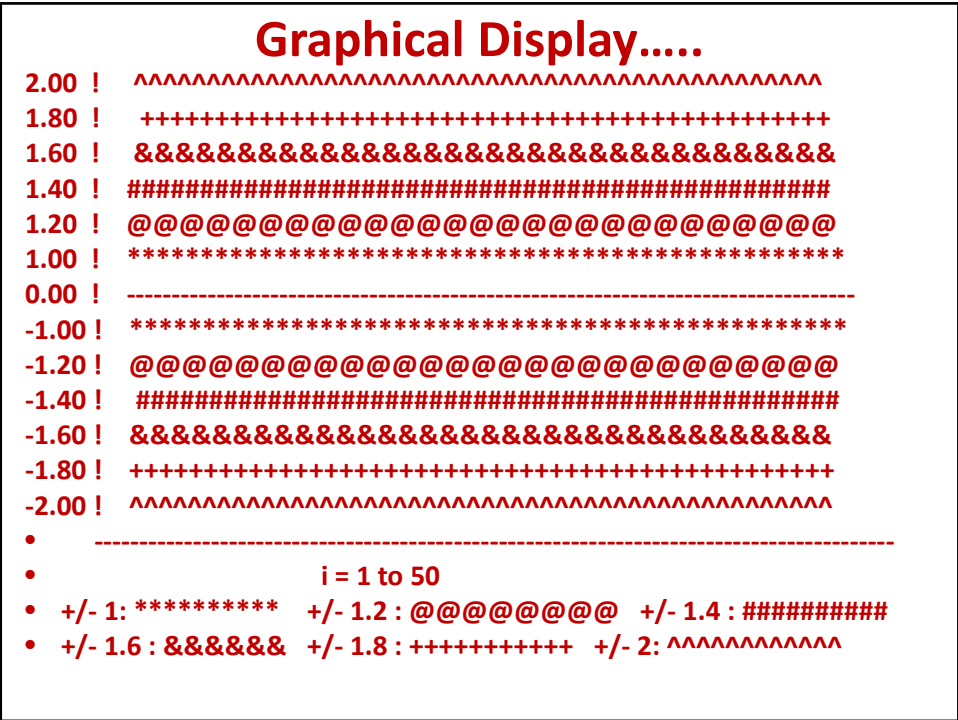
By changing the value of  $\eta$ , we can find out the extent of agreement between the 2 treatments.

We have incorporated an example to this effect.

### Illustrative Example – 1

Single Zone : Data on  $Z_i = Y_i(A) - Y_i(B)$ ,  $i=1$  to 50

Sl No.	Z	Sl. No.	Z	Sl. No.	Z	Sl No	Z	Sl No	Z
1	-1.3	11	2.9	21	2.7	31	2.4	41	2.3
2	-1.8	12	-1.5	22	1.4	32	-0.8	42	-1.9
3	0.7	13	-2.1	23	-2.8	33	-2.4	43	-2.2
4	3.4	14	0.8	24	-1.8	34	1.3	44	-1.7
5	1.9	15	2.5	25	1.2	35	-2.7	45	2.4
6	-0.3	16	1.7	26	-2.5	36	0.2	46	0.3
7	-2.2	17	2.8	27	1.4	37	2.5	47	-2.1
8	3.7	18	-3.2	28	2.7	38	-1.9	48	1.4
9	1.7	19	-2.4	29	-0.8	39	-2.8	49	0.7
10	2.1	20	0.4	30	-1.5	40	1.5	50	-1.8



### Non-parametric study of agreement

$\eta$	:	1.0	1.2	1.4	1.6	1.8	2.0	3.0
% coverage	:	18%	20%	30%	36%	46%	52%	92%

Decide on a criterion and conclude accordingly.

**Remark 1:** Very simple logic, computation and interpretation.

**Remark 2 :** This simple-minded computation does NOT reflect any dependence on TRUE TREATMENT EFFECTS' DIFFERENCE i.e., on  $\theta = \tau(A) - \tau(B)$ .

**Q.** What if we wish to use  $Z_{bar}$  in the study of agreement ?

Theory suggests : In that case we should use  $\eta/\sqrt{n}$  instead of  $\eta$ , 'n' being the sample size for  $Z_{bar}$ .

$|y_i(A) - y_i(B) = Z_i| < \eta$  to be changed to  $[- \eta^* < Z_{bar} < \eta^{**}]$   
 where  
 $\eta^{**} = \eta/\sqrt{n} + \theta(1 - 1/\sqrt{n})$  and  $\eta^* = \eta/\sqrt{n} - \theta(1 - 1/\sqrt{n})$ .

## **Assessment of Agreement based on Average Performance of Tr. Pairs**

**Purpose.....study dependence on the true treatment effects difference  $\theta = \tau(A) - \tau(B)$**

**Also zonal effects.....**

**Contemplate on a statistical model .....incorporate all sources of variation.....**

**Examine the coverage probability wrt the statistical model and then decide on extent of agreement.....**

## **Model Assumptions....**

**Model :  $Y_i(A) - Y_i(B)$  varies from pair to pair.....**

**Depends on  $\theta = \tau(A) - \tau(B)$ ...difference between true effects of Tr A and Tr B...in general these effects will interact with the zonal soil fertility properties....**

**$Y_i(A) = \mu + \tau(A) + s_iA + e_i(A)$**

**$Y_i(B) = \mu + \tau(B) + s_iB + e_i(B)$**

**$s_iA(B)$ : manifestation of excess soil fertility via Tr A (B)**

**$e$ 's are experimental errors.....uncorrelated with  $s$ 's**

**Assumption :  $|\tau(A) - \tau(B)|$  dominates both  $|\tau(A) - \tau(B)|$  and  $|e_i(A) - e_i(B)|$  for each  $i=1, 2, \dots, k$ .**



### Model Assumptions...contd.

- $si(A)-si(B) \sim N(0, \sigma^2(s)), ei(A)-ei(B) \sim N(0, \sigma^2(e))$
- $Z_i = Y_i(A) - Y_i(B) \sim N(\tau(A) - \tau(B), \sigma^2(z))$  where  $\sigma^2(z) = \sigma^2(s) + \sigma^2(e)$ .
- Recall :  $\theta = \tau(A) - \tau(B)$ ...
- Since  $Z_i$ 's have common unknown mean,  $\sigma^2(z)$  can easily be estimated by the usual sample variance of z-values.....
- $\sigma^2(z) \wedge = s^2(z) = \sum (Z_i - \bar{Z})^2 / (n-1)$ , n being the sample size [2n being total # of paired plots].

### Assessment of Agreement based on Average Performance of Tr. Pairs

- Need to evaluate  $P = \Pr[-\eta^* < \bar{Z} < \eta^{**}]$  where
  - $\eta^{**} = \eta / \sqrt{n} + \theta (1 - 1 / \sqrt{n})$  and
  - $\eta^* = \eta / \sqrt{n} - \theta (1 - 1 / \sqrt{n})$ .
  - $P = \Pr[-(\eta - \theta) / \sqrt{n} < \bar{Z} - \theta < (\eta - \theta) / \sqrt{n}]$
  - $\cong \Phi\{(\eta - \theta) / \sigma(z)\} - \Phi\{-(\eta - \theta) / \sigma(z)\}$
- $\Phi$  being the standard normal cdf.

It is further expected that  $\bar{Z}$  is close to  $\theta = \tau(A) - \tau(B)$ . Therefore, in the above, we may use  $\bar{z}$  [sample mean of z-values] as an approximation for  $\theta = \tau(A) - \tau(B)$  and  $s(z)$  as an approx. for  $\sigma(z)$ .

### Evaluation of $\Pr[-\eta^* < \bar{Z} < \eta^{**}]$

$$P \cong \Phi\{(\eta - \bar{z}) / s(z)\} - \Phi\{-(\eta - \bar{z}) / s(z)\}$$

Based on the above computation, we may determine the extent of agreement for given  $\eta$ .

Now we will consider an illustrative example by taking a global unclassified [homogeneous] sample of 50 pairs of plots and by making a choice of  $\eta$  as 1.0, 1.2, 1.4, 1.6, 1.8 and 2.0.

Same Illustrative Example 1 is used here.

### Illustrative Example – 2

Single Zone : Data on  $Z_i = Y_i(A) - Y_i(B)$ ,  $i=1$  to 50

Sl No	Z	Sl. No	Z	Sl. No	Z	Sl No	Z	Sl No	Z
1	-1.3	11	2.9	21	2.7	31	2.4	41	2.3
2	-1.8	12	-1.5	22	1.4	32	-0.8	42	-1.9
3	0.7	13	-2.1	23	-2.8	33	-2.4	43	-2.2
4	3.4	14	0.8	24	-1.8	34	1.3	44	-1.7
5	1.9	15	2.5	25	1.2	35	-2.7	45	2.4
6	-0.3	16	1.7	26	-2.5	36	0.2	46	0.3
7	-2.2	17	2.8	27	1.4	37	2.5	47	-2.1
8	3.7	18	-3.2	28	2.7	38	-1.9	48	1.4
9	1.7	19	-2.4	29	-0.8	39	-2.8	49	0.7
10	2.1	20	0.4	30	-1.5	40	1.5	50	-1.8

## Computations and Agreement Assessment for a Treatment Pair

$$\bar{z} = 0.9 \text{ and } s(z) = \sqrt{3.4435} = 1.8557$$

$$\Phi\{(\eta - \bar{z})/s(z)\} - \Phi\{(-\eta - \bar{z})/s(z)\}$$

$$= \Phi\{(\eta - 0.9) \times 0.5389\} - \Phi\{(-\eta - 0.9) \times 0.5389\}$$

$$\eta = 1 : \text{Prob.} = \Phi(0.05389) - \Phi(-1.0239) = 37\%$$

agreement according to  $\eta$  -criterion

$$\eta = 1.2 : \text{Prob.} = \Phi(0.1617) - \Phi(-1.1317) = 44\%$$

$$\eta = 1.4 : \text{Prob.} = \Phi(0.2694) - \Phi(-1.2395) = 50\%$$

$$\eta = 1.6 : \text{Prob.} = \Phi(0.3772) - \Phi(-1.3472) = 56\%$$

$$\eta = 1.8 : \text{Prob.} = \Phi(0.4850) - \Phi(-1.4550) = 61\%$$

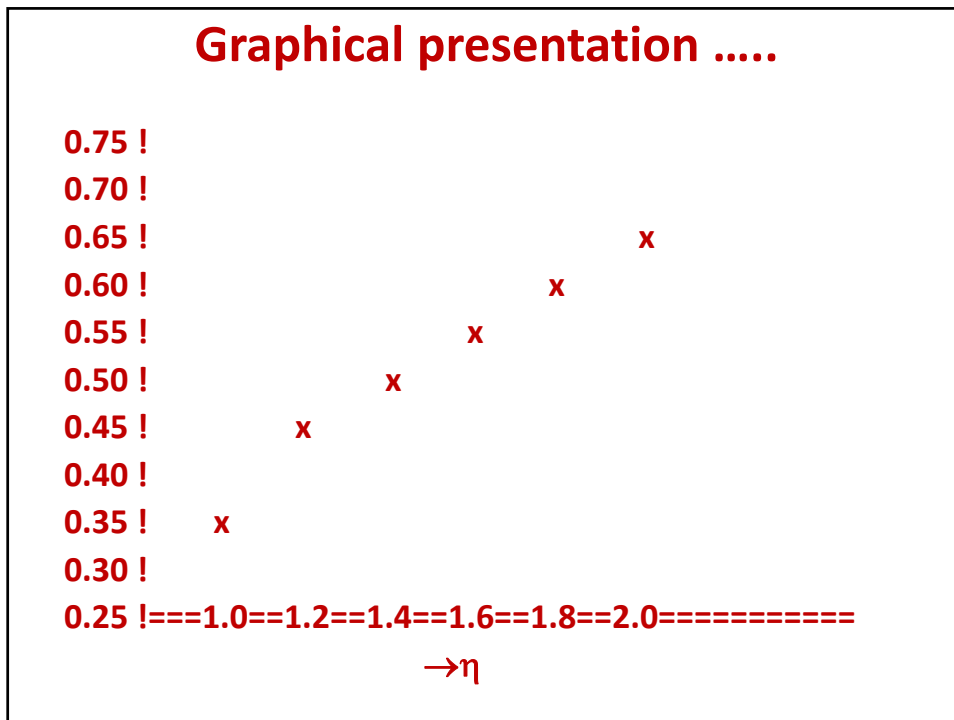
$$\eta = 2 : \text{Prob.} = \Phi(0.5927) - \Phi(-1.5628) = 65\%$$

$$\eta^{**} = \eta / \sqrt{n} + \theta (1 - 1 / \sqrt{n}) = 0.1414 \eta + 0.7727$$

$$\eta^* = \eta / \sqrt{n} - \theta (1 - 1 / \sqrt{n}) = 0.1414\eta - 0.7727$$

$$[\theta = 0.9]$$

$\eta$ :	1.0	1.2	1.4
$\eta^{**}$ :	0.9141	0.9423	0.9711
$\eta^*$ :	-0.6313	-0.6030	-0.5743
$\eta$ :	1.6	1.8	2.0
$\eta^{**}$ :	0.9989	1.0272	1.0555
$\eta^*$ :	-0.5465	-0.5182	-0.4899



### Illustrative Example – 3

**Zone - I : Data on  $Z_i = Y_i(A) - Y_i(B)$ ,  $i=1$  to 20**  
**Zone – II : Data on  $Z_i = Y_i(A) - Y_i(B)$ ,  $i=21$  to 50**

→ Zone – I ←		→ Zone – II ←		
Sl No.	Z	Sl. No.	Z	Sl No Z
1	-1.3	11	2.9	1 2.7
2	-1.8	12	-1.5	11 2.4
3	0.7	13	-2.1	21 2.3
4	3.4	14	0.8	2 1.4
5	1.9	15	2.5	12 -0.8
6	-0.3	16	1.7	22 -1.9
7	-2.2	17	2.8	3 -2.8
8	3.7	18	-3.2	13 -2.4
9	1.7	19	-2.4	3 1.4
10	2.1	20	0.4	4 -1.8
				14 1.3
				24 -1.7
				5 1.2
				15 -2.7
				25 2.4
				6 -2.5
				16 0.2
				26 0.3
				7 1.4
				17 2.5
				27 -2.1
				8 2.7
				18 -1.9
				28 1.4
				9 -0.8
				19 -2.8
				29 0.7
				10 -1.5
				20 1.5
				30 -1.8

## Non-parametric study of agreement

Antagonistic effect of soil.....

Use  $\eta(I) = 2.0 > 1.5 = \eta(II)$ . Vide Fig. 2.

Proportion [  $|y_i(A) - y_i(B)| < \eta(I)$  in Zone I

and/or  $|y_i(A) - y_i(B)| < \eta(II)$  in Zone II]

=22 %

## Zonal Analysis

- Zone I :  $n(I) = 20$ ,  $\bar{z} = 0.49$ ,  $s^2(z) = 4.7962$
  - Zone II :  $n(II) = 30$ ,  $\bar{z} = -0.1767$ ,  $s^2(z) = 3.8943$
  - Antagonistic effect of soil.....use  $\eta(I) = 2.0 > 1.5 = \eta(II)$ . Vide Fig. 2.
  - Need to evaluate
  - $\Pr[ -\eta^*(I) < \bar{Z}(I) < \eta^{**}(I) \ \& \ -\eta^*(II) < \bar{Z}(II) < \eta^{**}(II) ]$
  - where
  - $\eta^{**}(II) = \eta(II) / \sqrt{n(II)} + \theta^{\wedge} (1 - 1 / \sqrt{n(II)})$   
 $\eta^*(II) = \eta(II) / \sqrt{n(II)} - \theta^{\wedge} (1 - 1 / \sqrt{n(II)})$
  - $\eta^{**}(I) = \eta(I) / \sqrt{n(I)} + \theta^{\wedge} (1 - 1 / \sqrt{n(I)})$   
 $\eta^*(I) = \eta(I) / \sqrt{n(I)} - \theta^{\wedge} (1 - 1 / \sqrt{n(I)})$
- Need evaluation of  $\theta^{\wedge} =$  estimate of  $\tau(A) - \tau(B)$ .

- **Recall :  $Z_i = Y_i(A) - Y_i(B) \sim N(\theta, \sigma^2(z))$**
- **where  $\sigma^2(z) = \sigma^2(s) + \sigma^2(e)$ .**
- **For Zone I :  $Zbar(I) \sim N(\theta, \sigma^2(z(I)))$**
- **For Zone II :  $Zbar(II) \sim N(\theta, \sigma^2(z(II)))$**
- **As before,  $zbar(I)$  provides an estimate for  $\theta = \tau(A) - \tau(B)$  and in the same spirit,  $zbar(II)$  also provides an *independent* estimate for  $\theta = \tau(A) - \tau(B)$ . Therefore, we may combine the two by using standard weighting method.**
- **This leads to : Combined estimate of  $\theta$  given by**  

$$\theta^{\wedge}_c = [n(I) zbar(I) / s^2(z(I)) + n(II) zbar(II) / s^2(z(II))]$$

$$/ [n(I) / s^2(z(I)) + n(II) / s^2(z(II))]$$

**Combined Estimate for  $\theta = \tau(A) - \tau(B)$ .**

$$\theta^{\wedge}_c = [20 \times 0.49 / 4.7962 - 30 \times 0.1767 / 3.8943]$$

$$/ [20 / 4.7962 + 30 / 3.8943]$$

$$= 0.0574$$

**Further,  $\eta(I) = 2.0 > 1.5 = \eta(II)$ .**

**Therefore,**

- $\eta^{**}(II) = \eta(II) / \sqrt{n(II)} + \theta^{\wedge}_c (1 - 1 / \sqrt{n(II)}) = 0.3208$
- $\eta^{*}(II) = \eta(II) / \sqrt{n(II)} - \theta^{\wedge}_c (1 - 1 / \sqrt{n(II)}) = 0.2270$

- $\eta^{**}(I) = \eta(I) / \sqrt{n(I)} + \theta^{\wedge}_c (1 - 1 / \sqrt{n(I)}) = 0.4917$
- $\eta^{*}(I) = \eta(I) / \sqrt{n(I)} - \theta^{\wedge}_c (1 - 1 / \sqrt{n(I)}) = 0.4027$

- **We evaluate**

- **$Pr[0.4027 < Zbar(I) < 0.4917];$**

- **$0.2270 < Zbar(II) < 0.3208]$**

### Coverage Probability in 2 Zones

- Zone I :  $n(I) = 20$ ,  $\bar{z} = 0.49$ ,  $s^2(z) = 4.7962$

$$\begin{aligned} & \Phi\{(\eta(I) - \bar{z}(I))/s(z)(I)\} - \Phi\{(-\eta(I) - \bar{z}(I))/s(z)(I)\} \\ &= \Phi\{(2.0 - 0.49) \times 0.4566\} - \Phi\{(-2.0 - 0.49) \times 0.4566\} \\ &= \Phi(.6895) - \Phi(-1.1369) = 0.7549 - 0.1271 = 0.6278 \end{aligned}$$

- Zone II :  $n(II) = 30$ ,  $\bar{z} = -0.1767$ ,  $s^2(z) = 3.8943$

$$\begin{aligned} & \Phi\{(\eta(II) - \bar{z}(II))/s(z)(II)\} - \\ & \Phi\{(-\eta(II) - \bar{z}(II))/s(z)(II)\} \\ &= \Phi\{(1.5 + 0.1767) \times 0.5067\} - \Phi\{(-1.5 + 0.1767) \times \\ & 0.5067\} = \Phi(.8469) - \Phi(-.6705) \\ &= 0.8023 - 0.2514 = 0.5509 \end{aligned}$$

Over all....coverage prob. =  $0.6278 \times 0.5509 = 35\%$

### Unbalanced Data

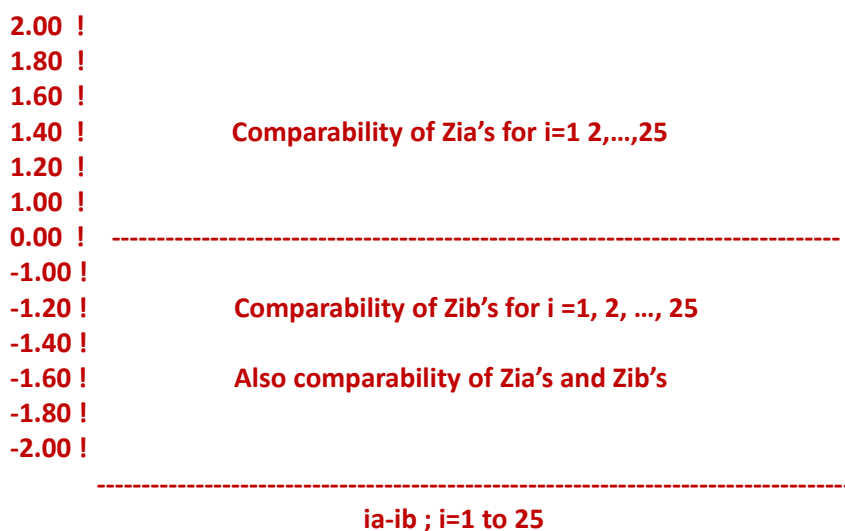
- There are 'k' triplets of experimental plots ....same size and shape within each triplet ....homogeneous in fertility gradient of the 3 plots in each triplet...
- In each triplet of plots...randomly apply Tr A to 2 plots and Tr B to the 3rd plot or otherwise....1 vs 2
- Generate 'yield' data  $y(A)$  and  $y(B)$  in due time...
- $\{\bar{y}_i(A), \bar{y}_i(B)\}$  OR  $\{y_i(A), y_i(B)\}$ ;  $i=1, 2, \dots, k$ .
- Treatments A and B are 'in agreement' provided  $|\bar{y}_i(A) - \bar{y}_i(B)| < \eta$  &  $|y_i(A) - y_i(B)| < \eta$  for *substantial* proportion of the k triplets;  $i=1, 2, \dots, k$ .

### Illustrative Example – 4

Single Zone : Data on  $Z_{ia} = \bar{Y}_i(A) - \bar{Y}_i(B)$  and  $Z_{ib} = \bar{Y}_i(A) - \bar{Y}_i(B)$ ; (ia, ib) for  $i = 1$  to 25

Sl No.	Z	Sl. No.	Z	Sl. No.	Z	Sl No	Z	Sl No	Z
1a	-1.3	6a	2.9	11a	2.7	16a	2.4	21a	2.3
1b	-1.8	6b	-1.5	11b	1.4	16b	-0.8	21b	-1.9
2a	0.7	7a	-2.1	12a	-2.8	17a	-2.4	22a	-2.2
2b	3.4	7b	0.8	12b	-1.8	17b	1.3	22b	-1.7
3a	1.9	8a	2.5	13a	1.2	18a	-2.7	23a	2.4
3b	-0.3	8b	1.7	13b	-2.5	18b	0.2	23b	0.3
4a	-2.2	9a	2.8	14a	1.4	19a	2.5	24a	-2.1
4b	3.7	9b	-3.2	14b	2.7	19b	-1.9	24b	1.4
5a	1.7	10a	-2.4	15a	-0.8	20a	-2.8	25a	0.7
5b	2.1	10b	0.4	15b	-1.5	20b	1.5	25b	-1.8

### Graphical Display.....





## Balanced Plan for Three Treatments

- There are 'k' triplets of experimental plots ....same size and shape within each triplet ....homogeneous in fertility gradient of the 3 plots in each triplet...
- In each triplet of plots...randomly apply Tr A, B and C each to one plot.....
- Generate 'yield' data  $y(A)$ ,  $y(B)$ ,  $Y(C)$  in due time...
- $\{y_i(A), y_i(B), y_i(C)\}; i=1, 2, \dots, k$ .
- Treatments A, B and C are '*in agreement*' provided  $|y_i(A) - y_i(B)| < \eta$ ,  $|y_i(A) - y_i(C)| < \eta$ ,  $|y_i(B) - y_i(C)| < \eta$ , for *substantial* proportion of the k triplets;  $i=1, 2, \dots, k$ .
- Analogous theory needs to be developed.

## SOIL HETEROGENEITY AND ITS TREATMENT – OPTIMUM PLOT SIZE DETERMINATION

Prof. Satyabrata Pal

### Introducing the Problem

The determination of optimum plot size is an essential activity in agricultural field experimentation, specially, in case of agronomic and horticultural field experiments. As soil heterogeneity is the governing factor in case of data emanating from field experiments, it is needed to determine the plot size (optimum in the sense that such plots identify the actual heterogeneity structure of the field), which ultimately lead to minimization of error variance by identifying the nature of embedded correlation in the field data. Smith was the prime scientist who considered the situation a hundred years ago.

**1. Smith (1938) in his remarkable paper suggests an empirical model relating the variability to plot size and shape.**

2. The fact in regard to the presence of association between adjacent plots is recognized by Smith (1938) and he includes a factor 'b' (called soil heterogeneity) in his proposed model.

3. His model has been the basis for analysis of uniformity studies for many decades and even till now. In view of the above facts, to understand in detail the nature of variation, association, similarity and dissimilarity among data points (yields) in different directions and their relation in the determination of optimum plot size and shape, it is important to discuss and comprehend the fathomless potentiality of his method in detail.

4. In what follows, the work of Smith in detail is discussed in both isotropic and anisotropic cases. The methods are explained on real-life data sets. T

5. The role of variogram in case of determination of optimum plot size and robust optimum plot size is also discussed.

### 6. Smith's Model - A Milestone – Theoretical Development

#### 7. Optimum plot size determination in case of isotropic field

8. According to Smith (1938) the regression of yield-variability on plot size is more easily interpreted (and understood) when the observations on variances of yields with respect to different plot sizes are plotted on double logarithmic paper – in fact, according to him a linear relationship is obtained. The above notion can be well described by an equation of the form,  $\log(V_x) = \log(V_1) - b \cdot \log(x)$ , where  $V_x$  is the variance of yields per unit area for plots of area  $x$  units and  $V_1$  is the variance of yields from plots of area one unit (unit area may be, 1m x 1m),  $b$  is a characteristic of the soil, and a measure of the correlation among contiguous units. The above equation can be written in the form,  $V_x = \frac{V_1}{x^b}$ .

9. If  $b = 1$ ,  $V_x = \frac{V_1}{x}$ , and the units making up the plot of  $x$  units are not correlated at all. Alternatively, if  $x$  units are perfectly correlated, then the value of  $b$  becomes zero ( $b = 0$ ), so that the use of larger plot sizes does not bring in any gain effectively. Under actual field conditions, the value of  $b$  lies between 0 and 1, and thus the implication of the above relationship points to the fact that the use of larger plots gives more information with the same number of plots.

10. It is important to note that since the variances are calculated from varying numbers of plots in Smith's formula, it may be appropriate to find out the regression equation (of  $V_x$  on  $x$ ) by attaching weight (equal to the inverse of variance) with respect to each point, equivalently, weights are taken as proportional to the reciprocal of the variances, and a first approximation of the variance of the logarithm of a variance satisfies the equation,

$\{\partial(\log_e s^2)\}^2 = \left\{ \frac{2 \cdot s \cdot \partial s}{s^2} \right\}^2 = \frac{2}{n}$ , where  $n$  is the number of degrees of freedom on which the estimate of variance is based, and s.d. of  $s$  (to a first approximation) is:  $\Delta(s) = (s/\sqrt{2n})$ .

11. Following Smith's formula if we consider a fixed area, the impact of Smith's formula can be described in the following table.

**12. Relationship between Size of the Unit and Variance of Treatment Means per Unit Area**

Size of the Unit	Number of Replicates	Variance of Treatment Means per Unit Area
$x$	$r$	$\frac{V_x}{r}$
$\frac{x}{2}$	$2r$	$\frac{V_x}{2r}$
$\frac{x}{3}$	$3r$	$\frac{V_x}{3r}$
:	:	:
$\frac{x}{n}$	$nr$	$\frac{V_x}{nr}$

Using the empirical relationship by Smith, the variance of treatment means per unit area with  $nr$  replicates is:

$$\frac{V_x}{nr} = \left(\frac{1}{nr}\right) (V_1) / \left(\frac{x}{n}\right)^b = \frac{(V_1)}{r \cdot x^b} \cdot \frac{1}{n^{(1-b)}}$$

With the increase in 'n', the value of the right hand side decreases as  $n$  increases ( $0 < b < 1$ ). The above observation implies that it is advantageous to use small (but how small or up to what size it can be regarded as acceptable) plot size, as  $n$  can be made as large as possible in case of a fixed total area. Thus the problem of determination of optimum plot size remains an interesting problem of research to Agriculturists and Statisticians involved directly with field experimentation.

13. The value of  $b$  i.e., the index of soil heterogeneity is used primarily to derive optimum plot size. The value of  $b$  indicates also the degree of correlation between adjacent experimental plots. Its value varies generally between unity and zero. The larger is the value of the index,  $b$ , lower is the correlation between adjacent plots, indicating that fertile spots are distributed randomly or in patches.

14. The values of 'b' lying in the range 0.3 to 0.7 do not greatly affect the increase in cost or in variance, when plots of sizes,  $\frac{1}{4}$  th time up to 4 times, of the optimum plot size are used. Based on the above results, plot sizes of  $\frac{1}{2}$  (half) time to 2 times of the optimum plot size can be recommended without any loss in efficiency. However, for plot sizes of  $\frac{1}{4}$  th time to 4 times of the optimum size, a loss in efficiency of 20% results because of the increased variance.

### 15. Determination of Optimum Plot Size in case of Non-isotropic field

16. Smith's empirical law relating the variance of crop yields per unit area to plot size is  $V_x = V_1 / x^b$  where  $V_x$  is the variance among plots (basic plots) with an area having  $x$  number of *units (unit-plots)* and  $V_1$  is the variance among unit-plots (plots of unit size, say, 1m x 1m size). The factor  $b$  is an index of heterogeneity. If the plots are spatially uncorrelated, then  $b$  will be 1. It can approach a limiting value of zero if no heterogeneity exists. More specifically, if  $x$  corresponds to an area, say,  $W$  and the size of the basic plot is  $w$  (then number of basic plots =  $x = (1/(w/W))$ ), we can rewrite the above equation as,  $V_W = V_w \cdot [w/W]^b$ , where  $V_W$  and  $V_w$  are the two corresponding variances. Smith's empirical law is applicable only under the broad assumption of an isotropic field (one directional trend of heterogeneity). However, in field condition, it is often encountered that the fertility pattern shows a directional trend (trend in two rectangular directions). To take anisotropy into account, a general variance relationship similar to Smith's law may be written as below:

$$V_{n,s} = V_1 / (n_1^{b_1} \cdot n_2^{b_2})$$

where  $n_1, n_2$  are the numbers of basic plots taken along the row direction and the column directions respectively;  $V_1$  is the variance of the basic plots;  $V_{n,s}$  is the variance of plots each of which has  $n = n_1 \cdot n_2$  basic plots;  $b_1$  and  $b_2$  are the indices which characterize the heterogeneity in the  $X, Y$  directions of a 2-D field, respectively. Such anisotropic models are used by Modjeska and Rawlings, Sethi, Zhang et al. etc.

17. For an isotropic field,  $b_1 = b_2$  resulting in  $V_{n,s} = V_1 / (n_1 \cdot n_2)^{b_1}$  which is essentially the same formula as given by Smith. For a completely uniform field ( $b_1 = b_2 = 0$ ); and for a field with no spatial correlation,  $b_1 = b_2 = 1$ . Using the logarithmic form we can rewrite the Smith's equation (equation 2.8) as,

$$\log(V_{n,s} / V_1) = -b_1 \log(n_1) - b_2 \log(n_2)$$

And the equation (2.8) or the equation (2.9) is used to compute the indices of heterogeneity, i.e.,  $b_1$  and  $b_2$  from available data.  $V_1$  is calculated from the basic units (the original data) while  $V_{n,s}$  is estimated from the reconstructed plots each of which consists of  $n = n_1.n_2$  basic units. During the reconstruction of the plots, if  $n_2$  is fixed (e.g.,  $n_2 = 1$ ) and  $n_1$  is varied i.e.,  $n_1 = 1, 2, 3, \dots$ ,  $V_{n,s}$  is a function of  $n_1$  only. Therefore, the second term i.e.,  $b_2 \log(n_2)$  is constant and we can compute  $b_1$  from the relationship,  $\log(V_{n,s}/V_1)$  vs.  $\log(n_1)$ . Similarly, if  $n_1$  is fixed and  $n_2$  is varied,  $b_2$  can be computed from the relationship,  $\log(V_{n,s}/V_1)$  vs.  $\log(n_2)$ . If the same number of units are taken along both  $X$  and  $Y$  directions, i.e.,  $n_1 = n_2$ , we have,  $\log(V_{n,s}/V_1) = -(b_1 + b_2) \log(n_1) = -b_s \log(n_1)$ . Thus  $b_s$  can be obtained from the linear regression of  $\log(V_{n,s}/V_1)$  on  $\log(n_1)$ . Now to verify whether the equation  $V_{n,s} = V_1 / (n_1^{b_1} . n_2^{b_2})$  is a reasonable mathematical form to characterize the heterogeneity in two directions, the sum of  $b_1$  and  $b_2$  computed from should be close to  $b_s$  independently computed from equation.

18. To study the effect of plot shapes on variances, Zhang et al. Have used an index called **relative difference of variance (RV), i.e.,  $RV = 100 \cdot (V_{n,s} - V_n) / V_n$**  .... (2.11), where  $V_{n,s}$  is computed using equation (2.8), and  $V_n = V_1 / (n_1 . n_2)^{0.5(b_1 + b_2)}$  .. (2.12). The equation, (2.12) represents the variance assuming that the field is isotropic. **Less RV value will indicate a more efficient plot.**

19. It is recalled that the optimum plot size has been defined as the size which balances between precision and sampling cost. The cost per plot is given by a linear relation  $K_1 + K_2 . n$ . Thus an objective function accounting for both cost and variance in an isotropic field can be expressed by  $C = (K_1 + K_2 . n) \cdot \frac{V_1}{n^b}$  ....,

where  $n$  is the number of units in the chosen plot and  $K_1, K_2$  are the cost components as defined earlier. This objective function is minimized when  $n_0 = \frac{K_1 . b}{K_2 . (1 - b)}$ , where  $n_0$  is the optimum plot size in terms of the

number of basic units. Let,  $K = \frac{K_2}{K_1}$ , we have,  $n_0 . K = \frac{b}{1 - b}$ . Again let,  $z = \frac{n}{n_0}$  (i.e. the ratio of plot size of

$n$  units to the most efficient size of  $n_0$  units), then the objective function,  $C$  (can be rewritten as,

$$C = K_1 \cdot \left[ 1 + \frac{b}{1 - b} \cdot z \right] \cdot \frac{V_1}{n^b}$$

20. Now if  $C$  is the cost related to variance for a specified plot size  $n$  and

$C_{\min}$  is the cost related to variance for the optimum plot size  $n_0$  then using Smith's (1938) definition the

relative cost is 
$$y = \frac{C}{C_{\min}} = b.z^{(1-b)} + (1-b).z^{-b}$$

20. In case of non-isotropic fields, the cost per plot may be given by  $K_1 + K_2.n_1.n_2$ . Here the objective function accounting for both cost and variance is  $C_s = (K_1 + K_2.n_1.n_2) \cdot \frac{V_1}{n_1^{b_1}.n_2^{b_2}}$ ,  $n_1.n_2$  is equal to the total number of basic units in a chosen plot.

21. When,  $b_1 \leq 0.5$ ,  $b_2 \leq 0.5$ ,  $C_s$  is a monotone increasing function of  $n_1$  &  $n_2$ . When  $b_1$  and  $b_2$  are greater than 0.5, the minimum value of  $C_s$  depends on the larger value of the heterogeneity indices. if,  $b_1 > b_2$ ,  $C_s$  has a minimum value, when  $n_1 = \frac{K_1.b_1}{[k_2.(1-b_1)]}$  and  $n_2 = 1$ . Similarly, if,  $b_2 > b_1$ ,  $C_s$  is minimum,

when  $n_2 = \frac{K_1.b_2}{[k_2.(1-b_2)]}$ , and  $n_1 = 1$ .

22. In what precedes, it is evident that Smith (1938) recognizes that the index of soil heterogeneity need not be the same for plots oriented in different directions in the field and that the longer axis of the plots covers the greatest variability. In Smith's analysis this can be taken into account by computing variances and estimating  $b$  separately for different plot orientations. Following the formulation of Modjeska et al., the two-dimensional version of Smith's model for variances of plots of size, ' $r \times c$ ' can be defined as,  $V_{rc} = \frac{V_1}{r^{b_1}.c^{b_2}}$ , where  $b_1$  and  $b_2$  are indices of soil heterogeneity across rows and across columns, respectively. If  $b_1 = b_2$ , Smith's one-dimensional model holds as a special case.

23. Zhang et al. (recalling equations, HAE propose that the effect of plot shapes on variances AND can be measured by an index (based on  $V_{n,s}$  and  $V_n$ ) called it the relative difference of variance (**RV**) i.e.,  $RV = 100 \cdot (V_{n,s} - V_n) / V_n \dots$ , where  $V_{n,s}$  is computed using equation (2.8), and,  $V_n = V_1 / (n_1.n_2)^{0.5(b_1+b_2)} \dots$ . All formulae presented above will be applied on two real-life data sets, the first one on jute crop and the other one on rice crop (as illustrations).

**24. Spatial Models**

25. The observations from field experiments have been taken under identical conditions and hence independence of observations is a general assumption included in the specification of the model and also independence is a convenient assumption in order to make the analysis / algebra tractable. However, the assumption of independence in real life data is found to be unrealistic very often. Fisher has proposed principles of randomizations, blocking and replication to take care of the spatial dependence of observations from agriculture

field experiments. Randomization controls unwanted bias and neutralizes (but does not remove) to some extent the effect of spatial correlation. Randomization does not neutralize the spatial correlation at spatial scales larger or smaller than the plot dimensions.

26. To measure the spatial dependence in a data set, the similarity among data values separated and situated at a particular distance or lag are examined. The easiest way to express the similarity or dissimilarity between the paired values is to plot them in a  $Z(s)$  vs  $Z(s+h)$  scatter-gram ('h' being the lag). The plot is called 'h scatter-gram'. If the difference between all the  $Z(s)$  and  $Z(s+h)$  values is small, then all the points (scatter of points) will be close to the  $45^\circ$  line and the variable is described as auto-correlated. Alternatively, the larger is the difference between the pairs, the more diffuse will be the scatter of points around the  $45^\circ$  or " $Z(s)$  vs  $Z(s+h)$ " line. For small values of  $h$  the scatter of points will, on average, be tighter; whereas when the value of  $h$  is large, the scatter diagram is typically more diffuse. The tightness or diffuseness of the cloud of points about the  $45^\circ$  line may be thought of as their moment of inertia about the line. If  $s_i$  and  $s_j$  are the co-ordinates for all  $i = 1$  to  $N$  points in an  $h$ -scatter-gram ( $j = i + h$ ), then the moment of inertia for all points is defined as moment of inertia =

$$\frac{1}{2N} \sum_{i=1}^N (s_i - s_j)^2$$

In other words, the moment of inertia summarizes the spread of the cloud in an  $h$ -scatter-gram. The  $h$ -scattergram can be useful models for measuring the degree of similarity or dissimilarity between samples separated by a common distance but such a concept is not practical because of the fact that too many  $h$ -scatter-grams would be required to adequately characterize the spatial similarity/dissimilarity for all samples and for all  $h$  values. Therefore, in order to achieve a meaningful summary of these  $h$ -scatter-grams, the concept of moment of inertia is called upon. Closely related to the moment of inertia is one of the most familiar tools in geo-statistics: the semi-variogram or simply the variogram.

### 27. Spatial modelling of uniformity data by Variogram (important reading)

28. Let the model be,  $Z(s) = \mu(s) + \delta(s)$ . The most common problems in statistical analysis involve **inference** on the **large scale variation**  $\mu(\cdot)$ . Usually  $\delta(\cdot)$  is assumed to be **white noise**. **Modelling and fitting spatial dependence parameters** from the data **allow** for **efficient estimation** of the **parameters in**  $\mu(\cdot)$ . Besag illustrated use of modelling **spatial or temporal correlation** in error process  $\delta(\cdot)$  for the efficient estimation of  $\mu(\cdot)$ . To illustrate and to **obtain a true model** for **spatial variation, responses** from the **uniformity trials** shall be **used**. By treating the uniformity data as a spatial data, such data can be written as  $\{Z(i(1.0), j(1.0)): i = 1, 2, \dots, R; j = 1, 2, \dots, C\}$ . Here each unit plot is a square plot of area  $1m \times 1m$ . We denote  $Z(i, j)$ , the  $(i, j)$ -th data-pair, where  $i = 1$  corresponds to the most Easterly row and  $j = 1$  corresponds to the most Northerly column. The two way layout (stochastic model) of the plots  $[Z(i, j) = a + r_i + c_j + \delta(i, j)]$  is proposed.

29. The residuals (after fitting the model) are to be used to estimate  $2\gamma(h) = \text{Var}(\delta(s+h) - \delta(s))$ . The **implication** of the above concepts (related to semi-variogram) is summarized below:

**nugget**: The height of the jump of the semi-variogram at the discontinuity existing at the origin.

**sill**: Limit of the variogram when lag distance tends to infinity.

**range:** The distance in which the difference of the variogram from the sill becomes negligible. In models with a fixed sill, it is the distance at which this is first reached; for models with an asymptotic sill, it is conventionally taken to be the distance when the semi-variance first reaches 95% of the sill.

### **Relevance of Spatial Models in the Context of Soil Heterogeneity**

Soil heterogeneity complicates the design analysis of field experiments. The three fundamental principles of design of experiments are employed to mitigate the causes and effects of the complications embedded there in owing to soil heterogeneity. Though the three principles assist to reduce the magnitude of the effect of the complication, but, unfortunately, cannot eliminate the effects prevailing due to such complications. The results (after analysis) obtained from a controlled experiment assume that residuals from the treatment are spatially independent and that the within block variation is random. However, experience indicates that the above expectations are rarely fulfilled in case of field experiments because of existence of strong spatial auto-correlation of soil properties. Thus soil spatial variation is regarded as an important source of external variation and to take care against its presence it is necessary to modify the standard procedure of analysis (and testing) of experimental data obtained from conduction of field experiments. In the presence of a significant spatial correlation over small distances, the assumption of independence between plots is violated. Obviously, in such a situation, a field researcher may face incongruous results: clear difference in crop yields between experimental plots but no significant treatment effect.

Recalling the above-mentioned phenomenon related to “existence of strong spatial auto-correlation of soil properties”, and to confront its (spatial auto correlation) existence, it is necessary to study and analyze the structure of spatial variability of crop yield data obtained from controlled uniformity trials. Using variogram parameters like nugget, sill and range it can be investigated to obtain the optimal experimental plot size and shape and the configuration of blocks (Fagroud and Meirvenne, 2002). Recalling the implied meanings of the variogram parameters, namely, “**range** is the separation-distance beyond which two observations are independent of each other; **sill** is the variogram value corresponding to the range; the discontinuity at the origin is called the **nugget effect** and arises from a combination of random errors and sources of variation at distances smaller than the shortest sampling interval (Goovaerts, 1998)”, the index, “nugget/sill ratio (**NSR**)” can be used as a criterion indicating the extent to which the experimental errors between plots are randomly distributed in space (Bhatti *et al.*, 1991; Ersboll, 1996).

#### **2.7.1 NSR as a Criterion to Determine the Optimum Plot Size**

The determination criteria refer to the proposition that the smaller is the value of NSR the less random errors remain between plots; and as a consequence, the more the plot configuration meets the above condition, the less are the chances that experimental errors can be considered to be random. Conversely, the largest NSR value points to the configuration which best satisfies the underlying hypothesis of classical ANOVA techniques (Fagroud *et al.*, 2002).

The block configuration can also be studied using a criterion based on NSR. It is widely observed that the ranges of the variogram models differ greatly, the criterion, “NSR” is standardized as, “NSR/Range” by Fagroud *et al* (2002) for studying the best block configuration.

From the plot of NSR/Range vs. number of plots per block, the criterion, NSR/Range indicates that the optimal number of plots per block should be neither too high (i.e., NSR small) nor too small (i.e., significant spatial



correlation) since it represents a compromise between two parameter requirements indicating a weak spatial dependence of the residual errors: a large NSR and a short range. The methods of calculation of NSR or (NSR/Range) are illustrated on real-life data-sets (the reader may find out the optimum plot size).

As sample support size increases, variance is expected to decrease. With respect to a variogram, the effect will correspond to a decrease in sill.

The commonly used variogram models ( $\gamma_u$ ) in terms of a Dimensionless Length  $h/a$  are given in the following table:

**Table 3: Variogram Models**

Name	Function	Effective range
Exponential	$1 - \exp(-h/a)$	$3.0 a$
Spherical	$1.5 (h/a) - 0.5 (h/a)^3, h < a$ $1, h \geq a$	$0.82 a$
Gaussian	$1 - \exp [-(h/a)^2]$	$1.7 a$
Michaelis-Menton	$(h/a) / [1 + (h/a)]$	$19.0 a$
Linear	$h/a$	None

Various parametric variogram models are presented in Journel and Huijbregts (1978). The following basic models (for interested readers) are generally considered:

(i) *Linear model:*  $\gamma(h; \theta) = \begin{cases} 0, & \dots \dots \dots h = 0 \\ c_o + b_l \|h\|, & \dots \dots \dots h \neq 0 \end{cases}$

$\theta = (c_o, b_l)'$ , where  $c_o \geq 0$ , and  $b_l \geq 0$ .

(ii) *Spherical model:*  $\gamma(h; \theta) = \begin{cases} 0, & \dots \dots \dots h = 0, \\ c_o + c_s \{ (3/2)(\|h\|/a_s) - (1/2)(\|h\|/a_s)^3 \}, & \dots 0 < \|h\| \leq a_s, \\ c_o + c_s, & \dots \dots \dots \|h\| \geq a_s, \end{cases}$

$\theta = (c_o, c_s, a_s)'$ , where  $c_o \geq 0, c_s \geq 0, a_s \geq 0$ .

(iii) *Exponential model:*  $\gamma(h; \theta) = \begin{cases} 0, & \dots \dots \dots h = 0 \\ c_o + c_e \{ 1 - \exp(\|h\|/a_e) \}, & \dots \dots \dots h \neq 0 \end{cases}$

$\theta = (c_o, c_e, a_e)'$ , where  $c_o \geq 0, c_e \geq 0, a_e \geq 0$ .

(iv) *Power model:*  $\gamma(h; \theta) = \begin{cases} 0, & \dots \dots \dots h = 0 \\ c_o + b_p \|h\|^\lambda, & \dots \dots \dots h \neq 0 \end{cases}$

$\theta = (c_o, b_p, \lambda)'$ , where  $c_o \geq 0, b_p \geq 0, 0 \leq \lambda < 2$ .

(v) *Gaussian model:*  $\gamma(h; \theta) = \begin{cases} 0, & \dots \dots \dots h = 0 \\ c_o + c_g \{ 1 - \exp[-(\|h\|/a_g)^2] \}, & \dots \dots \dots h \neq 0 \end{cases}$

$$\theta = (c_o, c_g, a_g)' , \text{ where } c_o \geq 0, c_g \geq 0, a_g \geq 0.$$

$$(vi) \text{ Michaelis-Menton model: } \gamma(h; \theta) = \begin{cases} 0, & \dots\dots\dots h = 0 \\ c_o + c_m (\|h\|/a_m)/(1 + \|h\|/a_m), & \dots\dots\dots h \neq 0 \end{cases}$$

$$\theta = (c_o, c_m, a_m)' , \text{ where } c_o \geq 0, c_m \geq 0, a_m \geq 0.$$

In the following Data Section, Nugget/sill ratio (NSR) has been calculated as a step to obtain the optimum plot size (O.P.S.) (best fitted Variogram is required for determination of O.P.S.).

**2.8 EXAMPLE - DATA SETS (with Description)**

In this Section we consider **two real-life data sets** emanated from field experiments, one on jute crop (at an ICAR Research Institute) and the other on rice crop (at a State Agricultural University) conducted in southern part of West Bengal and in northern part of West Bengal respectively.

**2.8.1 A UNIFORMITY TRIAL ON JUTE:** Under the project JST 3.8, at an experimental field of ICAR-CRIJAF, Barrackpore, West Bengal, India (lat: 22°45'N, long: 88°26'E) in the year 2001, a uniformity trial was conducted in an area of 41 x 31 m<sup>2</sup> with jute (*Corchorus Olitorius*), JRO 524. The crop was grown using 25 x 6 cm<sup>2</sup> spacing, uniform management practices and without application of fertilizers. After 120 days of sowing, leaving 1 m area from all sides of the field, the crop was harvested from an effective area of 40 x 30 m<sup>2</sup>, bundled separately from each basic unit of 1 m<sup>2</sup> area, retted for 20 days in water, dried and fibre weights were recorded for 1200 units in g. An analysis of Variance table on the above-mentioned data set is given below:

**Table 4: ANOVA Table for Jute Uniformity Trial Data**

SOV	DF	MS	F-Cal	Prob. Level
Row	39	9601.897	3.5511	4.03E-12
Column	29	15171.800	5.6111	6.78E-19
Error	1091	2703.887		
Total	1199	3764396		

From the ANOVA table it is revealed that the heterogeneity-status among columns is more pronounced (to a moderate extent) than that among rows, though both the classifications, rows and columns, are highly significant in respect of variation among rows and of variation among columns. Pictorial diagrams (two-dimensional and three-dimensional) are presented below: The contour diagram (2-dimensional) and surface plot (3-dimensional) related to the Jute data are presented below in Figure 1 and Figure II respectively. Salient features as revealed from the data-set are presented in the paragraph delineated under the following two Figures, Figure –I and Figure -II.

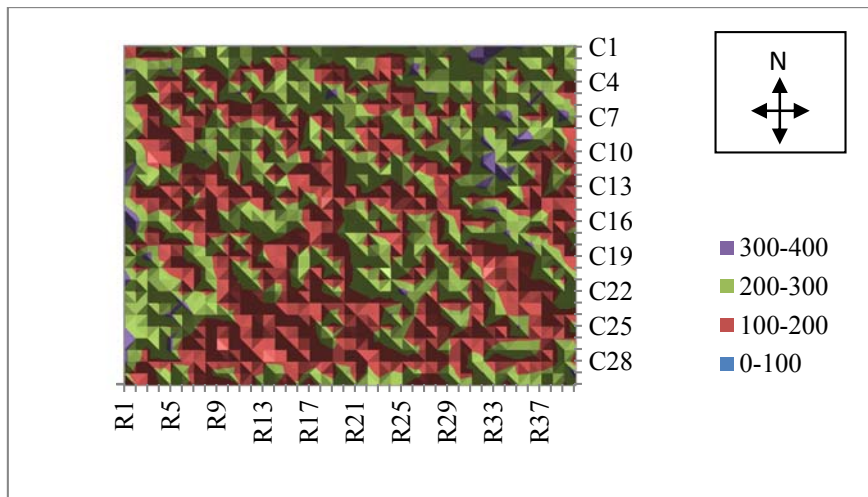


Figure – I

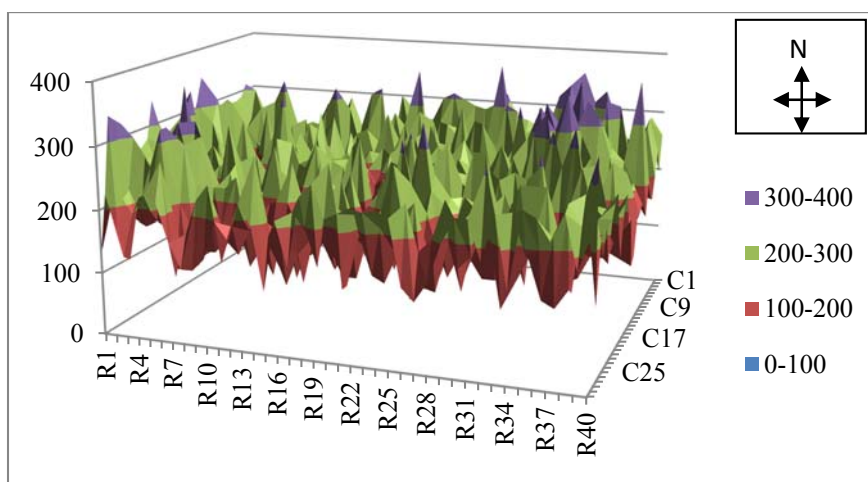


Figure – II

The above **two diagrams** reveal the following characteristics: (a) lowest yielding plots are rare, (b) moderate yielding plots (100-200 and 200-300) are in the maximum and such plots occur in patches, and (c) highest yielding plots are also rare. Moderately higher yielding (200-300) plots lie mainly along the western side and also along the north-eastern side of the experimental field.

**In Table 5, the Columns are self - expressive.** The contents of the Columns are calculated from the data obtained from the uniformity trial on Jute as described in the section 2.3.1. Referring to the equation, 2.6, and the related Columns, C<sub>13</sub> and C<sub>14</sub>, the value of

$$c = \frac{\sum_{i=1}^m w_i \cdot X_i \cdot Y_i}{\sum_{i=1}^m w_i \cdot X_i^2} = \frac{-13928.992 \text{ (from total of Col.13)}}{23650.539 \text{ (from total of Column 14)}} = -0.589; \text{ or } b = -c = 0.589. \text{ The value of index of}$$

heterogeneity, 'b' indicates that there exists moderate value of the coefficient of heterogeneity in the Jute field. Thus Smith's formula is regarded as the pioneering work in understanding the existence of soil heterogeneity index (nay, serial correlation of first order) in field data.

In Table 6,  $C_6$  gives the values of variance with respect to different plot sizes (same as in Table 5). The coefficient of variation (based on the columns  $C_7$  and  $C_8$ ) for each plot size is calculated (in cases of multiple values for the CV corresponding a plot size, the average value is considered).  $C_{10}$  of the Table 6 (same as in Table 5) gives the required CV (%) values. Best fitted models (curves) are then obtained based on the average CV (%) values.

Table – 5: Detailed calculations related to determination of optimum size of plots

SN	Width (W)	Length (L)	Size (x)	Mean	V(x)	$V_x=V(x)/x^2$	wi	X=log(x)	Y=log(Vx)-log(V1)	Xi * Yi	Xi^2	wi*Xi * Yi	wi*Xi^2	CV (%)	Average CV (%)
C <sub>1</sub>	C <sub>2</sub>	C <sub>3</sub>	C <sub>4</sub>	C <sub>5</sub>	C <sub>6</sub>	C <sub>7</sub>	C <sub>8</sub>	C <sub>9</sub>	C <sub>10</sub>	C <sub>11</sub>	C <sub>12</sub>	C <sub>13</sub>	C <sub>14</sub>	C <sub>15</sub>	C <sub>16</sub>
1	1	1	1	200.78	3139.61	3139.61	1199	0.000	0.000	0.000	0.000	0.000	0.000	27.907	27.907
2	2	1	2	401.56	8149.36	2037.34	599	0.693	-0.432	-0.300	0.480	-179.553	287.791	22.481	22.204
3	1	2		401.56	7753.20	1938.30	599	0.693	-0.482	-0.334	0.480	-200.244	287.791	21.928	
4	1	3	3	602.34	13027.54	1447.50	399	1.099	-0.774	-0.851	1.207	-339.393	481.573	18.949	18.949
5	4	1	4	803.12	20505.72	1281.61	299	1.386	-0.896	-1.242	1.922	-371.387	574.622	17.830	18.127
6	2	2		803.12	21891.68	1368.23	299	1.386	-0.831	-1.151	1.922	-344.278	574.622	18.423	
7	5	1	5	1003.90	32321.06	1292.84	239	1.609	-0.887	-1.428	2.590	-341.288	619.079	17.908	16.925
8	1	5		1003.90	25613.52	1024.54	239	1.609	-1.120	-1.802	2.590	-430.759	619.079	15.942	
9	2	3	6	1204.68	38919.11	1081.09	199	1.792	-1.066	-1.910	3.210	-380.141	638.870	16.376	15.911
10	1	6		1204.68	34620.87	961.69	199	1.792	-1.183	-2.120	3.210	-421.868	638.870	15.445	
11	8	1	8	1606.23	53758.37	839.97	149	2.079	-1.318	-2.742	4.324	-408.515	644.287	14.435	14.874
12	4	2		1606.23	60498.97	945.30	149	2.079	-1.200	-2.496	4.324	-371.915	644.287	15.313	
13	10	1	10	2007.79	77541.93	775.42	119	2.303	-1.398	-3.220	5.302	-383.186	630.926	13.869	14.221
14	5	2		2007.79	96480.17	964.80	119	2.303	-1.180	-2.717	5.302	-323.310	630.926	15.470	
15	2	5		2007.79	79641.93	796.42	119	2.303	-1.372	-3.159	5.302	-375.864	630.926	14.056	
16	1	10		2007.79	73356.22	733.56	119	2.303	-1.454	-3.348	5.302	-398.391	630.926	13.490	
17	4	3	12	2409.35	103185.18	716.56	99	2.485	-1.477	-3.671	6.175	-363.446	611.301	13.332	13.532
18	2	6		2409.35	109465.99	760.18	99	2.485	-1.418	-3.524	6.175	-348.910	611.301	13.732	
19	5	3	15	3011.69	162026.55	720.12	79	2.708	-1.472	-3.987	7.334	-315.008	579.349	13.365	12.722
20	1	15		3011.69	132344.27	588.20	79	2.708	-1.675	-4.535	7.334	-358.299	579.349	12.079	
21	8	2	16	3212.47	158181.33	617.90	74	2.773	-1.626	-4.507	7.687	-333.514	568.856	12.381	12.381
22	20	1	20	4015.58	189162.79	472.91	59	2.996	-1.893	-5.671	8.974	-334.577	529.490	10.831	11.639
23	10	2		4015.58	227711.94	569.28	59	2.996	-1.707	-5.115	8.974	-301.795	529.490	11.883	
24	4	5		4015.58	216606.01	541.52	59	2.996	-1.757	-5.265	8.974	-310.632	529.490	11.590	
25	2	10		4015.58	242061.94	605.15	59	2.996	-1.646	-4.932	8.974	-290.993	529.490	12.252	
26	8	3	24	4818.70	259079.40	449.79	49	3.178	-1.943	-6.175	10.100	-302.584	494.901	10.563	10.977
27	4	6		4818.70	301320.21	523.13	49	3.178	-1.792	-5.695	10.100	-279.064	494.901	11.392	
28	5	5	25	5019.48	368150.25	589.04	47	3.219	-1.673	-5.386	10.361	-253.158	486.975	12.088	12.088

National Workshop cum Training Programme on Statistical Tools for Research Data Analysis (Series II)

29	10	3	30	6023.38	390888.96	434.32	39	3.401	-1.978	-6.728	11.568	-262.385	451.158	10.380	10.504
30	5	6		6023.38	496783.83	551.98	39	3.401	-1.738	-5.912	11.568	-230.585	451.158	11.702	
31	2	15		6023.38	440871.01	489.86	39	3.401	-1.858	-6.319	11.568	-246.423	451.158	11.023	
32	1	30		6023.38	288056.91	320.06	39	3.401	-2.283	-7.766	11.568	-302.877	451.158	8.910	
33	20	2	40	8031.17	542658.07	339.16	29	3.689	-2.225	-8.209	13.608	-238.066	394.627	9.172	9.639
34	8	5		8031.17	591435.66	369.65	29	3.689	-2.139	-7.892	13.608	-228.858	394.627	9.576	
35	4	10		8031.17	666864.97	416.79	29	3.689	-2.019	-7.449	13.608	-216.017	394.627	10.168	
36	40	1		8031.17	606871.87	379.29	29	3.689	-2.114	-7.797	13.608	-226.101	394.627	9.700	
37	8	6	48	9637.40	776221.08	336.90	24	3.871	-2.232	-8.641	14.986	-207.378	359.669	9.142	9.243
38	10	5	50	10038.96	879991.26	352.00	23	3.912	-2.188	-8.560	15.304	-196.890	351.990	9.344	
39	5	10		10038.96	1128126.04	451.25	23	3.912	-1.940	-7.589	15.304	-174.539	351.990	10.580	
40	20	3	60	12046.75	960450.72	266.79	19	4.094	-2.465	-10.094	16.764	-191.789	318.509	8.135	9.180
41	10	6		12046.75	1192705.99	331.31	19	4.094	-2.249	-9.207	16.764	-174.941	318.509	9.066	
42	4	15		12046.75	1160087.57	322.25	19	4.094	-2.277	-9.321	16.764	-177.098	318.509	8.941	
43	2	30		12046.75	970277.04	269.52	19	4.094	-2.455	-10.052	16.764	-190.997	318.509	8.177	
44	5	15	75	15058.44	2043629.06	363.31	15	4.317	-2.157	-9.311	18.641	-139.666	279.611	9.493	8.322
45	8	10	80	16062.33	1786828.10	279.19	14	4.382	-2.420	-10.604	19.202	-148.460	268.830	8.322	8.317
46	40	2		16062.33	1782149.52	278.46	14	4.382	-2.423	-10.616	19.202	-148.621	268.830	8.311	
47	20	5	100	20077.92	2117815.72	211.78	11	4.605	-2.696	-12.417	21.208	-136.586	233.284	7.248	7.671
48	10	10		20077.92	2640283.90	264.03	11	4.605	-2.476	-11.401	21.208	-125.416	233.284	8.093	
49	8	15	120	24093.50	2871283.61	199.39	9	4.787	-2.757	-13.197	22.920	-118.773	206.281	7.033	7.064
50	4	30		24093.50	2534611.39	176.01	9	4.787	-2.881	-13.794	22.920	-124.147	206.281	6.608	
51	40	3		24093.50	3310455.83	229.89	9	4.787	-2.614	-12.516	22.920	-112.641	206.281	7.552	
52	10	15	150	30116.88	4433463.84	197.04	7	5.011	-2.768	-13.872	25.106	-97.101	175.745	6.991	6.894
53	5	30		30116.88	4190485.27	186.24	7	5.011	-2.825	-14.154	25.106	-99.078	175.745	6.797	
54	20	10	200	40155.83	7008354.17	175.21	5	5.298	-2.886	-15.290	28.072	-76.451	140.361	6.593	6.678
55	40	5		40155.83	7379154.17	184.48	5	5.298	-2.834	-15.017	28.072	-75.086	140.361	6.765	
56	8	30	240	48187.00	4412057.50	76.60	4	5.481	-3.713	-20.351	30.037	-81.405	120.150	4.359	4.359
57	20	15	300	60233.75	10198039.58	113.31	3	5.704	-3.322	-18.946	32.533	-56.839	97.599	5.302	4.950
58	10	30		60233.75	7672622.92	85.25	3	5.704	-3.606	-20.569	32.533	-61.708	97.599	4.599	

**Table 6: Calculations related to determination of optimum plot size (C<sub>4</sub> and C<sub>16</sub>)**

C <sub>1</sub>	C <sub>2</sub>	C <sub>3</sub>	C <sub>4</sub>	C <sub>5</sub>	C <sub>6</sub>	C <sub>7</sub>	C <sub>15</sub>	C <sub>4</sub>	C <sub>16</sub>
Serial No.	W	L	Size	Mean	V(x)	V <sub>x</sub> = V(x)/x <sup>2</sup>	CV (%) = 100 x (√V <sub>x</sub> ) /Mean	Xi= WxL	Average CV(%)
1	1	1	1	200.78	3139.61	3139.61	27.91	1	27.91
2	2	1	2	401.56	8149.36	2037.34	22.48	2	22.20
3	1	2	2	401.56	7753.20	1938.30	21.93	3	18.95
4	1	3	3	602.34	13027.54	1447.51	18.95	3	18.95
5	4	1	4	803.12	20505.72	1281.61	17.83	4	18.13
6	2	2	4	803.12	21891.68	1368.23	18.42	4	18.13
7	5	1	5	1003.90	32321.06	1292.84	17.91	5	16.93
8	1	5	5	1003.90	25613.52	1024.54	15.94	5	16.93
9	2	3	6	1204.68	38919.11	1081.09	16.38	6	15.91
10	1	6	6	1204.68	34620.87	961.69	15.45	6	15.91
11	8	1	8	1606.23	53758.37	839.97	14.44	8	14.87
12	4	2	8	1606.23	60498.97	945.30	15.31	8	14.87
13	10	1	10	2007.79	77541.93	775.42	13.87	10	14.22
14	5	2	10	2007.79	96480.17	964.80	15.47	10	14.22
15	2	5	10	2007.79	79641.93	796.42	14.06	10	14.22
16	1	10	10	2007.79	73356.22	733.56	13.49	10	14.22
17	4	3	12	2409.35	103185.20	716.56	13.33	12	13.53
18	2	6	12	2409.35	109466.00	760.18	13.73	12	13.53
19	5	3	15	3011.69	162026.50	720.12	13.37	15	13.55
20	1	15	15	3011.69	132344.30	588.20	12.08	15	13.55
21	8	2	16	3212.47	158181.30	617.90	12.38	16	12.38
22	20	1	20	4015.58	189162.80	472.91	10.83	16	12.38
23	10	2	20	4015.58	227711.90	569.28	11.88	20	11.64
24	4	5	20	4015.58	216606.00	541.52	11.59	20	11.64
25	2	10	20	4015.58	242061.90	605.16	12.25	20	11.64
26	8	3	24	4818.70	259079.40	449.79	10.56	24	10.98
27	4	6	24	4818.70	301320.20	523.13	11.39	24	10.98
28	5	5	25	5019.48	368150.30	589.04	12.09	25	12.09
29	10	3	30	6023.38	390889.00	434.32	10.38	25	12.09
30	5	6	30	6023.38	496783.80	551.98	11.70	30	10.50
31	2	15	30	6023.38	440871.00	489.86	11.02	30	10.50
32	1	30	30	6023.38	288056.90	320.06	8.91	30	10.50
33	20	2	40	8031.17	542658.10	339.16	9.17	40	9.64
34	8	5	40	8031.17	591435.70	369.65	9.58	40	9.64
35	4	10	40	8031.17	666865.00	416.79	10.17	40	9.64
36	8	6	48	9637.40	776221.10	336.90	9.14	48	9.14
37	10	5	50	10038.96	879991.30	352.00	9.34	50	9.96
38	5	10	50	10038.96	1128126.00	451.25	10.58	50	9.96
39	20	3	60	12046.75	960450.70	266.79	8.14	60	8.58
40	10	6	60	12046.75	1192706.00	331.31	9.07	60	8.58

41	4	15	60	12046.75	1160088.00	322.25	8.94		
42	2	30	60	12046.75	970277.00	269.52	8.18		
43	5	15	75	15058.44	2043629.00	363.31	9.49	75	9.49
44	8	10	80	16062.33	1786828.00	279.19	8.32	80	8.32
45	20	5	100	20077.92	2117816.00	211.78	7.25		
46	10	10	100	20077.92	2640284.00	264.03	8.09	100	7.67
47	8	15	120	24093.50	2871284.00	199.40	7.03	120	6.82
48	4	30	120	24093.50	2534611.00	176.02	6.61		
49	10	15	150	30116.88	4433464.00	197.04	6.99		
50	5	30	150	30116.88	4190485.00	186.24	6.80	150	6.89
51	20	10	200	40155.83	7008354.00	175.21	6.59	200	6.59
52	8	30	240	48187.00	4412058.00	76.60	4.36	240	4.36
53	20	15	300	60233.75	10198040.00	113.31	5.30		
54	10	30	300	60233.75	7672623.00	85.25	4.60	300	4.95

Different models have been tried and their corresponding precision coefficients (in terms of R<sup>2</sup> values) are calculated. Out of many models only two best-fitting models are presented (data: x – plot size; y - average CV (%)), one is logarithmic model and the other one is the power model. The diagrams of the above two fitted models are shown in the following Figure III.

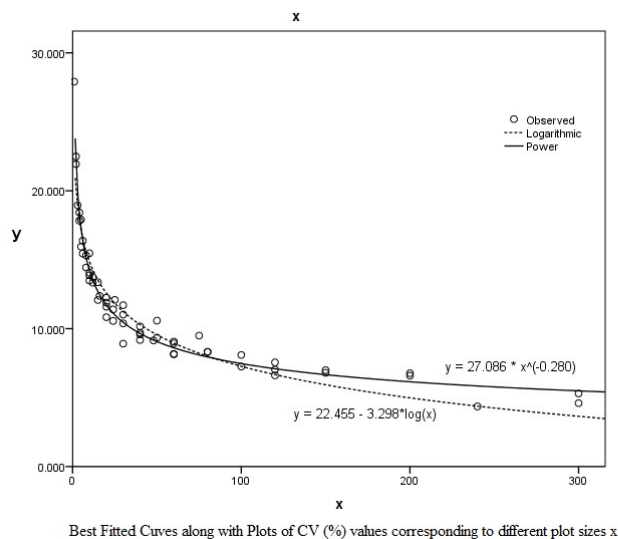


Figure III

The two best fitted models (Logarithmic and exponential) are compared with respect to their precision coefficients (Table 7). Both the models are found to be highly precise, however, the power model is slightly more precise in regard to its precision level, the values of Adjusted R-Square co-efficients are taken as precision levels). The validation tests (independence test and normality test) reveal conformation of independence and normality of errors obtained after fitting of the models on the data set on jute.

Table 7

Model	Value of R co-efficient	Value of R-Square co-efficient	Value of Adjusted R-Square co-efficient
Logarithmic	0.970	0.940	0.938
Power	0.983	0.966	0.965



Detailed Calculations: Power Model:  $y = a \cdot x^b + e$  (error), where,  $a = 27.086$  and  $b = -0.280$

The optimum plot size is:  $x_{opt} = \left\{ \frac{a^2 \cdot b^2 \cdot (1 + 2 \cdot b)}{(2 + b)} \right\}^{\frac{1}{2 \cdot (1 + b)}}$  (recalling equation: (2.21);

or,  $x_{opt} = 6.47$  sq. m... (2.36), values of ‘a’ and ‘b’ are given above. For final determination of the optimum **shape** of plots,  $y$  is further related to  $x$  as,  $y = a' \cdot x_1^{-b_1} \cdot x_2^{-b_2} \dots$  (Model 2), where  $x_1$  and  $x_2$  denote the number of units combined in row (E →W) direction and in column (N→S) direction respectively, to make a plot of size  $x$ . The equation of Model (2) is found as:  $y = 27.076 \cdot x_1^{0.270} \cdot x_2^{0.289}$ , i.e.,  $a' = 27.076$ ,  $b_1 = -0.270$ , and  $b_2 = -0.289$ . Adopting the procedure laid down in the theory (section 2.3) the value of  $x_{opt}$  ( $= x_{1opt} \cdot x_{2opt}$ ) is found as 7.837 sq m, where  $x_{1opt} = 1.835$  m and  $x_{2opt} = 4.271$  m. Thus the optimum plot size is 7.837 sq. m for the Jute crop.

### 2.5.2 A UNIFORMITY TRIAL ON RICE

#### Description of the Experiment

A uniformity trial with MW10 (a variety of rice) as Aus paddy was conducted (2000) at the RRS, Terai zone, Pundibari. The seedlings were transplanted in lines with hill to hill spacing of 20 cm. The distance between one line to the next line was kept at 20 cm leaving a border on each side. Uniform management practices were undertaken throughout the field. There was in total 22 rows, each of which was along N → S direction and 18 columns, each of which was along E →W direction. The field was harvested in basic units of size, 1 m x 1 m, as the field size is 22 m x 18 m i.e., 396 such basic units, each unit being of size 1 sq.m. The yields from each of these units were separately dried & weighed correct to the nearest gram.

Surface Diagram of Rice Data

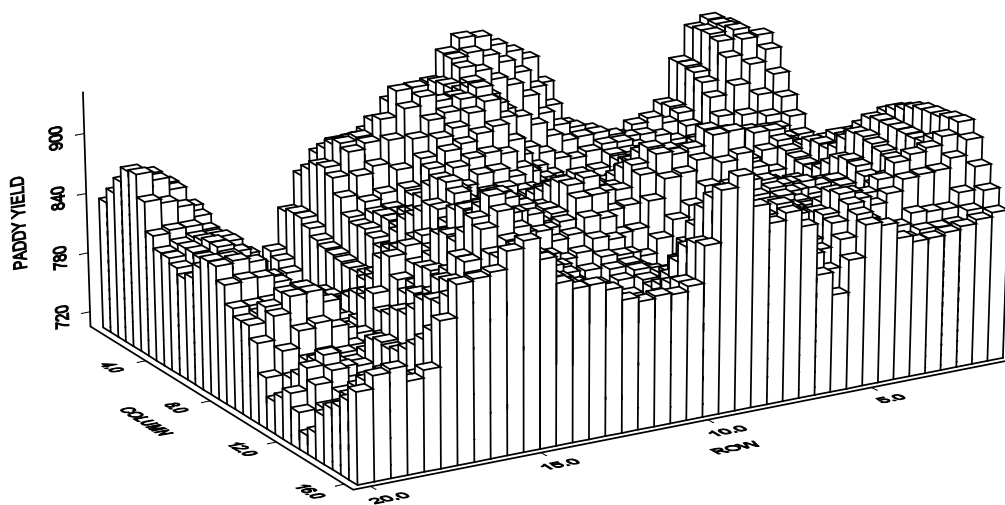


Figure –IV: Rice field fertility contour surface based on moving average of 3x3 basic units

Interpretation: The borders of Eastern, Northern & Southern ends of the Rice field are dominated by very high yields, the middle region shows a mixture of high and intermediate yields and Western borders are marked by low yields. Yields are not constant within any of these regions and appear to change along both row and column directions. Overall, a comprehensive visual presentation of the map of the actual yields on plots is displayed above.

**Simple Analyses on Rice Data:** The Rice uniformity data are analyzed (using two-way ANOVA). The results (given in Table 1.3.2) show that the Row mean square is more than four times higher than the Column mean square, indicating that the trend of soil fertility is more pronounced along the row i.e., E→W direction than that along the column i.e. N→S direction.

**Table 8: ANOVA Table for Rice Uniformity Data**

S.O.V.	D. F.	M.S.	F - Cal.
Row	21	1978.732	0.171
Column	17	490.308	0.042
Error	357	11519.33	
Total	395	13988.38	

**1. Calculation of Smith's optimum plot size (Rice Data-Set):** The basic units are combined by taking 1 to 9 units along N→S and 1 to 9 units across E→W in order to form plots of different sizes & shapes in case of rice data and also in case of jute data respectively. Based on the new sets of generated data between plot variance  $V_{(x)}$ , Variance per unit area  $V_x$  and coefficient of variability C.V. (%) are calculated for plots of various size and shapes. These are shown in the following tables:

**Table 9: Table showing different components – Mean, Variance and CV**

Plot Size	Width	Length	No. of Plots	Plot Mean	Plot Variance	Variance /sq.mt.	C.V. (%)
1	1	1	396	835.1843	13988.38	13988.380	14.161
2	2	1	198	1670.369	27349.12	6837.279	9.900
2	1	2	198	1670.369	35538.84	8884.711	11.286
3	3	1	126	2515.262	39960.83	4440.092	7.948
3	1	3	132	2505.553	60863.75	6762.639	9.846
4	4	1	90	3357.145	53770.79	3360.674	6.907
4	2	2	99	3340.737	80709.23	5044.327	8.504
4	1	4	88	3333.386	95908.78	5994.299	9.291
5	5	1	72	4196.431	62910.65	2516.426	5.977
5	1	5	66	4157.091	123822.5	4952.901	8.465
6	6	1	54	5050.056	79760.91	2215.581	5.592
6	3	2	63	5030.524	106220.4	2950.566	6.479
6	2	3	66	5011.106	136336.7	3787.132	7.369
6	1	6	66	5011.106	153177.6	4254.933	7.810
7	7	1	54	5868.944	107111.2	2185.944	5.576
7	1	7	44	5813.364	193632.8	3951.689	7.569
8	8	1	36	6807.583	89870.63	1404.229	4.404
8	4	2	45	6714.289	170132.4	2658.318	6.143
8	2	4	44	6666.773	236410	3693.907	7.293
8	1	8	44	6666.773	269960.9	4218.140	7.794
9	9	1	36	7575.083	92057.6	1136.514	4.005

National Workshop cum Training Programme on Statistical Tools for Research Data Analysis (Series II)

9	3	3	42	7545.786	187989.1	2320.853	5.746
9	1	9	44	7516.659	292946.6	3616.625	7.201
10	5	2	36	8392.861	169155.7	1691.557	4.900
10	2	5	33	8314.182	334565	3345.650	6.957
12	6	2	27	10100.11	240029.5	1666.872	4.851
12	4	3	30	10071.43	287783.7	1998.498	5.326
12	3	4	28	10034.93	334312.3	2321.613	5.762
12	2	6	33	10022.21	376188	2612.417	6.120
14	7	2	27	11737.89	306715.1	1564.873	4.718
14	2	7	22	11626.73	534253.7	2725.784	6.287
15	5	3	24	12589.29	263770.4	1172.313	4.080
15	3	5	21	12519.9	481429.6	2139.687	5.542
16	8	2	18	13615.17	286930.8	1120.823	3.934
16	4	4	20	13403.4	520879.2	2034.684	5.385
16	2	8	22	13333.55	707925.3	2765.333	6.310
18	9	2	18	15150.17	257340.2	794.260	3.348
18	6	3	18	15150.17	373504	1152.790	4.034
18	3	6	21	15091.57	535516.8	1652.830	4.849
18	2	9	22	15033.32	761731.1	2351.022	5.806
20	5	4	16	16754.25	513668.3	1284.171	4.278
20	4	5	15	16709.2	757961.1	1894.903	5.210
21	7	3	18	17606.83	508176.9	1152.329	4.049
21	3	7	14	17494.86	787478.1	1785.665	5.072
24	8	3	12	20422.75	332171.6	576.6868	2.822
24	6	4	12	20164.83	699194.2	1213.879	4.147
24	4	6	15	20142.87	727899.4	1263.714	4.236
24	3	8	14	20069.86	1052416	1827.111	5.112
25	5	5	12	20886.5	611211.6	977.9386	3.743
27	9	3	12	22725.25	366964.4	503.3805	2.666
27	3	9	14	22637.36	1224071	1679.110	4.887
28	7	4	12	23414.83	1006976	1284.408	4.285
28	4	7	10	23345.8	1167239	1488.825	4.627
30	6	5	9	25112	910596	1011.773	3.800
30	5	6	12	25178.58	589410.9	654.9011	3.049
32	8	4	8	27253.5	597357.7	583.3571	2.836
32	4	8	10	26806.8	1635726	1597.389	4.771
35	7	5	9	29213.11	1497438	1222.398	4.189
35	5	7	8	29182.25	905826.3	739.450	3.262
36	9	4	8	30247.25	569828.6	439.682	2.496
36	6	6	9	30300.33	886656	684.148	3.108
36	4	9	10	30214.3	1573305	1213.970	4.151
40	8	5	6	33976.33	673126.4	420.704	2.415
40	5	8	8	33508.5	1586304	991.440	3.759
42	7	6	9	35213.67	1250952	709.156	3.176
42	6	7	6	35068	1140448	646.512	3.045
45	9	5	6	37668	554096	273.627	1.976
45	5	9	8	37767.88	1761659	869.955	3.514
48	8	6	6	40845.5	107212.8	46.533	0.801
48	6	8	6	40329.67	1812224	786.555	3.338
49	7	7	6	40821.33	1871699	779.549	3.351
54	9	6	6	45450.5	279961.6	96.008	1.164
54	6	9	6	45450.5	2289459	785.136	3.329

56	8	7	4	47494.5	256170.7	81.687	1.065
56	7	8	6	46829.67	3119360	994.693	3.771
63	9	7	4	52602	149552	37.680	0.735
63	7	9	6	52820.5	3264122	822.404	3.420
64	8	8	4	54507	1319936	322.250	2.108
72	9	8	4	60494.5	1730773	333.868	2.174
72	8	9	4	61268.25	736341.3	142.041	1.401
81	9	9	4	68175.75	2025472	308.714	2.088

TABLE 10: Average C.V. (%) values for the same plot sizes (over different shapes)

Plot Size	No.of Plots	Plot Mean	Variance	Variance /sq.mt.	C.V. (%)
1	396	835.1843	13988.38	13988.38	14.161
2	198	1670.369	31443.98	7860.995	10.593
3	129	2510.408	50412.29	5601.366	8.897
4	92.3	3343.756	76796.267	4799.767	8.234
5	69	4176.761	93366.575	3734.664	7.221
6	62.3	5025.698	118873.9	3302.053	6.812
7	49	5841.154	150372	3068.817	6.573
8	42.3	6713.855	191593.48	2993.649	6.408
9	40.7	7545.843	190997.77	2357.997	5.650
10	34.5	8353.522	251860.35	2518.604	5.929
12	29.5	10057.17	309578.38	2149.85	5.515
14	24.5	11682.31	420484.4	2145.329	5.502
15	22.5	12554.6	372600	1656	4.811
16	20	13450.71	505245.1	1973.613	5.210
18	19.8	15106.31	482023.03	1487.726	4.509
20	15.5	16731.73	635814.7	1589.537	4.744
21	16	17550.85	647827.5	1468.997	4.561
24	13.3	20200.08	702920.3	1220.348	4.079
25	12	20886.5	611211.6	977.9386	3.743
27	13	22681.31	795517.7	1091.245	3.777
28	11	23380.32	1087107.5	1386.617	4.457
30	10.5	25145.29	750003.45	833.3371	3.425
32	9	27030.15	1116541.9	1090.373	3.803
35	8.5	29197.68	1201632.2	980.9241	3.725
36	9	30253.96	1009929.9	779.2669	3.252
40	7	33742.42	1129715.2	706.072	3.087
42	7.5	35140.84	1195700	677.8345	3.111
45	7	37717.94	1157877.5	571.7915	2.745
48	6	40587.59	959718.4	416.5444	2.070
49	6	40821.33	1871699	779.5499	3.351
54	6	45450.5	1284710.3	440.5728	2.247
56	5	47162.09	1687765.4	538.1904	2.419
63	5	52711.25	1706837	430.042	2.078
64	4	54507	1319936	322.25	2.108
72	4	60881.38	1233557.2	237.9548	1.788
81	4	68175.75	2025472	308.7139	2.088

From the above table it is observed that both variance per unit area and C.V. decrease with the increase in size of plot. For calculating Smith's heterogeneity coefficient  $b$  we consider only those plot shapes which fit exactly into the whole field. For rice data Smith's coefficient is 0.54 and its corresponding optimum plot size is 5. 51 sq. m. The moderate value of soil heterogeneity index for rice data indicate that the rice data are correlated (data are less random).

Recalling the discussion on determination of Optimum Plot Size in case of Non-isotropic field (Section 2.2), calculation of heterogeneity coefficients,  $b_1$ ,  $b_2$  and  $b_s$  are presented in the Table 11 (next page). The entries in the table speak for themselves. The discussion below the table also refers to the fact that the value of the Smith's co-efficient, 'b' (calculated for the rice data-set) is some sort averages of  $b_1$  and  $b_2$ .

**Table 11: Calculation Indices of heterogeneity  $b_1$ ,  $b_2$  and  $b_s$  for the rice data.**

Rice data							
Shape	Variance / Unit area	Log (var)	Plot size	Shape	Variance / Unit area	Log (size)	Log (var)
1X1	13988.38	2.16	1	1X1	13988.38	0	2.16
1X2	8884.71	1.78	2	2X1	6837.28	0.69	1.44
1X3	6762.64	1.43	3	3X1	4440.09	1.09	1.01
1X4	5994.29	1.31	4	4X1	3360.67	1.38	0.73
1X5	4952.90	1.12	5	5X1	2516.42	1.61	0.44
1X6	4254.93	0.97	6	6X1	2215.58	1.79	0.31
1X7	3951.69	0.89	7	7X1	2185.94	1.94	0.30
1X8	4218.14	0.96	8	8X1	1404.22	2.08	-0.13
1X9	3616.63	0.81	9	9X1	1136.51	2.19	-0.34
Intercept=	2.14			Intercept =	2.20		
Slope=	-0.61	<b><math>b_2=0.61</math></b>		Slope=	-1.08	<b><math>b_1=1.08</math></b>	
Corr. Coeff.	-0.99			Corr. Coeff.	-0.99		
					Variance per		
			Plot Size	Shape	Unit area	Log (Size)	Log (Var)
			1	1X1	13988.38	0	2.16
			2	2X2	5044.32	0.69	1.14
			3	3X3	2320.85	1.09	0.36
			4	4X4	2034.68	1.39	0.23
			5	5X5	977.94	1.61	-0.50
			6	6X6	684.15	1.79	-0.85
			7	7X7	779.55	1.95	-0.72
			8	8X8	322.25	2.10	-1.61
			9	9X9	308.71	2.20	-1.65
				Intercept ==	2.30		
				Slope =	-1.73	<b><math>b_s = 1.73</math></b>	
				Corr. Coeff.	-0.98		

The contents of the above tables display the values of  $b_1$ ,  $b_2$  and  $b_s$  for the rice data. In case of rice data the index of heterogeneity is nearly double along the row direction than that along the column direction. The sum of  $b_1$  and  $b_2$  is close to  $b_s$  for rice data. The value of Smith's co-efficient, 'b' calculated for rice data is some sort averages of  $b_1$  and  $b_2$ .

Zhang et al. (1994) proposed that the effect of plot shapes on variances can be measured by an index called relative difference of variance (**RV**) i.e.,  $\mathbf{RV} = 100 \cdot (V_{n,s} - V_n) / V_n \dots\dots\dots (2.11)$ ,  $V_{n,s}$  is computed using equation (2.8), and,  $V_n = V_1 / (n_1 \cdot n_2)^{0.5(b_1 + b_2)} \dots (2.12)$ . For the data set on Rice, the values of RV coefficient have been calculated and the findings (effects of plot shapes on variance) are evident from the following Figure IV. For the Rice data (Fig IV) the efficiency decreases considerably as we increase the number of units along column direction as compared to the number units taken along the row direction (from the diagrammatic representation of the rice-scenario below)

Note (See Fig – IV): Less RV value indicates a more efficient plot. In earlier discussion, optimum plot size has been defined as the size which balances between precision and sampling cost.

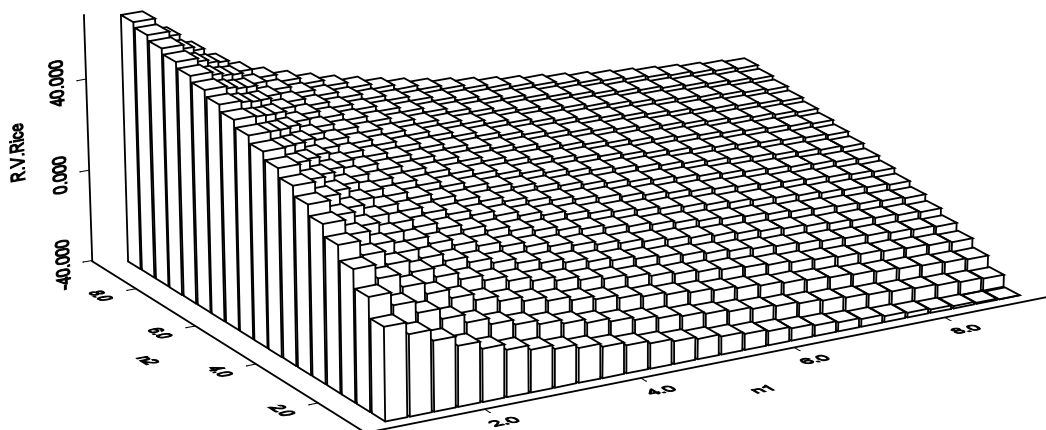


Fig.IV: Relative difference of variances R.V. as a function of plot shapes i.e. units along row and column directions (along  $n_1$  direction and along  $n_2$  direction) respectively for Rice data.

## Time Series Modeling- an overview

Dr K K Goswami

Former Principal Scientist, ICAR-CIFRI

## Expectation, mean & variance

---

- The *expectation* (E) of a variable is its mean value in the population
- $E(x) \equiv$  mean of  $x = \mu$
- $E([x - \mu]^2) \equiv$  mean of squared deviations about  $\mu$   
 $\equiv$  variance =  $\sigma^2$
- Can estimate  $\sigma^2$  from sample as

$$\text{Var}(x) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

## Covariance

- If we have 2 variables ( $x, y$ ) we can generalize variance

$$\sigma_x^2 = E[(x - \mu_x)(x - \mu_x)]$$

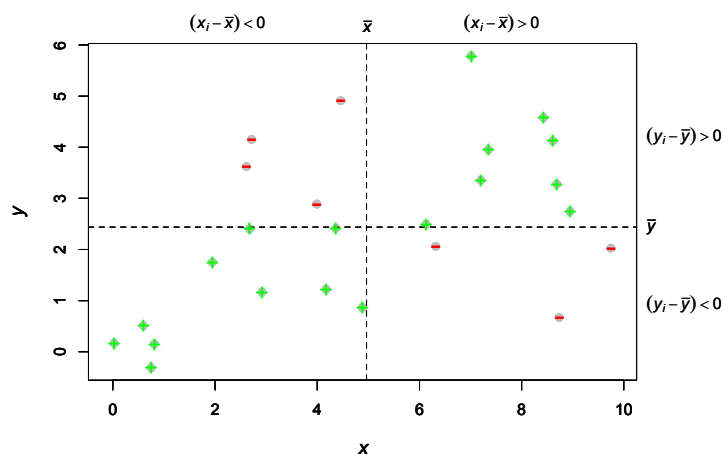
to *covariance*

$$\gamma(x, y) = E[(x - \mu_x)(y - \mu_y)]$$

- Can estimate  $\gamma$  from sample as

$$\text{Cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

## Graphical example of covariance





## Correlation

- *Correlation* is a dimensionless measure of the linear association between 2 variables  $x$  &  $y$
- It is simply the covariance standardized by the standard deviations

$$\rho(x, y) = \frac{E[(x - \mu_x)(y - \mu_y)]}{\sigma_x \sigma_y} = \frac{\gamma(x, y)}{\sigma_x \sigma_y} \in [-1, 1]$$

- Can estimate  $\gamma$  from sample as

$$\text{Cor}(x, y) = \frac{\text{Cov}(x, y)}{\text{sd}(x)\text{sd}(y)}$$

### Two Main Goals

There are two main goals of time series analysis: (a) identifying the nature of the phenomenon represented by the sequence of observations, and (b) forecasting (predicting future values of the time series variable). Both of these goals require that the pattern of observed time series data is identified and more or less formally described. Once the pattern is established, we can interpret and integrate it with other data (i.e., use it in our theory of the investigated phenomenon, e.g., seasonal commodity prices). Regardless of the depth of our understanding and the validity of our interpretation (theory) of the phenomenon, we can extrapolate the identified pattern to predict future events.

### **Systematic Pattern and Random Noise**

As in most other analyses, in time series analysis it is assumed that the data consist of a systematic pattern (usually a set of identifiable components) and random noise (error) which usually makes the pattern difficult to identify. Most time series analysis techniques involve some form of filtering out noise in order to make the pattern more salient.

### **Two General Aspects of Time Series Patterns**

Most time series patterns can be described in terms of two basic classes of components: trend and seasonality. The former represents a general systematic linear or (most often) nonlinear component that changes over time and does not repeat or at least does not repeat within the time range captured by our data (e.g., a plateau followed by a period of exponential growth). The latter may have a formally similar nature (e.g., a plateau followed by a period of exponential growth), however, it repeats itself in systematic intervals over time. Those two general classes of time series components may coexist in real-life data.

### **Trend Analysis**

There are no proven "automatic" techniques to identify trend components in the time series data; however, as long as the trend is monotonous (consistently increasing or decreasing) that part of data analysis is typically not very difficult. If the time series data contain considerable error, then the first step in the process of trend identification is smoothing.

**Smoothing.** Smoothing always involves some form of local averaging of data such that the nonsystematic components of individual observations cancel each other out. The most common technique is *moving average* smoothing which replaces each element of the series by either the simple or weighted average of  $n$  surrounding elements, where  $n$  is the width of the smoothing

**Fitting a function.** Many monotonous time series data can be adequately approximated by a linear function; if there is a clear monotonous nonlinear component, the data first need to be transformed to remove the nonlinearity. Usually a logarithmic, exponential, or (less often) polynomial function can be used.

### **Analysis of Seasonality**

Seasonal dependency (seasonality) is another general component of the time series pattern. It is formally defined as correlational dependency of order  $k$  between each  $i$ 'th element of the series and the  $(i-k)$ 'th element and measured by autocorrelation (i.e., a correlation between the two terms);  $k$  is usually called the *lag*. If the measurement error is not too large, seasonality can be visually identified in the series as a pattern that repeats every  $k$  elements.

*Stationarity* is a convenient assumption that allows us to describe the statistical properties of a time series. In general, a time series is said to be stationary if there is no systematic change in mean or variance, no systematic trend, and no periodic variations or seasonality

## Autocovariance function (ACVF)

- For stationary ts, we can define the *autocovariance function* (ACVF) as a function of the time lag (k)

$$\gamma_k = E[(x_t - \mu_x)(x_{t+k} - \mu_x)]$$

- Very “smooth” series have large ACVF for large k; “choppy” series have ACVF near 0 for small k
- Can estimate  $\gamma_k$  from sample as

$$c_k = \frac{1}{n} \sum_{t=1}^{n-k} (x_t - \bar{x})(x_{t+k} - \bar{x})$$

## Autocorrelation function (ACF)

---

- The *autocorrelation function* (ACF) is simply the ACVF normalized by the variance

$$\rho_k = \frac{\gamma_k}{\sigma^2} = \frac{\gamma_k}{\gamma_0}$$

- ACF measures the correlation of a time series against a time-shifted version of itself (& hence “auto”)
- Can estimate  $\gamma_k$  from sample as

$$r_k = \frac{c_k}{c_0} \quad \text{where} \quad -1 \leq r_k \leq 1 \quad \text{and} \quad r_k = r_{-k}$$

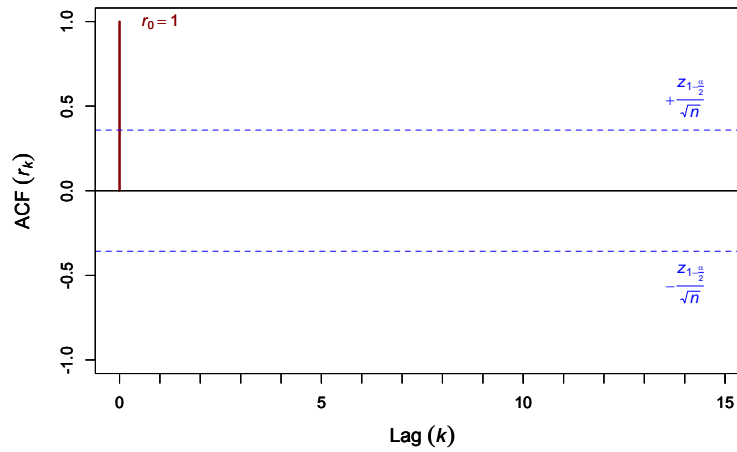
## Properties of the ACF

---

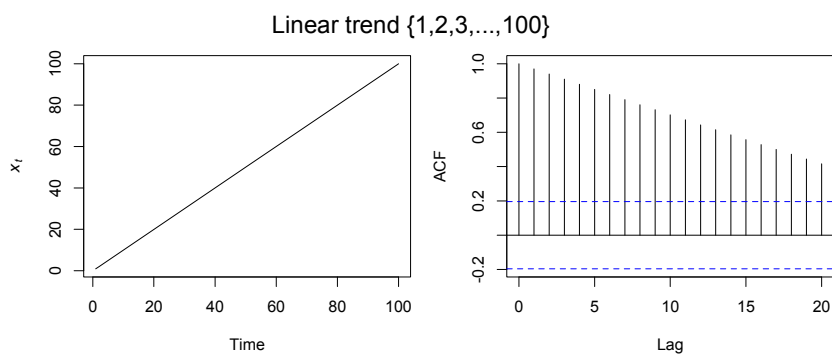
The ACF has several important properties, including

- 1)  $-1 \leq r_k \leq 1$ ,
- 2)  $r_k = r_{-k}$  (ie, it's an “even function”),
- 3)  $r_k$  of periodic function is itself periodic
- 4)  $r_k$  for sum of 2 indep vars is sum of  $r_k$  for each

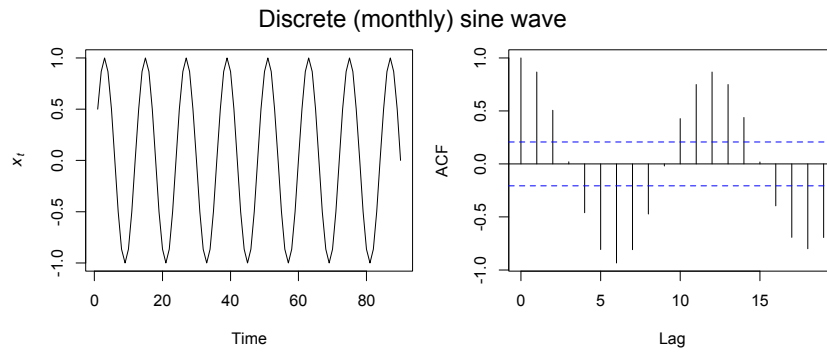
## The correlogram



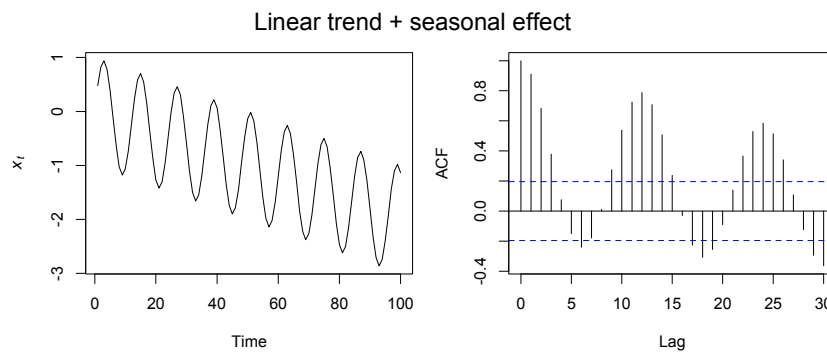
## Correlogram for deterministic trend



## Correlogram for sine wave

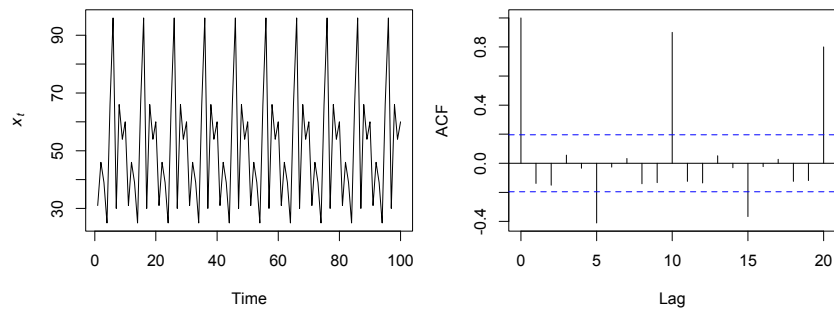


## Correlogram for trend + season



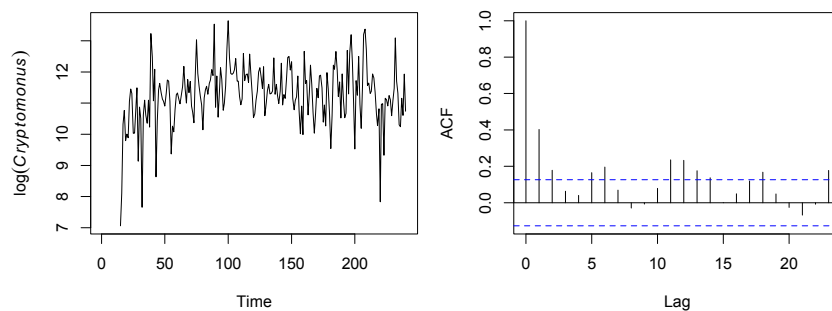
## Correlogram for random sequence

Random sequence of 10 numbers repeated 10 times



## Correlogram for real data

Lake Washington phytoplankton





## White noise (WN)

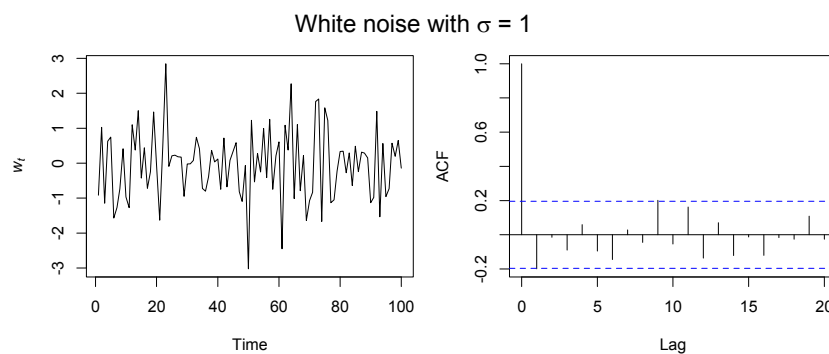
A time series  $\{w_t : t = 1, 2, 3, \dots, n\}$  is *discrete white noise* if the variables  $w_1, w_2, w_3, \dots, w_n$  are

- 1) *independent*, and
- 2) *identically distributed* with a mean of zero

WN has the following 2<sup>nd</sup>-order properties:

$$\mu_w = 0 \quad \gamma_k = \begin{cases} \sigma^2 & \text{if } k = 0 \\ 0 & \text{if } k \neq 0 \end{cases} \quad \rho_k = \begin{cases} 1 & \text{if } k = 0 \\ 0 & \text{if } k \neq 0 \end{cases}$$

## White noise



**Autoregressive process.** Most time series consist of elements that are serially dependent in the sense that one can estimate a coefficient or a set of coefficients that describe consecutive elements of the series from specific, time-lagged (previous) elements. This can be summarized in the equation:

$$x_t = \xi + \phi_1 * x_{(t-1)} + \phi_2 * x_{(t-2)} + \phi_3 * x_{(t-3)} + \dots + \varepsilon$$

Where:

$\xi$  is a constant (intercept), and

$\phi_1, \phi_2, \phi_3$  are the autoregressive model parameters.

Put in words, each observation is made up of a random error component (random shock, ) and a linear combination of prior observations.

**Stationarity requirement.** Note that an autoregressive process will only be stable if the parameters are within a certain range; for example, if there is only one autoregressive parameter then it must fall within the interval of  $-1 < \phi < 1$ . Otherwise, past effects would accumulate and the values of successive  $x_t$ 's would move towards infinity, that is, the series would not be stationary. If there is more than one autoregressive parameter, similar (general) restrictions on the parameter values can be defined

**Moving average process.** Independent from the autoregressive process, each element in the series can also be affected by the past error (or random shock) that cannot be accounted for by the autoregressive component, that is:

$$x_t = \mu + \varepsilon_t - \theta_1 * \varepsilon_{(t-1)} - \theta_2 * \varepsilon_{(t-2)} - \theta_3 * \varepsilon_{(t-3)} - \dots$$

Where:

$\mu$  is a constant, and

$\theta_1, \theta_2, \theta_3$  are the moving average model parameters.

Put in words, each observation is made up of a random error component (random shock,  $\varepsilon$ ) and a linear combination of prior random shocks.

**Invertibility requirement.** The moving average equation above can be rewritten (*inverted*) into an autoregressive form (of infinite order). However, analogous to the stationarity condition described above, this can only be done if the moving average parameters follow certain conditions, that is, if the model is *invertible*. Otherwise, the series will not be stationary.

### **ARIMA Methodology**

**Autoregressive moving average model.** The general model introduced by Box and Jenkins (1976) includes autoregressive as well as moving average parameters, and explicitly includes differencing in the formulation of the model. Specifically, the three types of parameters in the model are: the autoregressive parameters ( $p$ ), the number of differencing passes ( $d$ ), and moving average parameters ( $q$ ). In the notation introduced by Box and Jenkins, models are summarized as ARIMA ( $p, d, q$ ); so, for example, a model described as (0, 1, 2) means that it contains 0 (zero) autoregressive ( $p$ ) parameters and 2 moving average ( $q$ ) parameters which were computed for the series after it was differenced once.

### **Identification Phase**

**Number of parameters to be estimated.** Before the estimation can begin, we need to decide on (identify) the specific number and type of ARIMA parameters to be estimated. The major tools used in the identification phase are plots of the series, correlograms of auto correlation (ACF), and partial autocorrelation (PACF)

*One autoregressive ( $p$ ) parameter:* ACF - exponential decay; PACF - spike at lag 1, no correlation for other lags.

*Two autoregressive ( $p$ ) parameters:* ACF - a sine-wave shape pattern or a set of exponential decays; PACF - spikes at lags 1 and 2, no correlation for other lags.

*One moving average ( $q$ ) parameter:* ACF - spike at lag 1, no correlation for other lags; PACF - damps out exponentially.

*Two moving average ( $q$ ) parameters:* ACF - spikes at lags 1 and 2, no correlation for other lags; PACF - a sine-wave shape pattern or a set of exponential decays.

*One autoregressive ( $p$ ) and one moving average ( $q$ ) parameter:* ACF - exponential decay starting at lag 1; PACF - exponential decay starting at lag 1.

## The backward shift operator (**B**)

---

- Define the *backward shift operator* by

$$\mathbf{B}x_t = x_{t-1}$$

- Or, more generally as

$$\mathbf{B}^k x_t = x_{t-k}$$

- So, RW model can be expressed as

$$x_t = \mathbf{B}x_t + w_t$$

$$(1 - \mathbf{B})x_t = w_t$$

$$x_t = (1 - \mathbf{B})^{-1} w_t$$

## The difference operator ( $\nabla$ )

---

- Define the first *difference operator* as

$$\nabla x_t = x_t - x_{t-1}$$

- Differences of order  $d$  are then defined by

$$\nabla^d = (1 - B)^d$$

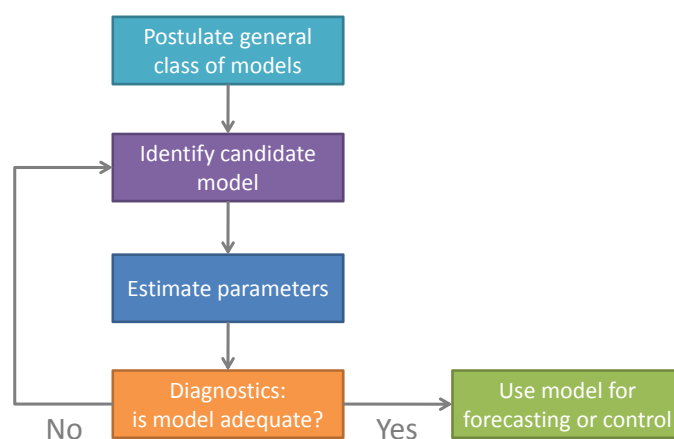
- So, first differencing a RW model yields WN

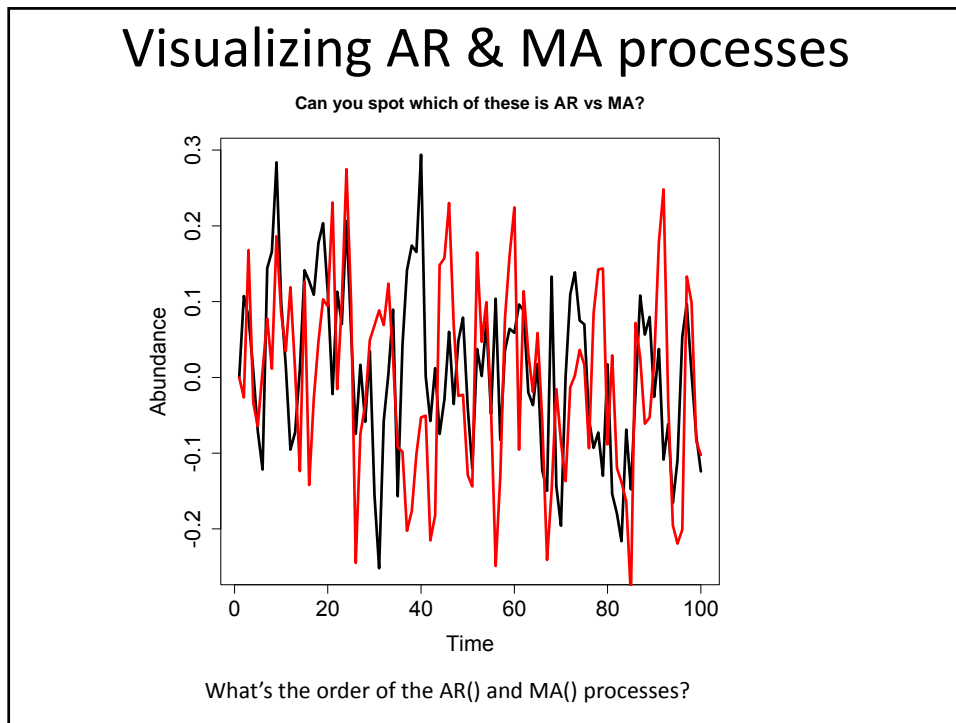
$$x_t - x_{t-1} = w_t$$

## Difference to remove trend/season

- Differencing is a very simple means for removing a trend or seasonal effect
- The 1<sup>st</sup>-difference removes a linear trend, a 2<sup>nd</sup>-difference would remove a quadratic trend, etc.
- For seasonal data, using a 1<sup>st</sup>-difference with *lag* = *period* removes both trend & seasonal effects
- Pro: no parameters to estimate
- Con: no estimate of stationary process

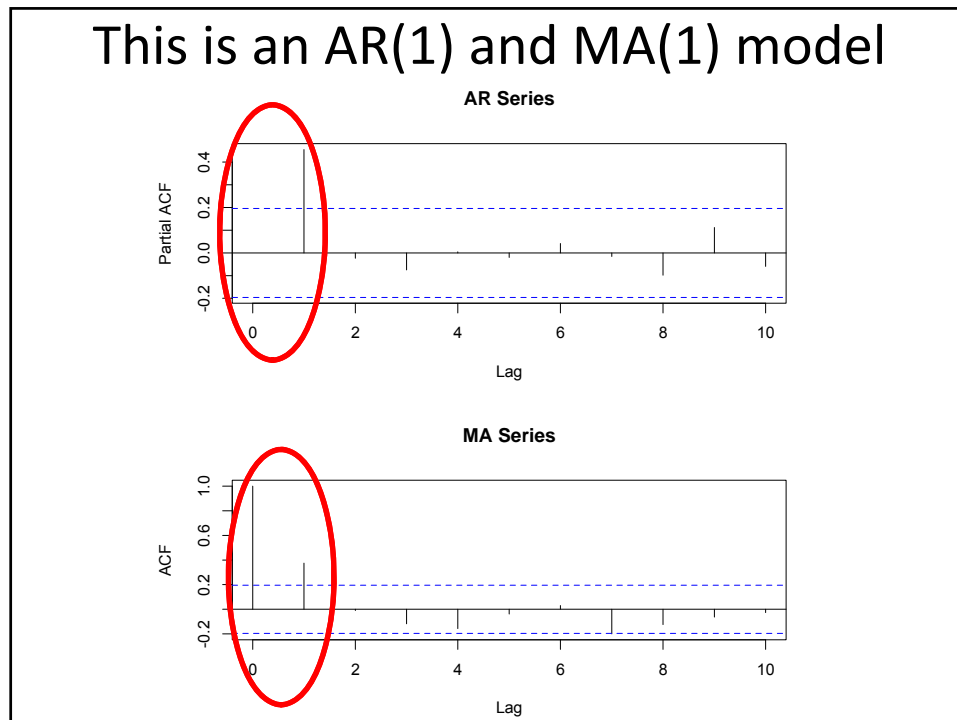
## Iterative approach to model building





### Remember Mark's lecture: use ACF & PACF for model ID

	ACF	PACF
AR( $p$ )	Tails off	Cuts off after lag- $p$
MA( $q$ )	Cuts off after lag- $q$	Tails off
ARMA( $p, q$ )	Tails off (after lag [ $q-p$ ])	Tails off (after lag [ $p-q$ ])



### Model selection tools to evaluate whether models are good

- Several candidate models might be built based on (1) hypotheses / mechanisms, (2) ARMA diagnostics
- Models can be evaluated by their ability to explain data
  - Schwarz or Bayesian Information Criterion (SIC=BIC)
- OR by the tradeoff in the ability to explain data, and ability to predict future data
  - Akaike's Information Criterion, AIC (or small sample AICc)



### **Statistics of Fit**

#### **Mean Square Error.**

The mean squared prediction error, MSE, calculated from the one-step-ahead forecasts.  $MSE = [1/n] SSE$ . This formula enables you to evaluate small holdout samples.

#### **Root Mean Square Error.**

The root mean square error (RMSE),  $\sqrt{MSE}$

#### **Mean Absolute Percent Error.**

The mean absolute percent prediction error (MAPE), .  
The summation ignores observations where  $y_t = 0$ .

#### **Mean Absolute Error.**

The mean absolute prediction error, .

#### **R-Square.**

The  $R^2$  statistic,  $R^2 = 1 - SSE / SST$ . If the model fits the series badly, the model error sum of squares,  $SSE$ , may be larger than  $SST$  and the  $R^2$  statistic will be negative.

#### **Akaike's Information Criterion.**

Akaike's information criterion (AIC),  $n \ln(MSE) + 2k$ .

#### **Schwarz Bayesian Information Criterion.**

Schwarz Bayesian information criterion (SBC or BIC),  
 $n \ln(MSE) + k \ln(n)$ .

#### **Amemiya's Prediction Criterion.**

Amemiya's prediction criterion,  $[1/n] SST \left( \frac{(n+k)}{(n-k)} \right) (1 - R^2) = \left( \frac{(n+k)}{(n-k)} \right) [1/n] SSE$ .

#### **Maximum Percent Error.**

The largest percent prediction error, .The summation ignores observations where  $y_t = 0$ .

#### **Minimum Percent Error.**

The smallest percent prediction error, .The summation ignores observations where  $y_t = 0$ .

## ANALYSIS OF COVARIANCE

**Premadhis Das**

**University of Kalyani**

Consider a dairy experiment where  $t$  fodders are to be compared for production of milk. We take  $n$  cows and allocate the 1<sup>st</sup> fodder to  $r_1$  cows chosen randomly, 2<sup>nd</sup> fodder to randomly chosen  $r_2$  cows and so on. Lastly the  $t$  th fodder is applied to  $r_t = n - (r_1 + r_2 + \dots + r_{t-1})$  cows left. Let  $y_{ij}$  denote the yield of the  $j$ th cow having  $i$ th fodder. The linear model for the observation is taken as

$$y_{ij} = \mu_i + e_{ij} \quad (1)$$

$$j = 1, 2, \dots, r_i \quad i = 1, 2, \dots, t. \quad \sum r_i = n.$$

where,

$\mu_i$  = effect due to  $i$ th fodder.

and  $e_{ij}$  = random error

We analyse the data to see if there is any differential effect due to the fodders. The model is known as ‘One Way Classification Model’.

Here we have ignored the possibility of differences among the cows due to the factors like ‘Breed of cows’, ‘Age of cows’, Location etc. which affect the yield of milk in addition to the effect of fodder. In model (1) we considered only the effect due to fodder. So all the effects due to the factors other than the fodder are accumulated in the error  $\{e_{ij}\}$ . So this will be inflated affecting the significance of  $F$ . It is well known that yield of milk depends on the age of cows, which is a quantitative variable.

For this we note the age of each cow together with the yield of milk. So here the observation are pair of values  $(y_{ij}, x_{ij})$ .  $x_{ij}$  is the age of  $j$ th cow having  $i$ th fodder,  $j = 1, 2, \dots, r_i$ ,  $i = 1, 2, \dots, t$ . Here  $x$ , the age is called the concomitant variable. We assume that the effect of age on the yield of milk  $y_{ij}$  is  $\beta x_{ij}$  where  $\beta$  is constant. So  $y_{ij}$  can be written as

$$y_{ij} = \mu_i + \beta x_{ij} + e'_{ij} \quad (2)$$

The 1<sup>st</sup> part ( $\mu_i$ ) containing the effect of fodder is the ‘Analysis of Variance’ (ANOVA) part. This part of the model is qualitative part. The 2<sup>nd</sup> part  $\beta x_{ij}$  is called

regression part.  $\beta$  is called the regression coefficient. This part is due to the quantitative factor 'Age'. The model (2) is called the 'Analysis of Covariance model for one-way classification. In the above model, the error gets reduced due to the introduction of the concomitant variable 'age'.

We can still control the error by introduction of another qualitative factor 'Breed' of cows. We take  $rt$  cows of  $r$  different breeds each containing  $t$  cows. We apply the  $t$  fodders to the  $t$  cows in each breed randomly. Let  $y_{ij}$  be the yield of milk for the cow of  $i$ th breed having  $j$ th fodder. Then ANOVA model for  $y_{ij}$  is given by

$$y_{ij} = \mu + \alpha_i + \theta_j + e''_{ij} \quad (3)$$

where  $\mu$  = general effect.

$\alpha_i$  = effect due to  $i$ th breed.

$\theta_j$  = effect due to  $j$ th fodder.

$e''_{ij}$  is the random error.

As the effect of breed is eliminated from  $e_{ij}$  of model (1), the error will get reduced. Again for further reduction of error we consider the regression of yield of milk ( $y$ ) on the age of cows ( $x$ ) together with the qualitative factors 'breed' and 'fodder'. Let  $x_{ij}$  be the age of ( $ij$ )th cow. So ANCOVA model for the two-way classification (or RBD) above is given by

$$y_{ij} = \mu + \alpha_i + \theta_j + \beta(x_{ij} - x_{00}) + e'''_{ij} \quad (4)$$

where

$$x_{00} = \frac{1}{rt} \sum_i \sum_j x_{ij}$$

The parametric restrictions are  $\sum_i \alpha_i = \sum_j \theta_j = 0$

The least-square estimators are

$$\hat{\mu} = y_{00},$$

$$\hat{\alpha}_i^* = (y_{i0} - y_{00}) - \hat{\beta}(x_{i0} - x_{00})$$

$$\hat{\theta}_j^* = (y_{0j} - y_{00}) - \hat{\beta}(x_{0j} - x_{00})$$

$$\text{and } \hat{\beta} = \frac{\sum_{i,j} (x_{ij} - x_{i0} - x_{0j} + x_{00})(y_{ij} - y_{i0} - y_{0j} + y_{00})}{\sum_{i,j} (x_{ij} - x_{i0} - x_{0j} + x_{00})^2}$$

Here  $\hat{\alpha}_i^*, \hat{\theta}_j^*$  are the corresponding estimators in the analysis of variance model minus adjustment factors due to the introduction of x.

The partitioning of the total sum of products of x and y is

$$\begin{aligned} \sum_i \sum_j (x_{ij} - x_{00})(y_{ij} - x_{00}) &= t \sum_i (x_{i0} - x_{00})(y_{i0} - y_{00}) + r \sum_j (x_{0j} - x_{00})(y_{0j} - y_{00}) + \\ &\quad \sum_i \sum_j (x_{ij} - x_{i0} - x_{0j} + x_{00})(y_{ij} - y_{i0} - y_{0j} + y_{00}) \end{aligned}$$

or, symbolically,

$$\text{total SP}_{xy} = B_{xy} + T_{xy} + E_{xy}.$$

Similarly, by replacing y by x or x by y

$$\sum_i \sum_j (x_{ij} - x_{00})^2 = t \sum_i (x_{i0} - x_{00})^2 + r \sum_j (x_{0j} - x_{00})^2 + \sum_i \sum_j (x_{ij} - x_{i0} - x_{0j} + x_{00})^2$$

or total  $SS_{xx} = B_{xx} + T_{xx} + E_x$  ;

and total  $SS_{yy} = B_{yy} + T_{yy} + E_{yy}$

The unrestricted residual SS obtained for the above model is

$$\begin{aligned} SSE^* &= \sum_i \sum_j (y_{ij} - \hat{\mu} - \hat{\alpha}_i^* - \hat{\theta}_j^* - \hat{\beta}(x_{ij} - x_{00}))^2 \\ &= \sum_i \sum_j (y_{ij} - y_{00})^2 - t \sum_i (y_{i0} - y_{00})^2 - (r \sum_j (y_{0j} - y_{00})^2) - \hat{\beta} E_{xy} \\ &= \text{total } SS_{yy} - B_{yy} - T_{yy} - \hat{\beta} E_{xy} \\ &= (\text{SSE for RBD}) - \hat{\beta} E_{xy}, \text{ with } df = (r-1)(t-1) - 1. \end{aligned}$$

Here  $\hat{\beta} E_{xy}$  is the reduction in error SS due to the regression of y on x.

Then

$$SSE^* = E_{yy} - \hat{\beta} E_{xy} \text{ with } df = (r-1)(t-1) - 1.$$

The null hypothesis to be tested is  $H_0$  : all  $\theta_j$  are equal, which means that the effects due to the folders after considering the regression of y on x are the same.

The restricted residual SS (i.e. the residual SS under  $H_0$ ) is

$$(SSE^*) = \text{minimum value of } \sum_i \sum_j [y_{ij} - \mu - \alpha_i - \beta(x_{ij} - x_{00})]^2$$

when minimised with respect to  $\mu$ ,  $\alpha_i$  and  $\beta$ .

$$= \sum_{i,j} (y_{ij} - y_{00})^2 - t \sum_i (y_{i0} - y_{00})^2 - \hat{\beta}^* E_{xy},$$

with  $df = r(t - 1) - 1$

$\hat{\beta}^*$  being the least-square estimator of  $\beta$  under  $H_0$  and being given by

where  $\hat{\beta}^* = E'_{xy} / E'_{xx}$ ,

$$E'_{xx} = E_{xx} + T_{xx},$$

$$E'_{yy} = E_{yy} + T_{yy},$$

and  $E'_{xy} = E_{xy} + T_{xy},$

The appropriate test statistics for testing  $H_0$  is

$$F_0 = \frac{(SSE^*) - SSE}{SSE^*} \times \frac{(r - 1)(t - 1) - 1}{(t - 1)}$$

and  $H_0$  is rejected at the level  $\alpha$  if the observed value of the above  $F_0$  exceeds  $F_{\alpha, (t-1), (r-1)-1}$ ; otherwise  $H_0$  is accepted at the level  $\alpha$ .

The corresponding analysis of covariance table is shown below :

**Table – 1**

**Analysis of Covariance for an RBD with One Concomitant Variable**

Source of variation	df	SS <sub>xx</sub>	SP <sub>xy</sub>	SS <sub>yy</sub>	Estimate of $\beta$	Adjusted	
						SS <sub>yy</sub>	df
Breed (B)	$r - 1$	B <sub>xx</sub>	B <sub>xy</sub>	B <sub>yy</sub>			
Fodder (T)	$t - 1$	T <sub>xx</sub>	T <sub>xy</sub>	T <sub>yy</sub>			
Error (E)	$(r - 1)(t - 1)$	E <sub>xx</sub>	E <sub>xy</sub>	E <sub>yy</sub>	$E_{xy}/E_{xx}$	SSE*	$(r - 1)(t - 1) - 1$
Fodder + error	$r(t - 1)$	E' <sub>xx</sub>	E' <sub>xy</sub>	E' <sub>yy</sub>	$E'_{xy}/E'_{xx}$	(SSE*)'	
Difference : SS due to Fodder after adjusting for age		-				(SSE*)' - SSE*	$t - 1$

**Table – 2**  
**Data Table**

Breed	Fodder						Total	
	F <sub>1</sub>		F <sub>2</sub>		F <sub>3</sub>			
	x	y	x	y	x	y	x	y
1	41	122	41	81	42	80	124	283
2	40	120	50	80	38	82	128	282
3	38	138	48	79	54	65	138	282
4	41	121	42	75	40	58	123	254
Total	160	501	179	315	174	285	513	1101

\*x values indicate age of cows in months, y values indicate yield of milk for 5 consecutive days.

(The data are fictitious)

The relevant computations are shown below :

$$T_{xx} = \frac{(160)^2 + (179)^2 + (174)^2}{4} - \frac{(513)^2}{12} = \frac{87917}{4} - \frac{263169}{12}$$

$$= 21,979.25 - 21,930.75 = 48.50$$

$$B_{xx} = \frac{(124)^2 + (128)^2 + (138)^2 + (123)^2}{3} - 21,930.75$$

$$= 65933 / 3 - 21,930.75 = 21,977.6667 - 21,930.75 = 45.9167.$$

$$\text{Total SS}_{xx} = (41)^2 + (40)^2 + \dots + (54)^2 + (40)^2 - 21,930.75$$

$$= 22,191 - 21,930.75 = 260.25$$

$$T_{yy} = \frac{(501)^2 + (315)^2 + (285)^2}{4} - \frac{(1101)^2}{12} = \frac{431451}{4} - \frac{1212201}{12}$$

$$= 107,862.75 - 101,016.75 = 6,846.00$$

$$B_{yy} = \frac{(283)^2 + (282)^2 + (282)^2 + (254)^2}{3} - 101,016.75$$

$$= 303653/3 - 101,016.75$$

$$= 101,217.6667 - 101,016.75 = 200.9167$$

$$\text{Total SS}_{yy} = (112)^2 + (120)^2 + \dots + (65)^2 + (58)^2 - 101,016.75$$

$$= 108,509 - 101,016.75 = 7,492.25$$

$$\text{Total SP}_{xy} = (41 \times 122) + \dots + (40 \times 58) - 513 \times 1101 / 12$$

$$= 46,418 - 564813 / 12 = 46,418 - 47,067.75 = - 649.75$$

$$T_{xy} = \frac{(160 \times 501) + (179 \times 315) + (174 \times 285)}{4} - 47,067.75$$

$$= 186135 / 4 - 47,067.75 = 46,533.75 - 47,067.75 = - 534.00$$

$$B_{xy} = \frac{(124 \times 283) + (128 \times 282) + (138 \times 282) + (123 \times 254)}{3}$$

$$- 47,067.75 = \frac{141346}{3} - 47,067.75 = 47,115.3333 - 47,067.75$$

$$= 47,5833$$

The SSs and SPs are entered in the following table

**Table 3**  
**ANCOVA Table**

Source of variation	df	SS <sub>xx</sub>	SP <sub>xy</sub>	SS <sub>yy</sub>	b	Adjusted	
						SS <sub>yy</sub>	df
Breed (B)	3	46.9167	47.5833	200.9167	-0.9909	283.4863	5
Fodder (T)	2	48.5000	-534.0000	6846.0000			
Error (E)	6	164.8333	-163.3333	445.3333			
Total	11	260.2500	- 649.7500	7492.2500			
Breed + error	8	213.3333	-697.3333	7291.3333	-3.2688	5011.8902	7
Difference : SS due to Fodder after adjusting for age			-			4728.4039	2

Since

$$F_0 = \frac{4728.4039 / 2}{283.4863 / 5} = \frac{2364.2019}{56.6972} = 41.6987$$

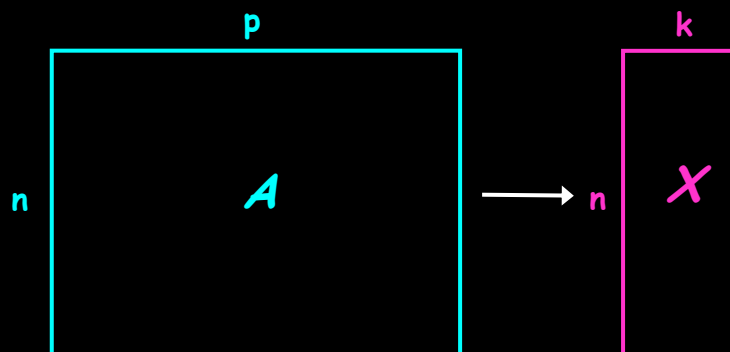
is greater than  $F_{.01;2,5} = 13.27$ , it would seem that there are real fodder differences after adjustment is made for the differences in the age of cows.

## Principal Component Analysis (PCA)

Asok K Nanda

### Data Reduction

- summarization of data with many ( $p$ ) variables by a smaller set of ( $k$ ) derived (synthetic, composite) variables.





## Data Reduction

- “Residual” variation is information in  $A$  that is not retained in  $X$
- balancing act between
  - clarity of representation, ease of understanding
  - oversimplification: loss of important or relevant information.

## Principal Component Analysis (PCA)

- probably the most widely-used and well-known of the “standard” multivariate methods
- invented by Pearson (1901) and Hotelling (1933)
- first applied in ecology by Goodall (1954) under the name “factor analysis” (“principal factor analysis” is a synonym of PCA).

## Principal Component Analysis (PCA)

- takes a data matrix of  $n$  objects by  $p$  variables, which may be correlated, and summarizes it by uncorrelated axes (principal components or principal axes) that are linear combinations of the original  $p$  variables
- the first  $k$  components display as much as possible of the variation among objects.

## Geometric Rationale of PCA

- objects are represented as a cloud of  $n$  points in a multidimensional space with an axis for each of the  $p$  variables
- the **centroid** of the points is defined by the mean of each variable
- the **variance** of each variable is the average squared deviation of its  $n$  values around the mean of that variable.

$$V_i = \frac{1}{n-1} \sum_{m=1}^n (X_{im} - \bar{X}_i)^2$$

## Geometric Rationale of PCA

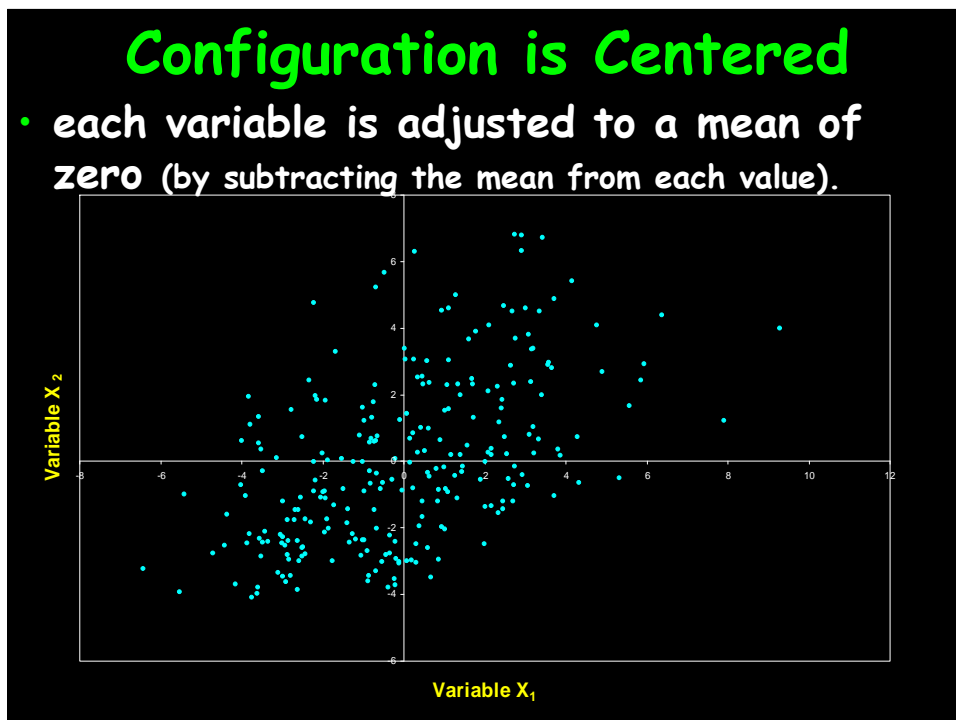
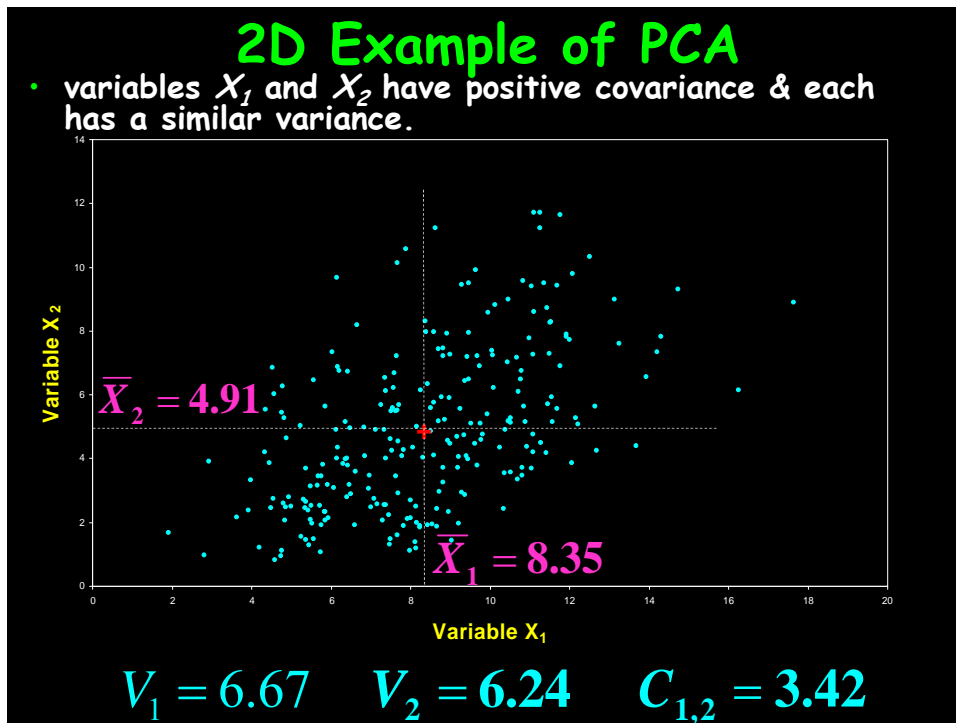
- degree to which the variables are linearly correlated is represented by their **covariances**.

$$C_{ij} = \frac{1}{n-1} \sum_{m=1}^n (X_{im} - \bar{X}_i)(X_{jm} - \bar{X}_j)$$

Covariance of variables  $i$  and  $j$   
 Sum over all  $n$  objects  
 Value of variable  $i$  in object  $m$   
 Mean of variable  $i$   
 Value of variable  $j$  in object  $m$   
 Mean of variable  $j$

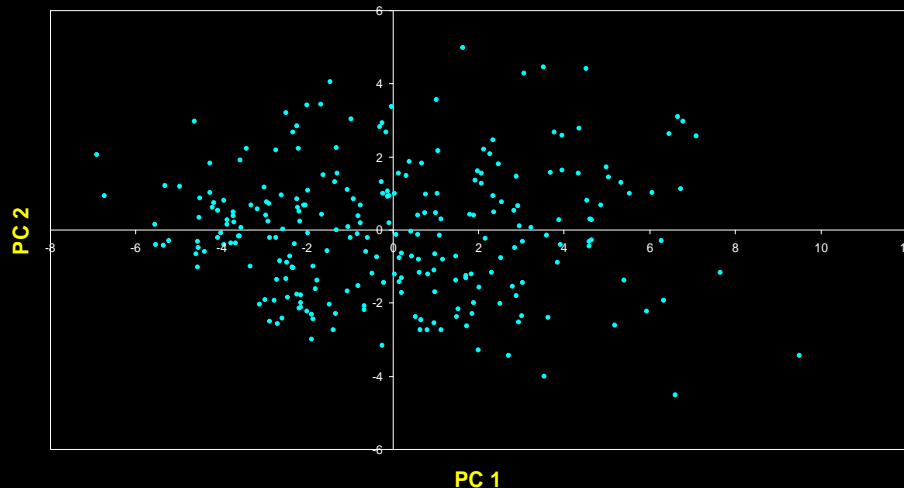
## Geometric Rationale of PCA

- objective of PCA is to **rigidly rotate** the axes of this  $p$ -dimensional space to new positions (**principal axes**) that have the following properties:
  - ordered such that **principal axis 1 has the highest variance**, axis 2 has the next highest variance, . . . . , and axis  $p$  has the lowest variance
  - covariance among each pair of the principal axes is zero (**the principal axes are uncorrelated**).



## Principal Components are Computed

- PC 1 has the highest possible variance (9.88)
- PC 2 has a variance of 3.03
- PC 1 and PC 2 have zero covariance.



## The Dissimilarity Measure Used in PCA is Euclidean Distance

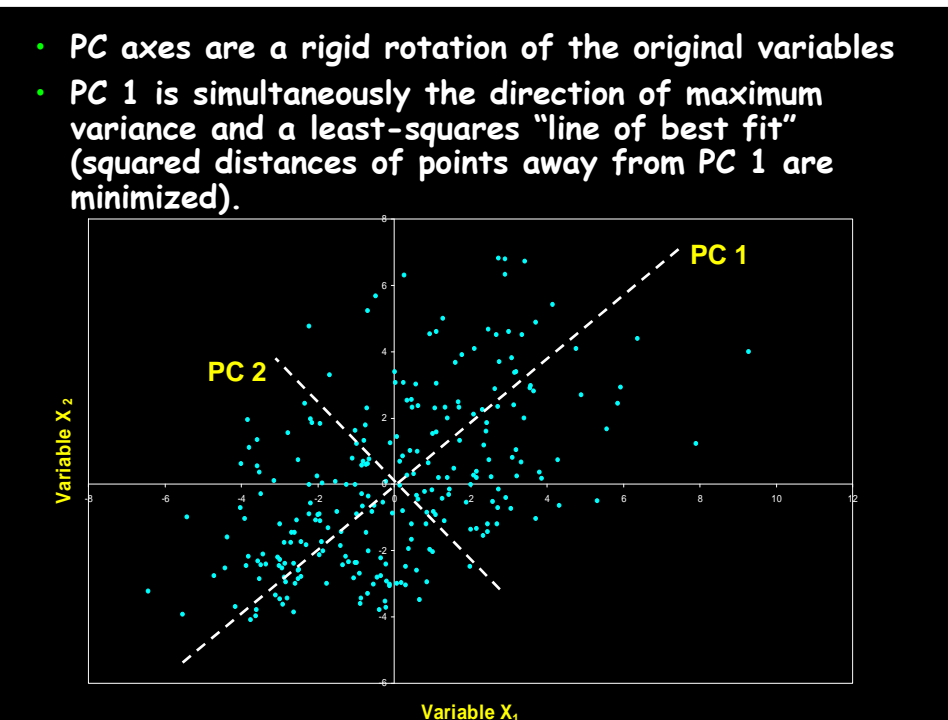
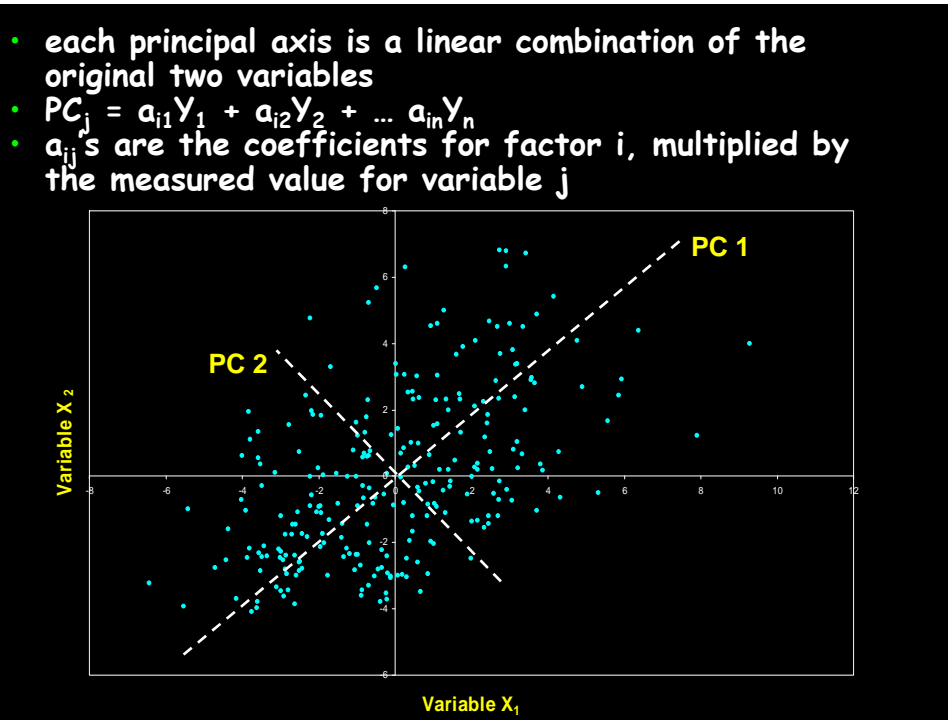
- PCA uses Euclidean Distance calculated from the  $p$  variables as the measure of dissimilarity among the  $n$  objects
- PCA derives the best possible  $k$  dimensional ( $k < p$ ) representation of the Euclidean distances among objects.

## Generalization to $p$ -dimensions

- In practice nobody uses PCA with only 2 variables
- The algebra for finding principal axes readily generalizes to  $p$  variables
- PC 1 is the direction of maximum variance in the  $p$ -dimensional cloud of points
- PC 2 is in the direction of the next highest variance, subject to the constraint that it has zero covariance with PC 1.

## Generalization to $p$ -dimensions

- PC 3 is in the direction of the next highest variance, subject to the constraint that it has zero covariance with both PC 1 and PC 2
- and so on... up to PC  $p$



## Generalization to $p$ -dimensions

- if we take the first  $k$  principal components, they define the  $k$ -dimensional "hyperplane of best fit" to the point cloud
- of the total variance of all  $p$  variables:
  - PCs 1 to  $k$  represent the maximum possible proportion of that variance that can be displayed in  $k$  dimensions
  - *i.e.* the squared Euclidean distances among points calculated from their coordinates on PCs 1 to  $k$  are the best possible representation of their squared Euclidean distances in the full  $p$  dimensions.

## Covariance vs Correlation

- using covariances among variables only makes sense if they are measured in the same units
- even then, variables with high variances will dominate the principal components
- these problems are generally avoided by standardizing each variable to unit variance and zero mean.

$$X'_{im} = \frac{(X_{im} - \bar{X}_i)}{SD_i}$$

Mean variable  $i$

Standard deviation of variable  $i$



## Covariance vs Correlation

- covariances between the standardized variables are **correlations**
- after standardization, each variable has a variance of 1.000
- correlations can be also calculated from the variances and covariances:

$$r_{ij} = \frac{C_{ij}}{\sqrt{V_i V_j}}$$

Correlation between variables  $i$  and  $j$  →  $r_{ij}$

Covariance of variables  $i$  and  $j$  →  $C_{ij}$

Variance of variable  $i$  →  $V_i$

Variance of variable  $j$  →  $V_j$

## The Algebra of PCA

- first step is to calculate the **cross-products matrix** of variances and covariances (or correlations) among every pair of the  $p$  variables
- square, symmetric matrix
- diagonals are the variances, off-diagonals are the covariances.

	$X_1$	$X_2$
$X_1$	6.6707	3.4170
$X_2$	3.4170	6.2384

Variance-covariance Matrix

	$X_1$	$X_2$
$X_1$	1.0000	0.5297
$X_2$	0.5297	1.0000

Correlation Matrix

## The Algebra of PCA

- in matrix notation, this is computed as

$$S = X'X$$

- where  $X$  is the  $n \times p$  data matrix, with each variable centered (also standardized by SD if using correlations).

	$X_1$	$X_2$
$X_1$	6.6707	3.4170
$X_2$	3.4170	6.2384

Variance-covariance Matrix

	$X_1$	$X_2$
$X_1$	1.0000	0.5297
$X_2$	0.5297	1.0000

Correlation Matrix

## Manipulating Matrices

- transposing: could change the columns to rows or the rows to columns

$$X = \begin{bmatrix} 10 & 0 & 4 \\ 7 & 1 & 2 \end{bmatrix} \quad X' = \begin{bmatrix} 10 & 7 \\ 0 & 1 \\ 4 & 2 \end{bmatrix}$$

- multiplying matrices
  - must have the same number of columns in the premultiplicand matrix as the number of rows in the postmultiplicand matrix

## The Algebra of PCA

- sum of the diagonals of the variance-covariance matrix is called the **trace**
- it represents the **total variance** in the data
- it is the mean squared Euclidean distance between each object and the centroid in  $p$ -dimensional space.

	$X_1$	$X_2$
$X_1$	6.6707	3.4170
$X_2$	3.4170	6.2384

Trace = 12.9091

	$X_1$	$X_2$
$X_1$	1.0000	0.5297
$X_2$	0.5297	1.0000

Trace = 2.0000

## The Algebra of PCA

- finding the principal axes involves eigenanalysis of the cross-products matrix ( $S$ )
- the eigenvalues (latent roots) of  $S$  are solutions ( $\lambda$ ) to the characteristic equation

$$|S - \lambda I| = 0$$

## The Algebra of PCA

- the eigenvalues,  $\lambda_1, \lambda_2, \dots, \lambda_p$  are the variances of the coordinates on each principal component axis
- the sum of all  $p$  eigenvalues equals the trace of  $S$  (the sum of the variances of the original variables).

	$X_1$	$X_2$
$X_1$	6.6707	3.4170
$X_2$	3.4170	6.2384

$$\lambda_1 = 9.8783$$

$$\lambda_2 = 3.0308$$

$$\text{Note: } \lambda_1 + \lambda_2 = 12.9091$$

$$\text{Trace} = 12.9091$$

## The Algebra of PCA

- each eigenvector consists of  $p$  values which represent the "contribution" of each variable to the principal component axis
- eigenvectors are uncorrelated (orthogonal)
  - their cross-products are zero.

### Eigenvectors

	$u_1$	$u_2$
$X_1$	0.7291	-0.6844
$X_2$	0.6844	0.7291

$$0.7291 * (-0.6844) + 0.6844 * 0.7291 = 0$$

### The Algebra of PCA

- coordinates of each object  $i$  on the  $k^{\text{th}}$  principal axis, known as the **scores** on PC  $k$ , are computed as

$$z_{ki} = u_{1k} x_{1i} + u_{2k} x_{2i} + \cdots + u_{pk} x_{pi}$$

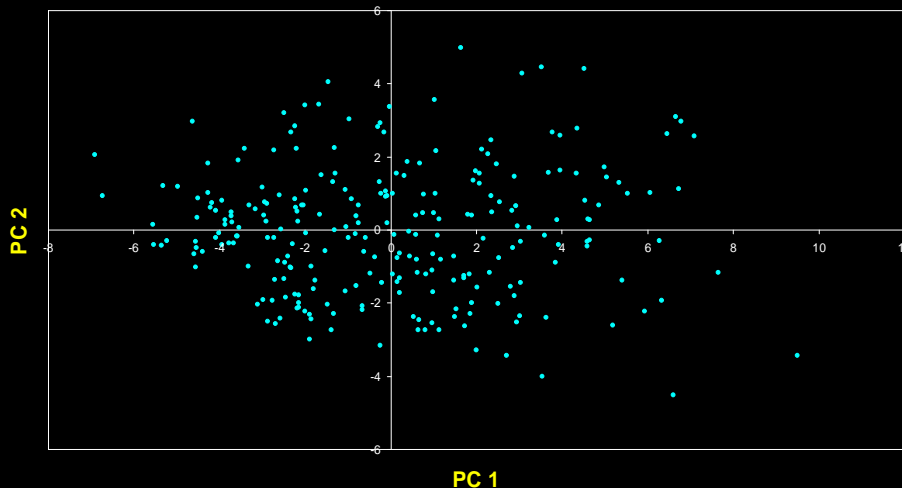
- where  $Z$  is the  $n \times k$  matrix of **PC scores**,  $X$  is the  $n \times p$  **centered data matrix** and  $U$  is the  $p \times k$  **matrix of eigenvectors**.

### The Algebra of PCA

- variance of the scores on each PC axis is equal to the corresponding eigenvalue for that axis
- the eigenvalue represents the variance displayed ("explained" or "extracted") by the  $k^{\text{th}}$  axis
- the sum of the first  $k$  eigenvalues is the variance explained by the  $k$ -dimensional ordination.

$\lambda_1 = 9.8783$     $\lambda_2 = 3.0308$    Trace = 12.9091

PC 1 displays ("explains")  
 $9.8783/12.9091 = 76.5\%$  of the total variance



## The Algebra of PCA

- The cross-products matrix computed among the  $p$  principal axes has a simple form:
  - all off-diagonal values are zero (the principal axes are uncorrelated)
  - the diagonal values are the eigenvalues.

	$PC_1$	$PC_2$
$PC_1$	9.8783	0.0000
$PC_2$	0.0000	3.0308

Variance-covariance Matrix  
of the PC axes

## A more challenging example

- data from research on habitat definition in the endangered Baw Baw frog
- 16 environmental and structural variables measured at each of 124 sites
- correlation matrix used because variables have different units



## Eigenvalues

Axis	Eigenvalue	% of Variance	Cumulative % of Variance
1	5.855	36.60	36.60
2	3.420	21.38	57.97
3	1.122	7.01	64.98
4	1.116	6.97	71.95
5	0.982	6.14	78.09
6	0.725	4.53	82.62
7	0.563	3.52	86.14
8	0.529	3.31	89.45
9	0.476	2.98	92.42
10	0.375	2.35	94.77

## Interpreting Eigenvectors

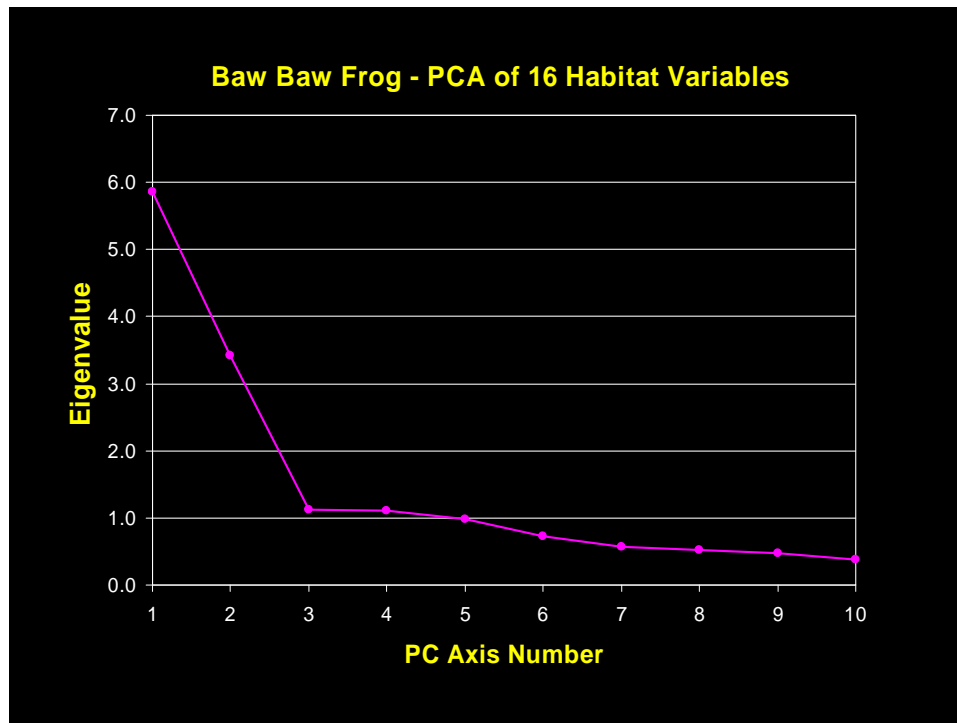
- correlations between variables and the principal axes are known as **loadings**
- each element of the eigenvectors represents the contribution of a given variable to a component

	1	2	3
Altitude	0.3842	0.0659	-0.1177
pH	-0.1159	0.1696	-0.5578
Cond	-0.2729	-0.1200	0.3636
TempSurf	0.0538	-0.2800	0.2621
Relief	-0.0765	0.3855	-0.1462
maxERht	0.0248	0.4879	0.2426
avERht	0.0599	0.4568	0.2497
%ER	0.0789	0.4223	0.2278
%VEG	0.3305	-0.2087	-0.0276
%LIT	-0.3053	0.1226	0.1145
%LOG	-0.3144	0.0402	-0.1067
%W	-0.0886	-0.0654	-0.1171
H1Moss	0.1364	-0.1262	0.4761
DistSWH	-0.3787	0.0101	0.0042
DistSW	-0.3494	-0.1283	0.1166
DistMF	0.3899	0.0586	-0.0175

## How many axes are needed?

- does the  $(k+1)^{th}$  principal axis represent more variance than would be expected by chance?
- several tests and rules have been proposed
- a common "rule of thumb" when PCA is based on correlations is that axes with eigenvalues  $> 1$  are worth interpreting



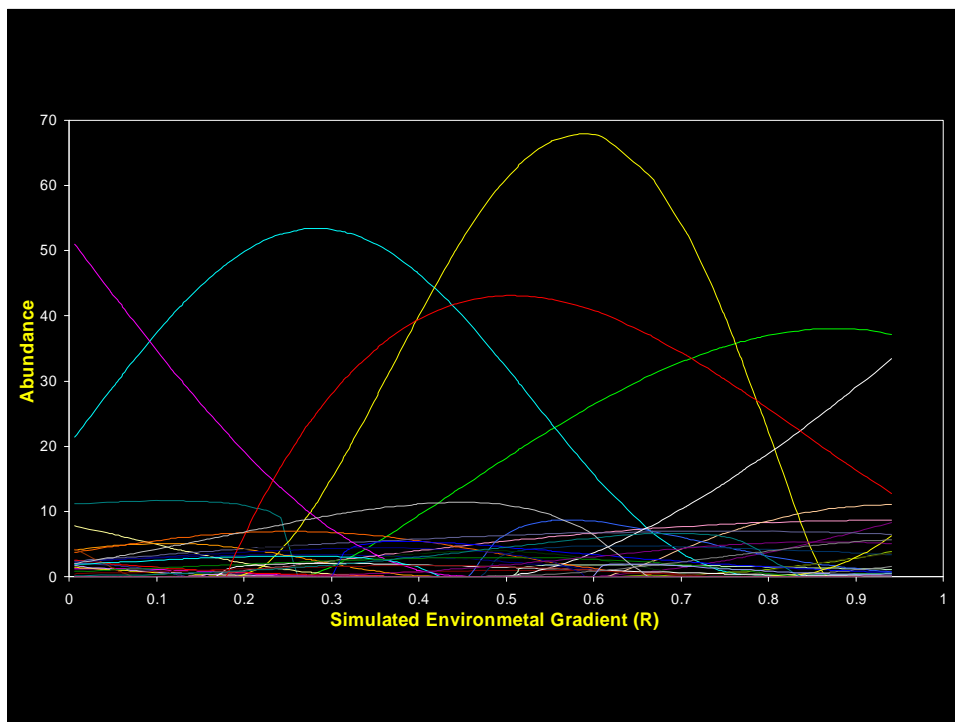


## What are the assumptions of PCA?

- assumes relationships among variables are **LINEAR**
  - cloud of points in  $p$ -dimensional space has linear dimensions that can be effectively summarized by the principal axes
- if the structure in the data is **NONLINEAR** (the cloud of points twists and curves its way through  $p$ -dimensional space), the principal axes will not be an efficient and informative summary of the data.

## When should PCA be used?

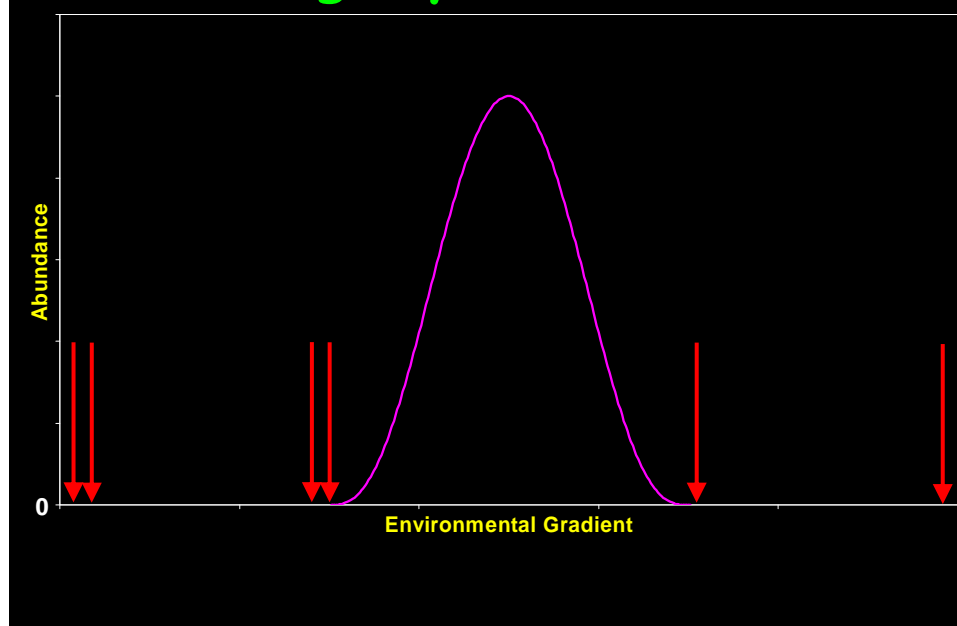
- In community ecology, PCA is useful for summarizing variables whose relationships are approximately linear or at least monotonic
  - *e.g.* A PCA of many soil properties might be used to extract a few components that summarize main dimensions of soil variation
- PCA is generally NOT useful for ordinating community data
- Why? Because relationships among species are highly nonlinear.

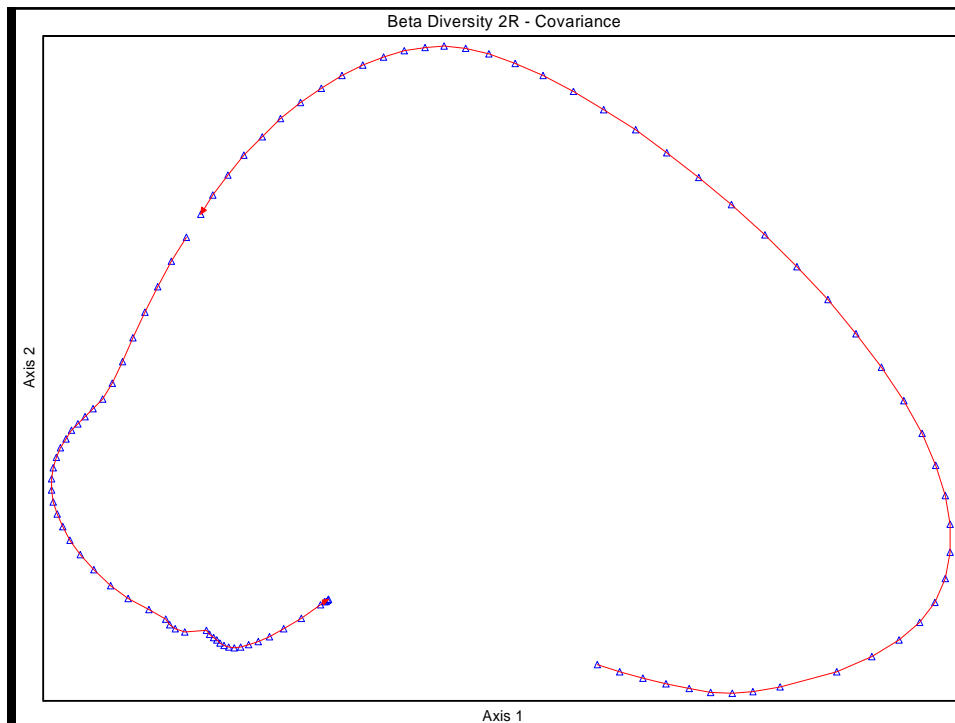


## The "Horseshoe" or Arch Effect

- community trends along environmental gradients appear as "horseshoes" in PCA ordinations
- none of the PC axes effectively summarizes the trend in species composition along the gradient
- SUs at opposite extremes of the gradient appear relatively close together.

## Ambiguity of Absence





## The "Horseshoe" Effect

- curvature of the gradient and the degree of infolding of the extremes increase with beta diversity
- PCA ordinations are not useful summaries of community data except when beta diversity is very low
- using correlation generally does better than covariance
  - this is because standardization by species improves the correlation between Euclidean distance and environmental distance.

What if there's more than one underlying ecological gradient?

### The "Horseshoe" Effect

- when two or more underlying gradients with high beta diversity a "horseshoe" is usually not detectable
- the SUs fall on a curved hypersurface that twists and turns through the  $p$ -dimensional species space
- interpretation problems are more severe
- PCA should NOT be used with community data (except maybe when beta diversity is very low).

## Impact on Ordination History

- by 1970 PCA was the ordination method of choice for community data
- simulation studies by Swan (1970) & Austin & Noy-Meir (1971) demonstrated the horseshoe effect and showed that the linear assumption of PCA was not compatible with the nonlinear structure of community data
- stimulated the quest for more appropriate ordination methods.

## **Bidhan Chandra Krishi Viswavidyalaya**

*Topic: Concepts & Methodological aspects in Farm Business Analysis*

**National Workshop cum Training Programme on Statistical  
Tools for Research Data Analysis**

**Organized**

**by**

**Society for Application of Statistics in Agriculture & Allied  
Sciences (SASAA) & Department of Agricultural Statistics,  
F/Ag, BCKV.**

*Speaker: Dr. A. K. Nandi, Head & Director in Charge, CCS,  
Ministry of Agriculture & Farmers' Welfare, GoI.*

*+91-9433174137, email: aknandibckv@rediffmail.com*

**Department of Agricultural Economics  
Faculty of Agriculture  
Bidhan Chandra Krishi Viswavidyalaya  
Mohanpur, Nadia.**

Dr. A.K.Nandi

## **Definition and Concepts of Economics**

**Wealth def. Adam Smith (1976): Science of *wealth*.**

**Welfare def. Alfred Marshall (1890): Study of *mankind* in  
the *ordinary business* of life.**

**Scarcity def. Lionel Robbins (1932): Economics is a  
science which studies human behavior as relationship  
between *ends* and *scarce means* which have *alternative  
uses*.**

- Science of choice
- Unlimited wants
- Scarce means
- Alternative uses of means

Dr. A.K.Nandi

### What we get from Economics

- **Problems of full employment of resources;**
- **Problems of allocation of resources;**
- **Problems of Methods of production**
- **Problems of Distribution of national income;**
- **Problems of Economic efficiency;**
- **Problems of Economic growth;**
- **Problems of Unemployment;**
- **Problems of Foreign exchange;**
- **Problems of Trade cycle;**

Dr. A.K.Nandi

#### Production

- What to produce?
- How to produce?
- How much to produce?
- For whom to produce?

#### Marketing

1. When to buy & sell?
2. Where to buy & sell?
3. How much to buy & sell?

Dr. A.K.Nandi



### *Arena of Farm Business Analysis*

- Agricultural production economics is an applied field of science, wherein, the principles of choice are applied to the use of capital, labour, land and management in farming industry.
- Study of Resource efficiency in production process (productivity)
  - ✓ Resource use
  - ✓ Resource allocation
  - ✓ Resource combination
  - ✓ Resource management
  - ✓ Resource administration
  - ✓ Resource efficiency

Dr. A.K.Nandi

### Objectives

- I. To determine and outline the conditions which gives the optimum use of land, labour, capital and management in the production of crops and livestock.
- II. To determine the extent to which the existing use of resources deviates from optimum use.
- III. To analyses the factors ,which condition production pattern and resource use.
- IV. To explain means and methods in getting the optimum use of resources from the existing ones.

Dr. A.K.Nandi

### Concepts

- **The firm** is a decision making unit or managerial unit (technical unit) of production.
- **A farm** is a firm which combines resources in the production of agricultural products on the lines of a business firm, i.e. Profit maximisation.
- **Economic unit and technical unit:** The application of input or measurement of output relate to a technical unit or an economic unit. Technical unit refers to a single fixed unit in production , for which technical coefficient are calculated, e.g. An acre of land or a unit of poultry birds, etc.
- Economic unit refers to aggregation of resources , for which cost & returns are worked out as a whole (economic returns) e.g. Farm holding.

Dr. A.K.Nandi

### Farm Management

- Improving practices on existing enterprises
- Reorganising existing & including new enterprises
- Determining time horizon of production
- Adopting farm practices for immediate & long term basis
- Determining the best size of farms
- Deciding capital & operational inputs
- Marketing opportunities
- Expectation of input & output prices
- Credit requirements

Dr. A.K.Nandi

### Farm Management Decisions

- **Strategic Management Decisions:** Size of the farm  
Programme on machinery & inputs Construction Irrigation  
conservation & reclamation
- **Operational Management Decisions:** Day-to-day
- **Administrative Decisions:**
  - Financing
  - Supervision
  - Accounting & book keeping
  - Policy adjustments
  - Home consumption & Market disposal
- **Marketing Decisions(when, where, how):**
  - Buying & Selling

Dr. A.K.Nandi

## FARM RESOURCE MANAGEMENT

Dr. A.K.Nandi

## Farm Resource Management

### Land management

Factors to be considered in Selecting a farm

- **Physical Factors:**
  - i) Climate, ii) Rainfall, iii) Topography, iv) Soil, v) water supply, vi) Drainage.
- **Economic Factors:** a) Transport and market facilities, b) institutional facilities, c) local taxes, d) land values and productivity
- **Social Factors:** i) accessibility to school and hospitals, ii) Type of neighbour and community, iii) Tradition and customs.

Dr. A.K.Nandi

### Size of farm

It is generally understood in terms of i) *physical area*, ii) volume of production and iii) *value of production*.

- **Factors affecting the size of the farm:**  
**Financial resources, Density of population, Climate Topography, Nature and sources of irrigation, Managerial capacity, Law of inheritance, State laws**

#### **## Value of a Farm and its estimation:**

- ❖ Farm Layout
- ❖ Cropping pattern (climate, population & labour availability, consumption habit, institutional arrangements, transport, communication, markets, etc.) & Cropping Scheme:
- ❖ Physical factors
- ❖ Economic factors
- ❖ Personal like & dislike

Dr. A.K.Nandi

### Value estimation (Land)

a) **Income capitalization method:** The value of land and any other asset on its current income in relation to the prevailing rate of interest is called capitalized value.

✓ **Methods:** Annual net income = Gross income - total cost. Rate of interest is a capitalisation rate ( $r$ ). When income is constant, the formula is  $V = I/r$ ,  $V$  = capitalised value of land,  $I$  = Net income / return per year,  $r$  = Rate of interest, When income rise or fall at a constant arithmetic rate then,  $V = (I/r) + / - (I/r^2) + / -$

$V$  = Capitalised value of land.

a) Comparison method:

b) Sale price method: Market rate of lands.

Dr. A.K.Nandi

### Types & Systems of Farming

- ✓ Crop & Livestock raising
- ✓ Mode of Economic & Social functioning

Farming is classified into two groups:

a) Types of Farming (enterprise & income):

- 1) **Specialised farming:** - 50% or more income is derived from one single source that means only one commodity in the market - single source of income
- 2) **Diversified farming:** no single product source of income equals as much as 50% of the total receipts is called as diversified or general farm.
- 3) **Mixed farming:** combination of crop production with a significant amount of livestock raising. It is also a type of diversified farming invariably devotes to livestock production as a complementary enterprise.
- 4) **Ranching:** natural vegetation & multiply under natural surroundings. Mainly public grazing lands.
- 5) **Dry farming:** Rainfed with scanty rainfall (20" or less) conservation of soil moisture.

b) System of farming(?)

Dr. A.K.Nandi

<b>Advantages &amp; Disadvantages of Specialised farming</b>		
Sl.no	Advantages	Disadvantages
1.	Better use of land	Other factors are not fully utilised.
2.	Better marketing	Risk in market & crop failure, returns are uneven in a year.
3.	Better management & confined on crop centered technology	Productive resources are not fully utilised & hampered the resource centered technology.
4.	Less equipment & labour	By product of the farm cannot be fully utilised
5.	Costly & efficient machinery can be used	Soil health can not be maintained & question of sustainability.
6.	Efficiency & skill are increased	Knowledge of farm enterprises become limited

Dr. A.K.Nandi

### Factors determining the type of farming

• **Two type of factors:**

a) **Physical**- Climate, soil, topography, etc.

b) **Economic**:-

i ) Marketing cost:

ii) Changes in relative values of farm products:

iii) Availability of labour & capital:

iv) Land values:

v) Cycles of over & under production:

vi) Competition between enterprises:

vii) Personal like & dislikes:

viii) Prevalence of pest & diseases:

Dr. A.K.Nandi

## Systems of farming

- **Systems of farming** is generally referred to the **method of agriculture** and the **type of ownership of land**. If the farming has been classified on the basis of **economic and social functioning**, it is called as systems of farming.
- ✓ Co-operative farming:
  - Cooperative better farming
  - Cooperative joint farming
  - Cooperative tenant farming
- ✓ **Collective farming**: Members worked together under a management committee elected by themselves.
- ✓ **Capitalist farming**: Landlordism and more capital intensive and profit oriented systems.
- ✓ **State farming**: State owned farming managed by Govt. officials.
- ✓ **Peasant farming**: family farming

Dr. A.K.Nandi

## Systems of farming

	Systems of farming	Type of ownership	Type of Operationship
1.	<b>Cooperative better farming</b>	Individual	Individual
	<b>Cooperative joint farming</b>	Individual	collective
	<b>Cooperative tenant farming</b>	collective	Individual
	<b>Cooperative collective farming</b>	collective	collective
2.	<b>Collective farming</b>	Society/State	Society/State
3.	<b>Capitalist farming</b>	Individual	Individual
4.	<b>State farming</b>	State	
5.	<b>Peasant farming</b>	Individual	Individual

Dr. A.K.Nandi

## Labour Management

Dr. A.K.Nandi

### Classification of Farm Labour

- Farm labour is classified into:
- Unpaid labour: Farmers own labour ,Family labour
- Paid labour: Permanent or attached labour, Casual hired labour/seasonal labour.
- In Bengal : Attached farm labour, Family labour, Casual hired labour, Contractual labour.
- Skilled labour ,Unskilled Labour
- Unit: man days.

Dr. A.K.Nandi



## Farm Business Analysis

Dr. A.K.Nandi

## Farm Business Analysis

**Farm business analysis is the name given to a technique based on computation and interpretation of a variety of efficiency measures for the farm under study**

**Farm Accountancy:** Science of recording books of business transaction

**Farm Book Keeping:** system of records

Dr. A.K.Nandi

## Farm Records and Accounts

With the help of farm records, the farmer knows:

- ❖ Which enterprises are making profit or losing money, i.e. to check on performance of different enterprises.
- ❖ Which enterprises are returning most over his capital investment.
- ❖ To guide future decisions.
- ❖ Whether to go in for specialization or diversification.
- ❖ To provide planning data for use in making or revising future plans.
- ❖ Whether addition of new activities will boost the rate of return on his capital.

Dr. A.K.Nandi

## Characteristics & Advantages of Good Records System

### Characteristics

- Easy to keep and be up to date.
- Give needed information for analysis.
- Provide the information when needed and serve a definite purpose.
- Permit the analysis of the information needed.

### Advantages

- Means to higher income
- Basis for diagnosis and planning
- Way to improve the managerial ability of the farmer
- Basis for credit acquisition and management
- Guide to better management and future decisions
- Basis for research
- Basis for policy formulation

Dr. A.K.Nandi

## Problems and Difficulties in Farm Accounting in India

- Subsistence nature of farming.
- Triple role of Indian farmer and difficulties of maintenance, farm manager, farm labour and family head.
- Illiteracy and lack of business awareness.
- Complicated nature of agri- business.
- Inadequate extension service in making farmers record oriented.
- Non availability of suitable and simplified farm record books under Indian conditions.
- Lack of record consciousness.
- Fear of taxation.

Dr. A.K.Nandi

## Types of Farm Records

Farm records system consists of three parts:

1. **Physical farm records:** Farm map, Land utilization records ,Production and disposal record for crops, livestock, poultry and others, Labour records, Machinery use records,Feed records, Stock and store register
2. **Financial farm records:** Firm Inventory, Farm cash or farm financial record, capital asset & sale register, cash sale register, credit sale register, purchase register, paid in kind register, wage register, borrowed & repayment register, non-farm income records, etc
3. **Supplementary farm records:** Sanction records, Auction records, Rainfall records, Hire register, Stationary register

Dr. A .K.Nandi

## Farm Efficiency measures

- Efficiency can be related to: *Operation of farm business as a whole, individual phase of business/ use of various resources.*
  - **Physical efficiency measures(technical)**
    - a) *Aggregate measures-* total area of the farm/ no. of livestock/total production.
    - b) **Ratio measures:**
      - i) **Land use efficiency:-** *yield/production/ crop yield index/intensity of cropping/percentage of land under selected crops.*
      - ii) **labour efficiency:** *crop acentage /man, /work equivalent*
- Machinery efficiency: horse power

Dr. A.K.Nandi

## Value efficiency measures (financial)

- Aggregate measures
- Ratio measures:

Dr. A.K.Nandi

# Farm Planning and Budgeting

Dr. A.K.Nandi

## **Farm Planning and Budgeting**

**Planning:** *Deliberate & conscious effort on the part of the farmer to think about farm programs in advance and adjust them according to new knowledge on technical development, changes in physical and economic situation, price structure etc.*

### **Farm Planning:**

*Adoption of business method* in every phase of farm activity.

**Actions:** *Integrated, coordinated and advance programme of actions which seek to present an opportunity to cultivators to improve his level of income.*

### **Why Farm Planning:**

*Underutilization or overutilization proper allocation of resources to obtain maximum net income and product.*

**Objective:** *Maximise net income sustained over a long period.*

Dr. A.K.Nandi

## Types & Stages of Farm Planning

### Types

1. **Simple farm Planning**- Part of land/for one enterprise/subsidy
2. **Complete farm Planning**- Whole farm/all enterprise/farm suss Organisation

### ***Essential elements of farm Planning:-***

Manipulation of limited resources among alternative opportunities in order to validity the net objectives of minimising profit.

### ***Three essents-----***

- a) An objectives b) Scarce resources c) alternative ways.

### **Stages of Farm Planning:-**

- ❖ Stage-1 : Adoption of Package of Practices (selected enterprise)
- ❖ Stage-2: Extension of stage 1 to all crop enter. no major changes.
- ❖ Stage -3: Major change structure & Organisation training required for farm Management.

Dr. A.K.Nandi

## Basic steps of Farm plan & Budgeting

- ✓ **Assessment of resources**
- ✓ Analysis of existing plan operations.
- ✓ **Identification of problems in pre cultivation/plants.**
- ✓ Discussion with other farmers/specie.
- ✓ Selection of final plan for unifrementation.

Dr. A.K.Nandi

## Fundamentals of Farm business management

- **Farm Budgeting:**

The budget is essentially a presentation of costs and returns accompanied by a statement showing the physical quantities of inputs and output associated with each value figure.

- **Objective:**

Measure the *returns expected* from the plan. Farm budgeting is a process of *estimating costs, returns, and net profits of a farm* or a particular enterprise. Planning and budgeting go side by side.

### Three common objectives of farm budgeting :

- To estimate the *profitability* of a particular pattern of Organisation.
- To determine the *change in profits* that are likely to follow a particular change in Organisation.
- To compare different *Organisational pattern* or *alternative changes* in Organisation on a profit basis.

Dr. A.K.Nandi

## Types of Farm budgeting

### **Two ways:**

- Partial budgeting(enterprise budgeting)
- Complete budgeting(Full budgeting)

- **Partial budgeting:**

It refers to estimating the outcome or returns for a *part of the business*, i.e one or a few activities.

*Small modifications* have to be made to the existing Organisations.

Increased revenue and added costs associated with the proposed change in Organisation on the credit side, and the details of increase in costs and reduction in revenue on the debit side.

The extra profit or likely to arise out of the proposed plan, is then desired from the difference between two totals.

- **Complete budgeting:**

This method is used to make out a plan for the whole farm i.e. extensive remodelling of the farm Organisations. In preparing the complete budget, all the physical data are included and all costs and receipt items have to be calculated.

Dr. A.K.Nandi

### Six main steps:

- *Listing available resources and stating objectives.*
- *Estimating crop areas and livestock numbers.*
- *Estimating physical inputs and outputs.*
- *Estimating factor and product prices, calculating cost and returns.*
- *Estimating fixed costs.*
- *Totals and layout of budget.*

Dr. A.K.Nandi

### How to work out a practical budget

**Four questions** are important in practical budget. Two of which relate to **financial losses** arising from the contemplated change (**Debit side**) and two of which relate to the consequential **financial gains** (**credit side**).

- A. Debits: What losses of present revenue occur?  
What extra new costs are increased?
- B. Credits: What extra new revenue is obtained?  
What present lost are no longer required?

#### **Elements of partial budget:**

- ❖ Added costs.
- ❖ Added returns.
- ❖ Reduced costs.
- ❖ Reduced returns.

(Total of added returns + reduced costs) – total of added costs + reduced returns) = Net income from the change made in the farm Organisation by partial budgeting.

Dr. A.K.Nandi



### Evaluation of present farm situation.

- **Resource position:** i) Land ii) Labour iii) Cattle or mechanical power availability. Iv) Capital V) Organisation vi) Irrigation vii) Other information.
- **Crop grown:** i) Risk in the farm Production. li) Weak points in the existing plan iii) Alternate farm plan.
- **Cause:** two factors: 1) time 2) use

Dr. A.K.Nandi

### Farm Inventory Analysis

Farm inventory is a List of all physical properties of a business along with their value at a particular point of time. Appreciation & depreciation are very common for the inventory based on the nature of asset and time.

**Asset:** It is a property with right and value and may be classified as **a) Fixed** - land ,building etc; with longer life time and used throughout the production /business process; **b) Working-** more liquid than land and building, eg. farm machinery, and equipments, breeding stock etc; and c) **Current-** cash in hand and other short-lived and generally used within a year.

**Methods of Valuation:** **a)** Cost minus depreciation, **b)** Cost or market price whichever is lower, **c)** Net selling price, **d)** Replacement cost minus depreciation, **e)** Income capitalisation method( $V=I/r$ ;  $V$ =value in rupees,  $I$ =net income per year,  $r$ =rate of interest).

**Depreciation:** The decline in value of capital equipment due to wear and tear is called depreciation.

**Depreciation:** Two categories

1. **Physical condition:** a) Wear and tear, b) Physical decay, c) Accidental, d) Deferred maintenance.
2. **Financial condition:** a) Inadequacy (reduction in efficiency), b) Obsolescence.

Dr. A.K.Nandi

### Methods of computation of depreciation

- ❖ ***Straight line method***- Annual dep.: Original cost-Junk value/ Expected life of the Asset
- ❖ ***Annual revaluation method***: Asset value at the beginning - Asset value at the end.
- ❖ ***Diminishing balance method***:
- ❖ ***Sun of the year-Digits method or reducing fraction method***:
- ❖ ***Compound interest method***: a) Sinking fund method. B) Annuity charging method.
- ❖ ***Insurance policy method***:
- ❖ ***Machine hour- basis method***:

Dr. A.K.Nandi

### Diminishing balance method:

Fixed rate of depreciation used for every year and applied to the value of the asset at the beginning of the year.

First Year      Original cost/Expected life =  $1000/10 =$   
Rs 100.

Rs. 100 is the 10% of the 1000/-

Rs  $1000-100=$  Rs.900 at the end of the year  
depreciated value.

2<sup>nd</sup> year. 10% of Rs 900/ =Rs 90

At the end of the 2<sup>nd</sup> year.  $900-90=$ Rs. 810.

Dr. A.K.Nandi

### Sinking fund method

In this system, a depreciation fund equal to the actual loss in the value of the assets of machine is estimated, taking into account, the interest on the so accumulation fund.

Let D= Let of depreciation/year

R= Rate of interest on accumulated fund in fraction number.

C= Total cost of maching

S= Junk Value.

N= No. of years.

$$D = \frac{R(C-S)}{(1+R)^N - 1}$$

Original/ purchase price 40,000/- life spen 15years.

Scrap/Junk Value 15,000/-

Rate of interest on depreciation fund is changed at 5%

$$D = \frac{0.05(40,000-15,000)}{(1+0.05)^{15} - 1} = \frac{1250}{2.08 - 1} = \frac{1250}{1.08}$$

Dr. A.K.Nandi

### Sum of the Year –Digits method (Reducing fraction method).

The annual depreciation is found out by multi physical a fraction timer the amount to be depreciated. (Cost salvage value)

Fraction of any year =  $\frac{\text{the years of life remaining at the beginning of accountability period.}}{\text{Sum of the years of the life of the assets.}}$

$$\text{Fraction for 1}^{\text{st}} \text{ year} = \frac{10}{(1+2+\dots+10)} = \frac{10}{55}$$

$$\text{2}^{\text{nd}} \text{ year } \frac{9}{(1+2+\dots+10)} = \frac{9}{55}$$

Rate of annual depreciation (Original cost-Junk value) x fraction for the particular year.

The value of the fracter is Rs 92,000/- life spem -10 years

$$\frac{92,000 - 92000 - 9200}{55} \times \frac{10}{55} = \frac{15054.55}{92000-15,054.55}$$

$$\frac{(92000-9200) \times 9}{55} = 13,549.09 \text{ Annual depreciation.}$$

$$76,941.55 - 13549.09 = 63,396.46.$$

Dr. A.K.Nandi

### **Cost concepts in Farm Management studies in India (Continued)**

*Cost concepts approach to farm costing is widely used in India. These cost concepts in brief, are Cost  $A_1$ ; Cost  $A_2$  Cost B and Cost C. the difference cost items that are to be included under each cost concepts are detailed below with their imputation procedures and examples.*

**Cost  $A_1$** : Consists of following items-

**Labour** : Casual hired labour, Attached Farm labour, Contractual labour,

**Bullock/Machine labour**

Hired bullock labour, Imputed value of own bullock labour, Hired machine labour. Imputed value of own machine labour.

**Material/other costs:**

Seeds, Manures and Fertilizers, plant protection chemicals, Irrigation charges, Interest on working capital, Depreciation charges, Land Revenue.

Dr. A.K.Nandi

### **Cost Concept (S. R. Sen's Committee)**

- **Cost  $A_1$** : All actual expenses in cash and kind incurred in production by owner operator.
- **Cost  $A_2$** : Cost  $A_2$ +rent paid for leased in land.
- **Cost  $B_1$** : Cost  $A_1$ +interest on value of owned capital assets (excluding land) and rent for leased in land.
- **Cost  $C_1$** : Cost  $B_1$  + imputed value of family labour.
- **Cost  $C_2$** : Cost  $B_2$  + imputed value of family labour

Dr. A.K.Nandi

## Cost $A_1$ & $A_2$

### Cost $A_1$

1. Casual hired labour
2. Attached labour
3. Hired bullock labour
4. Imputer value of own bullock labour
5. Hired machine labour
6. Imputed value of owned machine labour
7. Seeds
8. Manures & fertilizers
9. Plant protection chemicals
10. Irrigation charges
11. Interest on working capital
12. Depreciation
13. Land revenue

Cost  $A_2$  = Cost  $A_1$  + rent paid for leased in land.

Dr. A.K.Nandi

### **Cost concepts in Farm Management studies in India**

**Cost  $A_2$ :** Cost  $A_1$  + Rent paid for leased in land, if any.

**Cost B:** Cost  $A_1$  + imputed rental value of own land + interest on owned fixed capital.

**Cost C:** Cost B + Imputed value of family labour.

Cost C is the total cost of cultivation or gross cost.

Dr. A.K.Nandi

Rates of Return over different cost concepts

- **Gross return:** Value of the main product + Bye products. ( Prices of product has to be computed with average prices prevailed in the market at the point of time).
- **Farm business income:** Gross income –Cost A<sub>1</sub>
- **Family labour income:** Gross income –Cost B
- **Net Income:** Gross income –Cost C
- **Farm investment income:** Farm business income – wages of family labour.
- **Cost of production** = (Gross cost – value of by product)/ output.

Dr. A.K.Nandi

**Structure of different costs and their components**

Cost A2	=	Cost A1 + Rent Paid for leased in-land
Cost B1	=	Cost A1 + Interest on value of owned fixed capital assets (excluding land) Cost
Cost B2	=	Cost B1 + Rental value of owned land (net of land revenue) and rent paid for leased-in land
Cost C1	=	Cost B1 + imputed value of family labour
Cost C2	=	Cost B2 + Imputed value of family labour Cost C2*
Cost C2*).	=	Cost C2 + Additional value of human labour based on use of higher wage rate in consideration of statutory minimum wage rate. (This is an intermediate concept
Cost C3 =	=	Cost C2* + 10 percent of cost C2* to account for managerial input of the farmer
<p>Dr. A.K.Nandi</p> <p><b>Imputation procedure being currently used by Gov.</b></p>		

## Sampling Design as followed by DES

- Three stage Stratified random Sampling:
- First(Primary Stage): Tehsils (Block) based on different zones (Agro climatic) as probability proportional to area with replacement system under crop complex approach.
- Crop Complex:
- Second: Village/cluster of villages: Same procedure(200 farm families) If not covered then from second or third village)
- Third(ultimate): Operational Holding – Five Size groups ; 2 from each (< 1 ha., 1-2 ha, 2-4 ha, 4-6 ha, Above 6 ha. )

Dr. A.K.Nandi

### Agricultural Statistics West Bengal(BAE&S)

- **Objectives:**  
Estimates of Block wise productivity & production of 19 Major Crops grown in West Bengal.
- **Two purposes:**  
(1) to provide data of total production & productivity to the GoWB;  
(2) to provide block wise yield rate of selected Insured Crops to the Agriculture Department & Insurance Companies for coverage to the affected farmers/cultivators as a part of welfare measures.
- **Estimation of consumption of cereals in West Bengal started since 1977 to asses the quantity of Marketable Surplus of Rice and Wheat in 18 districts and Potato, Mustard and Maskalai in some selected districts of West Bengal as per requirement of Food and Supply Department, Government of West Bengal.**  
**Methodology: Crop cutting Experiments.**

Dr. A.K.Nandi

### Crop Cutting Experiments

- A stratified multi-stage sampling design was adopted for the yield estimation survey of 19 major crops in the state.
- Block as stratum, the first stage sampling unit was mouza, the second stage sampling unit was a plot with specified crop and the third and ultimate stage of sampling unit was a piece of land of specified shape and size, called 'cut' based on random sampling.
- Systems: A circular area 100 sq. ft. (i.e., 9.29 m<sup>2</sup>) divided into three concentric circles of radii 2 ft. (i.e., 60.96 cm.), 4 ft. (i.e., 121.92 cm.) and 5.625 ft (i.e., 171.45 cm.) respectively.
- The cut for the Mung, Arhar & Sugarcane was a square area of 225 sq. ft. (i.e., 20.9032 m<sup>2</sup>).
- Each A.I. is supplied with a crop cutting apparatus which is used to demarcate the area of a cut (for circular cuts) for performing the crop cutting experiments and R.S. maps and random number tables are made available to him for identification of plot and to locate the random point wherein the cut is take place.
- The *driage factor*, i.e., the ratio between the weight of the freshly harvested crop and that of the crop after it is dried, in respect of all crops . 500 grams of crop produced in a CCE was kept separated and allowed to dry for a number of days till the weight becomes stable.
- The weight of crop taken after getting it dry is called 'dry weight'. For Jute and Mesta driage experiments were performed not on CCEs but on separate samples drawn for purpose in proportion to area under the crop in the districts.
- After properly scrutinizing & codifying the estimates of block wise Yield Rates & its Standard Error are calculated by the estimation procedure meant for the purpose.

Dr. A.K.Nandi

### Season wise Time of harvesting-CCE (Crop year: 1 st July to 30<sup>th</sup> June)

Season	Crops	Duration
Bhadui	Jute, Autumn Paddy, & Maize	July to October
Winter	Winter Paddy, Mesta & Maskalai	November and December
Rabi	Wheat, Gram, Lentil , Pea, Khesari, Arhar, Linseed, Mustard & Sugarcane	January to March
Summer	Summer paddy, Til, Mung & Groundnut	April to June

Dr. A.K.Nandi



### Methodological aspects for perennials

• **TIME VALUE OF MONEY**

**Compounding factor for 1-** What an initial amount becomes when growing at compound interest.

$$s_n = (1+i)^n \quad S = P (1+i)^n$$

*S=Future sum, P= principal amount or initial amount, i+ rate of interest or compounding factor, n= number of years*

Utility:

- **Discount Factor-**How much 1 at a future date is worth today (Present worth of 1).

$$v_n = 1/(1+i)^n$$

- **Utility:** This factor permits determining the value today of an amount received or paid out in the future. The process of finding the present worth in project preparation generally referred to as discounting).

*Discount factor is the reciprocal of the compounding factor. The most common use of discounting in project evaluation is to find the present worth of future costs or future benefits.*

- **Example:** The net benefit of minor irrigation scheme in Nadia in the 14<sup>th</sup> year of the project was estimated to be Rs.173,831/ m. Discounted at an interest rate of 21%, what is the present worth at the beginning of the project?
- **ANS:** Rs.173831 /m. (actual value at the future time) X 0.069 (21% discount factor for 14<sup>th</sup> year) = Rs. 11994 /m.

Dr. A.K.Nandi

### COMPOUNDING FACTOR FOR 1 PER ANNUM

**Growth of equal year end deposits all growing at compound interest.**

$$s_n = (1+i)^n - 1/i \text{ (Recurring deposit).}$$

**Utility:** This factor permits calculating the value to which a constant amount deposited at the end of each year will grow by the end of a stated year at a stated interest rate.

- **Example:** On December 31<sup>st</sup> of each year for 9 years the amount of Rs.782 is deposited in to an account that earns 6% interest. What will be the balance at the end of the 9<sup>th</sup> year?
- **Ans:** Amount deposited at the end of each year from 1<sup>st</sup> to 9<sup>th</sup> years X 11.491 (6% compounding factor for 1 per annum for 9 years)= Rs. 8986/

Dr. A.K.Nandi

### **PRESENT WORTH OF AN ANNUITY FACTOR**

How much 1 received or paid annually is worth today. (Present worth of 1 per annum, discount factor for a stream of income).

$a_n = 1 - v_n/i$   $v_n =$  discount factor,  $n =$  number of yrs.,  $i =$  rate of interest.

**Utility:** This factor enables determination of the present worth of a constant amount received each year for some length of time in the future. Present worth of an annuity factor for a given number of years is the total discount factor for all yrs. through the last.

- **Example:** At the time of retirement a government employee who invests in a deposit scheme with an expectation that his daughter may expect to earn an incremental benefit of net income of Rs. 44150/year over 15 years of education period. Now what would be the estimated value of investment to earn per yrs. of the stipulated sum if 18% interest prevailed in banking system.
- **ANS:** Rs. 44,150 (annual amount received each year from 1 to 15 th. year) X 5.092 (18% present worth of an annuity factor for 15 years) = Rs. 22,4810 (amount to be invested at present).

Dr. A.K.Nandi

### **SINKING FUND FACTOR**

- **SINKING FUND FACTOR**- Level deposit required each year to reach 1 by a given year. (Replacement System).

Formula used:  $i / (1+i)^n - 1$  where,  $i =$  rate of interest,  $n =$  number of yrs.

- **Utility:** This factor permits calculating the level annual payment that must be set aside each year, to be invested at compound interest, in order to have a predetermined sum at a given time. It is used primarily to determine how much must be put in to a fund in order to have recovered the amount of a investment at the end of its useful life).
- **Example:** When a new irrigation scheme was installed, it was decided farmers should be charged an amount sufficient to replace the pumps when they wear out- that is the farmers will pay a replacement charge. It is decided that the money collected from the farmers should be invested in government bonds bearing 11% interest. To replace the pumps in 15 years will cost Rs. 875,000/-. How much should be the annual collection from the farmers?
- **Ans:** Rs. 875000 (Replacement cost in 15 yrs.) X 0.029065 (11% sinking fund factor for 15 yrs.) = Rs. 25,432 / (Annual amount to be collected from farmers and to be invested in govt. bonds at 11%).

Dr. A.K.Nandi

### **CAPITAL RECOVERY FACTOR-**

- **CAPITAL RECOVERY FACTOR-** Annual payment that will repay a loan in X years with compounded interest on the unpaid balance. (Partial payment factor).  

$$1/a_n = i/1-v_n$$
- **This factor is the reciprocal of the present worth of an annuity factor.**
- **Utility:** This factor permits calculating what constant annual payment would be necessary to repay a loan over a given period at a stated interest rate. The total payment is a varying combination of both interest and repayment of principal.
- **Example:** The ADB lends to Indian farmer at 8% interest to finance the tube wells. If a farmer borrows Rs. 8,790 for a tube well to be repaid in 10 years, what is the amount of combined interest and principal payment?
- **ANS:** Rs.8790 (amount initially borrowed at the beginning of the first year) X 0.149029 (8% capital recovery factor for a 10 year period) = Rs. 1,310/ (Amount of annual payment from end of first year through end of 10<sup>th</sup> year).

Dr. A.K.Nandi

### **Doubling Period (Rule of 72)**

#### **Growth Rate**

***Invested amount doubled itself with a stipulated interest.***

- **THUMB RULE:** Number of years = 72/i, i= rate of interest.

**More accurately :** 0.35+ 69/rate of interest= Numbers of years of doubling period of an investment.

- **Growth Rate:** Suppose initial amount is Rs 100/ and the final amount increased to Rs.1000/ in 10 yrs. What was the compound growth rate?

- **ANS:**  $1000 = 100(1+g)^{10}$

$$1000/100 = (1+g)^{10} ; \text{ or, } (1+g) = 10^{1/10}$$

$$g = 10^{1/10} - 1$$

$$g = 1.26 - 1 = 0.26 \text{ or, } 26\%$$

Dr. A.K.Nandi

## **Investment Criteria**

- The steps which are essential in project appraisal and or evaluation i.e. whether the project is worthwhile or not are:
- **Estimate the cost and benefit of the project,**
- **Asses the riskiness of project,**
- **Calculate the cost of capital,**
- **Compute the criterion of merit & judge whether the project is good or bad.**
- **Major criteria's are,**
- **i) Discounted-** Net present Value(NPV), Benefit cost ratio(B/C ratio), Internal rate of return (IRR),
- **ii) Undiscounted-** Pay- back period, Accounting rate of return.

Dr. A.K.Nandi

## **Discounted B/C ratio**

- **Benefit –cost ratio= Present worth of benefit/present worth of cost.**
- **Net present worth= Present worth of benefit -- present worth of cost.**
- *Net present worth (often referred to as net present value) is simply the net present worth of the net benefits of the project discounted at the opportunity cost of capital.*

Dr. A.K.Nandi

## Internal rate of return (IRR)

**Internal rate of return (IRR) = That discount rate, such that  
Present worth of benefit = present worth of cost.**

- IRR must be determined by trial & error. The measure represents the return over the life of projects to the resources engaged in the project. The cash flow is discounted to determine its present worth. By trial & error one discount rate is found which is too low and which leaves a positive present worth and another discount rate is found which is too high and which leaves a negative present worth of the cash flow stream.
- **IRR= Lower discount rate+ Difference between the two discount rates(present worth of the cash flow at the lower discount rate/Absolute difference between the present worth of the cash flow streams at the two discount rates)**

Yield Rate of Perennials=IRR/FRR/ERR

Dr. A.K.Nandi

## Comparison of Perennials with annuals

- Present worth of Benefit/ Annuity factor for the respective year= Annual constant flow over the total period.
- Rotation decision of Perennials- Where the IRR is higher or where the cumulative annual constant flow is highest.

Dr. A.K.Nandi

### Practical characteristics of a good farm plan

1. Efficient use of farm resources.
2. Balanced crop plan- combination with enterprise.
3. Different food/cash/Food & Crops> Mini.
4. Help to maintain & improve soil fertility.
5. Help to raise & stable farm earnings.
6. Improve distribution and use of lab, pow, water through own the year.
7. Avoid excess risk.
8. Provide flexibility.
9. Knowledge/train/ex.
10. Give consideration of efficient marketing farm
11. Credit- Use & repay
12. Modern Agricultural methods-update.

Dr. A.K.Nandi

*Thanks to everybody*

Dr. A.K.Nandi

## SPLIT-PLOT AND STRIP PLOT DESIGNS

**Prof. R. N. Panda**

**Professor (Retd.)**

**University of Kalyani**

### SPLIT-PLOT DESIGNS:

In field experiment, there may sometimes two types of factors : One of them requires large plots whereas other requires very small plots compared to the first. In this situation, the second factor can be accommodated in the experiment with a little extra cost.

Here the first factor A is said to be the whole plot (larger plot) treatment and second factor B is called the sub-plot (small plot) treatment.

A split-plot design may be performed in RBD or LSD.

### Split Plot Design in RBD

At first the levels of whole plot treatment are applied at random to whole-plots of each of the blocks and then each whole plots are split into some sub-plots, equal to the number of levels of sub-plot treatments. Then the levels of sub-plot treatment are allotted to the sub-plots within each whole-plot at random.

In this design main effect B and interaction effect  $A \times B$  will be tested more efficiently than the main effect A.

The split-plot design may be looked upon as a design confounding the main effect A where sub-plot will be treated as plot and whole plot will be treated as block.

**Example :** 1) A : Plant disease / Plant medicine

B : Varieties of a crop.

2) A : Feeds for fish

B : Varieties of fishes

3) A : Methods of plantation (or irrigation)

B : varieties of a crop

4) To study the Tensile Strength of paper

A : Pulp preparation method.

B : Different cooling temperatures

**Lay-out :** Suppose there are  $p$  levels of factor A and  $q$  levels of factor B. Let the experiment be conducted in an RBD with  $r$  blocks.

At first  $p$  levels of factor A are allotted at random to  $p$  whole plots of each block. Then each whole-plot is split into  $q$  sub-plots and the levels of B are allotted at random within  $q$  subplots of each whole plot.

**Analysis :** The model is

$$y_{ijk} = \text{observation due to } K\text{th sub-plot within } i\text{th whole plot in } i\text{th block}$$

$$= \mu + b_i + t_j + e_{1ij} + v_k + (vt)_{jk} + e_{2ijk} \quad (1)$$

$$i = 1, 2, \dots, r; \quad j = 1, 2, \dots, p; \quad K = 1, 2, \dots, q$$

$\mu$  = fixed but unknown constant,

$b_i$  = random effect due to  $i$ th block,

$t_j$  = fixed effect due to  $j$ th level of A with  $\sum_j t_j = 0$

$v_k$  = fixed effect due to  $k$ th level of B with  $\sum v_k = 0$

$vt_{jk}$  = fixed effect due to interaction effect of  $j$ th level of A and  $k$ th levels of B.

$e_{1ij}$  = whole plot error.

$e_{2ijk}$  = Sub-plot error.

The random components  $b_i$ ,  $e_{1ij}$  and  $e_{2ijk}$  are iid rv normal with zero means and variances  $\sigma_b^2$ ,  $\sigma_1^2$  and  $\sigma_2^2$ .

We are interested to test the following hypotheses :

$H_{01}$  (all whole-plot treatments are equally effective)

$$\equiv H_{01}(t_1 = t_2 = \dots = t_p = 0)$$

$H_{02}$  (all sub-plot treatments are equally effective)

$$\equiv H_{02}(v_1 = v_2 = \dots = v_q = 0)$$

$H_{03}$  ( $(vt)_{jk} = 0$  for all  $j, k$ ).

$$E(MS_{EI}) > E(MS_{EII}).$$



So sub-plot treatments and interaction effects are tested more efficiently than the whole-plot treatments].

Source of Variation	df	S. S.	M. S.	E(M. S.)	F <sub>0</sub>
Block	r - 1	SS <sub>Block</sub>	MS <sub>Block</sub>		
A	p - 1	SS <sub>A</sub>	MS <sub>A</sub>	$\sigma_2^2 + q\sigma_1^2 + \phi_1(\tau)$	MS <sub>A</sub> / MS <sub>E<sub>I</sub></sub>
Error I	(r - 1)(p - 1)	SS <sub>E<sub>I</sub></sub>	MS <sub>E<sub>I</sub></sub>	$\sigma_2^2 + q\sigma_1^2$	
B	q - 1	SS <sub>B</sub>	MS <sub>B</sub>	$\sigma_2^2 + \phi_2(v)$	MS <sub>B</sub> / MS <sub>E<sub>II</sub></sub>
AB	(p - 1)(q - 1)	SS <sub>AB</sub>	MS <sub>AB</sub>	$\sigma_2^2 + \phi_3(vt)$	MS <sub>AB</sub> / MS <sub>E<sub>II</sub></sub>
Error II	p(q - 1)(r - 1)	SS <sub>E<sub>II</sub></sub>	MS <sub>E<sub>II</sub></sub>	$\sigma_2^2$	
Total	rpq - 1	SS <sub>T</sub>	-		

$$\text{Here } SS_{\text{Block}} = \frac{\sum y_{i00}^2}{pq} - \text{C. F.}, \quad \text{C. F.} = \frac{y_{000}^2}{rpq}$$

$$SS_A = \frac{\sum y_{0j0}^2}{rq} - \text{C. F.}, \quad SS_B = \frac{\sum y_{00k}^2}{r_p} - \text{C. F.}$$

$$SS_{AB} = \frac{\sum X_{0jk}^2}{pq} - \text{C. F.} - SS_B - SS_A$$

$$SS_{E_I} = \frac{\sum y_{ij0}^2}{q} - \text{C. F.} - S_{\text{Block}} - SS_A$$

$$SS_T = \sum \sum y_{ijk}^2 - \text{C. F.}$$

SS<sub>E<sub>II</sub></sub> is obtained by subtraction.

	A	Total
	Y <sub>ij0</sub>	Y <sub>100</sub> ... Y <sub>r00</sub>
Total	Y <sub>010</sub> ... Y <sub>0p0</sub>	Y <sub>000</sub>

	B	
A		
	Y <sub>0jk</sub>	Y <sub>0j0</sub>
	Y <sub>001</sub> ... Y <sub>00q</sub>	Y <sub>000</sub>

$$SS_{\text{Block}}, SS_A, SS_{E_1}$$

$$SS_A, SS_B, SS_{A \times B}$$

**Interpretation :**

When  $H_{03}$  is rejected then there is no meaning of testing  $H_{01}$  and  $H_{02}$ .

If  $H_{03}$  is accepted then to find the best whole plot treatment and sub-plot treatment. We proceed as follows :

$$i) |\hat{t}_j - \hat{t}_j| = |y_{0j0} - y_{0j'0}| > t_{\alpha/2, p-1(r-1)} \sqrt{\frac{2MS_{EI}}{rq}}$$

$$ii) |\hat{v}_{K'} - \hat{v}_{K'}| = |y_{0k_0} - y_{0k'_0}| > t_{\alpha/2, p(r-1)(q-1)} \sqrt{\frac{2MS_{EII}}{r_p}}$$

If  $H_{03}$  is rejected,

(iii) At the same level of whole-plot treatment,

$$|\hat{v}_K - \hat{v}_{K'}| > t_{\alpha/2, p(r-1)(q-1)} \sqrt{\frac{2MS_{EII}}{r}}$$

Ex. : The following data on tensile strength of the paper were obtained from an experiment with 3 different pulp preparation methods ( $M_1, M_2, M_3$ ) and four different cooling temperatures ( $C_1, C_2, C_3, C_4$ ) in four replicates. Analyse the data

Block 1						Block 2					
M <sub>1</sub>		M <sub>2</sub>		M <sub>3</sub>		M <sub>2</sub>		M <sub>1</sub>		M <sub>3</sub>	
C <sub>1</sub>	94	C <sub>4</sub>	440	C <sub>2</sub>	250	C <sub>1</sub>	135	C <sub>2</sub>	160	C <sub>4</sub>	370
C <sub>3</sub>	220	C <sub>2</sub>	297	C <sub>1</sub>	147	C <sub>4</sub>	290	C <sub>4</sub>	95	C <sub>1</sub>	140
C <sub>2</sub>	185	C <sub>3</sub>	218	C <sub>3</sub>	248	C <sub>2</sub>	180	C <sub>3</sub>	124	C <sub>3</sub>	340
C <sub>4</sub>	110	C <sub>1</sub>	112	C <sub>4</sub>	275	C <sub>3</sub>	265	C <sub>1</sub>	71	C <sub>2</sub>	222

Block 3						Block 4					
M <sub>1</sub>		M <sub>2</sub>		M <sub>3</sub>		M <sub>1</sub>		M <sub>2</sub>		M <sub>3</sub>	
C <sub>1</sub>	78	C <sub>3</sub>	196	C <sub>2</sub>	235	C <sub>1</sub>	81	C <sub>3</sub>	246	C <sub>2</sub>	290
C <sub>3</sub>	135	C <sub>4</sub>	262	C <sub>3</sub>	260	C <sub>2</sub>	175	C <sub>4</sub>	191	C <sub>3</sub>	250
C <sub>4</sub>	130	C <sub>1</sub>	155	C <sub>1</sub>	115	C <sub>4</sub>	175	C <sub>1</sub>	145	C <sub>1</sub>	120
C <sub>2</sub>	145	C <sub>2</sub>	220	C <sub>4</sub>	483	C <sub>3</sub>	114	C <sub>2</sub>	323	C <sub>4</sub>	450

**STRIP-PLOT DESIGNS**

This design is used when both the factors each requiring large sized plots. Suppose there are  $p$  levels of factor A and  $q$  levels of factor B.

**Layout :** Suppose the experiment is conducted with  $r$  replicates. We divide each replicate into  $p$  rows and  $q$  columns. The levels of A are allotted randomly in  $p$  rows and the levels of B are allotted randomly in  $q$  columns. The randomization is done afresh for each of the replicates.

**Example :** 1) A : Dates of ploughing

B : Methods of ploughing

2) A : plant disease

B : plant medicine

**Analysis :** Here the model is

$$y_{ijk} = \mu + a_i + t_j + e_{1ij} + v_k + e_{2ik} + (tv)_{jk} + e_{3ijk}$$

where  $e_{1ij}$ 's,  $e_{2ik}$ 's and  $e_{3ijk}$ 's are errors, which are iidrv normal with zero means and variances  $\sigma_1^2$ ,  $\sigma_2^2$  and  $\sigma_3^2$  respectively.

**ANOVA TABLE**

Source of Variation	df	S. S.	
Replication	$r - 1$	$SS_R$	
A	$p - 1$	$SS_A$	
Error I ( $R \times A$ )	$(r - 1)(p - 1)$	$SS_{EI}$	$F_1 = \frac{MS_A}{MS_{EI}}$
B	$(q - 1)$	$SS_B$	
Error II ( $R \times B$ )	$(r - 1)(q - 1)$	$SS_{EII}$	$F_2 = \frac{MS_B}{MS_{EII}}$
AB	$(p - 1)(q - 1)$	$SS_{AB}$	

Error III (R × A × B)	(p - 1)(q - 1) (r - 1)	SS <sub>EIII</sub>	$F_3 = \frac{MS_{AB}}{MS_{EIII}}$
Total	pqr - 1		

**Example 2 :** The field plan and yield of a strip-plot experiment with 3 dates of planting (d<sub>1</sub>, d<sub>2</sub>, d<sub>3</sub>) and 3 methods of planting (m<sub>1</sub>, m<sub>2</sub>, m<sub>3</sub>) in 3 replicates are given below. Analyse the data

**Field plan and yield data**

Replicate I				Replicate II				Replicate III			
	m <sub>1</sub>	m <sub>2</sub>	m <sub>3</sub>		m <sub>1</sub>	m <sub>2</sub>	m <sub>3</sub>		m <sub>1</sub>	m <sub>2</sub>	m <sub>3</sub>
d <sub>1</sub>	290	70	220	d <sub>3</sub>	185	295	245	d <sub>3</sub>	175	245	295
d <sub>2</sub>	370	140	210	d <sub>4</sub>	220	125	135	d <sub>2</sub>	145	175	110
d <sub>3</sub>	95	135	240	d <sub>2</sub>	180	160	140	d <sub>1</sub>	80	190	250

**Solution :**

A \ R		Total (Y <sub>i00</sub> )
	Y <sub>ij0</sub>	
Total (y <sub>0j0</sub> )		Y <sub>000</sub>

B \ R		Y <sub>i00</sub>
	Y <sub>i0k</sub>	
Y <sub>00k</sub>		Y <sub>000</sub>

B \ A		Y <sub>0j0</sub>
	Y <sub>0jk</sub>	
Y <sub>00k</sub>		

$$SS_R = \frac{\sum y_{i00}^2}{pq} - C.F.,$$

$$SS_A = \frac{\sum y_{0j0}^2}{rq} - C.F.$$

$$SS_B = \frac{\sum y_{00k}^2}{rp} - C.F.$$

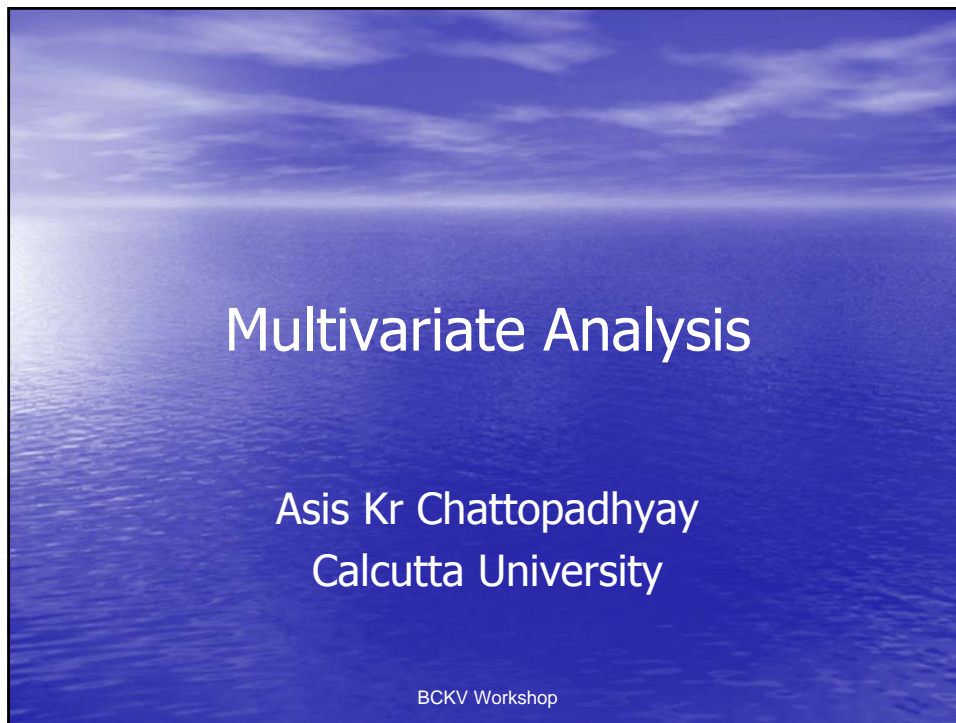
$$SS_{EI} = \frac{\sum \sum y_{ij0}^2}{q} - C.F. - SS_R - SS_A$$

$$SS_{EII} = \frac{\sum \sum y_{i0k}^2}{p} - C.F. - SS_R - SS_B$$

$$SS_{AB} = \frac{\sum \sum y_{0jk}^2}{r} - C.F. - SS_A - SS_B$$

$$SS_T = \sum \sum y_{ijk}^2 - C.F.$$

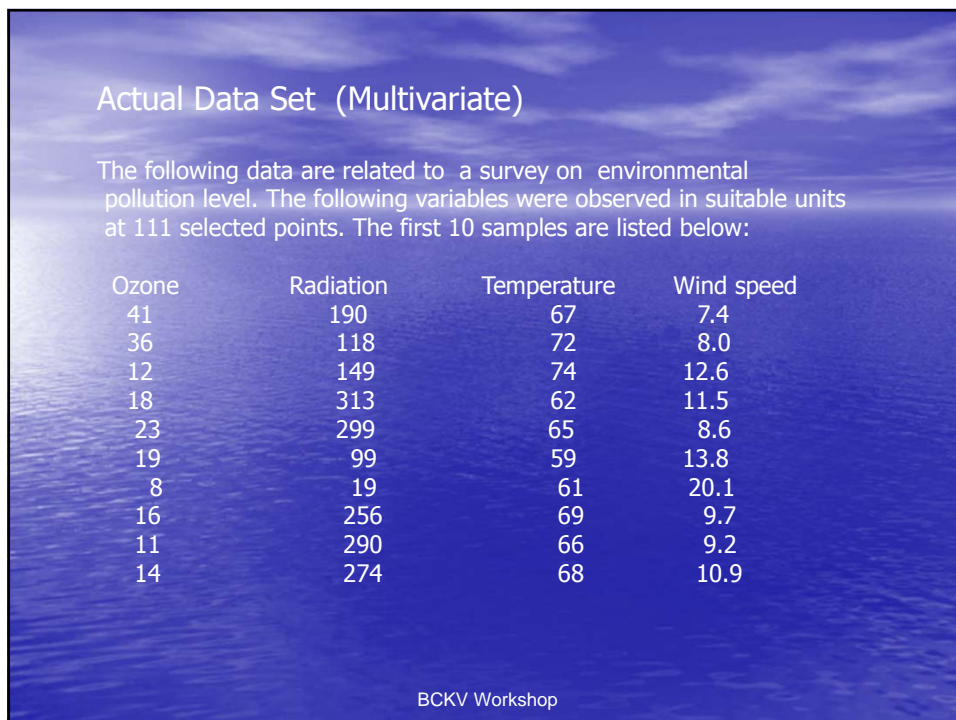
$SS_{E_{III}}$  is obtained by subtraction.



# Multivariate Analysis

Asis Kr Chattopadhyay  
Calcutta University

BCKV Workshop

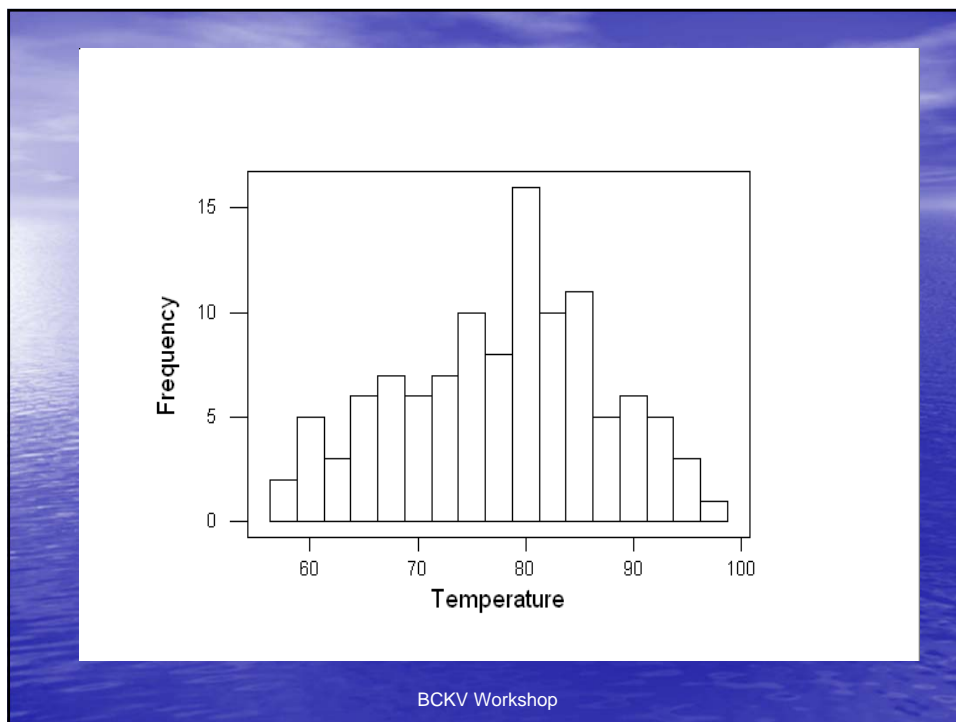
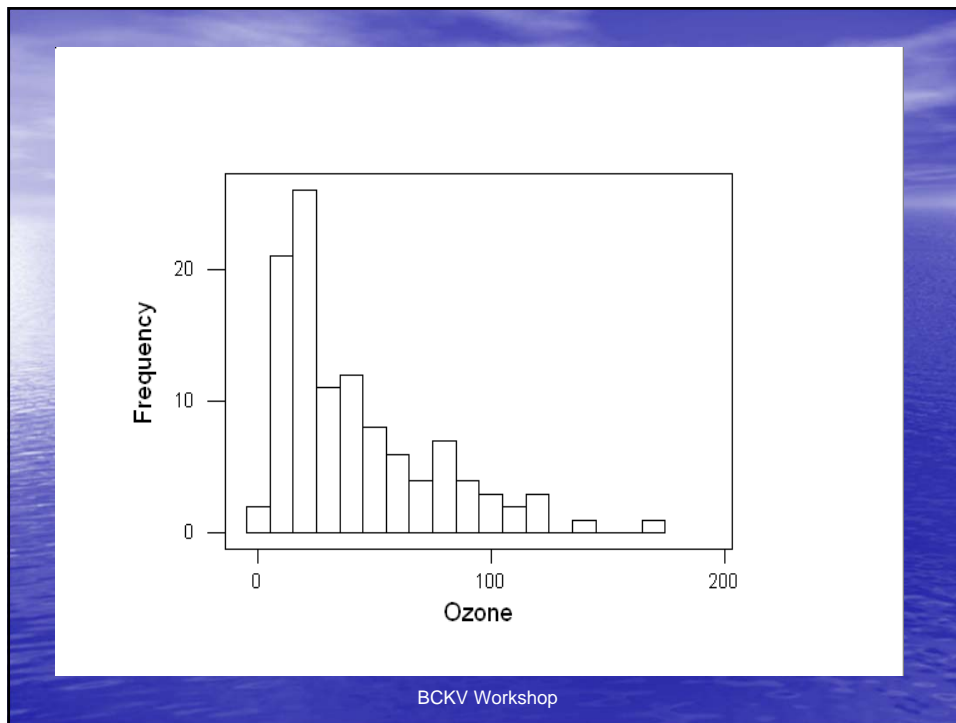


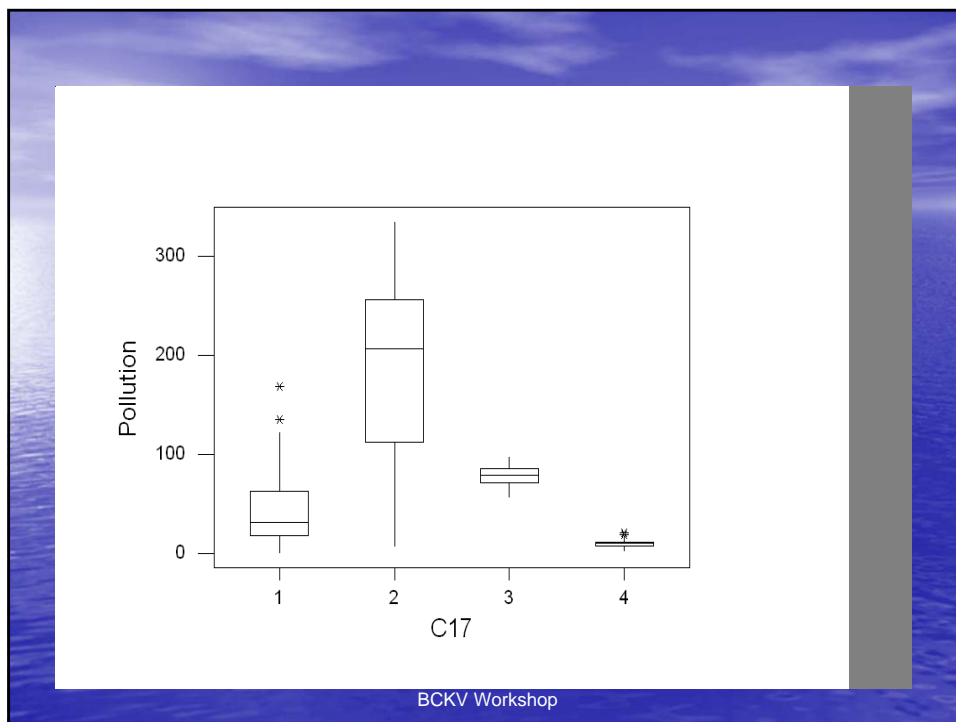
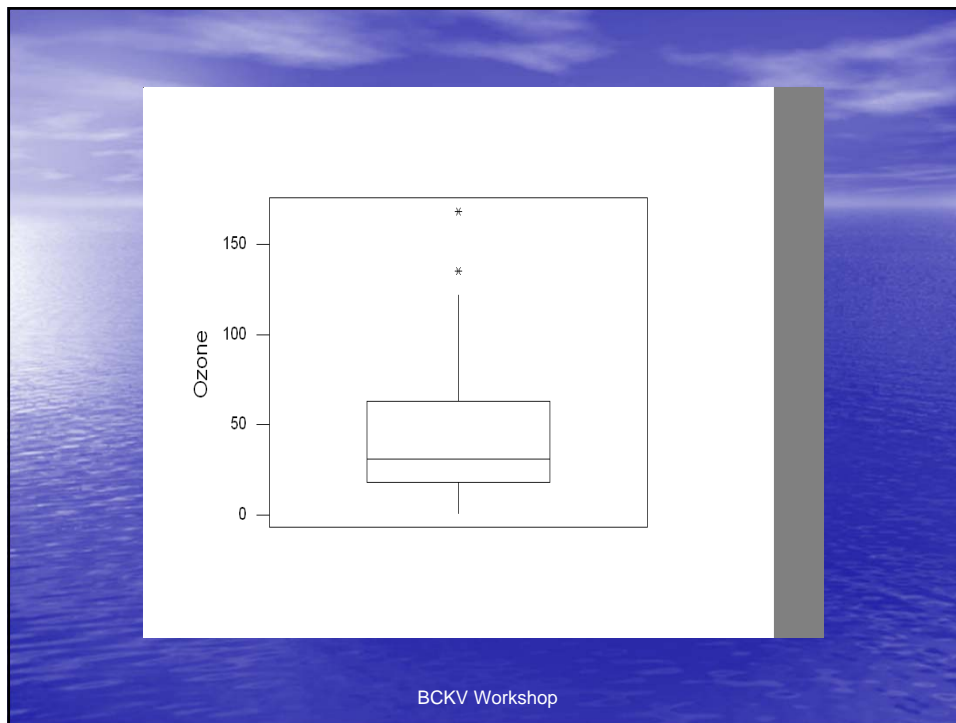
## Actual Data Set (Multivariate)

The following data are related to a survey on environmental pollution level. The following variables were observed in suitable units at 111 selected points. The first 10 samples are listed below:

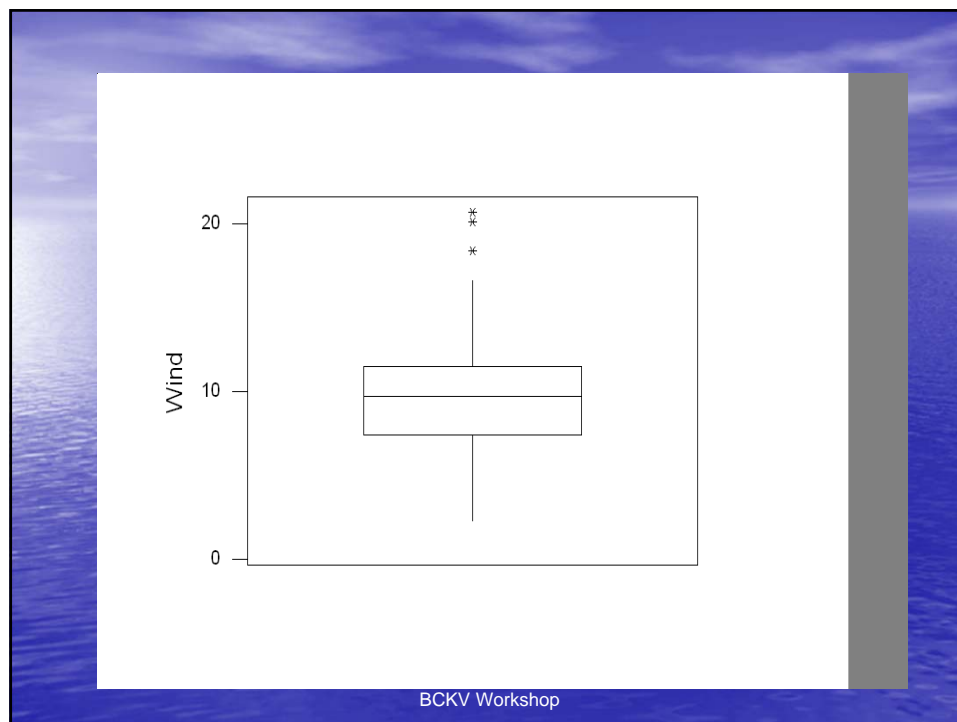
Ozone	Radiation	Temperature	Wind speed
41	190	67	7.4
36	118	72	8.0
12	149	74	12.6
18	313	62	11.5
23	299	65	8.6
19	99	59	13.8
8	19	61	20.1
16	256	69	9.7
11	290	66	9.2
14	274	68	10.9

BCKV Workshop









### Bivariate Data

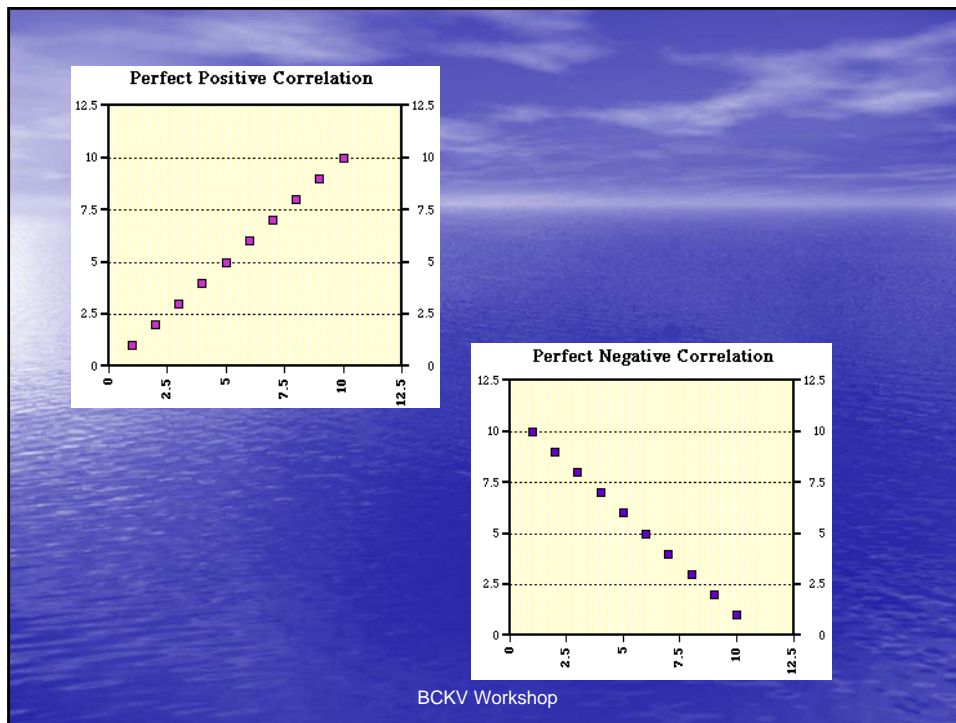
#### XY plots

(Scatter plots) are similar to line graphs in that they use horizontal and vertical axes to plot data points. However, they have a very specific purpose. Scatter plots show how much one variable is affected by another.

Scatter plots usually consist of a large body of data. The closer the data points come when plotted to making a straight line, the higher the correlation between the two variables or the stronger the relationship.

If the data points make a straight line going from the origin out to high x- and y-values, then the variables are said to have a **positive correlation**. If the line goes from a high-value on the y-axis down to a high-value on the x-axis, the variables have a **negative correlation**.

BCKV Workshop



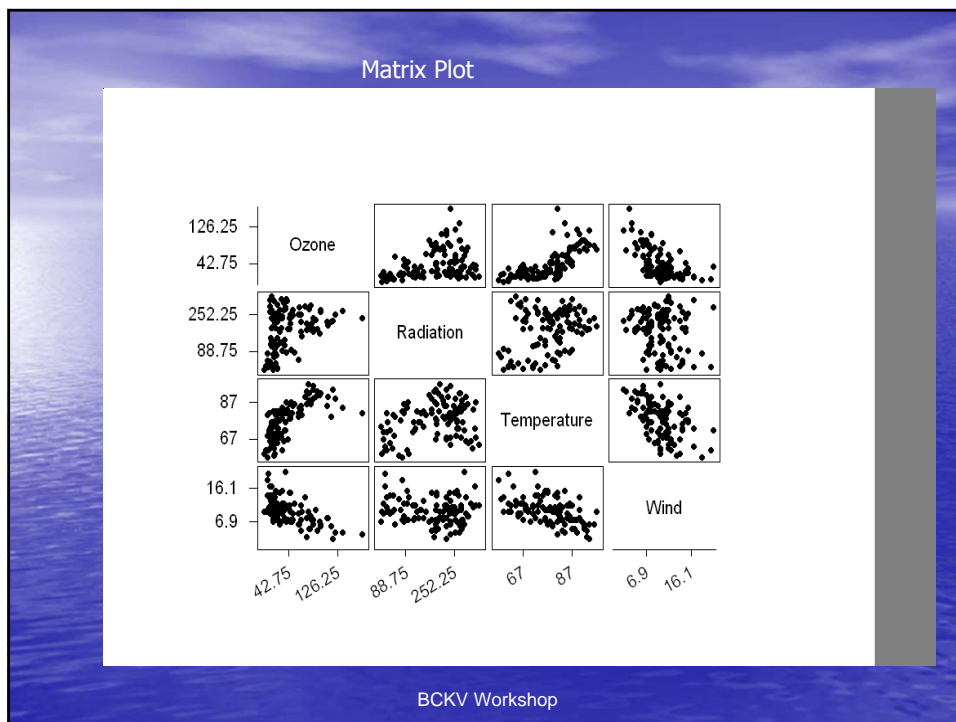
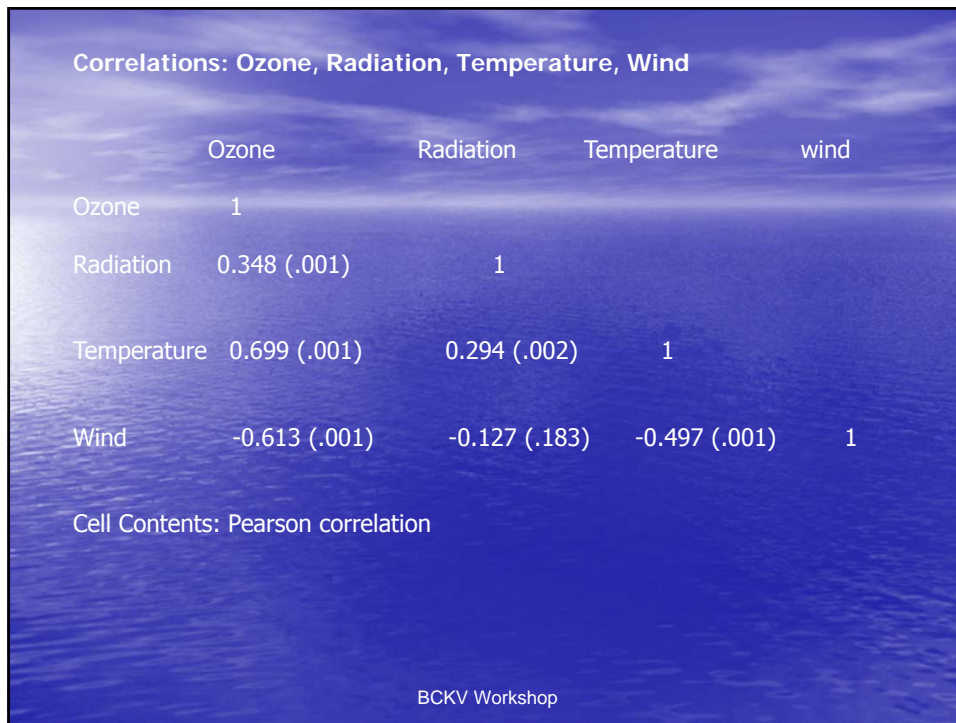
The correlation coefficient  $r$  (also called Pearson's product moment correlation after [Karl Pearson](#)) is calculated by

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

Data Set

- X1 Y1
- X2 Y2
- X3 Y3
- X4 Y4
  
- Xn Yn

BCKV Workshop



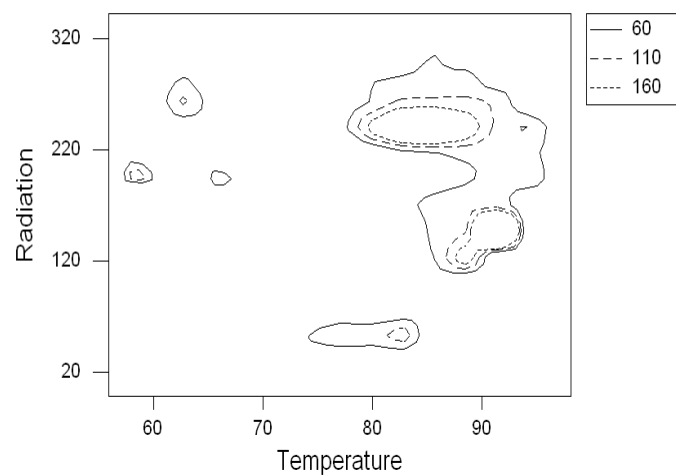
### Contour Plot

A contour plot is a graphical technique for representing a 3-dimensional surface by plotting constant  $z$  slices, called contours, on a 2-dimensional format.

That is, given a value for  $z$ , lines are drawn for connecting the  $(x,y)$  coordinates where that  $z$  value occurs

BCKV Workshop

### Contour Plot of Ozone



BCKV Workshop

#### Why Multivariate?

A single physical parameter usually does not take into account entire variation among the observations.

To understand the underlying process properly it is necessary to access the relative importance of the influence of various variables and to identify those which are most responsible.

Since many variables are responsible for the overall variation a Multivariate setup should be considered

#### How to reduce dimension?

It is always easier to explain variation in terms of a smaller number of Variables.

BCKV Workshop

#### PRINCIPAL COMPONENT ANALYSIS

A Principal Component analysis is concerned with explaining the variance-covariance structure of a set of variables through a few linear combinations of these variables.

##### Objectives:

- Data Reduction
- Interpretation

Generally the number of linear combinations (components) required to reproduce the total system variability is less than the total number of variables.

Those important linear combinations are called Principal components.

Thus the original data set of  $n$  measurements on  $p$  variables can be reduced to a data set consisting of  $n$  measurements on  $k$  ( $k \leq p$ ) principal components.

An analysis of principal components often reveals relationships that were not previously suspected and thereby allows interpretations that would not ordinarily result.

BCKV Workshop

We have  $p$  random variables  $X_1 X_2 X_3 \dots X_p$

Define  $X = (X_1 X_2 X_3 \dots X_p)'$

Let  $X_1 X_2 X_3 \dots X_p$  have covariance matrix  $\Sigma$  with characteristic roots  $\lambda_1 > \lambda_2 > \lambda_3 > \dots > \lambda_p > 0$ .

Consider the linear combinations

$$Y_1 = a_{11} X_1 + a_{12} X_2 + \dots + a_{1p} X_p = \mathbf{a}_1' \mathbf{X}$$

$$Y_2 = a_{21} X_1 + a_{22} X_2 + \dots + a_{2p} X_p = \mathbf{a}_2' \mathbf{X}$$

$$\dots$$

$$Y_p = a_{p1} X_1 + a_{p2} X_2 + \dots + a_{pp} X_p = \mathbf{a}_p' \mathbf{X}$$

Then  $\text{Var}(Y_i) = \mathbf{a}_i' \Sigma \mathbf{a}_i \quad i = 1, 2 \dots p$

$$\text{Cov}(Y_i Y_j) = \mathbf{a}_i' \Sigma \mathbf{a}_j$$

The principal components are those uncorrelated linear combinations whose variances are as large as possible.  
The first principal component is the linear combination with maximum variance.

BCKV Workshop

Result: Let  $\Sigma$  be the covariance matrix associated with the random vector  $\mathbf{X}' = (X_1, X_2, \dots, X_p)$ . Let  $\Sigma$  have the Eigen vector eigen value pairs  $(\lambda_i e_i)$   $i = 1, 2, \dots, p$  where  $\lambda_1 > \lambda_2 > \dots > \lambda_p$ .

Then the  $i$ th principal component is given by

$$Y_i = e_{i1} X_1 + e_{i2} X_2 + \dots + e_{ip} X_p$$

With  $\text{Var}(Y_i) = \lambda_i$

and  $\text{Cov}(Y_i Y_j) = 0$ .

$$\sum \text{Var}(X_i) = \sum \lambda_i = \sum \text{Var}(Y_i)$$

Hence the coefficients of linear combinations can be obtained as the eigen vectors of the covariance matrix  $\Sigma$

BCKV Workshop

### Eigen analysis of the Correlation Matrix

Eigenvalue	2.3602	0.8946	0.4758	0.2695
Proportion	0.590	0.224	0.119	0.067
Cumulative	0.590	0.814	0.933	1.000

Variable	PC1	PC2	PC3	PC4
Radiatio	0.317	0.899	-0.277	0.123
Temperat	0.553	-0.061	0.659	0.507
Wind spe	-0.497	0.430	0.690	-0.303
Ozone Co	0.589	-0.063	0.114	-0.798

BCKV Workshop

### HIERARCHICAL CLUSTERING

Hierarchical clustering techniques proceed by either a series of successive mergers (agglomerative) or successive divisions (divisive).

There are initially as many clusters as objects. The most similar objects are first grouped and these initial groups are merged according to their similarities. Eventually as the similarity decreases all subgroups are fused into a single cluster.

Otherwise an initial single group of objects is divided into two subgroups such that the objects in one subgroup are far from the objects in the other. These subgroups are then further divided into dissimilar subgroups. This process continues until each object forms a group.

BCKV Workshop

Suppose corresponding to each of  $n$  objects we have  $p$  continuous measurements which are positive or negative real numbers.

For observations with different units we may standardize the values.

The standardized values will be unit free.

But this is not always necessary.

Proximity Values:

A Collection of proximities must be available for all pairs of objects. There are two types of Proximity Measures:

Dissimilarities ( how far away the objects are)

Similarities (how much they resemble each other)

BCKV Workshop

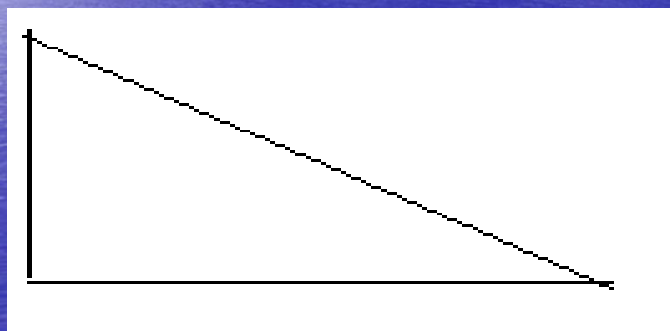
#### SIMILARITY MEASURES (DISTANCES)

Euclidian Distance

$$E(x,y) = \text{Sqrt}((x-y)'(x-y))$$

2. Manhattan Distance

$$M(x,y) = \text{abs}(x-y)' \mathbf{1}_{p \times 1}$$



BCKV Workshop



Dissimilarity in terms of correlation coefficient

Two Possible measures:

1.  $d(i,j) = (1 - r_{ij})/2$

2.  $d(i,j) = 1 - \text{abs}(r_{ij})$

BCKV Workshop

Dissimilarity for ordinal/nominal/binary data

Binary Data ( 0/1)

j \ i	1	0	Total
1	a	b	a+b
0	c	d	c+d

$s(i,j) = (a+b) / (a+b+c+d)$   
 $d(i,j) = (b+c) / (a+b+c+d)$

BCKV Workshop

#### DIFFERENT LINKAGES

Single linkage

Minimum distance or nearest neighbor

Complete linkage

Maximum linkage or furthest neighbor

Average linkage

Average distance.

Example of single linkage

The distances between two clusters (UV) and (W) is measures by  
 $D(uv)w = \text{Min} \{ D_{uw} \ D_{vw} \}$

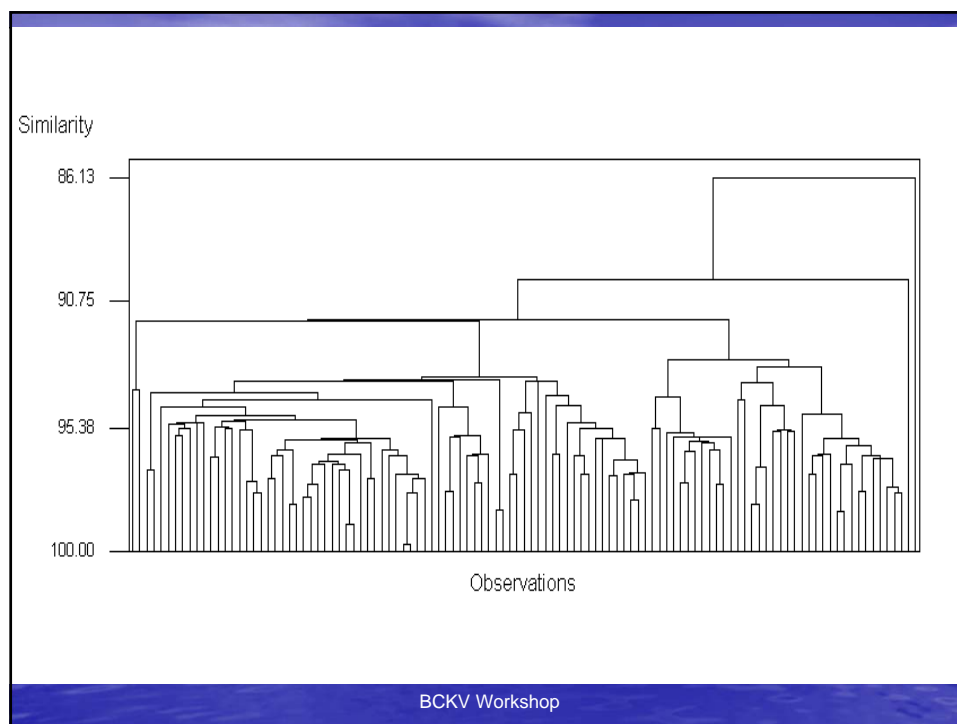
$D_{uw}$  = distance between nearest neighbors of clusters U and W

BCKV Workshop

Steps in agglomerative hierarchical clustering algorithm for grouping N objects (items or variables).

1. Start with N clusters and an  $N \times N$  symmetric matrix of distances  $D=(d_{ij})$
2. Search the distance matrix for nearest pair of clusters. Let the distance between most similar clusters U and V be  $D_{uv}$ .
3. Merge U and V. Label the newly formed cluster  $\{UV\}$ . Update the entries of the distance matrix by  
(a) deleting the rows and columns corresponding to the clusters U and V and (b) adding a row and column giving the distances between the clusters  $\{UV\}$  and the remaining clusters.
4. Repeat Steps 2 and 3 a total of N-1 times. All items will be in a single cluster after the process terminates

BCKV Workshop



### K-Means Clustering

Particularly used to group items rather than variables.

This method can be applied to much larger data sets.

Method starts with either an initial partition of items into k groups or an initial set of k seed points.

Mac Queen's Algorithm:

Given K

1. Partition the items into k initial Groups.
2. Proceed through the list of items, assigning an item to the cluster having the nearest centroid (mean) is nearest (Euclidian Distance)
3. Re calculate the centroids for the clusters receiving the new items and for the clusters losing the item.
4. Repeat steps 2 and 3 until no more reassignment take place.

BCKV Workshop

Cluster 1						
Variable	N	Mean	Median	TrMean	StDev	SE Mean
Radiatio	43	269.65	267.00	269.00	29.11	4.44
Temperat	43	77.02	78.00	76.97	8.30	1.27
Wind spe	43	10.507	10.900	10.395	3.440	0.525
Ozone Co	43	39.37	32.00	36.92	27.01	4.12
Cluster-2						
Variable	N	Mean	Median	TrMean	StDev	SE Mean
Radiatio	34	65.32	68.00	64.60	40.42	6.93
Temperat	34	72.09	72.50	72.10	8.91	1.53
Wind spe	34	11.035	10.300	10.783	3.342	0.573
Ozone Co	34	20.00	17.00	18.80	13.84	2.37
Cluster-3						
Variable	N	Mean	Median	TrMean	StDev	SE Mean
Radiatio	34	196.97	192.50	197.23	27.24	4.67
Temperat	34	84.47	85.00	84.73	7.44	1.28
Wind spe	34	8.124	8.000	8.043	3.308	0.567
Ozone Co	34	67.65	73.00	66.13	37.22	6.38

BCKV Workshop

### DISCRIMINANT ANALYSIS

Fisher's Method

Transform the multivariate observations into univariate observations by taking linear combination of the variables.

There is no assumption regarding normal distribution but population covariance matrices are assumed to be equal.

$y_{11} \ y_{12} \ \dots \ y_{1n_1}$  : observations from first population

$y_{21} \ y_{22} \ \dots \ y_{2n_2}$  : observations from second population

Separation =  $\frac{\text{abs}(\text{mean}(y_1) - \text{mean}(y_2))}{S_y}$

$S_y$  = Pooled covariance matrix =  $\frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{(n_1 + n_2 - 2)}$

BCKV Workshop

Objective is to select linear combination of X to achieve maximum separation of the sample means.

Fisher's linear discriminant function is

$$Y = (\text{mean}(X1) - \text{mean}(X2))' \text{inv}(\text{Sy}) X$$

Where  $X = (X1 \ X2)'$

Allocation rule:

Allocate an observation  $X_0$  to the first population

$$\text{If } Y_0 = (\text{mean}(X1) - \text{mean}(X2))' \text{inv}(\text{Sy}) X_0 \geq M$$

where

$$M = (\text{mean}(X1) - \text{mean}(X2))' \text{inv}(\text{Sy}) (\text{mean}(X1) + \text{mean}(X2))$$

Otherwise allocate it to the second population

BCKV Workshop

Predictors: Radiatio Temperat Wind spe Ozone Co

Group	1	2	3
Count	43	34	34

Summary of Classification

Put into	...True Group...		
Group	1	2	3
1	41	0	1
2	0	33	0
3	2	1	33
Total N	43	34	34
N Correct	41	33	33
Proportion	0.953	0.971	0.971

N = 111    N Correct = 107    Proportion Correct = 0.964

BCKV Workshop

### Summary of Misclassified Observations

Observation	True Group	Pred Group	Group	Squared Distance	Probability
33 **	2	3	1	17.309	0.002
			2	6.294	0.464
			3	6.016	0.534
34 **	1	3	1	14.83	0.190
			2	47.64	0.000
			3	11.93	0.810
40 **	1	3	1	5.366	0.455
			2	40.654	0.000
			3	5.007	0.545
90 **	3	1	1	2.843	0.654
			2	24.647	0.000
			3	4.114	0.346

BCKV Workshop

### Factor Analysis

#### Extension of Principal Component Analysis

Both are dimension reduction techniques and attempts to approximate the covariance matrix.

But the factor analysis model is more elaborate.

The main question is whether the data are consistent with a prescribed structure.

Here each group of variables represents a single underlying factor that is responsible for the observed correlations.

BCKV Workshop

For example for a group of students the test scores in Physics, Mathematics, 100 meter race and high jump may correspond to two underlying factors:

1. "Intelligence"
2. "Physical fitness"

Suppose the observable random vector  $X^{p \times 1}$  with  $p$  components has mean vector  $\mu^{p \times 1}$  and covariance matrix  $\Sigma^{p \times p}$  (or correlation matrix  $\rho^{p \times p}$ )

In the factor model we assume that  $X$  is linearly dependent on a few Unobservable random variables  $F_1, F_2, \dots, F_m$  called common factors and  $p$  additional sources of variation  $\epsilon_1, \epsilon_2, \dots, \epsilon_p$  called the errors (or specific factors).

BCKV Workshop

$$X_1 - \mu_1 = l_{11}F_1 + l_{12}F_2 + \dots + l_{1m}F_m + \epsilon_1$$

$$X_2 - \mu_2 = l_{21}F_1 + l_{22}F_2 + \dots + l_{2m}F_m + \epsilon_2$$

.

$$X_p - \mu_p = l_{p1}F_1 + l_{p2}F_2 + \dots + l_{pm}F_m + \epsilon_p$$

or in matrix notation

$$X^{p \times 1} = \mu^{p \times 1} + L^{p \times m} F^{m \times 1} + \epsilon^{p \times 1}$$

$l_{ij}$  = loading of the  $i^{\text{th}}$  variable

$L^{p \times m}$  = Factor loading matrix

BCKV Workshop

Orthogonal factor model  
Assumptions

$$E(F) = 0_{m \times 1} \quad E(FF') = I_{m \times m} \text{ (factors are standardised and uncorrelated)}$$

$$E(\varepsilon) = 0_{p \times 1} \quad E(\varepsilon\varepsilon') = \text{Diag}(\Psi_i)_{p \times p} \text{ (errors are uncorrelated)}$$

$$= \Psi_{p \times p} \text{ (say)}$$

$$E(\varepsilon F') = 0_{p \times m}$$

Under the above assumptions

Covariance matrix of  $X \quad \Sigma = LL' + \Psi = (\sigma_{ij})$

$$\sigma_{ii} = \text{Variance}(X_i) = l_{i1}^2 + l_{i2}^2 + \dots + l_{ip}^2 + \Psi_i$$

$$= h_i^2 + \Psi_i$$

BCKV Workshop

$h_i^2$  = communality of the  $i$ th variable  
= sum of squares of loadings of the  $i$ th variable on  $m$  common factors

Choice of  $L$  is not unique:

$$\Sigma = LL' = LTT'L' + \Psi = L_1L_1' + \Psi \text{ for any orthogonal matrix } T$$

$$X - \mu = LF + \varepsilon = LTT'F + \varepsilon = L_1F^* + \varepsilon$$

$F$  and  $F^*$  have the same statistical properties even if the loadings are different i.e.  $L$  and  $L_1$ .

This can be avoided by choosing orthogonal rotation  $T$  such that the final loading  $L$  satisfies the condition that  $L'\Psi^{-1}L$  is diagonal with positive diagonal elements (here  $L$  to be of full rank  $m$ ).

BCKV Workshop



Estimation:

The factor loadings (L) and specific variances ( $\Psi$ ) can be estimated by

1. Principal Component method
2. Maximum likelihood method

How to find the number of factors?

Let  $\Sigma$  has the eigen-value eigen vector pair  $(\lambda_i, e_i)$   $i=1,2,\dots,p$

Proportion of total sample variance due to the  $j$ th factor=  
( estimate of  $\lambda_j$  )/ total sample variance

$m$  is so chosen that a suitable proportion of the total sample variance has been explained.

BCKV Workshop

Factor rotation

Rotated loadings:  $L_1 = LT$   $T =$  Orthogonal rotation matrix

Since the original loadings may not be readily interpretable, it is usual practice to rotate them until a simple structure is achieved.

Example of simple structure: Suppose  $m=2$

Variable	original loadings		rotated loadings		Communality
Bengali	0.553	0.429	0.369	0.594	0.490
English	0.568	0.288	0.433	0.467	0.406
French	0.392	0.450	0.211	0.598	0.356
Physics	0.740	-0.273	0.789	0.001	0.623
Chemistry	0.724	-0.211	0.752	0.054	0.568
Mathematics	0.595	-0.132	0.604	0.083	0.372

BCKV Workshop

**Factor Analysis: Radiation, Temperature, Wind speed, Ozone Content**

Principal Component Factor Analysis of the Correlation Matrix

Unrotated Factor Loadings and Communalities

Variable	Factor1	Factor2	Factor3	Factor4	Communality
Radiatio	0.487	0.850	-0.191	0.064	1.000
Temperat	0.849	-0.058	0.454	0.263	1.000
Wind spe	-0.764	0.407	0.476	-0.157	1.000
Ozone Co	0.905	-0.060	0.078	-0.414	1.000
Variance	2.3602	0.8946	0.4758	0.2695	4.0000
% Var	0.590	0.224	0.119	0.067	1.000

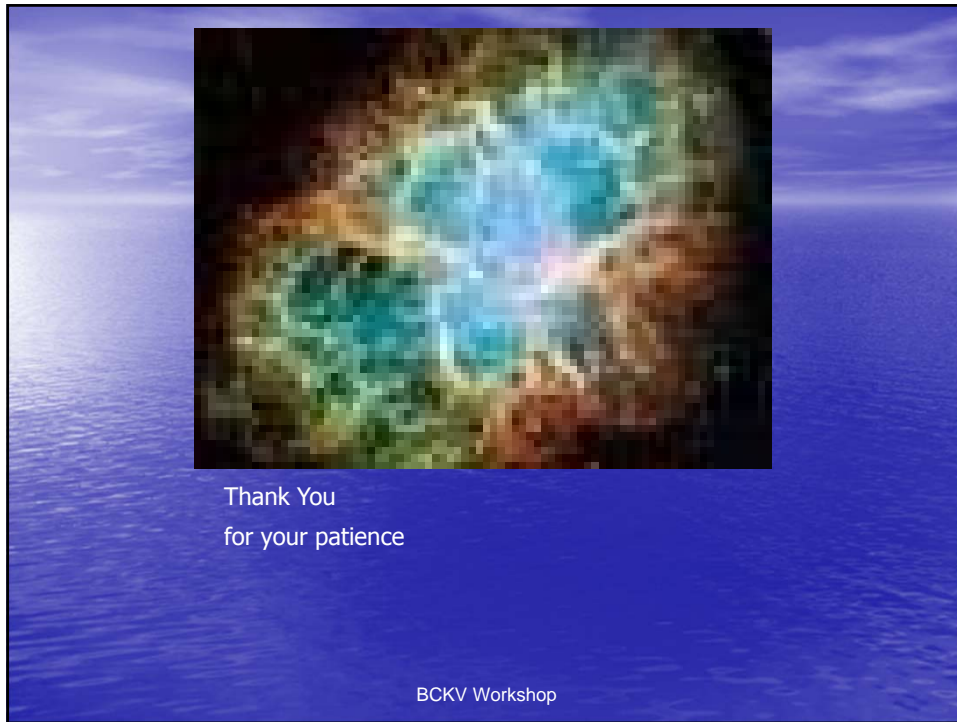
BCKV Workshop

Rotated Factor Loadings and Communalities

Varimax Rotation

Variable	Factor1	Factor2	Factor3	Factor4	Communality
Radiatio	0.036	0.118	0.984	-0.131	1.000
Temperat	0.231	0.913	0.140	-0.306	1.000
Wind spe	-0.942	-0.216	-0.035	0.254	1.000
Ozone Co	0.330	0.369	0.185	-0.849	1.000
Variance	1.0508	1.0307	1.0227	0.8959	4.0000
% Var	0.263	0.258	0.256	0.224	1.000

BCKV Workshop



**BCKV**  
**Workshop on Statistical Methods**  
**May 29 – June 10, 2017**

**Bikas K Sinha**  
**[bikassinha1946@gmail.com]**  
**Retired Professor**  
**Indian Statistical Institute, Kolkata**

\*\*\*\*\*

**Sample Size Determination**

**QUOTE OF THE DAY.....**

- **TEACH WHAT YOU KNOW TODAY.....**
- **USE WHAT YOU KNOW TODAY AND  
FIGURE OUT THE REST.....**
- **DON'T WAIT FOR WHEN YOU THINK ....**
  - **YOU WILL BE READY.....**

### **Agri. Expt.....Possible Scenarios**

There are  $N = 337$  small plots of approx. same shape and size in a large agricultural region.

Agri. Deptt wants to know about the *proportion* of plots with 'Low Soil Fertility [**LSF**]' / 'Medium Soil Fertility [**MSF**]' / 'High Soil Fertility [**HSF**]'.

Should be a straightforward question to 'Agri. Scientists' .....given enough time and resources.....

Possible Recommendation : Collect soil sample from *each* plot and send to laboratory for measuring soil fertility.....

Lab Analysis Results .....classified as

Category of soil :	LSF	MSF	HSF
Total Count (%) :	139 (41%)	82(24%)	116(35%)

### **Scientific Investigations.....**

Meaningful & Relevant questions are then raised by Agri. Deptt.....

AD : It seems....AS recommended carrying out 100% screening i.e., testing of all 337 soil samples !

Could not one do with a sample of, say 100 or, may be 150 plots out of the whole lot of 337 plots and thereby use results of only the tested 100/150 soil samples to derive a solution ?

Would there be a substantial loss of precision in estimating the proportion based on a sample of *adequate* size?

Statisticians may offer reasonable answers with possible explanations.....

## **Soil Fertility.....**

**Scenario : Totally new site ....no past experience in soil characteristic study for this site of 337 plots**

**Q1. What is the proportion of plots with LSF ? MSF ? HSF ?**

***Point Estimation Problem.....***

**Q2. What is the possible range of % of plots with Low SF ? Is it like 35 % - 45 % ? How do we decide on this ? And, how much confidence level can we attach to this sort of statement ?**

***Estimation through Confidence Interval .....***

**Q3. New Site :YET-Expert's Assessment of LSF : 35 % of plots**

**Do we accept the assessment outright ? Can we test the validity of this assessment? Is it tenable ?**

***Testing of Hypothesis problem.....***

***WE ARE RULING OUT 100% SCREENING OUTRIGHT !!!***

## **Statutory Warning !!!**

- **In GOD....we trust.....**
- ***All Others must bring DATA !!!***
- **\*\*\*\*\***
- ***Needed a random collection of 'soil samples' from a sample of plots for estimation of % LSF Category in the whole collection of 337 plots***  
***How many such plots are to be examined?***  
***50/75/100/ ..who will guide the experimenter ?***  
***Of course....a statistician is equipped with valid and adequate methodology to provide solutions...***

### Sample Size Determination...

- UNKNOWN % of LSF plots in the study site : 100P%

Estimate from a random collection of 'n' sample plots = 100p%.....this 'p' is sample proportion of LSF plots

Fact : Larger the sample size 'n', 'closer' is the sample proportion 'p' to population proportion 'P'.

*However, we can't increase 'n' indefinitely.....*

Be satisfied with  $[-d < P - p < d]$  for a given 'accuracy level d' ...naturally small.....say,  $d=0.10$ , or 0.05 or even 0.01. It is clear that smaller the 'd', higher the sample size 'n'.....

Fix a value of 'd' from practical considerations and figure out the value of 'n' so that the estimate 'p' covers the true unknown proportion P within the accuracy level 'd'.

### Sample Size Determination....

Even this can't be ensured all the time.....the statisticians usually attach a high-level of confidence to this value of 'n'....say on 95% of the occasions this should work...as stipulated.....

**Thumb Rule :  $n = 1/d^2$**

$d = 0.10 = 1/10 \dots n = 100$ ;

$d = 0.05 = 1/20 \dots n = 400^*$  [more than 337 ?]

**Interpretation :** If we base our study on a random sample of size  $n=100$  and come up with sample proportion of 23% i.e., 0.23, then we expect to capture the true and unknown popl. proportion in the range  $[p \pm d]$  i.e.,  $0.23 \pm 0.10$  i.e., between 13% & 33% with 95% chance.....so we may miss out 5% times.

## **Finite Population Correction [fpc]**

**In case the population has a finite collection of plots [as is usually the case], we make a correction to the sample size ....**

**This is called 'finite population correction' [fpc].**

**$n(i)$  = initial sample size &  $n(f)$  = final sample size**

**$n(f) = n(i) \times fpc$  where  $fpc = N / [N + n(i)]$**

**For  $n(i) = 400$  and  $N = 337$ ,**

**$n(f) = 400 \times 337 / [400 + 337] = 182.9 \dots 183$ .**

**For  $n(i) = 100$  and  $N = 337$ ,**

**$n(f) = 100 \times 337 / [100 + 337] = 77.12 \dots 78$ .**

**We draw a random sample of  $n(f)$  plots.....**

## **Illustrative Examples....**

**For a sample collection of 70 plots, there are 28 LSF plots.**

**So, sample proportion of LSF plots = 0.40 [40%]**

**(a) Estimate of Popl. Prop. of LSF plots = 40%**

**(b) 95% CI for  $P = 0.40 \pm 2 \times \sqrt{[0.4 \times 0.6 / 70]}$   
= [ 0.40  $\pm$  0.12 ] = [0.28, 0.52].**

**Unknown popl proportion of LSF plots is most likely between 28% & 52% and this is likely to hold with 95% confidence.....**

**Here we start with a given sample size and assess the performance of the sample proportion.**



## Sample Size Determination....

Be satisfied with

$$[-d < \text{Sample Prop.} - \text{Popl. Prop.} < d]$$

for a given 'accuracy level d' ...naturally small.....

say,  $d=0.10$ , or  $0.05$  or even  $0.01$ . It is clear that smaller the 'd', higher the sample size 'n'.....

Fix a value of 'd' from practical considerations and figure out the value of 'n' so that the sample prop. covers the true unknown popl. prop. within the accuracy level 'd'. Even this can't be ensured all the time.....the statisticians usually attach a high-level of confidence to this value of n.....

## More Examples....

Ex. 1. There are 138 small plots for cultivation and the farmers are worried about soil salinity. It is required to estimate the proportion of plots with *excessive soil salinity* within  $\pm 0.05$  accuracy. Determine the sample size.

Sol. Req'd.  $-0.05 < P-p < 0.05$  i.e.,  $d=0.05 = 1/20$

Therefore,  $n(I) = 1/d^2 = 400$ .

Hence,  $n(f) = 400 \times [138/(138+400)] = 103$ .

For 10% accuracy i.e.,  $d = 0.10$

$n(I)=100$  and  $n(f) = 58$ .

For higher d values....sample size decreases.....

### Have some prior knowledge about 'P' ?

**Ex. 2.** There are 230 small plots for cultivation and the farmers are worried about soil fertility. It is required to estimate the proportion of plots with *low soil fertility* within  $\pm 0.05$  accuracy. *Determine the sample size if it is believed that low soil fertility does not cover more than 20% of the plots.*

**Sol.** Req'd.  $-0.05 < P-p < 0.05$  i.e.,  $d=0.05 = 1/20$

However,  $P < 20\% = 0.20 = P^*$  [prior guess].

Hence,  $n(I) = 4P^*(1-P^*)/d^2 = 4 \times 0.2 \times 0.8 \times 400 = 256$

And hence  $n(f) = 256 \times [230/(230+256)] = 122$

For 10% accuracy i.e.,  $d = 0.10$

$n(I) = 4 \times 0.2 \times 0.8 \times 100 = 64$  and  $n(f) = 51$ .

For higher d values....sample size decreases.....

### Different Form of Prior Knowledge ?

**Ex. 3.** There are 320 small plots for cultivation and the farmers are worried about water scarcity. It is required to estimate the proportion of plots with *low water retention capacity* within  $\pm 0.05$  accuracy. *Determine the sample size if it is believed that such a proportion is likely to be in between 25 % and 40%.*

**Sol.** Req'd.  $-0.05 < P-p < 0.05$  i.e.,  $d=0.05 = 1/20$

However,  $25\% < P < 40\% = 0.40 = P^*$  [+ side prior guess].

Hence,  $n(I) = 4P^*(1-P^*)/d^2 = 4 \times 0.4 \times 0.6 \times 400 = 384$

And hence  $n(f) = 384 \times [320/(384+320)] = 175$ .

For 10% accuracy i.e.,  $d = 0.10$

$n(I) = 4 \times 0.4 \times 0.6 \times 100 = 96$  and  $n(f) = 74$ .

For higher d values....sample size decreases.....

### Formulae.....

- $n = 4P^*(1-P^*)/d^2$
- $0 \text{-----} P^* \text{-----} 1$
- **P\* : Prior Guess closest to 0.5**
- 1.  $P < 0.32 \text{.....} P^* = 0.32 \text{ -----} < 0.32$
- 2.  $P > 0.67 \text{.....} P^* = 0.67 \text{ 0.67} > \text{-----}$
- 3.  $P < 0.57 \text{.....} P^* = 0.5 \text{ -----} < 0.57$
- 4.  $0.43 < P < 0.61 \text{.....} P^* = 0.5 \text{ ---} < 0.43 \text{----} 0.61 > \text{---}$
- 5.  $P > 0.53 \text{.....} P^* = 0.53 \text{ >0.53} \text{-----}$
- 6.  $0.31 < P < 0.43 \text{.....} P^* = 0.43 \text{ <0.31} \text{-----} 0.43 >$
- **No Prior Knowledge.....use  $P^* = 0.5 \text{....} n = 1/d^2$**

### Sample Size Determination....contd.

**So much for determination of sample size (n) for estimation of unknown Popl Proportion (P) based on sample-based proportion (p) with a margin of error '+/- d' – with / without any prior knowledge.**

**We now study a different but related problem.**

## Understanding the Average Salinity Level of Plots

- **Background :** In an agricultural survey, there are  $N=107$  wheat-growing agricultural plots in a district and one variety of wheat has to be grown all over. The problem is to gather knowledge about the average salinity level  $[YBAR]$  of the plots for good production of wheat in the said district. One way would be to go 100 % and collect data on salinity from every plot and compute the average salinity level per plot i.e.,  $YBAR$ .
- An alternative is to select a sample of  $(n)$  plots and collect salinity data from the selected plots and compute  $ybar$ , the sample average salinity level of the plots. How close would this be to  $YBAR$  ? What should be the sample size  $(n)$  for  $ybar$  to be 'close enough' to  $YBAR$  ?
- We must qualify the phrase 'close enough' .....

## Sample Size Determination....

Population Mean  $YBAR(N)$  & Sample mean  $ybar(n)$

What is expected of  $ybar(n)$  ?

$| \text{sample mean} - \text{popl. mean} | < \text{an acceptable qty}$

Watch out :  $y$  is quantitative in nature and there is likely to be some unit of measurement attached to it.....ml/H etc

$| ybar(n) - YBAR(N) | < \text{an acceptable qty} = 0.01, 0.05, 0.10 ?$

These statements do NOT make any sense here.....

Needed a meaningful formulation like

$| ybar(n) - YBAR(N) | < \text{a fraction of } YBAR = dYBAR$   
where  $d = 0.01, 0.05, 0.10$  etc etc

***Relative Abs. Error not to exceed '+/- d'.....this is OK.***

## Sample Size Formula

- The defining equation in this formulation for sample size 'n' is given by :
- $d\sqrt{(n)} / cv = 2$  i.e.,  $n = 4 cv^2 / d^2$
- where cv = coeff. of variation which expresses sd of observations in terms of average of the observations. Usually, experimenters have some idea about the cv.
- Example : cv = 10% and d = 5% = 0.05 ..n=16;
- cv = 20% and d = 5% = 0.05 ..n = 64 etc etc
- Recall fpc to convert n(I) to n(f).

## Understanding the Variation in Salinity Levels of Plots

- Change the Background :
- In an agricultural survey, there are N=107 wheat-growing agricultural plots in a district and one variety of wheat has to be grown all over. The problem is to gather knowledge about the variation in the salinity level [SD] of the plots for good production of wheat in the said district. One way would be to go 100 % and collect data on salinity from every plot and compute the SD of salinity level per plot.

### Sampling Problem.....

- An alternative is to select a sample of (n) plots and collect salinity data from the selected plots and compute sd, the sample standard deviation of salinity levels of the plots. How close would this be to SD ? What should be the sample size (n) for sd to be 'close enough' to SD ?
- We must qualify the phrase 'close enough' meaningfully as
- $|sd(n) - SD(N)| < d \text{ times } SD(N)$  where  $d = 0.01, 0.05, 0.10$  etc etc

### Sample Size Determination....

- That means :
- $(1-d)SD(N) < sd(n) < (1+d)SD(N)$
- In other words :
- $(1-d) < sd(n)/SD(N) < 1+d$
- That is :  
 $(1-d)^2 < [sd(n)/SD(N)]^2 < (1+d)^2$   
Calls for an approx. solution to be derived from  
 $n(1-d)^2 < \text{Chi-Square}(n) < n(1+d)^2$   
which leads to :  $2 = \text{sqrt}(n/2)[2d+d^2]$  and  
 $2 = \text{sqrt}(n/2)[2d - d^2]$   
and hence finally to.... $n = 8/[(2d+d^2)(2d-d^2)]$  approx.

## Sample Size Determination

$$n = 8/[(2d+d^2)(2d-d^2)] = n(I)$$

**d = margin of rel. error**

**Recall :  $n(f) = n(I)$  times  $[N/(N+n(I))]$**

**Examples : Consider  $N = 107$**

**$d=0.01 : n(I) = 1992 \dots n(f) = 1992 \times 107 / 2099 = 102$**

**$d=0.05 : n = 800 \dots n(f) = 800 \times 107 / 907 = 95$**

**$d=0.10 : n=200 \dots n(f) = 200 \times 107 / 307 = 70$**

**Interpretation : For  $d = 0.10$ , if  $sd(n)=23.67$  is the sample sd, then we can say with 95% confidence that popl. SD(N) is likely to be between  $sd/(1+d)$ ,  $sd/(1-d)$  which yields :  $23.67/(1.1)$ ,  $23.67/0.9$  i.e., (21.52, 26.3).**

## **Application of Multiple Criteria Decision Making Approach in Agriculture**

Anurup Majumder  
Professor, Department of Agricultural Statistics,  
Faculty of Agriculture  
Bidhan Chandra Krishi Viswavidyalaya  
P.O. Krishi Viswavidyalaya, Dist. Nadia  
West Bengal, India, Pin: 741 252  
E. mail: [anurup.majumder@rediffmail.com](mailto:anurup.majumder@rediffmail.com)  
& [anurupbckv@gmail.com](mailto:anurupbckv@gmail.com)

### **Introduction:**

- Firstly, we may go to a common example. In any class room, someone is good in subject English, someone is good in Mathematics, etc. But if we want to select top students from the class, we take the help of grand total, which is a unique index for selection. The job of selection is very easy as all the subjects are judged by a single scale of 100 marks.
- Next, we start with an agricultural example. In India, several varieties of banana are cultivated with different genomes. The quality of the fruit and its acceptability to market of banana are governed by the joint contribution of different characters viz., phenotypes, genotypes and biochemical. These characters are different for different varieties and measured in different scales.
- Suppose I want judge the varieties and give ranks to the varieties according to their performances.



- Several authors, for many years, have tried for identification of a group of varieties suitable for a specific zone with the help of either using ANOVA technique ( Singh, 2005), herein after referred to as the **first method**, for selecting the better performers for each character under study. Or, clustering the varieties using Euclidean distance matrix and dendograms ( Mandal, 2005), herein after referred to as the **second method**, considering multiple characters under study at a time. The second method is comparatively better in the sense that the final decision includes all the characters at a time (using clustering technique), similar type of varieties being grouped in a single cluster.

- It may be observed that assessment on the basis of overall performance has not been done by any of the methods. A single index, representative of the whole set of characters under study is needed to evaluate the overall varietal performances comprehensively and easily. Such an index can usually be compiled as a function of the whole set of varietal characters. Statistical techniques can be used to construct a single integrated index with suitable robustness properties for scientific judgment about the varieties.
- Multiple criteria decision making (MCDM) is a body of techniques that can be used for making assessments in the presence of multiple characters. This approach of decision making with multiple indicators is not new. References in the area include Zeleny (1992), Hwang and Yoon ( 1981), Yoon and Hwang(1995) etc.. Filar, et. al. (2003) used the MCDM techniques called as TOPSIS (Technique for order of preference by similarity to ideal solution) method for environmental assessment based on multiple indicators.

- Again we return to the problem of ranking of banana varieties. Suppose, there are 27 varieties of desert banana which were selected from 30 banana germplasms of Mondal (2005), for the study. Mondal (2005) recorded the observations on 16 morphological and biochemical characters of the above 27 varieties. Let the varieties be denoted by  $S_1, S_2, \dots, S_{27}$  for the convenience and easy understanding of the technique. We also focus on the 16 characters for each variety and for simplicity, these characters are denoted as  $C_1, C_2, \dots, C_{16}$ .

- **The TOPSIS Method and Related Topics:** The basic principle employed by TOPSIS is that the best alternative should have the shortest distance from the **ideal alternative** and the farthest distance from the **negative –ideal alternative**, which is both **intuitive** and **effective**.
- 
- **MCDM Approach:** Suppose there are altogether  $K$  alternatives to be assessed and the best alternative is to be selected. Let the alternatives be denoted by  $S_1, S_2, \dots, S_K$ . There are also  $N$  criteria identified to assess the alternatives, which are denoted by  $C_1, C_2, \dots, C_N$ . The  $k$ th alternative's value on the  $n$ th criteria is obtained as  $x_{kn}$  and we write  $S_k = (x_{k1}, x_{k2}, \dots, x_{kN})$  and  $C_n = (x_{1n}, x_{2n}, \dots, x_{kn})$ ;  $k = 1, 2, \dots, K$  and  $n = 1, 2, \dots, N$ . In matrix form, it will be:

	<b>C1</b>	<b>C2</b>	<b>---</b>	<b>CN</b>
<b>S1</b>	$x_{11}$	$x_{12}$	$\dots$	$x_{1N}$
<b>S2</b>	$x_{21}$	$x_{22}$		$x_{2N}$
<b>•</b>	$\dots$			
<b>SK</b>	$x_{K1}$	$x_{K2}$		$x_{KN}$

- **The ideal solution:** The **ideal alternative**  $S_+ = (x_{+1}, x_{+2}, \dots, x_{+N})$  and the **negative-ideal alternative**  $S_- = (x_{-1}, x_{-2}, \dots, x_{-N})$  are formed by taking all the best values attained on each criteria by some alternatives and all the worst values attained on each criteria by some alternatives, respectively.
- **The TOPSIS procedure:** With the above notation and explanation, the TOPSIS procedure for assessing the ranking of the K alternatives based on their values on the N criteria can be described as follows.
- Firstly, the nth criteria vector  $C_n$  is normalised as  $TC_n$ , where
- $TC_n = C_n / |C_n| = (x_{1n} / |C_n|, x_{2n} / |C_n|, \dots, x_{kn} / |C_n|) = (t_{1n}, t_{2n}, \dots, t_{kn})$ ,  $n=1, 2, \dots, N$ . Where  $|C_n| = \sqrt{(\sum_{k=1}^K (x_{kn})^2)}$  is the euclidian length or norm of  $C_n$ , so the new criteria vectors have the same length and are thus unit free and directly comparable. Accordingly, the kth alternative vector  $S_k$ , the ideal solution  $S_+$  and the negative-ideal solution  $S_-$  are also transformed to  $TS_k$ ,  $TS_+$  and  $TS_-$ , respectively.

- Next,  $d(S_k, S_+)$  is defined as the weighted Euclidean distance of  $TS_k$  from  $TS_+$ ;
- $d(S_k, S_+) = \| w \bullet (TS_k - TS_+) \|$ , where  $\bullet$  is the vector product and  $w$  is the weight.
- $= \sqrt{(\sum_{n=1}^N (W_n (t_{kn} - t_{+n}))^2)}$ .
- Similarly,  $d(S_k, S_-)$  is defined as
- $d(S_k, S_-) = \sqrt{(\sum_{n=1}^N (W_n (t_{kn} - t_{-n}))^2)}$ .
- Finally, the  $K$  alternatives are ranked in order of performance by their relative closeness to the ideal solution  $S_+$ , which for the  $k$ th alter native is given below.
- $r(S_k, S_+) = d(S_k, S_+) / [d(S_k, S_+) + d(S_k, S_-)]$ .
- The assessment criteria of TOPSIS is based on the consideration that the smaller is the value of  $r(S_k, S_+)$ , the more is the preferred alternative.

- **Choice of weights:** To obtain the internal importance or weights, we use the entropy concept. It is a criterion for the amount of information ( or uncertainty ) represented by a discrete probability distribution,  $p_1, p_2, \dots, p_k$  and this measure of information was given by Shanon and Weaver (1947) as  $E(p_1, p_2, \dots, p_k) = - \phi_k \sum_{k=1}^K p_k \ln(p_k)$ , where  $\phi_k = 1 / \ln(p_k)$  is a positive constant ranges from 0 to 1.
- Now assuming that  $p_{kn} = x_{kn} / X_n$ , where  $X_n = x_{k1} + x_{k2} + \dots + x_{kN}$  as the probability distribution of  $C_n$  on the  $K$  alternatives, we may similarly define the entropy of  $C_n$  as  $E(C_n) = - \phi_k \sum_{k=1}^K p_k \ln(p_k) = - \phi_k \sum_{k=1}^K (x_{kn} / X_n) \ln(x_{kn} / X_n)$ ,  $n = 1, 2, \dots, N$  and lastly, define the weights as  $W_n = (1 - E(C_n)) / \sum_{j=1}^N (1 - E(C_j))$ ,  $n = 1, 2, \dots, N$ .

- **Application of the technique in Agriculture:** We have applied the technique in agriculture. The abovementioned 27 banana varieties are assessed on the basis of all 16 characteristics or parameters by the TOPSIS method. The result is shown in table 1. Earlier we also used the technique on nine groundnut varieties. The result of groundnut is shown in table 2.

**Table1: Performance of desert banana varieties considering all the plant morphological, fruit quality and biochemical characters.**

Varieties	Code	D(Sk,S+)	D(SK,S-)	R(Sk,S+)	Rank
AGNISWAR	S1	0.1597	0.1557	0.5063	11 <sup>th</sup>
AMRIT SAGA	S2	0.1694	0.1594	0.5153	10 <sup>th</sup>
KABULI	S3	0.1404	0.1656	0.4588	5 <sup>th</sup>
JAHAJI	S4	0.1430	0.1471	0.4929	7 <sup>th</sup>
ROBUSTA	S5	0.1358	0.1869	0.4209	3 <sup>rd</sup>
GIANT GOV'NOR	S6	0.1157	0.1962	0.3708	2 <sup>nd</sup>
MALBHOG	S7	0.1559	0.1303	0.5446	14 <sup>th</sup> (1)
MARTAMAN	S8	0.1874	0.1081	0.6342	25 <sup>th</sup>
MARTAMAN CLO	S9	0.1462	0.1494	0.4947	8 <sup>th</sup>
RAMPAL MA'MAN	S10	0.1910	0.1358	0.5845	18 <sup>th</sup>
KANAIBASI	S11	0.1720	0.1149	0.5995	21 <sup>st</sup>
DOODSAGAR	S12	0.1611	0.1241	0.5649	16 <sup>th</sup>
SABRI	s13	0.1670	0.1206	0.5808	17 <sup>th</sup>
CHAMPA	s14	0.1683	0.1044	0.6173	24 <sup>th</sup>

**Table1: Performance of desert banana varieties considering all the plant morphological, fruit quality and biochemical characters (Contd.)**

Varieties	Code	D(Sk,S+)	D(SK,S-)	R(Sk,S+)	Rank
CHINIA	S15	0.1641	0.1125	0.5933	20 <sup>th</sup>
KATALI CHAMPA	S16	0.1867	0.0954	0.6619	26 <sup>th</sup>
CHINI CHAMPA	S17	0.1865	0.1241	0.6004	22 <sup>nd</sup>
ATTIAKOLE	S18	0.1859	0.1276	0.5931	19 <sup>th</sup>
KALIBOW	S22	0.1533	0.1282	0.5446	14 <sup>th</sup> (2)
KATALI	S23	0.1616	0.1458	0.5257	12 <sup>th</sup>
KATALI CLONE	S24	0.1642	0.1417	0.5367	13 <sup>th</sup>
KALIBHOG	S25	0.1258	0.1670	0.4296	4 <sup>th</sup>
GERMAN KATALI	S26	0.1426	0.1447	0.4964	9 <sup>th</sup>
<b>BAGDA KATALI</b>	<b>S27</b>	<b>0.0967</b>	<b>0.2169</b>	<b>0.3083</b>	<b>1<sup>st</sup></b>
KRISNA KATALI	S28	0.1384	0.1458	0.4870	6 <sup>th</sup>
MANUA	S29	0.1770	0.1127	0.6110	23 <sup>rd</sup>
<b>MADHUBASH</b>	<b>S30</b>	<b>0.2250</b>	<b>0.0650</b>	<b>0.7758</b>	<b>27<sup>th</sup></b>

**Table 2: Performance of 9 Groundnut varieties.**

Name of the Variety	Distance from Ideal Solution	Distance from Negative Ideal Solution	Score of Ranking Index	Rank
D(S1,S+)	0.055	0.090	0.379	7.000
D(S2,S+)	0.047	0.107	0.306	6.000
D(S3,S+)	0.031	0.127	0.193	3.000
D(S4,S+)	0.118	0.040	0.745	8.000
<b>D(S5,S+)</b>	<b>0.139</b>	<b>0.024</b>	<b>0.851</b>	<b>9.000</b>
D(S6,S+)	0.021	0.126	0.146	2.000
D(S7,S+)	0.037	0.112	0.248	4.000
<b>D(S8,S+)</b>	<b>0.020</b>	<b>0.136</b>	<b>0.130</b>	<b>1.000</b>
D(S9,S+)	0.045	0.105	0.299	5.000
D(S1,S+)	0.055	0.090	0.379	7.000
D(S2,S+)	0.047	0.107	0.306	6.000

References:

- Filar, J.A.; Ross, N.P. and Wu, M.L. ( 2003) : Environmental assessment based on multiple indicators. *Cal. Statist. Assoc. Bull.* 54(March-June), pages: 93- 99.
- Hwang, C.L. and Yoon, K. ( 1981) : Multiple attribute decision making : methods and applications, a state of the art survey. Springer- Verleg, Berlin.
- Mandal, K.K. (2005): Characterization of diversity of Genus *Musa* in West Bengal. Ph.D. Thesis, Submitted under Faculty of Horticulture, BCKV, Mohanpur, India.
- Singh, D.B. (2005): Comparative performance of desert banana cultivars in Bay Islands. *The Horticultural Journal*; 18(2), pages: 80- 83.
- Sinha, B.K. and Shah, K. (2003): On some aspects of data integration techniques with Environmental applications. *Environmetrics*, 14, pages: 409- 414.
- Yoon K, Hwang C.L.(1995): Multiple Attribute Decision Making: An Introduction. Sage Publications.
- Zeleny, M. ( 1992) : Multiple criterion decision making. McGraw Hill , New York.







Manual

**NATIONAL WORKSHOP CUM TRAINING PROGRAMME**  
ON  
**STATISTICAL TOOLS FOR RESEARCH DATA ANALYSIS (Series II)**

**DURATION OF THE PROGRAMME: TWO WEEKS**

From 29<sup>th</sup> May, 2017 to 9th June, 2017

Organized By

**SOCIETY FOR APPLICATION OF STATISTICS IN AGRICULTURE  
AND ALLIED SCIENCES (SASAA)**

and

**DEPARTMENT OF AGRICULTURAL STATISTICS  
BIDHAN CHANDRA KRISHI VISWAVIDYALAYA**

**List of Sponsorship:**



**Venue**

**DEPARTMENT OF AGRICULTURAL STATISTICS,  
FACULTY OF AGRICULTURE,  
BIDHAN CHANDRA KRISHI VISWAVIDYALAYA,  
P.O.- KRISHI VISWAVIDYALAYA, NADIA, WEST BENGAL, India- 741252**