

# Reference Manual

## संदर्भ संहिता

मानव संसाधन प्रबंधन इकाई (एच.आर.एम.यूनिट), कृषि शिक्षा विभाग, भा०कृ०अनु०प०  
HRM Unit, Agricultural Education Division, ICAR

कृषि में आँकड़ों के विश्लेषण के लिए सांख्यिकीय तकनीकें  
Statistical Techniques for Data Analysis in Agriculture

04 से 13 अक्टूबर, 2021

04 – 13 October, 2021

पाठ्यक्रम समन्वयक:

डॉ. अजीत, डॉ. रंजीत कुमार पॉल, डॉ. सौमेन पाल

Course Coordinators:

Dr. Ajit, Dr. Ranjit Kumar Paul, Dr. Soumen Pal

भा०कृ०अनु०प०- भारतीय कृषि सांख्यिकी अनुसंधान संस्थान,  
लाइब्रेरी एवेन्यू, पूसा, नई दिल्ली-110012



ICAR- Indian Agricultural Statistics Research Institute  
Library Avenue, PUSA, New Delhi – 110012



हर कदम, हर डगर  
किसानों का हमसफर  
भारतीय कृषि अनुसंधान परिषद

Agrisearch with a human touch



कृषि में आँकड़ों के विश्लेषण के लिए सांख्यिकीय तकनीकें  
**Statistical Techniques for Data Analysis in Agriculture**

Under the aegis of  
**HRM Unit, Agricultural Education Division, ICAR**

04 – 13 October, 2021

Reference Manual

Course Coordinators:  
**Dr. Ajit, Dr. Ranjit Kumar Paul, Dr. Soumen Pal**

**ICAR- Indian Agricultural Statistics Research Institute**  
**Library Avenue, PUSA, New Delhi – 110012**

**Editors:**

Dr. Ajit

Dr. Ranjit Kumar Paul

Dr. Soumen Pal

**Disclaimer:** The information contained in this reference manual has been taken from various web resources. Respective URLs are mentioned in the content.

**Citation:**

Ajit, Ranjit Kumar Paul & Soumen Pal. (2021). **Statistical Techniques for Data Analysis in Agriculture**. Reference Manual, ICAR-Indian Agricultural Statistics Research Institute, New Delhi.

## आमुख

भा.कृ.अनु.प.- भा.कृ.सां.अ.सं. , सांख्यिकीय विज्ञान (सांख्यिकी, संगणक अनुप्रयोग और जैव सूचना विज्ञान) में प्रासंगिक कार्यों में संलग्न एक प्रमुख संस्थान है और कृषि अनुसंधान की गुणवत्ता को समृद्ध करने और नीतिगत निर्णय लेने के लिए कृषि विज्ञान में इनके विवेकपूर्ण संलयन में इसका प्रमुख योगदान है। 1930 में अपनी स्थापना के बाद से, तत्कालीन इंपीरियल काउंसिल ऑफ एग्रीकल्चरल रिसर्च के एक छोटे सांख्यिकीय खंड के रूप से, संस्थान का कद ऊंचा उठा और राष्ट्रीय और अंतरराष्ट्रीय स्तर पर अपनी उपस्थिति दर्ज कराने में सक्षम हुआ। संस्थान बहुत सक्रिय रूप से एड्वाइजरी सर्विस प्रदान कर रहा है जिसने संस्थान को राष्ट्रीय कृषि अनुसंधान और शिक्षा प्रणाली (NARES) और राष्ट्रीय कृषि सांख्यिकी प्रणाली (NASS) दोनों में अपनी उपस्थिति दर्ज कराने में सक्षम बनाया है। संस्थान ने एन.ए.आर.ई.एस. में एक उच्च स्तरीय सांख्यिकीय कंप्यूटिंग प्लेटफॉर्म बनाने में अग्रणी भूमिका निभाई है।

अनुसंधान कार्यों से लिए गए सांख्यिकीय रूप से मान्य और सार्थक निष्कर्ष गुणवत्तापूर्ण शोध की नींव रखते हैं और नीति नियोजन में विशेष रूप से विकासात्मक गतिविधियों और कार्यक्रम कार्यान्वयन में एक महत्वपूर्ण भूमिका निभाते हैं। इसलिए, यह आवश्यक है कि डेटा के संग्रह और विश्लेषण के लिए ठोस सांख्यिकीय पद्धति अपनाई जाए। संस्थान द्वारा आयोजित प्रशिक्षण कार्यक्रम कृषि विज्ञान के अनुसंधान और योजना में वास्तविक उपयोगकर्ताओं के लिए सांख्यिकीय तकनीकों में प्रगति का मूल्यांकन करने में बहुत उपयोगी हैं।

कृषि में आंकड़ों के विश्लेषण के लिए सांख्यिकीय तकनीकों पर वर्तमान प्रशिक्षण कार्यक्रम विशेष रूप से संकाय सदस्यों के और साथी प्रतिभागियों के बीच इंटरैक्टिव माध्यम से अधिकतम शैक्षणिक लाभ प्राप्त करने के लिए तैयार किया गया है। मुझे विश्वास है कि इस प्रशिक्षण कार्यक्रम से प्राप्त ज्ञान प्रतिभागियों को सांख्यिकीय प्रक्रियाओं की बेहतर समझ रखने में सक्षम करेगा, जिससे उन्हें उपयुक्त और आधुनिक सांख्यिकीय पद्धतियों का उपयोग करके डेटा सेट का विश्लेषण करने में भी लाभ होगा।

पाठ्यक्रम सामग्री सिद्धांत और अनुप्रयोग के मध्य समाहित है। विषयवस्तु पांच मॉड्यूल के अंतर्गत रखे गए हैं: (i) प्रारंभिक सांख्यिकी तकनीकें, (ii) सांख्यिकीय मॉडल (iii) सांख्यिकीय टूल/विश्लेषण (iv) प्रतिदर्श सर्वेक्षण और परीक्षण अभिकल्पना का विश्लेषण (v) मल्टीवेरियट और मशीन लर्निंग तकनीक।

इस पाठ्यक्रम में शामिल संकाय सदस्य कृषि सांख्यिकी और कंप्यूटर अनुप्रयोगों के क्षेत्र में सुस्थापित वैज्ञानिक हैं जो सॉफ्टवेयर पैकेजों के प्रतिपादन में विशेषज्ञता रखते हैं। संदर्भ संहिता में दिए गए व्याख्यान नोट्स विषय का विवरण प्रदान करते हैं। मुझे आशा है कि संदर्भ संहिता प्रतिभागियों के लिए काफी उपयोगी होगी। मैं इस अवसर पर पूरी फैकल्टी को उत्कर्ष कार्य करने के लिए धन्यवाद देता हूँ। मैं इस उपयोगी दस्तावेज को समय पर प्रकाशित करने के लिए इस प्रशिक्षण कार्यक्रम के पाठ्यक्रम समन्वयक डॉ. अजीत, डॉ. रंजीत कुमार पॉल और डॉ. सौमेन पाल को अपनी शुभकामनाएं देता हूँ। इस संदर्भ संहिता को और बेहतर बनाने के लिए आप सभी के सुझावों का स्वागत है।

नई दिल्ली  
01 अक्टूबर, 2021

(राजेन्द्र प्रसाद)  
निदेशक, भा.कृ.अनु.प.- भा.कृ.सां.अ.सं.



## FOREWORD

---

ICAR-IASRI is a premier Institute of relevance in Statistical Sciences (Statistics, Computer Applications and Bioinformatics) and their judicious fusion in agricultural sciences for enriching quality of agricultural research and informed policy decision making. Ever since its inception in 1930, as a small Statistical Section of the then Imperial Council of Agricultural Research, the Institute has grown in stature and made its presence felt both nationally and internationally. The Institute has been very actively pursuing advisory service that has enabled the institute to make its presence felt both in National Agricultural Research and Education System (NARES) and National Agricultural Statistics System (NASS). The Institute has taken a lead in creating a high end statistical computing environment in NARES.

Statistically valid and meaningful inferences drawn from research programmes lay the foundations of quality research and play a pivotal role in policy planning especially in developmental activities and in programme implementation. Therefore, it is essential that sound statistical methodologies be adopted for the collection and analysis of data. The training programmes organized by the Institute are very useful in appraising the advances in statistical techniques to the actual users in research and planning of agricultural sciences.

The present training programme on **Statistical Techniques for Data Analysis in Agriculture** has been especially designed to drive the maximum academic advantage through interaction with the faculty and among the fellow participants. I am sure that the knowledge assimilated from this training programme will enable the participants to have better understanding of statistical procedures, which will also benefit them in analyzing the data sets by using appropriate and modern statistical methodologies.

The course contents are intertwining of theory and application. The topics are covered under five modules: (i) Preliminary Statistics, (ii) Statistical Models (iii) Statistical Tool/analysis support (iv) Sample Survey and Analysis of Experimental Designs (v) Multivariate and Machine Learning Techniques.

The faculty for this course comprises of well-established scientists in the discipline of Agricultural Statistics and Computer Applications with expertise in handling software packages. The lecture notes given in the reference manual provide an exposition of the subject. I hope that the reference manual will be quite useful to the participants. I take this opportunity to thank the entire faculty for doing a wonderful job. I wish to complement Dr. Ajit, Dr. Ranjit Kumar Paul and Dr. Soumen Pal, Course Coordinators of this training programme, for bringing out this valuable document in time. We look forward to suggestions from every corner in improving this reference manual.

New Delhi  
01 October, 2021

*21/10/21*  
*11/10/21*  
(Rajender Parsad)  
DIRECTOR, ICAR-IASRI

## प्रस्तावना

भा.कृ.अनु.प. - भारतीय कृषि सांख्यिकी अनुसंधान संस्थान देश में कृषि सांख्यिकी, कंप्यूटर अनुप्रयोग और जैव सूचना विज्ञान विषयों में कार्य करने वाला एक प्रमुख संस्थान है। संस्थान परीक्षण अभिकल्पना, प्रतिदर्श सर्वेक्षण, सांख्यिकीय आनुवंशिकी, पूर्वानुमान तकनीक, जैव सूचना विज्ञान और संगणक अनुप्रयोगों पर विशेष जोर देने के साथ कृषि सांख्यिकी में अनुसंधान, शिक्षण और प्रशिक्षण कार्यक्रम आयोजित करने में कार्यरत है। संस्थान बहुत सक्रिय रूप से एड्वाइजरी सर्विस प्रदान कर रहा है जिसने संस्थान को राष्ट्रीय कृषि अनुसंधान और शिक्षा प्रणाली (NARES) और राष्ट्रीय कृषि सांख्यिकी प्रणाली (NASS) दोनों में अपनी उपस्थिति दर्ज कराने में सक्षम बनाया है। संस्थान ने कृषि अनुसंधान के लिए उपयोगी सांख्यिकीय सॉफ्टवेयर पैकेज विकसित करने में अग्रणी भूमिका निभाई है।

सांख्यिकीय तकनीकों में आधुनिक विकास के साथ प्रशिक्षित और पारंगत होने की मांग लगातार बढ़ रही है। कृषि के क्षेत्र में उन्नत सांख्यिकीय तकनीक इस संस्थान में अनुसंधान के महत्वपूर्ण विषयों में से एक है। सही तकनीक का चयन करने और अंतर्निहित सांख्यिकीय सिद्धांतों की प्रक्रिया को समझने में कृषि पेशेवरों और नीति निर्माताओं की भूमिका होती है। इसके अलावा, आर, पायथन आदि जैसे सॉफ्टवेयर कृषि प्रयोगों से निकले डेटा के विश्लेषण में प्रमुख भूमिका निभा रहे हैं। भा.कृ.अनु.प. - भारतीय कृषि सांख्यिकी अनुसंधान संस्थान 04-13 अक्टूबर 2021 (10 दिन) के दौरान भा.कृ.अनु.प.अथवा एस.ए.यू./सी.ए.यू./भा.कृ.अनु.प. वित्त पोषित केवीके में तकनीकी कर्मियों के लिए "कृषि में आंकड़ों के विश्लेषण के लिए सांख्यिकीय तकनीक" पर एक ऑनलाइन प्रशिक्षण कार्यक्रम का आयोजन कर रहा है। इस प्रशिक्षण कार्यक्रम की योजना कृषि डेटा के विश्लेषण के लिए सिद्धांत तथा अनुप्रयोगों दोनों की जरूरतों को समायोजित करने के लिए बनाई गई है। प्रशिक्षण कार्यक्रम का उद्देश्य प्रतिभागियों को उपयुक्त सांख्यिकीय तकनीकों का उपयोग करके कृषि आंकड़ों का विश्लेषण करने, सांख्यिकीय सॉफ्टवेयर पैकेजों के बारे में प्रतिभागियों को परिचित कराने और अनुसंधान गतिविधियों से जुड़े कौशल को बढ़ाने और उन्नत करने के लिए प्रशिक्षित करना है।

पाठ्यक्रम को कृषि सांख्यिकी के क्षेत्र में पारंपरिक और आधुनिक दोनों विषयों को शामिल करने के लिए संरचित किया गया है। इस पाठ्यक्रम के तहत विषय एक्सप्लोरेटरी डेटा विश्लेषण, सहसंबंध और प्रतिगमन, परिकल्पना परीक्षण, प्रतिगमन निदान, गैर-पैरामीट्रिक परीक्षण, लोजिस्टिक प्रतिगमन, रैखिक-समय श्रृंखला मॉडल, गैर-रेखीय समय-श्रृंखला मॉडल, सांख्यिकीय डेटा विश्लेषण के लिए एक्सेल, आर, पायथन, डिजाइन संसाधन सर्वर / एस.एस.सी.एन.ए.आर.एस, नमूनाकरण तकनीकों का अवलोकन, कृषि प्रयोगों के मूल डिजाइनों का विश्लेषण, स्प्लीट और स्ट्रिप प्लॉट डिजाइन, फैक्टोरियल प्रयोग, मशीन लर्निंग तकनीक: एएनएन, एसवीएम, रैंडम फॉरेस्ट आदि; क्लस्टर विश्लेषण, पीसीए और फैक्टर एनालिसिस शामिल हैं।

हम इस अवसर पर संस्थान के संकाय सदस्यों को धन्यवाद देना चाहते हैं जिन्होंने इस पाठ्यक्रम को सार्थक और सफल बनाने में अपना बहुमूल्य समय दिया जिससे इस मैनुअल को समय पर प्रकाशित करने में मदद मिली। हम इस प्रशिक्षण कार्यक्रम में अपने कर्मचारियों को प्रतिनियुक्त करने के लिए विभिन्न भा.कृ.अनु.प. संस्थानों और राज्य कृषि विश्वविद्यालयों के भी आभारी हैं। हम डॉ. राजेंद्र प्रसाद, निदेशक, भा.कृ.अनु.प.-भा.कृ.सां.अ.सं. के बहुमूल्य मार्गदर्शन और पाठ्यक्रम के सुचारू संचालन के लिए सभी आवश्यक सुविधाएं उपलब्ध कराने के लिए अत्यंत आभारी हैं। हम उन सभी के आभारी हैं जिन्होंने इस प्रशिक्षण मैनुअल को तैयार करने के लिए प्रत्यक्ष या अप्रत्यक्ष रूप से सहयोग दिया है।

(अजीत)

पाठ्यक्रम समन्वयक

(रंजीत कुमार पॉल)

पाठ्यक्रम समन्वयक

(सौमेन पाल)

पाठ्यक्रम समन्वयक



## PREFACE


---

The ICAR-Indian Agricultural Statistics Research Institute is a premier Institute in the disciplines of Agricultural Statistics, Computer Applications and Bioinformatics in the country. The Institute has been engaged in conducting research, teaching and organizing training programmes in Agricultural Statistics with special emphasis on Experimental Designs, Sampling Techniques, Statistical Genetics, Forecasting Techniques, Bioinformatics and Computer Applications. The Institute has been very actively pursuing advisory service that has enabled the institute to make its presence felt both in National Agricultural Research and Education System (NARES) and National Agricultural Statistics System (NASS). The Institute has taken a lead in developing Statistical Software Packages useful for Agricultural Research.

There is an ever-increasing demand to be trained and sensitized with recent developments in statistical techniques. Advanced statistical techniques in the domain of agriculture is one of the important subjects of research at the Institute. The agricultural professionals and policy makers have a role to play in selecting right technique and to understand the mechanisms of underlying statistical theories. Moreover, the software *like* R, Python *etc.* are playing major roles in analysing such data emerged from agricultural experiments. ICAR-Indian Agricultural Statistics Research Institute is organizing an online Training Programme on “Statistical Techniques for Data Analysis in Agriculture” for Technical personnel in ICAR or SAUs/CAUs/ICAR funded KVKs during 04–13 October 2021 (10 days). This training programme has been planned to accommodate the needs of both theory and applications for analyzing the agricultural data. The aim of the training programme is to train participants in analyzing agricultural data using appropriate statistical techniques, acquainting the participants about statistical software packages and to enhance and upgrade the skill associated with research activities.

The course is structured to include both conventional and recent topics in the field of Agricultural Statistics. The topics covered under this course include Exploratory Data Analysis, Correlation & Regression, Testing of Hypothesis, Regression Diagnostics, Non-Parametric Tests, Logistic Regression, Linear Time-Series Models, Nonlinear Time-Series Models, Excel for statistical data analysis, R, Python, Design Resources Server/ SSCNARS, Overview of Sampling Techniques, Analysis of Basic Designs of Experiments, Split and Strip Plot Designs, Factorial Experiments, Machine Learning Techniques: ANN, SVM, Random Forest *etc.*; Cluster Analysis, PCA and Factor Analysis.

We would like to take this opportunity to thank the faculty of the Institute who spared their valuable time in making this course meaningful and successful that helped in bringing out this manual in time. We are also thankful to the various ICAR Institutes and State Agricultural Universities for deputing their employees in this training programme. We are grateful to Dr. Rajender Parsad, Director, ICAR-IASRI for his valuable guidance and making all necessary facilities available for smooth conduct of the course. We are thankful to each and every one who supported directly or indirectly for preparing this training manual.

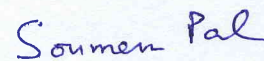


(Ajit)

Course Coordinator



(Ranjit Kumar Paul)  
Course Coordinator



(Soumen Pal)  
Course Coordinator

**Training Programme on  
“Statistical Techniques for Data Analysis in Agriculture”  
(04-10-2021 to 13-10-2021)**

**ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012  
Human Recourse Management Unit, ICAR**

**SCHEDULE OF DAILY LECTURES AND PRACTICALS**

<b>Date &amp; Time</b>	<b>Topic</b>	<b>Speaker</b>
<b>04-10-2021</b>	<b>Monday</b>	
10.00 - 10.30	Inaugural Function	
10.30 - 12.30	Exploratory Data Analysis	Dr. Ajit/ Md. Yeasin
14.30 - 16.30	Statistical Techniques in Excel	Mr. Upendra Pradhan
<b>05-10-2021</b>	<b>Tuesday</b>	
10.30 - 12.30	Overview of R Software	Dr. Soumen Pal
14.30 - 16.30	Correlation and Regression	Dr. R K Paul
<b>06-10-2021</b>	<b>Wednesday</b>	
10.30 - 12.30	Testing of Hypothesis	Mr. Prakash Kumar
14.30 - 16.30	Non parametric tests	Dr. Himadri Shekhar Roy
<b>07-10-2021</b>	<b>Thursday</b>	
10.30 - 12.30	Visual Representation of data and statistical graphs	Dr. Soumen Pal
14.30 - 16.30	Growth Models	Md. Yeasin
<b>08-10-2021</b>	<b>Friday</b>	
10.30 - 12.30	Time Series Models	Dr. R K Paul
14.30 - 16.30	Overview of Python Software	Md. Asraful
<b>11-10-2021</b>	<b>Monday</b>	
10.30 - 12.30	Overview of Sampling Techniques	Dr. Ankur Biswas
14.30 - 16.30	Tutorial-as per participants' interest	Dr. Ajit, Dr. R.K. Paul and Dr. Soumen Pal
<b>12-10-2021</b>	<b>Tuesday</b>	
10.30 - 12.30	Design Resources Server/Basic Designs	Dr. Rajender Parsad
14.30 - 16.30	Factorial Experiments including split and strip plot	Dr. S.K. Sarkar
<b>13-10-2021</b>	<b>Wednesday</b>	
10.30 - 12.30	Machine Learning Techniques	Dr. P K Meher
14.00 - 16.00	Multivariate Techniques	Dr. Samarendra Das
16.00 - 16.30	Valedictory Function	

## CONTENTS

<b>S. No.</b>	<b>Topic</b>	<b>Author</b>	<b>Page No.</b>
1.	Basic Statistical Technique	Md Yeasin, Ajit Gupta and Ranjit Kumar Paul	1-18
2.	An Introduction to Data Analysis Using Ms-Excel	Upendra Kumar Pradhan	19-42
3.	Basic Statistical Technique in Excel	Md Yeasin, Ajit Gupta and Ranjit Kumar Paul	43-49
4.	Overview on R Software and RStudio	Soumen Pal and B N Mandal	50-66
5.	Regression Analysis	Ranjit Kumar Paul	67-79
6.	Testing of Hypothesis	Prakash Kumar and Ranjit Kumar Paul	80-90
7.	Non-Parametric Tests	Himadri Shekhar Roy and Lalmohan Bhar	91-106
8.	Data Visualization Using R	Soumen Pal	107-114
9.	Nonlinear Growth Models	Ranjit Kumar Paul and Md Yeasin	115-128
10.	Linear Time Series Modelling	Ranjit Kumar Paul	129-146
11.	Overview of Python	Md. Ashraful Haque	147-160
12.	Overview of Sampling Methods	Ankur Biswas	161-172
13.	Design Resources Server	Rajender Parsad and V K Gupta	173-192
14.	Designs for Factorial Experiments	V.K.Gupta, Rajender Parsad, Sukanta Dash and Susheel Kumar Sarkar	193-224
15.	Multivariate Data Analysis Using R	Samarendra Das and A R Rao	225-231
16.	Machine Learning	Prabina Kumar Meher	232-250



---

---

# BASIC STATISTICAL TECHNIQUES

---

---

**Md Yeasin, Ajit Gupta, Ranjit Kumar Paul**  
*ICAR-Indian Agricultural Statistics Research Institute*  
*Library Avenue, New Delhi - 110 012*

[yeasin.iasri@gmail.com](mailto:yeasin.iasri@gmail.com) , [ajit@icar.gov.in](mailto:ajit@icar.gov.in) , [ranjit.paul@icar.gov.in](mailto:ranjit.paul@icar.gov.in)

---

---

## 1. Introduction

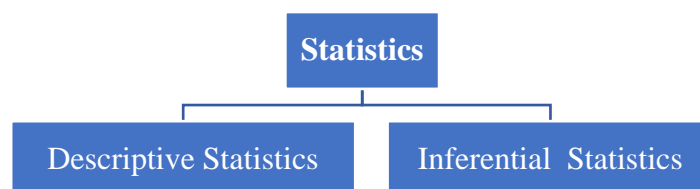
The word ‘Statistics’ has been derived from the Latin word ‘**Status**’ or the Italian word ‘**Statista**’ or the German word ‘**Statistik**’ each of which means ‘political state’. Statistics is a broad concept featuring applications in a wide range of areas. Statistics, in general, can be defined as the process for collecting, analyzing, interpreting, and making conclusions from data. In other terms, statistics is the approach established by scientists and mathematicians for analyzing and deriving conclusions from acquired data. Everything that has anything to do with the collection, processing, interpretation, and presentation of data falls within the scope of statistics.

**Definition of statistics:** Statistics is a branch of mathematics that deals with collecting, organizing, summarizing, presenting, and analyzing data as well as providing valid results and interpreting towards reasonable decisions.

Statisticians, in other words, give methodologies for

- **Design:** Planning and conducting out research projects.
- **Description:** Data summarization and exploration.
- **Inference:** Making predictions and inferences about the data

Statistics can be divided into two sections; one is descriptive statistics and another is inferential statistics.



**Descriptive statistics** helps describe, show or summarize data in a meaningful way. Descriptive statistics provides us with tools, tables, graphs, averages, ranges, correlations for organizing and summarizing data. Examples: measures of central tendency, measures of dispersion, skewness, kurtosis etc.

**Inferential statistics** helps to understand the properties of the population by observing the sample values. Inferential statistics deals with the estimation of parameters and test of hypothesis.

In this section we briefly discussed the descriptive statistics such as measures of central tendency, measures of dispersion, skewness, and kurtosis

### 2. Measures of central tendency

Central tendency is a statistical measure that determines a single value that accurately describes the center of the distribution. The objective of central tendency is to identify the single value that is the best representative for the entire set of data.

Different measure of central tendency are:

- Mean
  - Arithmetic mean
  - Geometric mean
  - Harmonic mean
- Median
- Mode
- Quartiles
- Deciles
- Percentiles

#### 2.1. Mean (Arithmetic mean: A.M.):

The mean is the most commonly used measure of central tendency. For computation of the mean data should be numerical values measured on an interval or ratio scale. To compute the mean, we add the observation of data sets and then divide by the number of observation.

$$\text{Mean} = \frac{\text{Sum of all observation}}{\text{Total number of observation}}$$

**2.1.1. Simple mean:** Let  $X_1, X_2, \dots, X_n$  are the  $n$  observation of a data set. The arithmetic mean is given by

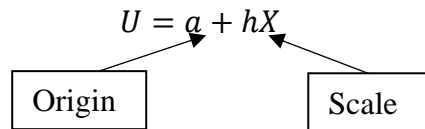
$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

**Mean for frequency distribution:** Let  $X_1, X_2, \dots, X_n$  are observations with corresponding frequencies are  $f_1, f_2, \dots, f_n$  and  $\sum_{i=1}^n f_i = N$ . The arithmetic mean is given by

$$\bar{X} = \frac{\sum_{i=1}^n f_i X_i}{N}$$

**Properties of mean:**

- It depends on change of origin as well as the change of scale.



Then  $\bar{U} = a + h\bar{X}$ .

- If are  $\bar{X}_1$  and  $\bar{X}_2$  the means of two sets of values with  $n_1$  and  $n_2$  observations respectively, then their combined mean is given by

$$\bar{X} = \frac{n_1 \bar{X}_1 + n_2 \bar{X}_2}{n_1 + n_2}$$

- Algebraic sum of deviations of set of values from their mean is zero.

$$\sum_{i=1}^n (X_i - \bar{X}) = 0$$

- The sum of squares of deviation of set of values about its mean is minimum

$$\sum_{i=1}^n (X_i - A)^2 \text{ is minimum when } A = \bar{X}$$

**Merits of mean:**

- Easy to understand
- Easy to calculate.
- It is rigidly defined.
- It is based on all observations.
- It is least affected by sampling fluctuations.
- It is capable of further mathematical treatment.

**Demerits of mean:**

- It is affected by extreme values.
- It cannot be calculated for open end class frequency distribution.
- It cannot be located graphically.
- It cannot be calculated for qualitative characteristic.

## Basic Statistical Techniques

- It cannot be calculated if any observations are missing in the data series.
- It is not suitable for highly skewed distribution.

### 2.1.2. Geometric mean (G.M.):

For  $n$  observations, Geometric mean is the  $n^{\text{th}}$  root of their product.

**For non-frequency data:** Let  $X_1, X_2, \dots, X_n$  are the  $n$  observation of a data set. The geometric mean is defined as

$$G = (X_1 * X_2 * \dots * X_n)^{1/n}$$

**For frequency distribution:** Let  $X_1, X_2, \dots, X_n$  are observations with corresponding frequencies are  $f_1, f_2, \dots, f_n$  and  $\sum_{i=1}^n f_i = N$ . The geometric mean is defined as

$$G = (X_1^{f_1} * X_2^{f_2} * \dots * X_n^{f_n})^{1/N}$$

Use of geometric mean:

- Measure average relative changes, averaging ratios and percentages
- Best average for construction of index number

### Merits of geometric mean:

- It is based on all observations.
- It is not affected by sampling fluctuations.
- It is capable of further mathematical treatment.

### Demerits of geometric mean:

- If any of the values is zero, it cannot be calculated.
- It is affected by extreme values.
- It cannot be calculated for open end class frequency distribution.
- It cannot be located graphically.
- It cannot be calculated for qualitative characteristic.
- It cannot be calculated if any observations are missing in the data series.

### 2.1.3. Harmonic mean (H.M.):

Harmonic mean is the reciprocal of the arithmetic mean of the reciprocals of the observations of the sets.

**For non-frequency data:** Let  $X_1, X_2, \dots, X_n$  are the  $n$  observation of a data set. The harmonic mean is defined as

## Basic Statistical Techniques

$$H = \frac{n}{\sum_{i=1}^n 1/X_i}$$

**For frequency data:** Let  $X_1, X_2, \dots, X_n$  are observations with corresponding frequencies are  $f_1, f_2, \dots, f_n$  and  $\sum_{i=1}^n f_i = N$ . The harmonic mean is defined as

$$H = \frac{N}{\sum_{i=1}^n f_i/X_i}$$

### Use of harmonic mean:

- Measure the change where the values of a variable are compared with a constant quantity of another variable like time, distance traveled within a given time, quantities purchased or sold over a unit.

### Merits of harmonic mean:

- It gives more weight to the small item and less weight to large values.
- It is based on all observations.
- It is not affected by sampling fluctuations.
- It is capable of further mathematical treatment.

### Demerits of harmonic mean:

- If any of the values is zero, it cannot be calculated.
- It is affected by extreme values.
- It cannot be calculated for open end class frequency distribution.
- It cannot be located graphically.
- It cannot be calculated for qualitative characteristics.
- It cannot be calculated if any observations are missing in the data series.

### Relation between A.M., G.M. and H.M.:

- For given two observations,  $A.M. \geq G.M. \geq H.M.$
- $G.M. = \sqrt{A.M. * H.M.}$
- $A.M. = \frac{G.M.^2}{H.M.}$
- $H.M. = \frac{G.M.^2}{A.M.}$

## 2.2. Median:

Median is the value situated in the middle position when all the observations are arranged in an ascending/descending order. The median is the central value of an ordered data series. It divides the data sets exactly into two parts. Fifty percent of observations are below the median



and 50% are above the median. Median is also known as 'positional average'. The Median is the 50<sup>th</sup> percentiles, 10<sup>th</sup> deciles, and 2<sup>nd</sup> quartiles. Median is also the intersect point of less than and more than ogive curve.

**Median for non-frequency data:**

**Step 1** Order the data from smallest to largest.

**Step 2** If the number of observations is odd, then  $(n + 1)/2$ <sup>th</sup> observation (in the ordered set) is the median. When the total number of observations is even, the median is given by the mean of  $n/2$ <sup>th</sup> and  $(n/2 + 1)$ <sup>th</sup> observation.

**Median for group frequency data:**

**Step 1** Obtain the cumulative frequencies for the data.

**Step 2** Mark the class corresponding to which a cumulative frequency is greater than  $N/2$ . That class is the median class.

**Step 3** Then median is evaluated by an interpolation formula

$$Median = l + \frac{h}{f} \left( \frac{N}{2} - C \right)$$

Where,  $l$  = lower limit of the median class

$N$  = Number of observations

$C$  = cumulative frequency of the class proceeding to the median class

$f$  = frequency of the median class

$h$  = magnitude of the median class

**Note:** Graphically, we can find the median by histogram.

**Use of median:**

- Qualitative data can be arranged in ascending or descending order of magnitude.
- Find average intelligence, honesty, etc.

**Merits of median:**

- It is rigidly defined.
- It is not affected by extreme values.
- It can be located graphically.
- It can be calculated for open end class frequency distribution.
- It can be calculated for data based on an ordinal scale.

**Demerits of median:**

## Basic Statistical Techniques

- It is not based on all observations.
- The calculation is more complex than the mean.
- It is not capable of further mathematical treatment.
- As compared to the mean, it is much affected by sampling fluctuations.

### 2.3. Mode:

Mode is defined as the value that occurs most frequently in the data. If in the data sets each observation occurs only once, then it does not have mode. When the data set has two or more values equal to the highest frequency than two or more mode are present in the datasets.

**Mode for ungroup frequency data:** The observation which has the highest frequency in the data sets.

**Mode for group (equal width) frequency data:**

**Step 1** Identify the modal class. Modal class is the class with the largest frequency.

**Step 2** Find mode by using interpolated formula.

$$mode = l + \frac{h(f_0 - f_{-1})}{(f_0 - f_{-1}) - (f_1 - f_0)}$$

Where,

$l$  = lower limit of the modal class

$f_0$  = frequency of the modal class

$f_{-1}$  = frequency of the preceding modal class

$f_1$  = frequency of the succeeding modal class

$h$  = magnitude of the modal class

**Note:** Graphically, we can find mode by histogram.

**Use of mode:**

- To find ideal consumer preferences for different kinds of products.
- The best measure for the average size of shoes or shirts.

**Merits of mode:**

- It is not affected by extreme values.
- It can be located graphically.
- It can be calculated for open end class frequency distribution.
- It can be calculated for data based on a nominal scale.

**Demerits of mode:**

- It is ill-defined.
- It is not based on all observations.
- The calculation is more complex than the mean.

## Basic Statistical Techniques

- It is not capable of further mathematical treatment.
- As compare to the mean, it is much affected by sampling fluctuations.

**Quartiles:** Quartiles are the three points that divide the whole data into four equal parts.

$$Q_i = l + \frac{h}{f} \left( \frac{iN}{4} - C \right)$$

**Deciles:** Deciles are the nine points that divide the whole data into ten equal parts.

$$D_i = l + \frac{h}{f} \left( \frac{iN}{10} - C \right)$$

**Percentiles:** Percentiles are the ninety-nine point that divides the whole data into hundreds of equal parts.

$$P_i = l + \frac{h}{f} \left( \frac{iN}{100} - C \right)$$

**Note: Median = 2nd Quartiles = 5th Deciles = 50th Percentiles**

**Empirical formula between mean median and mode:** If the data sets are asymmetric in nature, then

$$\mathbf{Mean - Mode = 3(Mean - Median)}$$

**The best measure of central tendency:**

According to prof. Yule, Mean is the best measure of central tendency. But there are some situations where the other measures of central tendency are preferred.

Scale	Use measure	Best measure
Interval	Mean, Median, Mode	Symmetrical data: Mean Asymmetrical data: Median
Ratio	Mean, Median, Mode	Symmetrical data: Mean Asymmetrical data: Median
Ordinal	Median, Mode	Median
Nominal	Mode	Mode

### 3. Measure of Dispersion

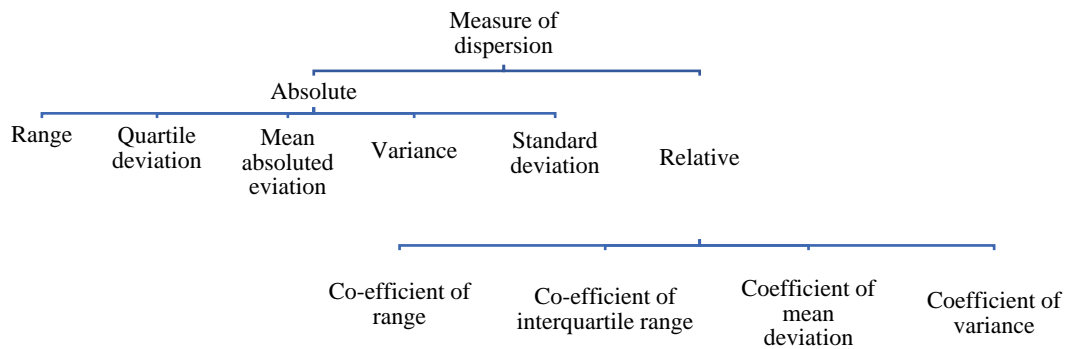
## Basic Statistical Techniques

The measure of central tendency such as mean, median, and mode only locate the center of the data. It does not infer anything about the spread of the data. Two data sets can have the same mean but they can be entirely different.

<b>Data 1</b>	38	42	41	44	45
<b>Data 2</b>	50	53	41	35	31

In the above example, two datasets have the same mean. So measures of central tendency are not adequate to describe data. Thus to describe data, one needs to know the measure of scatterness of observations. Dispersion is defined as deviation or scatterness of observations from their central values.

**Various measure of dispersion are:**



### 3.1.Range (R):

Range is the simplest measure of dispersion. It is defined as the difference between the highest value and lowest value of the variable. It is a crude measure of dispersion.

$$\text{Range} = \text{highest value } (H) - \text{lowest value } (L)$$

**Merits of range:**

- It is easy to understand and calculate.
- It is not affected by frequency of the data.

**Demerits of range:**

- It does not depend on all observations.
- It is very much affected by the extreme items.
- It cannot be calculated from open-end class intervals.
- It is not suitable for further mathematical treatment.
- It is the most unreliable measure of dispersion.

### 3.2. Quartile deviation (Q.D.):

Interquartile range is the difference between the first and third quartile. Hence the interquartile range describes the middle 50% of observations.

$$\text{Inter quartile range} = Q3 - Q1$$

Where,

$Q^3$  = first quartile of the data

$Q^1$  = third quartile of the data

Quartile deviation (Q.D.) is the half of the interquartile range.

$$\text{Quartile deviation (Q.D.)} = \frac{Q3 - Q1}{2}$$

#### Merits of Quartile deviation:

- It is easy to understand and calculate.
- It is not affected by extreme values
- It can be calculated for open end frequency data

#### Demerits of Quartile deviation:

- It does not depend on all observations.
- It is not suitable for further mathematical treatment.
- It is very much affected by sampling fluctuations.

### 3.3. Mean absolute deviation (MAD):

The absolute deviation of each value from the central value (mean is preferable) is calculated and the arithmetic mean of these deviations is called mean absolute deviation.

**For non-frequency data:** Let  $X_1, X_2, \dots, X_n$  are the n observations of a data set. The mean absolute deviation (MAD) about A is given by

$$MAD_A = \frac{\sum_{i=1}^n |X_i - A|}{n}$$

The mean absolute deviation (MAD) about mean is given by

$$MAD_{\bar{X}} = \frac{\sum_{i=1}^n |X_i - \bar{X}|}{n}$$

**For frequency data:** Let  $X_1, X_2, \dots, X_n$  are observations with corresponding frequencies are  $f_1, f_2, \dots, f_n$  and  $\sum_{i=1}^n f_i = N$ . The mean absolute deviation (MAD) about A is given by

$$MAD_A = \frac{\sum_{i=1}^n f_i |X_i - A|}{N}$$

The mean absolute deviation (MAD) about mean is given by



$$MAD_{\bar{X}} = \frac{\sum_{i=1}^n f_i |X_i - \bar{X}|}{N}$$

**Merits of mean absolute deviation about mean:**

- It is easy to understand and calculate.
- It is based on all observations.

**Demerits of mean absolute deviation about mean:**

- It is not suitable for further mathematical treatment.
- It does not take the sign of deviation under consideration.
- It is affected by extreme values.

**3.4. Standard deviation (S.D.):**

It is the best measure and the most commonly used measure of dispersion. It is defined as the positive square-root of the arithmetic mean of the square of the deviations of the given observation from their arithmetic mean. It takes into consideration the magnitude of all the observations and gives the minimum value of dispersion possible. It is also known as Root Mean Square Deviation about mean.

**For non-frequency data:** Let  $X_1, X_2, \dots, X_n$  are the n observation of a data set. The standard deviation A is given by

$$SD = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}}$$

**For frequency data:** Let  $X_1, X_2, \dots, X_n$  are observations with corresponding frequencies are  $f_1, f_2, \dots, f_n$  and  $\sum_{i=1}^n f_i = N$ . The standard deviation is given by

$$SD = \sqrt{\frac{\sum_{i=1}^n f_i (X_i - \bar{X})^2}{N}}$$

**Properties of standard deviation:**

- It is the independent of the change of origin but dependent on the change of scale  
Let  $U = a + hX$ , then  $sd(U) = |h| * sd(x)$
- If all observations are equal standard deviation is zero.
- It is never less than the quartile deviation and mean absolute deviation.

**Merits of standard deviation:**

- It is based on all observations.

- It is less affected by extreme values.
- It is suitable for further mathematical treatment.

**Demerits of standard deviation:**

- It is suitable for further mathematical treatment.
- It does not take the sign of deviation under consideration.
- It is affected by extreme values.
- It cannot be computed for open-end class data.

**3.5. Variance**

It is defined as the square of the standard deviation. Unit of the variance is the square of the actual observations, whereas unit of the standard deviation is same as actual observations.

**Relations between R, Q.D., M.D. and S.D.**

$$9QD = \frac{15}{2}MD = 6SD = R$$

**3.6. Coefficient of Variation (CV):**

The Coefficient of variation for a data set defined as the ratio of the standard deviation to the mean and expressed in percentage.

$$CV = \frac{SD}{mean} * 100\%$$

C.V is the relative measure of dispersion. It is the best measure among all the relative measure of dispersion. C.V is used to compare variability or consistency between two or more data series. If C.V. is greater indicate that the group is more variable, less stable, less uniform and less consistent. If the C.V. is less, it indicates that the group is less variable or more stable or more uniform and more consistent.

**Example:** Consider the data on score of Kohli and Smith in ODI cricket. The mean and standard deviation for Kohli are 55 and 5 respectively. The mean and standard deviation for Smith are 50 and 10 respectively. Find C.V. value for both the data and make compare them.

**Solution:**

For Kohli,  $CV = \frac{5}{55} * 100 = 9\%$

For Smith,  $CV = \frac{10}{50} * 100 = 20\%$

## Basic Statistical Techniques

The Smith is subject to more variation in score than Kohli. So Kohli is more consistent than Smith.

$$3.6. \text{ Coefficient of range} = \frac{H-L}{H+L} * 100\%$$

$$3.7. \text{ Coefficient of inter quartile range} = \frac{Q3-Q1}{Q3+Q1} * 100\%$$

$$3.8. \text{ Coefficient of mean deviation} = \frac{MAD}{\text{average from which it is calculated}} * 100\%$$

**Numerical Examples:** The marks of 10 students in statistics examination are as follows:

10,12,15,12,16, 20, 13,17,15,15

Find mean, median, mode, range and standard deviation.

**Solution:**

$X_i$	$f_i$	$f_i X_i$	$f_i (X_i - \bar{X})$	$(X_i - \bar{X})^2$	$f_i (X_i - \bar{X})^2$
10	1	10	-4.5	20.25	20.25
12	2	24	-5	6.25	12.5
13	1	13	-1.5	2.25	2.25
15	3	45	1.5	0.25	0.75
16	1	16	1.5	2.25	2.25
17	1	17	2.5	6.25	6.25
20	1	20	5.5	30.25	30.25
Total	10	145		67.75	74.5

$$\text{mean} = \frac{145}{10} = 14.5$$

$$\text{median} = 15$$

$$\text{mode} = 15$$

$$\text{range} = 20 - 10 = 10$$

$$SD = \frac{74.5}{10} = 7.45$$

### 4. Skewness and kurtosis:

We have discussed measures of central tendency and measure of dispersion which describe the location and scale parameter of the data sets. They do not give any idea about the shape of the data structure. The measure of skewness and kurtosis illustrate the shape of the data sets. The

measure of skewness gives the direction and the magnitude of the lack of symmetry and the measure of kurtosis gives the idea of the flatness of the curve.

#### 4.1. Skewness

Skewness measures the degree of asymmetry of the data. Skewness refers to the lack of symmetry.

Skewness is mainly three types: Positive skewness, Negative skewness, and Symmetric data.

##### Positive Skewness:

A data is said to be positive skew if the long tail is on the right side of the peak. The mean is on the right of the peak value. Here  $\text{Mean} > \text{Median} > \text{Mode}$ .

##### Negative Skewness:

A data is said to be negative skew if the long tail is on the left side of the peak. The mean is on the left of the peak value. Here  $\text{Mean} < \text{Median} < \text{Mode}$ .

##### Symmetric

The symmetrical distribution has zero skewness as all measures of a central tendency lies in the middle. When data is symmetrically distributed, the left-hand side, and right-hand side, contain the same number of observations. Here  $\text{Mean} = \text{Median} = \text{Mode}$ .

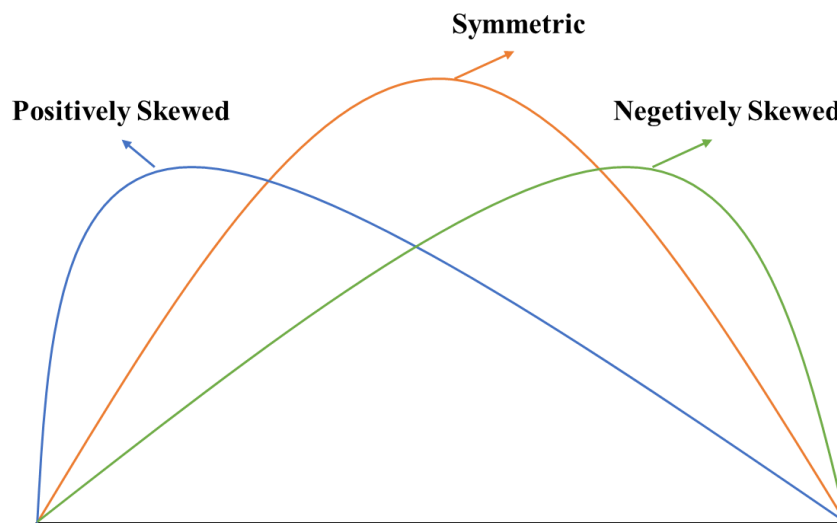


Figure 1. Skewness

##### The measure of Skewness:

$$\text{Pearson's first coefficient} = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}$$

$$\text{Pearson's second coefficient} = \frac{3 (\text{Mean} - \text{Median})}{\text{Standard Deviation}}$$

Interpretation:

1. If  $S_k = 0$ , then the frequency distribution is normal and symmetrical.
2. If  $S_k > 0$ , then the frequency distribution is positively skewed.
3. If  $S_k < 0$ , then the frequency distribution is negatively skewed.

#### 4.2. Kurtosis

Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. That is, data sets with high kurtosis tend to have heavy tails or outliers. Data sets with low kurtosis tend to have light tails or lack of outliers. A uniform distribution would be the extreme case.

**Types of kurtosis:** Leptokurtic or heavy-tailed distribution, Mesokurtic, Platykurtic or short-tailed distribution

##### Leptokurtic

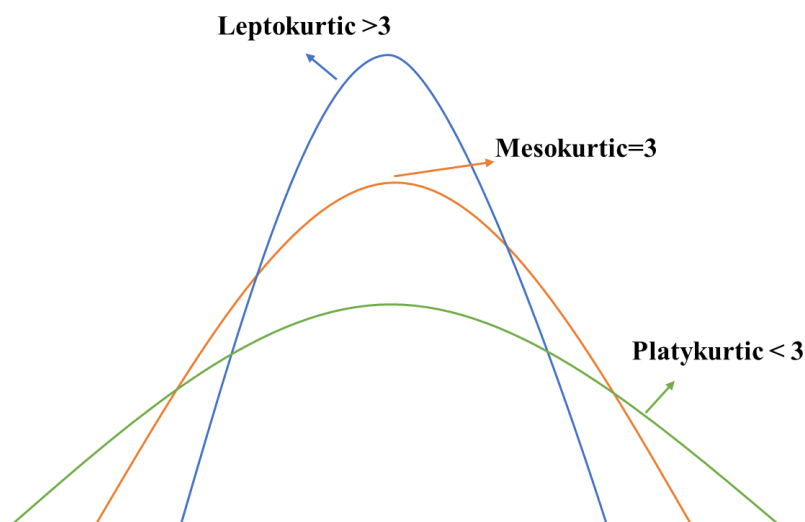
Leptokurtic indicates that distribution is peaked and possesses thick tails.

##### Platykurtic

Platykurtic having a lower tail and stretched around center tails means most of the data points are present in high proximity with mean. A platykurtic distribution is a flatter (less peaked) when compared with the normal distribution.

##### Mesokurtic

Mesokurtic is the same as the normal distribution. In Mesokurtic, distributions are moderate in breadth, and curves are a medium peaked height.



**Figure 2.** Kurtosis

$$\text{Measurement of Kurtosis } (\beta_2) = \frac{1}{N-1} \frac{\sum (y_i - \bar{y})^4}{s^4}, \quad \gamma_2 = \beta_2 - 3$$



## 5. Data presentation

There are three broad ways of presenting data. These are Textual presentation, Tabular presentation, and Graphic or diagrammatic presentation. We discussed only a few important diagrammatic presentations of data.

<b>Non dimensional diagram</b>	Pictograms
<b>Two dimensional diagram</b>	Bar diagrams, Pie diagrams, Histograms, Box Plot
<b>Three dimensional diagram</b>	Cubes, Cylinders diagrams

### 5.1.Bar Diagram

#### 5.1.1. Simple Bar Diagram

If the classification is based on attributes and if the attributes are to be compared with respect to a single character we use a simple bar diagram. Simple bar diagrams consist of vertical bars of equal width. The heights of these bars are proportional to the volume or magnitude of the attribute. All bars stand on the same baseline. The bars are separated from each other by equal intervals. The bars may be colored or marked.

#### 5.1.2. Multiple bar diagram

If the data is classified by attributes and if two or more characters or groups are to be compared within each attribute we use multiple bar diagrams. If only two characters are to be compared within each attribute, then the resultant bar diagram used is known as the double bar diagram. The multiple bar diagram is simply the extension of a simple bar diagram. For each attribute, two or more bars representing separate characters or groups are to be placed side by side. Each bar within an attribute will be marked or colored differently in order to distinguish them. The same type of marking or coloring should be done under each attribute. A footnote has to be given explaining the markings or colorings.

#### 5.1.3. Component bar diagram

This is also called a subdivided bar diagram. Instead of placing the bars for each component side by side, we may place this one on top of the other. This will result in a component bar diagram.

## 5.2. Histogram

Histograms is suitable for continuous class frequency distribution. We mark off class intervals along the x-axis and frequencies (frequency density for unequal frequency data) along the y-axis.

- For equal class intervals, the heights of the rectangles will be proportional to the frequencies, while for unequal class intervals, the heights will be equal (or proportional) to the frequency densities.
- A frequency polygon is a line graph obtained by connecting the midpoints of the tops of the rectangles in the histogram.

**Table 1.** Differences between bar diagrams and histograms

<b>Characteristics</b>	<b>Bar Diagrams</b>	<b>Histograms</b>
Frequency is measured by	Height of the bar	Area of the bar
Gaps between the bars	Yes	No
Width of the bar	Equal	May not be equal
Data types	Discrete and Continuous	Continuous only

## 5.3. Pie diagrams

When we are interested in the relative importance of the different components of a single factor, we use pie diagrams. For the pie diagram, one circle is used and the area enclosed by it being taken as 100. It is then divided into a number of sectors by drawing angles at the center, the area of each sector representing the corresponding percentage.

## 5.4. Box Plot

Minimum, maximum, and quartiles ( $Q_1$ , Median,  $Q_3$ ) together provide information on the center and variation of the variable in a nice compact way. Written in increasing order, they comprise what is called the five-number summary of the variable. A boxplot is based on the five-number summary and can be used to provide a graphical display of the center and variation of the observed values of the variable in a data set. It can tell you about your outliers and what their values are. It can also tell you if your data is symmetrical, how tightly your data is grouped, and if and how your data is skewed.

***N.B: Examples of graphical presentation have been given in our basic statistics with excel manual.***

## **6. Robust Estimate of Mean and Standard Deviation**

The mean and standard deviation provides a correct estimation only if the variable is normally distributed and without outliers. If the variable is skewed and/or has outliers, the mean and standard deviation will be excessively influenced by the extreme observations and provide faulty statistics of data. There are many alternatives to the mean and standard deviation. Alternatives to the mean include the well-known median and trimmed mean, Winsorized mean, and M-estimators and for standard deviation, the alternatives include the Inter-Quartile Range (IQR) and the Median Absolute Deviation (MAD), Trimmed standard deviation, the Winsorized standard deviation, and M-estimators. Median, IQR, MAD are already discussed in the previous section in detail. Here we only discussed the trimmed, Winsorized, and M estimators for mean and standard deviation.

### **6.1. Trimmed Mean and Standard Deviation**

A trimmed mean and standard deviation is similar to a “regular” mean but it trims any outliers from both the side. To obtain the 20% trimmed mean, the 20% lowest and 20 % highest values are removed and the mean is computed on the remaining observations. In our example, these values will be: 4, 4, 5, 5, 6, 6, and the 20% trimmed mean will be equal to 5.

### **6.2. Winsorized Mean and Standard Deviation**

The Winsorized technique is similar to the trimmed technique but the lowest (resp. highest) values are not removed but replaced by the lowest (resp. highest) untrimmed score. In our example, the values of the variables, also called Winsorized scores, will then be: 4, 4, 4, 4, 5, 5, 6, 6, 6, 6, and the 20% Winsorized mean will be equal to 5.

### **6.3. M estimators**

The trimmed mean all either take or drop observations. As for the Winsorized mean, it replaces values with less extreme values. In contrast, the M-estimators, weight each observation according to a function selected for its special properties. The weights depend on a constant that can be chosen by the researcher. The M-estimator solves this problem of assigning a zero value to many observations by downweighting the observations progressively. The only aspect of the M-estimator that could worry substantive researchers is that one must choose the degree of downweighting of the observations.

---

---

# AN INTRODUCTION TO DATA ANALYSIS USING MS-EXCEL

---

---

**Upendra Kumar Pradhan**

*ICAR-Indian Agricultural Statistics Research Institute*

*Library Avenue, New Delhi - 110 012*

[Upendra.Pradhan@icar.gov.in](mailto:Upendra.Pradhan@icar.gov.in)

---

---

## 1. Introduction

Microsoft Excel is a spreadsheet program. The term Excel means superior than any other to organizing data. Microsoft Excel is an excellent program for organizing, formatting and calculating any type of data. Excel displays data in a row-and-column format, with gridlines between the rows and columns, similar to accounting ledger books or graph paper. There are different version of excel are available, mainly **3.0, 3.1, 5.0, 95, 97, 2000, XP, 2003, 2007, 2010, 2013, 2016 and 2019**. An Excel workbook file can have multiple plies or worksheets. One can navigate through these separate worksheets by clicking on the tabs at the bottom of the workbook. Spreadsheets consist of rows and columns. There are a total of 1,048,576 rows by 16,384 columns present in a spreadsheet. The intersection of a row and a column is called a cell. By convention we refer to the column number first and the row number second. A cell in a worksheet can contain any combination of the following kinds of information:

- ✓ numbers
- ✓ text
- ✓ formulas
- ✓ built-in functions
- ✓ pointers to other cells

If you are entering text or numbers, you can just type those directly into the cell or the Formula Bar. If you are entering a formula, function, or pointer, you signal to Excel that you are not entering text or a number by preceding your entry with an equal sign (=). You can format the contents of cells by changing fonts, colours, borders, and other features. Right click on the cell or range to see the context-sensitive menu and then select **Format Cells**. It is also possible to embellish a worksheet by adding elements that are not attached to any specific cell. These include charts and graphs, clip art or other graphics, text boxes, drawings, word art, equations, and many other kinds of objects.

You can enter information into a cell of a worksheet in two different ways. You have to just type information directly into the cells of the worksheet or select the cell and then type into the formula bar. As an alternative, you can also use Excel's default data form. Assume that you have the following information and want to enter it into your worksheet.

**Table-1**

<b>Name</b>	<b>Variable 1</b>	<b>Variable2</b>	<b>variable3</b>
Lyle	88	92	89
Jorge	78	62	78
Bill	34	87	62
Bob	76	82	32
Jim	66	88	64

## 2. Data Structures and Descriptive Statistics

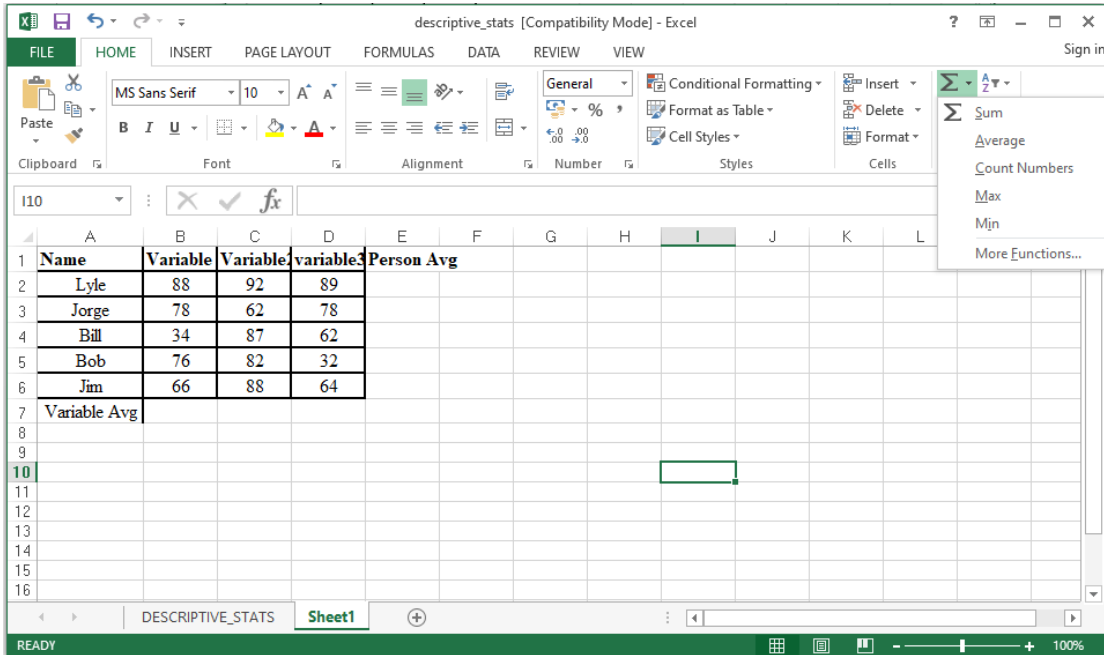
Excel provides many built-in functions and tools for data management and statistical analysis. Let us begin our exploration with a very useful tool called AutoSum. The

## MS-Excel

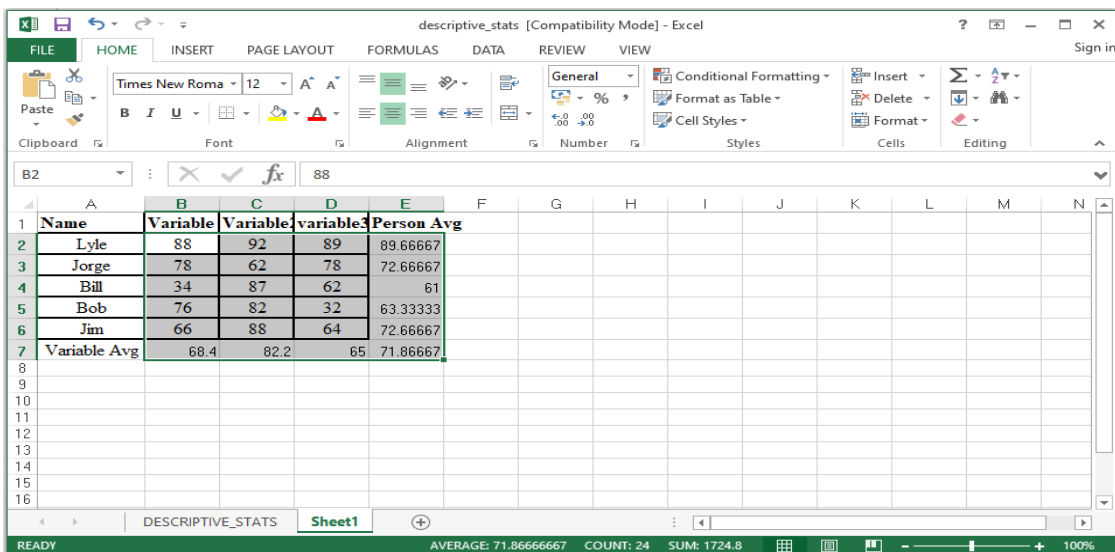
AutoSum Tool does much more than just add or average numbers. It provides the following summary statistics for a selected range of data.

- Sum
- Average
- Count
- Minimum
- Maximum

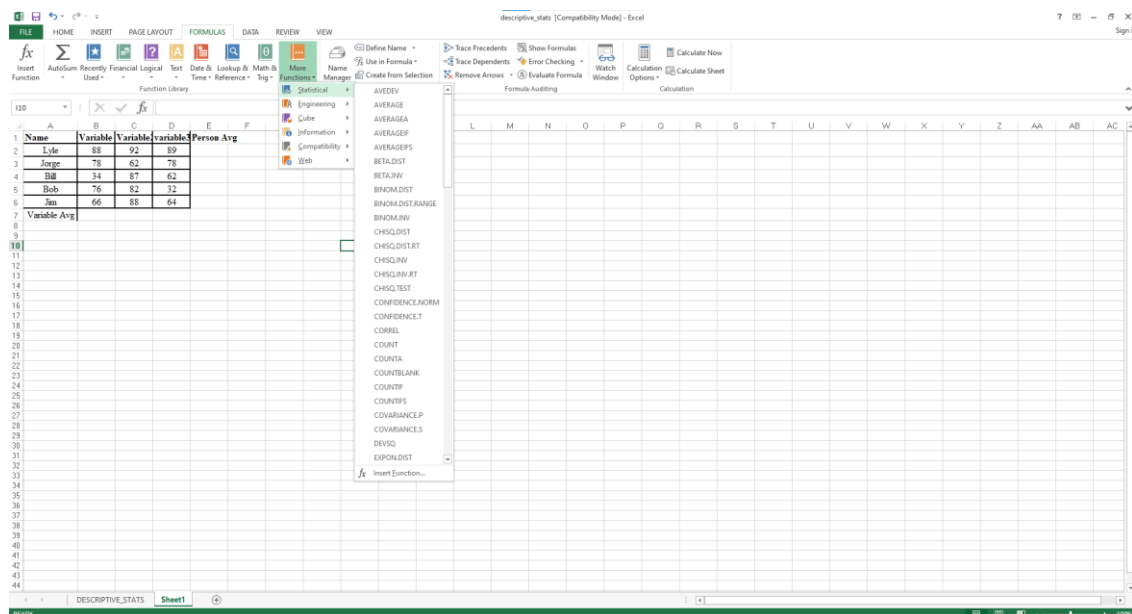
Return to the data from Table 1. Add a column for the individual averages and a row for the variable averages. Select the entire range of data and the empty row and column adjacent to the data as appeared below



Select the dropdown arrow beside the AutoSum icon (the Greek sigma) in the standard Toolbar. Select Average and all the averages will be calculated at the same time.



Many additional statistical functions are available via the Insert Function command. Select **Insert, Formulas**, click on *fx* in the formula bar, or select **More Functions** from the AutoSum menu. The Insert Function menu appears below.



There are many Excel add-ins and macros available that provide increased statistical functionality. We will concern ourselves with only one of these, the Analysis ToolPak provided by Microsoft with the Excel program. The Analysis ToolPak comes with Excel, but is not installed by default. To install the Analysis ToolPak, click the **Microsoft Office Button**, and then click **Excel Options**. Click **Add-ins**, and then in the **Manage box**, select Excel Add-ins. Click **Go**. In the Add-Ins available box, select the **Analysis ToolPak** checkbox, and then click **OK**. After you have installed the Analysis ToolPak, there will be a Data Analysis option available in the Data menu. Let us consider an example to perform the analysis using Analysis ToolPak.

**Example-1 :** An experiment was conducted to study the hybrid seed production of bottle gourd (*Lagenaria siceraria* (Mol) Standl) Cv. Pusa hybrid-3 under open field conditions during Kharif-2005 at Indian Agricultural Research Institute, New Delhi. The main aim of the investigation was to compare natural pollination {the male flowers pinched from the seed parent before the anthesis regularly with utmost care to avoid the chance selfing. The pollination is carried out by the natural pollinating agent} and hand pollination {the male flowers also pinched from the seed parent before the anthesis regularly. The female buds are covered with butter paper bag which contain 5-6 tiny hole to felicitate the ventilation and to avoid the built up of high temperature in size the butter paper bag. The butter paper bag is clipped/stippled. On the same day the male bud at pollen parent (male plant) are also covered with butter paper bag. On the next day the male bud are removed and the anthers are rubbed gently over all the three lobes. The female flower is again covered with butter paper bag and label is placed over the peduncle of pollinated female flower (plate 4, 5, 6 &7). The pollination is performed at noon (1-3pm)} under field conditions. The data were collected on 10 randomly selected plants from each of natural pollination and hand pollination on number of fruit set for the period of 45 days, fruit weight (kg), seed yield per plant (g) and seedling length (cm). The data obtained is as given below:

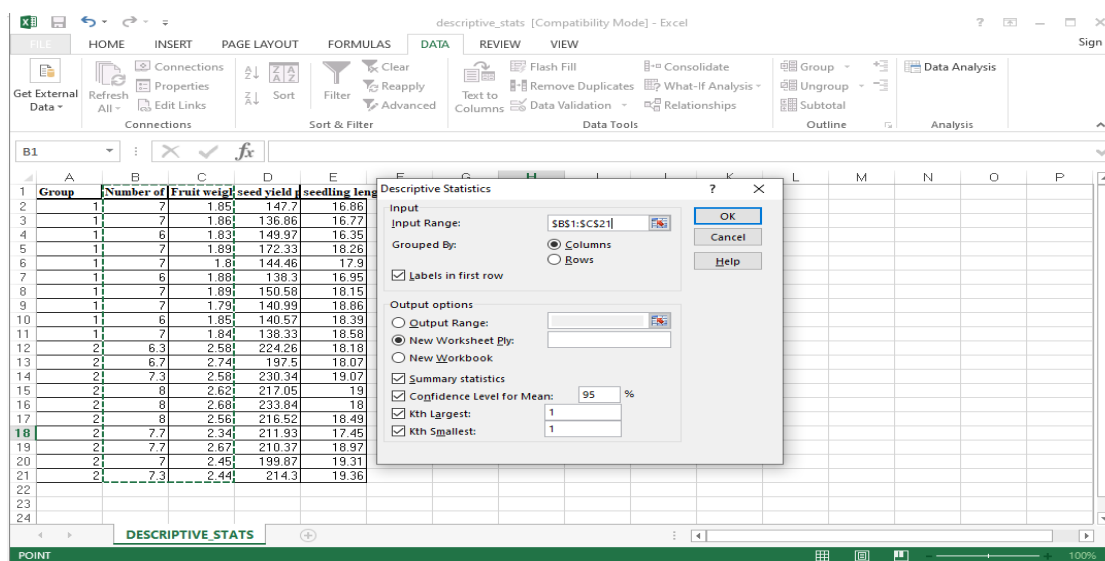
Group	Number of Fruit set	Fruit weight	seed yield per plant	seedling length
1	7	1.85	147.7	16.86
1	7	1.86	136.86	16.77
1	6	1.83	149.97	16.35
1	7	1.89	172.33	18.26
1	7	1.8	144.46	17.9
1	6	1.88	138.3	16.95
1	7	1.89	150.58	18.15
1	7	1.79	140.99	18.86
1	6	1.85	140.57	18.39
1	7	1.84	138.33	18.58
2	6.3	2.58	224.26	18.18
2	6.7	2.74	197.5	18.07
2	7.3	2.58	230.34	19.07
2	8	2.62	217.05	19
2	8	2.68	233.84	18
2	8	2.56	216.52	18.49
2	7.7	2.34	211.93	17.45
2	7.7	2.67	210.37	18.97
2	7	2.45	199.87	19.31
2	7.3	2.44	214.3	19.36

{Here 1 denotes natural pollination and 2 denotes the hand pollination}

**Descriptive Statistics in the Analysis ToolPak**

The “Descriptive Statistics” analysis tool generates a report of univariate Statistics for data in the input range, which includes information about the central tendency and variability of the entered data. From example-1 let us calculate Descriptive statistics for the characters such as Number of fruit set and Fruit weight of all.

The dialog box for the Descriptive Statistics tool is shown below. To access this tool, select **Data, Data Analysis, and Descriptive Statistics**. Select the desired range, or type in the name of the selected range.



The summary output from the descriptive statistics tool appears as.

	Number of Fruit set	Fruit weight
Mean	7.05	2.207
Standard Error	0.14226	0.084701022
Median	7	2.115
Mode	7	1.85
Standard Deviation	0.63619	0.378794487
Sample Variance	0.40474	0.143485263
Kurtosis	-0.5465	-1.99031995
Skewness	-0.1785	0.139517118
Range	2	0.95
Minimum	6	1.79
Maximum	8	2.74
Sum	141	44.14
Count	20	20
Confidence Level(95.0%)	0.29775	0.177281277

### 3. Tests of Significance Based on T – Distribution

#### 3.1. One-Sample *t* Test

The one-sample *t* test compares a given sample mean  $\bar{X}$  to a known or hypothesized value of the population mean  $\mu_0$  when the population standard deviation  $\sigma$  is unknown.. Excel does not have a built-in one-sample *t* test. However, the use of Excel functions and formulas makes the computations quite simple. The value of *t* can be calculated from the simple formula:

$$t = \frac{\bar{X} - \mu_0}{s_{\bar{X}}}$$

Where  $\bar{X}$  is the sample mean,  $\mu_0$  is the known or hypothesized population mean and  $s_{\bar{X}}$  is the standard error of mean which is

$$s_{\bar{X}} = \sqrt{\frac{s_{\bar{X}}^2}{n}} = \frac{s_{\bar{X}}}{\sqrt{n}}$$

Where n denote the sample size. From example-1 to test whether the mean of the population of Seed yield/plant (g) is 200 or not one can use the following steps in MS-EXCEL:

First compute sample mean and sample variance and then compute the test statistic. Sample mean and sample standard deviation can be obtained using functions “AVERAGE” and “STDEV” of MS-EXCEL as follows



## MS-Excel

Group	Number of Fruit	Fruit weight(g)	seed yield per plant(Kg)	seedling length(Cm)
1	7	1.85	147.7	16.86
1	7	1.86	136.86	16.77
1	6	1.83	149.97	16.35
1	7	1.89	172.33	18.26
1	7	1.8	144.46	17.9
1	6	1.88	138.3	16.95
1	7	1.89	150.58	18.15
1	7	1.79	140.99	18.86
1	6	1.85	140.57	18.39
1	7	1.84	138.33	18.58
2	6.3	2.58	224.26	18.18
2	6.7	2.74	197.5	18.07
2	7.3	2.58	230.34	19.07
2	8	2.62	217.05	19
2	8	2.68	233.84	18
2	8	2.56	216.52	18.49
2	7.7	2.34	211.93	17.45
2	7.7	2.67	210.37	18.97
2	7	2.45	199.87	19.31
2	7.3	2.44	214.3	19.36
			<b>=AVERAGE(D2:D21)</b>	

And the sample standard deviation comes from the function “STDEV”.

Group	Number of Fruit	Fruit weight(g)	seed yield per plant(Kg)	seedling length(Cm)
1	7	1.85	147.7	16.86
1	7	1.86	136.86	16.77
1	6	1.83	149.97	16.35
1	7	1.89	172.33	18.26
1	7	1.8	144.46	17.9
1	6	1.88	138.3	16.95
1	7	1.89	150.58	18.15
1	7	1.79	140.99	18.86
1	6	1.85	140.57	18.39
1	7	1.84	138.33	18.58
2	6.3	2.58	224.26	18.18
2	6.7	2.74	197.5	18.07
2	7.3	2.58	230.34	19.07
2	8	2.62	217.05	19
2	8	2.68	233.84	18
2	8	2.56	216.52	18.49
2	7.7	2.34	211.93	17.45
2	7.7	2.67	210.37	18.97
2	7	2.45	199.87	19.31
2	7.3	2.44	214.3	19.36
			<b>180.8035</b>	
			<b>=STDEV(D2:D21)</b>	

Null hypothesis for this example is  $H_0: \text{mean}=200$ , therefore the test value is 200.

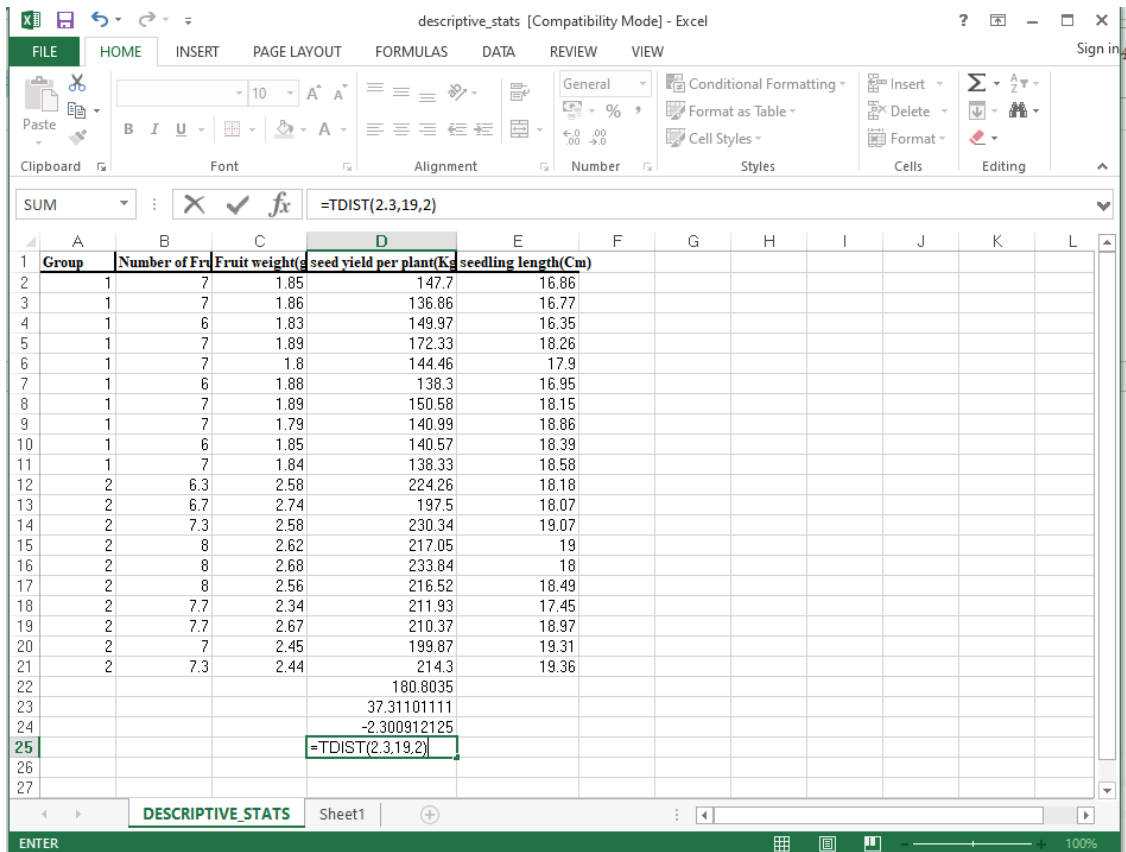
We now compute the value of test statistics as here  $n=20$ , sample mean =180.80, test value=200 and sample standard deviation = 37.31.

Group	Number of Fruit	Fruit weight(g)	seed yield per plant(Kg)	seedling length(Cm)
1	7	1.85	147.7	16.86
1	7	1.86	136.86	16.77
1	6	1.83	149.97	16.35
1	7	1.89	172.33	18.26
1	7	1.8	144.46	17.9
1	6	1.88	138.3	16.95
1	7	1.89	150.58	18.15
1	7	1.79	140.99	18.86
1	6	1.85	140.57	18.39
1	7	1.84	138.33	18.58
2	6.3	2.58	224.26	18.18
2	6.7	2.74	197.5	18.07
2	7.3	2.58	230.34	19.07
2	8	2.62	217.05	19
2	8	2.68	233.84	18
2	8	2.56	216.52	18.49
2	7.7	2.34	211.93	17.45
2	7.7	2.67	210.37	18.97
2	7	2.45	199.87	19.31
2	7.3	2.44	214.3	19.36
			180.8035	
			37.31101111	
			=(D22-200)/(D23/SQRT(20))	

Computed t-value= -2.30; Modulus value of the computed t-value is 2.30.

To find the p-value, use “TDIST” function by giving = TDIST(X, degrees of freedom, tail).

1. X is the modulus value of the computed t-value i.e., 2.30
2. Type in the  $df = n - 1=20-1=19$
3. If tails = 1, TDIST returns the one-tailed distribution. If tails = 2, TDIST returns the two-tailed distribution. In our case it is the two-tailed distribution i.e., 2.



Therefore the p-value= 0.03

### 3.2. Independent-Samples *t* Test

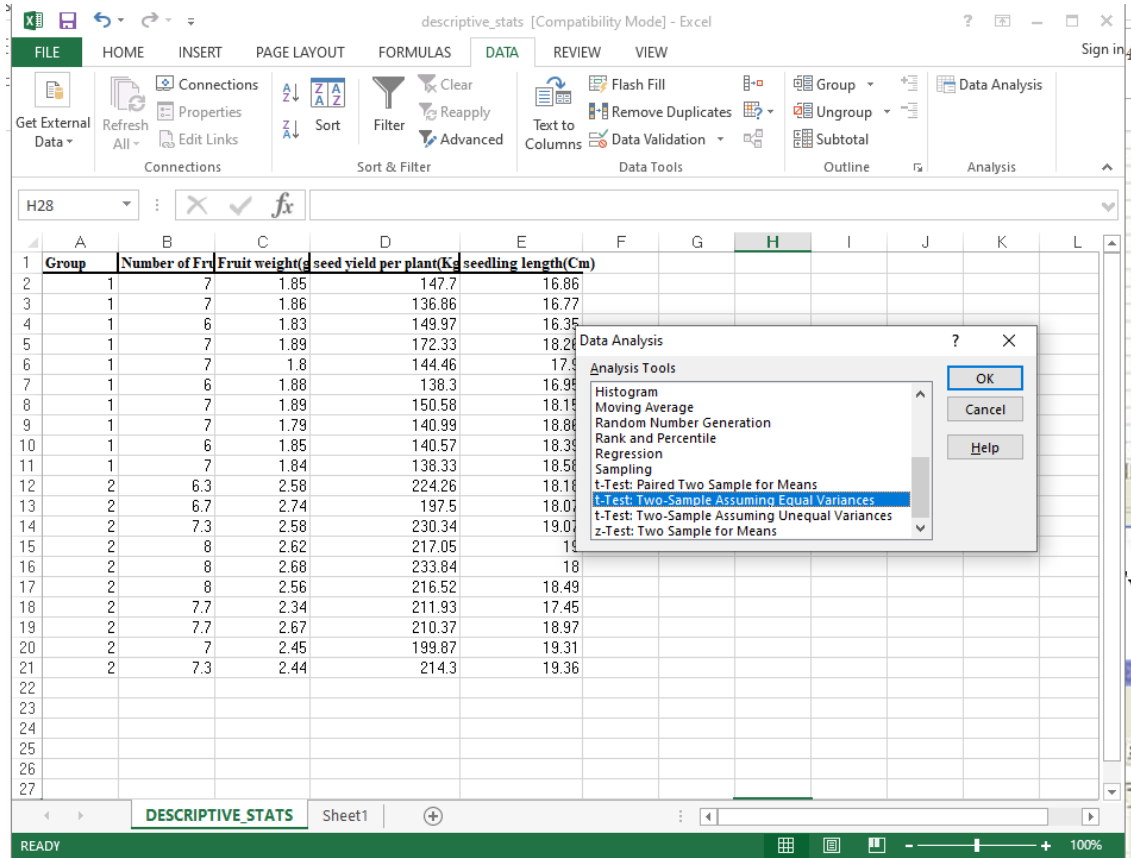
The independent-samples *t* test compares the means from two separate samples. It is not required that the two samples have the same number of observations. The Analysis ToolPak performs both one and two-tailed independent-samples *t* tests.

From example-1 one want to test whether the natural pollination and hand pollination under open field conditions are equally effective or are significantly different. To answer the question we follow the following steps:

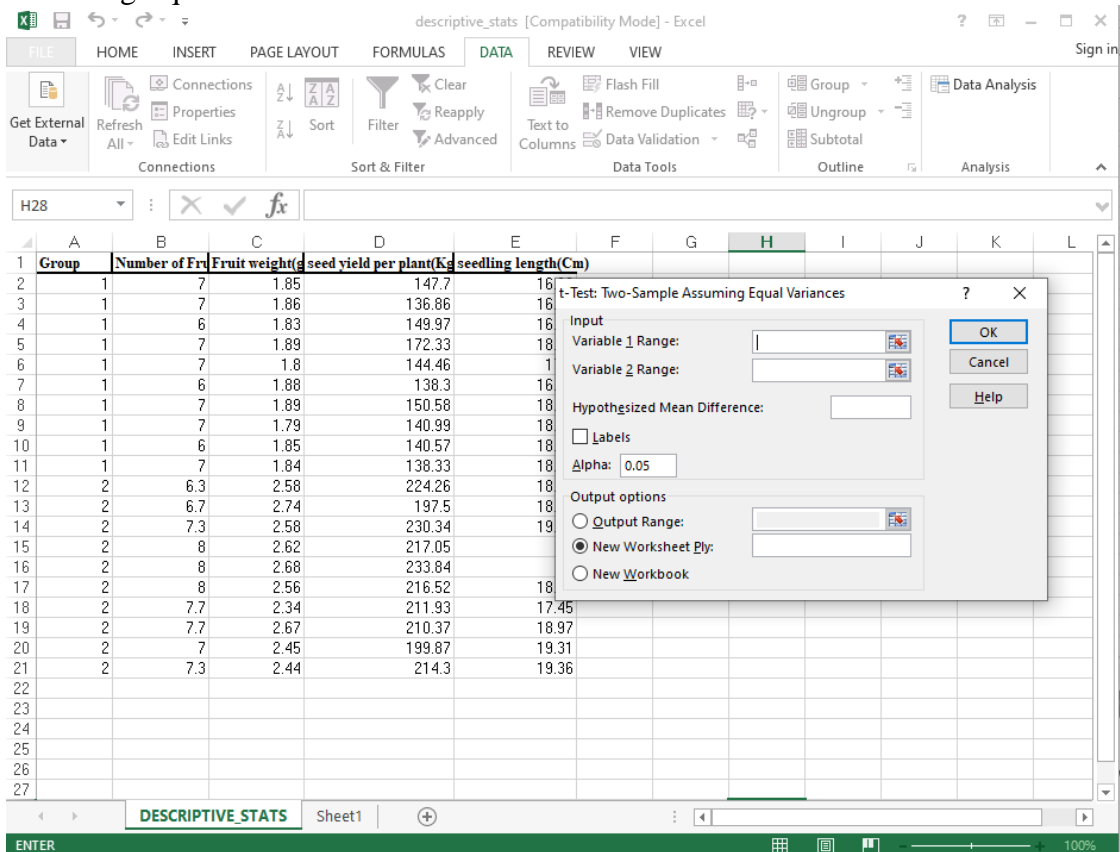
Once the data entry is complete, Choose Tools from the Menu Bar. Now select **Data** → **Data Analysis...**

In the Data Analysis dialog box select t-Test: Two-Sample Assuming Equal Variance. This selection displays the following screen.

# MS-Excel



Click **OK**. This displays the dialog box for the analysis of t-Test: Two-Sample Assuming Equal Variance



## MS-Excel

For the two groups select the variable number of fruit Set (45days) and select the range for Variable 1 Range: and Variable 2 Range: in the Input box. Now select Output Range: to get the output. This displays the following screen.

The screenshot shows the 't-Test: Two-Sample Assuming Equal Variances' dialog box in Microsoft Excel. The dialog box is open over a spreadsheet with the following data:

Group	Number of Fruit	Fruit weight (kg)	seed yield per plant (kg)	seedling length (Cm)
1	7	1.85	147.7	16.86
1	7	1.86	136.86	16.77
1	6	1.83	149.97	16.35
1	7	1.89	172.33	18.26
1	7	1.8	144.46	17.9
1	6	1.88	138.3	16.95
1	7	1.89	150.58	18.15
1	7	1.79	140.99	18.86
1	6	1.85	140.57	18.39
1	7	1.84	138.33	18.58
2	6.3	2.58	224.26	18.18
2	6.7	2.74	197.5	18.07
2	7.3	2.58	230.34	19.07
2	8	2.62	217.05	19
2	8	2.68	233.84	18
2	8	2.56	216.52	18.49
2	7.7	2.34	211.93	17.45
2	7.7	2.67	210.37	18.97
2	7	2.45	199.87	19.31
2	7.3	2.44	214.3	19.36

The dialog box settings are:

- Variable 1 Range: \$B\$2:\$B\$11
- Variable 2 Range: \$B\$12:\$B\$21
- Hypothesized Mean Difference: (empty)
- Labels:
- Alpha: 0.05
- Output options:
  - Output Range: \$G\$2
  - New Worksheet Ply:
  - New Workbook

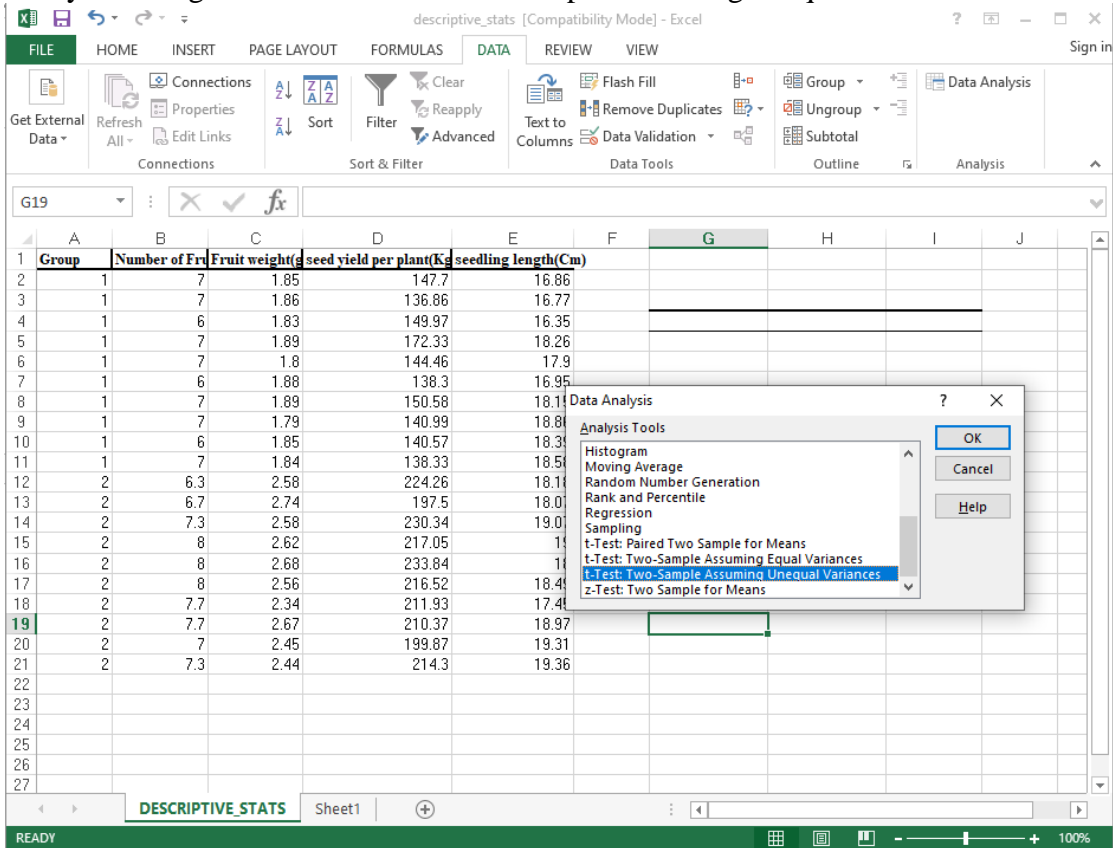
Click OK to get the output at the selected output range.

The screenshot shows the output of the t-Test: Two-Sample Assuming Equal Variances in Microsoft Excel. The output is displayed in the spreadsheet starting from cell G2. The output includes:

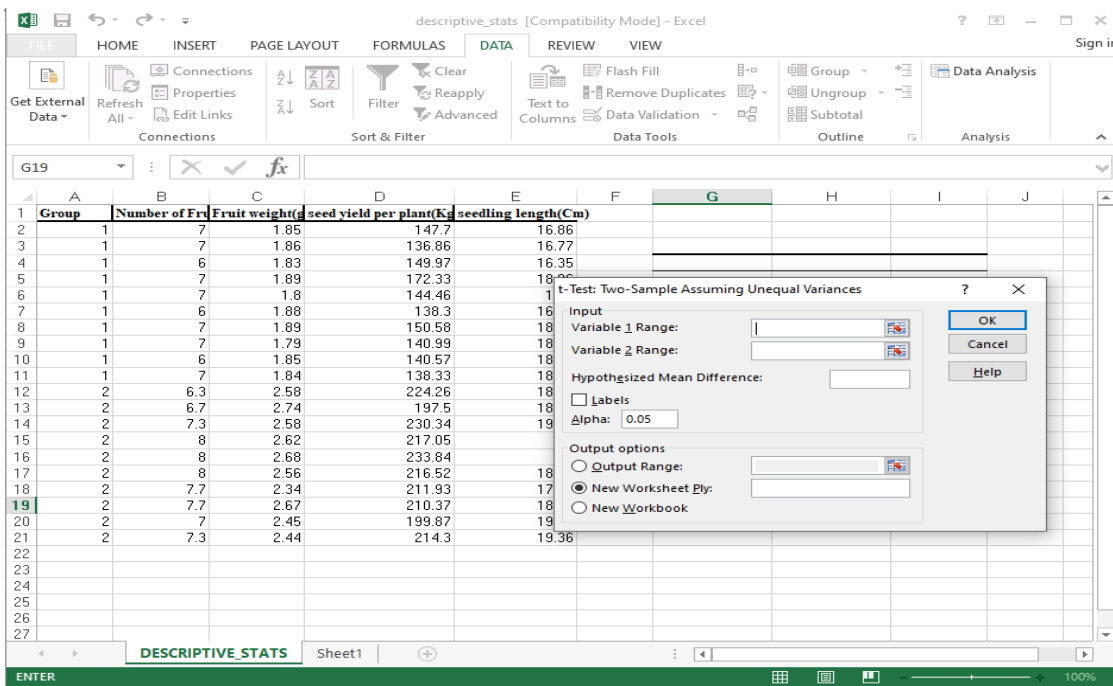
	Variable 1	Variable 2
Mean	6.7	7.4
Variance	0.233333333	0.348888889
Observations	10	10
Pooled Variance	0.291111111	
Hypothesized M	0	
df	18	
t Stat	-2.901039561	
P(T<=t) one-tail	0.004761993	
t Critical one-tail	1.734063607	
P(T<=t) two-tail	0.009523987	
t Critical two-tail	2.10092204	

Similarly one can perform the analysis for the other variables also.

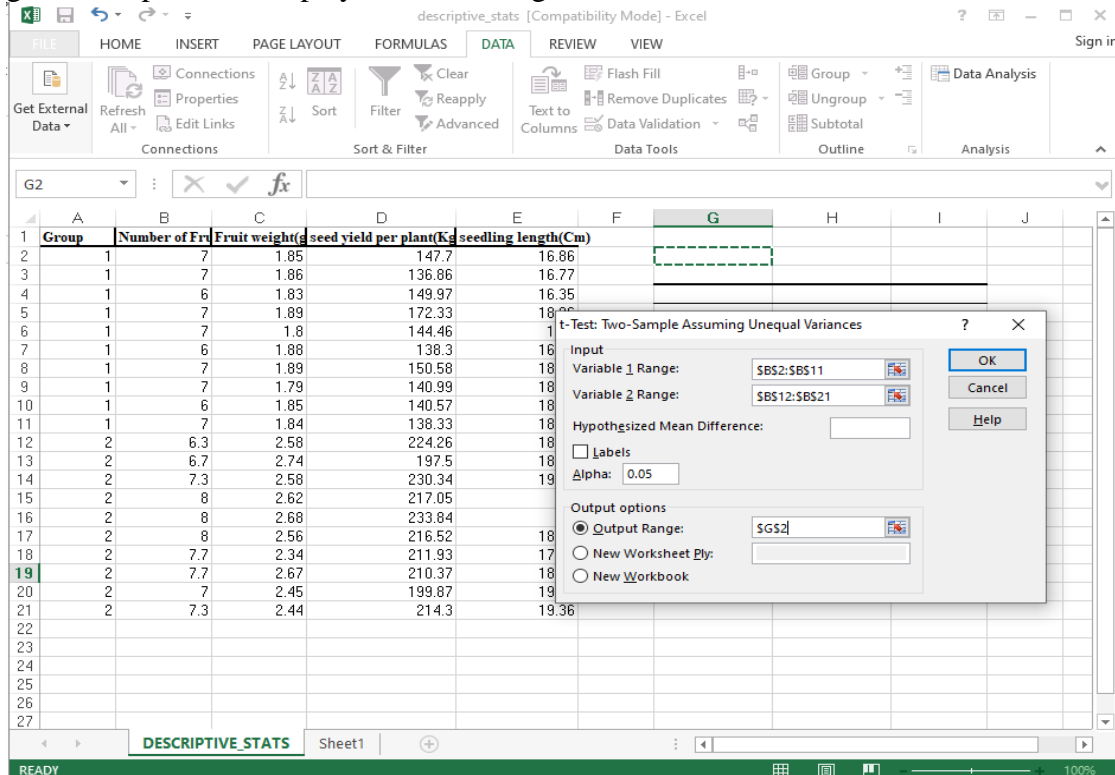
For the analysis of t-Test: Two-Sample Assuming Unequal Variances, in the Data Analysis dialog box select t-Test: Two-Sample Assuming unequal Variance.



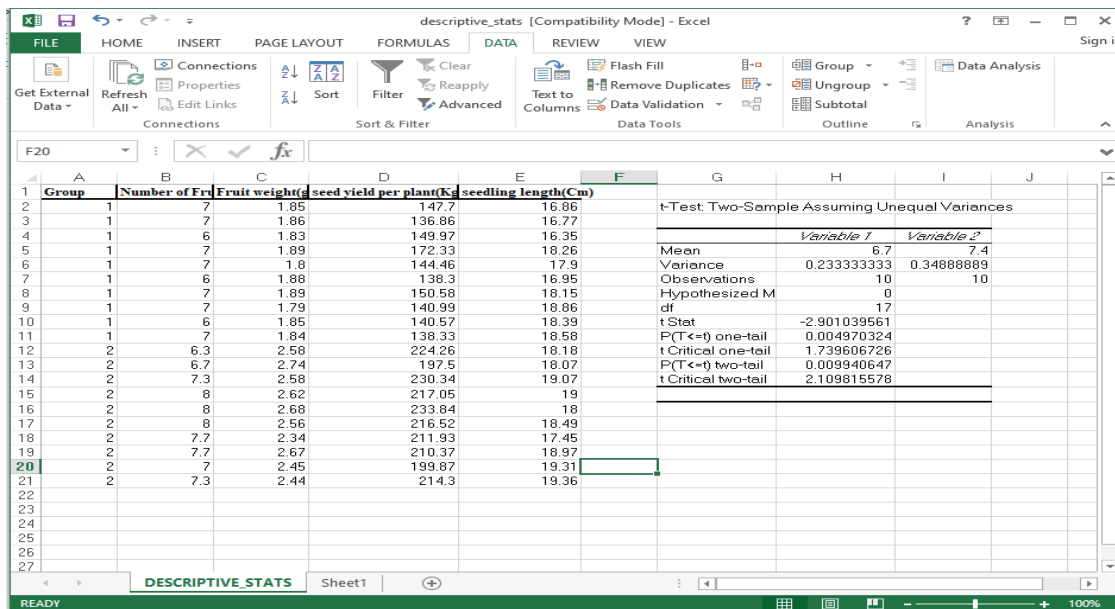
This selection displays the following screen.



For the two groups select the variable no. of fruit Set (45days) and select the range for Variable 1 Range: and Variable 2 Range: in the Input box. Now select Output Range: to get the output. This displays the following screen



Click OK to get the output at the selected output range

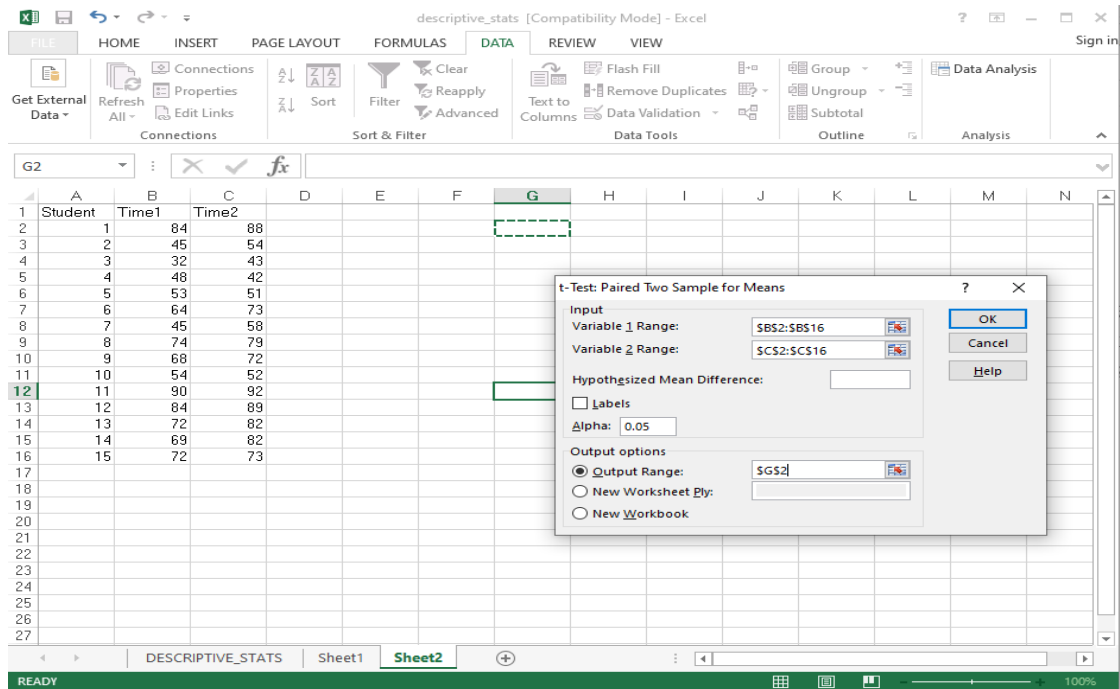


Similarly one can perform the analysis for the other variables also.

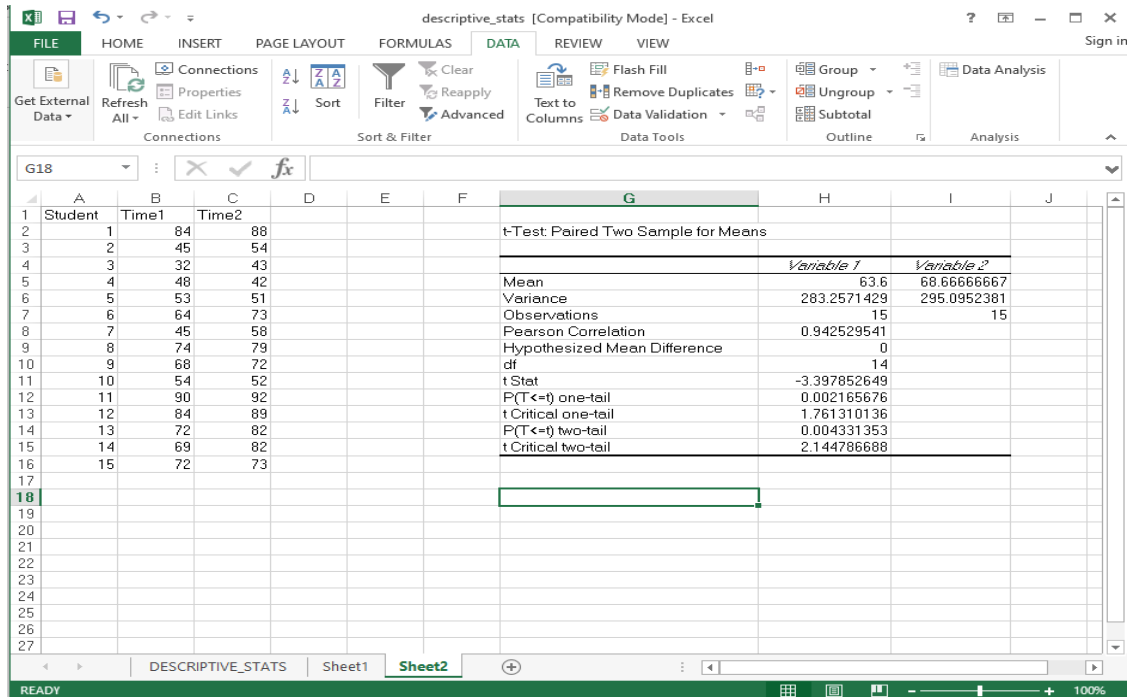
### 3.3. Paired-Samples *t* Test

The dependent or paired-samples *t* test is applied to data in which the observations from one sample are paired with or linked to the observations in the second sample.

To conduct the analysis, select **Data, Data Analysis, t-Test: Paired Two Sample for Means**. The dialog box is shown in below image. Enter the Variable1 and Variable2 ranges as shown. You may place the output in a new worksheet ply or in the same worksheet as the data.



The calculated  $t$  statistic is  $-3.398$  indicate that the Time2 ATS score is significantly higher than the ATS score for Time1. The statistics course led to significantly more positive attitudes toward statistics,  $t(14) = -3.398, p = .002$ , one-tailed.



#### 4. Correlation and Regression



To find the Pearson correlation between two variables, you may use the built-in function **CORREL** (Array1, Array2). You can also derive an inter correlation matrix for two or more variables by using the Correlation tool in the Analysis ToolPak. However, the Regression tool in the Analysis Tool Pak supplies the value of the correlation coefficient and also conducts an analysis of variance of the significance of the regression. Excel's Regression tool performs both simple and multiple regression analyses. To perform multiple regression analysis, supply the ranges of two or more predictor variables in the dialog box for the Regression tool.

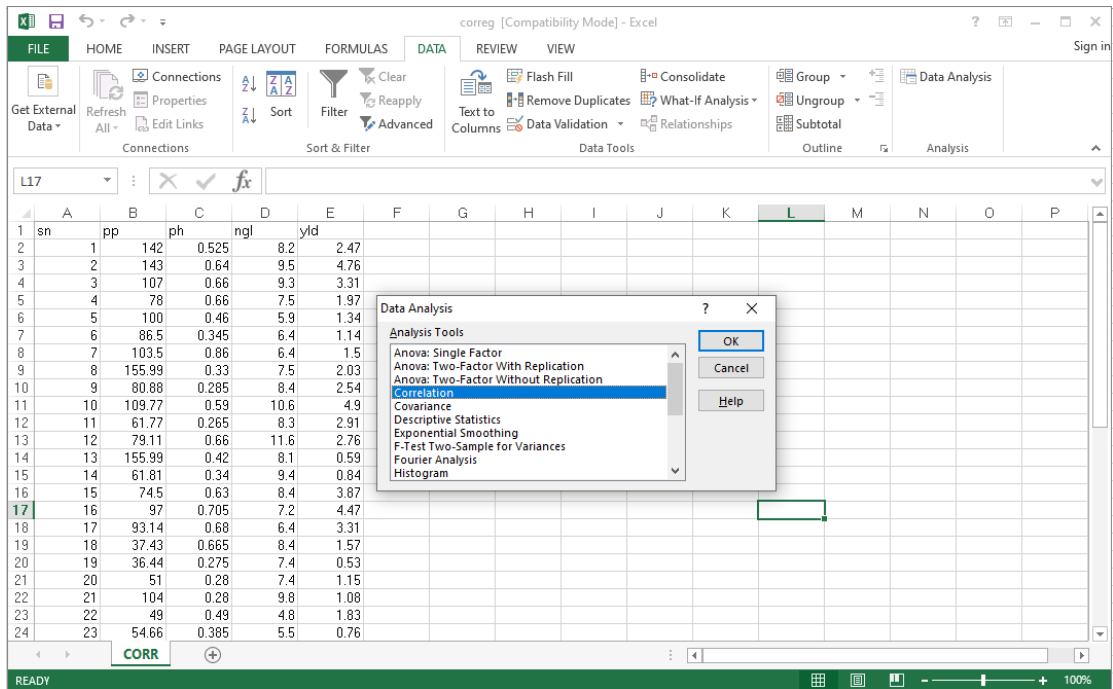
**Example-2:** The following data was collected through a pilot sample survey on Hybrid Jowar crop on yield and biometrical characters. The biometrical characters were average Plant Population (PP), average Plant Height (PH), average Number of Green Leaves (NGL) and Yield (kg/plot). We will use the Regression tool in the Analysis ToolPak.

sn	pp	ph	ngl	yld	sn	pp	ph	ngl	yld
1	142	0.525	8.2	2.47	24	55.55	0.265	5	0.43
2	143	0.64	9.5	4.76	25	88.44	0.98	5	4.08
3	107	0.66	9.3	3.31	26	99.55	0.645	9.6	2.83
4	78	0.66	7.5	1.97	27	63.99	0.635	5.6	2.57
5	100	0.46	5.9	1.34	28	101.77	0.29	8.2	7.42
6	86.5	0.345	6.4	1.14	29	138.66	0.72	9.9	2.62
7	103.5	0.86	6.4	1.5	30	90.22	0.63	8.4	2
8	155.99	0.33	7.5	2.03	31	76.92	1.25	7.3	1.99
9	80.88	0.285	8.4	2.54	32	126.22	0.58	6.9	1.36
10	109.77	0.59	10.6	4.9	33	80.36	0.605	6.8	0.68
11	61.77	0.265	8.3	2.91	34	150.23	1.19	8.8	5.36
12	79.11	0.66	11.6	2.76	35	56.5	0.355	9.7	2.12
13	155.99	0.42	8.1	0.59	36	136	0.59	10.2	4.16
14	61.81	0.34	9.4	0.84	37	144.5	0.61	9.8	3.12
15	74.5	0.63	8.4	3.87	38	157.33	0.605	8.8	2.07
16	97	0.705	7.2	4.47	39	91.99	0.38	7.7	1.17
17	93.14	0.68	6.4	3.31	40	121.5	0.55	7.7	3.62
18	37.43	0.665	8.4	1.57	41	64.5	0.32	5.7	0.67
19	36.44	0.275	7.4	0.53	42	116	0.455	6.8	3.05
20	51	0.28	7.4	1.15	43	77.5	0.72	11.8	1.7
21	104	0.28	9.8	1.08	44	70.43	0.625	10	1.55
22	49	0.49	4.8	1.83	45	133.77	0.535	9.3	3.28
23	54.66	0.385	5.5	0.76	46	89.99	0.49	9.8	2.69

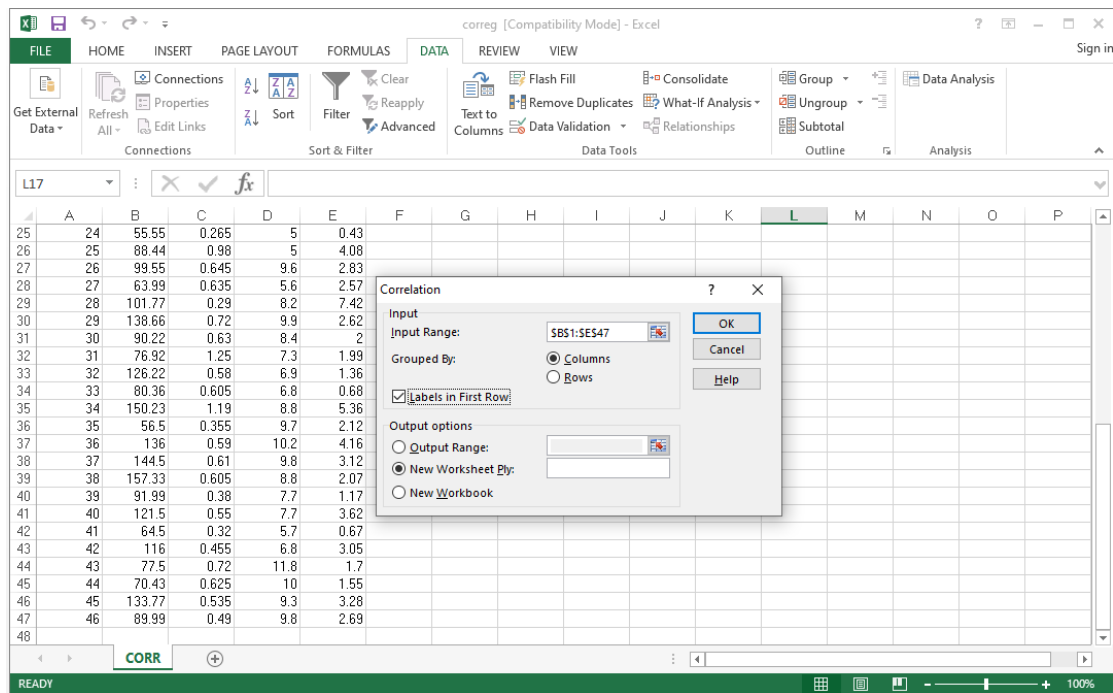
- Obtain correlation coefficient between each pair of the variables PP, PH, NGL and yield.
- Fit a multiple linear regression equation by taking yield as dependent variable and biometrical characters as explanatory variables.

## MS-Excel

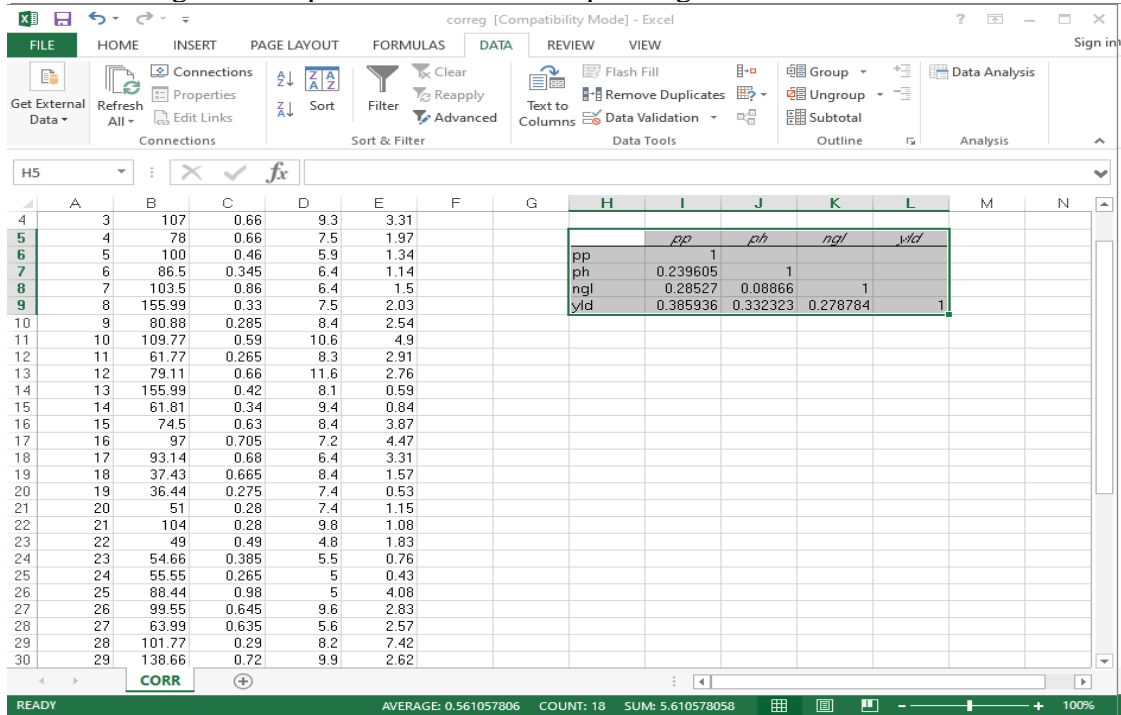
Once the data entry is complete. To obtain the Correlation coefficient between each pair of the variables click on **Data->Data Analysis->correlation**. The dialog box will appear as below



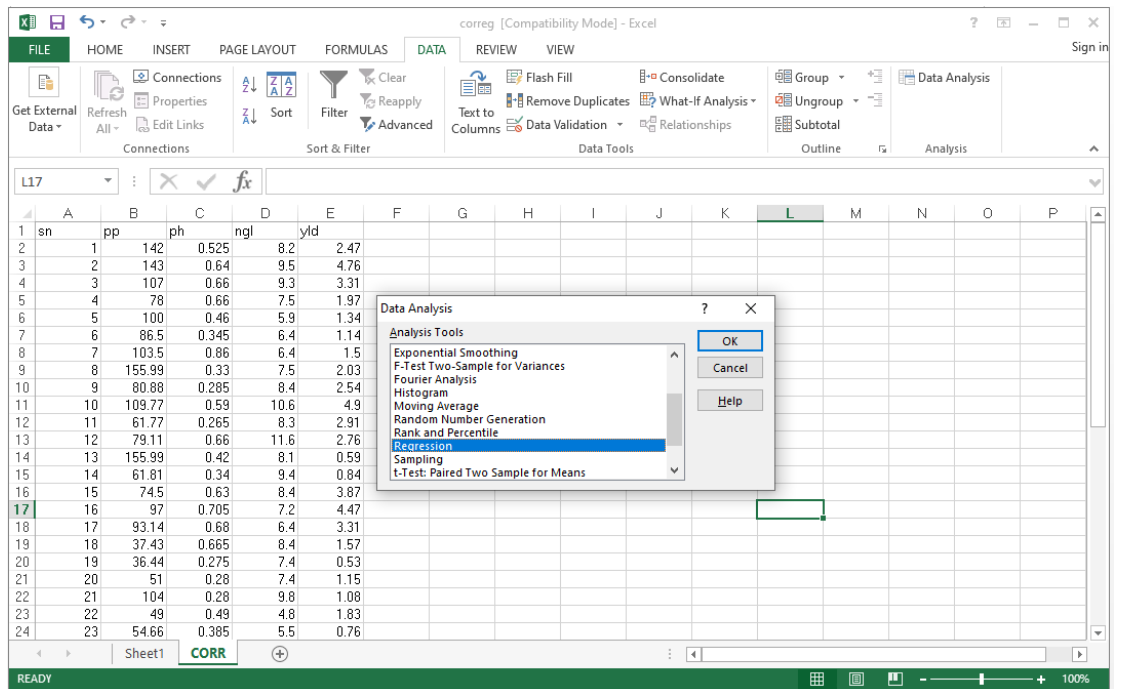
Click **OK**. This displays the dialog box for the analysis Correlation. For the two groups select the variable number of fruit Set (45days) and select the range for Variable 1 Range: and Variable 2 Range: in the Input box. Now select Output Range: to get the output. This displays the following screen.



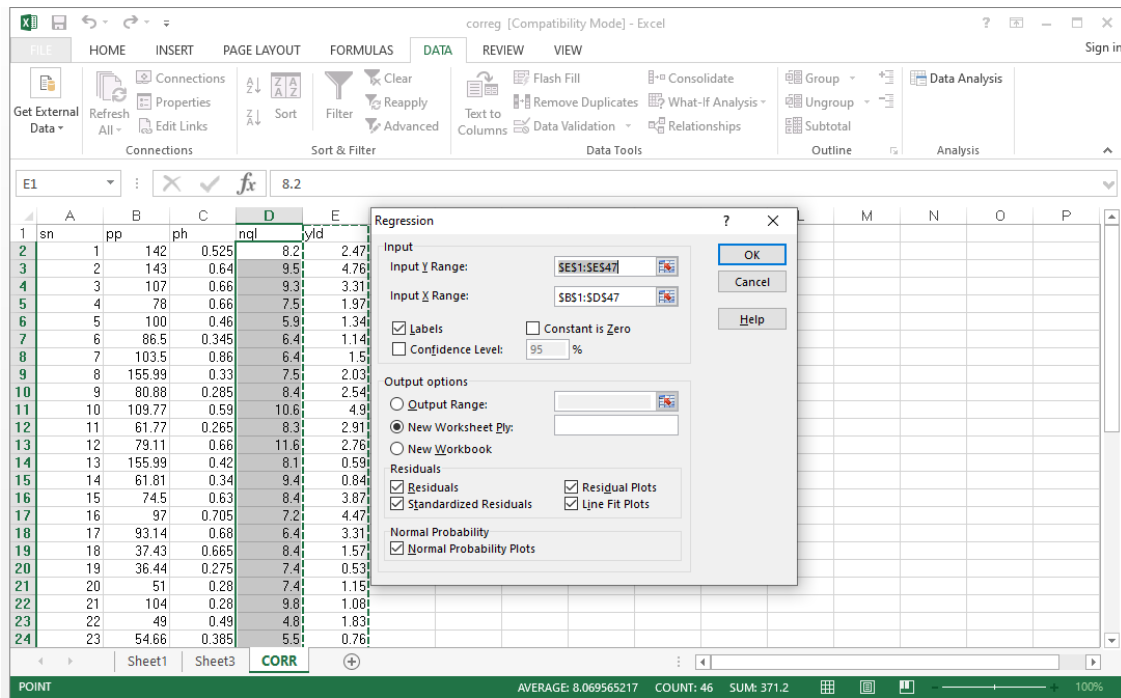
Click OK to get the output at the selected output range.



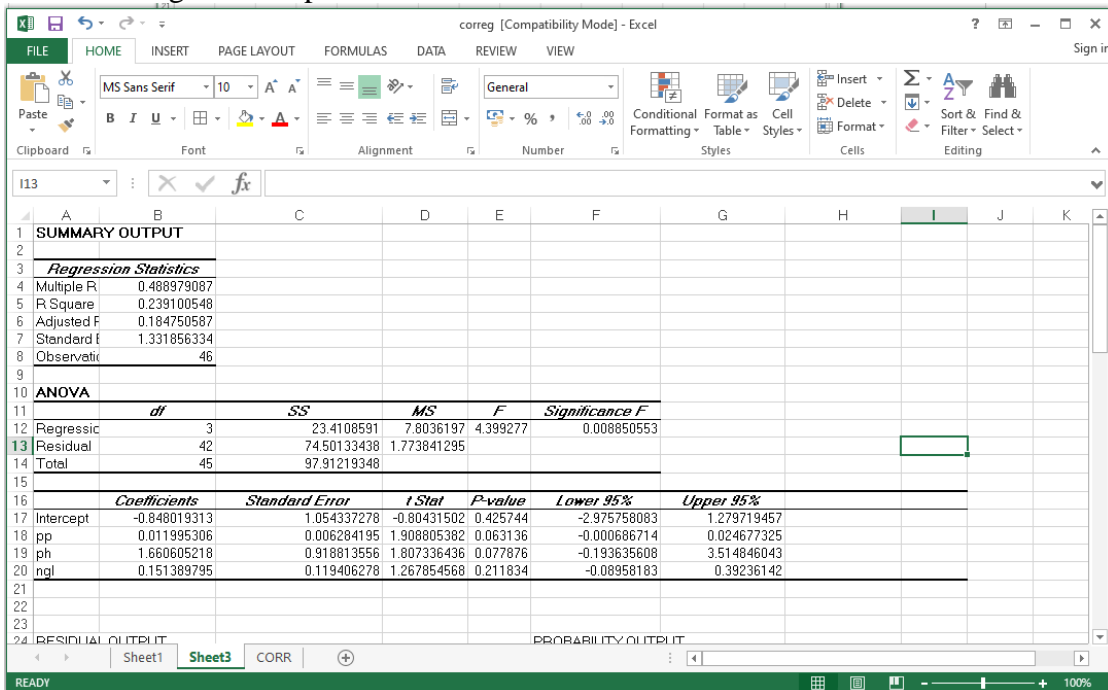
Once the data entry is complete. To obtain the multiple regression analysis between Yield as dependent variable and PP, Ph, ngl as independent variable, click on **Data->Data Analysis->Regression**. The dialog box will appear as below



Click Ok and select the range of dependent and independent variable. Also select a new worksheet for the output. Then click OK.



Click OK to get the output in a new worksheet.



The value “Multiple R” is the Pearson product-moment correlation between dependent and independent variables. The *F*-test of the significance of the multiple linear regression is mathematically and statistically identical to the *t*-test of the significance of the regression coefficient.

### 5. One-Way Between-Groups ANOVA

Excel’s Analysis ToolPak provides one-way between-groups analysis of variance (ANOVA), one-way within-subjects (repeated measures) ANOVA, and two-way

ANOVA for a balanced factorial design. The following hypothetical data will be used to illustrate the one-way between-groups analysis of variance (ANOVA).

**Example-3:** {Nigam, A.K. and Gupta V.K., 1979, *Handbook on Analysis of Agricultural experiments*, First Edition, I.A.S.R.I. Publication, New Delhi, pp16-20}. A feeding trial with 3 feeds namely (i) Pasture(control), (ii) Pasture and Concentrates and (iii) Pasture, Concentrates and Minerals was conducted at the Yellachihalli Sheep Farm, Mysore, to study their effect on wool yield of Sheep. For this purpose twenty-five ewe lambs were allotted at random to each of the three treatments and the three treatments and the weight records of the total wool yield (in gms) of first two clipping were obtained. The data for two lambs for feed 1, three for feed 2 and one for feed 3 are missing. The details of the experiment are given below:

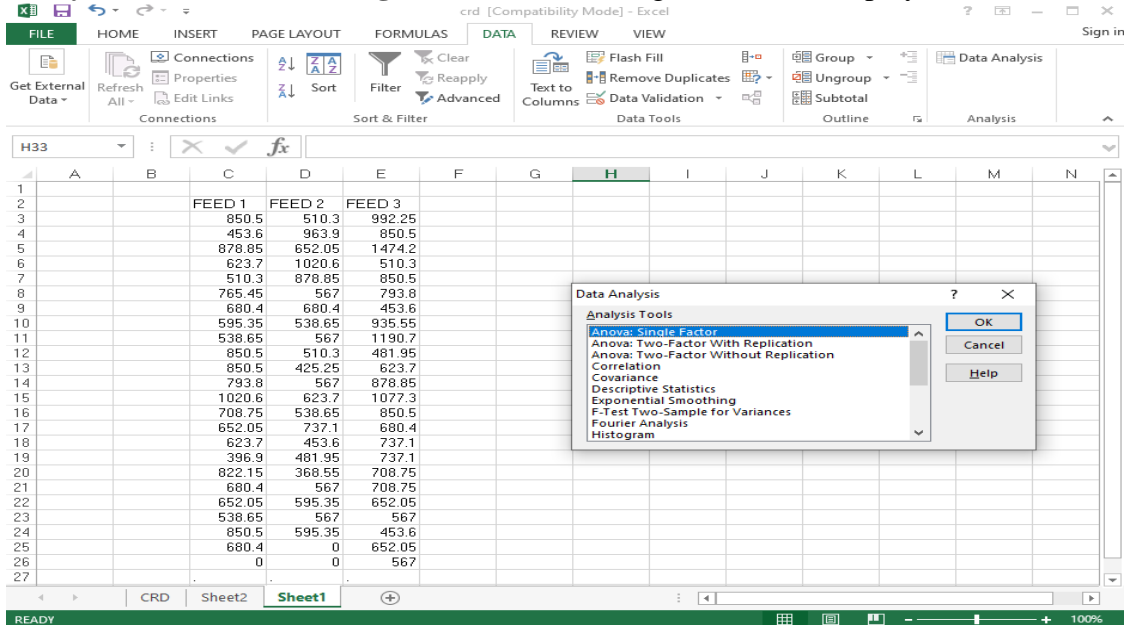
Yield (in gms)

FEED 1	FEED 2	FEED 3
850.50	510.30	992.25
453.60	963.90	850.50
878.85	652.05	1474.20
623.70	1020.60	510.30
510.30	878.85	850.50
765.45	567.00	793.80
680.40	680.40	453.60
595.35	538.65	935.55
538.65	567.00	1190.70
850.50	510.30	481.95
850.50	425.25	623.70
793.80	567.00	878.85
1020.60	623.70	1077.30
708.75	538.65	850.50
652.05	737.10	680.40
623.70	453.60	737.10
396.90	481.95	737.10
822.15	368.55	708.75
680.40	567.00	708.75
652.05	595.35	652.05
538.65	567.00	567.00
850.50	595.35	453.60
680.40	0	652.05
0	0	567.00

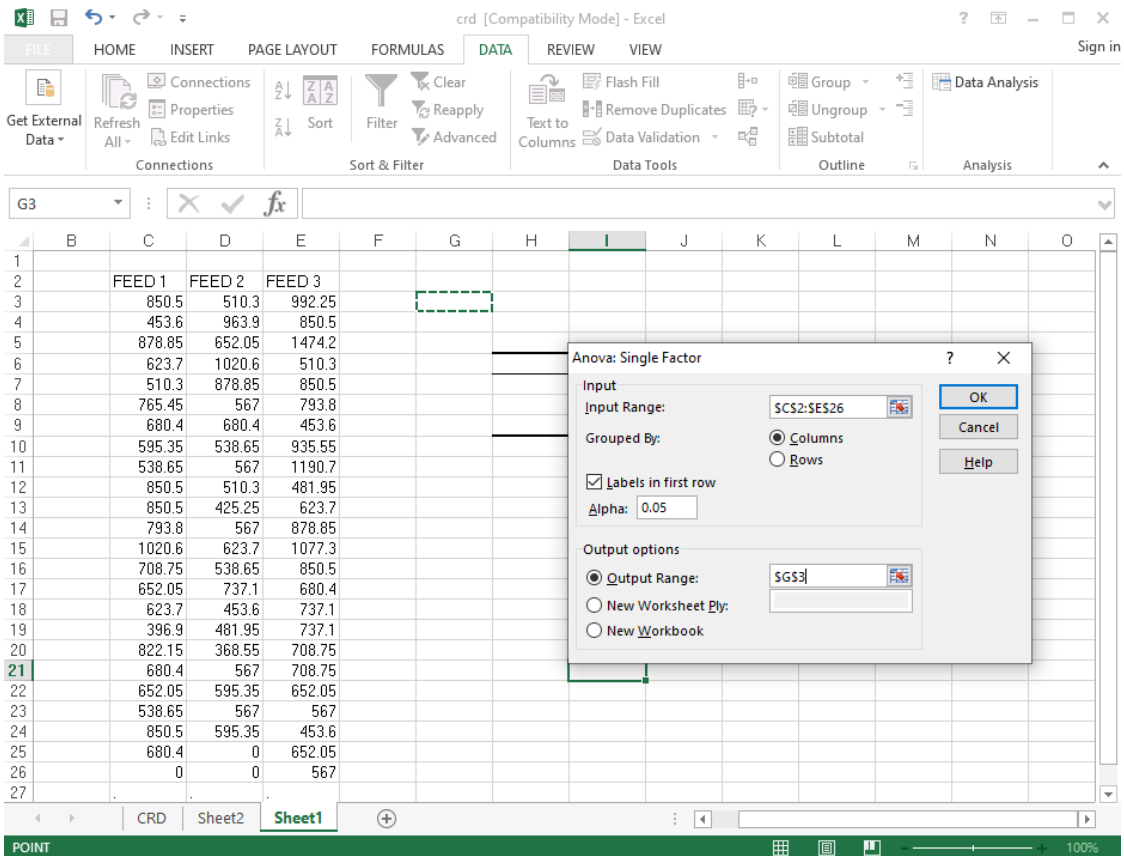
Where Feed 1- Pasture (control),  
 Feed 2- Pasture and Concentrates and  
 Feed 3- Pasture, Concentrates and Minerals.

- Perform the analysis of variance of the data to test whether there is any difference between treatment effects.

To conduct the one-way ANOVA using the Analysis ToolPak, select **Data, Data Analysis. Select Anova: Single Factor**. The dialog box will be displayed as follow



Click OK. The ANOVA tool Dialog box appears. Provide the input range including the column labels and click OK. The resulting descriptive statistics and ANOVA summary table are shown in Table-11



The resulting descriptive statistics and ANOVA summary table are shown in below image

The screenshot shows an Excel spreadsheet with the following data and ANOVA results:

Groups	Count	Sum	Average	Variance
FEED 1	24	18017.75	667.4063	42910.34
FEED 2	24	13409.55	558.7313	54057.62
FEED 3	24	18427.5	767.8125	59619.09

Source of Vari	SS	df	MS	F	Fvalue	Fcrit
Between C	524853.1	2	262426.6	5.021331	0.009206	3.129644
Within Gro	3606102	69	52262.35			
Total	4130955	71				

The significant  $F$ -ratio indicates that there is a difference among the feed for the Yield. Excel's standard functions and the Analysis ToolPak do not allow posthoc comparisons, but it is fairly easy to use Excel formulas to perform these comparisons.

## 6. Two-way ANOVA without Replication

The Analysis ToolPak tool that correctly performs one-way repeated-measures (within subjects) ANOVA is enigmatically labeled "Anova: Two-Factor without Replication." This tool produces a summary table in which the subject (row) variable is the within subjects source of variance, and the columns (conditions) factor is the between-groups source of variance.

**Example-4:** An initial varietal trial (Late Sown, irrigated) was conducted to study the performance of 20 new strains of mustard vis-a-vis four checks (Swarna Jyoti: ZC; Vardan: NC; Varuna: NC; and Kranti: NC) using a Randomized complete Block Design (RCB) design at Bhatinda with 3 replications. The seed yield in kg/ha was recorded. The details of the experiment are given below:

**Yield in kg/ha**

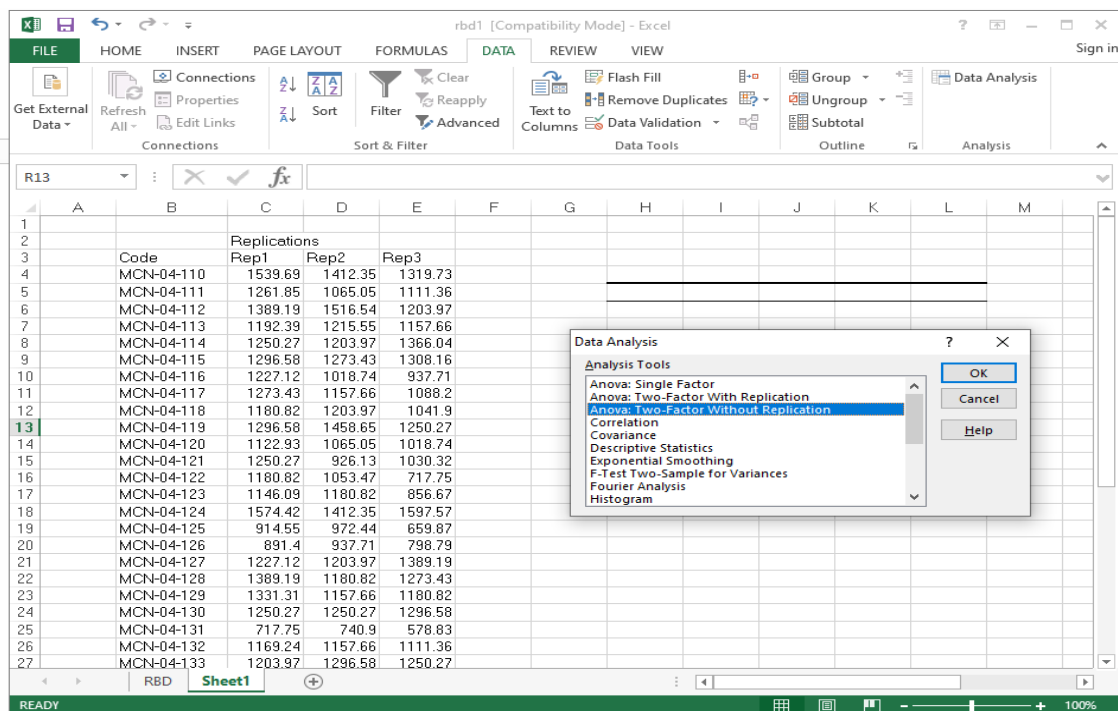
Strain	Code	Replications		
		1	2	3
RK-04-3	MCN-04-110	1539.69	1412.35	1319.73
RK-04-4	MCN-04-111	1261.85	1065.05	1111.36
RGN-124	MCN-04-112	1389.19	1516.54	1203.97
HYT-27	MCN-04-113	1192.39	1215.55	1157.66
PBR-275	MCN-04-114	1250.27	1203.97	1366.04
HUJM-03-03	MCN-04-115	1296.58	1273.43	1308.16
RGN-123	MCN-04-116	1227.12	1018.74	937.71
BIO-13-01	MCN-04-117	1273.43	1157.66	1088.20

RH-0115	MCN-04-118	1180.82	1203.97	1041.90
RH-0213	MCN-04-119	1296.58	1458.65	1250.27
NRCDR-05	MCN-04-120	1122.93	1065.05	1018.74
NRC-323-1	MCN-04-121	1250.27	926.13	1030.32
RRN-596	MCN-04-122	1180.82	1053.47	717.75
RRN-597	MCN-04-123	1146.09	1180.82	856.67
CS-234-2	MCN-04-124	1574.42	1412.35	1597.57
RM-109	MCN-04-125	914.55	972.44	659.87
BAUSM-2000	MCN-04-126	891.40	937.71	798.79
NPJ-99	MCN-04-127	1227.12	1203.97	1389.19
<b>SWARNA JYOTI(ZC)</b>	<b>MCN-04-128</b>	1389.19	1180.82	1273.43
<b>VARDAN(NC)</b>	<b>MCN-04-129</b>	1331.31	1157.66	1180.82
PR-2003-27	MCN-04-130	1250.27	1250.27	1296.58
<b>VARUNA(NC)</b>	<b>MCN-04-131</b>	717.75	740.90	578.83
PR-2003-30	MCN-04-132	1169.24	1157.66	1111.36
<b>KRANTI-(NC)</b>	<b>MCN-04-133</b>	1203.97	1296.58	1250.27

**Note:** Strains of mustard in bold are the four checks.

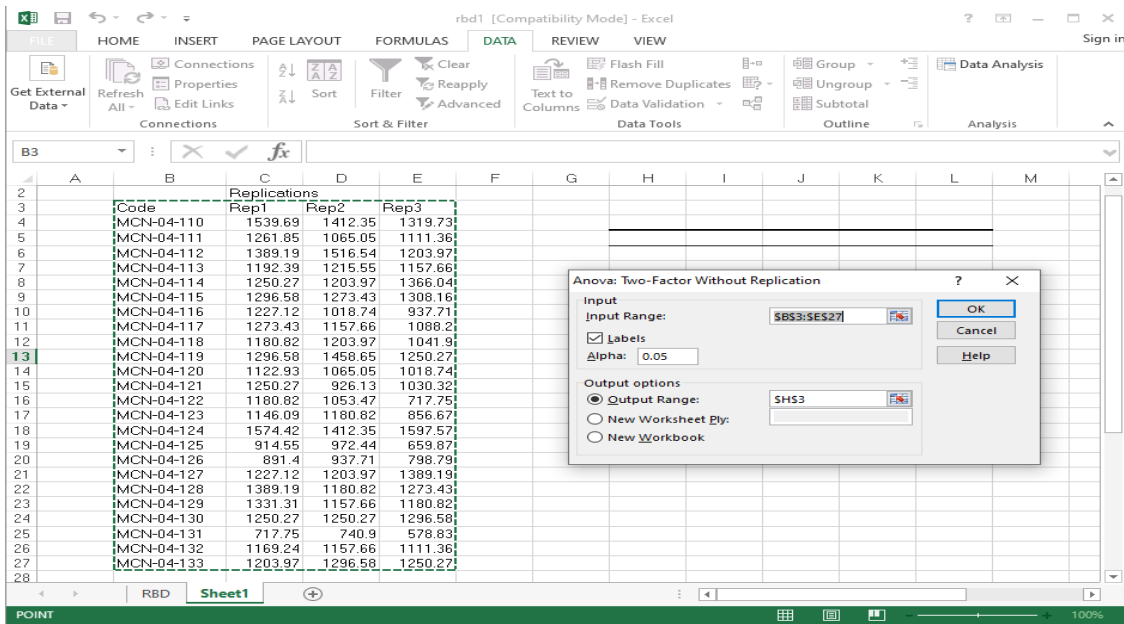
1. Perform the analysis of variance of the data to test whether there is any difference between treatment effects.

To conduct the Two-way ANOVA, use the “Anova: **Two-Factor without Replication** tool” in the Analysis ToolPak, as shown in Figure below.

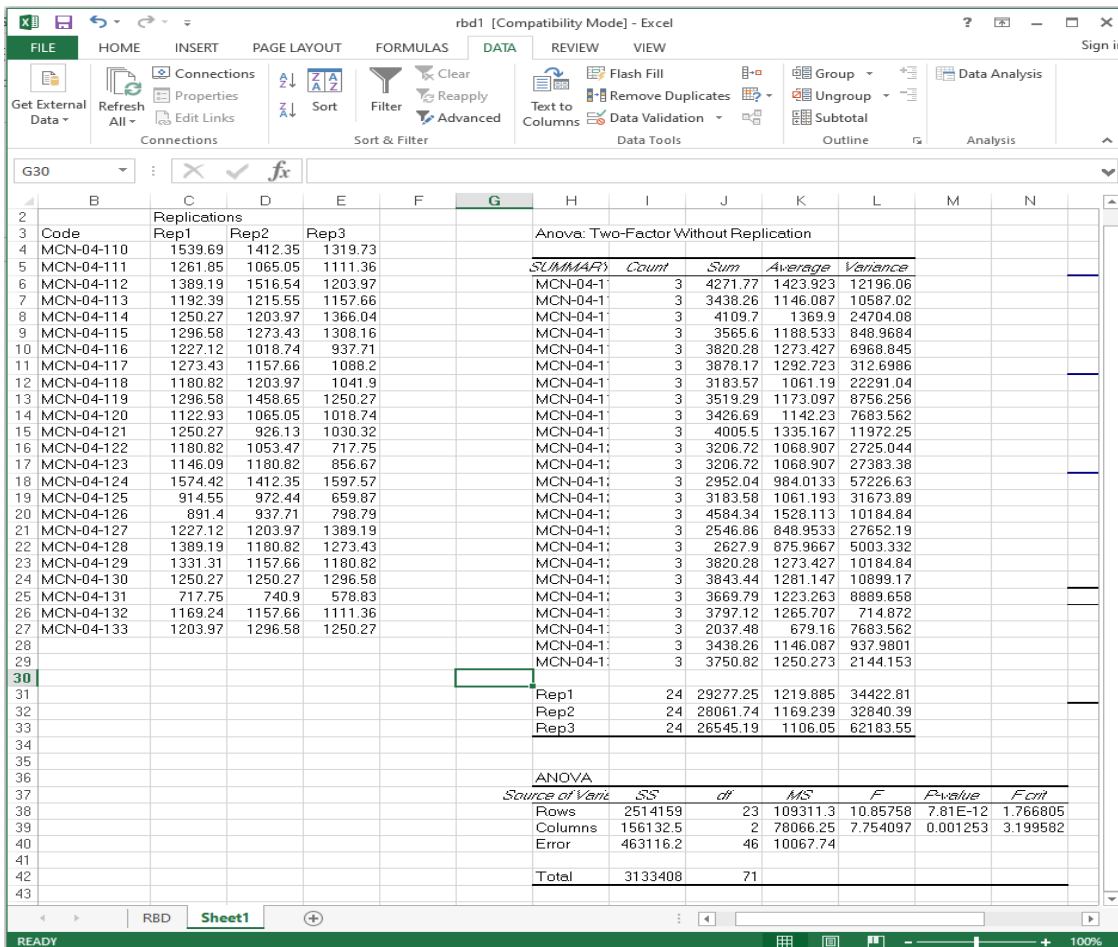




If you use data labels, be sure to include the column of code numbers in the input range. It is the “rows” variable.



The output is placed by default in a new worksheet ply, or optionally in the same worksheet with the data. The ANOVA summary table copied from Excel appears screen.



The usual test of interest is the *F*-ratio for “Columns,” the treatment comparison. It is

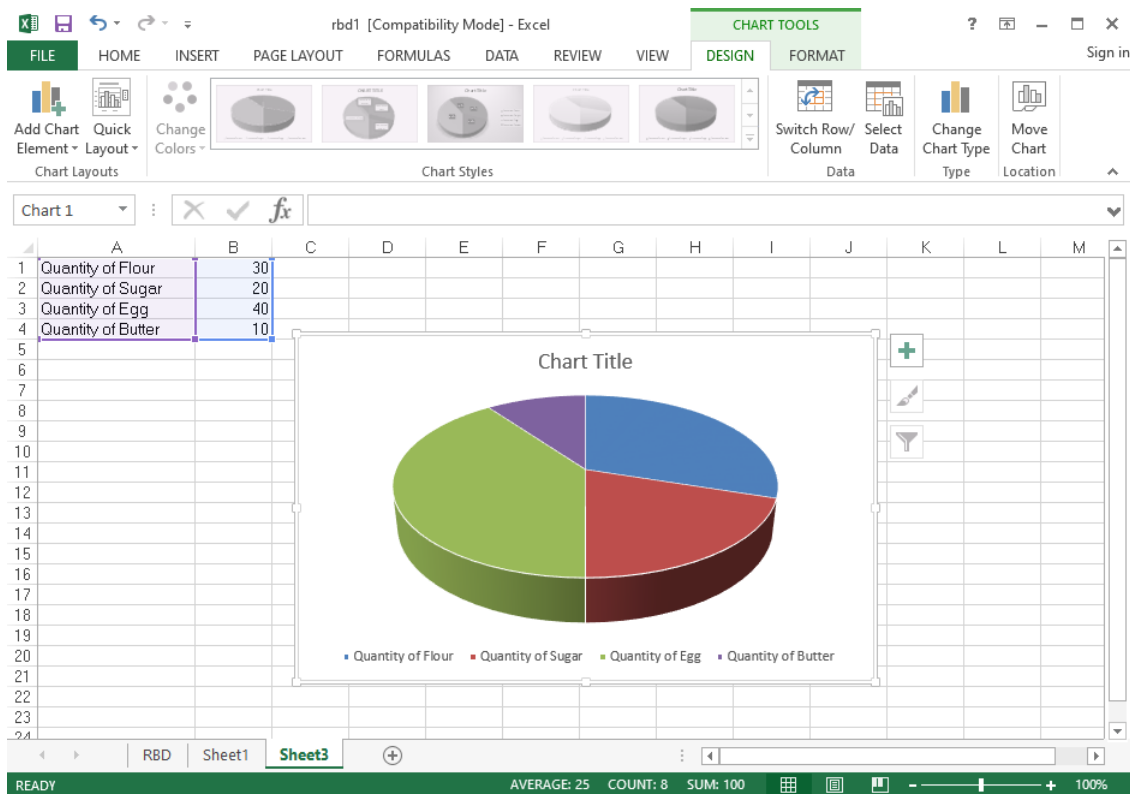
Typically of little interest to know whether the within-subjects or “Rows” *F-ratio* is significant. This test is telling us what we already know—that different treatments have different contribution towards yield.

## 7. Data Representation

Charts and graphs are easily constructed and modified in Excel, and can be displayed alongside the data for ease of visualization. The charts and graphs are dynamically linked to the data, and update automatically when the data are modified.

### *Pie Charts*

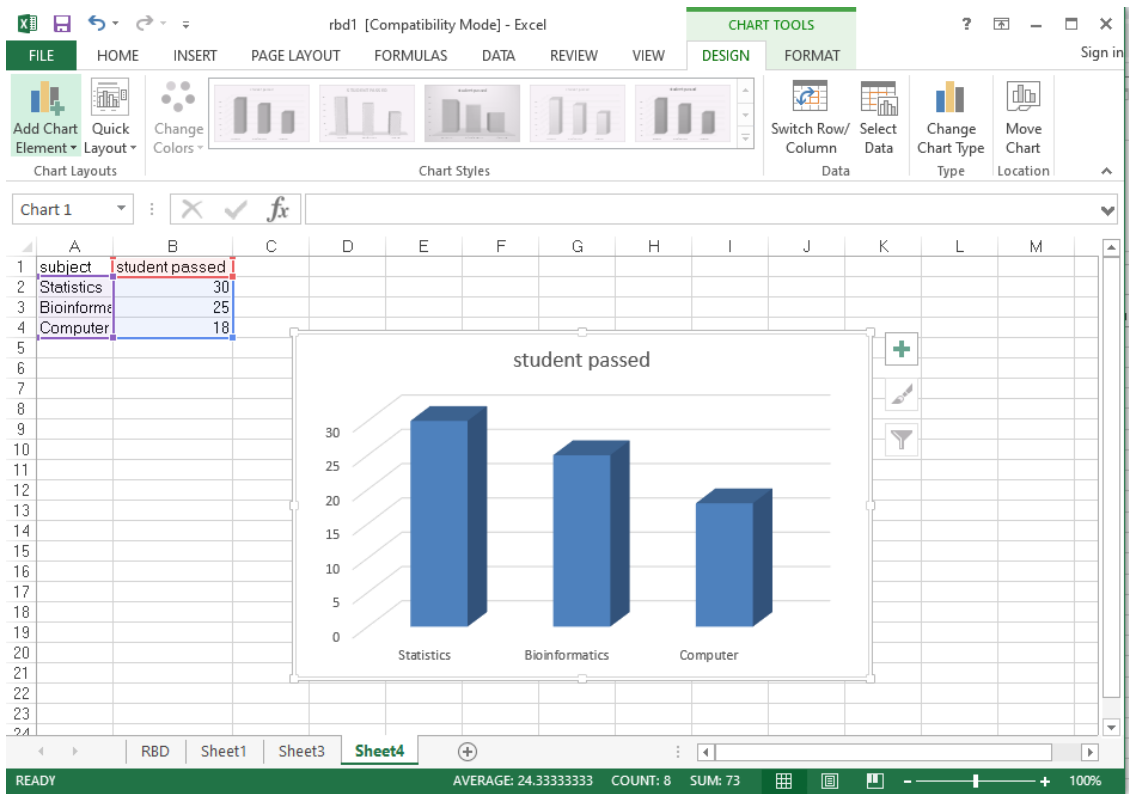
A pie chart divides a circle into relatively larger and smaller slices based on the relative frequencies or percentages of observations in different categories. To construct a pie chart in Excel, select the data, including the labels, and then select the chart icon in the Standard Toolbar (or select **Insert, Chart** from the Menu Bar). From the Chart Wizard, select Pie as the chart type and accept the defaults to place the new chart as an object in the current worksheet. To add percentages or labels, you can right-click on the pie chart and select “Format Data Series.” See the finished pie chart below



### *Bar Charts*

The same data will be used to produce a bar chart. Although Excel has a chart type labeled “bar,” it is horizontal in layout and should generally be avoided. To generate a bar chart in Excel, follow the steps above, but select “Column” as the chart type. You may ornament the bar chart by varying the colors of the bars. To do so, right-click on one of the bars and then select “Format Data Series.” Select the option “Vary colors by point” under the Options tab

## MS-Excel



Similarly we can construct Histograms, line graphs and scatter plots by using MS-Excel. Also one can visualized multiple series data in a single chart

---

---

# BASIC STATISTICAL TECHNIQUE IN EXCEL

---

---

**Md Yeasin, Ajit Gupta, Ranjit Kumar Paul**

*ICAR-Indian Agricultural Statistics Research Institute*

*Library Avenue, New Delhi - 110 012*

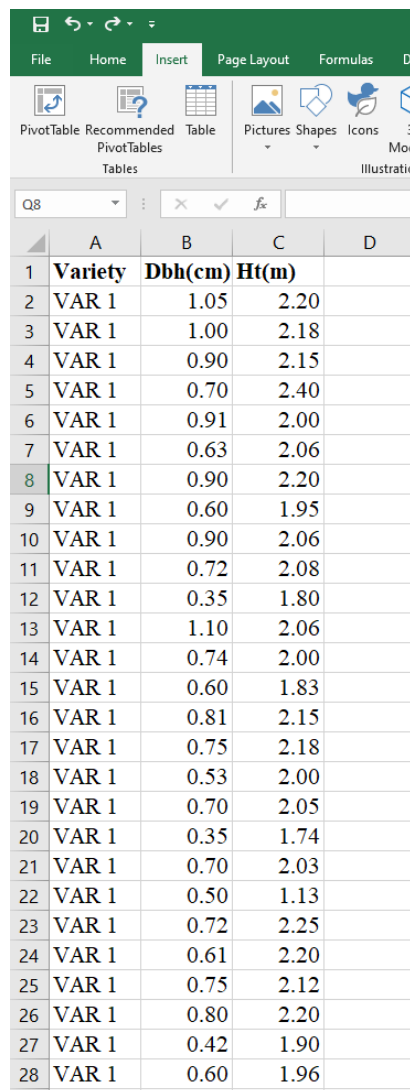
*[yeasin.iasri@gmail.com](mailto:yeasin.iasri@gmail.com) , [ajit@icar.gov.in](mailto:ajit@icar.gov.in) , [ranjit.paul@icar.gov.in](mailto:ranjit.paul@icar.gov.in)*

---

---

## 1. Data sets:

Data Description: We have two quantitative variables (diameter and height) for four variables (Var 1, var 2, var 3, and var 4) of forest plants.



	A	B	C	D
1	<b>Variety</b>	<b>Dbh(cm)</b>	<b>Ht(m)</b>	
2	VAR 1	1.05	2.20	
3	VAR 1	1.00	2.18	
4	VAR 1	0.90	2.15	
5	VAR 1	0.70	2.40	
6	VAR 1	0.91	2.00	
7	VAR 1	0.63	2.06	
8	VAR 1	0.90	2.20	
9	VAR 1	0.60	1.95	
10	VAR 1	0.90	2.06	
11	VAR 1	0.72	2.08	
12	VAR 1	0.35	1.80	
13	VAR 1	1.10	2.06	
14	VAR 1	0.74	2.00	
15	VAR 1	0.60	1.83	
16	VAR 1	0.81	2.15	
17	VAR 1	0.75	2.18	
18	VAR 1	0.53	2.00	
19	VAR 1	0.70	2.05	
20	VAR 1	0.35	1.74	
21	VAR 1	0.70	2.03	
22	VAR 1	0.50	1.13	
23	VAR 1	0.72	2.25	
24	VAR 1	0.61	2.20	
25	VAR 1	0.75	2.12	
26	VAR 1	0.80	2.20	
27	VAR 1	0.42	1.90	
28	VAR 1	0.60	1.96	

Here we are going to use the Data Analysis Tools Pack, Pivot Table and some inbuilt functions of excel. We have enlisted and presented below.

## 2. Data Analysis Tools Pack

### Add Data Analysis Tools Pack in excel

- Click the File tab, click Options, and then click the Add-Ins category.

## Basic Statistical Technique in Excel

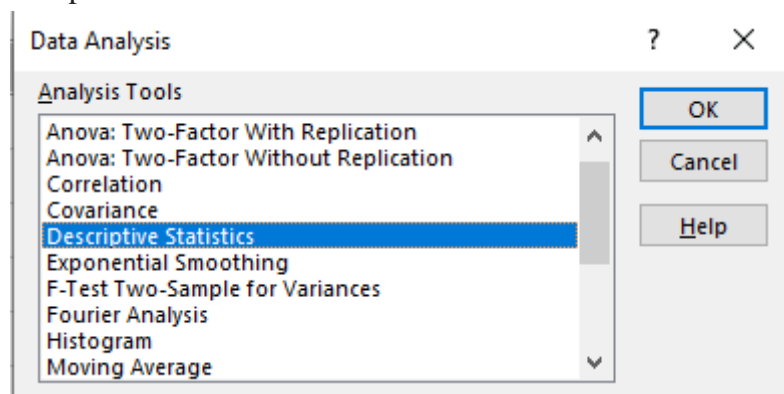
- In the Manage box, select Excel Add-ins and then click Go.
- In the Add-Ins box, check the Analysis ToolPak check box, and then click OK.
- If Analysis ToolPak is not listed in the Add-Ins available box, click Browse to locate it.
- If you are prompted that the Analysis ToolPak is not currently installed on your computer, click Yes to install it.

### The Analysis ToolPak contains the following tools:

- Correlation analysis tool
- Covariance analysis tool
- **Descriptive Statistics analysis tool**
- Exponential Smoothing analysis tool
- Fourier Analysis tool
- F-Test: Two-Sample for Variances analysis tool
- Histogram analysis tool
- Moving Average analysis tool
- Perform a t-Test analysis
- Random Number Generation analysis tool
- Rank and Percentile analysis tool
- Regression analysis tool
- Sampling analysis tool
- z-Test: Two-Sample for Means analysis tool

In this practical, Descriptive Statistics analysis tool will be used. Steps are as follows

- On the Data tab, in the Analysis group, click Data Analysis.
- Select Descriptive Statistics and click OK.



## Basic Statistical Technique in Excel

- Select the data range as the Input Range.
- Select the Output Range.
- Check in Summary statistics
- Click OK.

The screenshot shows the 'Descriptive Statistics' dialog box. In the 'Input' section, the 'Input Range' is '\$B\$1:\$C\$390', 'Grouped By' is 'Columns', and 'Labels in first row' is checked. In the 'Output options' section, 'New Worksheet Ply' is selected, 'Summary statistics' is checked, and 'Confidence Level for Mean' is set to 95%. There are also options for 'Kth Largest' and 'Kth Smallest', both set to 1.

<i>Dbh(cm)</i>		<i>Ht(m)</i>	
Mean	0.664087404	Mean	1.919254499
Standard Error	0.01436742	Standard Error	0.014493563
Median	0.63	Median	1.95
Mode	0.7	Mode	2
Standard Deviation	0.283369822	Standard Deviation	0.285857743
Sample Variance	0.080298456	Sample Variance	0.081714649
Kurtosis	1.815086722	Kurtosis	-0.475442555
Skewness	0.983837704	Skewness	-0.277022703
Range	1.71	Range	1.53
Minimum	0.23	Minimum	1.08
Maximum	1.94	Maximum	2.61
Sum	258.33	Sum	746.59
Count	389	Count	389
Confidence Level(95.0%)	0.02824774	Confidence Level(95.0%)	0.028495749

### 3. Pivot Table

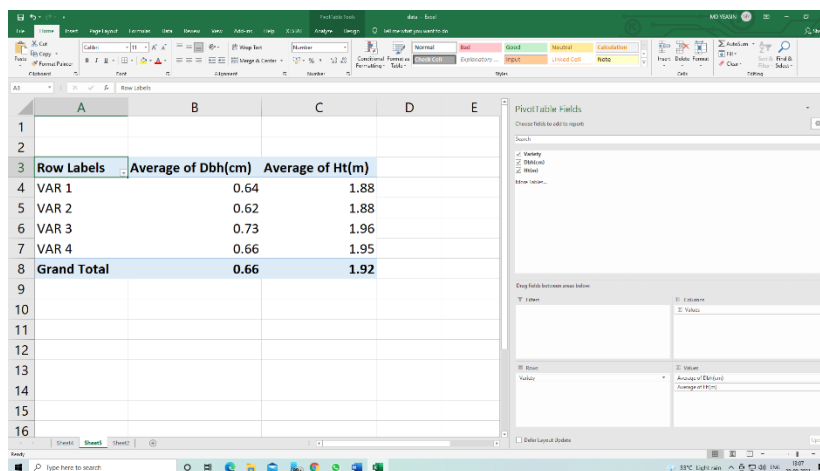
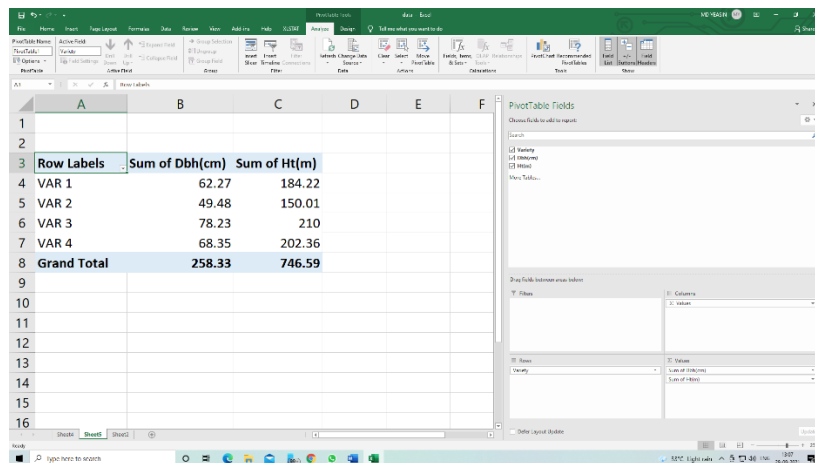
A PivotTable is a powerful tool to calculate, summarize, and analyze data that lets you see comparisons, patterns, and trends in your data.

#### Steps to create a pivot table:

- Select the cells you want to create a PivotTable from.
- Then, Select Insert > PivotTable.
- Choose the data that you want to analyze, in Select a table or range.

## Basic Statistical Technique in Excel

- Then choose where you want the PivotTable report to be placed, select New worksheet to place the PivotTable in a new worksheet or Existing worksheet, and then select the location you want the PivotTable to appear.
- Then select OK.
- After that build your pivot table.
- To add a field to your PivotTable, select the field name checkbox in the PivotTables Fields pane.
- To move a field from one area to another, drag the field to the target area.



### 4. List of excel functions

Function Name	Description
COUNT function	Counts how many numbers are in the list of arguments
COUNTBLANK function	Counts the number of blank cells within a range
COUNTIF function	Counts the number of cells within a range that meet the given criteria
FREQUENCY function	Returns a frequency distribution as a vertical array
LARGE function	Returns the k-th largest value in a data set
SMALL function	Returns the k-th smallest value in a data set
MAX function	Returns the maximum value in a list of arguments



## Basic Statistical Technique in Excel

MIN function	Returns the minimum value in a list of arguments
AVERAGE function	Returns the average of its arguments
<b>TRIMMEAN function</b>	Returns the mean of the interior of a data set
MEDIAN function	Returns the median of the given numbers
MODE.MULT function	Returns a vertical array of the most frequently occurring, or repetitive values in an array or range of data
MODE.SNGL function	Returns the most common value in a data set
PERCENTILE.EXC function	Returns the k-th percentile of values in a range, where k is in the range 0..1, exclusive.
QUARTILE.EXC function	Returns the quartile of the data set, based on percentile values from 0..1, exclusive
QUARTILE.INC function	Returns the quartile of a data set
PERCENTILE.INC function	Returns the k-th percentile of values in a range
PERCENTRANK.EXC function	Returns the rank of a value in a data set as a percentage (0..1, exclusive) of the data set
PERCENTRANK.INC function	Returns the percentage rank of a value in a data set
AVEDEV function	Returns the average of the absolute deviations of data points from their mean
COVARIANCE.P function	Returns covariance, the average of the products of paired deviations
COVARIANCE.S function	Returns the sample covariance, the average of the products deviations for each data point pair in two data sets
VAR.P function	Calculates variance based on the entire population
VAR.S function	Estimates variance based on a sample
STDEV.P function	Calculates standard deviation based on the entire population
STDEV.S function	Estimates standard deviation based on a sample
DEVSQ function	Returns the sum of squares of deviations
SKEW function	Returns the skewness of a distribution
SKEW.P function	Returns the skewness of a distribution based on a population
KURT function	Returns the kurtosis of a data set
CORREL function	Returns the correlation coefficient between two data sets

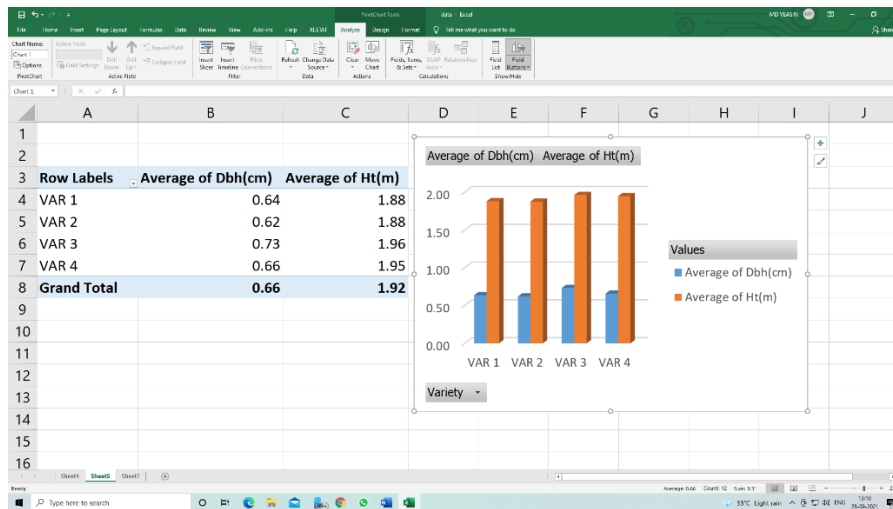
### 5. Chart

#### 5.1. Bar Chart

**Select data > click Insert > Insert Column or Bar Chart.**

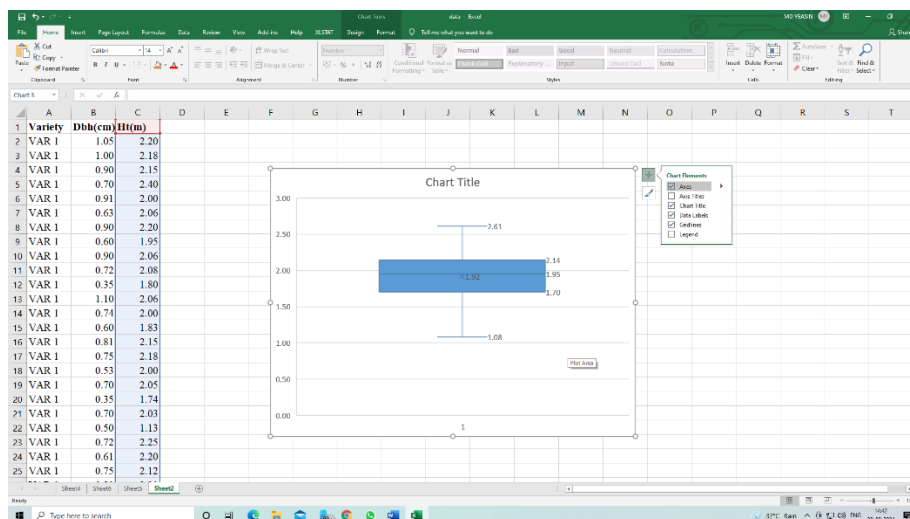
Various column charts are available, but to insert a standard bar chart, click the “Clustered Chart” option. This chart is the first icon listed under the “2-D Column or 3-D Column” section.

# Basic Statistical Technique in Excel



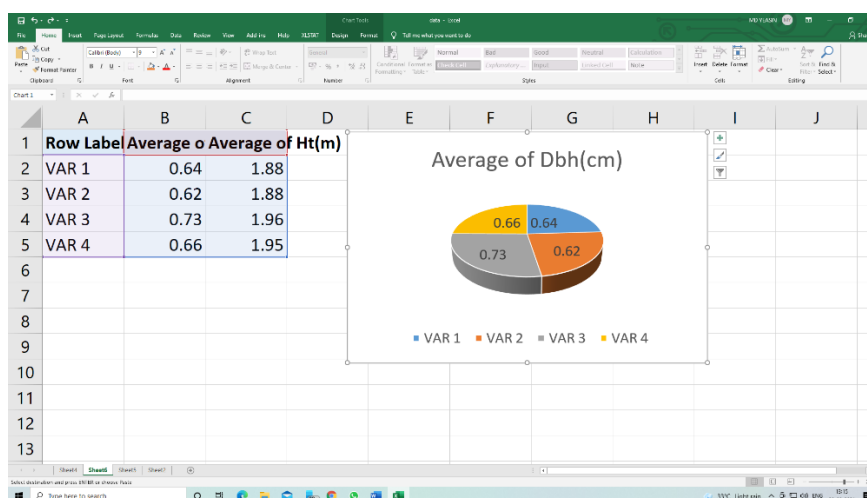
## 5.2. Histograms:

Select data > click Insert > Histogram > Insert Histogram Plot



## 5.3. Pie Charts

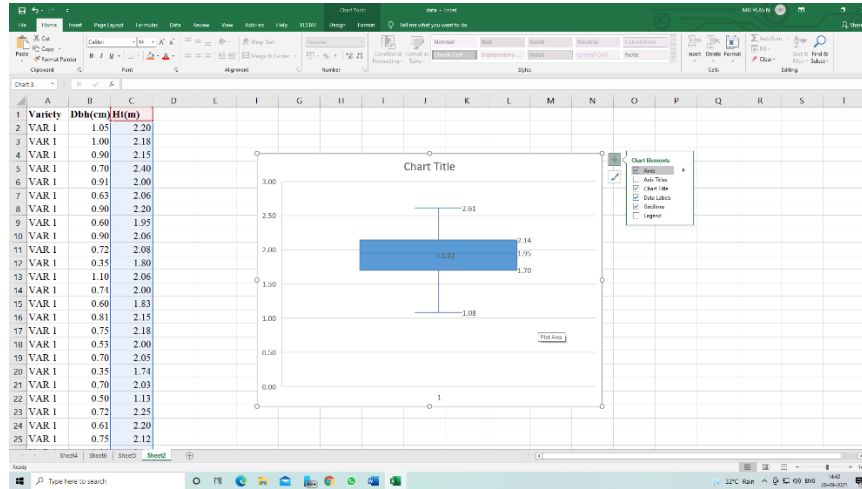
Select data > click Insert > Insert Pie Chart



# Basic Statistical Technique in Excel

## 5.4.Box Plots

Select data > click Insert > Histogram> Insert Box and Whisker Plot



## 6. Robust mean (Trimmed mean)

Formula for k trimmed mean:  $=TRIMMEAN(\text{data range}, .k)$

[we used formula  $=TRIMMEAN(C:C,0.2)$ ]

In the formula, it did not consider the 20% observation from both the side.

Dbh(cm)	Ht(m)
0.64214	1.9269

---

---

## OVERVIEW ON R SOFTWARE AND RSTUDIO

---

---

**Soumen Pal and B N Mandal**

*ICAR-Indian Agricultural Statistics Research Institute*

*Library Avenue, New Delhi - 110 012*

[soumen.pal@icar.gov.in](mailto:soumen.pal@icar.gov.in)

---

---

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

### **R environment**

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

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

### **Origin**

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

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

### **Availability**

Since R is an open source project, it can be obtained freely from the website [www.r-project.org](http://www.r-project.org). One can download R from any CRAN mirror out of several CRAN (Comprehensive R Archive Network) mirrors. Latest available version of R is *R version 4.1.1* and it has been released on 10.08.2021.

### **Installation**

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

## Usage

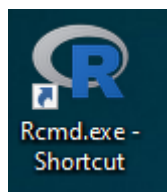
R can work under Windows, UNIX and Mac OS. In this note, we consider usage of R in Windows set up only.

## Difference with other packages

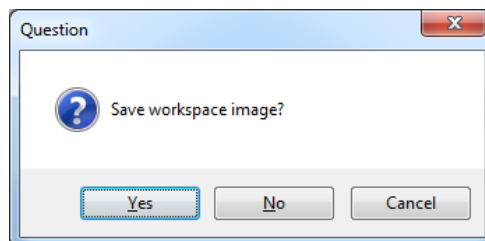
There is an important difference between R and the other statistical packages. In R, a statistical analysis is normally done as a series of steps, with intermediate results being stored in objects. Thus whereas SAS and SPSS will give large amount of output from a given analysis, R will give minimal output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

## Invoking R

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



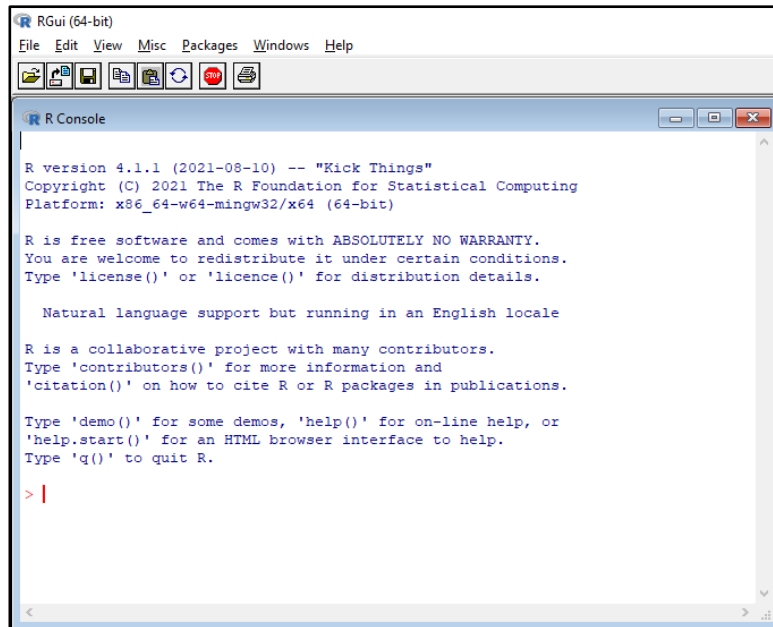
To quit R, type  $q()$  at the R prompt ( $>$ ) and press Enter key. A dialog box will ask whether to save the objects you have created during the session so that they will become available next time when R will be invoked.



## Windows of R

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

## Overview on R Software and RStudio



### R commands

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

### Executing commands from or diverting output to a file

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

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

For Windows Source is also available on the File menu.

The function *sink()*,

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

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

```
> sink()
```

restores it to the console once again.

### Simple manipulations of numbers and vectors

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

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

The function `c()` assigns the five numbers to the vector `x`. The assignment operator (`<-`) ‘points’ to the object receiving the value of the expression. One can use the ‘=’ operator as an alternative.

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

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

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

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

```
> 1/x
```

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

### The elementary arithmetic operators

- + addition
- subtraction
- \* multiplication
- / division
- ^ exponentiation

### Arithmetic functions

`log`, `exp`, `sin`, `cos`, `tan`, `sqrt`,

### Other basic functions

`max(x)` – maximum element of vector `x`,  
`min(x)` – minimum element of vector `x`,  
`range(x)` – range of the values of vector `x`,  
`length(x)` – the number of elements in `x`,  
`sum(x)` – the total of the elements in `x`,  
`prod(x)` – product of the elements in `x`  
`mean(x)` – average of the elements of `x`

var(x) – sample variance of the elements of (x)  
 sort(x) – returns a vector with elements sorted in increasing order.

**Logical operators**

< - less than  
 <= less than or equal to  
 > greater than  
 >= greater than or equal to  
 == equal to  
 != not equal to.

**Other objects in R**

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

**Matrix facilities**

A matrix is just an array with two subscripts. R provides many operators and functions those are available only for matrices. Some of the important R functions for matrices are

t(A) – transpose of the matrix A  
 nrow(A) – number of rows in the matrix A  
 ncol(A) – number of columns in the matrix A  
 A%% B– Cross product of two matrices A and B  
 A\*B – element by element product of two matrices A and B  
 diag (A) – gives a vector of diagonal elements of the square matrix A  
 diag(a) – gives a matrix with diagonal elements as the elements of vector a  
 eigen(A) – gives eigen values and eigen vectors of a symmetric matrix A  
 rbind (A,B) – concatenates two matrix A and B by appending B matrix below A matrix  
 (They should have same number of columns)  
 cbind(A, B) - concatenates two matrix A and B by appending B matrix in the right of A matrix  
 (They should have same number of rows)

**Data frame**

Data frame is an array consisting of columns of various mode (numeric, character, etc). Small to moderate size data frame can be constructed by *data.frame()* function. For example, following is an illustration how to construct a data frame from the car data\*:

Make	Model	Cylinder	Weight	Mileage	Type
Honda	Civic	V4	2170	33	Sporty
Chevrolet	Beretta	V4	2655	26	Compact
Ford	Escort	V4	2345	33	Small
Eagle	Summit	V4	2560	33	Small



## Overview on R Software and RStudio

Volkswagen	Jetta	V4	2330	26	Small
Buick	Le Sabre	V6	3325	23	Large
Mitsubishi	Galant	V4	2745	25	Compact
Dodge	Grand Caravan	V6	3735	18	Van
Chrysler	New Yorker	V6	3450	22	Medium
Acura	Legend	V6	3265	20	Medium

```
> Make<-c("Honda","Chevrolet","Ford","Eagle","Volkswagen","Buick","Mitsbusihi",
+ "Dodge","Chrysler","Acura")
> Model=c("Civic","Beretta","Escort","Summit","Jetta","Le Sabre","Galant",
+ "Grand Caravan","New Yorker","Legend")
```

Note that the plus sign (+) in the above commands are automatically inserted when the carriage return is pressed without completing the list. Save some typing by using *rep()* command. For example, *rep("V4",5)* instructs R to repeat V4 five times.

```
> Cylinder<-c(rep("V4",5),"V6","V4",rep("V6",3))
> Cylinder
[1] "V4" "V4" "V4" "V4" "V4" "V6" "V4" "V6" "V6" "V6"
> Weight<-c(2170,2655,2345,2560,2330,3325,2745,3735,3450,3265)
> Mileage<-c(33,26,33,33,26,23,25,18,22,20)
> Type<-
c("Sporty","Compact",rep("Small",3),"Large","Compact","Van",rep("Medium",2))
```

Now *data.frame()* function combines the six vectors into a single data frame.

```
> Car<-data.frame(Make,Model,Cylinder,Weight,Mileage,Type)
> Car
  Make      Model Cylinder Weight Mileage  Type
1  Honda     Civic      V4   2170     33 Sporty
2 Chevrolet Beretta     V4   2655     26 Compact
3   Ford     Escort     V4   2345     33 Small
4   Eagle   Summit     V4   2560     33 Small
5 Volkswagen Jetta     V4   2330     26 Small
6   Buick   Le Sabre     V6   3325     23 Large
7 Mitsbusihi Galant     V4   2745     25 Compact
8   Dodge Grand Caravan V6   3735     18 Van
9  Chrysler New Yorker  V6   3450     22 Medium
10  Acura   Legend     V6   3265     20 Medium

> names(Car)
[1] "Make"      "Model"     "Cylinder"  "Weight"    "Mileage"   "Type"
```

Just as in matrix objects, partial information can be easily extracted from the data frame:

```
> Car[1,]
  Make Model Cylinder Weight Mileage  Type
1 Honda Civic      V4   2170     33 Sporty
```

## Overview on R Software and RStudio

In addition, individual columns can be referenced by their labels:

```
> Car$Mileage
[1] 33 26 33 33 26 23 25 18 22 20
> Car[,5]           #equivalent expression
> mean(Car$Mileage) #average mileage of the 10 vehicles
[1] 25.9
> min(Car$Weight)
[1] 2170
```

*table()* command gives a frequency table:

```
> table(Car$Type)

Compact   Large   Medium   Small   Sporty   Van
         2         1         2         3         1         1
```

If the proportion is desired, type the following command instead:

```
> table(Car$Type)/10

Compact   Large   Medium   Small   Sporty   Van
      0.2     0.1     0.2     0.3     0.1     0.1
```

Note that the values were divided by 10 because there are that many vehicles in total. If you don't want to count them each time, the following does the trick:

```
> table(Car$Type)/length(Car$Type)
```

Cross tabulation is very easy, too:

```
> table(Car$Make, Car$Type)

           Compact Large Medium Small Sporty Van
Acura      0         0         1         0         0         0
Buick      0         1         0         0         0         0
Chevrolet  1         0         0         0         0         0
Chrysler   0         0         1         0         0         0
Dodge      0         0         0         0         0         1
Eagle      0         0         0         1         0         0
Ford       0         0         0         1         0         0
Honda      0         0         0         0         1         0
Mitsubishi 1         0         0         0         0         0
Volkswagen 0         0         0         1         0         0
```

What if you want to arrange the data set by vehicle weight? *order()* gets the job done.

```
> i<-order(Car$Weight);i
[1] 1 5 3 4 2 7 10 6 9 8
> Car[i,]
  Make      Model Cylinder Weight Mileage  Type
1  Honda      Civic         V4   2170     33 Sporty
5 Volkswagen  Jetta         V4   2330     26   Small
```

## Overview on R Software and RStudio

3	Ford	Escort	V4	2345	33	Small
4	Eagle	Summit	V4	2560	33	Small
2	Chevrolet	Beretta	V4	2655	26	Compact
7	Mitsubishi	Galant	V4	2745	25	Compact
10	Acura	Legend	V6	3265	20	Medium
6	Buick	Le Sabre	V6	3325	23	Large
9	Chrysler	New Yorker	V6	3450	22	Medium
8	Dodge	Grand Caravan	V6	3735	18	Van

### Creating/editing data objects

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

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

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

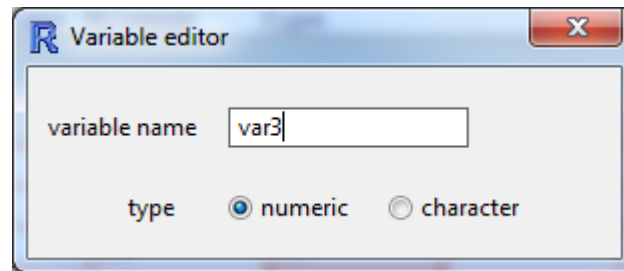
If you prefer entering the data.frame in a spreadsheet style data editor, the following command invokes the built-in editor with an empty spreadsheet.


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

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

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

You can also change the variable name by clicking once on the cell containing it. Doing so opens a dialog box:



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

```
> data1
```

### Reading data from files

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

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

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

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

By default numeric items (except row labels) are read as numeric variables and non-numeric variables, such as isNew in the example, as factors. This can be changed if necessary.

The function `read.table()` can then be used to read the data frame directly

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

### Reading comma delimited data

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

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

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

`read.csv(file.choose(), header=T)`

This reads a .CSV file, allowing you to select the file, the header is set explicitly. If you change to `header=F` then the first row will be treated like the rest of the data and not as a label.

### Storing variable names

Through `read.csv()` or `read.table()` functions, data along with variable labels is read into R memory. However, to read the variables' names directly into R, one should use `attach(dataset)` function. For example,

```
>attach(HousePrice)
```

causes R to directly read all the variables' names eg. Price, Floor, Area etc. it is a good practice to use the `attach(datafile)` function immediately after reading the `datafile` into R.

### Packages

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

```
> library()
```

To load a particular package, use a command like

```
> library(forecast)
```

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

### Standard packages

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

### Contributed packages and CRAN

There are a number of contributed packages for R, written by many authors. Various packages deal with various analyses. Most of the packages are available for download from CRAN (<https://cran.r-project.org/web/packages/>), and other repositories such as

Bioconductor (<http://www.bioconductor.org/>). The collection of available packages changes frequently. As on June07, 2019, the CRAN package repository contains 14346 available packages.

### Getting Help

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

```
> help(mean)
```

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

```
>?mean
```

will open the same window again.

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

```
> Help(mean)
```

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

### Further Readings

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

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

### RStudio

RStudio is an integrated development environment (IDE) that allows to interact with R more readily. RStudio is similar to the standard RGui, but is considerably more user friendly. It has more drop-down menus, windows with multiple tabs, and many customization options.

### Installation of RStudio

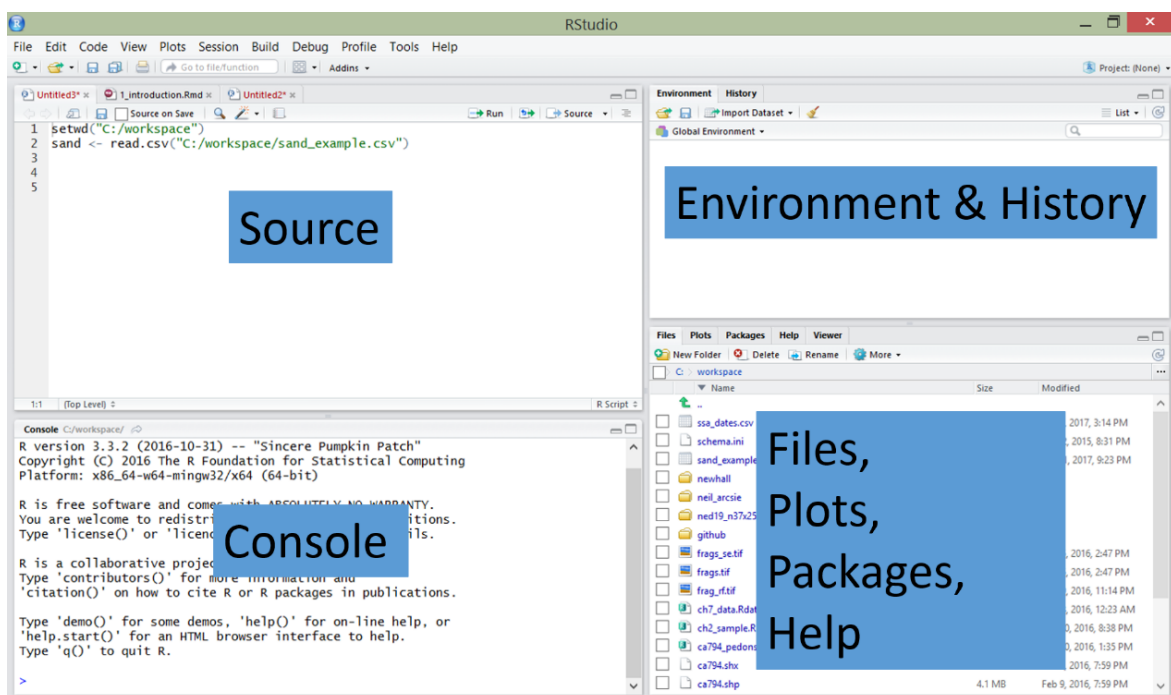
## Overview on R Software and RStudio

RStudio requires R 3.0.1+ that means R software should be pre-installed before using RStudio.

RStudio 1.2 requires a 64-bit operating system, and works exclusively with the 64 bit version of R. If you are on a 32 bit system or need the 32 bit version of R, you can use an older version of RStudio (<https://support.rstudio.com/hc/en-us/articles/206569407-Older-Versions-of-RStudio>).

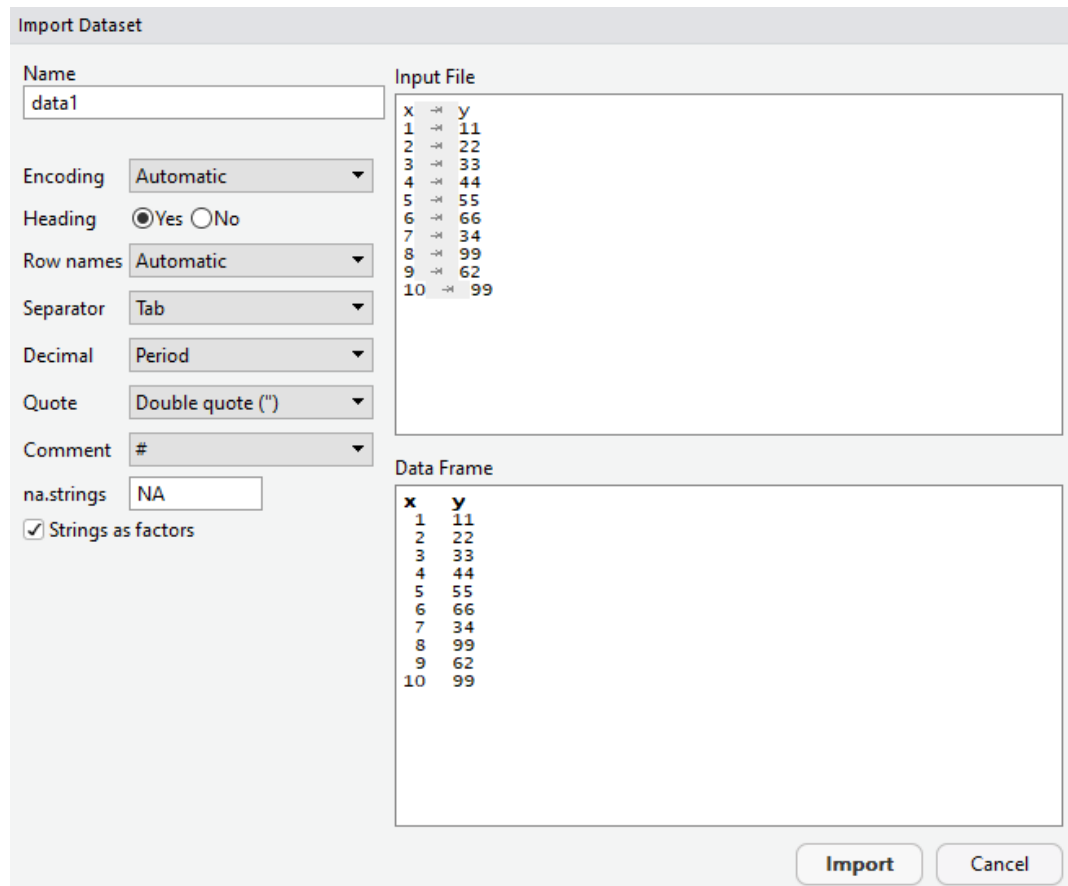
RStudio free desktop version can be downloaded from the following link:  
<https://www.rstudio.com/products/rstudio/download/#download>

The first time RStudio is opened, three windows are seen. A fourth window is hidden by default, but can be opened by clicking the **File** drop-down menu, then **New File**, and then **R Script**.



### Importing Data in R Studio

1. Click on the import dataset button in the top-right section under the environment tab. Select the file you want to import and then click open. The Import Dataset dialog will appear as shown below

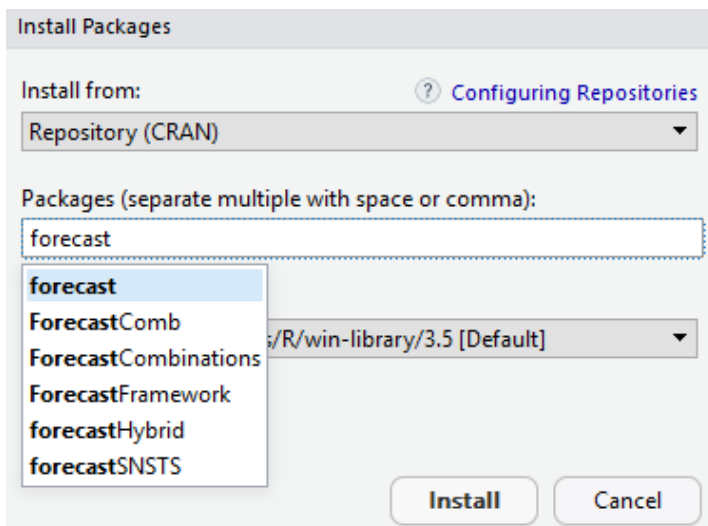
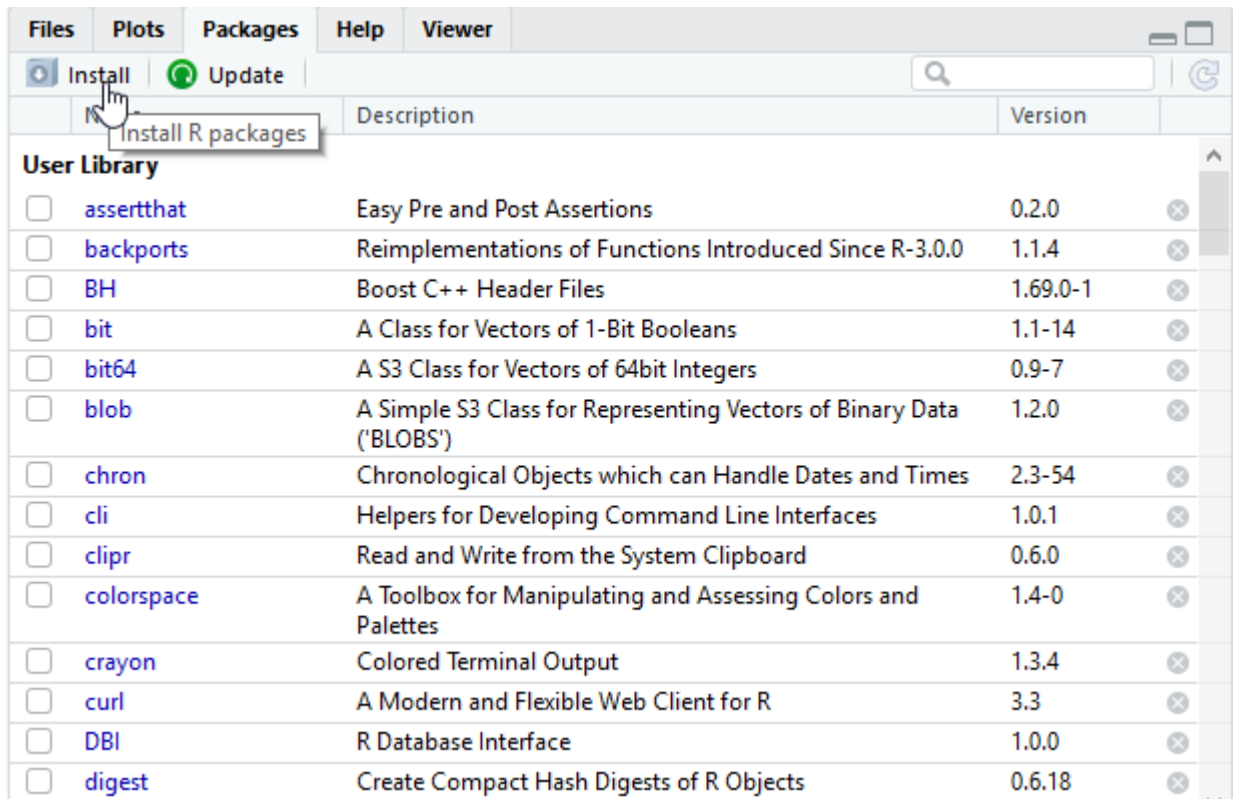


2. After setting up the preferences of separator, name and other parameters, click on the Import button. The dataset will be imported in R Studio and assigned to the variable name as set before.

## Installing Packages in RStudio

Within the **Packages** tab, a list of all the packages currently installed on the working computer and 2 buttons labeled either "Install" or "Update" are seen. To install a new package simply select the Install button. It is possible to install one or more than one packages at a time by simply separating them with a comma.





### Loading Packages in RStudio

Once a package is installed, it must be loaded into the R session to be used.

Name	Description	Version
Theory Group (Formerly: E1071), TU Wien		
<input type="checkbox"/> ellipsis	Tools for Working with ...	0.2.0.1
<input type="checkbox"/> fansi	ANSI Control Sequence Aware String Functions	0.4.0
<input type="checkbox"/> forcats	Tools for Working with Categorical Variables (Factors)	0.4.0
<input checked="" type="checkbox"/> forecast	Forecasting Functions for Time Series and Linear Models	8.5
<input type="checkbox"/> fracdiff	Fractionally differenced ARIMA aka ARFIMA(p,d,q) models	1.4-2
<input type="checkbox"/> ggplot2	Create Elegant Data Visualisations Using the Grammar of Graphics	3.1.0

## Writing Scripts in RStudio

RStudio's Source Tabs serve as a built-in text editor. Prior to executing R functions at the Console, commands are typically written down (or scripted). To write a script, simply open a new R script file by clicking File>New File>R Script. Within the text editor type out a sequence of functions.

- Place each function (e.g. read.csv()) on a separate line.
- If a function has a long list of arguments, place each argument on a separate line.
- A command can be executed from the text editor by placing the cursor on a line and typing Ctrl + Enter, or by clicking the Run button.
- An entire R script file can be executed by clicking the Source button.

The screenshot shows the RStudio Source Editor with the following R code:

```

1 # header is set explicitly
2 # If header=F then the first row will be treated like the rest of the data and not as
3 read.csv(file.choose(), header=T)
4
5 # to read the variables names directly into R, use attach(dataset) function.
6 attach(HousePrice)
7 #=====
8 # writing a function in R
9 #=====
10 avg=function(x)
11 {
12   sumx=0
13   for (i in 1:length(x))
14     sumx=sumx+x[i]
15   average=sumx/length(x)
16   return(average)
17 }

```

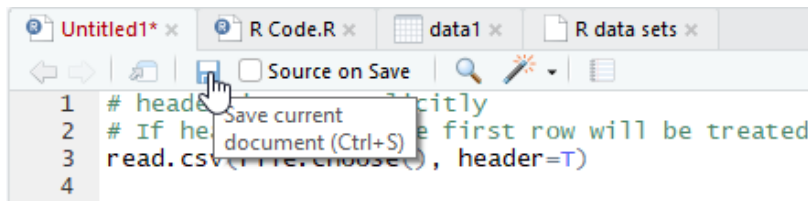
A tooltip is visible over the Run button (a blue arrow icon) in the toolbar, containing the text: "Run the current line or selection (Ctrl+Enter)".

## Saving R files in RStudio

In R, several types of files can be saved to keep track of the work performed. The file types include: script, workspace, history and graphics.

### ***R script (.R)***

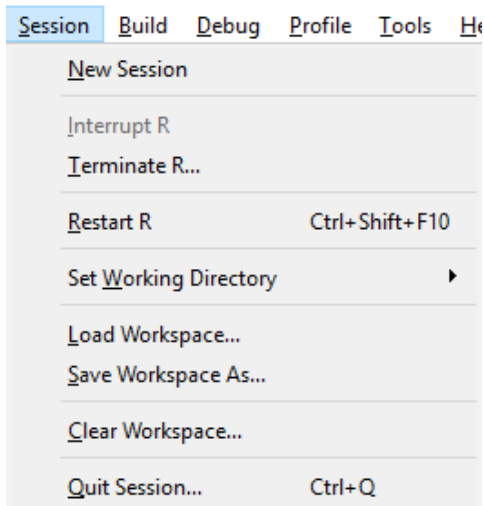
An R script is a text file of R commands that have been typed. To save R scripts in RStudio, click the save button from R script tab. Save scripts with the .R extension.



To open an R script, click the file icon.

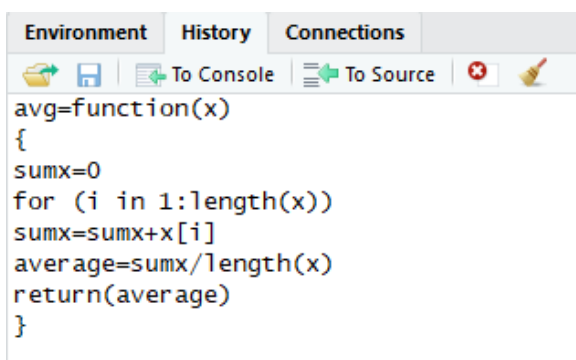
### ***Workspace (.Rdata)***

The R workspace consists of all the data objects created or loaded during the R session. It is possible to save or load the workspace at any time during the R session from the menu by clicking Session > Save Workspace As..., or the save button on the Environment Tab.



### ***R history (.Rhistory)***

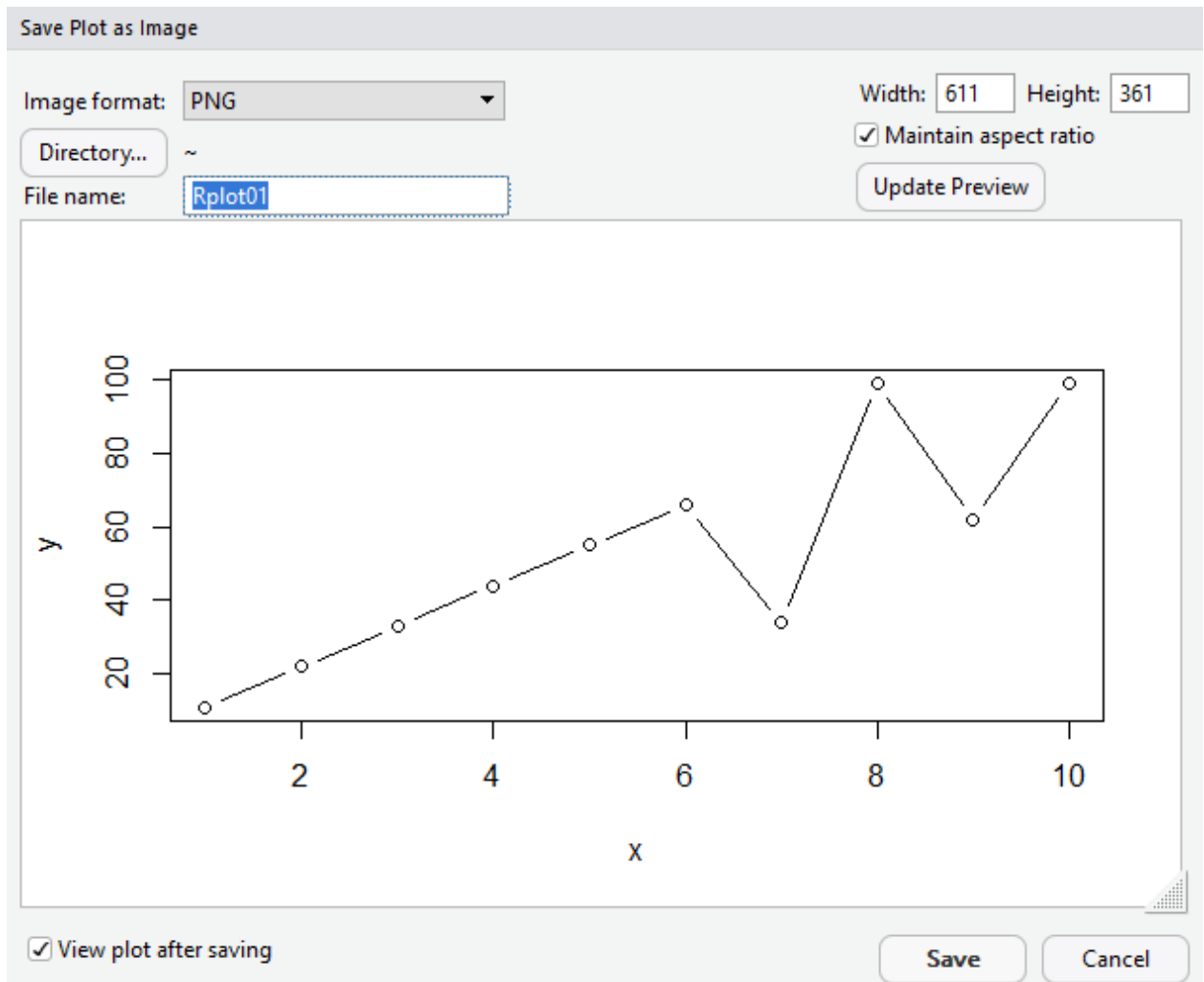
Rhistory file is a text file that lists all of the commands that have been executed. It does not keep a record of the results. To load or save R history from the History Tab click the **Open File** or **Save** button.



## R Graphics

Graphic outputs can be saved in various formats like pdf, png, jpeg, bmp etc.

To save a graphic: (1) Click the **Plots** Tab window, (2) click the **Export** button, (3) **Choose** desired format, (4) **Modify** the export settings as desired and (4) click **Save**.



## References

1. [http://ncss-tech.github.io/stats\\_for\\_soil\\_survey/chapters/1\\_introduction/1\\_introduction.html](http://ncss-tech.github.io/stats_for_soil_survey/chapters/1_introduction/1_introduction.html)
2. <http://web.cs.ucla.edu/~gulzar/rstudio/basic-tutorial.html>
3. <http://www.gardenersown.co.uk/Education/Lectures/R/index.htm>
4. <https://www.cran.r-project.org>
5. <https://www.rstudio.com/>
6. Matloff, N. (2011). *The art of R programming: A tour of statistical software design*. No Starch Press.
7. Venables, W. N., Smith, D. M. and R Development Core Team (2009). An introduction to R: Notes on R: A programming Environment for Data Analysis and Graphics, version 1.7. 1.

---

---

# REGRESSION ANALYSIS

---

---

**Ranjit Kumar Paul**

*ICAR-Indian Agricultural Statistics Research Institute*

*Library Avenue, New Delhi - 110 012*

[ranjit.paul@icar.gov.in](mailto:ranjit.paul@icar.gov.in)

---

---

## 1 Introduction

Regression analysis is a statistical methodology that utilizes the relation between two or more quantitative variables so that one variable can be predicted from the other, or others. This methodology is widely used in business, the social and behavioral sciences, the biological sciences including agriculture and fishery research. For example, fish weight at harvest can be predicted by utilizing the relationship between fish weights and other growth affecting factors like water temperature, dissolved oxygen, free carbon dioxide etc. There are other situations in fishery where relationship among variables can be exploited through regression analysis.

The use of multiple regression model often depends on the estimates of the individual regression coefficients. Some examples of inferences that are frequently made include

1. Identifying the relative effects of the regressor variables,
2. Prediction and/or estimation, and
3. Selection of an appropriate set of variables for the model.

A functional relation between two variables is expressed by a mathematical formula. If  $X$  denotes the independent variable and  $Y$  the dependent variable, a functional relation is of the form

$$Y = f(X)$$

Given a particular value of  $X$ , the function  $f$  indicates the corresponding value of  $Y$ . A statistical relation, unlike a function is not a perfect one. In general, the observations for a statistical relation do not fall directly on the curve of relationship.

A regression model that involves more than one regressor variables is called a multiple regression model. In other words, it is a linear relationship between a dependent variable and a group of independent variables. Multiple regression fits a model to predict a dependent ( $Y$ ) variable from two or more independent ( $X$ ) variables. Multiple linear regression models are often used as approximating functions. That is, true functional relationship between  $y$  and  $x_1, x_2, \dots, x_k$  is unknown, but over certain ranges of the regressor variables the linear regression model is an adequate approximation to the true unknown function. If the model fits the data well, the overall  $R^2$  value will be high, and the corresponding  $P$  value will be low ( $P$  value is the observed significance level at which the null hypothesis is rejected). In addition to the overall  $P$  value, multiple regressions also report an individual  $P$  value for each independent variable. A low  $P$  value here means that this particular independent variable significantly improves the fit of the model. It is calculated by comparing the goodness-of-fit of the entire model to the goodness-of-fit when that independent variable is omitted. If the fit is much worse when that variable is

omitted from the model, the P value will be low, telling that the variable has a significant impact on the model.

Depending on the nature of the relationships between  $X$  and  $Y$ , regression approach may be classified into two broad categories viz., linear regression models and nonlinear regression models. The response variable is generally related to other causal variables through some parameters. The models that are linear in these parameters are known as linear models, whereas in nonlinear models parameters appear nonlinearly. Linear models are generally satisfactory approximations for most regression applications. There are occasions, however, when an empirically indicated or a theoretically justified nonlinear model is more appropriate. In the present lecture we shall consider fitting of linear models only.

### 2. Linear Regression Models

We consider a basic linear model where there is only one predictor variable and the regression function is linear. Model with more than one predictor variable is straight forward. The model can be stated as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (1)$$

Where  $Y_i$  is the value of the response variable in the  $i^{\text{th}}$  trial  $\beta_0$  and  $\beta_1$  are parameters,  $X_i$  is a known constant, namely, the value of the predictor variable in the  $i^{\text{th}}$  trial,  $\varepsilon_i$  is a random error term with mean zero and variance  $\sigma^2$  and  $\varepsilon_i$  and  $\varepsilon_j$  are uncorrelated so that their covariance is zero.

Regression model (1) is said to be simple, linear in the parameters, and linear in the predictor variable. It is “simple” in that there is only one predictor variable, “linear in the parameters” because no parameters appears as an exponent or its multiplied or divided by another parameter, and “linear in predictor variable” because this variable appears only in the first power. A model that is linear in the parameters and in the predictor variable is also called first order model.

#### 2.1 Regression Parameters

The parameters  $\beta_0$  and  $\beta_1$  in regression model (1) are called regression coefficients,  $\beta_1$  is the slope of the regression line. It indicates the change in the mean of the probability distribution of  $Y$  per unit increase in  $X$ . The parameter  $\beta_0$  in  $Y$  intercept of the regression line. When the scope of the model includes  $X = 0$ ,  $\beta_0$  gives the mean of the probability distribution of  $Y$  at  $X = 0$ . When the scope of the model does not cover  $X = 0$ ,  $\beta_0$  does not have any particular meaning as a separate term in the regression model.

#### 2.2 Method of Ordinary Least Squares

To find “good” estimates of the regression parameters  $\beta_0$  and  $\beta_1$ , we employ the method of least squares. For each observations  $(X_i, Y_i)$  for each case, the method of least squares considers the deviation of  $Y$  from its expected value,  $Y_i - \beta_0 - \beta_1 X_i$ . In particular, the method of least squares requires that we consider the sum of the  $n$  squared deviations. This criterion is denoted by  $Q$ :

$$Q = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2 \quad (2)$$

## Regression Analysis

According to the method of least squares, the estimators of  $\beta_0$  and  $\beta_1$  are those values  $b_0$  and  $b_1$ , respectively, that minimize the criterion  $Q$  for the given observations.

Using the analytical approach, it can be shown for regression model (1) that the values of  $b_0$  and  $b_1$  that minimizes  $Q$  for any particular set of sample data are given by the following simultaneous equations:

$$\begin{aligned}\sum_{i=1}^n Y_i &= nb_0 + b_1 \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i Y_i &= b_0 \sum_{i=1}^n X_i + b_1 \sum_{i=1}^n X_i^2.\end{aligned}$$

These two equations are called normal equations and can be solved for  $b_0$  and  $b_1$ :

$$\begin{aligned}b_1 &= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ b_0 &= \frac{1}{n} \left( \sum_{i=1}^n Y_i - b_1 \sum_{i=1}^n X_i \right) = \bar{Y} - b_1 \bar{X},\end{aligned}$$

where  $\bar{X}$  and  $\bar{Y}$  are the means of the  $X_i$  and the  $Y_i$  observations, respectively.

### 2.3 Properties of Fitted Regression Line

Once the parameters estimates are obtained, the fitted line would be

$$\hat{Y}_i = b_0 + b_1 X_i \quad (3)$$

The  $i$ th residual is the difference between the observed value  $Y_i$  and the corresponding fitted value  $\hat{Y}_i$ , i.e.,  $e_i = Y_i - \hat{Y}_i$ .

The estimated regression line (3) fitted by the method of least squares has a number of properties worth noting.

1. The sum of the residuals is zero,  $\sum_{i=1}^n e_i = 0$ .
2. Sum of the squared residuals,  $\sum_{i=1}^n e_i^2$  is a minimum.
3. Sum of the observed values  $Y_i$  equals the sum of the fitted values  $\hat{Y}_i$ ,  $\sum_{i=1}^n Y_i = \sum_{i=1}^n \hat{Y}_i$ .
4. Sum of the weighted residuals is zero, weighted by the level of the predictor variable in the  $i$ th trial:  $\sum_{i=1}^n X_i e_i = 0$ .
5. Sum of the weighted residuals is zero, weighted by the fitted value of the response variable in the  $i$ th trial:  $\sum_{i=1}^n \hat{Y}_i e_i = 0$ .

6. The regression line always goes through the points  $(\bar{X}, \bar{Y})$ .

## 2.4 Estimation of Error Term Variance $\sigma^2$

The variance  $\sigma^2$  of the error terms  $\varepsilon_i$  in regression model (1) needs to be estimated to obtain an indication of the variability of the probability distribution of  $Y$ . In addition, a variety of inferences concerning the regression function and the prediction of  $Y$  require an estimate of  $\sigma^2$ .

Denote by  $SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n e_i^2$ , is the error sum of squares or residual sum of squares.

Then an estimate of  $\sigma^2$  is given by,

$$\hat{\sigma}^2 = \frac{SSE}{n-p}, \quad (4)$$

where  $p$  is the total number of parameters involved in the model. We also denote this quantity by MSE.

## 2.5 Inferences in Linear Models

Frequently, we are interested in drawing inferences about  $\beta_1$ , the slope of the regression line. At times, tests concerning  $\beta_1$  are of interest, particularly one of the form:

$$H_0 = \beta_1 = 0$$

$$H_1 = \beta_1 \neq 0$$

The reason for interest in testing whether or not  $\beta_1 = 0$  is that, when  $\beta_1 = 0$ , there is no linear association between  $Y$  and  $X$ . For normal error regression model, the condition  $\beta_1 = 0$  implies even more than no linear association between  $Y$  and  $X$ .  $\beta_1 = 0$  for the normal error regression model implies not only that there is no linear association between  $Y$  and  $X$  but also that there is no relation of any kind between  $Y$  and  $X$ , since the probability distribution of  $Y$  are then identical at all levels of  $X$ .

An explicit test of the alternatives is based on the test statistic:

$$t = \frac{b_1}{s(b_1)},$$

where  $s(b_1)$  is the standard error of  $b_1$  and calculated as  $s(b_1) = \sqrt{\frac{MSE}{\sum_{i=1}^n (X_i - \bar{X})^2}}$ .

The decision rule with this test statistic when controlling level of significance at  $\alpha$  is

if  $|t| \leq t(1 - \alpha/2; n - p)$ , conclude  $H_0$ ,

if  $|t| > t(1 - \alpha/2; n - p)$ , conclude  $H_1$ .

Similarly testing for other parameters can be carried out.

## 2.6 Prediction of New Observations



The new observation on  $Y$  to be predicted is viewed as the result of a new trial, independent of the trials on which the regression analysis is based. We denote the level of  $X$  for the new trial as  $X_h$  and the new observation on  $Y$  as  $Y_h$ . Of course, we assume that the underlying regression model applicable for the basic sample data continues to be appropriate for the new observation.

The distinction between estimation of the mean response, and prediction of a new response, is basic. In the former case, we estimate the mean of the distribution of  $Y$ . In the present case, we predict an individual outcome drawn from the distribution of  $Y$ . Of course, the great majority of individual outcomes deviate from the mean response, and this must be taken into account by the procedure for predicting  $Y_{h(\text{new})}$ . We denote by  $\hat{Y}_h$ , the predicted new observation and by  $\sigma^2(\hat{Y}_h)$  the variance of  $\hat{Y}_h$ . An unbiased estimator of  $\sigma^2(\hat{Y}_h)$  is given by  $\hat{\sigma}^2(\hat{Y}_h) = \hat{\sigma}^2 + s^2(\hat{Y}_h)$ , where  $s^2(\hat{Y}_h)$  is the estimate of variance of prediction at  $X_h$  and given by

$$s^2(\hat{Y}_h) = \hat{\sigma}^2 \left( \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right). \quad (5)$$

Confidence interval of  $\hat{Y}_h$  can be constructed by using t-statistic namely,

$$\hat{Y}_h \pm t(1 - \alpha/2; n - p) \sigma^2(\hat{Y}_h).$$

## 2.7 Measure of Fitting, $R^2$

There are times when the degree of linear association is of interest in its right. Here we describe one descriptive measure that is frequently used in practice to describe the degree of linear association between  $Y$  and  $X$ .

Denote by  $SSTO = \sum_{i=1}^n (Y_i - \bar{Y})^2$ , total sum of squares which measures the variation in the

observation  $Y_i$ , or the uncertainty in predicting  $Y$ , when no account of the predictor variable  $X$  is taken. Thus  $SSTO$  is a measure of uncertainty in predicting  $Y$  when  $X$  is not considered. Similarly,  $SSE$  measures the variation in the  $Y_i$  when a regression model utilizing the predictor variable  $X$  is employed. A natural measure of the effect of  $X$  in reducing the variation in  $Y$ , i.e., in reducing the uncertainty in predicting  $Y$ , is to express the reduction in variation ( $SSTO - SSE = SSR$ ) as a proportion of the total variation:

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} \quad (6)$$

The measure  $R^2$  is called coefficient of determination,  $0 \leq R^2 \leq 1$ . In practice  $R^2$  is not likely to be 0 or 1 but somewhere between these limits. The closer it is to 1, the greater is said to be the degree of linear association between  $X$  and  $Y$ .

## 2.8 Diagnostics and Remedial Measures

When a regression model is considered for an application, we can usually not be certain in advance that the model is appropriate for that application, any one, or several, of the features of the model, such as linearity of the regression function or normality of the error terms, may not be appropriate for the particular data at hand. Hence, it is important to examine the aptness of the

## Regression Analysis

model for the data before inferences based on that model are undertaken. We should consider following six important types of departures from linear regression model with normal errors:

- (i) The linearity of regression function.
- (ii) The constancy of error variance.
- (iii) The independency of error terms.
- (iv) Presence of one or a few outlier observations.
- (v) The normal distribution of error terms.
- (vi) One or several important predictor variables have been omitted from the model.
- (vii) Presence of Multicollinearity.

### **Practical on Regression Analysis using R:**

#### **Example dataset**

**Example:** An experiment was conducted to study the hybrid seed production of bottle gourd (*Lagenaria siceraria (Mol) Standl*) Cv. Pusa hybrid-3 under open field conditions during Kharif-2005 at Indian Agricultural Research Institute, New Delhi. The main aim of the investigation was to compare natural pollination and hand pollination under field conditions. The data were collected on 10 randomly selected plants from each of natural pollination and hand pollination. The data were collected on number of fruit set (NFS) for the period of 45 days, fruit weight (FW) (kg), seed yield per plant (SY)(g) and seedling length (SL) (cm). The data obtained is as given below: {Here 1 denotes natural pollination and 2 denotes the hand pollination }

Group	NFS	FW	SY	SL
1	7.0	1.85	147.70	16.86
1	7.0	1.86	136.86	16.77
1	6.0	1.83	149.97	16.35
1	7.0	1.89	172.33	18.26
1	7.0	1.80	144.46	17.90
1	6.0	1.88	138.30	16.95
1	7.0	1.89	150.58	18.15
1	7.0	1.79	140.99	18.86
1	6.0	1.85	140.57	18.39
1	7.0	1.84	138.33	18.58
2	6.3	2.58	224.26	18.18
2	6.7	2.74	197.50	18.07

## Regression Analysis

2	7.3	2.58	230.34	19.07
2	8.0	2.62	217.05	19.00
2	8.0	2.68	233.84	18.00
2	8.0	2.56	216.52	18.49
2	7.7	2.34	211.93	17.45
2	7.7	2.67	210.37	18.97
2	7.0	2.45	199.87	19.31
2	7.3	2.44	214.30	19.36

```
#importing the csv file
```

```
FruitData <-read.csv(file.choose(),header=TRUE)
```

```
attach(FruitData)
```

```
#finding correlation coefficient matrix among the variables in a dataset
```

```
cor(FruitData)
```

```
#estimating and testing correlation coefficient between any two variables say between SY and SL
```

```
> cor(FruitData)
```

```
      Group      NFS      FW      SY      SL
Group 1.0000000 0.5644427 0.9723643 0.9567793 0.5139970
NFS    0.5644427 1.0000000 0.5411999 0.5998586 0.4250208
FW     0.9723643 0.5411999 1.0000000 0.9411130 0.4887463
SY     0.9567793 0.5998586 0.9411130 1.0000000 0.4921407
SL     0.5139970 0.4250208 0.4887463 0.4921407 1.0000000
```

```
cor.test(SY, SL)
```

```
> cor.test(SY, SL)
```

```
      Pearson's product-moment correlation
```

```
data: SY and SL
```

```
t = 2.3986, df = 18, p-value = 0.02751
```

```
alternative hypothesis: true correlation is not equal to 0
```

```
95 percent confidence interval:
```

```
 0.06343489 0.76751101
```

```
sample estimates:
```

```
      cor
```

```
0.4921407
```

## Regression Analysis

#Finding Spearman's Rank correlation coefficient between any variables

```
cor.test(SY, SL, method="s")
```

#partial correlating coefficient

```
library(ppcor)
```

```
ppcor(FruitData, method = c("pearson"))
```

```
> ppcor(FruitData, method = c("pearson"))
```

```
$estimate
```

	Group	NFS	FW	SY	SL
Group	1.000000000	0.002229428	0.72832273	0.50105532	0.15380435
NFS	0.002229428	1.000000000	-0.07335929	0.26169618	0.19266422
FW	0.728322734	-0.073359293	1.00000000	0.17226110	-0.04076243
SY	0.501055324	0.261696183	0.17226110	1.00000000	-0.04062804
SL	0.153804353	0.192664221	-0.04076243	-0.04062804	1.00000000

```
$p.value
```

	Group	NFS	FW	SY	SL
Group	0.000000000	0.9932245	0.000915098	0.04047686	0.5556110
NFS	0.993224497	0.0000000	0.779626444	0.31027830	0.4587834
FW	0.000915098	0.7796264	0.000000000	0.50853463	0.8765615
SY	0.040476860	0.3102783	0.508534632	0.00000000	0.8769655
SL	0.555611009	0.4587834	0.876561451	0.87696551	0.0000000

#regression model of SY on SL

```
reg<-lm(SY ~SL)
```

```
summary(reg)
```

```
> reg<-lm(SY ~SL)
```

```
> summary(reg)
```

```
Call:
```

```
lm(formula = SY ~ SL)
```

```
Residuals:
```

```
    Min       1Q   Median       3Q      Max
```

-54.638	-20.712	0.753	21.029	56.131
---------	---------	-------	--------	--------

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-197.342	157.832	-1.250	0.2272
SL	20.836	8.687	2.399	0.0275 *

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 33.37 on 18 degrees of freedom
```

```
Multiple R-squared:  0.2422,    Adjusted R-squared:  0.2001
```

```
F-statistic: 5.753 on 1 and 18 DF,  p-value: 0.02751
```

#regression model of SY on SL, FW and NFS

```
reg1<-lm(SY ~SL+FW+NFS)
```

## Regression Analysis

```
summary(reg1)
```

```
> reg1<-lm(SY ~SL+FW+NFS)
> summary(reg1)

Call:
lm(formula = SY ~ SL + FW + NFS)

Residuals:
    Min       1Q   Median       3Q      Max
-26.160  -6.226  -1.820   10.397   18.854

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -71.2001    65.0731  -1.094   0.290
SL             0.6792     3.9812   0.171   0.867
FW            85.2960     9.9705   8.555 2.3e-07 ***
NFS             7.2949     5.7217   1.275   0.221
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.02 on 16 degrees of freedom
Multiple R-squared:  0.8975,    Adjusted R-squared:  0.8782
F-statistic: 46.68 on 3 and 16 DF,  p-value: 3.887e-08
```

```
#regression model of SY on SL without intercept
reg<-lm(SY ~-1+SL)
```

```
#prediction using regression model
lm.predict<-predict(reg ,interval="confidence")
```

```
lm.predict
```

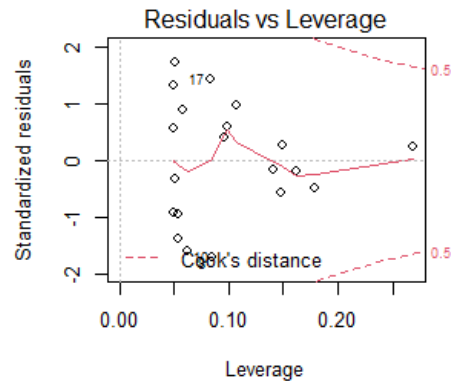
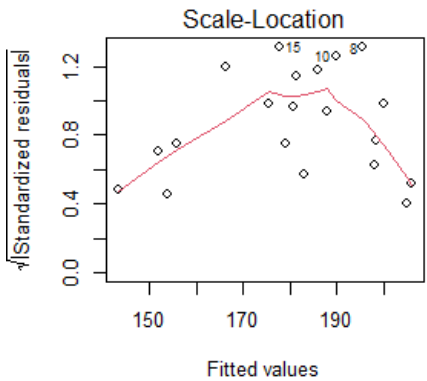
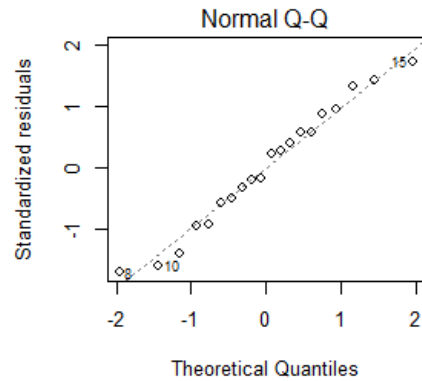
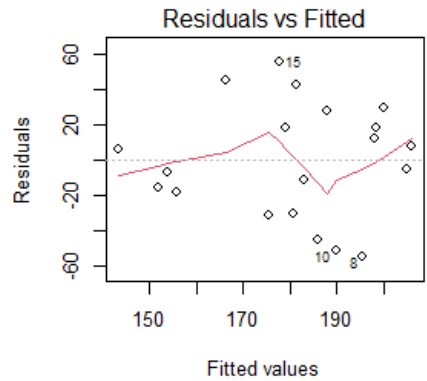
## Regression Analysis

```
> lm.predict<-predict(reg ,interval="confidence")
> lm.predict
```

	fit	lwr	upr
1	153.9561	125.6938	182.2183
2	152.0808	122.4379	181.7238
3	143.3296	106.9544	179.7049
4	183.1267	167.3187	198.9348
5	175.6257	159.3064	191.9451
6	155.8313	128.9204	182.7423
7	180.8348	165.1582	196.5113
8	195.6284	175.2723	215.9846
9	185.8354	169.5511	202.1198
10	189.7943	172.2509	207.3377
11	181.4598	165.7728	197.1469
12	179.1679	163.4260	194.9097
13	200.0040	177.0128	222.9953
14	198.5455	176.4716	220.6194
15	177.7093	161.8003	193.6184
16	187.9191	171.0490	204.7891
17	166.2494	146.0438	186.4550
18	197.9204	176.2285	219.6124
19	205.0047	178.6397	231.3697
20	206.0465	178.9424	233.1507

```
#Residual diagnostics of the fitted model
par(mfrow = c(2, 2))
plot(reg)
```

## Regression Analysis



```
#For plotting cook's distance  
plot(reg, 4)
```

```
#distribution fitting in r  
library(MASS)  
fitdistr(SY,"normal")
```

```
library(vcd)## loading vcd package  
gf<-goodfit(SY,type="normal",method="MinChisq")  
summary(gf)
```

```
library(olsrr)  
# Fit the full model  
full.model <- lm(SY ~ SL+FW+NFS, data = FruitData)
```

```
# Fit the model with stepwise selection procedure  
ols_step_both_p(full.model)
```

## Regression Analysis

```
> ols_step_both_p(full.model)
```

### Stepwise Selection Summary

Step	variable	Added/ Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	FW	addition	0.886	0.879	1.8380	163.1258	12.9602

```
# Fitting of robust regression model
```

```
library(MASS)
```

```
reg1<-rlm(SY~ SL+FW+NFS)
```

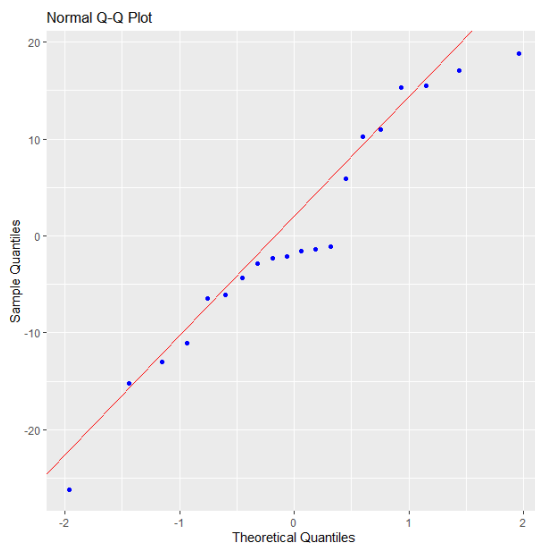
```
summary(reg1)
```

```
library(car)
```

```
vif(full.model)
```

```
# Residual QQ Plot
```

```
ols_plot_resid_qq(full.model)
```



```
# Residual Normality Test
```

```
ols_test_normality(full.model)
```

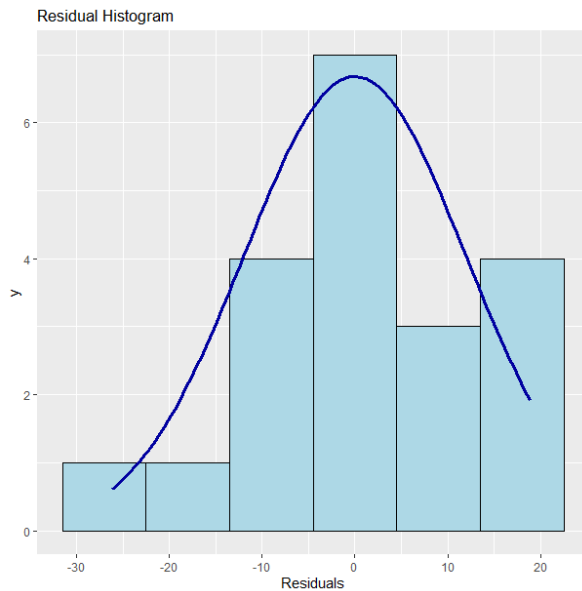


```
> ols_test_normality(full.model)
```

Test	statistic	pvalue
shapiro-wilk	0.9529	0.4125
kolmogorov-smirnov	0.1865	0.4370
Cramer-von Mises	2.1167	0.0000
Anderson-Darling	0.4274	0.2825

```
# Residual vs Fitted Values Plot
ols_plot_resid_fit(full.model)
```

```
# Residual Histogram
ols_plot_resid_hist(full.model)
```



**Some Selected References**

Belsley, D.A., Kuh, E. and Welsch, R.E. (2004). Regression diagnostics – Identifying influential data and sources of collinearity, New York.: Wiley

Barnett, V. and Lewis, T. (1984). Outliers in Statistical Data, New York: Wiley Ltd.

Chatterjee, S. and Price, B (1977). Regression analysis by example, New York: John Wiley & sons

Draper, N.R. and Smith, H. (1998). Applied Regression analysis, New York: Wiley Eastern Ltd.

Kleinbaum, D.G. & Kupper, L.L. (1978). Applied Regression analysis and other multivariate methods, Massachusetts: Duxbury Press

Montgomery, D.C., Peck, E. and Vining, G. (2003). Introduction to linear regression analysis, 3rd Edition, New York: John Wiley and Sons Inc.

---

---

## TESTING OF HYPOTHESIS

---

---

**Prakash Kumar and Ranjit Kumar Paul**  
*ICAR-Indian Agricultural Statistics Research Institute*  
*Library Avenue, New Delhi - 110 012*  
[prakash.kumar@icar.gov.in](mailto:prakash.kumar@icar.gov.in) [ranjit.paul@icar.gov.in](mailto:ranjit.paul@icar.gov.in)

---

---

**Purpose:** A hypothesis test allows us to draw conclusions or make decisions regarding population from sample data. In the following cases, we are making decisions regarding the population mean (or means).

**Hypothesis Testing** – a) the standard approach to assessing whether an observed value of a variable or an observed relationship between two or more variables derived from sample data is “real,” that is holds true in the population or is a result of mere chance. b) is an inferential statistics approach, that allows the researcher to use characteristics derived from sample data to make inferences about population characteristics. c) involves comparing empirically observed findings with theoretical expected findings. d) estimates the statistical significance of findings. e) involves posing opposing hypotheses about a population characteristic or the relationship between two or more population characteristics and then testing those Directional Hypothesis

Using the directional hypothesis, the direction of the hypothesis can be specified like, if the user wants to know the sample mean is lower or greater than another mean sample of the data. hypotheses. f) asks how often the observed results could be expected to occur by chance, if the answer is relatively frequently, then chance would remain a viable explanation of the effect but if relatively rarely, then chance would not be a viable explanation.

**Null Hypothesis** – a) in hypothesis testing it is the claim or statement about a population parameter that is assumed to be valid or true unless the observed data contradicts this assumption. b) the hypothesis that two variables are not related or that two statistics (e.g. means or proportions) are the same. c) symbolized as  $H_0$ . d) hypothesis test can either reject the null hypothesis, in which case the alternative hypothesis may be true, or fail to reject the null hypothesis.

**Alternative Hypothesis** (research hypothesis) - a) in hypothesis testing it is the opposite claim or statement about a population parameter from the null hypothesis. b) the hypothesis that two variables are related or that two statistics (e.g. means or proportions) are different. c) the hypothesis that the researcher expects to be supported, although this perspective is controversial. d) symbolized as  $H_1$  or  $H_A$ .

**Type I Error** (alpha, false positive, false alarm) – a) falsely rejecting a true null hypothesis. b) mistakenly concluding that a difference exists in a population parameter when the sample difference was merely a result of chance. c) considered the more serious form of error and more important error to avoid. d) A court finding a person guilty of a crime that they did not actually commit.

**Type II Error** (beta) – a) falsely rejecting a true alternative hypothesis or falsely failing to reject a false null hypothesis. b) the inverse of type I error, so the greater the risk of committing one then the lower the risk of committing the other. c) a court finding a person innocent of a crime that they actually committed.

## Testing of Hypothesis

**Significance Level** (alpha level) – a) the probability of making a Type I Error. b) 0.05 is probably the most common significance level and corresponds to a situation in which Type I Error is committed only one time in 20. c) the inverse of the confidence level, so significance level of 0.05 corresponds to a 95% confidence level.

**Critical Value** – a) the points in a test statistic sampling distribution that define a statistically significant result that is a result unlikely to have occurred merely due to chance. b) the value of a test statistic that result in rejecting or failing to reject the null hypothesis, that is the zone of rejection. c) found in tables for a test statistic (like chi-square , t-test, and z-test) and not calculated from observed data.

**Critical Region** (Zone of Rejection) – a) the critical region of a hypothesis test is the set of all outcomes which, if they occur, will result in rejecting the null hypothesis.

**Degrees of Freedom** (*df*) – a) the number of values that are free to vary or the number of independent pieces of information when calculating a test statistic. b) the number of independent scores or observations used to calculate a test statistic minus the number of statistics estimated as intermediate steps in the estimation of the statistic itself. c) an important but difficult to understand concept in inferential statistics but fortunately one with straightforward practical applications and simple equations.

### One Sample Hypothesis Tests

Applied to determine if the population mean is consistent with a specified value or standard

Two tests

- z- test
- t-test

#### Assumptions: z-test

- The underlying distribution is normal or the Central Limit Theorem can be assumed to hold
- The sample has been randomly selected
- The population standard deviation is known or the sample size is at least 25.

#### Assumptions: the t- test

- The underlying distribution is normal or the Central Limit Theorem can be assumed to hold
- The sample has been randomly selected

### Two Samples Hypothesis Tests

Applied to compare the values of two population means.

Two tests

- z- test
- t-test

#### Assumptions: z-Test

- The underlying distribution is normal or the CLT can be assumed to hold
- The samples have been randomly and independently selected from two populations
- The population standard deviations are known or the sample size of each sample is at least 25.

#### Assumptions: t-Test

- The underlying distribution is normal or the CLT can be assumed to hold
- The samples have been randomly and independently selected from two populations,

## Testing of Hypothesis

- The variability of the measurements in the two populations is the same and can be measured by a common variance.

### The Logic of Hypothesis Tests

Assume a population distribution with a specified population mean.

State the hypothesized population mean (this statement is referred to as the null hypothesis).

This mean is stated as the null hypothesis and is designated  $H_0$ .

For example,  $\mu = 10$

- State the logical alternative to this hypothesis. This is called the alternate hypothesis and is designated  $H_a$ .

For example,  $\mu \neq 10$ .

(Note the alternate hypothesis can have other forms since the concept of not equal can imply  $\mu > 10$  or  $\mu < 10$ .)

- Draw a random sample from the population.
  - Calculate the sample mean. This sample mean represents one point on the distribution of sample means.
  - Determine the “relative position” of the calculated mean (sample mean) on the distribution of sample means.
- If the sample mean is “close” to the specified population mean, we do not have evidence to reject the hypothesized population mean.
  - If the calculated sample mean is “not close” to the specified population mean, we conclude that our sample could not have been drawn from the hypothesized distribution, and thus, we reject the null hypothesis.

### Example 1: One Sample Hypothesis Test

Large Sample: Sample size:  $n > 30$

1. The scores on an aptitude test required for entry into a certain job position have a mean of 500 and a standard deviation of 120. If a random sample of 36 applicants has a mean of 546, is there evidence that their mean score is different from the mean that is expected from all applicants?

Null and Alternative Hypothesis

$H_0: \mu = 500$

$H_1: \mu \neq 500$

Convert 546 to a z-score to compare it to the assumed population mean.

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{546 - 500}{\frac{120}{\sqrt{36}}} = \frac{46}{20} = 2.3$$

This means that 546 is 2.3 standard deviations from the hypothesized mean.

Using the z-table, we find that the probability that a value is to the right of 2.3 or to the left of -2.3 is  $2 * (.0107) = 0.0214$ . This value is called the p value  $p = 0.0214$ .

This probability is considered very small (values less than 0.05 are typically considered small). Thus, if the mean is really 500, it is unlikely that we would get a sample mean that is 2.3 standard deviations from it. Thus, we conclude that the population mean is not 500; that is we reject the null hypothesis and accept the alternate, concluding that the mean is not 500. The probability that we are rejecting a true null hypothesis is 0.0294 (the value of p).

## Testing of Hypothesis

Let's construct a 95% confidence interval estimate of the population mean.

$$546 \pm 1.96 * \left( \frac{120}{\sqrt{36}} \right) = 546 \pm 39.2$$

The lower limit of the interval is  $546 - 39.2 = 506.8$

The upper limit of the interval is  $546 + 39.2 = 585.2$

Thus, we conclude that the actual mean score for the population from which this sample was drawn falls between 507 and 585.

### 2. For sample size is small in problem 1.

Approach the problem the same way as in 1, using the t-distribution

$$t = \frac{\frac{\bar{x} - \mu}{s}}{\frac{1}{\sqrt{n}}} = \frac{546 - 500}{\frac{120}{\sqrt{16}}} = \frac{46}{30} = 1.3$$

The degrees of freedom is  $16 - 1 = 15$

Using the t-table with 15 degrees of freedom, we find the closest t-value to 1.53 is 1.753 and that the associated probability is 0.05.

Find is  $2 * (0.05) = 0.1$ . We then write the  $p$  value as  $p < 0.1$ .

### Paired Samples Test

The paired t-test is commonly used to compare a sample group's scores before and after an intervention. This test is used when the samples are dependent. A paired t-test is used to compare two population means where you have two samples in which observations in one sample can be paired with observations in the other sample.

- Find the difference in the paired values
- Treat the difference scores as one sample.
- Apply a one-sample test.

Example 3: Compare the difference in cost wheat per kg for 2016 vs. 2015.

<u>2016</u>	<u>2015</u>	<u>Difference</u>
18.36	18.41	-0.05
32.82	31.34	1.48
23.58	37.36	-13.78
17.52	16.58	0.94
19.12	21.35	-2.23
14.85	14.59	0.26
30.50	31.00	-0.50
25.06	26.21	-1.15
30.89	31.52	-0.63
35.74	35.21	0.53
19.33	19.55	-0.22
30.92	25.75	5.17
<u>34.30</u>	<u>33.91</u>	<u>0.39</u>

Do a one-sample test on the difference values.

$$t = -0.63$$

## Testing of Hypothesis

$$P = 0.54$$

Do not reject  $H_0$ ; we have no evidence to conclude that there has been a change in the cost for 2016 over 2015.

### Performing a Paired t-test in Excel:

To compare two paired values (such as in a before-after situation) where both observations are taken from the same or matched subjects, you can perform a paired t- For example, suppose your data contained the variables BEFORE and AFTER, (before and after weight on a diet), for 8 subjects. The hypotheses for this test are:

$H_0$ :  $mLoss = 0$  (The average weight loss was 0)

$H_a$ :  $mLoss \neq 0$  (The weight loss was different than 0)

For example, the following weight loss data is used in this example

#### (DIET.XLS)

Before	After
162	168
170	136
184	147
164	159
172	143
176	161
159	143
170	145

1. To perform a paired t-test, select **Tools/ Data Analysis / t-test: Paired two sample for means**.

2. In the **t-test: Paired two sample for means** dialog box: For the Input Range for Variable 1, highlight the 8 values of Score in group “Before” (values from 162 to 170). For the input range for Variable 2, highlight the eight values of Score in group “After” (values from 168 to 145). For now, leave the other items at their default selections. This dialog box is shown below. Click OK. This dialog box is shown below:

## Testing of Hypothesis

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Before	After												
2	162	168												
3	170	136												
4	184	147												
5	164	159												
6	172	143												
7	176	161												
8	159	143												
9	170	145												
10														
11														
12														
13														
14														
15														
16														

**t-Test: Paired Two Sample for Means**

Input

Variable 1 Range:

Variable 2 Range:

Hypothesized Mean Difference:

Labels

Alpha:

Output options

Output Range:

New Worksheet Ply:

New Workbook

3. The results are shown in the output below:

t-Test: Paired Two Sample for Means

	<i>Variable 1</i>	<i>Variable 2</i>
Mean	169.625	150.25
Variance	65.125	121.9286
Observations	8	8
Pearson Correlation	-0.17675	
Hypothesized Mean Difference	0	
df	7	
t Stat	3.706873	
P(T<=t) one-tail	0.003793	
t Critical one-tail	1.894579	
P(T<=t) two-tail	0.007586	
t Critical two-tail	2.364624	

Thus, the two-tail p-value for this t-test is  $p=0.008$  (.007585988) and  $t=3.71$ .

Excel actually does a poor job providing what you need to report the results of this test – for a more complete understanding, you need to realize that the paired t-test is actually a test on the DIFFERENCE between the two values. Thus, to make this a better analysis, first calculate the difference between BEFORE and AFTER, creating the following new column called “DIFF” using a formula such as =A2-B2 in cell C2 and copying the formula for the appropriate remaining cells in the worksheet. Notice also that the average difference is calculated (19.38).

Before	After	Diff
162	168	-6
170	136	34
184	147	37
164	159	5
172	143	29
176	161	15
159	143	16

## Testing of Hypothesis

170      145      25  
Average Diff=              19.375

Look back up at the original hypotheses – what you are testing is that the average loss is different than zero (0). Thus, the t-test is actually testing to determine if the value 19.38 is sufficiently different from 0 to claim significance. Thus, the number you are interested in most is the average difference (loss) and not as much as the individual means of Before and After.

Therefore to report these results properly, you need the mean difference and standard deviation. You can get this by calculating descriptive statistics on the difference values. (**Tools/Data Analysis/ Descriptive Statistics**) – choose the Summary Statistics and 95% confidence interval options. The results in the following output:

<i>Column1</i>	
Mean	19.375
	5.22677
Standard Error	7
Median	20.5
Mode	#N/A
Standard Deviation	14.7835
	6
Sample Variance	218.553
	6
Kurtosis	-0.5242
	-
Skewness	0.57529
Range	43
Minimum	-6
Maximum	37
Sum	155
Count	8
Confidence Level(95.0%)	12.3593
	6

Notice that the mean divided by the standard error ( $19.375/5.227 = 3.71$ ) is same as the value of the “t Stat” in the previous table. Another piece of information that is usually reported is a 95% confidence interval. Using the Confidence Level (95%) value of 12.359 in the table, the confidence interval is the mean plus or minus this value. Thus, a 95% C.I. about Mean Difference is (7.01, 31.74).

It can be stated in more precise manner that the mean weight loss ( $M=19.38$ ,  $SD =14.784$ ,  $N=8$ ) was significantly greater than zero,  $t(7)=3.71$ , two-tail  $p = 0.008$ , providing evidence that the diet is effective in producing weight loss. A 95% C.I. about mean weight loss is (7.01, 31.74).



## Testing of Hypothesis

**NOTE:** The test could have also been performed as a one-tail test. If so, use the appropriate t-statistic and p-value from the Excel table.

**NOTE:** Also, you could do this test using a hypothesized value of the difference other than zero – although zero is almost always used. Excel provides the opportunity to enter another hypothesized value to test in the paired t-test dialog box.

### Hands-On Hypothesis Testing Programming with R

#### One Sample T-Testing

One sample T-Testing approach collects a huge amount of data and tests it on random samples. To perform T-Test in R, normally distributed data is required. This test is used to test the mean of the sample with the population. For example, the height of persons living in an area is different or identical to other persons living in other areas.

```
# Defining sample observations
x <- rnorm(100)
# One Sample T-Test
t.test(x, mu = 5)
# you can understand the parameter using in t-test
help("t.test")
# hypothesis test with an example
x = c(38.7, 39.6, 37.9, 40.6, 40.5, 37.7, 41.2, 37.5, 39.1)
t.test(x,mu=40, alternative="less")
```

#### Two Sample T-Testing

In two sample T-Testing, the sample vectors are compared. If var.equal = TRUE, the test assumes that the variances of both the samples are equal.

```
# Defining sample vector
x <- rnorm(100)
y <- rnorm(100)
# Two Sample T-Test
t.test(x, y)
ttest=t.test(x,y)
names(ttest)
ttest$statistic
```

### **Directional Hypothesis**

Using the directional hypothesis, the direction of the hypothesis can be specified like, if the user wants to know the sample mean is lower or greater than another mean sample of the data.

```
# Defining sample vector
x <- rnorm(100)
# Directional hypothesis testing
t.test(x, mu = 2, alternative = 'greater')
```

### **Non-parametric one sample Wilcoxon signed rank test**

This type of test is used when comparison has to be computed on one sample and the data is non-parametric. It is performed using `wilcox.test()` function in R programming.

```
# Define vector
x <- rnorm(100)
# one sample test
wilcox.test(x, exact = FALSE)
```

### **Non-parametric two sample Wilcoxon signed rank test**

This type of test is used when comparison has to be computed on two samples of data. The basic way of using `wilcox.test()` command is to specify the two samples you want to compare as separate vectors, as shown in the following command:

```
# Define vectors
x <- rnorm(100)
y <- rnorm(100)
# Two sample test
wilcox.test(x, y)
```

### **Correlation Test**

This test is used to compare the correlation of the two vectors provided in the function call or to test for the association between the paired samples.

```
#correlation Coefficient
count = c(9,25,15,2,14,25,24,47)
speed = c(2,3,5,9,14,24,29,34)
cor(count, speed)
```

## Testing of Hypothesis

```
cor(count, speed, method = 'spearman')  
# Using given dataset in R  
cor.test(count, speed)  
cor.test(count, speed, method = 'spearman', exact = F)
```

### **Test the equality of the variances of this two groups**

```
x = rnorm(50,0,1)
```

```
y = rnorm(50,-0.27,0.4)
```

```
var.test(x,y)
```

F test to compare two variances

As  $p.value < \alpha = 0.05$  ,we accept H 1 and the variance of the 2 groups are different .

### **Test of normality**

Test de Shapiro-Wilk (hypothèse nulle ="normalité") :

```
x = rnorm(100)
```

```
shapiro.test(x)
```

### **Distribution tests / Kolmogorov-Smirnov test**

#ks.test(x,"distribution") to test if x follows the distribution (ex : punif, pnorm ...)

```
x=runif(20) ;y=rnorm(20)
```

```
ks.test(x,y)
```

#ks.test(x,y) to test if x and y follow the same distributions

```
ks.test(x,"pnorm")
```

we accept,H 0 : x and y follow the same distributions if  $p.value > \alpha$

### **ANOVA**

Suppose we want to study the effect of the "sport" factor at 2 levels on the "weight" variable.

```
sport=rbinom(100,1,0.7)
```

```
facsport=as.factor(sport)
```

```
weight=round(rnorm(100,80,20),0)
```

```
anova(aov(weight~ facsport ))
```

## Testing of Hypothesis

### Analysis of Variance Table

Response : poids

Df Sum Sq Mean Sq F value Pr(>F)

facsport 1 4.00 4.000 0.0677 0.7986

Residuals 14 827.75 59.125

As p.value >  $\alpha = 0.05$  we accept  $H_0$

Decision It is concluded that sport has an effect on weight

### References:

- Balluerka, N., Gómez, J., & Hidalgo, D. (2005). The controversy over null hypothesis significance testing revisited. *Methodology: European Journal of Research Methods for the Behavioral & Social Sciences*, **1(2)**, 55-70.
- Cowles, M., & Davis, C. (1982). On the origins of the .05 level of statistical significance. *American Psychologist*, **37(5)**, 553-558.
- Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory & Psychology*, **5(1)**, 75-98.
- Hubbard, R., & Lindsay, R. M. (2008). Why p values are not a useful measure of evidence in statistical significance testing. *Theory & Psychology*, **18(1)**, 69-88.
- Pollard, P., & Richardson, J. T. (1987). On the probability of making Type I errors. *Psychological Bulletin*, **102(1)**, 159-163.
- Thompson, B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. *Journal of Experimental Education. Special Issue: Statistical significance testing in contemporary practice*, **61(4)**, 361-377.
- Thompson, B. (1997). Rejoinder: Editorial policies regarding statistical significance tests: Further comments. *Educational Researcher*, **26(5)**, 29-32.
- Vacha-Haase, T., & Thompson, B. (1998). Further comments on statistical significance tests. *Measurement & Evaluation in Counseling & Development*, **31(1)**, 63-67.

---

---

## NON-PARAMETRIC TESTS

---

---

**Himadri Shekhar Roy and Lalmohan Bhar**  
*ICAR-Indian Agricultural Statistics Research Institute*  
*Library Avenue, New Delhi - 110 012*  
[himadri.roy@icar.gov.in](mailto:himadri.roy@icar.gov.in)

---

---

**Introduction:** Statistical inference plays a major role in statistical investigation as it draws conclusions for a given hypothesis. In testing of hypothesis, the main concern is to test the assumed hypothesis about some feature(s) of one or more populations. Almost all large and small sample tests such as t, F and  $\chi^2$  are based on the assumptions that the parent population (from which the sample is drawn) has a specific distribution, such as normal distribution. The distributions are usually defined through some parameters. In case of nonparametric tests, no such assumptions are required and hence nonparametric tests are also known as distribution-free tests. The cost of fewer assumptions is that nonparametric tests are generally less powerful than their parametric counterparts (i.e., when the alternative is true, they may be less likely to reject  $H_0$ ). The inferences drawn from tests based on the parametric tests such t, F and  $\chi^2$  may be seriously affected when the parent population distributions is not normal. These effects could be more if when sample size is small. Thus when there is doubt about the distribution of the parent population, a nonparametric method should be used. In many situations particularly in social and behavioral sciences observations are difficult or impossible to take on numerical scales. Nonparametric tests are well suited under such situations. However, certain assumptions associated with N.P. tests are: (i) Sample observations are independent, (ii) variables under study are continuous, (iii) p.d.f. is continuous, (iv) Lower order moments exist. Some advantages and Drawbacks of the N.P tests are given below-

Advantages	Drawback
1. N.P. methods are readily comprehensible, very simple and easy to apply and do not require complicated sample theory.	1. N.P. test can be used only if the measurements are nominal or ordinal. Even in that case, if a parametric test exists it is more powerful than the N.P. test.
2. No assumption is made about the form of frequency function of the parent population from which sampling is done	2. So far no N.P. test exists for testing interactions in "Analysis of Variance" model.
3. No parametric techniques will apply to data which are mere classification	3. N.P. tests are designed to test statistical hypothesis only and not

	for estimating the parameters
4. Since, the socio-economic data are not in general, normally distributed; N.P. tests have found application in Psychometry, Sociology and Educational Statistics.	
5. N.P. tests are available to deal with the data which are given in ranks or whose seemingly numerical scores have the strength of ranks.	

First step in statistical testing is formulation of a hypothesis. A hypothesis is a statement about the population. Its plausibility is evaluated on the basis of information obtained by sampling from the population. A test generally involves two hypotheses. An assertion about the population in favour of the ‘existing’ situations is taken as null hypothesis and denoted as  $H_0$ . The negation of the null hypothesis is known as alternative hypothesis and denoted as  $H_1$ .  $H_1$  plays a decisive role in classifying a test as one-sided or two sided. We first develop a statistic  $T$ (say) on the basis of the sample observations. The statistic  $T$  decides whether to reject or accept the null hypothesis. Usually  $T$  follows some distribution. Based on this distribution the range of  $T$  is divided into two groups; the critical region and the region of acceptance. If the sample point falls in critical region, we reject null hypothesis. The size of the critical region depends the risk we wish to incur which ultimately gives the significance level of the test. It is denoted by  $\alpha$ . It represents the probability of rejecting the null hypothesis when it is true, also known as Type I error. Type II error is the probability of rejecting  $H_0$  when  $H_1$  is true and denoted by  $\beta$ . Commonly used significance levels are 5% and 1% ( $\alpha = .05$  and  $.01$ ). Finally, a conclusion is drawn on the basis of the value of  $T$  falling or not falling in the critical region. Some commonly used nonparametric tests are discussed in the sequel.

**1. Run Test for Randomness:**

Run test is used for examining whether or not a set of observations constitutes a random sample from an infinite population. Test for randomness is of major importance because the assumption of randomness underlies statistical inference. In addition, tests for randomness are important for time series analysis. Departure from randomness can take many forms.

$H_0$ : Sample values come from a random sequence

$H_1$ : Sample values come from a non-random sequence

**Test statistic:** Let  $r$  be the number of runs (a Run is a sequence of sign of same kind bounded by signs of other kind). For finding the number of runs, the observations are listed in their order of occurrence. Each observation is denoted by a '+' sign if it is more than the previous observation and by a '-' sign if it is less than the previous observation. Total number of runs up (+s) and down (-) is counted. Too few runs indicate that the sequence is not random (has persistency) and too many runs also indicate that the sequence is not random (is zigzag).

**Critical value:** Critical value for the test is obtained from the table for a given value of  $n$  and at desired level of significance ( $\alpha$ ). Let this value is  $r_c$ .

**Decision rule:** If  $r_c$  (lower)  $\leq r \leq r_c$  (upper) accept  $H_0$ . Otherwise reject  $H_0$ .

**Tied values:** If an observation is equal to its preceding observation denote it by zero. While counting the number of runs ignore it and reduce the value of  $n$  accordingly.

**Large sample sizes:** When sample size is greater than 25 the critical value  $r_c$  can be obtained using a normal distribution approximation.

The critical values for two-sided test at 5% level of significance are

$$r_c \text{ (lower)} = \mu - 1.96 \sigma ; \quad r_c \text{ (upper)} = \mu + 1.96 \sigma$$

For one-sided tests, these are

$$r_c \text{ (left tailed)} = \mu - 1.65 \sigma , \text{ if } r \leq r_c , \text{ reject } H_0$$

$$r_c \text{ (right tailed)} = \mu + 1.65 \sigma , \text{ if } r \geq r_c , \text{ reject } H_1$$

where,  $\mu = \frac{2n-1}{3}$  and  $\sigma = \sqrt{\frac{16n-29}{90}}$

**Example 1:** Data on value of imports of selected agricultural production inputs from U.K. by a county (in million dollars) during recent 12 years is given below: Is the sequence random?

5.2    5.5    3.8    2.5    8.3    2.1    1.7    10.0    10.0    6.9    7.5    10.6

$H_0$ : Sequence is random.  $H_1$ : Sequence is not random.

5.2    5.5    3.8    2.5    8.3    2.1    1.7    10.0    10.0    6.9    7.5    10.6

---

+    -    -    +    -    -    +    0    -    +    +

Here  $n = 11$ , the number of runs  $r = 7$ . Critical values for  $\alpha = 5\%$  (two sided test) from the table are  $r_c$  (lower) = 4 and  $r_c$  (upper) = 10. Since  $r_c$  (lower)  $\leq r \leq r_c$  (upper), *i.e.*, observed  $r$  lies between 4 and 10 the  $H_0$  is accepted. The sequence is random.

**Analysis using R:**

```
> install.packages("randtests")
> library(randtests)
> data<-c(5.2, 5.5, 3.8, 2.5, 8.3, 2.1, 1.7, 10.0, 10.0, 6.9, 7.5, 10.6)
> runs.test(data)
```

**Result:**

**Runs Test**

```
data: data
statistic = -1.8166, runs = 4, n1 = 6, n2 = 6, n = 12, p-value = 0.06928
alternative hypothesis: nonrandomness
```

**2. WALD-WOLFOWITZ Two-Sample Run Test**

Wald–Wolfowitz run test is used to examine whether two random samples come from populations having same distribution. This test can detect differences in averages or spread or any other important aspect between the two populations. This test is efficient when each sample size is moderately large (greater than or equal to 10).

$H_0$ : Two sample come from populations having same distribution

$H_1$ : Two sample come from populations having different distributions

**Test statistic:** Let  $r$  denotes the number of runs. To obtain  $r$ , list the  $n_1 + n_2$  observations from two samples in order of magnitude. Denote observations from one sample by  $x$ 's and other by  $y$ 's. Count the number of runs.

**Critical Value:** Difference in location results in few runs and difference in spread also result in little number of runs. Consequently, critical region for this test is always one-sided. The critical value to decide whether or not the number of runs are few, is obtained from the table. The table gives critical value  $r_c$  for  $n_1$  (size of sample 1) and  $n_2$  (size of sample 2) at 5% level of significance.

**Decision rule:** If  $r \leq r_c$  reject  $H_0$ .

**Tie:** In case  $x$  and  $y$  observations have same value place the observation  $x(y)$  first if run of  $x(y)$  observation is continuing.

**Large sample sizes:** For sample sizes larger than 20 critical values  $r_c$  is given below



$r_c = \mu - 1.96\sigma$  at 5% level of significance

$$\mu = 1 + \frac{2n_1n_2}{n_1 + n_2} \text{ and } \sigma = \sqrt{\frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}}$$

**Example 2:** To determine if a new hybrid seeding produces a bushier flowering plant, following data was collected. Examine if the data indicate that new hybrid produces larger shrubs than the current variety?

**Shrubs Girth (in inches)**

<b>Hybrid</b>	$x$	31.8	32.8	39.2	36.0	30.0	34.5	37.4
<b>Current variety</b>	$y$	35.5	27.6	21.3	24.8	36.7	30.0	

$H_0$ :  $x$  and  $y$  populations are identical

$H_1$ : There is some difference in girth of  $x$  and  $y$  shrubs.

Consider the combined ordered data.

21.3 24.8 27.6 30.0 30.0 31.8 32.8 34.5 35.5 36.0 36.7 37.4 39.2  
 $y$   $y$   $y$   $y$   $x$   $x$   $x$   $x$   $y$   $x$   $y$   $x$   $x$

Test statistic  $r = 6$  (total number of runs). For  $n_1 = 7$  and  $n_2 = 6$ , critical value  $r_c$  at 5% level of significance is 3. Since  $r > r_c$ , we accept  $H_0$  that  $x$  and  $y$  have identical distribution.

**Analysis using R:**

```
> install.packages("DescTools")
> library(DescTools)
> RunsTest(x, y = NULL, alternative = c("two.sided", "less", "greater"), exact = NULL, na.rm = FALSE, ...)
> A <- c(31.8, 32.8, 39.2, 36.0, 30.0, 34.5, 37.4)
> B <- c(35.5, 27.6, 21.3, 24.8, 36.7, 30.0)
> RunsTest(A, B, exact=TRUE)#Exact P-value
> RunsTest(A, B, exact=FALSE)##Exact P-value
```

**Result:**

wald-wolfowitz Runs Test

data: A and B

runs = 8, m = 7, n = 6, p-value = 0.796

alternative hypothesis: true number of runs is not equal the expected number

**3. Median Test for Two Samples**

To test whether or not two samples come from same population median test is used. It is more efficient than the run test but each sample should be of size 10 at least. In this case, the hypothesis to be tested is

$H_0$  : Two samples come from populations having same distribution.

$H_1$  : Two samples come from populations having different distribution.

Test Statistic:  $\chi^2$  (Chi-square). To test the value of test statistics two samples of sizes  $n_1$  and  $n_2$  are combined. Median  $M$  of the combined sample of size  $n = n_1 + n_2$  is obtained. Number of observations below and above the median  $M$  for each sample is determined. This is then analyzed as a  $2 \times 2$  contingency table in the manner given below.

	Number of observations		
	Sample 1	Sample 2	Total
Above Median	a	b	a+b
Below Median	c	d	c+d
	a+c= $n_1$	b+d = $n_2$	n = a+b+ c+d

**Test Statistic:** 
$$\chi^2 = \frac{(ad - bc)^2(a + b + c + d)}{(a + c)(b + d)(a + b)(c + d)}$$

**Decision rule:** if  $\chi^2 \geq \chi_c^2$  reject  $H_0$  otherwise accept it.

**Tie:** ties are ignored and  $n$  is adjusted accordingly.

**Note:** This test can be extended to  $k$  samples. Number of observations below and above the combined median  $M$  from a  $2 \times k$  contingency table.

**Example 3:** Perform a median test on the problem of example 1 for the testing that the two samples come from same population.

$H_0$  : x and y populations are identical.

$H_1$  : There is some difference in girth of x and y shrubs.

## Non-Parametric Tests

	Number of observations		Total
	x	y	
Above M	4	2	6
Below M	2	4	6
	6	6	12

Seventh value 32.8 is the median of combined ordered sequence.

$$\chi^2 = \frac{12(16-4)^2}{6.6.6.6} = \frac{4}{3} = 1.33$$

Since  $\chi^2 = 1.33 < \chi^2_C = 3.84$ ,  $H_0$  is accepted. It is concluded that two samples come from the same population. There is no significance difference in the girth of hybrid and current variety of shrub.

Note: This example is simple to demonstrate test procedure. In real situation  $n$  should be at least 20 and each cell frequency at least 5.

Analysis using R:

```
> library(MASS)
```

```
> chisq.test(x, y = NULL, correct = TRUE, p = rep(1/length(x), length(x)), rescale.p = FALSE,
  simulate.p.value = FALSE,...)
```

### 4. Sign Test for Matched Pairs:

In many situations, comparison of effect of two treatments is of interest but observations occur in pairs. Thus the two samples are not truly random. Because of such pair-wise dependence ordinary two sample tests are not appropriate. In such situations when one member of the pair is associated with the treatment A and the other with treatment B, nonparametric sign test has wide applicability. It can be applied even when qualitative data are available. As the name suggests it is based on the signs of the response differences  $D_i$ . If  $i$ th pairs of observations is denoted by  $(x_i, y_i)$  where  $x$  is the effect of treatment A and  $y$  to B then  $D_i = x_i - y_i$ . The hypothesis to be tested is  $H_0$  : No difference in the effect of treatments A and B.  $H_1$  : A is better than B.

**Test Statistic:** Let  $S$  be the number of '-' signs.

**Critical value:** Critical value  $S_c$  corresponding to  $n$  the number of pairs is given in Table 3. Significance level is given by  $\alpha_1$  as critical region is one sided (left tailed).

**Decision rule:** If  $S \leq S_c$  reject  $H_0$ , other wise accept  $H_0$ .

**Tie:** In case two values of a pair are equal, reject that pair and reduce the number of observations accordingly.

**Note:** In case alternative  $H_1$  is that there is some difference in effect of A and B, S represents either the number of negative signs or the number of positive signs whichever turn out to be smaller. A critical region is two sided and significant level is given by  $\alpha_2$  for finding  $S_c$ .

**Example 4:** In a market study, two brands of lemonade were compared. Each of 50 judges tasted two samples, one of brand A and one of brand B with the following results. 35 preferred brand A, 10 preferred B, and 5 could not tell the difference. Thus  $n = 45$  and  $S = 10$ . Assuming  $\alpha_1 = 5\%$ , critical value  $S_c = 16$  from Table 3. Since  $S < S_c$ , we reject  $H_0$  of no difference in favour of the alternative  $H_1$  that the brand A is preferred.

#### Analysis using R:

The `binom.test()` function performs an exact test of a simple null hypothesis about the probability of success in a Bernoulli experiment from summarized data or from raw data.

```
> binom.test(x, n = NULL, p = 0.5, alternative = c("two.sided", "less", "greater"), conf.level = 0.95,...)
```

```
> binom.test(10, 45)#10=number of success, 45=total number of trial
```

#### Result:

```
Exact binomial test

data: 10 and 45
number of successes = 10, number of trials = 45, p-value =
0.0002471
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.1120459 0.3708876
sample estimates:
probability of success
      0.2222222
```

Another example, here values of the variables given, Hence-

```
>x <- c(1.83, 0.50, 1.62, 2.48, 1.68, 1.88, 1.55, 3.06, 1.30)
>y <- c(0.878, 0.647, 0.598, 2.05, 1.06, 1.29, 1.06, 3.14, 1.29)
>SignTest(x, y)
```

**Result:**

**Dependent-samples Sign-Test**

```

data: x and y
S = 7, number of differences = 9, p-value = 0.1797
alternative hypothesis: true median difference is not equal to 0
96.1 percent confidence interval:
-0.080 0.952
sample estimates:
median of the differences
0.49
    
```

**5. WILCOXON Signed Rank Test for Matched Pairs**

In situations where there is some kind of pairing between observations in the two samples ordinary two sample tests are not appropriate. Signed rank tests are useful in such situations. When observations are measured data, signed rank test is more efficient than sign test as it takes account of the magnitude of the observed differences, if the difference between the response of the two treatments A and B is to be tested the test hypothesis is  
 $H_0$  : No difference in the effect of treatments A and B.  $H_1$  :  
 Treatment A is better than B.

**Test Statistic:** T represents the sum of ranks with negative signs. For calculating T, obtain the differences  $D_i = x_i - y_i$  where  $x_i$ 's are response of treatment A and  $y_i$ 's of treatment B. Rank the absolute values of differences. Smallest give rank 1. Ties are assigned average ranks. Assign to each rank sign of observed difference. Obtain the sum of negative ranks.

**Critical value:**  $T_c$  is given in Table 4 for  $n$  no. of pairs. Significance level is given by  $\alpha_1$  as critical region is one sided.

**Decision rule:**  $T \leq T_c$  reject  $H_0$ , other wise accept it.

**Tie:** Discard the pair for which difference = 0 and reduce  $n$  accordingly. Equal differences are assigned average ranks.

**Example 5:** Blood pressure reading of ten patients before and after medication for reducing the blood pressure are as follows. Test the null hypothesis of no effect against the alternative that medication is effective.

Patient		1	2	3	4	5	6	7	8	9	10
Before treatment	$x$	86	84	78	90	92	77	89	90	90	86
After treatment	$y$	80	80	92	79	92	82	88	89	92	83
Differences		6	4	-14	11	0	-5	1	1	-2	3
Rank		7	5	9	8	Discard	6	1.5	1.5	3	4
Sign		+	+	-	+	Discard	-	+	+	-	+

Rank sum of negative differences = 3+6+9 = 18. Therefore value of test statistic  $T = 18$ . For  $n = 9$  and  $\alpha_1 = 5\%$   $T_c = 8$  from table 4. Since  $T > T_c$  null hypothesis of no effect of medication is accepted.

**Analysis using R:**

```
>library(MASS)
> x<-c(86,84,78,90,92,77,89,90,90,86)
> y<-c(80,80,92,79,92,82,88,89,92,83)
> wilcox.test(x, y, paired = TRUE)
```

**Result:**

wilcoxon signed rank test with continuity correction

data: x and y  
 $V = 27$ , p-value = 0.6353  
 alternative hypothesis: true location shift is not equal to 0

**Warning messages:**

1: In wilcox.test.default(x, y, paired = TRUE) :  
 cannot compute exact p-value with ties  
 2: In wilcox.test.default(x, y, paired = TRUE) :  
 cannot compute exact p-value with zeroes

**6. KOLMOGOROV-SMIRNOV Test**

In situations where there is unequal number of observations in two samples Kolmogorov- Smirnov test is appropriate. This test is used to test whether there is any significance difference between two treatments A and B (say). The test hypothesis is

$H_0$  : No difference in the effect of treatments A and B.

$H_1$  : There is some difference in the effect of treatments A and B.

**Test Statistic:** The test statistic is  $D_{m,n} = \sup |F_m(x) - G_n(x)|$ , both F and G are the sample empirical distribution with sample size m and n.  $F(x_i)$  is calculated as the average number of sample observations of the first sample that are less than  $x_i$ . Similarly  $G(x_i)$  is calculated.  $D_{m,n}$  is largest value of the absolute difference between  $F(x)$  and  $G(x)$ .

**Critical value:** Tabulated value of  $D_{m,n}$  is available for different values of m, n and for different level of significance. is given in Table 4 for n no. of pairs. Significance level is given by  $\alpha_1$  as critical region is one sided.

**Decision rule:** If the calculated value of  $D_{m,n}$  is greater than the Tabulated value of  $D_{m,n}$ ,  $H_0$  is rejected otherwise it is accepted.

**Example 6:** The following data represent the lifetimes (hours) of batteries for different brands:

Brand A	40	30	40	45	55	30
Brand B	50	50	45	55	60	40

Are these brands different with respect to average life?

## Non-Parametric Tests

We first calculate the sample empirical distributions of two samples:

x	$F_6(x)$	$G_6(x)$	$ F_6(x) - G_6(x) $
30	2/6	0	2/6
40	4/6	1/6	3/6
45	5/6	2/6	3/6
50	5/6	4/6	1/6
55	1	5/6	1/6
60	1	1	0

$D_{6,6} = \sup|F_6(x) - G_6(x)| = 3/6$  , From Table the critical value for  $m = n = 6$  at  $\alpha=0.05$  level is 4/6. Since the calculated value of  $D_{m,n}$  is not greater than the Tabulated value,  $H_0$  is not rejected and it is concluded that the average length of life for two brands is the same.

### Analysis using R:

```
require(graphics)
require(dgof)
x <- c(40,30,40,45,55,30)
y <- c(50,50,45,55,60,40)
ks.test(x, y)
```

### Result:

**Two-sample Kolmogorov-Smirnov test**

```
data: x and y
D = 0.5, p-value = 0.4413
alternative hypothesis: two-sided
```

### Some Selected references:

1. Bhattacharya, G.K. and Johnson, R.A. *Statistics concepts and Methods*. New York, John Wiley and Sons. pp 505-521.
2. Gupta, S.C. and Kapoor, V.K. *Fundamentals of Mathematical Statistics*, Sultan Chand & Sons. pp-18.41-18.48.
3. Neave, H.R. and Worthington, P.L. *Distribution free tests*. London Unwin Hyman, pp 161-164,328,337-341.
4. Neter, J.W.W. and Whitmore, G.A. *Applied Statistics*. London, Allyn and Bacon Inc. pp 360- 388.
5. Ostle, B. *Statistics in Reasersch*. Ames. Iowa, USA. The Iowa State University. pp 466-473

## Non-Parametric Tests

### PRACTICAL EXERCISES

1. Maximum level of a lake each year for a period of 20 years is given below. It is desired to test (a) whether the sequence is generated by a random process, or (b) the process contains a trend. The presence of trend will have significant environmental policy implications.

Year	Level(Above 190 meters)	Year	Level(Above 190 meters)
1	6.6	11	6.0
2	6.5	12	5.8
3	6.4	13	5.9
4	6.5	14	5.6
5	6.4	15	5.5
6	6.4	16	5.3
7	6.3	17	5.1
8	6.2	18	5.3
9	6.1	19	5.4
10	5.9	20	5.2

2. Seasonal rainfall at two meteorological observations of a district is given below. Examine by using Run test and Median test whether the rainfall of two observations can be considered as same.

Year	Seasonal rainfall(cm) observations	
	A	B
1985	25.34	24.31
1986	49.35	45.13
1987	39.62	42.83
1988	42.90	46.94
1989	57.66	57.50
1990	24.89	30.70
1991	50.63	48.37
1992	38.47	38.45
1993	43.25	44.00
1994	50.83	50.00
1995	22.02	

3. An experiment was performed to determine if self-fertilized and cross fertilized plants have different growth rates. Pairs of plants one self and other cross fertilized were planted in 15 pots. Their heights were measured after specified period of time.

- (a) Perform the sign test to determine whether there is any difference in the growth rates of self-fertilized and cross fertilized plants.



## Non-Parametric Tests

(b) Perform Wilcoxon signed rank test to determine if crossed plants have a higher growth rate.

Pair	Plant height (cms)		Pair	Plant height (cms)	
	Crossed fertilized	Self-fertilized		Crossed fertilized	Self-fertilized
1	45.5	40.0	9	41.2	41.2
2	40.0	42.3	10	42.7	42.0
3	42.8	41.2	11	43.3	42.0
4	41.6	41.3	12	41.0	40.7
5	37.9	36.7	13	46.0	43.5
6	42.5	38.0	14	39.2	40.6
7	44.1	39.8	15	44.3	42.5
8	40.7	38.9			

**Table 1: Critical values for runs up and down test**

n	$\alpha_1 = 5\%$ $\alpha_2 = 10\%$		$\alpha_1 = 2.5\%$ $\alpha_2 = 5\%$		$\alpha_1 = 1\%$ $\alpha_2 = 2\%$		$\alpha_1 = 0.5\%$ $\alpha_2 = 1\%$	
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
3	-	-	-	-	-	-	-	-
4	-	-	-	-	-	-	-	-
5	1	-	1	-	-	-	-	-
6	1	-	1	-	1	-	1	-
7	2	-	2	-	1	-	1	-
8	2	-	2	-	2	-	1	-
9	3	8	3	-	3	-	2	-
10	3	9	3	-	3	-	2	-
11	4	10	4	10	3	-	3	-
12	4	11	4	11	4	-	3	-
13	5	12	5	12	4	12	4	-
14	6	12	5	13	5	13	4	13
15	6	13	6	14	5	14	4	14
16	7	14	6	14	6	15	5	15
17	7	15	7	15	6	16	6	16
18	8	15	7	16	7	16	6	17
19	8	16	8	17	7	17	7	18
20	9	17	8	17	8	18	7	18
21	10	18	9	18	8	19	8	19
22	10	18	10	19	9	20	8	20
23	1	19	10	20	10	20	9	21
24	1	20	11	20	10	21	10	22
25	12	21	11	21	11	22	10	22

$\alpha_1$  : Significance level for one sided test     $\alpha_2$  : Significance level for two sided test

Source: Distribution Free Tests by H.R. Neave and P.L. Worthington. London, Unwin Hyman.

Non-Parametric Tests

**Table 2: Critical values for the two sample run test.**

n1 \ n2	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2											2	2	2	2	2	2	2	2	2
3					2	2	2	2	2	2	2	2	2	2	3	3	3	3	3
4				2	2	2	3	3	3	3	3	3	3	3	4	4	4	4	4
5			2	2	3	3	3	3	3	4	4	4	4	4	4	4	5	5	5
6		2	2	3	3	3	3	4	4	4	4	5	5	5	5	5	5	6	6
7		2	2	3	3	3	4	4	5	5	5	5	5	6	6	6	6	6	6
8		2	3	3	3	4	4	5	5	5	6	6	6	6	6	7	7	7	7
9		2	3	3	4	4	5	5	5	6	6	6	7	7	7	7	8	8	8
10		2	3	3	4	5	5	5	6	6	7	7	7	7	8	8	8	8	9
11		2	3	4	4	5	5	6	6	7	7	7	8	8	8	9	9	9	9
12	2	2	3	4	4	5	6	6	7	7	7	8	8	8	9	9	9	10	10
13	2	2	3	4	5	5	6	6	7	7	8	8	9	9	9	10	10	10	10
14	2	2	3	4	5	5	6	7	7	8	8	9	9	9	10	10	10	11	11
15	2	2	3	4	5	6	6	7	7	8	8	9	9	10	10	11	11	11	12
16	2	3	4	4	5	6	6	7	8	8	9	9	10	10	11	11	11	12	12
17	2	3	4	4	5	6	7	7	8	9	9	10	10	11	11	11	12	12	13
18	2	3	4	5	5	6	7	8	8	9	9	10	10	11	11	12	12	13	13
19	2	3	4	5	6	6	7	8	8	9	10	10	11	11	12	12	13	13	13
20	2	3	4	5	6	6	7	8	9	9	10	10	11	12	12	13	13	13	14

Significance level 5%

Source: Statistics in Research by Borten Ostle. Ames. Iowa USA. Iowa State University Press.

Non-Parametric Tests

**Table 3: Critical values for the Sign test (Matched pairs)**

n	$\alpha_1$	5 %	2.5 %	1 %	0.5 %	n	$\alpha_1$	5 %	2.5 %	1 %	0.5 %
	$\alpha_2$	10 %	5 %	2 %	1 %		$\alpha_2$	10 %	5 %	2 %	1 %
1	-		-	-	-	26	8		7	6	6
2	-		-	-	-	27	8		7	7	6
3	-		-	-	-	28	9		8	7	6
4	-		-	-	-	29	9		8	7	7
5	0		-	-	-	30	10		9	8	7
6	0		0	-	-	31	10		9	8	7
7	0		0	0	-	32	10		9	8	8
8	1		0	0	0	33	11		10	9	8
9	1		1	0	0	34	11		10	9	9
10	1		1	0	0	35	12		11	10	9
11	2		1	1	0	36	12		11	10	9
12	2		2	1	1	37	13		12	10	10
13	3		2	1	1	38	13		12	11	10
14	3		2	2	1	39	13		12	11	11
15	3		3	2	2	40	14		13	12	11
16	4		3	2	2	41	14		13	12	11
17	4		4	3	2	42	15		14	13	12
18	5		4	3	3	43	15		14	13	12
19	5		4	4	3	44	16		15	13	13
20	5		5	4	3	45	16		15	14	13
21	6		5	4	4	46	16		15	14	13
22	6		5	5	4	47	17		16	15	14
23	7		6	5	4	48	17		16	15	14
24	7		6	5	5	49	18		17	15	15
25	7		7	6	5	50	18		17	16	15

$\alpha_1$  : Significance level for one sided test

$\alpha_2$  : Significance level for two sided test

Source: Distribution Free Tests by H.R. Neave and P.L. Worthington. London, Unwin Hyman.

Non-Parametric Tests

**Table 4: Critical values for the Wilcoxon signed rank test**

n	$\alpha_1$	5 %	2.5 %	1 %	0.5 %	n	$\alpha_1$	5 %	2.5 %	1 %	0.5 %
	$\alpha_2$	10 %	5 %	2 %	1 %		$\alpha_2$	10 %	5 %	2 %	1 %
1	-	-	-	-	-	26	110	-	98	84	75
2	-	-	-	-	-	27	119	-	107	92	83
3	-	-	-	-	-	28	130	-	116	101	91
4	-	-	-	-	-	29	140	-	126	110	100
5	0	-	-	-	-	30	151	-	137	120	109
6	2	-	0	-	-	31	163	-	147	130	118
7	3	-	2	0	-	32	175	-	159	140	128
8	5	-	3	1	0	33	187	-	170	151	138
9	8	-	5	3	1	34	200	-	182	162	148
10	10	-	8	5	3	35	213	-	195	173	159
11	13	-	10	7	5	36	227	-	208	185	171
12	17	-	13	9	7	37	241	-	221	198	182
13	21	-	17	12	9	38	256	-	235	211	194
14	25	-	21	15	12	39	271	-	239	224	207
15	30	-	25	19	15	40	286	-	264	238	220
16	35	-	29	23	19	41	302	-	279	252	233
17	41	-	34	27	23	42	319	-	294	266	244
18	47	-	40	32	27	43	336	-	310	281	261
19	53	-	46	37	32	44	353	-	327	296	276
20	60	-	52	43	37	45	371	-	343	312	291
21	67	-	58	49	42	46	389	-	361	328	307
22	75	-	65	55	48	47	407	-	278	345	322
23	83	-	73	62	54	48	426	-	296	362	339
24	91	-	81	69	61	49	446	-	415	379	355
25	100	-	89	76	68	50	466	-	434	397	373

$\alpha_1$  : Significance level for one sided test

$\alpha_2$  : Significance level for two sided test

Source: Distribution Free Tests by H.R. Neave and P.L. Worthington. London, Unwin Hyman.

---

---

## DATA VISUALIZATION USING R

---

---

Soumen Pal

ICAR-Indian Agricultural Statistics Research Institute

Library Avenue, New Delhi - 110 012

[soumen.pal@icar.gov.in](mailto:soumen.pal@icar.gov.in)

---

---

### Dataset

An experiment was conducted to study the hybrid seed production of bottle gourd (*Lagenaria siceraria (Mol.) Standl.*) Cv. Pusa hybrid-3 under open field conditions during Kharif-2005 at ICAR-Indian Agricultural Research Institute, New Delhi. The main aim of the investigation was to compare natural pollination and hand pollination under field conditions. The data were collected on 10 randomly selected plants from each of natural pollination and hand pollination. The data were collected on number of fruit set for the period of 45 days, fruit weight (kg), seed yield per plant (g) and seedling length (cm). The data obtained is as given below:

{ Here Group 1 denotes natural pollination and Group 2 denotes the hand pollination }

Group	No of Fruit set (45 days)	Fruit Weight(Kg)	Seed Yield per plant(g)	Seedling length(cm)
1	7	1.85	147.70	16.86
1	7	1.86	136.86	16.77
1	6	1.83	149.97	16.35
1	7	1.89	172.33	18.26
1	7	1.80	144.46	17.90
1	6	1.88	138.3	16.95
1	7	1.89	150.58	18.15
1	7	1.79	140.99	18.86
1	6	1.85	140.57	18.39
1	7	1.84	138.33	18.58
2	6.3	2.58	224.26	18.18
2	6.7	2.74	197.50	18.07
2	7.3	2.58	230.34	19.07
2	8	2.62	217.05	19
2	8	2.68	233.84	18
2	8	2.56	216.52	18.49
2	7.7	2.34	211.93	17.45
2	7.7	2.67	210.37	18.97
2	7	2.45	199.87	19.31
2	7.3	2.44	214.30	19.36

## Importing and Attaching the dataset

```
prac.data = read.table(file.choose(),header=T)
attach(prac.data)
```

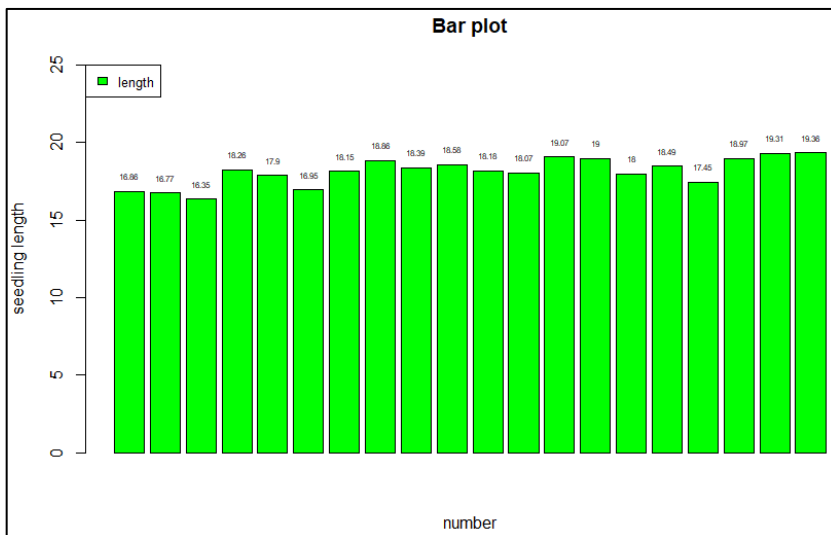
The commonly used diagrams and graphs are:

1. Bar Chart
2. Pie Chart
3. Line Graph
4. Scatter Plot
5. Histogram
6. Box Plot

## Bar Chart

A bar chart represents data in rectangular bars with length of the bar proportional to the value of the variable.

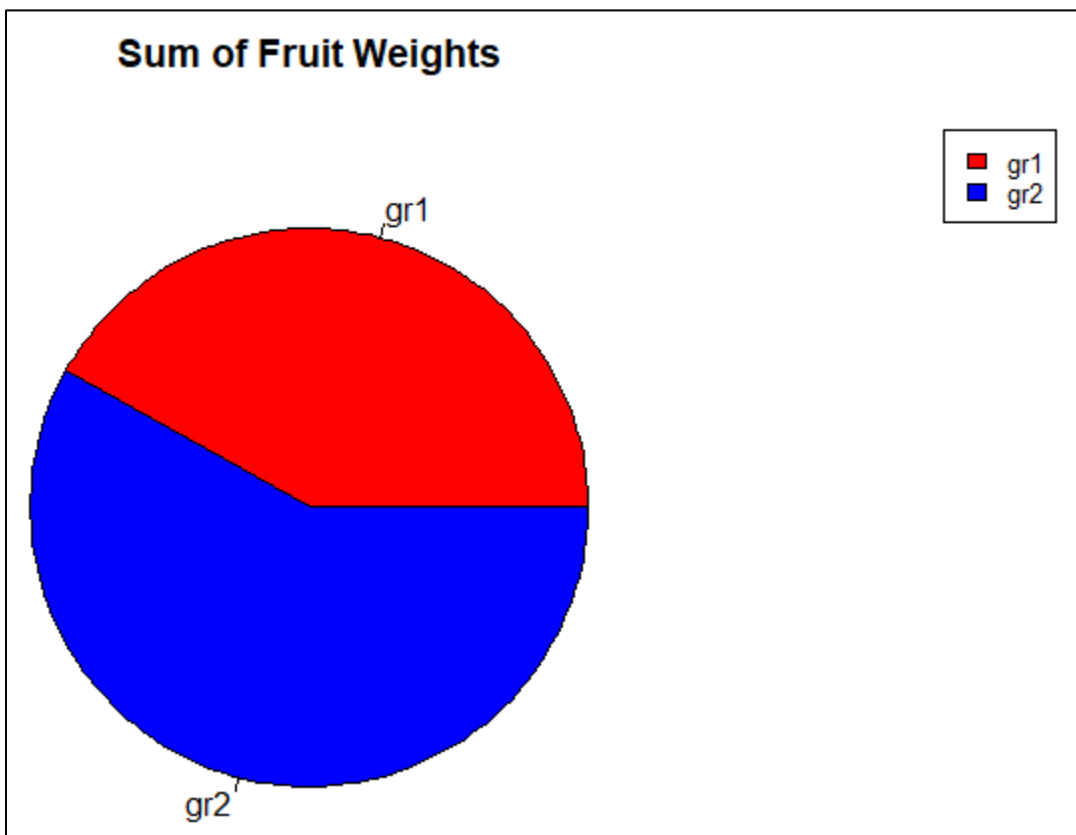
```
bp = barplot(sl, main="Bar plot", xlab='number',ylab='seedling
length',ylim=c(0,25),col='green')
legend("topleft", "length", cex = 0.8, fill = 'green')
## Add text at top of bars
#text(x = bp, y = sl, label = sl, pos = 3, cex = 0.5, col = "red")
# cex: character expansion factor relative to current par("cex").
text(bp,sl+1,labels=as.character(sl),cex = 0.5)
```



## Pie Chart

Pie chart is used to compare the relation between the whole and its components.

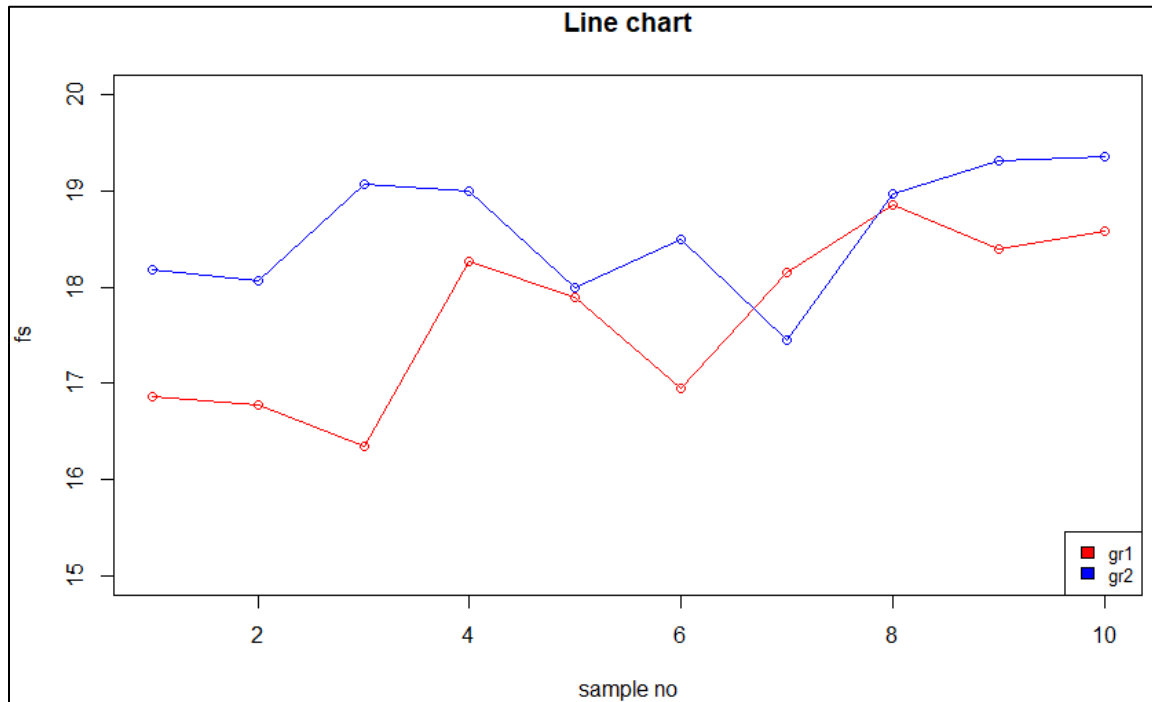
```
# divide the prac.data into 2 groups.  
gr1 = prac.data[c(1:10),]  
gr2 = prac.data[c(11:20),]  
  
wtsum <- c(sum(gr1$fw), sum(gr2$fw))  
labels <- c("gr1", "gr2")  
pie(wtsum, labels, main = "Sum of Fruit Weights", col = c('red',  
'blue'))  
legend("topright", c("gr1", "gr2"), cex = 0.8, fill = c('red',  
'blue'))
```



## Line Graph

Line Graphs are used to display quantitative values over a continuous interval or time period.

```
# Plot the line chart.
plot(gr1$s1,type = "o",col = "red", xlab = "sample no", ylab =
"fs", ylim=c(15,20),
      main = "Line chart")
lines(gr2$s1, type = "o", col = "blue")
legend("bottomright", c('gr1','gr2'), cex = 0.8, fill = c('red',
'blue'))
```



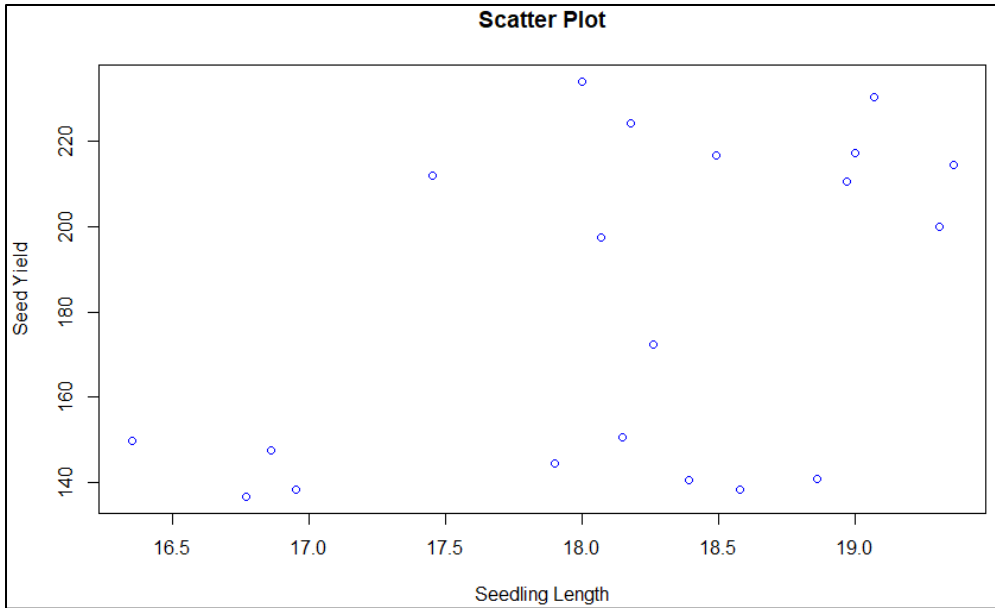
## Scatter Plot

A scatter plot, also known as a scatter graph or a scatter chart, is a two-dimensional data visualization that uses dots to represent the values obtained for two different variables - one plotted along the x-axis and the other plotted along the y-axis.

```
plot(s1,sy, main="Scatter Plot", xlab="Seedling Length",
ylab="Seed Yield",col="blue", type="p", pch=1)
```



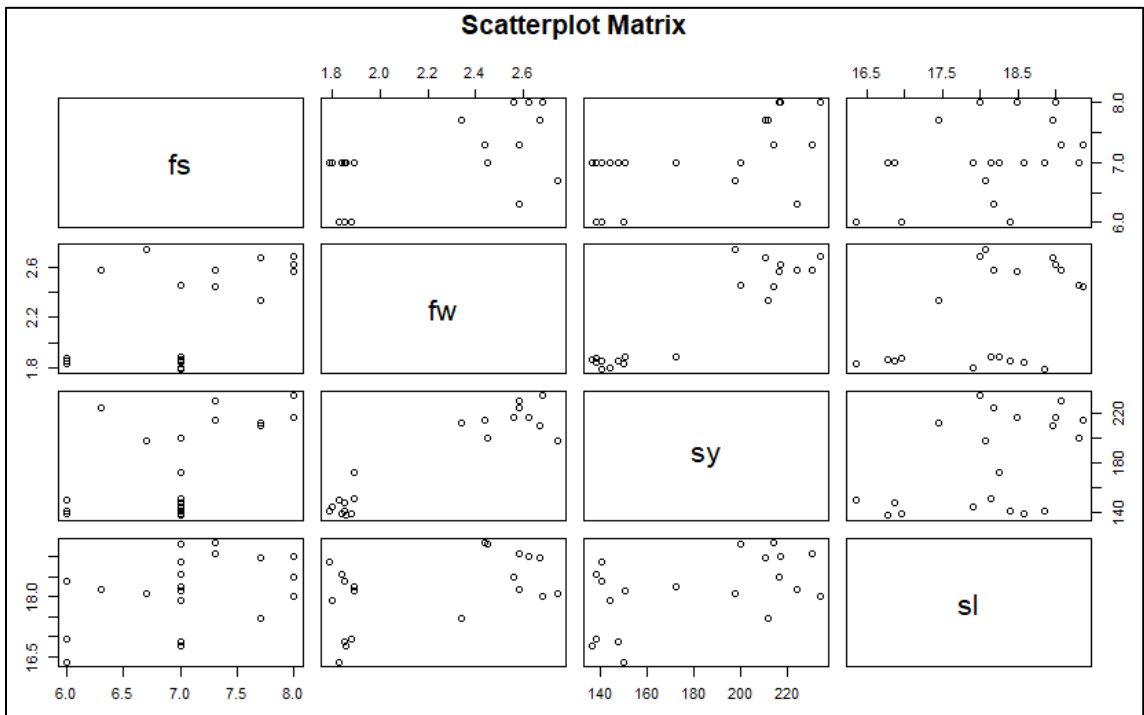
## Data Visualization using R



### Scatter Plot Matrix

For a set of data variables, the scatter plot matrix shows all the pairwise scatterplots of the variables on a single view with multiple scatterplots in a matrix format.

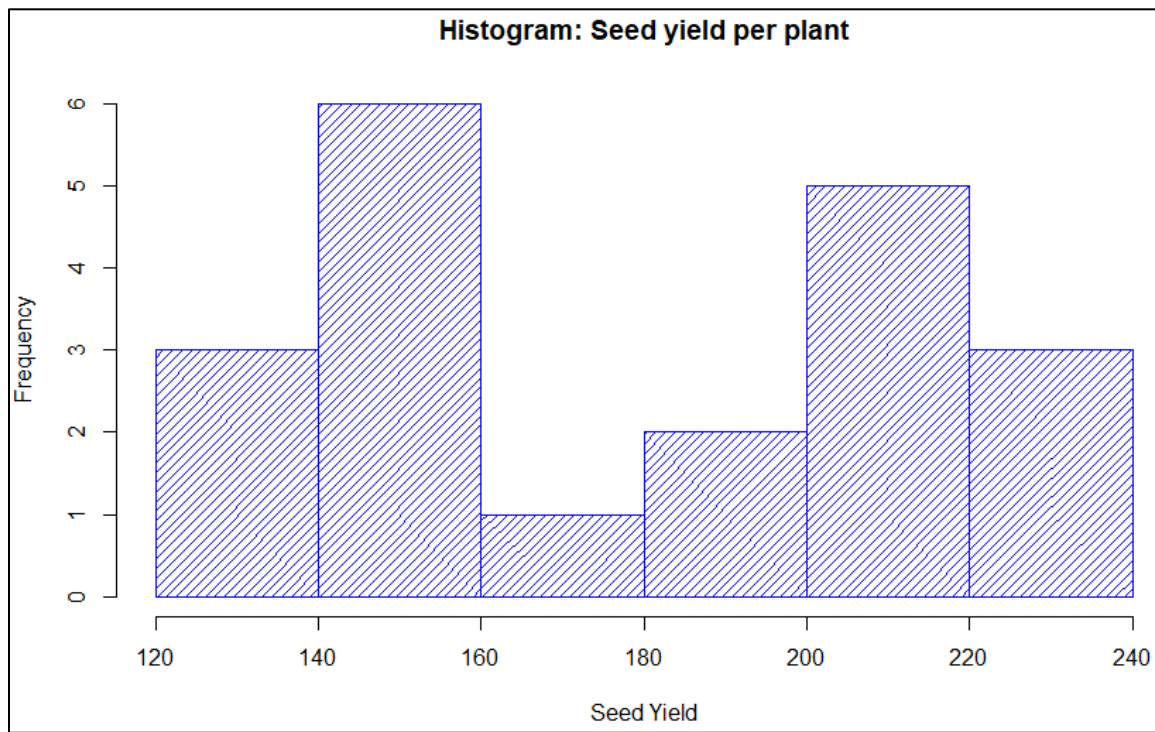
```
pairs(~fs+fw+sy+sl,data = prac.data, main = "Scatterplot Matrix")
```



## Histogram

Histograms show the number of observations that fall within specified divisions (i.e., bins).

```
hist(sy, main = 'Histogram: Seed yield per plant', xlab='Seed  
Yield', ylab='Frequency', col = 'blue', density = 20) # the  
density of shading lines, in lines per inch.
```



## Box Plot

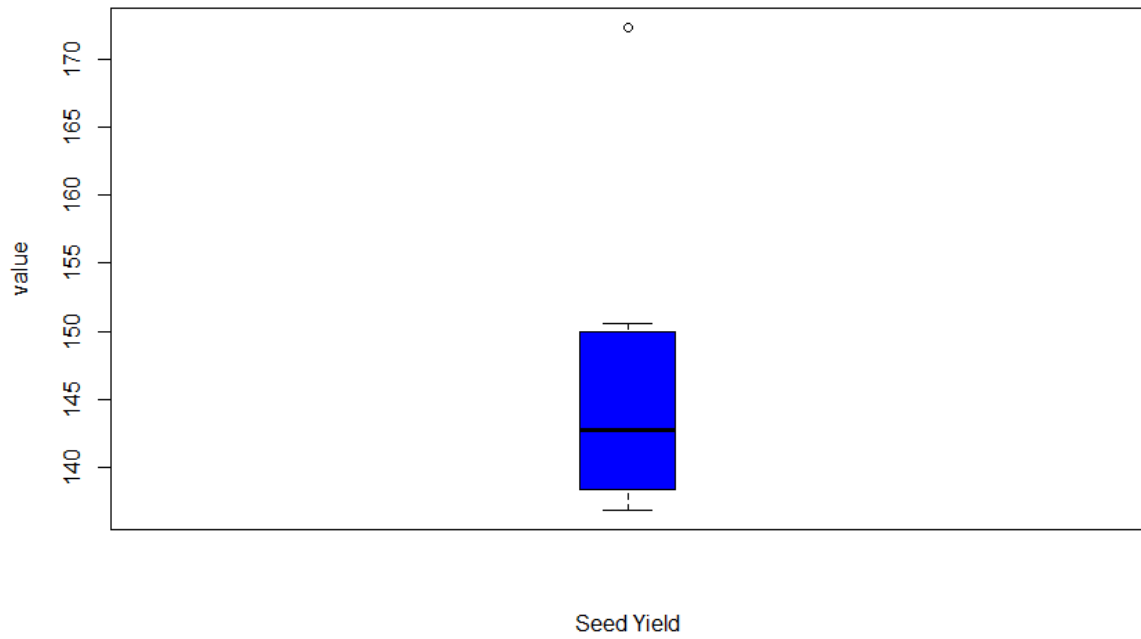
A Box Plot is the visual representation of the statistical five number summary of a given data set. The five number summary includes: Minimum, First Quartile, Median (Second Quartile), Third Quartile, and Maximum.

# Individual Box Plot.

```
boxplot(gr1$sy, main="boxplot", xlab="Seed  
Yield", ylab="value", col="blue", boxwex=0.2) # for gr1.
```

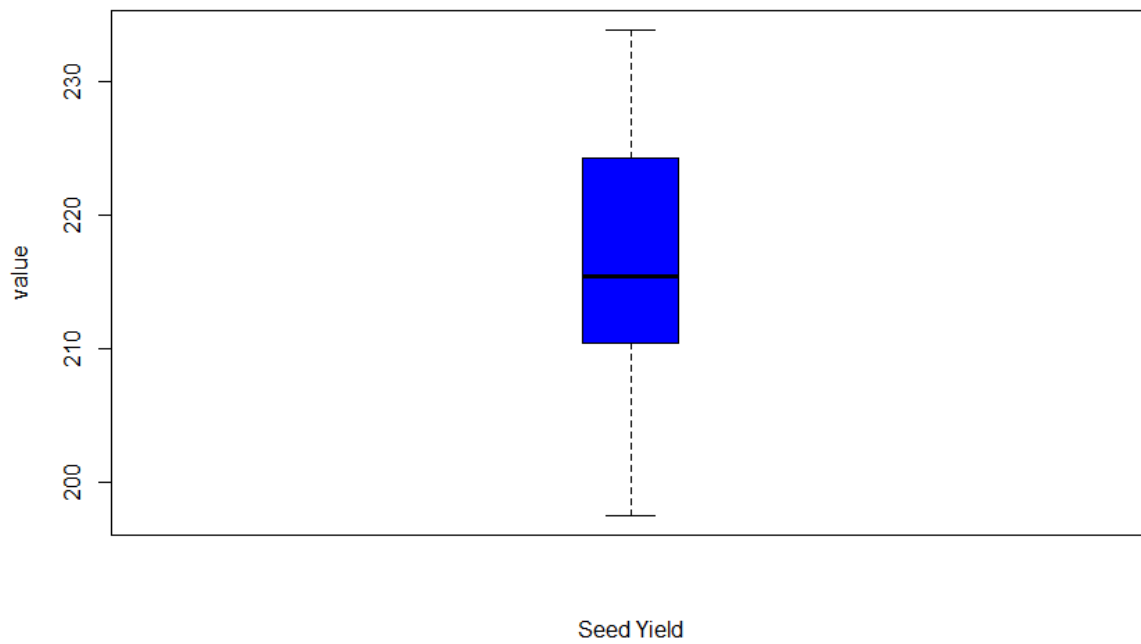
## Data Visualization using R

**boxplot**



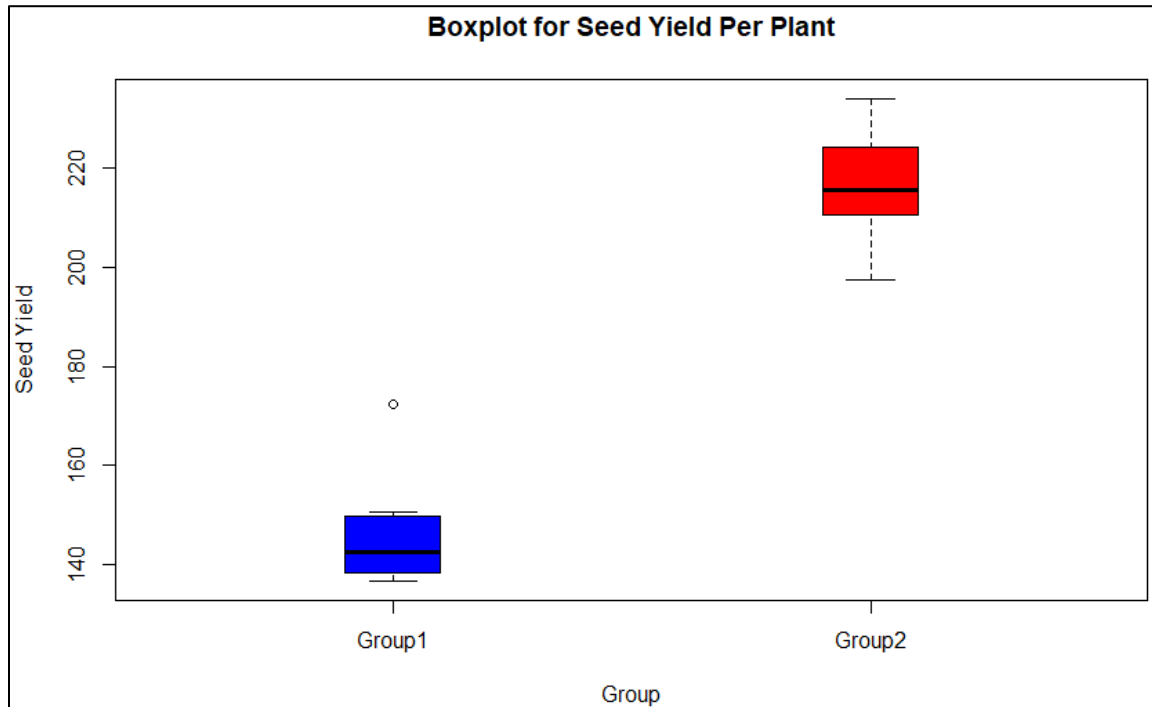
```
boxplot(gr2$sy, main="boxplot", xlab="Seed Yield", ylab="value", col="blue", boxwex=0.2) # for gr2.
```

**boxplot**



## Data Visualization using R

```
# Boxplots in a single diagram.  
boxplot(gr1$sy,gr2$sy, main = 'Boxplot for Seed Yield Per Plant',  
xlab = 'Group', ylab = 'Seed Yield', col = c('blue', 'red'),  
boxwex=0.2, names = c('Group1', 'Group2'))
```



## References

1. <https://www.cran.r-project.org>
2. Chang, W. (2018). *R graphics cookbook: practical recipes for visualizing data*. O'Reilly Media.
3. Healy, K. (2018). *Data visualization: a practical introduction*. Princeton University Press.

---

---

## NONLINEAR GROWTH MODELS

---

---

**Ranjit Kumar Paul and Md Yeasin**

*ICAR-Indian Agricultural Statistics Research Institute*

*Library Avenue, New Delhi - 110 012*

[yeasin.iasri@gmail.com](mailto:yeasin.iasri@gmail.com) , [ranjit.paul@icar.gov.in](mailto:ranjit.paul@icar.gov.in)

---

---

**1. Linear Model.** A mathematical model is an equation or a set of equations which represents the behaviour of a system (France and Thornley, 2006). It can be either 'linear' or 'nonlinear'. A linear model is one in which all the parameters appear linearly. Some examples of linear model are:

(a) *Multiple linear regression*

$$Y = a_0 + a_1X_1 + \dots + a_pX_p + \varepsilon,$$

where  $Y$  is response variable,  $X_i$  are explanatory (or predictor) variables and  $\varepsilon$  is the error term.

(b) *Polynomial models with one predictor variable*

$$Y = a + bX + \varepsilon \quad \text{(First-order model)}$$

$$Y = a + bX + cX^2 + \varepsilon \quad \text{(Second-order or Quadratic or Curvilinear model)}$$

Above models are very widely used in Agriculture, Industry, Education, Medicine, etc. 'Method of least squares' is generally employed for estimation of parameters. However, if polynomial models of a certain order are fitted to data by applying this procedure and later it is decided to add an extra term of a higher order, then estimates of all the parameters in the model have to be computed afresh.

**2. Nonlinear Models.** It is well recognized that any type of statistical inquiry in which principles from some body of knowledge enter seriously into the analysis is likely to lead to a 'Nonlinear model' (Seber and Wild, 2003). Such models play a very important role in understanding the complex inter-relationships among variables. A 'nonlinear model' is one in which at least one of the parameters appears nonlinearly. More formally, in a 'nonlinear model', at least one derivative with respect to a parameter should involve that parameter. Examples of a nonlinear model are:

$$Y(t) = \exp(at+bt^2) \quad (1a)$$

$$Y(t) = at + \exp(-bt). \quad (1b)$$

**Note.** Some authors use the term 'intrinsically linear' to indicate a nonlinear model which can be transformed to a linear model by means of some transformation. For example, the model given by eq. (1a) is 'intrinsically linear' in view of the transformation

$$X(t) = \log_e Y(t).$$

**3. Some Important Nonlinear Growth Models.** Those models, which describe the growth behaviour over time, are applied in many fields. In the area of population biology, growth occurs in plants, animals, organisms, etc. The type of model needed in a specific situation depends on the type of growth that occurs. In general, growth models are mechanistic in nature,

rather than empirical. In the former, the parameters have meaningful biological interpretation; the latter is just like a ‘black-box’ where some input is given and some output is taken out. A mechanistic model usually arises as a result of making assumptions about the type of growth, writing down differential or difference equations that represent these assumptions, and then solving these equations to obtain a growth model. The utility of such models is that, on one hand, they help us to gain insight into the underlying mechanism of the system and on the other hand, they are of immense help in efficient management. We now discuss briefly some well-known nonlinear growth models:

(i) *Malthus Model.* If  $N(t)$  denotes the population size or biomass at time  $t$  and  $r$  is the intrinsic growth rate, then the rate of growth of population size is given by

$$dN/dt = rN. \quad (2)$$

Integrating, we get

$$N(t) = N_0 \exp(rt), \quad (3)$$

where  $N_0$  denotes the population size at  $t=0$ . Thus this law entails an exponential increase for  $r>0$ . Furthermore,  $N(t) \rightarrow \infty$  as  $t \rightarrow \infty$ , which cannot happen in reality.

**Note.** The parameter  $r$  is assumed to be positive in all models.

(ii) *Monomolecular Model.* This model describes the progress of a growth situation in which it is believed that the rate of growth at any time is proportional to the resources yet to be achieved, i.e.

$$dN/dt = r(K-N) \quad (4)$$

where  $K$  is the carrying size of the system. Integrating eq. (4), we get

$$N(t) = K - (K - N_0) \exp(-rt) \quad (5)$$

(iii) *Logistic Model.* This model is represented by the differential equation

$$dN/dt = rN(1-N/K) \quad (6)$$

Integrating, we get

$$N(t) = K/[1 + (K/N_0 - 1) \exp(-rt)] \quad (7)$$

The graph of  $N(t)$  versus  $t$  is elongated S-shaped and the curve is symmetrical about its point of inflexion.

(iv) *Gompertz Model.* This is another model having a sigmoid type of behaviour and is found to be quite useful in biological work. However, unlike the logistic model, this is not symmetric about its point of inflexion. The differential equation for this model is

$$dN/dt = rN \log_e(K/N) \quad (8)$$

Integration of this equation yields

$$N(t) = K \exp[\log_e(N_0/K) \exp(-rt)]. \quad (9)$$

(v) *Richards Model*. This model is given by

$$dN / dt = rN(K^m - N^m) / (mK^m), \quad (10)$$

which, on integration, gives

$$N(t) = K N_0 / [N_0 + (K^m - N_0^m) \exp(-rt)]^{1/m}. \quad (11)$$

Evidently, the last three models are particular cases of this model when  $m = -1, 1, 0$  respectively. However, unlike the earlier models, this model has four parameters.

(vi) *Bass (on Mixed influence) Model*. This model is widely used in “Marketing Research”. It is a combination of Monomolecular and Logistic models and is given by the differential equation

$$dN/dt = (a + bN) (C - N) \quad (12)$$

where  $C$  is carry capacity. The solution is

$$N(t) = \frac{c(a+bd) - a(c-d)e^{-(a+bc)t}}{(a+bd) + b(c-d)e^{-(a+bc)t}}.$$

**4. Fitting of Nonlinear Models.** The above models have been posed deterministically. Obviously this is unrealistic and so we replace these deterministic models by statistical models by adding an error term on the right hand side and making appropriate assumptions about them. This results in a ‘Nonlinear statistical model’. As in linear regression, in non-linear case also, parameter estimates can be obtained by the ‘Method of least squares’. However, minimization of residual sum of squares yield normal equations which are nonlinear in the parameters. Since it is not possible to solve nonlinear equations exactly, the next alternative is to obtain approximate analytic solutions by employing iterative procedures. Three main methods of this kind are:

- i) Linearization (or Taylor Series) method
- ii) Steepest Descent method
- iii) Levenberg-Marquardt’s method

The details of these methods along with their merits and demerits are given in Draper and Smith (1998). The linearization method uses the results of linear least square theory in a succession of stages. However, neither this method nor the Steepest descent method, is ideal. The latter method is able to converge on true parameter values even though initial trial values are far from the true parameter values, but this convergence tends to be very slow at the later stages of the iterative process. On the other hand, the linearization method will converge very rapidly provided the vicinity of the true parameter values has been reached, but if initial trial values are too far removed, convergence may not occur at all.

The most widely used method of computing nonlinear least squares estimators is the Levenberg-Marquardt's method. This method represents a compromise between the other two methods and combines successfully the best features of both and avoids their serious disadvantages. It is good in the sense that it almost always converges and does not 'slow down' at the latter part of the iterative process. We now discuss this method in some detail.

Let us consider the model

$$y_i = f(x_i, \boldsymbol{\theta}) + \varepsilon_i, i = 1, 2, \dots, n, \quad (12)$$

where  $x_i$  and  $y_i$  are respectively the  $i^{\text{th}}$  observations of explanatory and response variables,  $\boldsymbol{\theta} = (\theta_1 \ \theta_2 \ \dots \ \theta_p)'$  are parameters, and error terms  $\varepsilon_i$  are independent and follow

$N(0, \sigma^2)$ . The residual sum of squares is

$$S(\boldsymbol{\theta}) = \sum_{i=1}^n [y_i - f(x_i, \boldsymbol{\theta})]^2. \quad (13)$$

Let  $\boldsymbol{\theta}_0 = (\theta_{10} \ \theta_{20} \ \dots \ \theta_{p0})'$  be the vector of initial parameter values. Then the algorithm

for obtaining successive estimates is essentially given by

$$(H + \tau I)(\boldsymbol{\theta}_0 - \boldsymbol{\theta}_1) = \mathbf{g}, \quad (14)$$

where

$$\mathbf{g} = \left. \frac{\partial S(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta} = \boldsymbol{\theta}_0}, \quad H = \left. \frac{\partial^2 S(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right|_{\boldsymbol{\theta} = \boldsymbol{\theta}_0}, \quad \text{I is the identity matrix and } \tau \text{ is a suitable multiplier.}$$

**Note** (i) Marquardt scaled the various quantities appearing in (14) by the standard deviations of the derivatives with respect to the parameter values.

ii) Now a days most of the standard statistical packages contain computer programmes to fit nonlinear statistical models based on Levenberg-Marquardt algorithm. For example, SPSS has NLR option, SAS has NLIN option, IMSL has RNSSQ option to accomplish the task. In SAS package, one more procedure for nonlinear estimation viz. Does not use derivatives (DUD) procedure is also available.

**Choice of Initial Values.** All the procedures for nonlinear estimation require initial values of the parameters and the choice of good initial values is very crucial. However, there is no standard procedure for getting initial estimates. The most obvious method for making initial guesses is by the use of prior information. Estimates calculated from previous experiments, known values for similar systems, values computed from theoretical considerations all these form ideal initial guesses. Some other methods are:

(i) *Linearization.* After ignoring the error term, check the form of the model to see if it could be transformed into a linear form by means of some transformation. In such cases, linear regression can be used to obtain initial values.



(ii) *Solving a system of equations.* If there are  $p$  parameters, substitute for  $p$  sets of observations into the model ignoring the error. Solve these equations for the parameters, if possible. Widely separated  $x_i$  often work best.

(iii) *Using properties of the model.* Consider the behaviour of the response function as the  $x_i$  go to zero or infinity, and substitute in for observations that most nearly represent those conditions in the scale and context of the problem, solve, if possible, the resulting equations.

(iv) *Graphical method.* Sometimes a visual estimate can be obtained by plotting the data.

**5. Goodness of Fit of a Model.** This is generally assessed by the coefficient of determination,  $R^2$ . However, as pointed out by Kvalseth (1985), eight different expressions for  $R^2$  appear in the literature. Huang and Draper (2003) have suggested yet another measure based on  $R^2$ . One of the most frequent mistakes occurs when the fits of a linear and a nonlinear model are compared by using the same  $R^2$  expression but different variables. Thus, for example, a power model or an exponential model may first be linearized by using a logarithmic transformation and then fitted to data by using ordinary least squares method. The  $R^2$ -value is then often calculated using the data points  $(\log_e y_i, \log_e \hat{y}_i)$ . The  $R^2$  is generally interpreted as a measure of goodness of fit of even the original nonlinear model, which is incorrect. Scott and Wild (1991) have given a real example where two models are identical for all practical purposes and yet have very different values of  $R^2$  calculated on the transformed scales.

Kvalseth (1985) has emphasized that, although  $R_1^2$  given by

$$R_1^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \quad (15)$$

is quite appropriate even for nonlinear models, other summary statistics like

$$\text{Mean Absolute Error (MAE)} = \frac{\sum |y_i - \hat{y}_i|}{n},$$

$$\text{Mean Squared Error (MSE)} = \frac{\sum (y_i - \hat{y}_i)^2}{(n - p)},$$

should also be computed. Here  $n$  is the total number of observed values and  $p$  denotes the number of model parameters.

**6. Examination of Residuals.** Uncritical use and sole reliance on the above statistics may fail to reveal important data characteristics and model inadequacies. Additional detailed analysis of the residuals is strongly recommended to decide about the suitability of a model. Two important assumptions made in the model are:

- i) errors are independent
- ii) errors are normally distributed.

These assumptions can be verified by examining the residuals. If the fitted model is correct, the residuals should exhibit tendencies that tend to confirm or at least should not exhibit a denial of the assumptions. The principal ways of plotting the residuals are: (a) in time sequence, (b) against fitted values. We now discuss the tests for the assumptions (i) and (ii) above.

**(i) Test for independence of errors (Run test).** We test

$H_0$ : Errors are independent

against

$H_1$ : Errors are not independent.

Replace a residual by '+' or '-' sign according as it is positive or negative. Let  $m$  be the number of pluses and  $n$  be the number of minuses in the series of residuals. The test is based on the number of runs ( $r$ ), where a run is defined as a sequence of symbols of one kind separated by symbols of another kind (Siegel and Castellan, 1988). A good large sample approximation to the sampling distribution of the number of runs is the normal distribution with

$$\text{Mean } (\mu) = 2mn/(m+n) + 1$$

and

$$\text{Variance } (\sigma^2) = 2mn(2mn - m - n)/(m+n)^2(m+n-1)^{-1}$$

Therefore, for large samples the required test statistic is

$$Z = (r+h-\mu)/\sigma \sim N(0,1),$$

where

$$h = \begin{cases} 0.5, & \text{if } r < \mu \\ -0.5, & \text{if } r > \mu. \end{cases}$$

$H_0$  is rejected at level  $\alpha$  if  $|Z| > Z_{\alpha/2}$ , where

$$Z_{\alpha} = P\{Z > Z_{\alpha}\} = \alpha$$

**(ii) Test for normality (Shapiro-Wilk test (n<50)).** We test

$H_0$ : Errors are normally distributed

against

$H_1$ : Errors are not normally distributed.

The required test statistic  $W$  is defined as

$$W = S^2/b,$$

where

$$S^2 = \sum a(k) \{x_{(n+1-k)} - x_{(k)}\}, \quad b = \sum (x_i - \bar{X})^2$$

In the above, the parameter  $k$  takes the values

$$k = \begin{cases} 1, 2, \dots, n/2, & \text{when } n \text{ is even} \\ 1, 2, \dots, (n-1)/2, & \text{when } n \text{ is odd} \end{cases}$$

and  $x_{(k)}$  is the  $k^{\text{th}}$  order statistic of the set of residuals. The values of coefficients  $a(k)$  for different values of  $n$  and  $k$  are given in Table 5 of Shapiro-Wilk (1965).  $H_0$  is rejected at level  $\alpha$  if calculated value of  $W$  is less than the tabulated value, which is given in Table 6 of the above article.

## 7. Growth models using r software

## Nonlinear Growth Models

In this section, we will demonstrate the Monomolecular, Logistic and Gompertz models with an example using R software.

**Monomolecular Model:**  $y(t) = K - (K - y_0)\exp(-rt)$ .

**Logistic Model:**  $y(t) = K/[1 + (K/y_0 - 1)\exp(-rt)]$ .

**Gompertz Model:**  $y(t) = K \exp[\ln(y_0/K) \exp(-rt)]$ .

Here,  $y(t)$  denotes the production/ productivity/ area etc. at time  $t$ ,  $r$  is the intrinsic growth rate,  $K$  is the carrying capacity of the system and  $y_0$  is the value of  $y(t)$  at time  $t_0$ .

For logistic and Gompertz models, the annual growth rates pertaining to the period  $(t_i, t_{i+1})$ ,  $[i = 0, 1, \dots, n - 1]$ , where  $n$  denotes the number of data points] respectively are:

$$r^M_t = r[K/y(t) - 1]$$

$$r^L_t = r[1 - y(t)/K]$$

$$r^G_t = r \ln[K/y(t)]$$

**Example:**

Following data contains the total food grain production of India during the period 1980 to 2009:

Year	Production (Million Tonnes)
1980-81	129.59
1981-82	133.30
1982-83	129.52
1983-84	152.37
1984-85	145.54
1985-86	150.44
1986-87	143.42
1987-88	140.35
1988-89	169.92
1989-90	171.04
1990-91	176.39
1991-92	168.38
1992-93	179.48
1993-94	184.26
1994-95	191.50
1995-96	180.42
1996-97	199.44
1997-98	192.26
1998-99	203.61
1999-00	209.80
2000-01	196.81
2001-02	212.85
2002-03	174.77

## Nonlinear Growth Models

2003-04	213.19
2004-05	198.36
2005-06	208.60
2006-07	217.28
2007-08	230.78
2008-09	233.88

(Source: *Agricultural Statistics at a Glance 2009*, Directorate of Economics and Statistics, Ministry of Agriculture, Govt. of India.)

R-code with results has been given below:

```
library(readxl)
library(nlstools)
library(modelr)

data <- read_excel("D:/Training Mannual/Non-linear/Excel.xlsx")
attach(data)
y<-data$Production
t<-data$Year

#-----Monomolecula model Fitting-----#
Model<-nls((y~k-(k-y0)*exp(-r*t)),start=list(k=200, y0=100,r=0.5))
summary(Model)

##
## Formula: y ~ k - (k - y0) * exp(-r * t)
##
## Parameters:
##      Estimate Std. Error t value Pr(>|t|)
## k  307.53314   87.91667   3.498  0.0017 **
## y0 127.52410    5.81721  21.922 <2e-16 ***
## r    0.02673    0.01861   1.437  0.1627
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.42 on 26 degrees of freedom
##
## Number of iterations to convergence: 9
```

## Nonlinear Growth Models

```
## Achieved convergence tolerance: 9.921e-06
```

```
rsquare(Model, data)
```

```
## [1] 0.8912371
```

```
Predicted<-predict(Model)
```

```
Actual<-data$Production
```

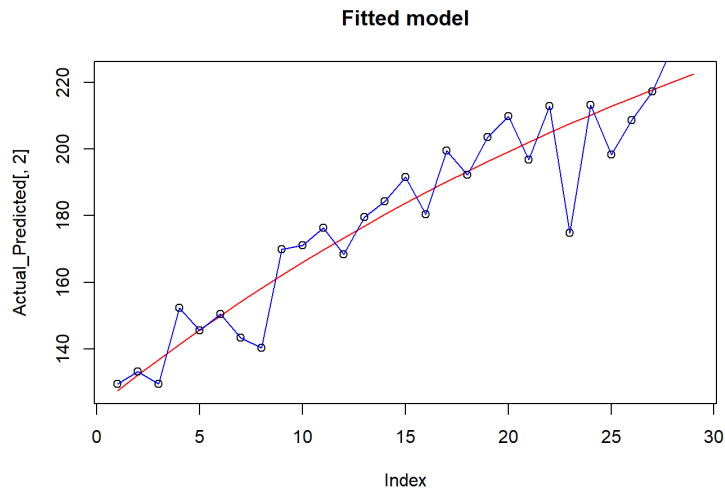
```
Actual_Predicted<-cbind(Actual, Predicted)
```

```
Actual_Predicted
```

```
##      Actual Predicted
## [1,] 129.59 127.5241
## [2,] 133.30 132.2727
## [3,] 129.52 136.8961
## [4,] 152.37 141.3975
## [5,] 145.54 145.7801
## [6,] 150.44 150.0472
## [7,] 143.42 154.2016
## [8,] 140.35 158.2465
## [9,] 169.92 162.1847
## [10,] 171.04 166.0190
## [11,] 176.39 169.7521
## [12,] 168.38 173.3868
## [13,] 179.48 176.9256
## [14,] 184.26 180.3710
## [15,] 191.50 183.7255
## [16,] 180.42 186.9915
## [17,] 199.44 190.1714
## [18,] 192.26 193.2674
## [19,] 203.61 196.2818
## [20,] 209.80 199.2166
## [21,] 196.81 202.0739
## [22,] 212.85 204.8560
## [23,] 174.77 207.5646
## [24,] 213.19 210.2017
## [25,] 198.36 212.7693
## [26,] 208.60 215.2692
```

## Nonlinear Growth Models

```
## [27,] 217.28 217.7031
## [28,] 230.78 220.0728
## [29,] 233.88 222.3800
plot(Actual_Predicted[,2],type="l",col="red",main="Fitted model")
points(Actual_Predicted[,1], cex = 1.0,col="black")
lines(Actual_Predicted[,1],col="blue")
```



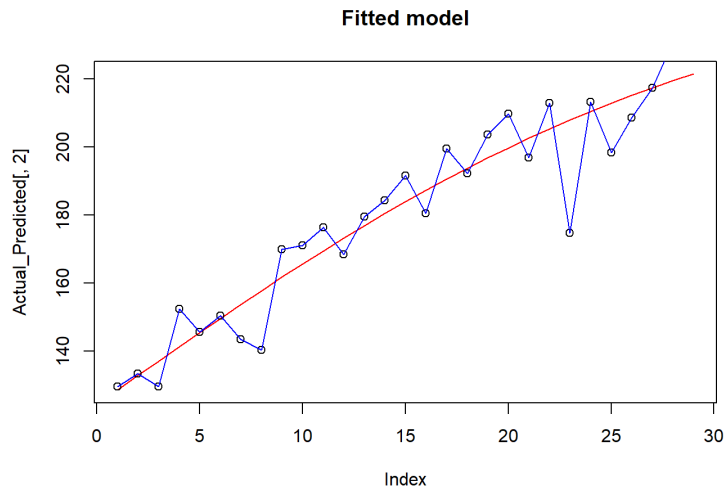
```
#Calculation of Compound Growth Rate

summary<-summary(Model)
coefficient<-summary[["coefficients"]]
coefficient
##      Estimate  Std. Error  t value  Pr(>|t|)
## k  307.53314440  87.91667308  3.498007  1.704821e-03
## y0 127.52409646  5.81721033  21.921864  2.724242e-18
## r   0.02673415  0.01860641  1.436824  1.626934e-01
r<-coefficient[3,1]
r
## [1] 0.02673415
k<-coefficient[1,1]
k
## [1] 307.5331
gr_rate <- r*(k/Actual_Predicted[,1]-1)
```

## Nonlinear Growth Models

```
growth_rate <-mean(gr_rate)
percentage_growth_rate <- growth_rate*100
percentage_growth_rate
## [1] 2.014404
#-----Logistic model Fitting-----#
Model<-nls((y~k/(1+(k/y0-1)*exp(-r*t))),start=list(k=200, y0=100,r=0.5))
summary(Model)
##
## Formula: y ~ k/(1 + (k/y0 - 1) * exp(-r * t))
##
## Parameters:
##      Estimate Std. Error t value Pr(>|t|)
## k  255.02130   29.33704   8.693  3.6e-09 ***
## y0 128.54229    5.35041  24.025 < 2e-16 ***
## r    0.06673    0.02015   3.312  0.00272 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.46 on 26 degrees of freedom
##
## Number of iterations to convergence: 7
## Achieved convergence tolerance: 1.592e-06
rsquare(Model, data)
## [1] 0.8904272
Predicted<-predict(Model)
Actual<-data$Production
Actual_Predicted<-cbind(Actual, Predicted)
plot(Actual_Predicted[,2],type="l",col="red",main="Fitted model")
points(Actual_Predicted[,1], cex = 1.0,col="black")
lines(Actual_Predicted[,1],col="blue")
```

## Nonlinear Growth Models



```
#Calculation of Compound Growth Rate
```

```
summary<-summary(Model)
```

```
coefficient<-summary[["coefficients"]]
```

```
coefficient
```

```
##      Estimate  Std. Error  t value  Pr(>|t|)
## k  255.02130226  29.33703921  8.692810 3.599819e-09
## y0 128.54229236   5.35041140 24.024749 2.804873e-19
## r   0.06673483   0.02014755  3.312306 2.723417e-03
```

```
r<-coefficient[3,1]
```

```
k<-coefficient[1,1]
```

```
gr_rate <- r*(k/Actual_Predicted[,1]-1)
```

```
growth_rate <-mean(gr_rate)
```

```
percentage_growth_rate <- growth_rate*100
```

```
percentage_growth_rate
```

```
## [1] 3.03031
```

```
#-----Gompertz model Fitting-----#
```

```
Model<-nls(y~k*exp(log(y0/k)*exp(-r*t)),start=list(k=200, y0=100,r=0.5))
```

```
summary(Model)
```

```
##
```

```
## Formula: y ~ k * exp(log(y0/k) * exp(-r * t))
```

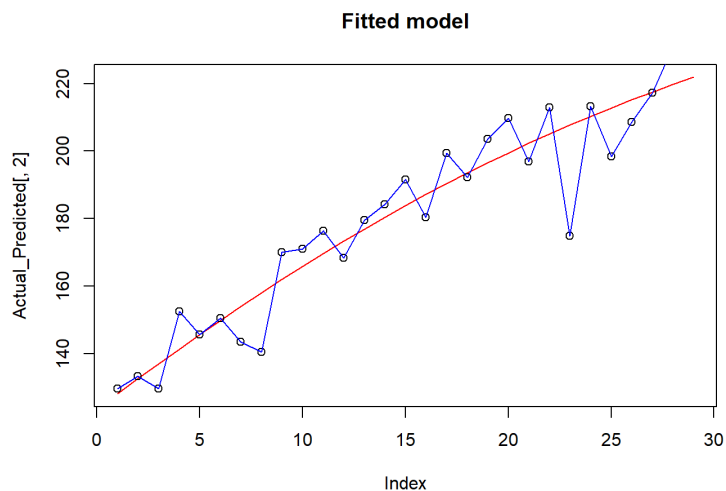
```
##
```

```
## Parameters:
```



## Nonlinear Growth Models

```
##      Estimate Std. Error t value Pr(>|t|)
## k  271.96089   44.47076   6.115 1.83e-06 ***
## y0 128.03080    5.56619   23.001 < 2e-16 ***
## r    0.04674    0.01929    2.423  0.0227 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.44 on 26 degrees of freedom
##
## Number of iterations to convergence: 7
## Achieved convergence tolerance: 5.362e-06
rsquare(Model, data)
## [1] 0.8909237
Predicted<-predict(Model)
Actual<-data$Production
Actual_Predicted<-cbind(Actual, Predicted)
plot(Actual_Predicted[,2],type="l",col="red",main="Fitted model")
points(Actual_Predicted[,1], cex = 1.0,col="black")
lines(Actual_Predicted[,1],col="blue")
```



```
#Calculation of Compound Growth Rate
```

```
summary<-summary(Model)
```

## Nonlinear Growth Models

```
coefficient<-summary[["coefficients"]]
coefficient
##           Estimate  Std. Error  t value    Pr(>|t|)
## k  271.96089364  44.47075502   6.115500 1.829013e-06
## y0 128.03080089   5.56619495  23.001494 8.280989e-19
## r    0.04674349   0.01929183   2.422968 2.266056e-02
r<-coefficient[3,1]
k<-coefficient[1,1]
gr_rate <- r*(k/Actual_Predicted[,1]-1)
growth_rate <-mean(gr_rate)
percentage_growth_rate <- growth_rate*100
percentage_growth_rate
## [1] 2.574017
```

## References

- Draper, N. R. and Smith, H. (1998): *Applied Regression Analysis*, 3rd Ed. John Wiley
- France, J. and Thornley, J.H.M. (2006): *Mathematical Models in Agriculture*, 2<sup>nd</sup> Ed. Butterworths
- Huang, Y. and Draper, N. R. (2003). Transformations, regression geometry and  $R^2$ . *Compu. Statist. Data Anal.*, **42**, 647 – 64
- Kvalseth, T. O. (1985): Cautionary note about  $R^2$ . *Amer. Statistician*, **39**, 279-85
- Prajneshu and Chandran, K. P. (2005): Computation of compound growth rates in agriculture: Revisited. *Ag. Econ. Res. Rev.*, **18**, 317-24
- Ratkowsky, D. A. (1990): *Handbook of Nonlinear Regression Models*. Marcel Dekker
- Scott, A. and Wild, C. J. (1991): Transformations and  $R^2$ . *Amer. Statistician*, **45**, 127-28
- Seber, G. A. F. and Wild, C. J. (2003): *Nonlinear Regression*, 2<sup>nd</sup> Ed. John Wiley

---

---

# LINEAR TIME SERIES MODELLING

---

---

**Ranjit Kumar Paul**

*ICAR-Indian Agricultural Statistics Research Institute*

*Library Avenue, New Delhi - 110 012*

[ranjit.paul@icar.gov.in](mailto:ranjit.paul@icar.gov.in)

---

---

## 1. Introduction

A data set containing observations on a single phenomenon observed over multiple time periods is called time-series. In time-series data, both the *values* and the *ordering* of the data points have meaning. For many agricultural products, data are usually collected over time. Analysis of time series has been a part of statistics for long. Some methods have also been developed for its analysis to suit the distinct features of time series data, which differ both from cross section and panel or pooled data. Various approaches are available for time series modeling. Some of the tools and models which can be used for time series analysis, modeling and forecasting are briefly discussed. Various statistical approaches viz. regression, time series, stochastic and, of late, machine learning approaches are in vogue for statistical modeling. However, the same cannot be claimed to be complete and exhaustive. Every approach has its own advantages and limitations. These models typically utilize a host of empirical data and attempt to forecast market behavior and estimate future values of key variables by using past values of core economic indicators.

Forecasting plays a crucial role in business, Industry, government and institutional planning because many important decisions depend on the anticipated future values of certain variables. Forecast can be made in many different ways, the choice of the method depending on the purpose and importance of the forecasts as well as the costs of alternative methods. The most widely used technique for analysis of time-series data is; undoubtedly, the Box Jenkins' Autoregressive integrated moving average (ARIMA) methodology. In this presentation, we shall talk about 'Univariate' Box-Jenkins models, also referred to as ARIMA models. Univariate or single series means that forecasts are based only on past values of the variable being forecast, they are not based on any other data series. The time-series data refer to observations on a variable that occur in a time sequence. One characteristic of such data is that the successive observations are dependent. Each observation of the observed data series,  $Y_t$ , may be considered as a realization of a stochastic process  $\{Y_t\}$ , which is a family of random variables  $\{Y_t, t \in T\}$ , where  $T = \{0, \pm 1, \pm 2, \dots\}$ , and apply standard time-series approach to develop an ideal model which will adequately represent the set of realizations and also their statistical relationships in a satisfactory manner.

We denote by  $Y_t$ , the observation made at time  $t$  ( $t = 1, 2, \dots, n$ ). Thus, a time-series involving  $n$  points may be represented as sequence of  $n$  observations  $Y_1, Y_2, \dots$ .

,  $Y_n$ . The statistical analysis of time series data differs from the classical regression analysis. Time series data typically violates the assumption that the error terms/successive observations are uncorrelated with each other. This effect, known as autocorrelation, biases the standard error associated with regression slope parameters estimates and makes the relevant t-test invalid. Contrary to statistical independence of observations, in the Box-Jenkins method, we suppose that the time sequenced observations ( $Y_1, Y_2, \dots, Y_{t-1}, Y_t, Y_{t+1}, \dots$ ) may be statistically related to others in the same series. Our goal is to find a good way of stating that statistical relationship. That is, we want to find a good model that describes how the observations in a single time-series are related to each other.

The Box-Jenkins models are specially suited to short term forecasting because most ARIMA models place greater emphasis on the recent past rather than the distant past. The Box-Jenkins method applies to both discrete data as well as to continuous data. However, the data should be available at equally spaced discrete time intervals. Also, building of a ARIMA model requires a minimum of about 40-50 observations.

## **2. Time series models and components**

Time series (TS) data refers to observations on a variable that occurs in a time sequence. Mostly these observations are collected at equally spaced, discrete time intervals. The TS movements of such chronological data can be decomposed into trend, periodic (say, seasonal), cyclical and irregular variations. One or two of these components may overshadow the others in some series. A basic assumption in any TS analysis/modeling is that some aspects of the past pattern will continue to remain in the future.

TS models have advantages over other statistical models in certain situations. They can be used more easily for forecasting purposes because historical sequences of observations upon study variables are readily available from published secondary sources. These successive observations are statistically dependent and TS modeling is concerned with techniques for the analysis of such dependencies. Thus in TS modeling, the prediction of values for the future periods is based on the pattern of past values of the variable under study, but not generally on explanatory variables which may affect the system. There are two main reasons for resorting to such TS models. First, the system may not be understood, and even if it is understood it may be extremely difficult to measure the cause and effect relationship, second, the main concern may be only to predict what will happen and not to know why it happens. Many a time, collection of information on causal factors (explanatory variables) affecting the study variable(s) may be cumbersome /impossible and hence availability of long series data on explanatory variables is a problem. In such situations, the TS models are a boon for forecasters. Hence, if TS models are put to use, say, for instance, for forecasting purposes, then they are especially applicable only in the 'short term'.

A detailed discussion regarding various TS components has been done by Croxton *et al.* (1979). A good account on exponential smoothing methods is given in Makridakis *et al.* (1998). A practical treatment on ARIMA modeling along with several case studies can be found in Pankratz (1983). A reference book on ARIMA and related topics with a more rigorous theoretical flavour is by Box *et al.* (1994). Paul (2010), Paul and Das (2010, 2013), Paul *et al.* (2013, 2014) applied ARIMA model in the field of agriculture as well as livestock's and fisheries.

An important step in analyzing TS data is to consider the types of data patterns, so that the models most appropriate to those patterns can be utilized. Four types of TS components can be distinguished.

They are

- (i) Horizontal – when data values fluctuate around a constant value
- (ii) Trend – when there is long term increase or decrease in the data
- (iii) Seasonal – when a series is influenced by seasonal factor and recurs on a regular periodic basis
- (iv) Cyclical – when the data exhibit rises and falls that are not of a fixed period

Note that many data series include combinations of the preceding patterns. After separating out the existing patterns in any TS data, the pattern that remains unidentifiable form the 'random' or 'error' component. Time plot (data plotted over time) and seasonal plot (data plotted against individual seasons in which the data were observed) help in visualizing these patterns while exploring the data.

Trend analysis of TS data is usually done to analyse a variable over time to detect or investigate long-term changes. Trend is 'long-term' behaviour of a TS process usually in relation to the mean level. The trend of a TS may be studied because the interest lies in the trend itself, or may be to eliminate the trend statistically in order to have insight into other components such as periodic variations in the series. A periodic movement is one which recurs with some degree of regularity, within a definite period. The most frequently studied periodic movement is that which occurs within a year and which is known as seasonal variation. Sometimes the TS data are de-seasonalized for the purpose of making the other movements (particularly trend) more readily discernible. Climatic conditions directly affect the production system in agriculture and hence in turn their patterns of prices and thus are primarily responsible for most of the seasonal variations exhibited in such series.

Over a period of time, a TS is very likely to show a tendency to increase or to decrease otherwise termed as an upward or downward trend respectively. One should not lose sight of the underlying factors that sometimes may cause such trend like growth in population, price changes etc. Technological developments and adoption patterns have been affecting agriculture so as to increase output enormously. Not always keeping pace with them, but induced by them, have been changes in the main variable (read here price of agricultural commodities) under study. Not all historical series show upward trends. Some, say, plant disease incidence, exhibit a generally downward trend. This particular declining trend is attributable to better and more widely available advisory and extension services or due to good government policies. An economic series may have a downward trend because a better or cheaper substitute may be available.

Many techniques such as time plots, auto-correlation functions, box plots and scatter plots abound for suggesting relationships with possibly influential factors. For long and erratic series, time plots may not be helpful. Alternatives could be to go for smoothing or averaging methods like moving averages, exponential smoothing methods etc. In fact, if the data contains considerable error, then the first step in the process of trend identification is smoothing.

### 3. Stationarity of a TS process

A TS is said to be stationary if its underlying generating process is based on a constant mean and constant variance with its autocorrelation function (ACF) essentially constant through time. Thus, if we consider different subsets of a realization (TS 'sample') the different subsets will typically have means, variances and autocorrelation functions that do not differ significantly.

A statistical test for stationarity or test for unit root has been proposed by Dickey and Fuller (1979). The test is applied for the parameter  $\rho$  in the auxiliary regression

$$\Delta_1 y_t = \rho y_{t-1} + \alpha_1 \Delta_1 y_{t-1} + \varepsilon_t$$

where  $\Delta_1$  denotes the differencing operator i.e.  $\Delta_1 y_t = y_t - y_{t-1}$ .

The relevant null hypothesis is  $\rho = 0$  i.e. the original series is non stationary and the alternative is  $\rho < 0$  i.e. the original series is stationary. Usually, differencing is applied until the acf shows an interpretable pattern with only a few significant autocorrelations.

#### 3.1 Autocorrelation functions

##### (i) Autocorrelation

Autocorrelation refers to the way the observations in a TS are related to each other and is measured by the simple correlation between current observation ( $Y_t$ ) and observation from  $p$  periods before the current one ( $Y_{t-p}$ ). That is for a given series  $Y_t$ , autocorrelation at lag  $p$  is the correlation between the pair ( $Y_t, Y_{t-p}$ ) and is given by

$$r_p = \frac{\sum_{t=1}^{n-p} (Y_t - \bar{Y})(Y_{t+p} - \bar{Y})}{\sum_{t=1}^n (Y_t - \bar{Y})^2}$$

It ranges from  $-1$  to  $+1$ . Box and Jenkins has suggested that maximum number of useful  $r_p$  are roughly  $N/4$  where  $N$  is the number of periods upon which information on  $y_t$  is available.

**(ii) Partial autocorrelation**

Partial autocorrelations are used to measure the degree of association between  $y_t$  and  $y_{t-p}$  when the  $y$ -effects at other time lags  $1,2,3,\dots,p-1$  are removed.

**(iii) Autocorrelation function(ACF) and partial autocorrelation function(PACF)**

Theoretical ACFs and PACFs (Autocorrelations versus lags) are available for the various models chosen (say, see Pankratz, 1983) for various values of orders of autoregressive and moving average components i.e.  $p$  and  $q$ . Thus compare the correlograms (plot of sample ACFs versus lags) obtained from the given TS data with these theoretical ACF/PACFs, to find a reasonably good match and tentatively select one or more ARIMA models. The general characteristics of theoretical ACFs and PACFs are as follows:- (here ‘spike’ represents the line at various lags in the plot with length equal to magnitude of autocorrelations)

Model	ACF	PACF
AR	Spikes decay towards zero	Spikes cutoff to zero
MA	Spikes cutoff to zero	Spikes decay to zero
ARMA	Spikes decay to zero	Spikes decay to zero

**4. Description of ARIMA models**

**4.1 Autoregressive (AR) Model**

A stochastic model that can be extremely useful in the representation of certain practically occurring series is the autoregressive model. In this model, the current value of the process is expressed as a finite, linear aggregate of previous values of the process

and a shock  $\varepsilon_t$ . Let us denote the values of a process at equally spaced time epochs  $t, t-1, t-2, \dots$  by  $y_t, y_{t-1}, y_{t-2}, \dots$ , then  $y_t$  can be described by the following expression:

$$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \dots + \varphi_p y_{t-p} + \varepsilon_t$$

If we define an autoregressive operator of order  $p$  by

$$\varphi(B) = 1 - \varphi_1 B - \varphi_2 B^2 - \dots - \varphi_p B^p,$$

where  $B$  is the backshift operator such that  $By_t = y_{t-1}$ , the autoregressive model can be written as  $\varphi(B)y_t = \varepsilon_t$ .

#### 4.2 Moving Average (MA) Model

Another kind of model of great practical importance in the representation of observed time-series is the finite moving average process. MA ( $q$ ) model is defined as

$$y_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \dots - \theta_q \varepsilon_{t-q}.$$

If we define a moving average operator of order  $q$  by

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q,$$

where  $B$  is the backshift operator such that  $By_t = y_{t-1}$ , the moving average model can be written as  $y_t = \theta(B)\varepsilon_t$ .

#### 4.3 Autoregressive Moving Average (ARMA) Model

To achieve greater flexibility in fitting of actual time-series data, it is sometimes advantageous to include both autoregressive and moving average processes. This leads to the mixed autoregressive-moving average model

$$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \dots + \varphi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \dots - \theta_q \varepsilon_{t-q}$$

or

$$\varphi(B)y_t = \theta(B)\varepsilon_t.$$

This is written as ARMA( $p, q$ ) model. In practice, it is frequently true that adequate representation of actually occurring stationary time-series can be obtained with autoregressive, moving average, or mixed models, in which  $p$  and  $q$  are not greater than 2 and often less than 2.

#### 4.4 Autoregressive Integrated Moving Average (ARIMA) Model

A generalization of ARMA models which incorporates a wide class of non-stationary time-series is obtained by introducing the differencing into the model. The simplest example of a non-stationary process which reduces to a stationary one after differencing is Random Walk. A process  $\{y_t\}$  is said to follow an Integrated ARMA model, denoted by ARIMA ( $p, d, q$ ), if  $\nabla^d y_t = (1 - B)^d \varepsilon_t$  is ARMA ( $p, q$ ). The model is written as



$$\varphi(B)(1-B)^d y_t = \theta(B)\varepsilon_t$$

where  $\varepsilon_t \sim WN(0, \sigma^2)$ ,  $WN$  indicating White Noise. The integration parameter  $d$  is a nonnegative integer. When  $d = 0$ ,  $ARIMA(p, d, q) \equiv ARMA(p, q)$ .

The ARIMA methodology is carried out in three stages, viz. identification, estimation and diagnostic checking. Parameters of the tentatively selected ARIMA model at the identification stage are estimated at the estimation stage and adequacy of tentatively selected model is tested at the diagnostic checking stage. If the model is found to be inadequate, the three stages are repeated until satisfactory ARIMA model is selected for the time-series under consideration. An excellent discussion of various aspects of this approach is given in Box *et al.* (2007). Most of the standard software packages, like SAS, SPSS, R and EViews contain programs for fitting of ARIMA models.

#### 4.5 Seasonal Autoregressive Integrated Moving Average (SARIMA) Model

The fundamental fact about seasonal time-series with period  $S$  is that observations, which are  $S$  intervals apart, are similar. Therefore, the operation  $L(y_t) = y_{t-1}$  plays a particularly important role in the analysis of seasonal time-series. In general, the order of SARIMA model is denoted by  $(p, d, q) \times (P, D, Q)_S$ , and the model is represented as follows:

$$\phi_p(L)\Phi_P(L^S)\nabla^d\nabla_S^D y_t = \theta_q(L)\Theta_Q(L^S)\varepsilon_t$$

where  $\phi_p(L)$ ,  $\theta_q(L)$  are polynomials in  $L$  of degrees  $p$  and  $q$  respectively and  $\Phi_P(L^S)$ ,  $\Theta_Q(L^S)$  are polynomials in  $L^S$  of degrees  $P$  and  $Q$  respectively. For estimation of parameters, iterative least squares method is used.

### 5. Model building

#### (i) Identification

The foremost step in the process of modeling is to check for the stationarity of the series, as the estimation procedures are available only for stationary series. If the original series is non stationary then first of all it should be made stationary.

The next step in the identification process is to find the initial values for the orders of seasonal and non-seasonal parameters,  $p, q$ , and  $P, Q$ . They could be obtained by looking for significant autocorrelation and partial autocorrelation coefficients (see section 5 (iii)). Say, if second order auto correlation coefficient is significant, then an AR (2), or MA (2) or ARMA (2) model could be tried to start with. This is not a hard and fast rule, as sample autocorrelation coefficients are poor estimates of population autocorrelation coefficients. Still they can be used as initial values while the final models are achieved

after going through the stages repeatedly. Note that usually up to order 2 for p, d, or q are sufficient for developing a good model in practice.

**(ii) Estimation**

At the identification stage one or more models are tentatively chosen that seem to provide statistically adequate representations of the available data. Then we attempt to obtain precise estimates of parameters of the model by least squares as advocated by Box and Jenkins. Standard computer packages like SAS, SPSS etc. are available for finding the estimates of relevant parameters using iterative procedures. The methods of estimation are not discussed here for brevity.

**(iii) Diagnostics**

Different models can be obtained for various combinations of AR and MA individually and collectively. The best model is obtained with following diagnostics.

**(a) Low Akaike Information Criteria (AIC)/ Bayesian Information Criteria (BIC)/ Schwarz-Bayesian Information Criteria (SBC)**

AIC is given by  $(-2 \log L + 2 m)$  where  $m=p+ q+ P+ Q$  and L is the likelihood function. Since  $-2 \log L$  is approximately equal to  $\{n (1+\log 2\pi) + n \log \sigma^2\}$  where  $\sigma^2$  is the model MSE, Thus AIC can be written as  $AIC=\{n (1+\log 2\pi) + n \log \sigma^2 + 2 m\}$  and because first term in this equation is a constant, it is usually omitted while comparing between models. As an alternative to AIC, sometimes SBC is also used which is given by  $SBC = \log \sigma^2 + (m \log n) /n$ .

**(b) Plot of residual ACF**

Once the appropriate ARIMA model has been fitted, one can examine the goodness of fit by means of plotting the ACF of residuals of the fitted model. If most of the sample autocorrelation coefficients of the residuals are within the limits  $\pm 1.96 / \sqrt{N}$  where N is the number of observations upon which the model is based then the residuals are white noise indicating that the model is a good fit.

**(c) Non-significance of auto correlations of residuals via Portmonteau tests (Q-tests based on Chisquare statistics)-Box-Pierce or Ljung-Box texts**

After tentative model has been fitted to the data, it is important to perform diagnostic checks to test the adequacy of the model and, if need be, to suggest potential improvements. One way to accomplish this is through the analysis of residuals. It has been found that it is effective to measure the overall adequacy of the chosen model by examining a quantity  $Q$  known as Box-Pierce statistic (a function of autocorrelations of residuals) whose approximate distribution is chi-square and is computed as follows:

$$Q = n \sum r^2(j)$$

where summation extends from 1 to  $k$  with  $k$  as the maximum lag considered,  $n$  is the number of observations in the series,  $r(j)$  is the estimated autocorrelation at lag  $j$ ;  $k$  can be any positive integer and is usually around 20.  $Q$  follows Chi-square with  $(k-m_1)$  degrees of freedom where  $m_1$  is the number of parameters estimated in the model. A modified  $Q$  statistic is the Ljung-box statistic which is given by

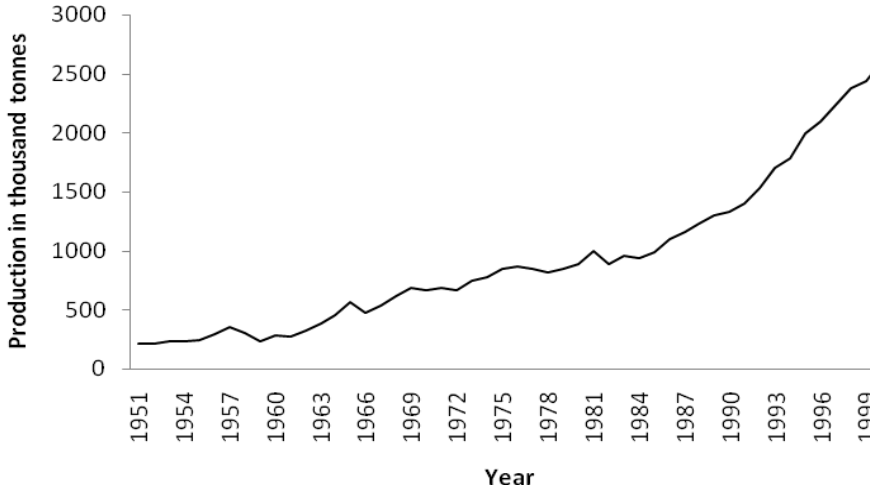
$$Q = n(n+2) \sum r^2(j) / (n-j)$$

The  $Q$  Statistic is compared to critical values from chi-square distribution. If model is correctly specified, residuals should be uncorrelated and  $Q$  should be small (the probability value should be large). A significant value indicates that the chosen model does not fit well.

All these stages require considerable care and work and they themselves are not exhaustive.

## 6. An Illustration

All-India data of inland fish production during the period 1951 to 2008 are obtained from handbook of fishery, ministry of agriculture, Govt. of India and the website [www.indiastat.com](http://www.indiastat.com) and the same are exhibited in Fig. 1. From the total 56 data points, first 50 data points corresponding to the period 1951 to 2000 are used for building the model and remaining are used for validation purpose. A perusal of the data shows that, there is a linear trend in the inland fish production.



**Fig. 1 Inland fish production**

**Fitting of ARIMA Model**

From the estimated autocorrelation function (acf), reported in Table 1, it is found that it decays very slowly thereby requires to be differenced so that the resulting series depicts a pattern for a possible ARMA modelling. Further, in this situation it becomes difficult for selection of order of ARIMA model. The test for unit root proposed by Dickey and Fuller (1979) is applied for the parameter  $\rho$  in the auxiliary regression

$$\Delta_1 y_t = \rho y_{t-1} + \alpha_1 \Delta_1 y_{t-1} + \varepsilon_t$$

The relevant null hypothesis is  $\rho = 0$  and the alternative is  $\rho < 0$ . In the present situation the estimate of  $\rho$  is 0.061 with calculated  $t$ -statistic is 5.55 which is greater than the critical value of  $t$  at 5% level of significance i.e. -1.95 (Franses, 1998) resulting the acceptance of null hypothesis. Thus there is presence of unit root and so differencing is required. Usually, differencing is applied until the acf shows an interpretable pattern with only a few significant autocorrelations. On taking the second difference of the original series, it is seen that only a few acfs, reported in Table 1, are high making it easier to select the order of the model.

**Table 1.** Sample autocorrelation functions (Acf) and partial autocorrelation functions (Pacf) of the original and differenced series

Lag	Acf of the series	Pacf of the series	Acf of the differenced series	Pacf of the differenced series	Acf of the double differenced series	Pacf of the double differenced series
1	0.912	0.912	0.227	0.227	-0.564	-0.564
2	0.829	-0.013	0.346	0.31	0.131	-0.275

3	0.743	-0.065	0.221	0.112	-0.077	-0.219
4	0.661	-0.026	0.203	0.058	-0.075	-0.336
5	0.584	-0.018	0.328	0.231	0.106	-0.243
6	0.509	-0.035	0.242	0.107	0	-0.134
7	0.447	0.024	0.196	-0.017	-0.048	-0.194
8	0.383	-0.048	0.157	-0.02	0.182	0.122
9	0.325	-0.013	-0.101	-0.299	-0.203	0.081
10	0.277	0.018	-0.037	-0.206	0.038	-0.013

The appropriate model is chosen on the basis of minimum Akaike information criterion (AIC) and Bayesian information criterion (BIC) values. Using eqs.(3) and (4), the AIC and BIC values are respectively computed and listed in table 2. A perusal of table 2 shows that the AIC and BIC values are minimum for ARIMA (1,2,2) but the corresponding values for ARIMA (1,2,1) model do not differ much from that of ARIMA(1,2,2). As because ARIMA (1,2,1) is more parsimonious than ARIMA (1,2,2), the ARIMA(1,2,1) model is selected for modelling and forecasting of the inland fish production in India. The estimates of parameters of above model are reported in Table 3.

**Table 2.** AIC and BIC values for different ARIMA models

Criteria	ARIMA(1, 2,0)	ARIMA(1, 2,1)	ARIMA(2, 2,0)	ARIMA(2, 2,1)	ARIMA(2, 2,2)	ARIMA(1, 2,2)
AIC	425.87	413.22	422.97	414.32	412.93	414.27
BIC	443.34	430.69	440.44	431.79	430.41	431.75

**Table 3.** Estimates of parameters along with their SE for fitted ARIMA(1,2,1) model

Parameter	Estimate	Standard error
AR1	-0.141	0.171
MA1	0.823	0.107
Constant	2.623	1.469

The graph of fitted model along with data points is exhibited in Fig. 2. A perusal of fig. 2 indicates that the fitted ARIMA(1,2,1) model is able to capture the trend present in the inland fish production in India very well.

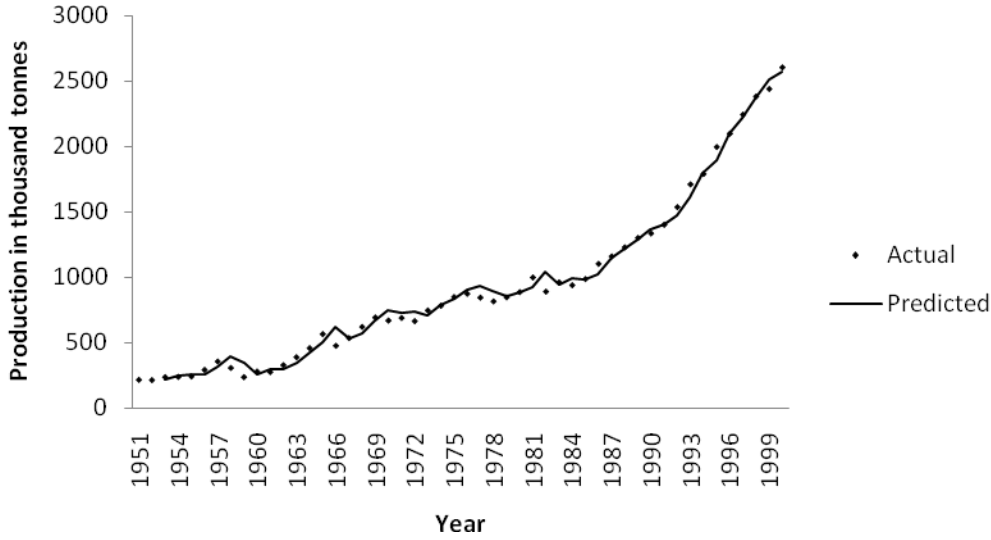


Fig.2 Fitted ARIMA(1,2,1) model along with the data points

One-step ahead forecasts of inland fish production along with their corresponding standard errors, upper confidence interval and lower confidence interval for the year, 2001 to 2008 in respect of above fitted model are reported in Table 4. The attractive feature for fitted ARIMA model is that all the forecast values except for 2008, lie within one standard error of forecasts.

**Table 4.** Forecasts of inland fish production (in tonnes) for fitted models

Years	Actual	Forecasts by ARIMA(1,2,1)	SE of Forecast	Lower Confidence Limit	Upper Confidence Limit
2001	2823.0	2727.09	59.037	2609.3	2844.87
2002	2845.0	2860.68	84.010	2691.15	3030.2
2003	3126.0	2996.05	108.299	2774.81	3217.3
2004	3210.0	3134.17	132.228	2861.08	3407.27
2005	3458.0	3274.9	156.400	2948.71	3601.09
2006	3525.0	3418.25	181.033	3037.4	3799.11
2007	3755.0	3564.23	206.247	3126.96	4001.5
2008	4200.0	3712.83	232.103	3217.32	4208.33

The out of sample forecast of inland fish production in India for the year 2009 and 2010 have been found out as 4360 and 4610 thousand tonnes. For measuring the accuracy in fitted time series model, Mean absolute error (MAE), Mean absolute percentage error

(MAPE) and Relative mean absolute prediction error (RMAPE) are computed by using the formulae given in eqs. 5, 6 and 7. The MAE, MAPE and RMAPE values for fitted ARIMA(1,2,1) model are respectively computed as 160.64, 0.044 and 4.43.

$$MAE = 1/8 \sum_{i=1}^8 |y_{t+i} - \hat{y}_{t+i}| \quad (5)$$

$$MAPE = 1/8 \sum_{i=1}^8 \left\{ |y_{t+i} - \hat{y}_{t+i}| / y_{t+i} \right\} \quad (6)$$

$$RMAPE = 1/8 \sum_{i=1}^8 \left\{ |y_{t+i} - \hat{y}_{t+i}| / y_{t+i} \right\} \times 100 \quad (7)$$

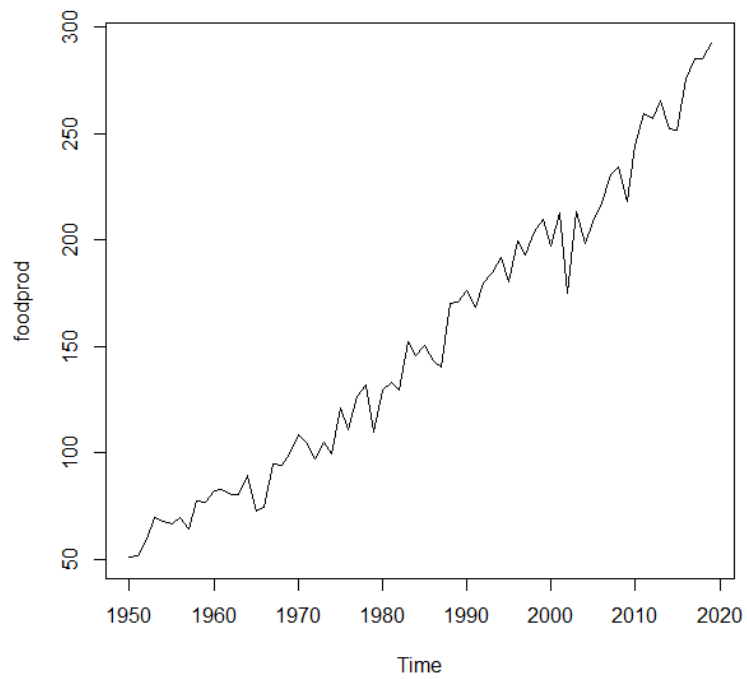
### Conclusion

The ARIMA models being stochastic in nature emphasized variations in data using empirically based methods to determine the proper form of the model that is best suited for short-term forecasting. The more realistic forecast intervals for India's inland fish production data obtained through ARIMA approach could be of immense help to planners in formulating appropriate strategies. These in turn would also benefit the farmers in production of optimum quantities of fish. All this would ultimately result in efficient management of India's inland fish production scenario through sound statistical technique.

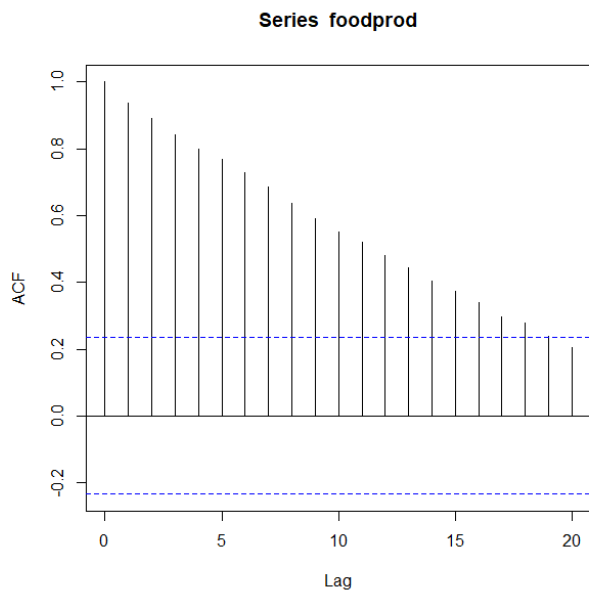
### R code for analyzing time series data

```
#importing data
food<-read.table("C:/Users/Ranjit/Desktop/prod.txt",header=TRUE) #making time series
data
foodprod<-ts(food,frequency=1,start=c(1950))
#plotting time series
plot.ts(foodprod)
```

## Linear Time Series Modelling

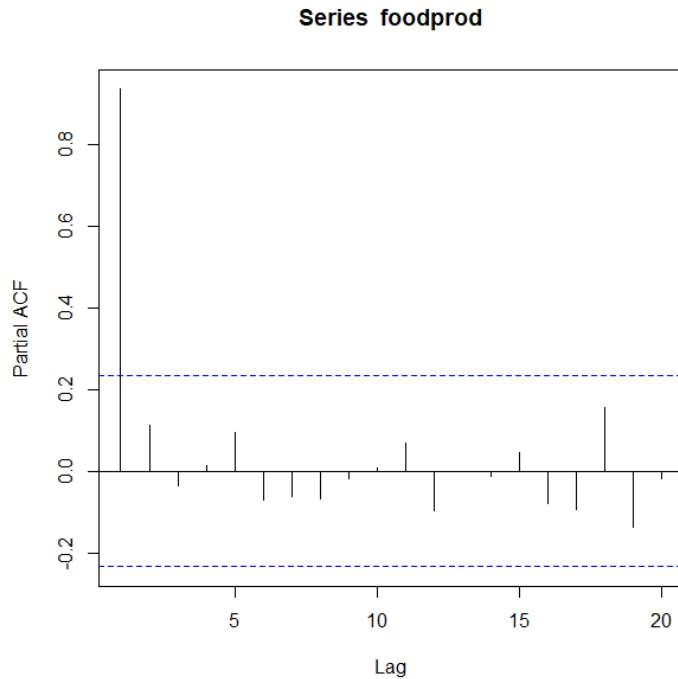


```
#plot of acf and pacf  
acf<-acf(foodprod,lag.max=20) #to get acf values
```





```
pacf<-pacf(foodprod,lag.max=20)
```



Testing stationarity of data

```
library(tseries)
```

```
adf.test(foodprod)
```

```
> adf.test(foodprod)
```

Augmented Dickey-Fuller Test

data: foodprod

Dickey-Fuller = -1.5077, Lag order = 4, p-value = 0.7754

alternative hypothesis: stationary

Testing stationarity of first differenced data

```
> adf.test(diff(foodprod))
```

Augmented Dickey-Fuller Test

data: diff(foodprod)

Dickey-Fuller = -5.6322, Lag order = 4, p-value = 0.01

alternative hypothesis: stationary

#fitting arima model

```
library("forecast")
```

```
foodprodarima<-auto.arima(foodprod, trace=TRUE)
```

```
summary(foodprodarima)
```

```
> foodprodarima<-auto.arima(foodprod, trace=TRUE)
```

```
ARIMA(2,1,2) with drift      : 531.5978
ARIMA(0,1,0) with drift     : 549.5408
ARIMA(1,1,0) with drift     : 530.9133
ARIMA(0,1,1) with drift     : 525.1123
ARIMA(0,1,0)                : 552.5413
ARIMA(1,1,1) with drift     : 527.19
ARIMA(0,1,2) with drift     : 527.2232
ARIMA(1,1,2) with drift     : 529.4065
ARIMA(0,1,1)                : 544.7801
```

```
Best model: ARIMA(0,1,1) with drift
```

```
#forecasting
```

```
foodprodforecast<-forecast.Arima(foodprodarima,h=10)
```

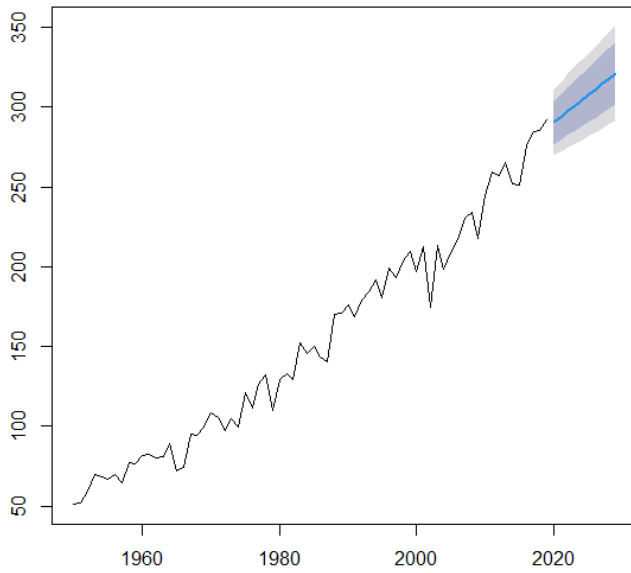
```
> foodprodforecast<-forecast(foodprodarima,h=10)
```

```
> foodprodforecast
```

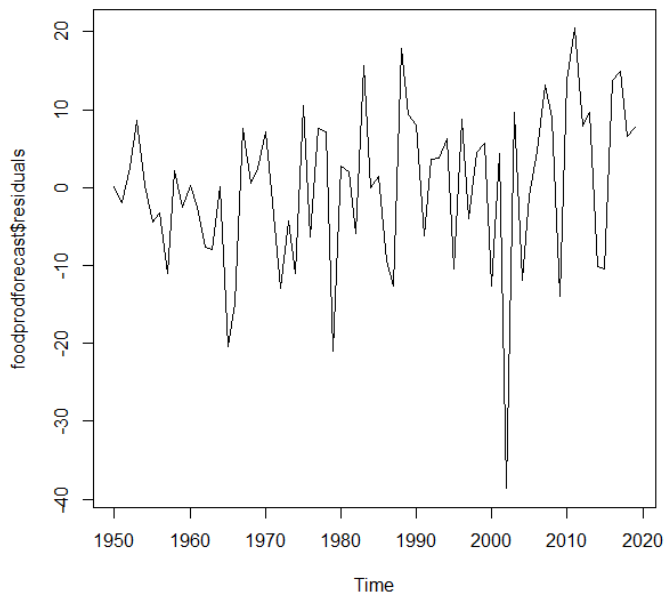
Point	Forecast	Lo 80	Hi 80	Lo 95	Hi 95
2020	290.3624	276.9151	303.8096	269.7965	310.9282
2021	293.7794	279.5573	308.0016	272.0286	315.5303
2022	297.1965	282.2396	312.1534	274.3219	320.0712
2023	300.6136	284.9564	316.2709	276.6679	324.5593
2024	304.0307	287.7031	320.3583	279.0598	329.0016
2025	307.4478	290.4764	324.4192	281.4922	333.4034
2026	310.8649	293.2731	328.4566	283.9606	337.7692
2027	314.2820	296.0910	332.4729	286.4613	342.1026
2028	317.6991	298.9281	336.4701	288.9913	346.4068
2029	321.1161	301.7825	340.4498	291.5479	350.6844

```
plot.forecast(foodprodforecast)
```

Forecasts from ARIMA(0,1,1) with drift

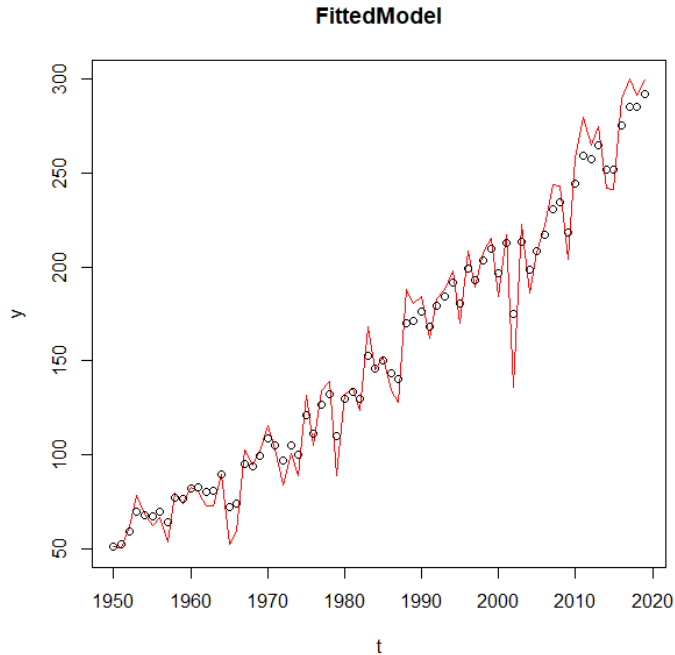


```
plot.ts(foodprodforecast$residuals)
```



```
#finding residuals of fitted arima model  
residual<-resid(foodprodarima)  
#computing predicted values  
pred<-residual+foodprod  
#plotting observed vs predicted values
```

```
obs_Vs_pred = cbind(foodprod,pred,residual)
plot(obs_Vs_pred[,2],type='l',col='red',ylab='y',xlab='t',main='FittedModel')
points(obs_Vs_pred[,1], cex = 1.0,col="black")
leg.txt = c('Observed','Estimated')
```



## References

- Box, G.E.P., Jenkins, G.M. and Reinsel, G.C. (1994). *Time series analysis : Forecasting and control*, Pearson Education, Delhi.
- Croxtan, F.E., Cowden, D.J. and Klein, S. (1979). *Applied General Statistics*, New Delhi: Prentice Hall of India Pvt. Ltd.
- Makridakis, S., Wheelwright, S.C. and Hyndman, R.J. (1998). *Forecasting Methods and Applications*, 3<sup>rd</sup> Edition, John Wiley, New York.
- Pankratz, A. (1983). *Forecasting with univariate Box – Jenkins models: concepts and cases*, John Wiley, New York.
- Paul, R. K. and Das, M. K. (2010). Statistical modelling of inland fish production in India. *Journal of the Inland Fisheries Society of India*, **42**, 1-7.
- Paul, R. K. (2010). Stochastic Modeling of Wholesale Price of Rohu in West Bengal, India. *Interstat*.
- Paul, R. K. and Das, M. K. (2013). Forecasting of average annual fish landing in Ganga Basin. *Fishing chimes*, **33** (3), 51-54
- Paul, R. K., Panwar, S., Sarkar, S. K., Kumar, A. Singh, K. N., Farooqi, S. and Chaudhary, V. K. (2013). Modelling and Forecasting of Meat Exports from India. *Agricultural Economics Research Review*, **26** (2), 249-256.
- Paul, R. K., Alam, W. and Paul, A. K. (2014). Prospects of livestock and dairy production in India under time series framework. *Indian Journal of Animal Sciences*, **84**(4), 130-134.

---

---

## Overview of Python

---

---

**Md. Ashraful Haque**

*ICAR-Indian Agricultural Statistics Research Institute*

*Library Avenue, New Delhi - 110 012*

[ashraful.haque@icar.gov.in](mailto:ashraful.haque@icar.gov.in)

---

---

Python is the one of the most popular programming languages now-a-days. It is a high-level, interpreted, interactive, object-oriented programming language. Python language was created by Guido van Rossum in 1991 at the National Research Institute for Mathematics and Computer Science in the Netherlands. Python programming language is mainly used for-

- Data handling and visualization
- Analysis of variety of data such as numerical, textual, image, videos, audio etc.
- Performing complex mathematical computations
- Server-side scripting for developing web applications
- Standalone software development etc.

### Why Python?

Python is very easy learn language. It can work in any system irrespective of the operating system. Syntax of python language is very simple and allows programmers to write programs in very few lines. Python runs on an interpreter system, which means that the code is being executed as soon as it is written. And last but not the least that python has a very large and mature community for the developers. There are lots of blogs, tutorials, documents, guide videos available online for the python developers.

### Python Installation:

Most of the latest computer systems have python already installed. To check if you have python installed on a Windows PC, search in the start bar for Python or run the following on the Command Line (cmd.exe):

```
C:\your\python\installation\folder>python --version
```

If not, then one can download the latest version of python (latest version is 3.9.2) from <https://www.python.org/downloads/> for the particular operating system and follow the guidelines while installation.

### Getting Started with Python:

Any python script or file is saved with .py file extension. Let's us write the first python program that prints 'Hello, Everyone!!!'. So, first open a text editor and write the following code in it:

e.g.

```
print("Hello, Everyone!!!")
```

## Overview of Python

Now save it as 'first.py'. Now open command prompt, go to the python installation folder and type the following command:

```
C:\your\python\installation\path>python /your/program/path/first.py
```

The output should read:

```
Hello, Everyone!!!
```

### Python from Command Line:

In case of python it is possible to run the code as a command line itself using the command prompt.

Type the following on the Windows, Mac or Linux command line:

```
C:\your\python\installation\path>python
```

From there one can write any python code, including our first example from earlier in the :

```
C:\your\python\installation\path>python
Python 3.6.4 (v3.6.4:d48eceb, Dec 19 2017, 06:04:45) [MSC v.1900 32 bit
(Intel)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>>
```

Which will write "Hello, Everyone!!!" in the command line:

```
C:\your\python\installation\path>python
Python 3.6.4 (v3.6.4:d48eceb, Dec 19 2017, 06:04:45) [MSC v.1900 32 bit
(Intel)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> print("Hello, Everyone!!!")
Hello, Everyone!!!
```

Whenever you are done in the python command line, you can simply type the following to quit the python command line interface:

```
exit()
```

### Python Syntax:

The major syntactical rules of python programs has been provided below-

### *Execution of code*

a. python can be executed directly from command line.

```
>>> print("Hello, Everyone!!!")
Hello, Everyone!!!
```

b. Python can also be executed using a file with '.py' extension

```
C:\your\python\installation\path>python /your/program/path/first.py
```

### *Indentation*

The indentation refers to the spaces at the beginning of a program line. Indentation is very important and stricter in python. Python uses indentation as a block of code.

e.g.

```
if 5 > 2:
    print("Five is greater than two!")
```

### *Comments*

In python, comments can be included in the code by using '#' symbol. Comments can be used in the beginning, middle, or in the end of the code. Comments can be multiline. For multiline comments one can use triple quotes (""").

### **Variables in Python:**

In python, the variables are simple storage structures for storing data values. There is no requirement of *type* declaration for the variables in python. The *type* of any variable can be acquired by *type()* function.

e.g.

```
x = 5
y = "python"
print(type(x))
print(type(y))
```

In python variables names -

- are case sensitive
- Must starts with a letter or underscore
- Can be alphanumeric

## Overview of Python

Python variables can store different types of data.

Text Type:	str
Numeric Types:	int, float, complex
Sequence Types:	list, tuple, range
Mapping Type:	dict
Set Types:	set, frozenset
Boolean Type:	bool
Binary Types:	bytes, bytearray, memoryview

### Operators in Python:

Python divides the operators in the following groups:

Arithmetic operators	+, -, /, *, %, **, //
Assignment operators	=, +=, -=, *=, /=
Comparison operators	==, !=, >, <, >=, <=
Logical operators	And, or, not
Identity operators	is, is not,
Membership operators	in, in not
Bitwise operators	&,  , ^, ~, >>, <<

### Data structures in python:

There are mainly four types of built-in data structures in python which are used for storing collection of data. These data structures are List, Tuple, Set and Dictionary.

**a. Lists-** List are used to store more than one data in single variable. Items in the list are indexed (starting from 0), ordered and changeable. Lists allow duplicate values. Lists are created by using the square brackets and items can be access by mentioning the index number inside the square brackets.

e.g.



## Overview of Python

```
list1 = ["apple", "banana", "cherry"]
list2 = [1, 5, 7, 9, 3]
list3 = [True, False, False]
```

N.B. Python does not have built-in support of Arrays, so Lists can be used as Arrays in python.

**b. Tuple-** Tuples are used to store more than one data in single variable. Items in the tuples are indexed (starting from 0), ordered and non-changeable. Tuples allow duplicate values. Tuples are created by using the round brackets and items can be access by mentioning the index number inside the square brackets.

e.g.

```
tuple1 = ("apple", "banana", "cherry")
tuple2 = (1, 5, 7, 9, 3)
tuple3 = (True, False, False)
```

**c. Set-** Sets are used to store more than one data in single variable. Items in the sets are unindexed, unordered and non-changeable. Sets doesn't allow duplicate values. Sets are created by using curly brackets. Items can't be access by mentioning the index number inside the square brackets. For accessing the items in the Sets one can use any loop structure.

e.g.

```
set1 = {"apple", "banana", "cherry"}
set2 = {1, 5, 7, 9, 3}
set3 = {True, False, False}
```

**d. Dictionary-** Dictionaries are used to store data in key:value pair. Items in the sets are ordered and changeable. But dictionary doesn't allow duplicate values. Dictionaries are created by using curly brackets having key:value pair. Items ca be access by mentioning the key name inside the square bracket.

e.g.

```
dictionary1 = {
    "brand": "Ford",
    "model": "Mustang",
    "year": 1964
}
x = dictionary1 ["model"]
```

### Control and Loops Structures in Python:

There is mainly one control structure that is *'if...else'* and two loop structure such as *'while'* and *'for'* loop.

a. **'if.. else' structure-** 'if...else' structures are used to implement the logical conditions of the program. Syntax of 'if...else' structure is given below-

e.g.

```
a = 33
b = 200
if b > a:
    print("b is greater than a")
else:
    print("error")
```

N.B. Indentation in the control and loop structures are very crucial in case of python programming language.

b. **'while' loop-** With the 'while' loop, one can execute a set of statements as long as a condition is true.

e.g.

```
i = 1
while i < 6:
    print(i)
    i += 1
```

c. **'for' loop-** A 'for' loop is used for iterating over a sequence (that is either a list, a tuple, a dictionary, a set, or a string).

e.g. 1-

```
fruits = ["apple", "banana", "cherry"]
for x in fruits:
    print(x)
```

e.g:2- Looping through the letters of a strings

```
for x in "banana":
    print(x)
```

### Python Functions:

A function is a block of code that contains a set of statements and runs only when it is called explicitly. One can pass data, known as parameters, into a function. A function can return data as a result.

e.g.

```
def my_function(str):
    print(str + "! Welcome to the class.")
my_function("Bob")
```

### Packages and PIP:

Package or module is a python object with arbitrarily named attributes that one can bind and reference. Packages allows us to logically locate the python code. Simply a package or module is file containing a set of python codes. Packages are also referred as library

Packages or modules or libraries can be imported by using the *'import'* keyword.

e.g.

```
import os
```

```
import sys
```

PIP is a package manager available in python. PIP is used to install, upgrade, or uninstall a packages in python environment.

```
C:\your\python\installation\path>pip install numpy
```

### Some important packages or modules in Python:

#### NumPy:

NumPy is python library or packages used for working with arrays. NumPy was created by Travis Oliphant in 2005 and it is open source.

In python, the concepts of arrays is served by the List data structure but it is too slow in processing. NumPy provides a 50x faster access speed for the array objects in python than the List. NumPy has a lots of applications in the domain of -

- Arrays
- Matrices
- Linear Algebra
- Fourier Transformation

#### Creating Arrays

The object of NumPy that deals with the arrays is known as *'ndarray'*. One can create a *'ndarray'* object by using *array()* function. One can pass any type of array-like object in the *array()* function.

e.g.

```
import numpy as np
array_var = np.array([1, 2, 3, 4, 5])
```

Array can be of 0, 1, 2 or 3 dimensions.

e.g.

```
import numpy as np
array0 = np.array(42) #0 dimension
array1 = np.array([1, 2, 3, 4, 5, 6, 7, 8]) # 1 dimension
array2 = np.array([[1, 2, 3], [4, 5, 6]]) # 2 dimension
```

## Overview of Python

```
array3 = np.array([[[1, 2, 3], [4, 5, 6]], [[1, 2, 3], [4, 5, 6]]]) #3  
dimension
```

### *Accessing Array elements*

Array elements can be accessed by its index number

```
print(array1 [2]) #accessing the 3rd item from the array 'array1'
```

### *Slicing an Array*

Slicing in python means taking elements from one given index to another given index.

```
print(array1 [1:3]) #slicing from 2nd item to the 4th element  
print(array1 [2:]) #slicing from 3rd item to the last element  
print(array1 [:6]) #slicing from beginning to the 5th element
```

### *Properties and functions:*

dtype- returns the type of values stored in the array object

shape- gives the number of elements in each dimension of the array object

reshape- allows to change the shape of the array either by adding adding/removing dimensions or changing the number of elements in each dimension

concatenate()- joins two or more arrays axis wise.

array\_split()- splitting an array into two or more parts

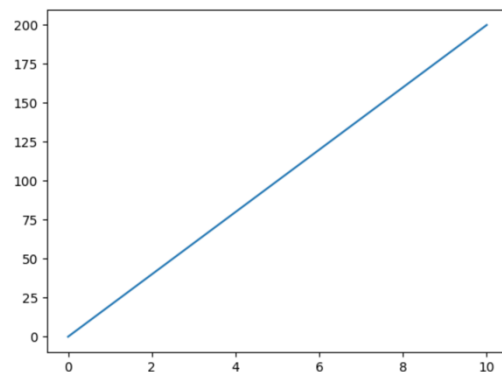
## **Matplotlib:**

Matplotlib is a low level graph plotting library in python that serves as a visualization utility. Matplotlib was created by John D. Hunter. Matplotlib is open source and we can use it freely.

Most of the Matplotlib utilities lies under the pyplot submodule, and are usually imported under the plt alias.

e.g. Draw a line in a diagram from position (0,0) to position (10, 200):

```
import matplotlib.pyplot as plt  
import numpy as np  
xpoints = np.array([0, 10])  
ypoints = np.array([0, 200])  
plt.plot(xpoints, ypoints)  
plt.show()
```



### *Properties and functions:*

marker- keyword argument to emphasize each point in the plot

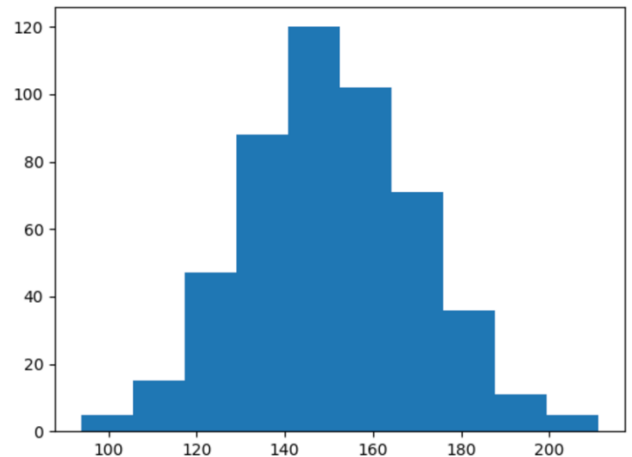
linestyle/ls- keyword argument to change the style of the plotted line

## Overview of Python

`xlabel()`- functions for setting a label for x-axis  
`ylabel()`- function for setting a label for y-axis  
`title()` - function for giving the title for the plot  
`grid()` -function to add grid lines to the plot  
`scatter()`-function to draw a scatter plot  
`bar()`- function to draw bar graphs  
`hist()`- function to create histograms

e.g.

```
import matplotlib.pyplot as plt
import numpy as np
x = np.random.normal(150, 20, 250)
plt.hist(x)
plt.show()
```



### **Pandas:**

Pandas is a one of the most popular python package providing high-performance data manipulation and analysis tool using its powerful data structures. The name Pandas is derived from the word 'Panel Data' – an Econometrics from Multidimensional data. Pandas is well suited for many different kinds of data:

- Fast and efficient DataFrame object with default and customized indexing.
- Tools for loading data into in-memory data objects from different file formats.
- Data alignment and integrated handling of missing data.
- Reshaping and pivoting of date sets.
- Label-based slicing, indexing and subsetting of large data sets.
- Columns from a data structure can be deleted or inserted.
- Group by data for aggregation and transformations

There are mainly two data structures of pandas which handle the majority of typical use cases in finance, statistics, social science and Engineering are Series (1-dimensional) and DataFrame (2-dimensional).

### **DataFrame**

A Data frame is a two-dimensional data structure, i.e., data is aligned in a tabular fashion in rows and columns.

Features of DataFrame:

- Potentially columns are of different types
- Size – Mutable

## Overview of Python

- Labelled axes (rows and columns)
- Can Perform Arithmetic operations on rows and columns

e.g.1

```
import pandas as pd
data = [1,2,3,4,5]
df = pd.DataFrame(data)
print df
```

```
0
0  1
1  2
2  3
3  4
4  5
```

e.g. 2

```
import pandas as pd
data = [['Alex',10],['Bob',12],['Clarke',13]]
df = pd.DataFrame(data,columns=['Name','Age'],dtype=float)
print df
```

```
      Name  Age
0  Alex  10.0
1  Bob   12.0
2  Clarke 13.0
```

### ***Importing data files using pandas***

Pandas provides the means for datafiles to be imported to the python environment. External files in any format (.csv, .xls, .txt, .pdf, etc.) can be imported using pandas.

e.g. 1: .csv file can be imported by read\_csv() function

```
data = pd.read_csv('/content/sample_data/california_housing_test.csv')
```

e.g. 2: .xls file can be imported by read\_excel() function

```
data = pd.read_excel('/content/sample_data/shishamharvesteddata.xls')
```

### ***Measure of central tendency***

Mean, Median and Mode of the dataset can be calculated using mean(), median() and mode() functions available in Pandas

e.g.:

```
## mean
data[].mean()
## median
data[].median()
## mode
data[].mode()
```

### *Description statistics*

Description statistics can be calculated by `describe()` function available in Pandas

e.g.:

```
data[['dbhcm', 'Branchkg', 'Stemkg']].describe()
```

output:

	dbhcm	Branchkg	Stemkg
<b>count</b>	42.000000	42.000000	42.000000
<b>mean</b>	18.927701	27.347262	91.985714
<b>std</b>	4.520851	14.871299	36.560946
<b>min</b>	10.828025	7.630000	20.640000
<b>25%</b>	16.037675	16.015000	66.947500
<b>50%</b>	19.135000	24.550000	96.705000
<b>75%</b>	21.702500	35.378750	114.047500
<b>max</b>	29.681529	70.235000	171.460000

### *Boxplot*

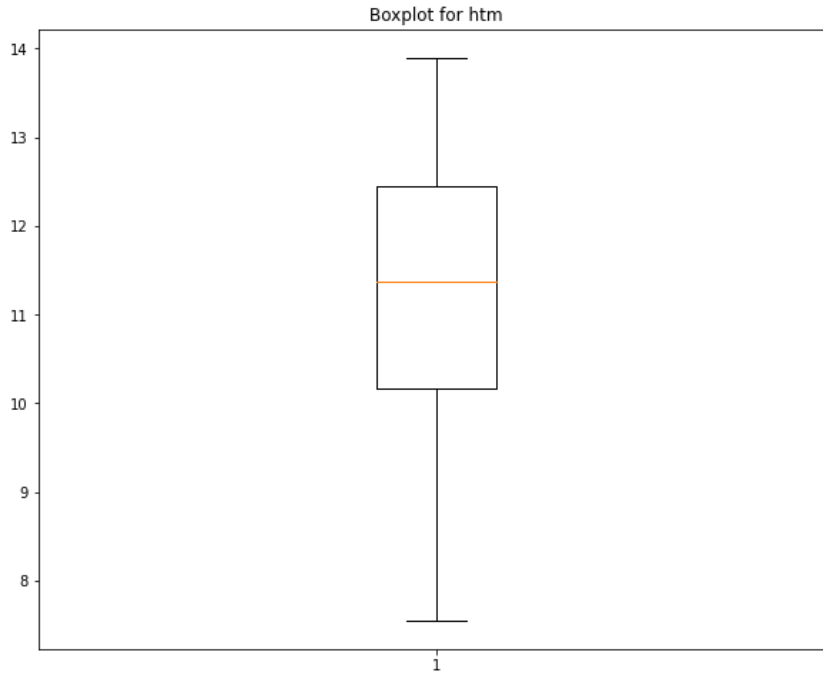
The boxplots can be drawn with the help of `pyplot.boxplot` function available with `matplotlib`.

e.g.:

```
## Boxplot
from matplotlib import pyplot as plt
fig = plt.figure(figsize=(10,8))
plt.boxplot(data['htm'])
plt.title('Boxplot for htm')
plt.show()
```

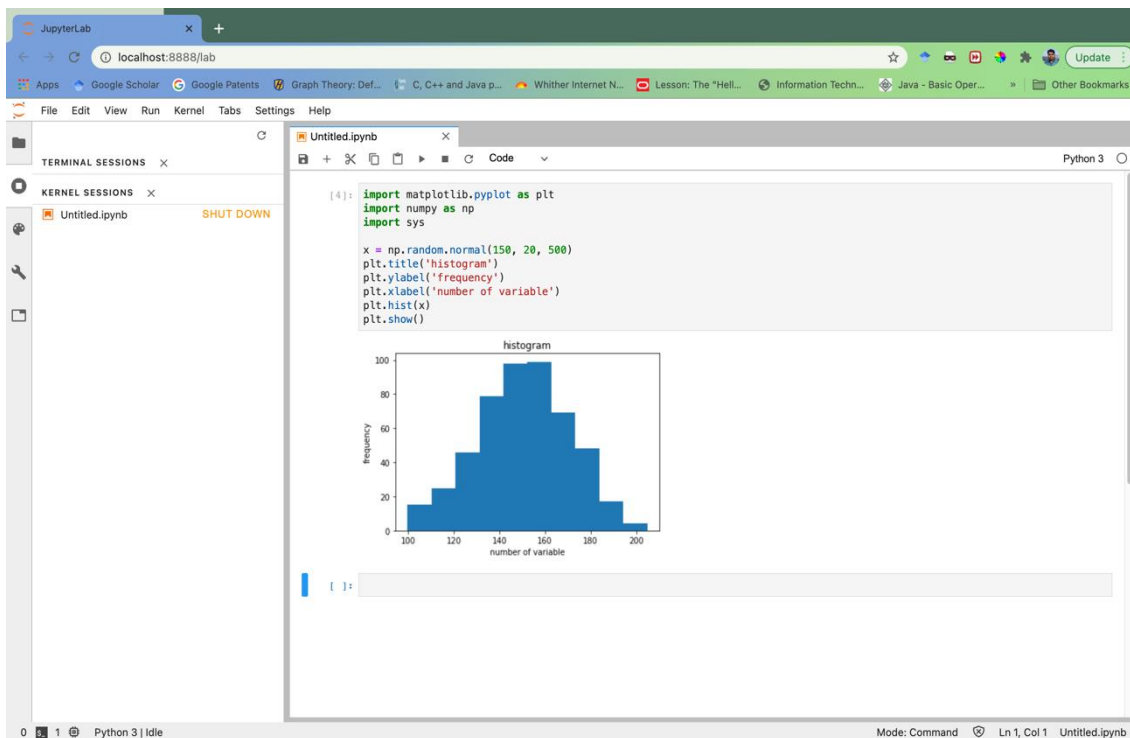
Output:

# Overview of Python



## Jupyter Notebook

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

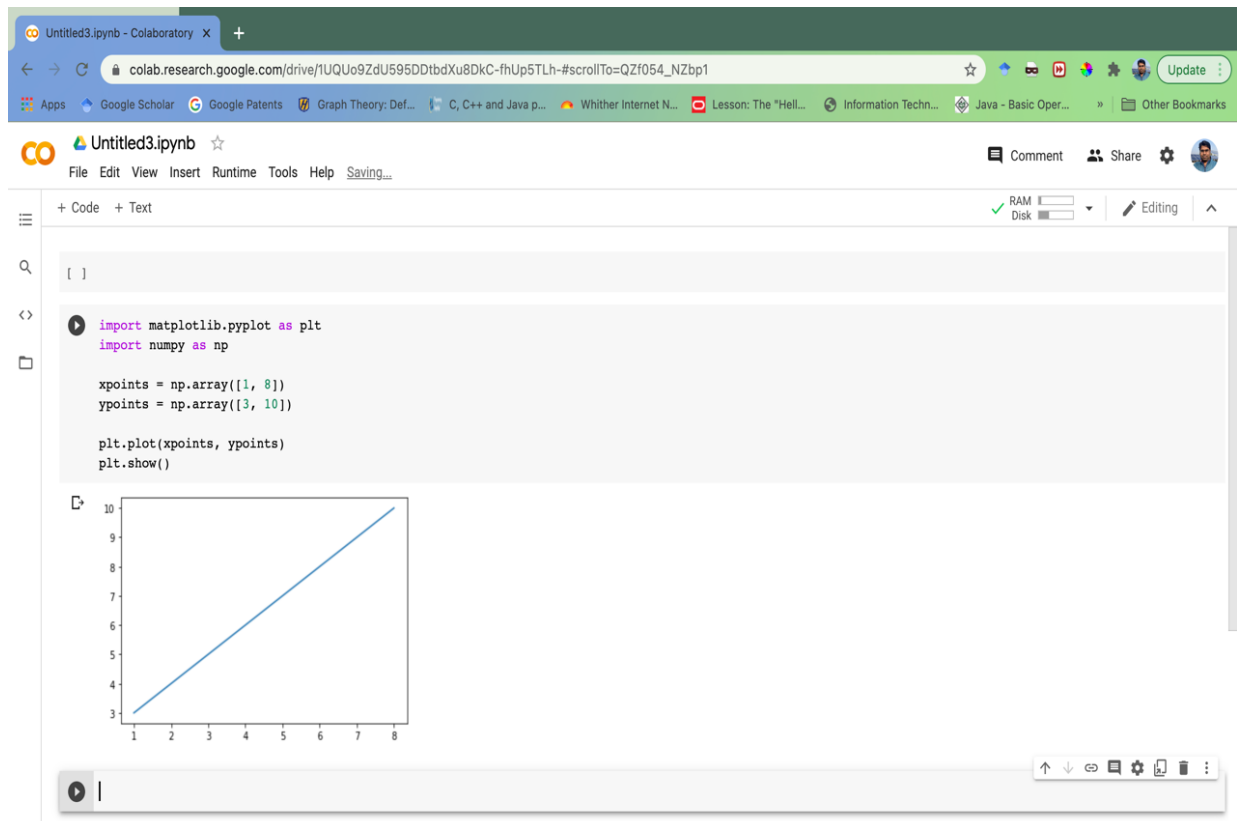




# Overview of Python

## Google Colab:

Colab is a Python development environment that runs in the browser. This facility is provided by the google in free of cost. Jupyter notebook is used for the programming purpose in colab. It allows the user to install any python library at any time. It also provide high computing programming environment that is Graphics Processing Unit (GPU) on free of cost to the users. One can upload his/her data to the google drive and analyse the data using the colab environment.



## Overview of Python

### References:

1. [https://colab.research.google.com/github/tensorflow/examples/blob/master/courses/udacity\\_intro\\_to\\_tensorflow\\_for\\_deep\\_learning/l01c01\\_introduction\\_to\\_colab\\_and\\_python.ipynb#scrollTo=F8YVA\\_634OFk](https://colab.research.google.com/github/tensorflow/examples/blob/master/courses/udacity_intro_to_tensorflow_for_deep_learning/l01c01_introduction_to_colab_and_python.ipynb#scrollTo=F8YVA_634OFk)
2. <https://docs.python.org/3/tutorial/>
3. <https://numpy.org/>
4. <https://pandas.pydata.org/>
5. <https://www.guru99.com/python-tutorials.html>
6. <https://www.programiz.com/python-programming>
7. <https://www.tutorialspoint.com/python/index.htm>
8. <https://www.w3schools.com/python/default.asp>

---

---

# OVERVIEW OF SAMPLING METHODS

---

---

**Ankur Biswas**

*ICAR-Indian Agricultural Statistics Research Institute*

*Library Avenue, New Delhi - 110 012*

[ankur.biswas@icar.gov.in](mailto:ankur.biswas@icar.gov.in)

---

---

## **1. Introduction**

The need to gather information arises in almost every conceivable sphere of human activity. Many of the questions that are subject to common conservation and controversy require numerical data for their resolution. The data collected and analyzed in an objective manner and presented suitably serve as a basis for taking policy decisions in different fields of daily life.

The important users of statistical data, among others, include government, industry, business, research institutions, public organizations and international agencies and organizations. To discharge its various responsibilities, the government needs variety of information regarding different sectors of economy, trade, industrial production, health and mortality, population, livestock, agriculture, forestry, environment and available resources. The inferences drawn from the data help in determining future needs of the nation and also in tackling social and economic problems of people. For instance, the information on cost of living for different categories of people, living in various parts of the country is of importance in shaping its policies in respect of wages and price levels. Data on agricultural production are of immense use to the state for planning to feed the nation. In case of industry and business, the information is to be collected on labour, cost and quality of production, stock and demand and supply positions for proper planning of production levels and sales campaigns.

### **1.1 Complete enumeration**

One way of obtaining the required information at regional and country level is to collect the data for each and every unit (person, household, field, factory, shop etc. as the case may be) belonging to the population which is the aggregate of all units of a given type under consideration and this procedure of obtaining information is termed as complete enumeration. The effort, money and time required for the carrying out complete enumeration to obtain the different types of data will, generally, be extremely large. However, if the information is required for each and every unit in the domain of study, a complete enumeration is clearly necessary. Examples of such situations are preparation of “voter list” for election purposes and

recruitment of personnel in an establishment, etc. But there are many situations, where only summary figures are required for the domain of study as a whole or for group of units.

### **1.2 Need for sampling**

An effective alternative to a complete enumeration can be sample survey where only some of the units selected in a suitable manner from the population are surveyed and an inference is drawn about the population on the basis of observations made on the selected units. It can be easily seen that compared to sample survey, a complete enumeration is time-consuming, expensive, has less scope in the sense of restricted subject coverage and is subject to greater coverage, observational and tabulation errors. In certain investigations, it may be essential to use specialized equipment or highly trained field staff for data collection making it almost impossible to carry out such investigations. It is of interest to note that if a sample survey is carried out according to certain specified statistical principles, it is possible not only to estimate the value of the characteristic of the population as a whole on the basis of the sample data, but also to get a valid estimate of the sampling error of the estimate. There are various steps involved in the planning and execution of the sample survey. One of the principal steps in a sample survey relates to methods of data collection.

### **1.3 Methods of data collection**

The different methods of data collection are:

- i. Physical observation or measurement
- ii. Personal interview
- iii. Mail enquiry
- iv. Telephonic enquiry
- v. Web-based enquiry
- vi. Method of Registration
- vii. Transcription from records

The first six methods relate to the collection of primary data from the units/ respondents directly, while the last one relates to the extraction of secondary data, collected earlier generally by one or more of the first six methods. These methods have their respective merits and therefore sufficient thought should be given in selection of an appropriate method(s) of data

collection in any survey. The choice of the method of data collection should be arrived at after careful consideration of accuracy, practicability and cost from among the alternative methods.

### **1.4. Various concepts and definitions**

#### *i. Element:*

An element is a unit about which we require information. For example, a field growing a particular crop is an element for collecting information on the yield of a crop.

#### *ii. Population*

The collection of all units of a specified type in a given region at a particular point or period of time is termed as a population or universe. Thus, we may consider a population of persons, families, farms, cattle in a region or a population of trees or birds in a forest or a population of fish in a tank etc. depending on the nature of data required.

#### *iii. Sampling unit*

Elementary units or group of such units which besides being clearly defined, identifiable and observable, are convenient for the purpose of sampling are called sampling units. For instance, in a family budget enquiry, usually a family is considered as the sampling unit since it is found to be convenient for sampling and for ascertaining the required information. In a crop survey, a farm or a group of farms owned or operated by a household may be considered as the sampling unit.

#### *iv. Sampling frame*

A list of all the sampling units belonging to the population to be studied with their identification particulars or a map showing the boundaries of the sampling units is known as sampling frame. Examples of a frame are a list of farms and a list of suitable area segments like villages in India or counties in the United States. The frame should be up to date and free from errors of omission and duplication of sampling units.

#### *v. Random sample*

One or more sampling units selected from a population according to some specified procedures are said to constitute a sample. The sample will be considered as random or probability sample, if its selection is governed by ascertainable laws of chance. In other words, a random or probability sample is a sample drawn in such a manner that each unit in the population has a predetermined probability of selection. For example, if a population consists of the  $N$  sampling

units  $U_1, U_2, \dots, U_i, \dots, U_N$  then, we may select a sample of  $n$  units by selecting them unit by unit with equal probability for every unit at each draw with or without replacing the sampling units selected in the previous draws.

### *vi. Non-random sample*

A sample selected by a non-random process is termed as non-random sample. A non-random sample, which is drawn using certain amount of judgment with a view to get a representative sample, is termed as judgment or purposive sample. In purposive sampling units are selected by considering the available auxiliary information more or less subjectively with a view to ensuring a reflection of the population in the sample. This type of sampling is seldom used in large-scale surveys mainly because it is not generally possible to get strictly valid estimates of the population parameters under consideration and of their sampling errors due to the risk of bias in subjective selection and the lack of information on the probabilities of selection of the units.

### *vii. Population parameters*

Suppose a finite population consists of the  $N$  units  $U_1, U_2, \dots, U_N$  and let  $Y_i$  be the value of the variable  $y$ , the characteristic under study, for the  $i^{\text{th}}$  unit  $U_i$ , ( $i=1, 2, \dots, N$ ). For instance, the unit may be a farm and the characteristic under study may be the area under a particular crop. Any function of the values of all the population units is known as a population parameter or simply a parameter. Some of the important parameters usually required to be estimated in surveys are population total and population mean.

### *viii. Statistic, estimator and estimate*

Suppose, a sample of  $n$  units is selected from a population of  $N$  units, according to some probability scheme and let, the sample observations be denoted by  $y_1, y_2, \dots, y_n$ . Any function of these values which is free from unknown population parameters is called a statistic. An estimator is a statistic obtained by a specified procedure for estimating a population parameter. The estimator is a random variable and its value differs from sample to sample and the samples are selected with specified probabilities. The particular value, which the estimator takes for a given sample, is known as an estimate.

### *ix. Sampling and non-sampling error*

The error arises due to drawing inferences about the population on the basis of observations on a part (sample) of it, is termed sampling error. The sampling error is non-existent in a complete

enumeration survey since the whole population is surveyed. On the contrary, the errors other than sampling errors such as those arising through non-response, incompleteness and inaccuracy of response are termed non-sampling errors and are likely to be more wide-spread and important in a complete enumeration survey than in a sample survey. Non-sampling errors arise due to various causes right from the beginning stage when the survey is planned and designed to the final stage when the data are processed and analyzed.

The sampling error usually decreases with increase in sample size (number of units selected in the sample) while the non-sampling error is likely to increase with increase in sample size.

As regards the non-sampling error, it is likely to be more in the case of a complete enumeration survey than in the case of a sample survey since it is possible to reduce the non-sampling error to a great extent by using better organization and suitably trained personnel at the field and tabulation stages in the latter than in the former.

### **2. Simple Random Sampling**

Simple random sampling (SRS) can be regarded as the basic form of probability sampling applicable to situations where there is no previous information available on the population structure. Simple random sampling is a method of selecting  $n$  units out of the  $N$  such that every

one of the  $\binom{N}{n}$  distinct samples has an equal chance of being drawn. In practice a simple

random sample is drawn unit by unit. The units in the population are numbered from 1 to  $N$ . A series of random numbers between 1 and  $N$  is then drawn, either by means of a table of random numbers or by means of a computer program that produces such a table. At any draw the process used must give an equal chance of selection to any number in the population not already drawn. The units that bear these numbers constitute the sample. Since a number that has been drawn is removed from the population for all subsequent draws, this method is also called random sampling without replacement. In case of a random sampling with replacement, at any draw all  $N$  members of the population are given an equal chance of being drawn, no matter how often they have already been drawn. The with-replacement assumption simplifies the estimation under complex sampling designs and is often adopted, although in practice sampling is usually carried out under a without replacement type scheme. Obviously, the difference between with replacement and without replacement sampling becomes less important when the population size is large and the sample size is noticeably smaller than it.

## 2.1 Procedure of selecting a random sample

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

- i) Lottery method
- ii) Use of random number tables

### *i) Lottery Method:*

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

### *ii) Use of Random Number Tables:*

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

- Tippett's random number Tables
- Fisher and Yates Tables
- Kendall and Smith Tables
- A million random digits Table

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



sample through the use of random numbers is, therefore, modified and one of these modified procedures is:

- **Remainder Approach:**

Let  $N$  be an  $r$ -digit number and let its  $r$ -digit highest multiple be  $N'$ . A random number  $k$  is chosen from 1 to  $N'$  and the unit with serial number equal to the remainder obtained on dividing  $k$  by  $N$  is selected, *i.e.* the selected number is reduced mod ( $N$ ). If the remainder is zero, the last unit is selected. As an illustration, let  $N = 123$ , then highest three-digit multiple of 123 is 984. For selecting a unit, one random number from 001 to 984 has to be selected. Let the random number selected be 287. Dividing 287 by 123 gives the remainder as 41. Hence, the unit with serial number 41 is selected in the sample. Suppose that another random number selected is 245. Dividing 245 by 123 leaves 122 as remainder. So the unit bearing the serial number 122 is selected. Similarly, if the random number selected is 369, then dividing 369 by 123 leaves remainder as 0. So the unit bearing serial number 123 is selected in the sample.

## 2.2 Estimation of Population Total

Let  $Y$  be the character of interest and  $Y_1, Y_2, \dots, Y_i, \dots, Y_N$  be the values of the character from  $N$  units of the population. Further, let  $y_1, y_2, \dots, y_i, \dots, y_n$  be the sample of size  $n$  selected by simple random sampling without replacement. For the total  $Y = \sum_{i=1}^N Y_i$  we have an estimator

$$\hat{Y} = N \sum_{i=1}^n y_i / n = N \bar{y}_n$$

*i.e.*, the sample mean  $\bar{y}_n$  multiplied by the population size  $N$ .

The estimator can be expressed as

$$\hat{Y} = \sum_{i=1}^n w_i y_i = (N/n) \sum_{i=1}^n y_i, \text{ where } w_i = N/n.$$

The constant  $N/n$  is the sampling weight and is the inverse of the sampling fraction  $n/N$ .

The estimator has the statistical property of unbiasedness in relation to the sampling design.

Variance of the estimator  $\hat{Y}$  of the population total is given by

$$V_{\text{SRS}}(\hat{Y}) = \frac{N^2}{n} \left(1 - \frac{n}{N}\right) \sum_{i=1}^N (Y_i - \bar{Y})^2 / (N-1)$$

where  $\bar{Y} = \sum_{i=1}^N Y_i / N$  is the population mean and  $S^2 = \sum_{i=1}^N (Y_i - \bar{Y})^2 / (N-1)$  is the population mean square.

An unbiased estimator of variance of the estimator  $\hat{Y}$  of the total,  $V_{SRS}(\hat{Y})$  is given by

$$\begin{aligned} \hat{V}_{SRS}(\hat{Y}) &= N^2 \left(1 - \frac{n}{N}\right) \sum_{i=1}^n (y_i - \bar{y}_n)^2 / n(n-1) \\ &= N^2 \left(1 - \frac{n}{N}\right) s^2 / n \end{aligned}$$

where  $\bar{y}_n = \sum_{i=1}^n y_i / n$  is the sample mean and  $s^2$  is an unbiased estimator of the population mean square  $S^2$ .

### 3. Use of Auxiliary Information

In sampling theory if the auxiliary information, related to the character under study, is available on all the population units, then it may be advantageous to make use of this additional information in survey sampling. One way of using this additional information is in the sample selection with unequal probabilities of selection of units. The knowledge of auxiliary information may also be exploited at the estimation stage. The estimator can be developed in such a way that it makes use of this additional information. Ratio estimator, difference estimator, regression estimator, generalized difference estimators are the examples of such estimators. Obviously, it is assumed that the auxiliary information is available on all the sampling units. In case the auxiliary information is not available then it can be obtained easily without much burden on the cost.

Another way the auxiliary information can be used is at the stage of planning of survey. An example of this is the stratification of the population units by making use of the auxiliary information.

### 4. Sampling with Varying Probability

Under certain circumstances, selection of units with unequal probabilities provides more efficient estimators than equal probability sampling, and this type of sampling is known as unequal or varying probability sampling. In the most commonly used varying probability sampling scheme, the units are selected with probability proportional to a given measure of size (PPS) where the size measure is the value of an auxiliary variable  $x$  related to the characteristic  $y$  under study and this sampling scheme is termed as probability proportional to

size sampling. For instance, the number of persons in some previous period may be taken as a measure of the size in sampling area units for a survey of socio-economic characters, which are likely to be related to population. Similarly, in estimating crop characteristics the geographical area or cultivated area for a previous period, if available, may be considered as a measure of size, or in an industrial survey, the number of workers may be taken as the size of an industrial establishment.

Since a large unit, that is, a unit with a large value for the study variable  $y$ , contributes more to the population total than smaller units, it is natural to expect that a scheme of selection which gives more chance of inclusion in a sample to larger units than to smaller units would provide estimators more efficient than equal probability sampling. Such a scheme is provided by pps sampling, size being the value of an auxiliary variable  $x$  directly related to  $y$ . It may appear that such a selection procedure would give biased estimators as the larger units are over-represented and the smaller units are under-represented in the sample. This would be so, if the sample means is used as an estimator of population mean. Instead, if the sample observations are suitably weighted at the estimation stage taking into consideration their probabilities of selection, it is possible to obtain unbiased estimators. Mahalanobis (1938) has referred to this procedure in the context of sampling plots for a crop survey and this procedure has been discussed in detail by Hansen and Hurwitz (1943).

### **5. Stratified Random Sampling**

The basic idea in stratified random sampling is to divide a heterogeneous population into sub-populations, usually known as strata, each of which is internally homogeneous in which case a precise estimate of any stratum mean can be obtained based on a small sample from that stratum and by combining such estimates, a precise estimate for the whole population can be obtained. Stratified sampling provides a better cross section of the population than the procedure of simple random sampling. It may also simplify the organization of the field work. Geographical proximity is sometimes taken as the basis of stratification. The assumption here is that geographically contiguous areas are often more alike than areas that are far apart. Administrative convenience may also dictate the basis on which the stratification is made. For example, the staff already available in each range of a forest division may have to supervise the survey in the area under their jurisdiction. Thus, compact geographical regions may form the strata. If the characteristic under study is known to be correlated with a supplementary variable for which actual data or at least good estimates are available for the units in the population, the stratification may be done using the information on the supplementary variable.

## Overview of Sampling Methods

For instance, the volume estimates obtained at a previous inventory of the forest area may be used for stratification of the population.

In stratified sampling, the variance of the estimator consists of only the 'within strata' variation. Thus the larger the number of strata into which a population is divided, the higher, in general, the precision, since it is likely that, in this case, the units within a stratum will be more homogeneous. For estimating the variance within stratum, there should be a minimum of 2 units in each stratum. The larger the number of strata the higher will, in general, be the cost of enumeration. So, depending on administrative convenience, cost of the survey and variability of the characteristic under study in the area, a decision on the number of strata will have to be arrived at.

### **6. Cluster Sampling**

A sampling procedure presupposes division of the population into a finite number of distinct and identifiable units called the sampling units. The smallest units into which the population can be divided are called the elements of the population, and group of elements the clusters. A cluster may be a class of students or cultivators' fields in a village. When the sampling unit is a cluster, the procedure of sampling is called cluster sampling.

For many types of population a list of elements is not available and the use of an element as the sampling unit is, therefore, not feasible. The method of cluster or area sampling is available in such cases. Thus, in a city a list of all the houses may be available, but that of persons is rarely so. Again, list of farms are not available, but those of villages or enumeration districts prepared for the census are. Cluster sampling is, therefore, widely practiced in sample surveys.

For a given number of sampling units cluster sampling is more convenient and less costly than simple random sampling due to the saving time in journeys, identification and contacts etc., but cluster sampling is generally less efficient than simple random sampling due to the tendency of the units in a cluster to be similar. In most practical situations, the loss in efficiency may be balanced by the reduction in the cost and the efficiency per unit cost may be more in cluster sampling as compares to simple random sampling.

### **7. Multistage Sampling**

Cluster sampling is a sampling procedure in which clusters are considered as sampling units and all the elements of the selected clusters are enumerated. One of the main considerations of adopting cluster sampling is the reduction of travel cost because of the nearness of elements in the clusters. However, this method restricts the spread of the sample over population which

## Overview of Sampling Methods

results generally in increasing the variance of the estimator. In order to increase the efficiency of the estimator with the given cost it is natural to think of further sampling the clusters and selecting more number of clusters so as to increase the spread of the sample over population. This type of sampling which consists of first selecting clusters and then selecting a specified number of elements from each selected cluster is known as sub-sampling or two stage sampling, since the units are selected in two stages. In such sampling designs, clusters are generally termed as first stage units (fsu's) or primary stage units (psu's) and the elements within clusters or ultimate observational units are termed as second stage units (ssu's) or ultimate stage units (usu's). It may be noted that this procedure can be easily generalized to give rise to multistage sampling, where the sampling units at each stage are clusters of units of the next stage and the ultimate observational units are selected in stages, sampling at each stage being done from each of the sampling units or clusters selected in the previous stage. This procedure, being a compromise between uni-stage or direct sampling of units and cluster sampling, can be expected to be (i) more efficient than uni-stage sampling and less efficient than cluster sampling from considerations of operational convenience and cost, and (ii) less efficient than uni-stage sampling and more efficient than cluster sampling from the view point of sampling variability, when the sample size in terms of number of ultimate units is fixed.

It may be mentioned that multistage sampling may be the only feasible procedure in a number of practical situations, where a satisfactory sampling frame of ultimate observational units is not readily available and the cost of obtaining such a frame is prohibitive or where the cost of locating and physically identifying the usu's is considerable. For instance, for conducting a socio-economic survey in a region, where generally household is taken as the usu, a complete and up-to-date list of all the households in the region may not be available, whereas a list of villages and urban blocks which are group of households may be readily available. In such a case, a sample of villages or urban blocks may be selected first and then a sample of households may be drawn from each selected village and urban block after making a complete list of households. It may happen that even a list of villages is not available, but only a list of all tehsils (group of villages) is available. In this case a sample of households may be selected in three stages by selecting first a sample of tehsils, then a sample of villages from each selected tehsil after making a list of all the villages in the tehsil and finally a sample of households from each selected village after listing all the households in it. Since the selection is done in three stages, this procedure is termed as three stage sampling. Here, tehsils are taken as first stage

units (fsu's), villages as second stage units (ssu's) and households as third or ultimate stage units (tsu's).

### **8. Systematic Sampling**

In all other sampling methods, the successive units (whether elements or clusters) are selected with the help of random numbers. But a method of sampling in which only the first unit is selected with the help of random number while the rest of the units are selected according to a pre-determined pattern, is known as systematic sampling. The systematic sampling has been found very useful in forest surveys for estimating the volume of timber, in fisheries surveys for estimating the total catch of fish, in milk yield surveys for estimating the lactation yield etc.

### **9. Conclusion**

Simple random sampling and probability proportional size designs are most important uni-stage design. In most of the practical situations, complex sampling designs are utilized on the basis of these uni-stage sampling designs. Stratified random sampling, multistage sampling, multiphase sampling, etc. are efficient complex designs widely used in agricultural and socio-economic surveys.

### **References**

- Cochran, W.G. (1977). *Sampling techniques*. Wiley Eastern Ltd.
- Des Raj, (1968). *Sampling theory*. Tata-Mcgraw-Hill Publishing Company Ltd.
- Hansen, M.H. and Hurwitz, W.H. (1943). On the theory of sampling from finite populations. *Ann. Math. Statist.*, **14**, 333-362.
- Hansen, M.H., Hurwitz, W.H. and Madow, W.G. (1993). *Sample survey methods and theory*. Vol. 1 and Vol. 2, John Wiley & Sons, Inc.
- Murthy, M.N. (1977). *Sampling theory and methods*. Statistical Publishing Society.
- Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S. and Ashok, C. (1984). *Sampling theory of surveys with applications*. Indian Society of Agricultural Statistics.

---

# DESIGN RESOURCES SERVER<sup>1</sup>

(<https://drs.icar.gov.in>)

---

**Rajender Parsad and V.K. Gupta**

*ICAR-Indian Agricultural Statistics Research Institute*

*Library Avenue, New Delhi - 110 012*

[rajender.parsad@icar.gov.in](mailto:rajender.parsad@icar.gov.in) , [vkgupta1751@yahoo.co.in](mailto:vkgupta1751@yahoo.co.in)

---

## 1. Introduction

Design Resources Server is developed to popularize and disseminate the research in Design of Experiments among the scientists of National Agricultural Research System (NARS) in particular and researchers all over the globe in general and is hosted at <https://drs.icar.gov.in>. The home page of the server is



Design Resources Server is matter-of-factly a virtual, mobile library on design of experiments created with an objective to advise and help the experimenters in agricultural sciences, biological sciences, animal sciences, social sciences and industry in planning and designing their experiments for making precise and valid inferences on the problems of their interest. This also provides support for analysis of data generated so as to meet the objectives of the study. The server also aims at providing a platform to the researchers in design of experiments for disseminating research and also strengthening research in newer emerging areas so as to meet the challenges of agricultural research. The purpose of this server is to spread advances in theoretical, computational, and statistical aspects of Design of Experiments among the

---

<sup>1</sup> With the active help and support from A. Dhandapani, Alka Arora, Rakesh Saini and Subhash Chand. The details of contributors can be seen at respective links.

mathematicians and statisticians in academia and among the practicing statisticians involved in advisory and consultancy services.

This server works as an e-advisory resource for the experimenters. The actual layout of the designs is available to the experimenters online and the experimenter can use these designs for their experimentation. It is expected that the material provided at this server would help the experimenters in general and agricultural scientists in particular in improving the quality of research in their respective sciences and making their research globally competitive.

Design Server is open to everyone from all over the globe. Anyone can join this and add information to the site to strengthen it further with the permission of the developers. The Server contains a lot of useful information for scientists of NARS. The material available on the server has been partitioned into 4 components:

- **Useful for Experimenters:** Electronic Books, online generation of randomized layout of designs, online analysis of data, analysis of data using various softwares, statistical genomics.
- **Useful for Statisticians:** Literature and catalogues of BBB designs, designs for making test treatments-control treatment comparisons, designs for bioassays, designs for factorial experiments (supersaturated designs, block designs with factorial treatment structure), experiments with mixtures, Online generation of Hadamard matrices, MOLS and orthogonal arrays.
- **Other Useful Links:** Discussion Board, Ask a Question, Who-is-where, important links.
- **Site Information:** Feedback, How to Quote Design Resources Server, Copyright, disclaimer, contact us and site map.

The major components are Useful for Experimenters and Research Statisticians. The scientists, however, can use either of the parts or parts of their choice. A brief description of all the above four components is given in the sequel.

## 2. Useful for Experimenters

This link has been designed essentially to meet the requirements of the experimenters whose prime interest is in designing the experiment and then subsequently analyzing the data generated so as to draw statistically valid inferences. To meet this end, the link contains the following sub-links:

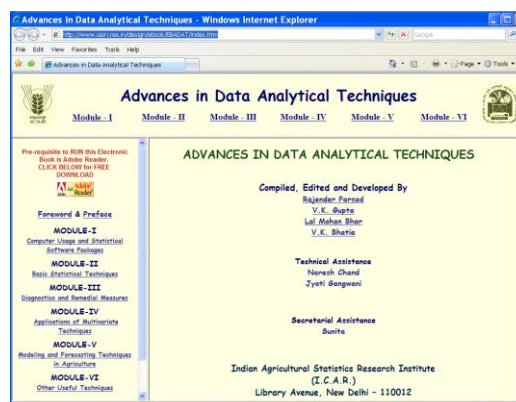
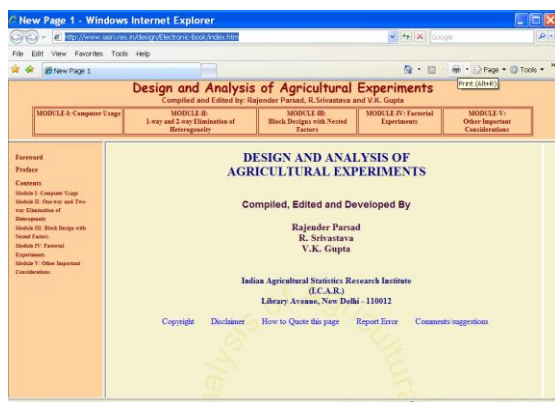
### 2.1 E-Learning

This is an important link that provides useful and important reading material on use of some statistical software packages, designing experiments, statistical analysis of data and other useful topics in statistics in the form of two electronic books viz.

1. Design and Analysis of Agricultural Experiments  
<https://drs.icar.gov.in/Electronic-Book/index.htm>
2. Advances in Data Analytical Techniques  
<https://drs.icar.gov.in/ebook/EBADAT/index.htm>

The screen shots of cover pages of these books are shown below:





The coverage of topics in these electronic books is very wide and almost all the aspects of designing an experiment and analysis of data are covered. The chapters are decorated with solved examples giving the steps of analysis. The users can have online access to these electronic books. This provides good theoretical support and also reading material to the users.

## 2.2 Online Design Generation-I

This link is very useful for experimenters because it helps in generation of randomized layout of the following designs:

**Basic Designs:** Generates of randomized layout of completely randomized design and randomized complete block design both for single factor and multifactor experiments and Latin square designs for single factor experiments. The field book can be created as a .csv file or a text file. This is available at

[https://drs.icar.gov.in/Basic Designs/generate\\_designs.htm](https://drs.icar.gov.in/Basic Designs/generate_designs.htm).

**Augmented Designs:** A large number of germplasm evaluation trials are conducted using augmented designs. The experimenters generally compromise with the randomization of treatments in the design. Further, experimenters also need to know the optimum replication number of controls in each block so as to maximize the efficiency per observation. Online software for generation of randomized layout of an augmented randomized complete block design for given number of test treatments, control treatments and number of blocks with given block sizes, not necessarily equal, is developed and is available at

<https://drs.icar.gov.in/Augmented Designs/home.htm>.

The design can be generated with optimum replication of control treatments in each block so as to maximize efficiency per observation.

**Resolvable Block Designs:** Resolvable block designs are an important class of incomplete block designs wherein the blocks can be formed together into sets with the blocks within each set constituting a complete replication. In the class of resolvable block designs, square lattice designs are very popular among experimenters. One can generate square lattice designs with three replications using

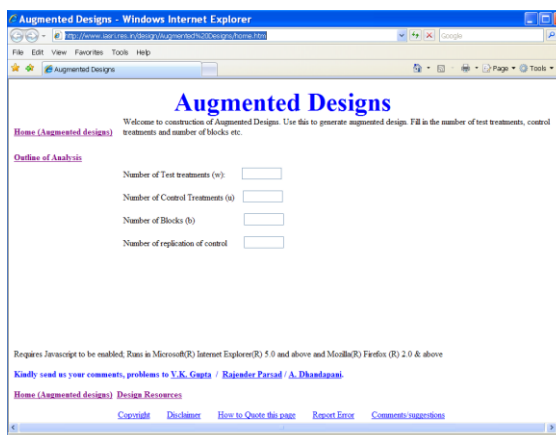
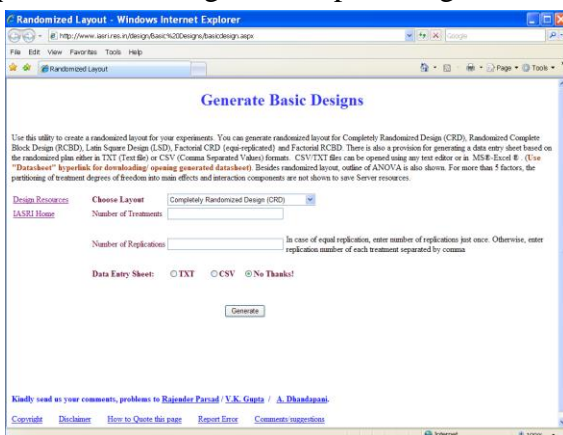
[https://drs.icar.gov.in/WebHadamard/square\\_lattice.htm](https://drs.icar.gov.in/WebHadamard/square_lattice.htm).

Another important class of resolvable block designs is the alpha designs. These designs are available when the number of treatments is a composite number. Literature on alpha designs is available at

<https://drs.icar.gov.in/Alpha/Home.htm>.

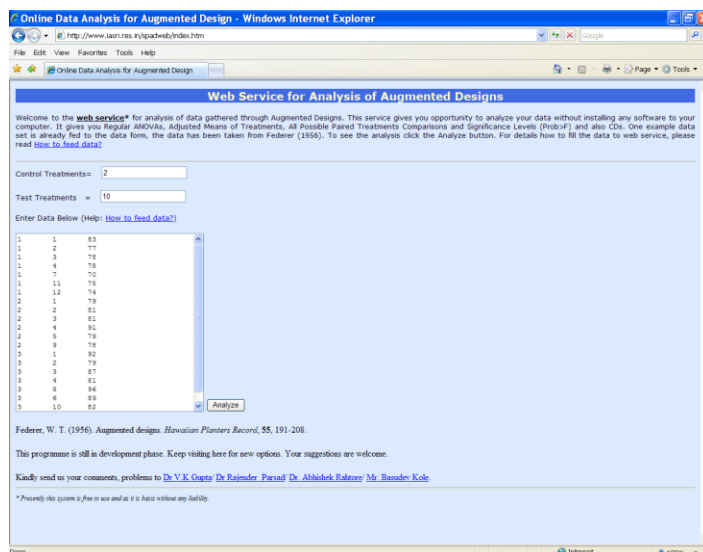
This link also provides randomized layout of alpha designs for  $6 \leq v (=sk, \text{the number of treatments}) \leq 150$ ,  $2 \leq r$  (number of replications)  $\leq 5$ ,  $3 \leq k$  (block size)  $\leq 10$  and  $2 \leq s \leq 15$  along with the lower bounds to A- and D- efficiencies of the designs.

The screen shots for generation of randomized layout of basic designs, augmented designs, square lattice designs and alpha designs are



### 2.3 Online Analysis of Data

This link together with Analysis of Data forms the backbone of the Design Resources Server. This particular link targets at providing online analysis of data generated to the experimenter. At present an experimenter can perform online analysis of data generated from augmented randomized block designs. This is available at <https://drs.icar.gov.in/spadweb/index.htm>.



## 2.4 Analysis of Data

This is the most important link of the server because it targets at providing steps of analysis of data generated from designed experiments using several statistical packages like SAS, SPSS, GenStat, MINITAB, SYSTAT, SPAD, SPFE, SPAR 2.0, MS-Excel, etc. Some real life examples of experiments are given and the questions to be answered are listed. Steps for preparation of data files, the commands and macros to be used for analysis of data and the treatment contrasts to be used for answering specific questions, etc. are given, which the user can use without any difficulty. The data files and result files can also be downloaded. This is available at

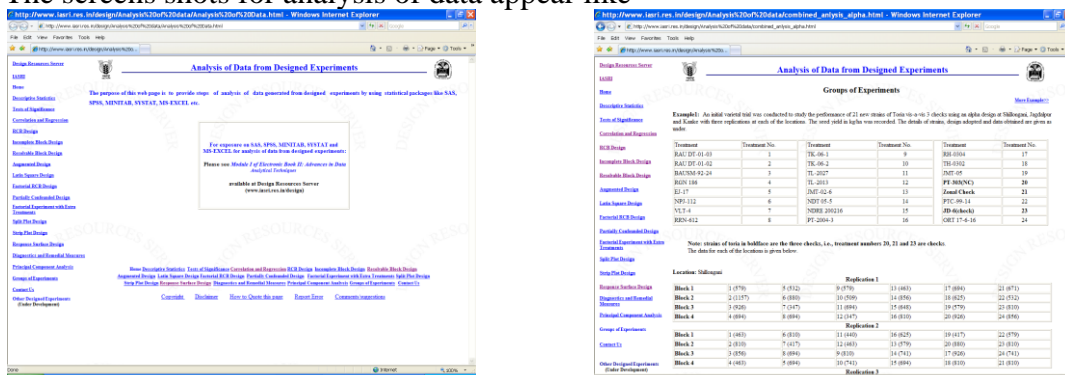
<https://drs.icar.gov.in/Analysis> of data/Analysis of Data.html.

The following analysis can be performed using this link:

- Analysis of data generated from completely randomized designs, randomized complete block design; incomplete block design; resolvable incomplete block design; Latin square design; factorial experiments both without and with confounding; factorial experiments with extra treatments; split and strip plot designs; cross over designs using SAS and SPSS; steps of analysis of augmented design using SAS, SPSS and SPAD
- Response surface design using SAS and SPSS
- SAS code for analysis of groups of experiments conducted in different environments (locations or season / year), each experiment conducted as a complete block or an incomplete block design. Using this code, one can analyze the data for each of the environments separately, test the homogeneity of error variances using Bartlett's  $\chi^2$ -test, perform combined analysis of data considering both environment effects as fixed and environment effects as random (both through PROC GLM and PROC MIXED) and prepare site regression or GGE biplots
- SAS Macro for performing diagnostics (normality and homogeneity of errors) in experimental data generated through randomized complete block designs and then applying remedial measures such as Box-Cox transformation and applying the non-parametric tests if the errors remain non-normal and / or heterogeneous even after transformation

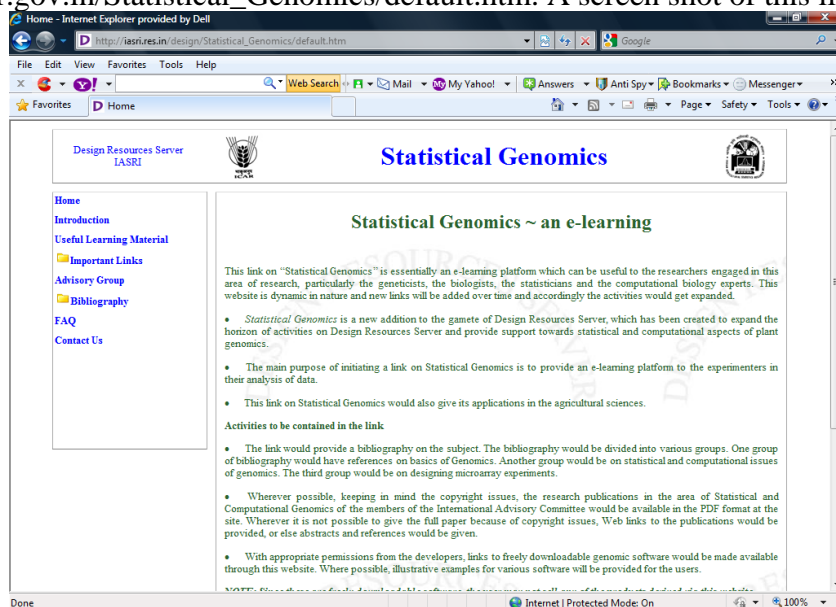
- SAS codes are also available for obtaining descriptive statistics, generating discrete frequency distribution, grouped frequency distribution, histogram, testing the normality of a given variable (overall groups or for each of the groups separately)
- correlation and regression using SAS and SPSS
- Tests of significance based on Student's *t*-distribution using SAS, SPSS and MS-EXCEL
- SAS and SPSS codes for performing principal component analysis, cluster analysis and analysis of covariance
- SAS and SPSS codes for fitting non-linear models

The screens shots for analysis of data appear like

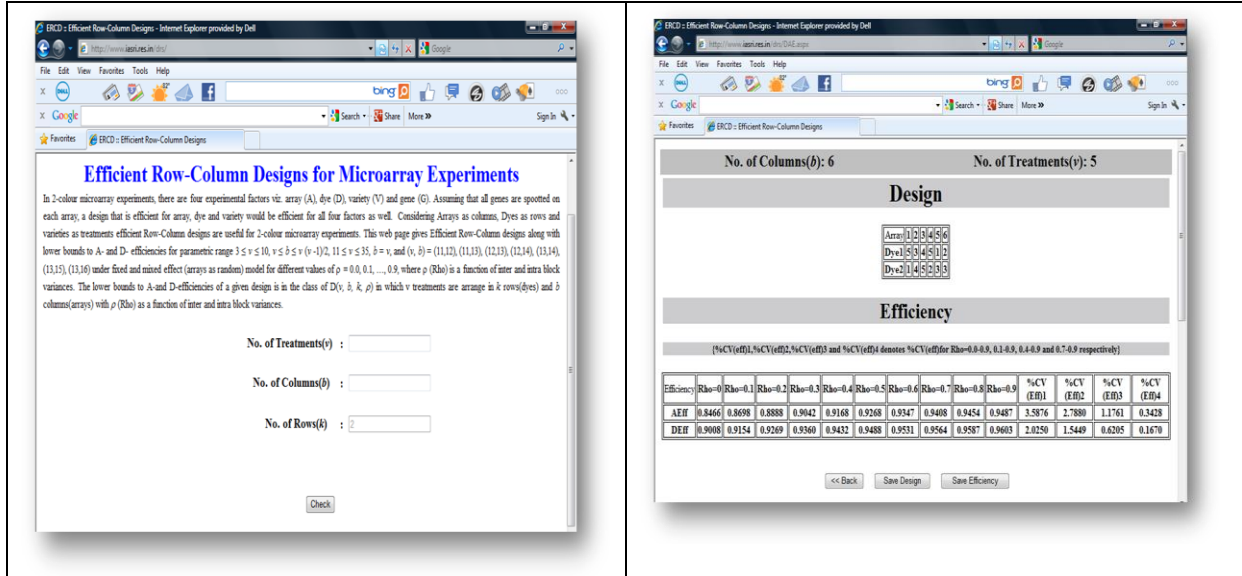


## 2.6 Statistical Genomics

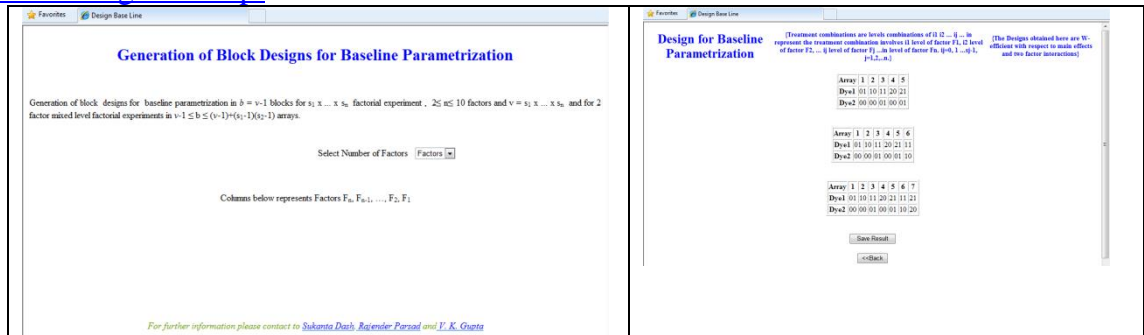
A link on Statistical Genomics has been initiated essentially as an e-learning platform which can be useful to the researchers particularly the geneticists, the biologists, the statisticians and the computational biology experts. It contains the information on some public domain softwares that can be downloaded free of cost. A bibliography on design and analysis of microarray experiments is also given. These are hosted at [https://drs.icar.gov.in/Statistical\\_Genomics/default.htm](https://drs.icar.gov.in/Statistical_Genomics/default.htm). A screen shot of this link is



A new link ‘Catalogue and Generation of Row-Column Designs’ (<https://drs.icar.gov.in/drs/>) has been initiated for generation of Row-Column Designs with two rows along with lower bounds to A- and D- efficiencies for parametric range  $3 \leq v \leq 10, v \leq b \leq v(v-1)/2, 11 \leq v \leq 35, b = v$ , and  $(v, b) = (11,12), (11,13), (12,13), (12,14), (13,14), (13,15), (13,16)$  under fixed and mixed effects models.



Another link on online generation of block designs with block size 2 for factorial experiments with baseline parameterization in  $b = v-1$  blocks (where  $v$  is the number of treatment combinations  $v = s_1 \times s_2 \times \dots \times s_n, 2 \leq n \leq 10$  factors and  $v = s_1 \times s_2 \times \dots \times s_n$  and for 2 factor mixed level factorial experiments in  $v-1 \leq b \leq (v-1) + (s_1-1)(s_2-1)$  arrays and made available at <https://drs.icar.gov.in/dbp/>. Some screen shots are



## 2.7 Modules based on R-Software

A new application for R Package based modules have been developed at <http://drsr.icar.gov.in/> Following modules of online generation of designs are available

### Single Factor Experiments

For generation of (i) Balanced Incomplete Latin Square Designs; (ii) Incomplete block designs up to 30 treatments and up to block size 10 and (iii) Position Balanced Block Designs.

### Factorial Experiments

Generation of (a) Orthogonal and Nested Orthogonal Latin Hypercube Designs viz. (i) 1st order OLH design; (ii) 2nd order OLH design; (iii) Nested OLH design and (iv) OLH design with good



space filling property and (b) Incomplete Split Plot Designs for three situations namely (i) when blocks are complete with respect to whole plot treatments and whole plots are incomplete with respect to subplot treatments, (ii) when blocks are incomplete with respect to whole plot treatments and whole plots are complete with respect to subplot treatments and (iii) when blocks are incomplete with respect to whole plot treatments and whole plots are incomplete with respect to subplot treatments.

### 3. Useful for Research Statisticians

This link is useful for researchers engaged in conducting research in design of experiments and can be used for class room teaching also. The material on this link is divided into the following sub-links:

#### 3.1 Block Designs

This link provides some theoretical considerations of balanced incomplete block (BIB) designs, binary variance balanced block (BBB) designs with 2 and 3 distinct block sizes, partially balanced incomplete block (PBIB) designs, designs for test treatments-control treatment(s) comparisons, etc. for research statisticians. The link also gives a catalogue of designs and a bibliography on the subject for use of researchers. At present the following material is available on this link:

- General method of construction of BBB designs; general methods of construction of block designs for making test treatments - control treatment(s) comparisons; bibliography
- Catalogue of BIB designs for number of replications  $r \leq 30$  for symmetric BIB designs and  $r \leq 20$  for asymmetric BIB designs
- Catalogue of BBB designs with 2 and 3 distinct block sizes for number of replications  $r \leq 30$ . The catalogue also gives the resolvability status of the designs along with the efficiency factor of the designs
- 6574 block designs for making all possible pair wise treatment comparisons for  $v \leq 35$  (number of treatments),  $b \leq 64$  (number of blocks),  $k \leq 34$  (block size)

Some screen shots on block designs are given below:

The screenshot shows the 'Design Resources Server' website. The main content area displays a 'Catalogue of BIB design for r=30 for symmetric and r=20 for asymmetric data'. Below the title, there is a search criteria section and a table of designs. The table has columns for S, v, b, r, k, A, m, EF, Type, and Resolvability. The table lists 19 different design configurations with their respective parameters and efficiency factors.

S	v	b	r	k	A	m	EF	Type	Resolvability
1	3	3	2	2	1	6	0.75	ur	
2	3	6	4	2	2	12	0.75	Rep02(1)	RESO 2
3	3	9	6	2	3	18	0.75	Rep03(1)	RESO 2
4	3	12	8	2	4	24	0.75	Rep04(1)	RESO 2
5	3	15	10	2	5	30	0.75	Rep05(1)	RESO 2
6	3	18	12	2	6	36	0.75	Rep06(1)	RESO 2
7	3	21	14	2	7	42	0.75	Rep07(1)	RESO 2
8	3	24	16	2	8	48	0.75	Rep08(1)	RESO 2
9	3	27	18	2	9	54	0.75	Rep09(1)	RESO 2
10	3	30	20	2	10	60	0.75	Rep10(1)	RESO 2
11	4	6	3	2	1	12	0.6667	8b2	RESO 1
12	4	12	6	2	2	24	0.6667	Rep02(11)	RESO 1
13	4	18	9	2	3	36	0.6667	Rep03(11)	RESO 1
14	4	24	12	2	4	48	0.6667	Rep04(11)	RESO 1
15	4	30	15	2	5	60	0.6667	Rep05(11)	RESO 1
16	4	36	18	2	6	72	0.6667	Rep06(11)	RESO 1
17	4	4	1	3	2	12	0.8333	ur	
18	4	8	6	3	4	24	0.8333	Rep02(17)	RESO 3
19	4	12	9	3	6	36	0.8333	Rep03(17)	RESO 3

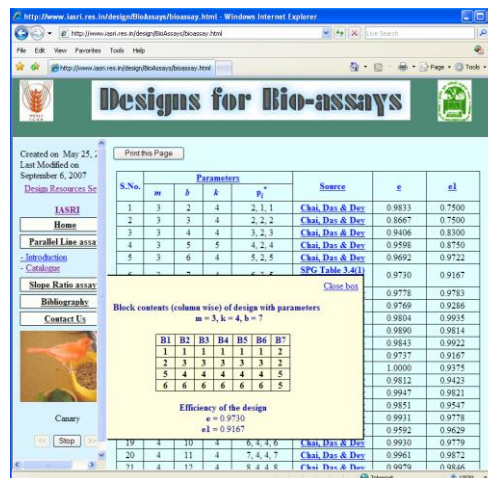
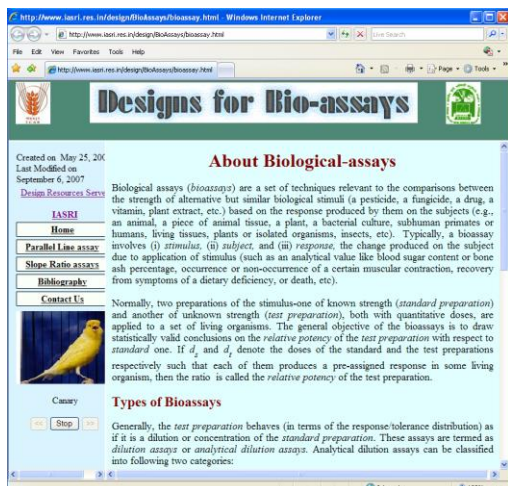
#### 3.2 Designs for Bioassays

Designs for biological assays help in estimation of relative potency of the test preparation with respect to the standard one. The material uploaded includes contrasts of interest in parallel line assays and slope ratio assays. This link provides some theoretical considerations

The screenshot shows the 'Block Designs for Making All Possible Pairwise Treatment Comparisons' website. The page has a title bar and a main heading. Below the heading, there are input fields for 'No. of Treatments (v)', 'No. of Blocks (b)', and 'Block Size (k)'. There is a 'Display Design' button. A large text box contains a detailed description of the page's purpose and the algorithm used to generate the designs. At the bottom, there are links for 'Copyright', 'Disclaimer', 'How to Order this page', 'Report Error', and 'Comments/Feedback'.

of designs for bioassays along with a catalogue of designs and a bibliography on the subject for use of researchers. Literature on bioassays is available at <https://drs.icar.gov.in/BioAssays/bioassay.html>.

Some screen shots of this link are displayed below:



### 3.3 Designs for Factorial Experiments

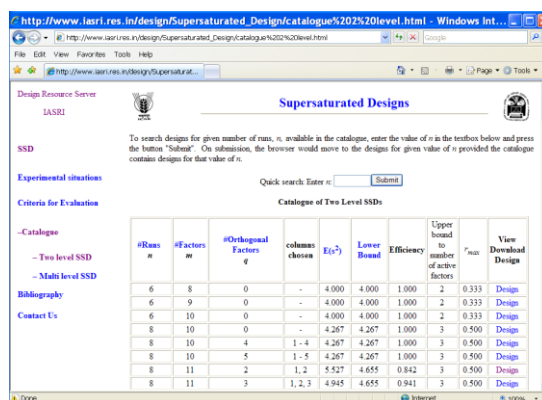
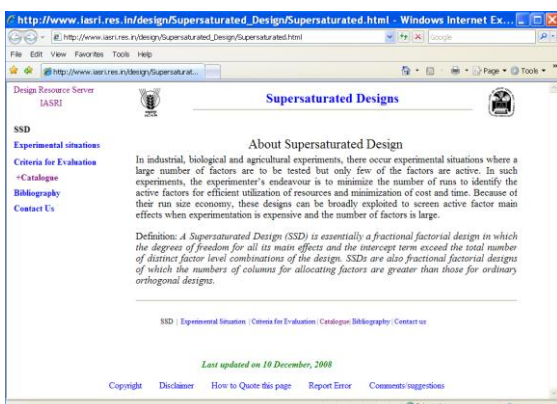
Factorial experiments are most popular among agricultural scientists. To begin with, material on block designs with factorial treatment structure and supersaturated designs is available on this link.

#### ➤ Supersaturated Designs

Supersaturated designs are fractional factorial designs in which the degrees of freedom for all its main effects and the intercept term exceed the total number of distinct factor level combinations of the design. These designs are useful when the experimenter is interested in identifying the active factors through the experiment and experimental resources are scarce. Definition of supersaturated designs, experimental situations in which supersaturated designs are useful, efficiency criteria for evaluation of supersaturated designs, catalogue of supersaturated designs for 2-level factorial experiments and asymmetrical factorial experiments and bibliography on supersaturated designs has been uploaded on the Server. The complete details of the runs can be obtained by clicking on the required design in the catalogue.

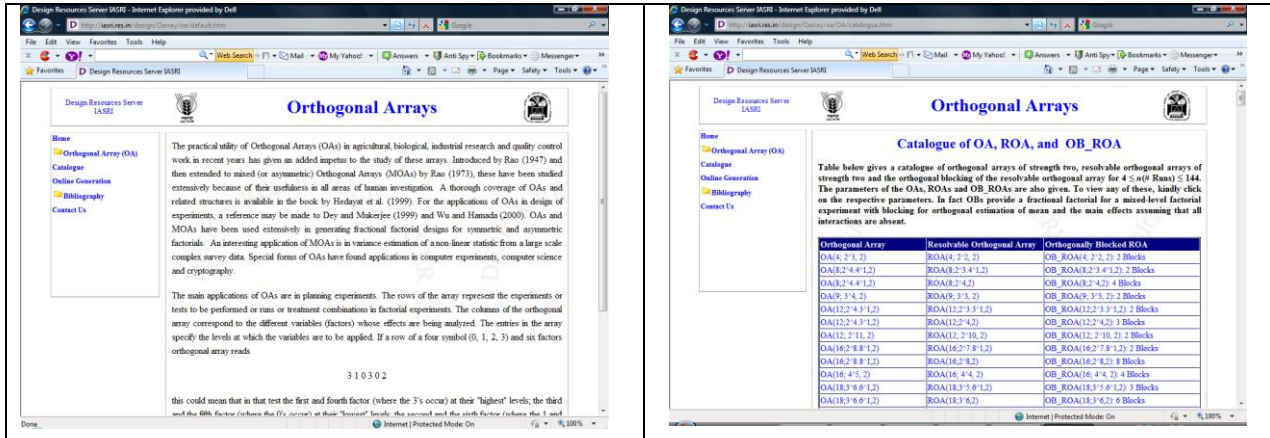
[https://drs.icar.gov.in/Supersaturated\\_Design/Supersaturated.html](https://drs.icar.gov.in/Supersaturated_Design/Supersaturated.html).

Some screen shots of supersaturated designs are



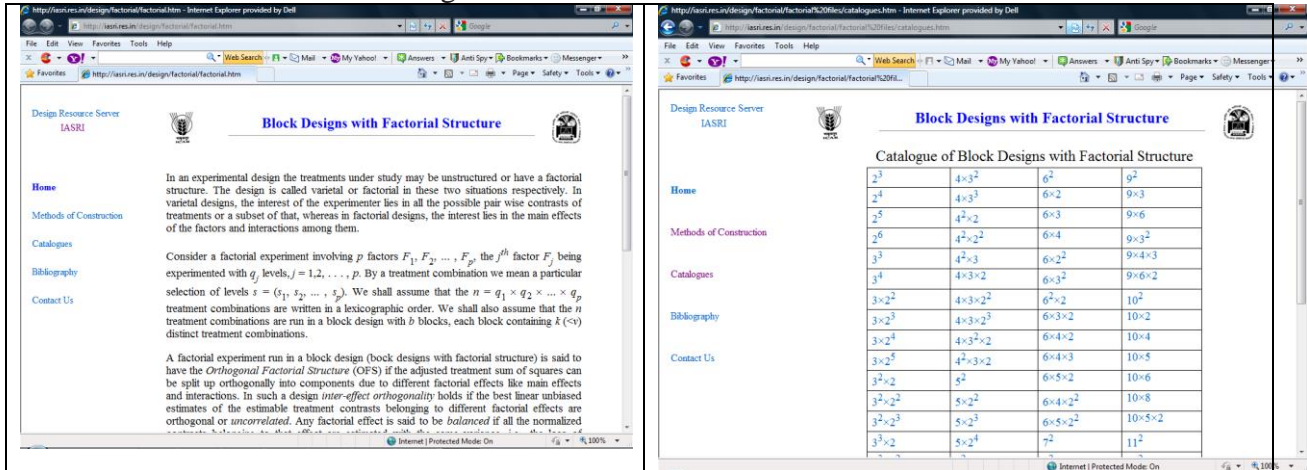
➤ **Mixed Orthogonal arrays**

Definitions of Orthogonal arrays(OAs), mixed OA, Resolvable OA,  $\alpha$ -resolvable OA, resolvable MOA, construction of OAs, blocking in OAs, generation of orthogonal arrays of strength two, resolvable orthogonal arrays of strength two and the orthogonal blocking of the resolvable orthogonal array for  $4 \leq n(\# \text{ Runs}) \leq 144$ , and bibliography on OAs.



➤ **Block Designs with Factorial Treatment Structure**

Block designs with factorial treatment structure have useful applications in designs for crop sequence experiments. The link on block designs with factorial Treatment Structure provides a bibliography with 232 references on the subject. Catalogues of block designs with factorial treatment structure in 3-replications for number of levels for any factor at most 12 permitting estimation of main effects with full efficiency and controlling efficiency for interaction effects are also given at this link. URL for this link is <https://drs.icar.gov.in/factorial/factorial.htm>. Some screen shots for block designs with factorial treatment structure are



➤ **Row-Column Designs in 2 Rows for Orthogonal parameterization**

Row-column designs are useful for the experimental situations in which there are two cross classified sources of heterogeneity in the experimental material. In many experimental situations due to practical considerations, it may not be possible to accommodate more than two experimental units in a column of a row-column design. Row-column designs with two rows and with factorial treatment structure have also been found useful in many agricultural



experimental situations including two-colour microarray experiments. When the design is non-orthogonal in a row-column set up, it would be desirable that it permits orthogonal estimation of all factorial effects with high efficiency. This may require a large number of columns. Due to cost and time considerations, it may not be possible to run a design in number of runs that are required for orthogonal estimation of all the factorial effects. The experimenter may, however, be interested in orthogonal estimation of all the main effects and two factor interactions based on an orthogonal parameterization. A module for on-line generation of Row Column Designs with equal replication of each treatment for Factorial Experiments in Two Rows for  $2^n$  ( $n < 10$ ) factorial experiments for orthogonal estimation of main effects and two factor interactions. Here, for each factor, two designs are generated, one in which main effects are estimated with more efficiency. Online generation module also generates designs with unequal replications in the same parametric range for orthogonal estimation of main effects and two factor interactions. Here one can obtain designs with fewer number of columns as compared to the minimum number of replications required for orthogonal estimation of main effects and two factor interactions in equi-replicated designs. It is made available at

[https://drs.icar.gov.in/Row\\_Column\\_design\\_OP\\_2\\_rows/Default.aspx](https://drs.icar.gov.in/Row_Column_design_OP_2_rows/Default.aspx).

Some screen shots for row-column designs in two rows for orthogonal parameterisation are given in the sequel

**Summary of Design requirements for  $2^7$  (2, 5 x 3) factorial experiments in Row-Column set up with two rows**

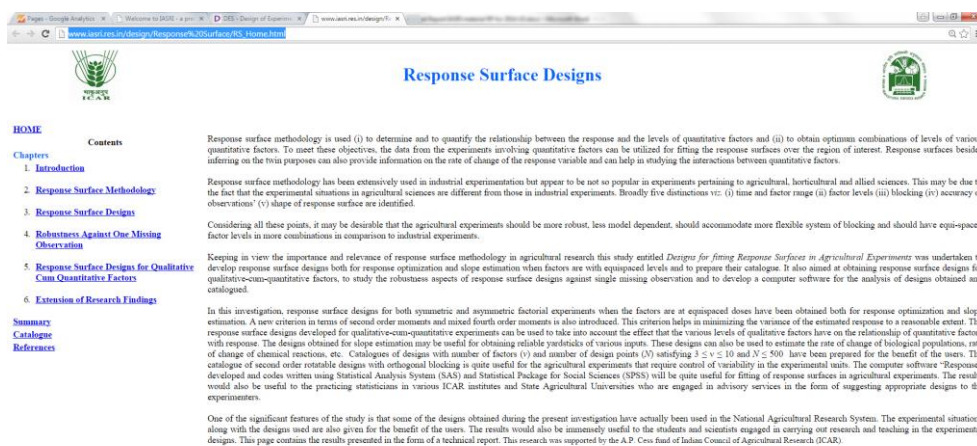
Number of Factorial Effects	2	3	4	5	6	7	8
Minimum number of Replicates	2	2	3	3	3	3	4
Number of Columns	2 <sup>n</sup>	4	8	24	48	96	192

Level	00000	00001	00010	00011	00100	00101	00110	00111	01000	01001	01010	01011	01100	01101	01110	01111
Row 1	00000	00001	00010	00011	00100	00101	00110	00111	01000	01001	01010	01011	01100	01101	01110	01111
Row 2	00000	00010	00001	00011	00100	00110	00101	00111	01000	01010	01001	01011	01100	01110	01101	01111
Row 3	00000	00010	00001	00011	00100	00110	00101	00111	01000	01010	01001	01011	01100	01110	01101	01111
Row 4	00000	00010	00001	00011	00100	00110	00101	00111	01000	01010	01001	01011	01100	01110	01101	01111
Row 5	00000	00010	00001	00011	00100	00110	00101	00111	01000	01010	01001	01011	01100	01110	01101	01111
Row 6	00000	00010	00001	00011	00100	00110	00101	00111	01000	01010	01001	01011	01100	01110	01101	01111
Row 7	00000	00010	00001	00011	00100	00110	00101	00111	01000	01010	01001	01011	01100	01110	01101	01111
Row 8	00000	00010	00001	00011	00100	00110	00101	00111	01000	01010	01001	01011	01100	01110	01101	01111
Row 9	00000	00010	00001	00011	00100	00110	00101	00111	01000	01010	01001	01011	01100	01110	01101	01111
Row 10	00000	00010	00001	00011	00100	00110	00101	00111	01000	01010	01001	01011	01100	01110	01101	01111
Row 11	00000	00010	00001	00011	00100	00110	00101	00111	01000	01010	01001	01011	01100	01110	01101	01111
Row 12	00000	00010	00001	00011	00100	00110	00101	00111	01000	01010	01001	01011	01100	01110	01101	01111
Row 13	00000	00010	00001	00011	00100	00110	00101	00111	01000	01010	01001	01011	01100	01110	01101	01111
Row 14	00000	00010	00001	00011	00100	00110	00101	00111	01000	01010	01001	01011	01100	01110	01101	01111
Row 15	00000	00010	00001	00011	00100	00110	00101	00111	01000	01010	01001	01011	01100	01110	01101	01111
Row 16	00000	00010	00001	00011	00100	00110	00101	00111	01000	01010	01001	01011	01100	01110	01101	01111

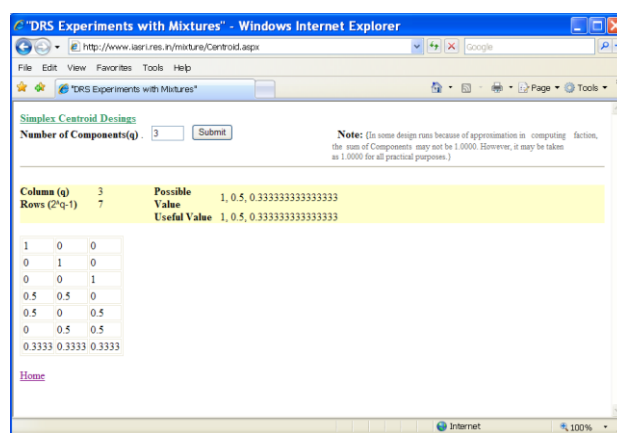
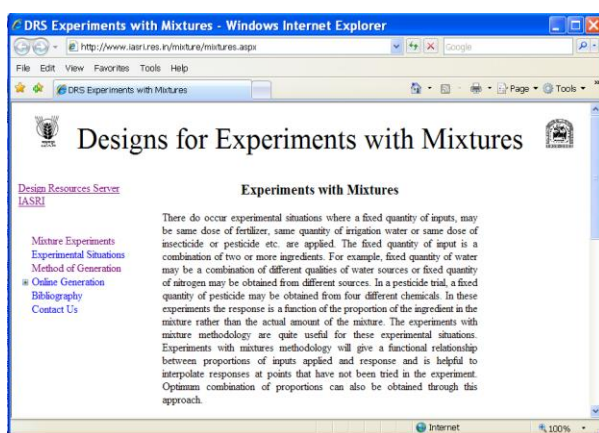
### 3.4 Response Surface Designs

These designs are used (i) to determine and to quantify the relationship between the response and the levels of quantitative factors and (ii) to obtain optimum combinations of levels of various quantitative factors. To meet these objectives, the data from the experiments involving quantitative factors can be utilized for fitting the response surfaces over the region of interest. Response surfaces besides inferring on the twin purposes can also provide information on the rate of change of the response variable and can help in studying the interactions between quantitative factors. The literature, layout of designs and bibliography on this topic is made available at [https://drs.icar.gov.in/Response%20Surface/RS\\_Home.html](https://drs.icar.gov.in/Response%20Surface/RS_Home.html) A screen shot of this link is as given below:



### 3.5 Experiments with Mixtures

Experiments with mixtures are quite useful for the experiments where a fixed quantity of inputs (may be same dose of fertilizer, same quantity of irrigation water or same dose of insecticide or pesticide etc.) are applied as a combination of two or more ingredients. In these experiments the response is a function of the proportion of the ingredients in the mixture rather than the actual amount of the mixture. A bibliography of experiments with mixtures and online generation of simplex centroid designs are available on this page <http://https://drs.icar.gov.in/mixture/mixtures.aspx>. Some screen shots of experiments with mixtures are:



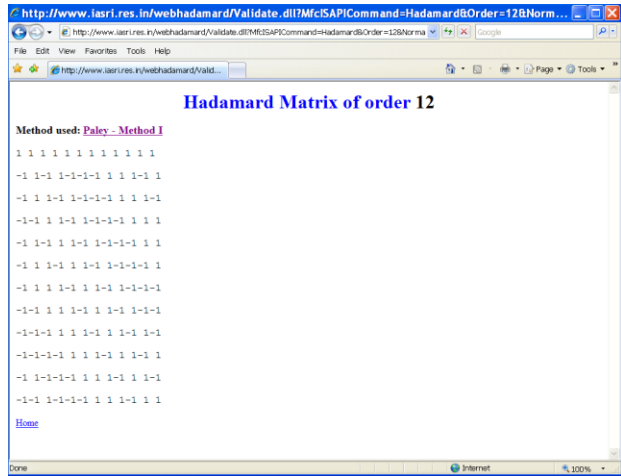
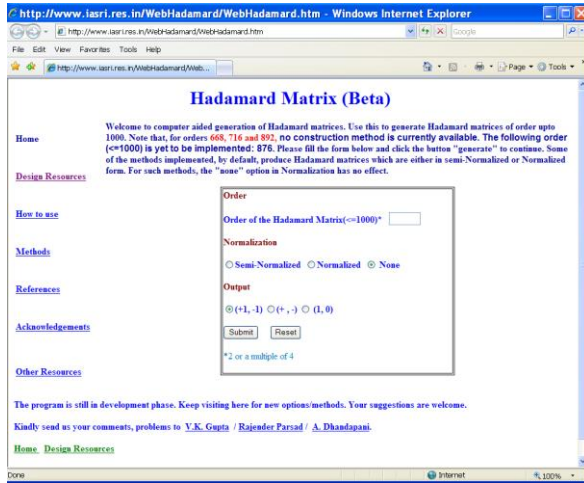
### 3.6 Online Design Generation- II

This link is helpful in generation of the following:

#### Hadamard matrix

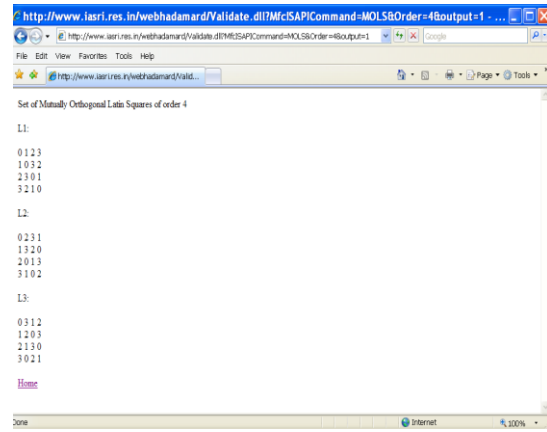
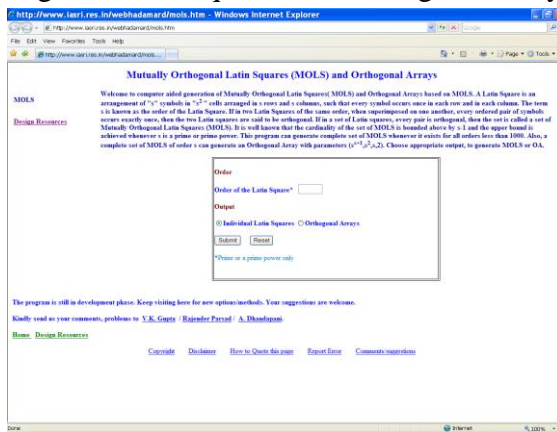
Hadamard matrices have a tremendous potential for applications in many fields particularly in fractional factorial plans, supersaturated designs, variance estimation from large scale complex survey data, generation of incomplete block designs, coding theory, etc. One can generate Hadamard matrices for all permissible orders up to 1000 except 668, 716, 876 and 892 using the URL <https://drs.icar.gov.in/WebHadamard/WebHadamard.htm>. Methods implemented produce Hadamard matrices in semi-normalized or normalized form. "None" option is also available. Hadamard matrix can be generated in (0,1); (+1,-1); or (+,-) form.

The method of generation of Hadamard matrix is also given. The screen shots for generation of Hadamard matrices are



### Mutually Orthogonal Latin Squares and Orthogonal arrays

Using this link one can generate complete set of mutually orthogonal Latin squares of order  $s$ ,  $s$  being a prime or prime power less than 1000. One can also generate an orthogonal array with parameters  $(s^{s+1}, s^2, s, 2)$  by choosing the output option as orthogonal arrays. The URL of this link is <https://drs.icar.gov.in/WebHadamard/mols.htm>. Some screen shots of mutually orthogonal Latin squares and orthogonal arrays are



### Balanced Incomplete Latin square Designs

Latin square designs are widely used in comparative experiments where two crossed blocking factors are present and each blocking factor has  $v$  levels, where  $v$  is also equal to the number of Latin letters in the Latin square or the number of treatments in the design. In this arrangement each Latin letter appears in each row and each column precisely once. However, it may not always be possible to accommodate all the  $v$  Latin letters or treatments precisely once in each row and / or each column, leading thereby to a situation where each row and / or each column may have less than  $v$  Latin letters or treatments appearing in them. In other words, there would be empty nodes in the Latin square arrangement and the Latin square design, meaning thereby that the row-column design is a non-orthogonal design with treatments versus rows and / or treatments versus columns classifications as non-orthogonal. Balanced incomplete Latin square designs have been introduced for such situations. A

balanced incomplete Latin square design with parameters  $v$  and  $r$  is an incomplete Latin square of order  $v$  such that each row and each column has  $r < v$  non-empty cells and  $v - r$  empty cells and each of the  $v$  symbols appears exactly  $r$  times in the whole square. Henceforth, we shall denote a balanced incomplete Latin square with  $v$  symbols and  $r$  replications of each symbol as BILS ( $v, r$ ). Here the term ‘balanced’ implies that each row and column has same number of non-empty cells and each symbol has same number of replications in the whole square. This balance is neither related to pair-wise balance nor variance balance. Here, construction of BILS ( $v, r$ ) is done by removing the  $v - r$  disjoint transversals from a Latin square of order  $v$  via a pair of orthogonal Latin squares (Ai et al., 2013). Here, a transversal in a Latin Square of order  $v$  is a set of  $v$  cells such that only one cell is allowed in each row and in each column, and furthermore, each symbol can appear in each cell once. A module for online generation of Balanced Latin square designs for all values of  $3 < v < 21$  (except  $v = 6, 10, 14, 18$ ) have been developed and made available at [https://drs.icar.gov.in/BILS\\_Design/Default.aspx](https://drs.icar.gov.in/BILS_Design/Default.aspx). Some screen shots for balanced incomplete Latin Squares are given in the sequel



### 3.7 Workshop Proceedings

Proceedings of 3 dissemination workshops are available for the stakeholders

1. Design and Analysis of On-Station and On-Farm Agricultural Experiments
2. Design and Analysis of Bioassays
3. Outliers in Designed Experiments

### 4. Other Useful Links

The purpose of this component is to develop a network of scientists in general and a network of statisticians in particular around the globe so that interesting and useful information can be shared among the peers. It also attempts to provide a sort of advisory to the scientists. Some other useful and important links available on world wide web are also provided.

#### 4.1 Discussion Board

The purpose of discussion board is to create a network of scientists and also to provide a platform for sharing any useful piece of

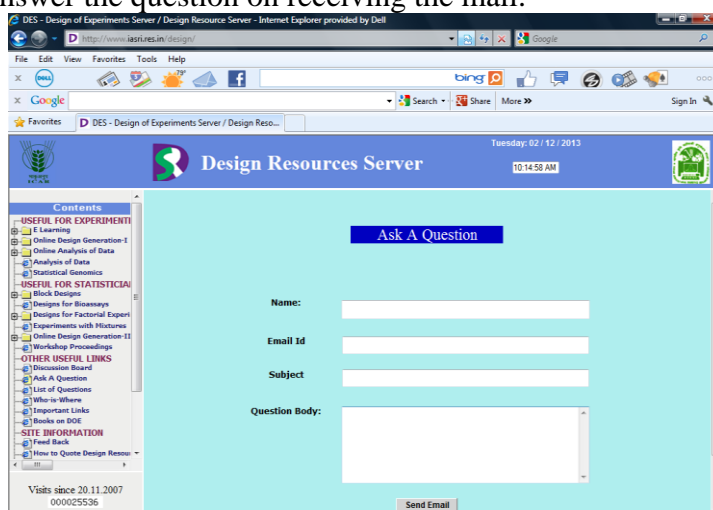




research or idea with scientists over the globe. The user can use this board for learning and disseminating information after registering on the discussion board. The information can be viewed by anybody over the globe. In case there are some queries or some researchable issues, then other peers can also respond to these queries. This helps in creating a network of scientists. Number of registered participants so far is 78 (23: Agricultural Research Statisticians; 37: Experimenters; One Vice-Chancellor and 17 ISS Officers). (<https://drs.icar.gov.in/MessageBoard/MessageBoard.asp> ).

### 4.2 Ask a Question

The ultimate objective of this server is to provide e-learning and e-advisory services. At present this is being achieved through the link “Ask a Question”. Once a user submits a question, a mail is automatically generated for Dr. Rajender Parsad, Dr. V.K. Gupta and Mrs. Alka Arora, who answer the question on receiving the mail.



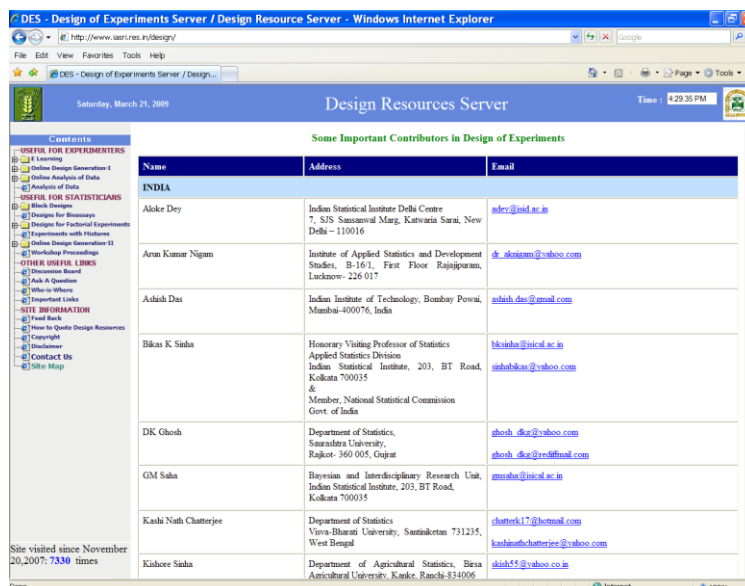
### 4.3 List of Questions

A list of questions answered through ‘Ask a Question’ has been provided for archiving and ready reference of users.

Design Resources Server			
Tuesday, 01/12/2013 2:56:02 PM			
Narany Kumar	narany_kumar@yahoo.com	I analyzed a response surface design having three factors each at three levels. There are four independent and eight dependent variables given over the 12 trials. I am not sure that there is lack of fit for almost all dependent variables in this situation, what should I need to do to make lack of fit insignificant?	Please let me know how do you analyze the data? Do you use software using MINITAB or Minitab? Also, can you share the "yes, used and which of one parameter" I shall be able to comment after receiving your data.
Rajar	rajarp186@gmail.com	DETAILS ABOUT PRINCIPLE COMPONENT ANALYSIS	Please see lecture notes on Substrate Technology, An Overview in Module IV of Electronic Book on Advances in Data Analytical Techniques available at <a href="http://ind.res.in/design_ebook/EBAAAT">http://ind.res.in/design_ebook/EBAAAT</a>
Narany Kumar	narany_kumar@yahoo.com	Respected sir, I have two factors (Temperature and Humidity) each at three levels. It is replicated three times. We have 27 design points. I want to analyze using Response surface design. Can I use independent central order rotatable design or central composite design or Box-Behnken design. Can I call this full factorial and analyze the response?	Check the experiment is conducted, the design cannot be changed. Design has to be done before the experiment is conducted. You can conduct it a factorial
David Brandt	vdbrandt@kellcometh.net	How would you analyze data where the response is not necessary the complete in full factorial and the two level factors are determined from combinatorial (one or two) level based type (substantial or practical)?	Please let me know how the data is collected. Is it a planned experiment or a screen test? If designed experiment, which design was used? Is it a screen

### 4.4 Who-is-where

Addresses of important contributors in Design of Experiments including their E-mail addresses have been linked to Design Resources Server. The list includes experts from USA, Canada, Australia, UK, China, Japan, Mexico, New Zealand, Oman, Syria, Taiwan, Vietnam and India. This information is useful for all the researchers in Design of Experiments in establishing linkages with their counterparts over the globe.



#### 4.5 Important Links

This gives links to other important sites that provide useful material on statistical learning in general and Design of Experiments in particular. Some links are as given below:

Sr. No.	Important Links
1.	Design Resources: <a href="http://www.designtheory.org">www.designtheory.org</a>
2.	Sankhya : <a href="http://sankhya.isical.ac.in">http://sankhya.isical.ac.in</a>
3.	Statistics Glossary <a href="http://www.cas.lancs.ac.uk/glossary_v1.1/main.html">http://www.cas.lancs.ac.uk/glossary_v1.1/main.html</a>
4.	Free Encyclopedia on Design of Experiments: <a href="http://en.wikipedia.org/wiki/Design_of_experiments">http://en.wikipedia.org/wiki/Design_of_experiments</a>
5.	Important Contributors to Statistics: <a href="http://en.wikipedia.org/wiki/Statistics#Important_contributors_to_statistics">http://en.wikipedia.org/wiki/Statistics#Important_contributors_to_statistics</a>
6.	Electronic Statistics Text Book: <a href="http://www.statsoft.com/textbook/stathome.html">http://www.statsoft.com/textbook/stathome.html</a>
7.	On-line construction of Designs: <a href="http://biometrics.hri.ac.uk/experimentaldesigns/website/hri.htm">http://biometrics.hri.ac.uk/experimentaldesigns/website/hri.htm</a>
8.	GENDEX: <a href="http://www.designcomputing.net/gendex/">http://www.designcomputing.net/gendex/</a>
9.	The Electronic Journal of Combinatorics: <a href="http://www.combinatorics.org">www.combinatorics.org</a>
10.	Annals of Combinatorics: <a href="http://www.combinatorics.net">www.combinatorics.net</a>
11.	Journal of Quality Technology: <a href="http://www.asq.org/pub/jqt/index.html">http://www.asq.org/pub/jqt/index.html</a>
12.	Indian Council of Agriculture Research (ICAR): <a href="http://www.icar.org.in">http://www.icar.org.in</a>
13.	Hadamard Matrices 1. <a href="http://www.research.att.com/~njas/hadamard">http://www.research.att.com/~njas/hadamard</a> 2. <a href="http://www.uow.edu.au/~jennie/WILLIAMSON/williamson.html">http://www.uow.edu.au/~jennie/WILLIAMSON/williamson.html</a>
14.	Biplots : <a href="http://www.ggebiplot.com">http://www.ggebiplot.com</a>
15.	Free Statistical Softwares: <a href="http://freestatistics.altervista.org/en/stat.php">http://freestatistics.altervista.org/en/stat.php</a>
16.	Learning Statistics: <a href="http://freestatistics.altervista.org/en/learning.php">http://freestatistics.altervista.org/en/learning.php</a>
17.	Free Math Software: <a href="http://freestatistics.altervista.org/en/math.php">http://freestatistics.altervista.org/en/math.php</a>
18.	Statistical Calculators: <a href="http://www.graphpad.com/quickcalcs/index.cfm">http://www.graphpad.com/quickcalcs/index.cfm</a>
19.	Current Index to Statistics: <a href="http://www.statindex.org/">http://www.statindex.org/</a>

20.	SAS Online Doc 9.1.3: <a href="http://support.sas.com/onlinedoc/913/docMainpage.jsp">http://support.sas.com/onlinedoc/913/docMainpage.jsp</a>
21.	University of South California: Courses in Statistics: <a href="http://www.stat.sc.edu/curricula/courses/">http://www.stat.sc.edu/curricula/courses/</a>
22.	Course on Introduction to Experimental Design: <a href="http://www.stat.sc.edu/~grego/courses/stat506">http://www.stat.sc.edu/~grego/courses/stat506</a>
23.	Course on Experimental design: <a href="http://www.stat.sc.edu/~grego/courses/stat706/">http://www.stat.sc.edu/~grego/courses/stat706/</a>
24.	Current Index to Statistics: <a href="http://www.statindex.org/">http://www.statindex.org/</a>
25.	International Indian Statistical Association: <a href="http://www.intindstat.org/">http://www.intindstat.org/</a>
26.	Statistical Calculators: <a href="http://www.graphpad.com/quickcalcs/index.cfm">www.graphpad.com/quickcalcs/index.cfm</a>
27.	Journal of Indian Society of Agricultural Statistics: <a href="http://www.isas.org.in/jisas">www.isas.org.in/jisas</a>
28.	Free Statistical Softwares: <a href="http://statpages.org/javasta2.html">http://statpages.org/javasta2.html</a>
29.	Teaching Statistics Video: <a href="http://www.youtube.com/user/sarjinder1">http://www.youtube.com/user/sarjinder1</a>
30.	Statistical Services Centre at University of Reading: <a href="http://www.reading.ac.uk/ssc/home.html">http://www.reading.ac.uk/ssc/home.html</a>
31.	Rothamsted Experimental Station: <a href="http://www.rothamsted.ac.uk/">http://www.rothamsted.ac.uk/</a>
32.	Indian Statisticians: <a href="http://www.en.wikipedia.org/wiki/category:Indian_statisticians">http://www.en.wikipedia.org/wiki/category:Indian_statisticians</a>
33.	Evolution of Statistics in India: <a href="http://library.isical.ac.in/jspui/bitstream/10263/3935/1/Evolution%20of%20Statistics%20in%20India.pdf">http://library.isical.ac.in/jspui/bitstream/10263/3935/1/Evolution%20of%20Statistics%20in%20India.pdf</a>

#### 4.6 Books on Design of Experiments

A list of books on Design of Experiments has been provided for the benefit of the visitors of this web resource, the faculty, the researchers in Design of Experiments and the students. No claim is being made for this list to being exhaustive. New additions would be made to it from time to time.

#### 5. Site Information

This link provides information about the site on the following aspects (i) Feedback from stakeholders, (ii) How to Quote Design Resources Server, (iii) Copyright, (iv) Disclaimer, (v) Contact us, and (vi) Sitemap.

##### 5.1 Feedback/ Comments

The feedback / comments received from the users visiting the site have been put on the server so that every user can benefit from the experience of other users. More importantly, the feedback helps in improving the contents of the site and their presentation too. We have received feedback from 19 researchers (6: Design Experts from India; 7: Experts from abroad; 4: Experimenters and 2: Agricultural Research Statisticians). The first feedback was received from Dr K Rameash, Entomologist working at ICAR Research Complex for NEH Region, Sikkim Centre, Tadong, Gangtok.

##### 5.2 How to quote Design Resources Server

To Quote Design Resources Server, use:

**Design Resources Server.** *Indian Agricultural Statistics Research Institute (ICAR), New Delhi 110 012, India.* <https://drs.icar.gov.in> (accessed last on <date>).

If referring to a particular page, then the site may be quoted as

Authors' name in 'Contact Us' list on that page. Title of page: Design Resources Server. *Indian Agricultural Statistics Research Institute (ICAR), New Delhi 110 012, India.* <https://drs.icar.gov.in> (accessed last on <date>).

For example, page on alpha designs may be cited as  
Parsad, R., Gupta, V.K. and Dhandapani, A. Alpha Designs: Design Resources Server. *Indian Agricultural Statistics Research Institute (ICAR), New Delhi 110 012, India.* <https://drs.icar.gov.in> (accessed last on 21.03.2009).

### **5.3 Copyright**

This website and its contents are copyright of "IASRI (ICAR)" - © "ICAR" 2008. All rights reserved. Any redistribution or reproduction of part or all of the contents in any form, other than the following, is prohibited:

- print or download to a local hard disk extracts for personal and non-commercial use only.
- transmit it or store it in any other website or other form of electronic retrieval system.
- except with express written permission of the authors, distribution or commercial exploitation of the contents.

### **5.4 Disclaimer**

The information contained in this website is for general information purposes only. The information is provided by "IASRI" and whilst "IASRI" endeavours to keep the information up-to-date and correct, no representations or warranties of any kind, express or implied, about the completeness, accuracy, reliability, suitability or availability with respect to the website or the information, products, services, or related graphics contained on the website are made for any purpose. Any reliance placed on such information is, therefore, strictly at user's own risk.

In no event will "IASRI" be liable for any loss or damage including without limitation, indirect or consequential loss or damage, or any loss or damage whatsoever arising from loss of data or profits arising out of or in connection with the use of this website.

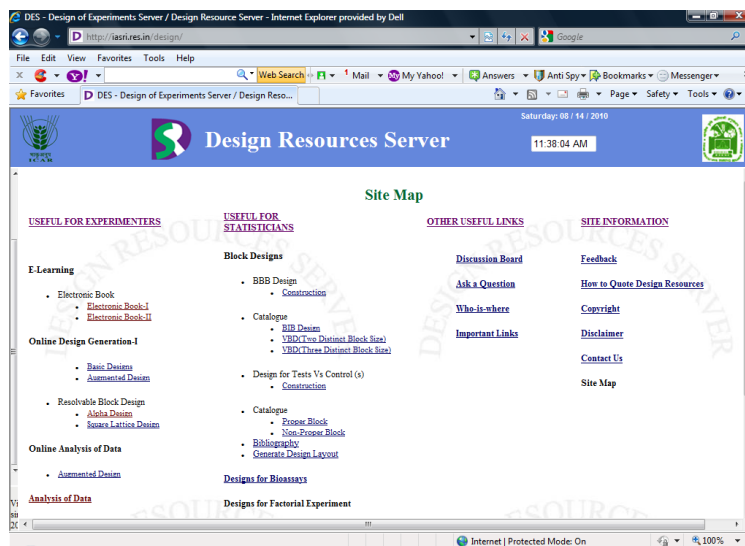
Through this website users are able to link to other websites which are not under the control of "IASRI". The inclusion of any links does not necessarily imply a recommendation or endorsement the views expressed within them.

Every effort is made to keep the website running smoothly. However, "IASRI" takes no responsibility for and will not be liable for the website being temporarily unavailable due to technical issues beyond our control.

### **5.5 Site Map**



This link gives a map of the various links available on the server. A user can access any of the links through this map also. A snap shot of the site map is given below:



## 6. Some Information on the Usage of the Server

- Design Resources Server is a copyright of IASRI (ICAR): L-46452/2013 dated June 07, 2013. The Server was registered under Google Analytics on May 26, 2008.
- External links of the server are also available at:
  - [http://en.wikipedia.org/wiki/Design\\_of\\_experiments](http://en.wikipedia.org/wiki/Design_of_experiments)
  - [http://en.wikipedia.org/wiki/Hadamard\\_matrix](http://en.wikipedia.org/wiki/Hadamard_matrix)
- The server has been cited at:
  - [https://dspace.ist.utl.pt/bitstream/2295/145675/1/licao\\_21.pdf](https://dspace.ist.utl.pt/bitstream/2295/145675/1/licao_21.pdf) for lecture presentation on Unitary operators.
  - Chiarandini, Marco (2008). DM811-Heuristics for Combinatorial Optimization. Laboratory Assignment, Fall 2008. Department of Mathematics and Computer Science, University of Southern Denmark, Odense.
  - <http://support.sas.com/techsup/technote/ts723.html>
  - Warren F. Kuhfeld. Orthogonal Arrays. Analytics Division SAS, Document No. 273 (<http://support.sas.com/techsup/technote/ts723.html>).
  - Electronic text material in “New and Restructured Post-Graduate Curricula & Syllabi on Statistical Sciences (Statistics/Agricultural Statistics; Bio-Statistics, Computer Application) of Education Division, Indian Council of Agricultural Research, New Delhi, 2008.
  - Jingbo Gao, Xu Zhu, Nandi, A.K. (2009). Nonredundant precoding and PARR reduction in MIMO OFDM systems with ICA based blind equalization. IEEE transactions on Wireless Communications, 8(6), 3038-3049.
- Server is also linked at
  - ICARDA Intranet: Biometric Services
  - CG Online learning resources- [http://learning.cgiar.org/moodle/Experimental Designs and Data Analysis](http://learning.cgiar.org/moodle/Experimental_Designs_and_Data_Analysis)

## 7. Future Directions

The Design Resources Server created and being strengthened at IASRI aims to culminate into an expert system on design of experiments. To achieve this end, the materials available on various links need to be strengthened dynamically. Besides this, the following additions need to be made to the server in the near future:

- Online generation of
  - balanced incomplete block designs, binary balanced block designs and partially balanced incomplete block designs
  - block designs with nested factors
  - designs for crop sequence experiments
  - efficient designs for correlated error structures
  - online generation of row-column designs
  - designs for factorial experiments; fractional factorial plans
- designs for microarray experiments
- designs for computer experiments
- designs for fitting response surfaces; designs for experiments with mixtures
- split and strip plot designs
- field book of all the designs generated
- labels generation for preparing seed packets
- online analysis of data

The success of the server lies in the hands of users. It is requested that the scientists in NARS use this server rigorously and send their comments for further improvements to Dr. Rajender Parsad ([rajender.parsad@icar.gov.in](mailto:rajender.parsad@icar.gov.in)) / Dr. V.K. Gupta ([vk Gupta1751@yahoo.co.in](mailto:vk Gupta1751@yahoo.co.in)). The comments/ suggestions would be helpful in making this server more meaningful and useful.

---

---

## DESIGNS FOR FACTORIAL EXPERIMENTS

---

---

V.K. Gupta, Rajender Parsad, Sukanta Dash and Susheel Kumar Sarkar  
ICAR-Indian Agricultural Statistics Research Institute  
Library Avenue, New Delhi - 110 012  
*vk Gupta\_1751@yahoo.co.in; rajender.parsad@icar.gov.in; sukanta.dash@icar.gov.in;*  
*susheel.sarkar@icar.gov.in*

---

---

### 1. Introduction

Suppose that one wants to conduct an experiment to study the performance of a new crop or tree species on the basis of yield in an area where it has never been grown before. A sample of pertinent questions that arise for planning the experiment must be answered is given below:

1. What should be the best crop variety?
2. When should the crop be planted (Date of sowing)?
3. Should it be sown directly or transplanted? If sown directly, what would be the seeding rate and if transplanted, what would be the age of the seedlings?
4. Should the seed be drilled or broadcast?
5. Must we use fertilizer? If yes, how much of the major elements are needed?
6. Have we to add minor elements?
7. Is irrigation necessary?
8. What should be the plant-to-plant and line-to-line spacing?

This problem may be investigated by varying a single factor at a time using designs for single factor experiments (like completely randomized designs, randomized complete block designs, incomplete block designs, row-column designs, etc.). For example, an experiment may be conducted with varieties of the crop as treatments to pick the best variety. Using the best variety, another experiment may be conducted to obtain the date of sowing. Then using the best variety and the optimum date of sowing another experiment may be conducted to find the optimum level for the other factors one at a time. The soundness of this approach rests on the assumption that the response to different varieties is independent of amount of nitrogen given *i.e.* the factors act independent of each other. But then this is a big assumption and such situations are very rare.

To make the exposition simple, let us take two factors *viz.* irrigation and nitrogen fertilizer. It is known that for most of the crops, higher level of irrigation up to certain limit is required to secure an adequate response from a higher dose of manure. The two factors are not independent but interact with each other. Thus, ***interaction is the failure of the differences in response to changes in levels of one factor, to retain the same order and magnitude of performance through out all the levels of other factors or the factors are said to interact if the effect of one factor changes as the levels of the other factor(s) changes.***

In practice the experimenter deals with simultaneous variation in more than one factor. It may be required to find the combination of most suitable level of irrigation and the optimum dose of a nitrogenous fertilizer. Consider the results of a trial designed to measure the effects of nitrogen and irrigation, both alone and in combinations when applied to rice crop. The yield of rice crop in q/ha is given as

**Rice yield in q/ha**

Nitrogen → Irrigation ↓	0 kg N/ha (N <sub>0</sub> )	60 kg N/ha (N <sub>1</sub> )	Mean
5 cm irrigation (I <sub>0</sub> )	N <sub>0</sub> I <sub>0</sub> 10.0	N <sub>1</sub> I <sub>0</sub> 30.0	20.0
10 cm irrigation (I <sub>1</sub> )	N <sub>0</sub> I <sub>1</sub> 20.0	N <sub>1</sub> I <sub>1</sub> 40.0	30.0
Mean	15.0	35.0	

Effect of nitrogen at I<sub>0</sub> level of irrigation = 30.0 - 10.0 = 20.0 q/ha

Effect of nitrogen at I<sub>1</sub> level of irrigation = 40.0 - 20.0 = 20.0 q/ha

Effect of irrigation at N<sub>0</sub> level of nitrogen = 20.0 - 10.0 = 10.0 q/ha

Effect of irrigation at N<sub>1</sub> level of nitrogen = 40.0 - 30.0 = 10.0 q/ha

As effect of nitrogen (irrigation) is same at all the levels of irrigation (nitrogen) hence, there is **no interaction** between nitrogen and irrigation. Consider the results of another trial designed to measure the effects of nitrogen and irrigation, both alone and in combinations when applied to rice crop. The yield of rice crop in q/ha is given as

**Rice yield in q/ha**

Nitrogen → Irrigation ↓	0 kg N/ha (N <sub>0</sub> )	60 kg N/ha (N <sub>1</sub> )	Mean
5 cm irrigation (I <sub>0</sub> )	N <sub>0</sub> I <sub>0</sub> 10.0	N <sub>1</sub> I <sub>0</sub> 30.0	20.0
10 cm irrigation (I <sub>1</sub> )	N <sub>0</sub> I <sub>1</sub> 20.0	N <sub>1</sub> I <sub>1</sub> 50.0	35.0
Mean	15.0	40.0	

Effect of nitrogen at I<sub>0</sub> level of irrigation = 30.0 - 10.0 = 20.0 q/ha

Effect of nitrogen at I<sub>1</sub> level of irrigation = 50.0 - 20.0 = 30.0 q/ha

Effect of irrigation at N<sub>0</sub> level of nitrogen = 20.0 - 10.0 = 10.0 q/ha

Effect of irrigation at N<sub>1</sub> level of nitrogen = 50.0 - 30.0 = 20.0 q/ha

As effect of nitrogen (irrigation) is not same at all the levels of irrigation (nitrogen) hence, nitrogen and irrigation are **interacting**.

These effects as explained above are called as simple effects of the factors and average of these simple effects is called **main effect** of the factor. Thus,

$$\text{Main effect of Nitrogen} = \frac{20.0 + 30.0}{2} = 25.0 \text{ q/ha}$$

$$\text{Main effect of Irrigation} = \frac{10.0 + 20.0}{2} = 15.0 \text{ q/ha}$$

Interaction of Irrigation and Nitrogen is the difference between simple effects, e.g., simple effect of Irrigation at N<sub>1</sub> level of Nitrogen minus the simple effect of Irrigation at N<sub>0</sub> level of Nitrogen = 20.0 - 10.0 = 10.0 q/ha. It may also be defined as the simple effect of Nitrogen at I<sub>1</sub> level of Irrigation minus the simple effect of Nitrogen at I<sub>0</sub> level of Irrigation = 30.0 - 20.0 = 10.0 q/ha.

If interactions exist, which is generally true, the experiments should be planned in such a way that these can be estimated and tested. It is now clear that it is not possible to

estimate interactions from the experiments in which levels of only one factor are studied at a time. For this purpose, we must use multi-level, multi-factor experiments.

## 2. What are factorial experiments?

**Definition:** A treatment arrangement in which the treatments consist of all combination of all levels of two or more factors. It is just an arrangement of treatments, not a design. One can use this approach with a variety of designs.

Also factorial experiments can be defined as experiments in which the effects (main effects and interactions) of more than one factor are studied together. In general if there are  $n$  factors, say,  $F_1, F_2, \dots, F_n$  and the  $i^{\text{th}}$  factor has  $s_i$  levels,  $i=1, \dots, n$ , then the total number of treatment combinations is  $\prod_{i=1}^n s_i$ . Factorial experiments are of two types.

1. **Symmetrical Factorial Experiments:** In these experiments the number of levels of all factors is same *i.e.*,  $s_i = s \quad \forall i = 1, \dots, n$ .
2. **Asymmetrical Factorial Experiments:** In these experiments the number of levels of all the factors are not same *i.e.* there are at least two factors for which the number of levels  $s_i$ 's are different.

Factorial experiments have many advantages over single factor experiments.

### Advantages:

- More precision on each factor than with single factor experiments due to hidden replications.
- Provide an opportunity to study not only the individual effects of the factors but also their interactions.
- Good for exploratory work where we wish to find most important factor or the optimal level of factor or combination of levels of more than one factor.
- These experiments have the further advantage of economizing the experimental resources. When the experiments are conducted factor by factor a large number of experimental units are required for getting the same precision of estimation as one would have got when all the factors are experimented together in the same experiment, *i.e.*, factorial experiment. There is thus a considerable amount of saving of resources. Moreover, factorial experiments also enable us to study interactions which the experiments conducted factor by factor do not allow us to study.

### Disadvantages:

- This approach is more complex than that of single factor experiments
- With a number of factors each at several levels, the experiment can become very large.

### 2.1 Symmetrical factorial experiments

The simplest symmetrical factorial experiments are  $2^n$  factorial experiments in which all the  $n$  factors have 2 levels each. Consider the  $2^2$  factorial experiment with 2 factors say  $A$

and  $B$  each at two levels, say  $0$  and  $1$ . There will be  $4$  treatment combinations that can be written as

- $00 = a_0 b_0 = (1)$ ;  $A$  and  $B$  both at first levels
- $10 = a_1 b_0 = a$ ;  $A$  at second level and  $B$  at first level
- $01 = a_0 b_1 = b$ ;  $A$  at first level and  $B$  at second level
- $11 = a_1 b_1 = ab$ ;  $A$  and  $B$  both at second level.

We denote the treatment combinations by small letters ( $1$ ),  $a$ ,  $b$ ,  $ab$  indicating the presence of low or high level of the factor and treatment totals by  $[1]$ ,  $[a]$ ,  $[b]$ ,  $[ab]$ . The following table gives the responses due to Factor  $A$  and Factor  $B$ .

Factor A → Factor B ↓	$a_0$ or $0$	$a_1$ or $1$	Response due to A
$b_0$ or $0$	$[1]$ or $[a_0 b_0]$	$[a]$ or $[a_1 b_0]$	$[a] - [1]$ or $[a_1 b_0] - [a_0 b_0]$
$b_1$ or $1$	$[b]$ or $[a_0 b_1]$	$[ab]$ or $[a_1 b_1]$	$[ab] - [b]$ or $[a_1 b_1] - [a_0 b_1]$
Response Due to B	$[b] - [1]$ or $[a_0 b_1] - [a_0 b_0]$	$[ab] - [a]$ or $[a_1 b_1] - [a_1 b_0]$	

The responses  $[a] - [1]$  and  $[ab] - [b]$  are called simple effects of the factor  $A$  at  $0$  and  $1$  levels, respectively of the factor  $B$ . If the factors  $A$  and  $B$  are independent, the responses  $[a] - [1]$  and  $[ab] - [b]$ , both provide the estimate of the response due to  $A$  (except for the experimental error). The average of these two simple effects is known as Main Effect of factor  $A$ . Thus the main effect of factor  $A$  is

$$A = \frac{1}{2} \{ [a_1 b_1] - [a_0 b_1] + [a_1 b_0] - [a_0 b_0] \} \text{ or } A = \frac{1}{2} \{ [ab] - [b] + [a] - [1] \} \quad (1)$$

This is simplified by writing it in the form  $A = \frac{1}{2} (a - 1)(b + 1)$ , where the right hand side is to be expanded algebraically and then the treatment combinations are to be replaced by corresponding treatment totals. From (1) we find that  $A$  is a linear function of the four treatments totals with the sum of the coefficients of the linear function equal to zero ( $\frac{1}{2} -$

$\frac{1}{2} + \frac{1}{2} - \frac{1}{2} = 0$ ). Such a linear function among the treatment totals with sum of coefficients equal to zero is called a contrast (or a comparison) of the treatment totals. Similarly the main effect of factor  $B$  is

$$B = \frac{1}{2} \{ [a_1 b_1] + [a_0 b_1] - [a_1 b_0] - [a_0 b_0] \} \text{ or } B = \frac{1}{2} \{ [ab] + [b] - [a] - [1] \} \quad (2)$$

This is simplified by writing it in the form  $B = \frac{1}{2} (a + 1)(b - 1)$  where the right hand side is to be expanded algebraically and then the treatment combinations are to be replaced by corresponding treatment totals. From (2), we find that  $B$  is a linear function of the four treatments totals with the sum of the coefficients of the linear function equal to zero ( $\frac{1}{2} +$

$\frac{1}{2} - \frac{1}{2} - \frac{1}{2} = 0$ ), hence a contrast.

Consider now the difference of two simple effects of  $A$

$$= \{ [ab] - [b] - [a] + [1] \} \tag{3}$$

Had the two factors been independent, then (3) would be zero. If not then this provides an estimate of interdependence of the two factors and it is called the interaction between  $A$  and  $B$ . The interaction between  $A$  and  $B$  is defined as

$$AB = \frac{1}{2} (a - 1)(b - 1)$$

where the expression on the right hand side is to be expanded algebraically and then the treatment combinations are to be replaced by the corresponding treatment totals. It is easy to verify that  $AB$  is a contrast of the treatment totals. The coefficients of the contrasts  $A$  and  $AB$  are such that the sum of the products of the corresponding coefficients of the contrasts  $A$  and  $AB$  is equal to zero *i.e.*  $(\frac{1}{2})(\frac{1}{2}) + (-\frac{1}{2})(-\frac{1}{2}) + (\frac{1}{2})(-\frac{1}{2}) + (-\frac{1}{2})(\frac{1}{2}) = 0$ . Thus the contrasts  $A$  and  $AB$  are orthogonal contrasts. It is easy to verify that the interaction of the factor  $B$  with factor  $A$ , *i.e.*,  $BA$  is the same as the interaction  $AB$  and hence the interaction does not depend on the order of the factors. It is also easy to verify that the main effect  $B$  is orthogonal to both  $A$  and  $AB$ .

The above three orthogonal contrasts defining the main effects and interaction can be easily obtained from the following table, which gives the signs with which to combine the treatment totals and also the divisor for obtaining the corresponding sum of squares. Main effects and interactions are expressed in terms of individual treatment totals.

<i>Treatment Totals</i> → <i>Factorial Effect</i> ↓	$[1]$	$[a]$	$[b]$	$[ab]$	<i>Divisor</i>
$M$	+	+	+	+	$4r$
$A$	-	+	-	+	$4r$
$B$	-	-	+	+	$4r$
$AB$	+	-	-	+	$4r$

Here  $r$  denotes the replication number. The rule to write down the signs of the main effect is to give a plus sign to the treatment combinations containing the corresponding small letter and a minus sign where the corresponding small letter is absent. The signs of interaction are obtained by multiplying the corresponding signs of the two main effects. The first line gives the general mean

$$M = \frac{1}{4} \{ [ab] + [a] + [b] + [1] \}$$

Consider now the  $2^3$  factorial experiment with 3 factors  $A$ ,  $B$ , and  $C$  each at two levels, say  $0$  and  $1$ . The 8 treatment combinations are written as

- $000 = a_0 b_0 c_0 = (1)$ ;       $A, B$  and  $C$ , all three at first level
- $100 = a_1 b_0 c_0 = a$  ;       $A$  at second level and  $B$  and  $C$  at first level
- $010 = a_0 b_1 c_0 = b$  ;       $A$  and  $C$  both at first level and  $B$  at second level
- $110 = a_1 b_1 c_0 = ab$ ;       $A$  and  $B$  both at second level and  $C$  at first level
- $001 = a_0 b_0 c_1 = c$  ;       $A$  and  $B$  both at first level and  $C$  at second level.
- $101 = a_1 b_0 c_1 = ac$ ;       $A$  and  $C$  both at second level and  $B$  at first level
- $011 = a_0 b_1 c_1 = bc$ ;       $A$  at first level and  $B$  and  $C$  both at second level

$III = a_1 b_1 c_1 = abc$ ;  $A, B$  and  $C$ , all three at second level

In a three factor experiment there are 3 main effects  $A, B$ , and  $C$ ; 3 first order or two factor interactions  $AB, AC$ , and  $BC$ ; and one second order or three factor interaction  $ABC$ . The main effects and interactions may be written as

$$A = \frac{1}{4}(a-1)(b+1)(c+1), B = \frac{1}{4}(a+1)(b-1)(c+1), C = \frac{1}{4}(a+1)(b+1)(c-1)$$

$$AB = \frac{1}{4}(a-1)(b-1)(c+1), AC = \frac{1}{4}(a-1)(b+1)(c-1), BC = \frac{1}{4}(a+1)(b-1)(c-1)$$

$$ABC = \frac{1}{4}(a-1)(b-1)(c-1).$$

These main effects and interactions are mutually orthogonal as may be verified from the following table of signs:

<i>Treatment Totals</i> → <i>Factorial Effect</i> ↓	[1]	[a]	[b]	[ab]	[c]	[ac]	[bc]	[abc]	Divisor
<i>M</i>	+	+	+	+	+	+	+	+	8r
<i>A</i>	-	+	-	+	-	+	-	+	8r
<i>B</i>	-	-	+	+	-	-	+	+	8r
<i>AB</i>	+	-	-	+	+	-	-	+	8r
<i>C</i>	-	-	-	-	+	+	+	+	8r
<i>AC</i>	+	-	+	-	-	+	-	+	8r
<i>BC</i>	+	+	-	-	-	-	+	+	8r
<i>ABC</i>	-	+	+	-	+	-	-	+	8r

The rule for obtaining the signs of main effects and two factor interactions is the same as that stated for a  $2^2$  experiment. The signs of  $ABC$  may be obtained by multiplying the signs of  $AB$  and  $C$  or  $AC$  and  $B$  or  $BC$  and  $A$  or  $A, B$  and  $C$ .

Incidentally, it may be remarked that the method of representing the main effects and interactions, which is due to Yates, is very useful and quite straightforward. For example, if the design is  $2^4$  then

$$A = \frac{1}{2^3}(a-1)(b+1)(c+1)(d+1), AB = \frac{1}{2^3}(a-1)(b-1)(c+1)(d+1),$$

$$ABC = \frac{1}{2^3}(a-1)(b-1)(c-1)(d+1), \text{ and } ABCD = \frac{1}{2^3}(a-1)(b-1)(c-1)(d-1)$$

By this rule the main effect or interaction of any design of the series  $2^n$  can be written out directly without first obtaining the simple effects and then expressing the main effects or interactions. For example,

$$A = \frac{1}{2^{n-1}}(a-1)(b+1)(c+1)(d+1)(e+1) \dots, AB = \frac{1}{2^{n-1}}(a-1)(b-1)(c+1)(d+1)(e+1) \dots,$$

$$ABC = \frac{1}{2^{n-1}}(a-1)(b-1)(c-1)(d+1)(e+1) \dots,$$

$$\text{and } ABCD = \frac{1}{2^{n-1}}(a-1)(b-1)(c-1)(d-1)(e+1) \dots$$



In case of a  $2^n$  factorial experiment, there will be  $2^n (=v)$  treatment combinations. We shall have  $n$  main effects;  $\binom{n}{2}$  first order or two factor interactions;  $\binom{n}{3}$  second order or three factor interactions;  $\binom{n}{4}$  third order or four factor interactions and so on,  $\binom{n}{r}$ ,  $(r-1)^{th}$  order or  $r$  factor interactions and  $\binom{n}{n}$ ,  $(n-1)^{th}$  order or  $n$  factor interaction. Using these  $v$  treatment combinations, the experiment may be laid out using any of the suitable experimental designs viz. completely randomized design or block designs or row-column designs, etc.

### 2.1.1 Steps of Analysis:

**Step 1:** Obtain the sum of squares ( $S.S.$ ) due to treatments,  $S.S.$  due to replications (in case randomized block design is used),  $S.S.$  due to rows and columns (in case a row-column design is used), total  $S.S.$  and  $S.S.$  due to error as per established procedures. In case a completely randomized design is used, there will be no  $S.S.$  due to replications.

**Step 2:** In order to study the main effects and interactions, the treatment sum of squares is divided into different components viz. main effects and interactions each with single  $d.f.$  We can obtain the  $S.S.$  due to these factorial effects by dividing the squares of the factorial effect totals by  $r.2^n$ .

**Step 3:** Obtain mean squares ( $M.S.$ ) by dividing each  $S.S.$  by corresponding respective degrees of freedom.

**Step 4:** After obtaining the different  $S.S.$ , the usual ANOVA table is prepared and the different effects are tested against error mean square and conclusions drawn.

**Step 5:** Obtain the standard errors ( $S.E.$ ) for difference of means for all levels of single factor averaged over levels of all other factors and means for all level combinations of two factors averaged over levels of all other factors, using the following expressions.

$S.E$  estimate of difference between means for all levels of single factor averaged over levels of all other factors =  $\sqrt{\frac{2MSE}{r.2^{n-1}}}$

$S.E$  estimate of difference between means for all level combinations of two factors averaged over levels of all other factors =  $\sqrt{\frac{2MSE}{r.2^{n-2}}}$

In general,  $S.E.$  estimate for testing the difference between means for all level combinations of  $p$ - factors averaged over levels of all other factors

$$= \sqrt{\frac{2MSE}{r.2^{n-p}}} \quad \forall p=1,2,\dots,n.$$

The critical differences are obtained by multiplying the  $S.E.$  estimate by the student's  $t$  value at  $\alpha\%$  level of significance and at error  $d.f.$

Please note that when we say the critical difference for a factorial main effect, we actually mean to say that the critical difference for testing the pairwise difference between levels of that factor averaged over levels of other factors. Similarly, the critical difference for the interaction effect involving  $p$  factors means that the critical difference for testing the pairwise difference between the treatment combinations of levels of those factors averaged over levels of other factors.

The ANOVA for a  $2^n$  factorial experiment with  $r$  replications conducted using a randomized complete block design will be

ANOVA				
Source of variation	Degrees of freedom	S.S.	M.S.	F-calculated
Replications	$r-1$	$SSR$	$MSR = SSR/(r-1)$	$MSR/MSE$
Treatments	$2^n - 1$	$SST$	$MST = SST/(2^n - 1)$	$MST/MSE$
A	$1$	$SSA = [A]^2/r2^n$	$MSA = SSA$	$MSA/MSE$
B	$1$	$SSB = [B]^2/r2^n$	$MSB = SSB$	$MSB/MSE$
AB	$1$	$SSAB = [AB]^2/r2^n$	$MSAB = SSAB$	$MSAB/MSE$
C	$1$	$SSC = [C]^2/r2^n$	$MSC = SSC$	$MSC/MSE$
AC	$1$	$SSAC = [AC]^2/r2^n$	$MSAC = SSAC$	$MSAC/MSE$
	$:$	$:$	$:$	$:$
Error	$(r-1)(2^n-1)$	$SSE$	$MSE = SSE/(r-1)(2^n-1)$	
Total	$r.2^n-1$	$TSS$		

**Example 1:** Analyze the data of a  $2^3$  Factorial Experiment conducted using a RCBD with three replications. The three factors are the fertilizers viz, Nitrogen ( $N$ ), Phosphorus ( $P$ ) and Potassium ( $K$ ). The purpose of the experiment is to determine the effect of different kinds of fertilizers on potato crop yield. The yields under 8 treatment combinations for each of the three randomized blocks are given below:

**Block-I**

$npk$	(1)	$k$	$np$	$p$	$n$	$nk$	$Pk$
450	101	265	373	312	106	291	391

**Block-II**

$p$	$nk$	$k$	$np$	(1)	$npk$	$pk$	$N$
324	306	272	338	106	449	407	89

**Block-III**

$p$	$npk$	$nk$	(1)	$n$	$k$	$pk$	$Np$
323	471	334	87	128	279	423	324

**Analysis:**

**Step 1:** To find the sum of squares due to blocks (replications), due to treatments and total S.S., arrange the data in the following table

Blocks↓	Treatment Combinations→								Total
	(1)	$n$	$p$	$np$	$k$	$nk$	$pk$	$npk$	
$B_1$	101	106	312	373	265	291	391	450	2289 ( $B_1$ )
$B_2$	106	89	324	338	272	306	407	449	2291 ( $B_2$ )

$B_3$	87	128	323	324	279	334	423	471	2369 ( $B_3$ )
<b>Total</b>	294	323	959	1035	816	931	1221	1370	6949 ( $G$ )
	( $T_1$ )	( $T_2$ )	( $T_3$ )	( $T_4$ )	( $T_5$ )	( $T_6$ )	( $T_7$ )	( $T_8$ )	

Grand Total,  $G = 6949$ ; Number of observations ( $n$ ) = 24 = ( $r.2^n$ )

$$\text{Correction Factor (C.F.)} = \frac{G^2}{n} = \frac{(6949)^2}{24} = 2012025.042$$

$$\begin{aligned} \text{Total S.S. (TSS)} &= (101^2 + 106^2 + \dots + 449^2 + 471^2) - C.F. \\ &= 352843.958 \end{aligned}$$

$$\begin{aligned} \text{Block (Replication) S.S. (SSR)} &= \sum_{j=1}^r \frac{B_j^2}{2^3} - C.F. \\ &= \frac{[(2289)^2 + (2291)^2 + (2369)^2]}{8} - C.F. \\ &= 520.333 \end{aligned}$$

$$\begin{aligned} \text{Treatment S.S. (SST)} &= \sum_{i=1}^v \frac{T_i^2}{r} - C.F. \\ &= \frac{(294)^2 + (323)^2 + (959)^2 + (1035)^2 + (816)^2 + (931)^2 + (1221)^2 + (1370)^2}{3} - C.F. \\ &= \frac{7082029}{3} - 2012025.042 = 348651.2913 \end{aligned}$$

$$\begin{aligned} \text{Error S.S. (SSE)} &= \text{Total S.S.} - \text{Block S.S.} - \text{Treatment S.S.} \\ &= 352843.958 - 520.333 - 348651.2913 = 3672.3337 \end{aligned}$$

**Step 2:** Calculation of main effect totals and interactions totals is made by using the following contrasts

$$\begin{aligned} N &= [npk] - [pk] + [nk] - [k] + [np] - [p] + [n] - [1] = 369 \\ P &= [npk] + [pk] - [nk] - [k] + [np] + [p] - [n] - [1] = 2221 \\ K &= [npk] + [pk] + [nk] + [k] - [np] - [p] - [n] - [1] = 1727 \\ NP &= [npk] - [pk] - [nk] + [k] + [np] - [p] - [n] + [1] = 81 \\ NK &= [npk] - [pk] + [nk] - [k] - [np] + [p] - [n] + [1] = 159 \\ PK &= [npk] + [pk] - [nk] - [k] - [np] - [p] + [n] + [1] = -533 \\ NPK &= [npk] - [pk] - [nk] + [k] - [np] + [p] + [n] - [1] = -13 \end{aligned}$$

We now obtain factorial effects (main effects and interactions) and S.S. due to factorial effects

$$\text{Factorial effects} = \frac{\text{Factorial effect Total}}{r.2^{n-1} (= 12)}$$

$$\text{Factorial effect SS} = \frac{(\text{Factorial effect Total})^2}{r.2^n (= 24)}$$

Factorial Effects:

$$N = 30.75, P = 185.083, K = 143.917, NP = 6.75, NK = 13.25, PK = -44.417, NPK = -1.083$$

SS due to Factorial effects

$$\text{SS due to } N = 5673.375; \quad \text{SS due to } P = 205535.042$$

SS due to  $K = 124272.0417$ ; SS due to  $NP = 273.375$   
 SS due to  $NK = 1053.375$ ; SS due to  $PK = 11837.0417$   
 SS due to  $NPK = 7.04166$ .

**Step 3:** We now obtain  $M.S.$  by dividing  $S.S.$  's by respective  $d.f.$

**Step 4:** Construct ANOVA table as given below:

ANOVA				
Source of Variation	Degrees of Freedom (d.f)	Sum of Squares (S.S)	Mean Squares (M.S.)	Variance Ratio F
Replications	$r-1 = 2$	520.333	260.167	0.9918
Treatments	$2^3-1=7$	348651.291	49807.3273	189.8797*
N	$(s-1)=1$	5673.375	5673.375	21.6285*
P	1	205535.042	205535.042	783.5582*
K	1	124272.042	124272.042	473.7606*
NP	1	273.375	273.375	1.0422
NK	1	1053.375	1053.375	4.0158
PK	1	11837.041	11837.041	45.1262*
NPK	1	7.0412	7.0412	0.02684
Error	$(r-1)(2^n-1)=14$	3672.337	262.3098	
Total	$r.2^n-1=23$	352843.958		

(\* indicates significance at 5% level of significance).

**Step 5:**  $S.E$  estimate of difference between means of levels of single factor averaged over

levels of all other factors =  $\sqrt{\frac{MSE}{r.2^{n-2}}} = 6.612$

$S.E$  estimate of difference between means for all level combinations of two factors

averaged over levels of all other factors =  $\sqrt{\frac{MSE}{r.2^{n-3}}} = 9.351$ .

$t_{0.05}$  at 14  $d.f.$  = 2.145. Accordingly critical differences ( $C.D.$ ) can be calculated.

## 2.2 Experiments with factors each at three levels

When factors are taken at three levels instead of two, the scope of an experiment increases. It becomes more informative. A study to investigate if the change is linear or quadratic is possible when the factors are at three levels. The more the number of levels the better, yet the number of the levels of the factors cannot be increased too much as the size of the experiment increases too rapidly with them. Let us begin with two factors  $A$  and  $B$ , each at three levels say 0, 1 and 2 ( $3^2$ -factorial experiment). The treatment combinations are

- 00 =  $a_0b_0 = (1)$  ; A and B both at first levels
- 10 =  $a_1b_0 = a$  ; A is at second level and B is at first level
- 20 =  $a_2b_0 = a^2$  ; A is at third level and B is at first level
- 01 =  $a_0b_1 = b$  ; A is at first level and B is at second level
- 11 =  $a_1b_1 = ab$  ; A and B both at second level
- 21 =  $a_2b_1 = a^2b$  ; A is at third level and B is at second level
- 02 =  $a_0b_2 = b^2$  ; A is at first level and B is at third level
- 12 =  $a_1b_2 = ab^2$  ; A is at second level and B is at third level
- 22 =  $a_2b_2 = a^2b^2$  ; A and B both at third level

Any standard design can be adopted for the experiment. The main effects  $A, B$  can respectively be divided into linear and quadratic components each with  $1 d.f.$  as  $A_L, A_Q, B_L$  and  $B_Q$ . Accordingly  $AB$  can be partitioned into four components as  $A_L B_L, A_L B_Q, A_Q B_L, A_Q B_Q$ , each with one  $df$ . The coefficients of the treatment combinations to obtain the above effects are given as

Treatment totals → Factorial effects ↓	[1]	[a]	[a <sup>2</sup> ]	[b]	[ab]	[a <sup>2</sup> b]	[b <sup>2</sup> ]	[ab <sup>2</sup> ]	[a <sup>2</sup> b <sup>2</sup> ]	Divisor
$M$	+1	+1	+1	+1	+1	+1	+1	+1	+1	$9r=rx3^2$
$A_L$	-1	0	+1	-1	0	+1	-1	0	+1	$6r=rx2x3$
$A_Q$	+1	-2	+1	+1	-2	+1	+1	-2	+1	$18r=6x3$
$B_L$	-1	-1	-1	0	0	0	+1	+1	+1	$6r=rx2x3$
$A_L B_L$	+1	0	-1	0	0	0	-1	0	+1	$4r=rx2x2$
$A_Q B_L$	-1	+2	-1	0	0	0	+1	-2	+1	$12r=rx6x2$
$B_Q$	+1	+1	+1	-2	-2	-2	+1	+1	+1	$18r=rx3x6$
$A_L B_Q$	-1	0	+1	+2	0	-2	-1	0	+1	$12r=rx2x6$
$A_Q B_Q$	+1	-2	+1	-2	+4	-2	+1	-2	+1	$36r=rx6x6$

The rule to write down the coefficients of the linear (quadratic) main effects is to give a coefficient as  $+1 (+1)$  to those treatment combinations containing the third level of the corresponding factor, coefficient as  $0(-2)$  to the treatment combinations containing the second level of the corresponding factor and coefficient as  $-1(+1)$  to those treatment combinations containing the first level of the corresponding factor. The coefficients of the treatment combinations for two factor interactions are obtained by multiplying the corresponding coefficients of two main effects. The various factorial effect totals are given as

$$\begin{aligned}
 [A_L] &= +1[a^2b^2]+0[ab^2]-1[b^2]+1[a^2b]+0[ab]-1[b]+1[a^2]+0[a]-1[1] \\
 [A_Q] &= +1[a^2b^2]-2[ab^2]+1[b^2]+1[a^2b]-2[ab]+1[b]+1[a^2]-2[a]+1[1] \\
 [B_L] &= +1[a^2b^2]+1[ab^2]+1[b^2]+0[a^2b]+0[ab]+0[b]-1[a^2]-1[a]-1[1] \\
 [A_L B_L] &= +1[a^2b^2]+0[ab^2]-1[b^2]+0[a^2b]+0[ab]+0[b]-1[a^2]+0[a]-1[1] \\
 [A_Q B_L] &= +1[a^2b^2]-2[ab^2]+1[b^2]+0[a^2b]+0[ab]+0[b]-1[a^2]+2[a]-1[1] \\
 [B_Q] &= +1[a^2b^2]+1[ab^2]+1[b^2]-2[a^2b]-2[ab]-2[b]-1[a^2]-1[a]-1[1] \\
 [A_L B_Q] &= +1[a^2b^2]+0[ab^2]-1[b^2]-2[a^2b]+0[ab]+2[b]+1[a^2]+0[a]-1[1] \\
 [A_Q B_Q] &= +1[a^2b^2]-2[ab^2]+1[b^2]-2[a^2b]+4[ab]-2[b]+1[a^2]-2[a]+1[1]
 \end{aligned}$$

The sum of squares due to various factorial effects is given by

$$\begin{aligned}
 SSA_L &= \frac{[A_L]^2}{r.2.3}; & SSA_Q &= \frac{[A_Q]^2}{r.6.3}; & SSB_L &= \frac{[B_L]^2}{r.3.2}; & SSA_L B_L &= \frac{[A_L B_L]^2}{r.2.2}; \\
 SSA_Q B_L &= \frac{[A_Q B_L]^2}{r.6.2}; & SSB_Q &= \frac{[B_Q]^2}{r.3.6}; & SSA_L B_Q &= \frac{[A_L B_Q]^2}{r.2.6}; & SSA_Q B_Q &= \frac{[A_Q B_Q]^2}{r.6.6};
 \end{aligned}$$

If a randomized complete block design is used with  $r$ -replications then the outline of analysis of variance is

ANOVA			
Source of Variation	D.F.	S.S.	M.S.
Replications	$r-1$	$SSR$	$MSR=SSR/(r-1)$
Treatments	$3^2-1=8$	$SST$	$MST=SST/8$

Designs for Factorial Experiments

A		2	SSA	$MSA=SSA/2$
	$A_L$		1 $SSA_L$	$MSA_L=SSA_L$
	$A_Q$		1 $SSA_Q$	$MSA_Q=SSA_Q$
B		2	SSB	$MSB=SSB/2$
	$B_L$		1 $SSB_L$	$MSB_L=SSB_L$
	$B_Q$		1 $SSB_Q$	$MSB_Q=SSB_Q$
AB		4	SSAB	$MSAB=SSAB/2$
	$A_L B_L$		1 $SSA_L B_L$	$MSA_L B_L=SSA_L B_L$
	$A_Q B_L$		1 $SSA_Q B_L$	$MSA_Q B_L=SSA_Q B_L$
	$A_L B_Q$		1 $SSA_L B_Q$	$MSA_L B_Q=SSA_L B_Q$
	$A_Q B_Q$		1 $SSA_Q B_Q$	$MSA_Q B_Q=SSA_Q B_Q$
Error		$(r-1)(3^2-1)$ $=8(r-1)$	SSE	$MSE=SSE/8(r-1)$
<b>Total</b>		$r.3^2-1=9r-1$	<b>TSS</b>	

In general, for  $n$  factors each at 3 levels, the sum of squares due to any linear (quadratic) main effect is obtained by dividing the square of the linear (quadratic) main effect total by  $r.2.3^{n-1}$  ( $r.6.3^{n-1}$ ). Sum of squares due to a  $p$ -factor interaction is given by taking the square of the total of the particular interaction component divided by  $r.(a_1 a_2 \dots a_p).3^{n-p}$ , where  $a_1, a_2, \dots, a_p$  are taken as 2 or 6 depending upon whether the effect of a particular factor is linear or quadratic.

**Example 2:** A  $3^2$  experiment was conducted to study the effects of the two factors, viz., Nitrogen (N) and Phosphorus (P) each at three levels 0,1,2 on sugar beets. Two replications of nine plots each were used. The table shows the plan and the percentage of sugar (approximated to nearest whole number).

**Plan and percentage of sugar of a  $3^2$  experiment**

Replication	Treatment		% of sugar	Replication	Treatment		% of sugar
	N	P			N	P	
I	0	1	14	II	1	2	20
	2	0	15		1	0	19
	0	0	16		1	1	17
	2	1	15		0	0	18
	0	2	16		2	1	19
	1	2	18		0	1	16
	1	1	17		0	2	16
	1	0	19		2	2	19
	2	2	17		2	0	16

Analyze the data.

**Analysis:**

**Step 1:** In order to obtain the sum of squares due to replications, due to treatments and total sum of squares arrange the data in a Replication  $\times$  Treatment combinations table as follows:

Repl.	Treatment Combinations									Total
	1	n	n <sup>2</sup>	p	np	n <sup>2</sup> p	p <sup>2</sup>	np <sup>2</sup>	n <sup>2</sup> p <sup>2</sup>	
	00	10	20	01	11	21	02	12	22	
1	16	19	15	14	17	15	16	18	17	147 ( $R_1$ )
2	18	19	16	16	17	19	16	20	19	160 ( $R_2$ )
<b>Total</b>	34	38	31	30	34	34	32	38	36	307 ( $G$ )
	( $T_1$ )	( $T_2$ )	( $T_3$ )	( $T_4$ )	( $T_5$ )	( $T_6$ )	( $T_7$ )	( $T_8$ )	( $T_9$ )	

Grand Total = 307, Number of observations ( $n$ ) =  $r.3^2 = 18$ .

$$\text{Correction Factor (C.F.)} = \frac{(307)^2}{18} = 5236.0556$$

$$\text{Total S.S. (TSS)} = 16^2 + 18^2 + \dots + 17^2 + 19^2 - 5236.0556 = 48.9444$$

$$\begin{aligned} \text{Replication SS (SSR)} &= \frac{R_1^2 + R_2^2}{9} - C.F. \\ &= \frac{147^2 + 160^2}{9} - 5236.0556 = 9.3888 \end{aligned}$$

$$\begin{aligned} \text{Treatment SS (SST)} &= \frac{\text{Sum}(\text{treatment totals})^2}{r} - C.F. \\ &= \frac{34^2 + 38^2 + \dots + 38^2 + 36^2}{2} - 5236.0556 = 32.4444 \end{aligned}$$

$$\text{Error SS} = \text{Total SS} - \text{Replication SS} - \text{Treatment SS} = 7.1112$$

**Step 2:** Obtain various factorial effects totals

$$\begin{aligned} [N_L] &= +1[n^2p^2] + 0[np^2] - 1[p^2] + 1[n^2p] + 0[np] - 1[p] + 1[n^2] + 0[n] - 1[1] = 5 \\ [N_Q] &= +1[n^2p^2] - 2[np^2] + 1[p^2] + 1[n^2p] - 2[np] + 1[p] + 1[n^2] - 2[n] + 1[1] = -23 \\ [P_L] &= +1[n^2p^2] + 1[np^2] + 1[p^2] + 0[n^2p] + 0[np] + 0[p] - 1[n^2] - 1[n] - 1[1] = 3 \\ [N_L P_L] &= +1[n^2p^2] + 0[np^2] - 1[p^2] + 0[n^2p] + 0[np] + 0[p] - 1[n^2] + 0[n] + 1[1] = 7 \\ [N_Q P_L] &= +1[n^2p^2] - 2[np^2] + 1[p^2] + 0[n^2p] + 0[np] + 0[p] - 1[n^2] + 2[n] - 1[1] = 3 \\ [P_Q] &= +1[n^2p^2] + 1[np^2] + 1[p^2] - 2[n^2p] - 2[np] - 2[p] + 1[n^2] + 1[n] + 1[1] = 13 \\ [N_L P_Q] &= +1[n^2p^2] + 0[np^2] - 1[p^2] - 2[n^2p] + 0[np] + 2[p] + 1[n^2] + 0[n] - 1[1] = -7 \\ [N_Q P_Q] &= +1[n^2p^2] - 2[np^2] + 1[p^2] - 2[n^2p] + 4[np] - 2[p] + 1[n^2] - 2[n] + 1[1] = -11 \end{aligned}$$

**Step 3:** Obtain the sum of squares due to various factorial effects

$$SS_{N_L} = \frac{[N_L]^2}{r.2.3} = \frac{5^2}{12} = 2.0833; \quad SS_{N_Q} = \frac{[N_Q]^2}{r.6.3} = \frac{(-23)^2}{36} = 14.6944;$$

$$SS_{P_L} = \frac{[P_L]^2}{r.3.2} = \frac{3^2}{12} = 0.7500; \quad SS_{N_L P_L} = \frac{[N_L P_L]^2}{r.2.2} = \frac{7^2}{8} = 6.1250;$$

$$SS_{N_Q P_L} = \frac{[N_Q P_L]^2}{r.6.2} = \frac{3^2}{24} = 0.375; \quad SS_{P_Q} = \frac{[P_Q]^2}{r.3.6} = \frac{13^2}{36} = 4.6944;$$

$$SS_{N_L P_Q} = \frac{[N_L P_Q]^2}{r.2.6} = \frac{(-7)^2}{24} = 2.0417; \quad SS_{N_Q P_Q} = \frac{[N_Q P_Q]^2}{r.6.6} = \frac{(-11)^2}{72} = 1.6806;$$

**Step 4:** Construct the ANOVA table as given above and test the significance of the various factorial effects:

ANOVA				
Source of Variation	D.F.	S.S.	M.S.	F
Replications	1	9.3888	9.3888	10.5623*
Treatments	8	32.4444	4.0555	4.5624*
N	2	16.7774	8.3887	9.4371*
N <sub>L</sub>	1	2.0833	2.0833	2.3437
N <sub>Q</sub>	1	14.6944	14.6944	16.5310*

<i>P</i>		2	5.4444	2.7222	3.0624
	<i>P<sub>L</sub></i>	1	0.7500	0.7500	0.8437
	<i>P<sub>Q</sub></i>	1	4.6944	4.6944	5.2811
<i>NP</i>		4	10.2223	2.5556	2.875
	<i>N<sub>L</sub>P<sub>L</sub></i>	1	6.1250	6.1250	6.8905*
	<i>N<sub>Q</sub>P<sub>L</sub></i>	1	0.3750	0.3750	0.4219
	<i>N<sub>L</sub>P<sub>Q</sub></i>	1	2.0417	2.0417	2.2968
	<i>N<sub>Q</sub>P<sub>Q</sub></i>	1	1.6806	1.6806	1.8906
<i>Error</i>		8	7.1112	0.8889	
<i>Total</i>		17	48.9444		

(\* indicates the significance at 5% level of significance)

### 2.3 Yates Algorithm

We now describe below a general procedure of computing the factorial effects:

**Step 1:** Write the treatment combinations in the lexicographic order, *i.e.*, first vary the levels of the first factor from 0 to  $s_1 - 1$  by keeping fixed the levels of other  $n - 1$  factors at level 0. Then vary the levels of the second factor from 1 to  $s_2 - 1$  levels in each of the first  $s_1$  treatment combinations by keeping the levels of factors 3 to  $n$  factors at 0 levels so as to get  $s_1 \times s_2$  treatment combinations; then vary the levels of the third factor from 1 to  $s_3 - 1$  by keeping the levels of factors 4 to  $n$  at 0 levels in the earlier  $s_1 \times s_2$  treatment combinations

and repeat the process till you get all the  $\prod_{i=1}^n s_i$  treatment combinations. For example, if

there are three factors, first factor at 3 levels, second factor at 4 levels and third factor at 5 levels. Then  $3 \times 4 \times 5 = 60$  treatment combinations are:

000, 100, 200, 010, 110, 210, 020, 120, 220, 030, 130, 230, 001, 101, 201, 011, 111, 211, 021, 121, 221, 031, 131, 231, 002, 102, 202, 012, 112, 212, 022, 122, 222, 032, 132, 232, 003, 103, 203, 013, 113, 213, 023, 123, 223, 033, 133, 233, 004, 104, 204, 014, 114, 214, 024, 124, 224, 034, 134, 234.

Write all these treatment combinations in the first column and in the second column write the corresponding treatment totals.

**Step 2:** Divide the observations in the second column in groups such that each group has  $s_1$  observations. Then we add the observations in each of these  $s_1$  groups in the third column, then we repeat the process of linear component of the main effect of the first factor with these groups and append the third column, repeat the process for quadratic effects and so on till the polynomial upto the order of  $s_1 - 1$ . For example, if the factor is at two levels, then we make the groups of two observations each, and first half of the third column is filled with sum of observations in these groups and second half with the differences of the second observation and the first observation in each group. If the factor is at three levels, we make the groups of three observations each, and one third column is filled with the sum of observations in these groups, next one third by using the linear component, say  $-1, 0, 1$ , *i.e.*, by taking the difference of the third observation and first observation in each group and rest one third is filled by using the quadratic component  $1, -2, 1$ , *i.e.*, by adding the first and third observation in each group and subtracting the twice of the second observation from this sum. If the factor is at four levels, we make the groups of four observations each, the first quarter of the next column is filled by sum of these observations in each of the groups, next quarter is filled by using the linear component  $-3, -1, 1, 3$ , *i.e.*, by adding the third observation and 3 times the fourth



observation from each group and then subtracting the sum of second observation and three times the first observation from this sum. Next quarter is filled using the quadratic component  $1, -1, -1, 1$ , i.e. first observation minus second observation minus third observation plus fourth observation of each of the groups and last quarter is filled by using the cubic component say  $-1, 3, -3, 1$ , i.e.  $[-(\text{first observation}) + 3(\text{second observation}) - 3(\text{third observation}) + \text{fourth observation}]$  from each group, and so on.

In the third column, divide the observations into groups such that each group contains  $s_2$  observations and then use these groups to obtain the fourth column as in second column. In the fourth column divide the observations into groups of  $s_3$  observations each and so on. Repeat the process for all the  $n$  factors.

If all the factors are at same levels, then perform same operation on all the  $n$  columns.

For illustration, the various factorial effect totals in the Example 2, where each of the three factors is at 2 levels each, can be obtained as follows:

Treatment combinations (1)	Treatment totals (2)	Operation as per first factor (3)	Operation as per second factor (4)	Operation as per third factor (5)
000 (1)	294 = I	617 = I + II	2611	6949 = G
100 n	323 = II	1994 = III + IV	4338	369 = [N]
010 p	959 = III	1747 = V + VI	105	2221 = [P]
110 np	1035 = IV	2591 = VII + VIII	264	81 = [NP]
001 k	816 = V	29 = II - I	1377	1727 = [K]
101 nk	931 = VI	76 = IV - III	844	159 = [NK]
011 pk	1221 = VII	115 = VI - V	47	-533 = [PK]
111 npk	1370 = VIII	149 = VIII - VII	34	-13 = [NPK]

For Example 2, the various factorial effects totals can be obtained as given in the following table

Treatment combinations (1)	Treatment totals (2)	Operation as per first factor (3)	Operation as per second factor (4)
00 (1)	34 = I	103 = I + II + III	307 = G
10 (n)	38 = II	98 = IV + V + VI	5 = $N_L$
20 ( $n^2$ )	31 = III	106 = VII + VIII + IX	-23 = $N_Q$
01 (p)	30 = IV	-3 = III - I	3 = $P_L$
11 (np)	34 = V	4 = VI - IV	7 = $N_L P_L$
21 ( $n^2 p$ )	34 = VI	4 = IX - VII	3 = $N_Q P_L$
02 ( $p^2$ )	32 = VII	-11 = III - 2II + I	13 = $P_Q$
12 ( $np^2$ )	38 = VIII	-4 = VI - 2V + IV	-7 = $N_L P_Q$
22 ( $n^2 p^2$ )	36 = IX	-8 = IX - 2VIII + VII	-11 = $N_Q P_Q$

**Remark:** The analysis demonstrated so far is computationally feasible for the situation when large number of factors is experimented with smaller number of levels. However, usual tabular method of analysis can be employed for the situations when there are few factors with more number of levels.

### 3. Confounding in Factorial Experiments

When the number of factors and/or levels of the factors increase, the number of treatment combinations increase very rapidly and it is not possible to accommodate all these treatment combinations in a single homogeneous block. For example, a  $2^5$  factorial would have 32 treatment combinations and blocks of 32 plots are quite big to ensure homogeneity within them. In such a situation it is desirable to form blocks of size smaller than the total number of treatment combinations (incomplete blocks) and, therefore, have more than one block per replication. The treatment combinations are then allotted randomly to the blocks within the replication and the total number of treatment combinations is grouped into as many groups as the number of blocks per replication.

A consequence of such an arrangement is that the block contrasts become identical to some of the interaction component contrasts. For example, consider a  $2^4$  factorial experiment to be conducted in two blocks of size 8 each per replication. The two blocks in a single replication are the following:

<b>Block - I</b>	<b>Block - II</b>
<b>treatment combination</b>	<b>treatment combination</b>
A B C D	A B C D
0 0 0 0 (1)	1 0 0 0 a
1 1 0 0 ab	0 1 0 0 b
1 0 1 0 ac	0 0 1 0 c
1 0 0 1 ad	0 0 0 1 d
0 1 1 0 bc	1 1 1 0 abc
0 1 0 1 bd	1 1 0 1 abd
0 0 1 1 cd	1 0 1 1 acd
1 1 1 1 abcd	0 1 1 1 bcd

It may easily be verified that the block contrast is identical with the contrast for the interaction ABCD, *i.e.*,  $0000+1100+1010+1001+0110+0101+0011+1111-1000-0100-0010-0001-1110-1101-1011-0111$ . Thus, the interaction ABCD gets confounded with block effects and it is not possible to separate out the two effects.

Evidently the interaction confounded has been lost but the other interactions and main effects can now be estimated with better precision because of reduced block size. This device of reducing the block size by taking one or more interactions contrasts identical with block contrasts is known as **confounding**. Preferably only higher order interactions with three or more factors are confounded, because these interactions are less important to the experimenter. As an experimenter is generally interested in main effects and two factor interactions, these should not be confounded as far as possible. The designs for such confounded factorials are incomplete block designs. However usual incomplete block designs for single factor experiments cannot be adopted, as the contrasts of interest in two kinds of experiments are different. The treatment groups are first allocated at random to the different blocks. The treatments allotted to a block are then distributed at random to its different units.

When there are two or more replications in the design and if the same set of interaction components is confounded in all the replications, then confounding is called **complete** and if different sets of interactions are confounded in different replications, confounding is called **partial**. In complete confounding all the information on confounded interactions is

lost. However, in partial confounding, the information on confounded interactions can be recovered from those replications in which these are not confounded.

**Advantages of Confounding**

- It reduces the experimental error considerably by stratifying the experimental material into homogeneous subsets or subgroups. The removal of the variation among incomplete blocks (freed from treatments) within replications results in smaller error mean square as compared with a RCB design, thus making the comparisons among some treatment effects more precise.

**Disadvantages of Confounding**

- In the confounding scheme, the increased precision is obtained at the cost of sacrifice of information (partial or complete) on certain relatively unimportant interactions.
- The confounded contrasts are replicated fewer times than are the other contrasts and as such there is loss of information on them and these can be estimated with a lower degree of precision as the number of replications for them is reduced.
- An indiscriminate use of confounding may result in complete or partial loss of information on the contrasts or comparisons of greatest importance. As such the experimenter should confound only those treatment combinations or contrasts that are of relatively less or of no importance at all.
- The algebraic calculations are usually more difficult and the statistical analysis is complex, especially when some of the units (observations) are missing. In this package, the attempt has been made to ease this problem.

**3.1 Confounding in 2<sup>3</sup> Experiment**

To make the exposition simple, we consider a small factorial experiment 2<sup>3</sup>. Let the three factors be A, B, C each at two levels.

Effects→	A	B	C	AB	AC	BC	ABC
Treat. Combinations↓							
(1)	-	-	-	+	+	+	-
(a)	+	-	-	-	-	+	+
(b)	-	+	-	-	+	-	-
(ab)	+	+	-	+	-	-	-
(c)	-	-	+	+	-	-	+
(ac)	+	-	+	-	+	-	-
(bc)	-	+	+	-	-	+	-
(abc)	+	+	+	+	+	+	+

The various effects are given by

$$\begin{aligned}
 A &= (abc) + (ac) + (ab) + (a) - (bc) - (c) - (b) - (1) \\
 B &= (abc) + (bc) + (ab) + (b) - (ac) - (c) - (a) - (1) \\
 C &= (abc) + (bc) + (ac) + (c) - (ab) - (b) - (a) - (1) \\
 AB &= (abc) + (c) + (ab) + (1) - (bc) - (ac) - (b) - (a) \\
 AC &= (abc) + (ac) + (b) + (1) - (bc) - (c) - (ab) - (a) \\
 BC &= (abc) + (bc) + (a) + (1) - (ac) - (c) - (ab) - (b) \\
 ABC &= (abc) + (c) + (b) + (a) - (bc) - (ac) - (ab) - (1)
 \end{aligned}$$

Suppose that the experimenter decides to use two blocks of 4 units (plots) per replication and that the highest order interaction ABC is confounded. Thus, in order to confound the interaction ABC with blocks all the treatment combinations with positive sign are allocated at random in one block and those with negative signs in the other block. Thus

the following arrangement gives ABC confounded with blocks and hence the entire information is lost on ABC in this replication.

**Replication I**

Block 1: (1) (ab) (ac) (bc)  
 Block 2 : (a) (b) (c) (abc)

We observe that the contrast estimating ABC is identical to the contrast estimating block effects. If the same interaction ABC is confounded in all the other replications, then the interaction is said to be completely confounded and we cannot recover any information on the interaction ABC through such a design. For the other six factorial effects viz. A, B, C, AB, AC, BC there are two treatment combinations with a positive sign and two treatment combinations with a negative sign in each of the two blocks and hence these differences are not influenced among blocks and can thus be estimated and tested as usual without any difficulty.

Similarly if we want to confound AB, then the two blocks will consists of

Block 1 (abc) (c) (ab) (1)  
 Block 2 (bc) (ac) (b) (a)

Here interaction AB is confounded with block effects whereas all other effects A, B, C, AC, BC and ABC can be estimated orthogonally.

**3.2 Partial confounding**

When different interactions are confounded in different replications, the interactions are said to be partially confounded. Consider again the 2<sup>3</sup> factorial experiment with each replicate divided into two blocks of 4 units each. It is not necessary to confound the same interaction in all the replications and several factorial effects may be confounded in one single experiment. For example, the following plan confounds the interaction ABC, AB, BC and AC in replications I, II, III and IV respectively.

Rep. I		Rep. II		Rep. III		Rep. IV	
Block 1	Block 2	Block 3	Block 4	Block 5	Block 6	Block 7	Block 8
(abc)	(ab)	(abc)	(ac)	(abc)	(ab)	(abc)	(ab)
(a)	(ac)	(c)	(bc)	(bc)	(ac)	(ac)	(bc)
(b)	(bc)	(ab)	(a)	(a)	(b)	(b)	(a)
(c)	(1)	(1)	(b)	(1)	(c)	(1)	(c)

In the above arrangement, the main effects A, B and C are orthogonal to block contrasts. The interaction ABC is completely confounded with blocks in replication I, but in the other three replications the interaction ABC is orthogonal to blocks and consequently an estimate of ABC may be obtained from replicates II, III and IV. Similarly it is possible to recover information on the other confounded interactions AB (from replications I, III, IV), BC (from replications I, II, IV) and AC (from replications I, II, III). Since the partially confounded interactions are estimated from only a portion of the observations, they are determined with a lower degree of precision than the other effects.

**3.3 Construction of a Confounded Factorial**

Given a set of interactions confounded, the blocks of the design can be constructed and vice-versa i.e., if the design is given the interactions confounded can be identified.

**3.4 Given a set of interactions confounded, how to obtain the blocks?**

The blocks of the design pertaining to the confounded interaction can be obtained by solving the equations obtained from confounded interaction. We illustrate this through an example.

**Example 3:** Construct a design for  $2^5$  factorial experiment in  $2^3$  plots per block confounding interactions ABD, ACE and BCDE.

Let  $x_1, x_2, x_3, x_4$  and  $x_5$  denote the levels (0 or 1) of each of the 5 factors A, B, C, D and E. Solving the following equations would result in different blocks of the design.

For interaction ABD:  $x_1 + x_2 + x_4 = 0, 1$

For interaction ACE :  $x_1 + x_3 + x_5 = 0, 1$

The interactions ABD and ACE are independent and BCDE is a generalized interaction. In other words, a solution of the above two equations will also satisfy the equation  $x_1 + x_2 + x_3 + x_4 = 0, 1$ . Treatment combinations satisfying the following solutions of above equations will generate the required four blocks

(0, 0)          (0, 1)          (1, 0)          (1, 1)

The solution (0, 0) will give the key block (A key block is one that contains one of the treatment combination of factors, each at lower level).

There will be  $\frac{2^5}{2^3} = 4$  blocks per replication. The key block is as obtained below

A	B	C	D	E	
1	1	1	0	0	abc
1	1	0	0	1	abe
1	0	1	1	0	acd
1	0	0	1	1	ade
0	1	1	1	1	bcde
0	1	0	1	0	bd
0	0	1	0	1	ce
0	0	0	0	0	(1)

Similarly we can write the other blocks by taking the solutions of above equations as (0, 1) (1, 0) and (1, 1).

**3.5 Given a block, how to find the interactions confounded?**

The first step in detecting the interactions confounded in blocking is to select the key block. If the key block is not given, it is not difficult to obtain it. Select any treatment combination in the given block; multiply all the treatment combinations in the block by that treatment combination and we get the key block. From the key block we know the number of factors as well as the block size. Let it be  $n$  and  $k$ . We know then that the given design belongs to the  $2^n$  factorial in  $2^f$  plots per block. The next step is to search out a unit matrix of order  $r$ . From these we can find the interaction confounded. We illustrate this through an example.

**Example 4:** Given the following block, find out the interactions confounded.

(acde), (bcd), (e), (abec), (ad), (bde), (ab), (c)

Since the given block is not the key block we first obtain the key block by multiplying every treatment combination of the given block by e. We get the following block:

(acd), (bcde), (1), (abc), (ade), (bd), (abe), (ce)

This is the key block as it includes (1). It is obvious that the factorial experiment involves five factors and has  $2^3 (=8)$  plots per block. Hence, the given design is  $(2^5, 2^3)$ .

	A	B	C	D	E
	1	0	1	1	0
	0	1	1	1	1
	0	0	0	0	0
	1	1	1	0	0
*	1	0	0	1	1
*	0	1	0	1	0
	1	1	0	0	1
*	0	0	1	0	1

\* indicates the rows of a unit matrix of order 3.

A	B	C	D	E
1	0	0	1(= $\alpha_1$ )	1(= $\beta_1$ )
0	1	0	1(= $\alpha_2$ )	0(= $\beta_2$ )
0	0	1	0(= $\alpha_3$ )	1(= $\beta_3$ )

The interaction confounded is  $A^{\alpha_1}B^{\alpha_2}C^{\alpha_3}D$ ,  $A^{\beta_1}B^{\beta_2}C^{\beta_3}E$ . Here ABD and ACE are independent interactions confounded and BCDE is obtained as the product of these two and is known as generalized interaction.

### 3.6 General rule for confounding in $2^n$ series

Let the design be  $(2^n, 2^r)$  i.e.  $2^n$  treatment combinations arranged in  $2^r$  plots per block. Number of treatment combinations =  $2^n$ , Block size =  $2^r$ , Number of blocks per replication =  $2^{n-r}$ , Total number of interactions confounded =  $2^{n-r} - 1$ , Number of independent interactions confounded =  $n - r$ , Generalized interactions confounded =  $(2^{n-r} - 1) - (n - r)$ .

### 3.7 Analysis

For carrying out the statistical analysis of a  $(2^n, 2^r)$  factorial experiment in  $p$  replications, the various factorial effects and their S.S. are estimated in the usual manner with the modification that for **completely confounded** interactions neither the S.S due to confounded interaction is computed nor it is included in the ANOVA table. The confounded component is contained in the  $(p2^{n-r} - 1)$  d.f. due to blocks. The splitting of the total degrees of freedom is as follows:

Source of Variation	Degrees of Freedom
Replication	$p - 1$
Blocks within replication	$p(2^{n-r} - 1)$
Treatments	$(2^n - 1) - (2^{n-r} - 1)$
Error	By subtraction
Total	$p2^n - 1$

The  $d.f$  due to treatment has been reduced by  $2^{n-r}-1$  as this is the total  $d.f$  confounded per block.

### 3.8 Partial Confounding

In case of partial confounding, we can estimate the effects confounded in one replication from the other replications in which it is not confounded. In  $(2^n, 2^r)$  factorial experiment with  $p$  replications, following is the splitting of  $d.f$ 's.

Source of Variation	Degrees of Freedom
Replication	$p-1$
Blocks within replication	$p(2^{n-r}-1)$
Treatments	$2^n-1$
Error	By subtraction
Total	$p2^n-1$

The S.S. for confounded effects are obtained from only those replications where the given effect is not confounded. From practical point of view, the S.S. for all the effects including the confounded effects is obtained as usual and then some adjustment factor (A.F) is applied to the confounded effects. The adjusting factor for any confounded effect is computed as follows:

- (i) Note the replication in which the given effect is confounded
- (ii) Note the sign of (1) in the corresponding algebraic expression of the effect to the confounded. If the sign is positive then

$$A.F = [\text{Total of the block containing (1) of replicate in which the effect is confounded}] - [\text{Total of the block not containing (1) of the replicate in which the effect is confounded}] = T_1 - T_2.$$

If the sign is negative, then  $A.F = T_2 - T_1$ .

This adjusting factor will be subtracted from the factorial effects totals of the confounded effects obtained.

**Example 5:** Analyze the following  $2^3$  factorial-experiment conducted in two blocks of 4 plots per replication, involving three fertilizers N, P, K, each at two levels:

Replication I		Replication II		Replication III	
Block 1	Block 2	Block 3	Block 4	Block 5	Block 6
(np)	(p)	(1)	(np)	(pk)	(n)
101	88	125	115	75	53
(npk)	(n)	(npk)	(k)	(nk)	(npk)
111	90	95	95	100	76
(1)	(pk)	(nk)	(pk)	(1)	(p)
75	115	80	90	55	65
(k)	(nk)	(p)	(n)	(np)	(k)
55	75	100	80	92	82

**Step 1:** Identify the interactions confounded in each replication. Here, each replication has been divided into two blocks and one effect has been confounded in each replication. The effects confounded are

Replication I  $\rightarrow$  NP; Replicate II  $\rightarrow$  NK; Replicate III  $\rightarrow$  NPK

**Step 2:** Obtain the blocks S.S. and Total S.S.

$$\text{S.S. due to Blocks} = \sum_{i=1}^6 \frac{B_i^2}{4} - \text{C.F} = 2506$$

$$\text{Total S.S.} = \sum (\text{Obs.})^2 - \text{C.F} = 8658$$

**Step 3:** Obtain the sum of squares due to all the factorial effects other than the confounded effects.

Treatment Combinations	Total Yield	Factorial Effects	Sum of Squares (S.S) = [Effect] <sup>2</sup> / 2 <sup>3</sup> .r
(1)	255	G=0	
n	223	[N]=48	96 = S <sub>N</sub> <sup>2</sup>
p	253	[P]=158	1040.17 = S <sub>P</sub> <sup>2</sup>
np	308	[NP]=66	-
k	232	[K]=10	4.17 = S <sub>K</sub> <sup>2</sup>
nk	255	[NK]=2	-
pk	280	[PK]=-8	2.67 = S <sub>PK</sub> <sup>2</sup>
npk	282	[NPK]=-108	-

Total for the interaction NP is given by

$$[\text{NP}] = [\text{npk}] - [\text{pk}] - [\text{nk}] + [\text{k}] + [\text{np}] - [\text{p}] - [\text{n}] + [1]$$

Here the sign of (1) is positive. Hence the adjusting factor (A.F) for NP, which is to be obtained from replicate 1 is given by

$$\text{A.F. for NP} = (101 + 111 + 75 + 55) - (88 + 90 + 115 + 75) = -26$$

Adjusted effect total for NP becomes, [NP\*] = [NP] - (-26) = 66 + 26 = 92.

It can easily be seen that the total of interaction NP using the above contrast from replications II and III also gives the same total *i.e.* 92.

Similarly A.F. for NK =20, A.F. for NPK = -46

Hence adjusted effect totals for NK and NPK are respectively [NK\*] =-18 and [NPK\*] = -62.

$$S_{\text{NP}}^2 = \text{S.S. due to NP} = \frac{1}{16} [\text{NP*}]^2 = 529; S_{\text{NK}}^2 = \text{S.S. due to NK} = \frac{1}{16} [\text{NK*}]^2 = 20.25$$

$$S_{\text{NPK}}^2 = \text{S.S. due to NPK} = \frac{1}{16} [\text{NPK*}]^2 = 240.25$$

$$\text{Treatment S.S.} = S_N^2 + S_P^2 + S_K^2 + S_{\text{NP}}^2 + S_{\text{NK}}^2 + S_{\text{PK}}^2 + S_{\text{NPK}}^2 = 1932.7501$$

**ANOVA**

Source	d.f.	Sum of Squares	M.S.	F
Blocks	5	2506	501	1.31
Treatments	7	1932.75	276.107	-
N	1	96.00	96.00	-
P	1	1040.16	1040.16	2.71
NP	1	529.00	529.00	1.3



K	1	4.41	4.41	-
NK	1	20.25	20.25	-
PK	1	2.66	2.66	-
NPK	1	240.25	240.25	-
Error	11	4219.24	383.57	
<b>Total</b>	<b>23</b>	<b>8658</b>		

‘-’ indicates that these ratios are less than one and hence these effects are non-significant.

From the above table it is seen that effects due to blocks, main effects due to factor N, P, and K or interactions are not significant.

**4. Confounding in 3<sup>n</sup> Series**

The concept of confounding here also is the same as in 2<sup>n</sup> series. We shall illustrate the principles of confounding in 3<sup>n</sup> in 3<sup>r</sup> plots per block with the help of a 3<sup>3</sup> experiments laid out in blocks of size 3<sup>2</sup>(=9). Let the three factors be A, B and C and the confounded interaction be ABC<sup>2</sup>. The three levels of each of the factor are denoted by 0, 1 and 2 and a particular treatment combination be x<sub>i</sub> x<sub>j</sub> x<sub>k</sub> , i, j, k = 0, 1, 2.

Number of blocks per replication = 3<sup>n-r</sup> = 3; Block size = 3<sup>r</sup> = 9; Degrees of freedom confounded per replication = 3<sup>n-r</sup> - 1 = 2.

Number of interactions confounded per replicate =  $\frac{3^{n-r} - 1}{3 - 1} = 1$ .

The treatment combinations in 3 blocks are determined by solving the following equations mod(3)

$$x_1 + x_2 + 2x_3 = 0 ; \quad x_1 + x_2 + 2x_3 = 1 ; \quad x_1 + x_2 + 2x_3 = 2$$

Block I			Block II			Block III		
A	B	C	A	B	C	A	B	C
1	0	1	1	0	0	1	0	2
0	1	1	0	1	0	0	1	2
1	1	2	1	1	1	1	1	0
2	0	2	2	0	1	2	0	0
0	2	2	0	2	1	0	2	0
2	1	0	2	1	2	2	1	1
1	2	0	1	2	2	1	2	1
2	2	1	2	2	0	2	2	2
0	0	0	0	0	2	0	0	1

**5. Confounding in s<sup>n</sup> Factorial Experiments in s<sup>r</sup> experimental units per block**

s<sup>n</sup> Factorial Experiments in s<sup>r</sup> experimental units per block are represented by (s<sup>n</sup>, s<sup>r</sup>) factorial experiments. For generation of (s<sup>n</sup>, s<sup>r</sup>), s should be a prime or prime power, i.e., s = p<sup>m</sup>, where p is prime and m is a positive integer. For the factorial experiments of the type (s<sup>n</sup>, s<sup>r</sup>) there will be s<sup>n-r</sup> blocks per replication with (s<sup>r</sup>) experimental units per block. The total number of degrees of freedom confounded per replication is s<sup>n-r</sup> - 1, while the total number of interaction components confounded per replication is  $\frac{s^{n-r} - 1}{s - 1}$  as each interaction component has (s - 1) degrees of freedom. The total number of

independent interaction components to be confounded is  $n-r$  and rest are generalized interaction components. For the  $(n-r)$  independent interaction components confounded, we have the following set of  $(n-r)$  equations as:

$$\begin{aligned} \sum_{j=1}^n p_{j1}x_j &= 0, 1, \alpha_2, \alpha_3, \dots, \alpha_{s-1} \pmod{s} \\ \sum_{j=1}^n p_{j2}x_j &= 0, 1, \alpha_2, \alpha_3, \dots, \alpha_{s-1} \pmod{s} \\ &\vdots \\ \sum_{j=1}^n p_{jk}x_j &= 0, 1, \alpha_2, \alpha_3, \dots, \alpha_{s-1} \pmod{s} \\ &\vdots \\ \sum_{j=1}^n p_{j(n-r)}x_j &= 0, 1, \alpha_2, \alpha_3, \dots, \alpha_{s-1} \pmod{s} \end{aligned}$$

where  $p_{jk}$ 's and  $0, 1, \alpha_2, \alpha_3, \dots, \alpha_{s-1}$  are the elements of the Galois Field  $s$  and  $x_1, x_2, \dots, x_n$  are the variates corresponding to the  $n$ -factors and denote the levels of the corresponding factors in the different treatment combinations. If  $m > 1$ , then mod  $s$  in the above equations should be replaced by mod  $\{p, p(x)\}$ , where  $p(x)$  is the minimal function for  $GF(p^m)$  and  $x$  is the primitive root of the  $GF(p^m)$ . These equations result into  $s^{n-r}$  different sets. Solution of each set gives one block. For example, if one wants to generate a  $(3^4, 3^2)$  factorial experiment, then the number of independent interaction components to be confounded are  $4-2=2$ . These two independent interactions are represented by:

$$\begin{aligned} p_{11}x_1 + p_{21}x_2 + p_{31}x_3 + p_{41}x_4 &= 0, 1, 2 \pmod{3} \\ p_{12}x_1 + p_{22}x_2 + p_{32}x_3 + p_{42}x_4 &= 0, 1, 2 \pmod{3} \end{aligned}$$

These sets of equations give rise to 9 combinations viz. the left hand sides satisfying (0,0); (0,1); (0,2); (1,0); (1,1); (1,2); (2,0); (2,1) and (2,2). The treatment combinations in 9 blocks in one replication are those satisfy the above combinations. The block containing the treatment combinations satisfying

$$\begin{aligned} p_{11}x_1 + p_{21}x_2 + p_{31}x_3 + p_{41}x_4 &= 0 \pmod{3} \text{ and} \\ p_{12}x_1 + p_{22}x_2 + p_{32}x_3 + p_{42}x_4 &= 0 \pmod{3} \text{ is the key block.} \end{aligned}$$

For the situations, where  $s$  is a prime power, we make use of the concept of minimal functions. For example, if one wants to generate a  $(4^2, 4)$ -factorial experiment, then the number of levels for each of the two factors is a prime power, i.e.  $4 = 2^2$ . The minimal function for  $GF(4)$  is  $p(x) = x^2 + x + 1$  and the elements of the  $GF(4)$  are  $0, 1, x, x+1$ . The total number of treatment combinations is 16 and are given by

A	0	0	0	0	1	1	1	1	x	x	x	x	x+1	x+1	x+1	x+1
B	0	1	x	x+1	0	1	x	x+1	0	1	x	x+1	0	1	x	x+1

Here  $n = 2$  and  $r = 1$ , therefore, the number of blocks per replication is 4 and number of experimental units in each block is also 4. The number of independent interaction

components to be confounded is  $n - r = I$ . Let the experimenter is interested in confounding the interaction component AB. Therefore, the block contents can be obtained from the solution of

$$x_1 + x_2 = 0, 1, x, x + 1 \pmod{2, x^2 + x + 1}$$

The block contents obtained through the solution of the above equations are

Block - I		Block - II		Block - III		Block - IV	
A	B	A	B	A	B	A	B
0	0	0	1	0	x	0	x+1
1	1	1	0	x	0	1	x
x	x	x	x+1	1	x+1	x	1
x+1	x+1	x+1	x	x+1	1	x+1	0

Similarly, we can get the block contents, if the other interaction components are confounded.

The above discussion relates to the methods of construction of symmetrical factorial experiments with confounding. The loss of information on the confounded interaction components depends upon the number of replications in which these are confounded. The designs in which the loss of information is equally distributed over the different components of the interaction of given orders (order of an interaction is one less than the number of factors involved in the interaction) may be desirable. A design with the above characterization is a **balanced confounded design**. This design in case of symmetrical factorials is defined as:

### 6. Balanced Confounded Design

A partially confounded design is said to be balanced if all the interactions of a particular order are confounded in equal number of replications.

#### How to construct a Balanced confounded Factorial Design?

Let us take the example of a  $(2^5, 2^3)$ - factorial experiment. The interest is in constructing a design for a  $(2^5, 2^3)$ - factorial experiment achieving balance over three and four factor interactions. In this case,  $s = 2$ ,  $n = 5$  and  $r = 3$ . Therefore, the total number of treatment combinations is 32, the block size is 8 and the number of blocks per replicate is  $32/8$ . The number of degrees of freedom confounded is  $2^{5-3} - 1 = 3$ . Each interaction component has 1 degree of freedom. Therefore, the number of interaction components to be confounded is 3. The number of independent interactions to be confounded is  $5 - 3 = 2$  and one is the generalized interaction component.

The number of 3 factor interactions =  $({}^5C_3) = 10$  viz. ABC, ABD, ABE, ACD, ACE, ADE, BCD, BCE, BDE, and CDE and number of four factor interactions =  $({}^5C_4) = 5$  viz. ABCD, ABCE, ABDE, ACDE, and BCDE. Therefore, to achieve balance total number of degrees of freedom to be confounded is  $10 + 5 = 15$ . As each interaction component has 1 d.f., therefore, number of degrees of freedom to be confounded are 15. The number of degrees of freedom confounded in one replication is 3. Therefore, the number of replications required is  $15/3 = 5$ . The balance can be achieved by confounding the following interactions in different replications:

- Replication – I: ABD, ACE and BCDE
- Replication – II: ACD, BCE and ABDE
- Replication – III: ADE, BCD and ABCE
- Replication – IV: ABE, CDE and ABCD
- Replication – V: ABC, BDE and ACDE

The block contents may be obtained following the above procedure. The confounding in asymmetrical factorials is somewhat different from symmetrical factorials. When an interaction component is confounded in a replication in these designs, it is not necessary that it is completely confounded with the blocks in the sense that the block contrasts and the interaction contrasts become identical. These two sets of contrasts although not identical, yet are dependent so that the contrasts for obtaining a confounded interaction from the treatment totals are not free from block effects. Therefore, more than one replication is needed in obtaining balanced confounded designs for asymmetrical factorial experiment. A design is said to be Balanced confounded factorial experiment (BFE) if (i) any contrasts of a confounded interaction component is estimable independently of any other contrasts belonging to any other confounded interaction and (ii) the loss of information of each degrees of freedom of any confounded interaction is same. To be more specific, BFE may be defined as:

A factorial experiment will be called a balanced factorial experiment if

- (i) Each treatment is replicated the same number of times.
- (ii) Each of the blocks has the same number of plots.
- (iii) Estimates of the contrasts belonging to different interactions are uncorrelated with each other.
- (iv) Complete balance is achieved over each of the interactions, i.e., all the normalized contrasts belonging to the same interaction are estimated with the same variance.

Several methods of construction of designs for balanced factorial experiments are available in literature based on pseudo factors or pairwise balanced block designs. We shall not be presenting these methods here. The user may refer to standard textbooks for the same. Further, it is known that an *extended group divisible* (EGD) design, if existent, has orthogonal factorial structure with balance. In other words, an EGD design is a balanced confounded factorial experiment. Therefore, the vast literature on the methods of construction of extended group divisible designs may be used for the construction of BFE. The conditions of equal replications and equal block sizes may now be relaxed.

Generation of a design for factorial experiments is easy. But when the number of factors or the number of levels become large it becomes difficult to generate the layout of the design. To circumvent this problem, IASRI has developed a statistical package SPFE (Statistical Package for Factorial Experiments). This package is essentially for symmetrical factorial experiments. There is a provision of generation of designs as well as the randomized layout of the designs including totally and partially confounded designs. The design is generated once the independent interactions to be confounded are listed. One can give different number of independent interactions to be confounded in different replications (The package is also capable of generating the design for factorial experiments by simply giving the number of factors along with the number of levels and the block size. In this case the package will itself determine the number of blocks per replication and the layout by keeping the higher interactions confounded). Provision has also been made in this package for analyzing the data generated from the experiments using these designs. The data generated are analyzed as a general block design and the contrast analysis is carried out to obtain the sum of squares due to main effects and interactions. Separate modules have been developed for generating the probabilities using  $\chi^2$ ,  $F$  and  $t$  distributions for testing the levels of significance.

This package deals with only symmetrical factorial experiments. However, in practice an experimenter encounters situations where one has to use various factors with unequal number of levels. The generation of the design for asymmetrical factorial experiments is, however, a tedious job. We, therefore, give below a catalogue of designs commonly used. In this catalogue A, B, C, etc. denote the factors and a, b, c, etc. denote the blocks within replications.

*Plan 1.* Balanced group of sets for  $3 \times 2^2$  factorial, blocks of 6 units each

**BC, ABC confounded**

Replication I				Replication II				Replication III			
Block-1		Block-2		Block-1		Block-2		Block-1		Block-2	
0	0	1	0	0	0	0	0	0	0	0	0
0	1	0	0	1	1	0	1	0	0	1	0
1	0	0	1	0	1	1	0	0	1	0	1
1	1	1	1	1	1	0	1	1	1	1	0
2	0	0	2	0	1	2	0	1	2	0	0
2	1	1	2	1	0	2	1	0	2	1	1

*Plan 2.* Balanced group of sets for  $3 \times 2^3$  factorial, blocks of 6 units

**BC, BD, CD**

**ABC, ABD, ACD confounded**

Replication I											
Block-1			Block-2			Block-3			Block-4		
0	1	0	0	0	0	0	0	1	0	0	1
0	0	1	1	0	1	1	1	0	0	1	0
1	0	1	0	1	0	0	1	0	1	0	0
1	1	0	1	1	1	1	0	1	1	0	1
2	0	0	1	2	0	1	0	2	1	0	0
2	1	1	0	2	1	0	1	2	0	1	1

Replication II											
Block-1			Block-2			Block-3			Block-4		
0	0	1	0	0	0	0	1	0	0	0	0
0	1	0	1	0	1	1	0	0	1	1	1
1	0	0	1	1	0	1	0	0	1	0	0
1	1	1	0	1	1	0	1	1	0	1	1
2	1	0	0	2	0	0	0	2	0	0	1
2	0	1	1	2	1	1	1	2	1	1	0

Replication III											
Block-1			Block-2			Block-3			Block-4		
0	0	0	1	0	0	1	0	0	0	0	0
0	1	1	0	0	1	0	1	1	0	1	1
1	1	0	0	1	0	0	0	1	1	0	1
1	0	1	1	1	1	1	1	0	1	1	0
2	0	1	0	2	0	0	1	2	0	0	0
2	1	0	1	2	1	1	0	2	1	1	1

*Plan 3.* Balanced group of sets for  $3^2 \times 2$  factorial, blocks of 6 units

**AB, ABC Confounded**

Replication I						Replication II					
Block-1		Block-2		Block-3		Block-1		Block-2		Block-3	
1	0	0	2	0	0	2	0	0	0	0	0
2	1	0	0	1	0	0	1	0	1	1	0
0	2	0	1	2	0	1	2	0	2	2	0
2	0	1	0	0	1	1	0	1	2	0	1
0	1	1	1	1	1	2	1	1	0	1	1
1	2	1	2	2	1	0	2	1	1	2	1

Replication III						Replication IV					
Block-1		Block-2		Block-3		Block-1		Block-2		Block-3	
1	0	0	2	0	0	2	0	0	0	0	0
0	1	0	1	1	0	1	1	0	2	1	0
2	2	0	0	2	0	0	2	0	1	2	0
2	0	1	0	0	1	1	0	1	2	0	1
1	1	1	2	1	1	0	1	1	1	1	1
0	2	1	1	2	1	2	2	1	0	2	1

Plan 4. Balanced group of sets for  $4 \times 2^2$  factorial, blocks of 8 units

**ABC Confounded**

Replication I			Replication II			Replication III		
Block-1		Block-2	Block-1		Block-2	Block-1		Block-2
0	0	0	0	0	0	0	0	0
0	1	1	0	1	1	0	1	1
1	0	0	1	0	1	1	0	1
1	1	1	1	1	0	1	1	0
2	0	1	2	0	1	2	0	1
2	1	0	2	1	0	2	1	1
3	0	1	3	0	0	3	0	1
3	1	0	3	1	1	3	1	0

Plan 5. Balanced group of sets for  $4 \times 3 \times 2$  factorial, blocks of 12 units

**AC, ABC confounded**

Replication I			Replication II			Replication III		
Block-1		Block-2	Block-1		Block-2	Block-1		Block-2
0	0	0	0	0	1	0	0	0
0	1	1	0	1	0	0	1	1
0	2	1	0	2	1	0	2	0
1	0	0	1	0	1	1	0	0
1	1	1	1	1	0	1	1	1
1	2	1	1	2	1	1	2	0
2	0	1	2	0	0	2	0	1
2	1	0	2	1	1	2	1	0
2	2	0	2	2	0	2	2	1
3	0	1	3	0	0	3	0	1
3	1	0	3	1	1	3	1	0
3	2	0	3	2	0	3	2	1

**A<sup>2</sup>C, A<sup>2</sup>BC confounded**

Replication IV			Replication V			Replication VI		
Block-1		Block-2	Block-1		Block-2	Block-1		Block-2
0	0	1	0	0	0	0	0	0
0	1	0	0	1	1	0	1	0
0	2	0	0	2	0	0	2	1

Designs for Factorial Experiments

1	0	0	1	0	1	1	0	1	1	0	0	1	0	0
1	1	1	1	1	0	1	1	0	1	1	1	1	1	0
1	2	1	1	2	0	1	2	1	1	2	0	1	2	1
2	0	0	2	0	1	2	0	1	2	0	0	2	0	0
2	1	1	2	1	0	2	1	0	2	1	1	2	1	0
2	2	1	2	2	0	2	2	1	2	2	0	2	2	1
3	0	1	3	0	0	3	0	0	3	0	1	3	0	1
3	1	0	3	1	1	3	1	1	3	1	0	3	1	1
3	2	0	3	2	1	3	2	0	3	2	1	3	2	0

**A<sup>3</sup>C, A<sup>3</sup>BC confounded**

Replication VII				Replication VIII				Replication IX			
Block-1		Block-2		Block-1		Block-2		Block-1		Block-2	
0	0	0	0	0	0	1	0	0	0	0	0
0	1	1	0	0	1	0	0	1	1	0	0
0	2	1	0	0	2	1	0	2	0	0	1
1	0	1	1	1	0	0	1	0	1	0	1
1	1	0	1	1	1	1	1	1	0	1	1
1	2	0	1	1	2	1	1	2	1	1	0
2	0	0	2	0	1	2	0	0	2	0	0
2	1	1	2	1	0	2	1	1	2	1	0
2	2	1	2	2	1	2	2	0	2	2	1
3	0	1	3	0	0	3	0	1	3	0	1
3	1	0	3	1	1	3	1	0	3	1	1
3	2	0	3	2	1	3	2	1	3	2	0

Plan 6. Balanced group of sets for  $3 \times 2^3$  factorial, blocks of 12 units

**ABC, ABCD confounded**

Replication I								Replication II							
Block-1				Block-1				Block-1				Block-1			
0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
0	0	1	1	0	0	1	0	0	0	1	1	0	0	1	0
0	1	0	1	0	1	0	0	0	1	0	1	0	1	0	0
0	1	1	0	0	1	1	1	0	1	1	0	0	1	1	1
1	0	0	1	1	0	0	0	1	0	0	1	1	0	0	0
1	0	1	0	1	0	1	1	1	0	1	0	1	1	1	1
1	1	0	0	1	1	0	1	1	1	0	0	1	1	0	1
1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0
2	0	0	1	2	0	0	0	2	0	0	0	2	0	0	1
2	0	1	0	2	0	1	1	2	0	1	1	2	0	1	0
2	1	0	0	2	1	0	1	2	1	0	1	2	1	0	0
2	1	1	1	2	1	1	0	2	1	1	0	2	1	1	1

**Replication III**

Block-1				Block-2			
0	0	0	0	0	0	0	1
0	0	1	1	0	0	1	0
0	1	0	1	0	1	0	0
0	1	1	0	0	1	1	1
1	0	0	0	1	0	0	1
1	0	1	1	1	0	1	0

Designs for Factorial Experiments

1	1	0	1	1	1	0	0
1	1	1	0	1	1	1	1
2	0	0	1	2	0	0	0
2	0	1	0	2	0	1	1
2	1	0	0	2	1	0	1
2	1	1	1	2	1	1	0

Example : An experiment was conducted at Crop Research Center, G.B.P.U.A.T., Pantnagar, Uttar Pradesh on bengal gram in rabi season 2003 using a factoria experiment with three factors viz., Farmyard Manure (2 levels:0 and 50 q/ha), Phosphorus (3 levels:0, 20 and 40 kg/ha) and Phosphorus Solublizing Bacteria (2 levels: control and 20 gm/kg of seed as seed inoculation). The main objective of the experiment was to study the effect of Farmyard Manure (FYM), Phosphorus (P) and Phosphate Solublizing Bacteria (PSB) on productivity of bengal gram. The experiment was conducted in a randomized complete block design in 4 replications with 12 plots per replication of net plot size as 5.00×1.80m<sup>2</sup>. The yield (in kg/plot) are given as below.

REP	FYM	P	PSB	Yield
1	1	1	1	0.7
1	1	1	2	1.13
1	1	2	1	1.23
1	1	2	2	1.25
1	1	3	1	1.25
1	1	3	2	1.25
1	2	1	1	0.83
1	2	1	2	1.23
1	2	2	1	1.18
1	2	2	2	0.88
1	2	3	1	1.63
1	2	3	2	1.48
2	1	1	1	0.98
2	1	1	2	1.13
2	1	2	1	1.18
2	1	2	2	1.13
2	1	3	1	1.26
2	1	3	2	1.25
2	2	1	1	0.93
2	2	1	2	0.88
2	2	2	1	1.5
2	2	2	2	1.3
2	2	3	1	1.38
2	2	3	2	1.43
3	1	1	1	0.9
3	1	1	2	1.1
3	1	2	1	1.1
3	1	2	2	0.88
3	1	3	1	1.35
3	1	3	2	1.35



## Designs for Factorial Experiments

3	2	1	1	1.1
3	2	1	2	1.03
3	2	2	1	1.3
3	2	2	2	0.88
3	2	3	1	1.38
3	2	3	2	1.43
4	1	1	1	0.73
4	1	1	2	1.25
4	1	2	1	1.43
4	1	2	2	1.25
4	1	3	1	1.1
4	1	3	2	1.75
4	2	1	1	0.98
4	2	1	2	1.38
4	2	2	1	1.35
4	2	2	2	1.43
4	2	3	1	1.3
4	2	3	2	1.5

R code for the analysis is , where fact is name of dataset

```
install.packages("car")
library(car)
Yield=fact$Yield
REP=as.factor(fact$REP)
P=as.factor(fact$P)
FYM=as.factor(fact$FYM)
PSB=as.factor(fact$PSB)
Yield <- lm(Yield~REP+FYM*P*PSB)
summary(Yield)
par(mfrow=c(1,2))
plot(Yield)
plot(Yield, which=1)
plot(Yield, which=2)
summary.aov(Yield)
```

and the results is

```
> summary.aov(Yield)
              Df    Sum Sq   Mean Sq  F value    Pr(>F)
REP           3    0.1328    0.0443    1.756    0.17473
FYM           1    0.0660    0.0660    2.618    0.11519
P             2    1.0552    0.5276   20.924   1.35e-06 ***
PSB           1    0.0469    0.0469    1.859    0.18197
FYM:P         2    0.0135    0.0067    0.267    0.76747
FYM:PSB       1    0.0481    0.0481    1.909    0.17637
P:PSB         2    0.3380    0.1690    6.702    0.00361 **
FYM:P:PSB     2    0.0014    0.0007    0.028    0.97256
Residuals    33    0.8321    0.0252
```

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

```
> Anova(Yield, type="III")
Anova Table (Type III tests)
```

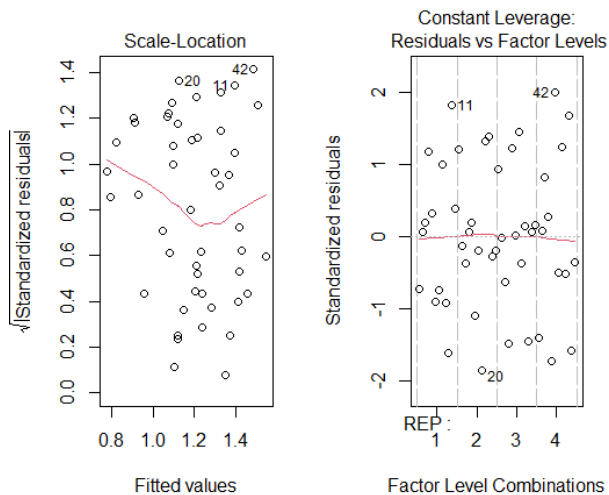
Response: Yield

	Sum Sq	Df	F value	Pr(>F)
(Intercept)	2.03097	1	80.5456	2.26e-10 ***
REP	0.13285	3	1.7562	0.1747268
FYM	0.03511	1	1.3925	0.2464175
P	0.44832	2	8.8898	0.0008158 ***
PSB	0.21125	1	8.3779	0.0066848 **
FYM:P	0.00730	2	0.1448	0.8657813
FYM:PSB	0.02402	1	0.9528	0.3361092
P:PSB	0.19056	2	3.7786	0.0332922 *
FYM:P:PSB	0.00140	2	0.0278	0.9725632
Residuals	0.83210	33		

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

```
plot(Yield)
plot(Yield, which=1)
plot(Yield, which=2)
```



---

---

# MULTIVARIATE DATA ANALYSIS USING R

---

---

**Samarendra Das and A. R. Rao**

*ICAR-Indian Agricultural Statistics Research Institute*

*Library Avenue, New Delhi - 110 012*

[samarendra.das@icar.gov.in](mailto:samarendra.das@icar.gov.in)

---

---

## **Background**

Multivariate statistical techniques simultaneously analyze measurements on multiple variables for each individual under investigation and are widely used in plant breeding programs. The main purpose of multivariate data analysis is to study the relationships among the (multiple) variables and perform several analyses on the collected samples. The multivariate techniques are efficient compared to the univariate counterparts due to their ability to consider inter-variable relationships. In this tutorial, we will focus on important multivariate data analytical techniques including: (1) Principal Component Analysis; (2) Factor Analysis; (3) Cluster Analysis; (4) Discriminant Analysis with real data examples. These four types of multivariate analytical techniques are extensively used in Agricultural experimental data analysis.

### **1. Principal Components Analysis and Factor Analysis**

Principal Components Analysis (PCA) and Factor Analysis (FA) are usually viewed as attempts to approximate the relationships among a set of (*i.e.*, multiple) variables. PCA is concerned with explaining the variance-covariance structure through a few *linear* combinations of the original variables. Whereas FA is concerned with explaining covariance relationships among original variables in terms of a few underlying, but unobservable, random quantities called factors. Factors which are generated are thought to be representative of the underlying processes that have created the correlations among variables. FA is considered as an extension of PCA and its model is also considered to be more elaborative than PCA model. Many a time, PCA and FA together called as Common Factor Analysis (CFA).

The key underlying base to Common Factor Analysis (PCA and FA) is that the chosen variables can be transformed into linear combinations of factors. Factors may either be associated with 2 or more of the original variables (common factors) or associated with an individual variable (unique factors). Loadings relate the specific association between factors and original variables. Therefore, it is necessary to find the loadings, then solve for the factors, which will approximate the relationship between the original variables and underlying factors. The loadings are derived from the magnitude of eigenvalues associated to individual variables. The difference between PCA and FA is that for the purposes of matrix computations PCA assumes that all variance is common, with all unique factors set equal to zero; while FA assumes that there is some unique variance. The level of unique variance is dictated by the FA model which is chosen. Accordingly, PCA is a model of a closed system, while FA is a model of an open system. Rotation in CFA attempts to put the factors in a simpler position with respect to the original variables, which aids in the interpretation of factors. Rotation places the factors into positions that only the variables, which are distinctly related to a factor, will be associated. Varimax, quartimax, and equimax are all orthogonal rotations, while oblique rotations are non-orthogonal. The varimax rotation maximizes the variance of the loadings, and is also the most commonly used rotation method. To analyze data with either PCA or FA, three key decisions must be made. They are (i) the factor extraction method (ii) the number of factors to extract and (iii) the transformation method to be used.

**For example:** Foresters measure data on several characters (*e.g.*, variables) of tree species, such as, growth, volume, yield, forest potential, height, collar diameter, diameter at breast height, crown

## Multivariate Data Analysis Using R

diameter, *etc.* The example data is shown below. Here, the main idea is to illustrate CFA approach on this example data using R. The main purpose of this lecture isto focus more on CFA approach using public statistical R software than dealing with its theory.

### Principal Components Analysis

*Example 1:* The following data pertains to variables, such as Height, Collar diameter, Diameter Breast Height (DBH), and Crown diameter of 36 trees of a particular species. We perform PCA on this data using R and provide hand-on to interpret the obtained results.

Obs.	Height	Collar diameter	DBH	Crown diameter
1	4.00	10.50	6.90	15.13
2	3.80	7.00	4.30	2.63
3	4.90	10.30	7.30	21.71
4	3.00	9.10	5.80	5.24
5	3.80	9.80	6.40	7.57
6	4.00	10.90	6.50	8.67
7	5.30	11.10	6.90	13.09
8	4.50	10.30	6.50	10.55
9	4.40	10.30	6.30	11.53
10	4.70	13.70	9.10	20.66
11	5.20	14.90	10.50	23.19
12	5.30	14.90	9.40	18.59
13	3.60	9.30	6.30	10.21
14	3.30	6.40	3.70	4.67
15	5.00	9.70	6.20	12.34
16	3.70	8.10	5.10	5.89
17	3.80	9.20	5.60	5.36
18	4.00	10.30	7.80	7.54
19	4.90	12.10	8.00	12.93
20	5.50	12.70	8.70	17.79
21	5.30	13.60	9.00	12.76
22	4.80	14.90	10.00	25.62
23	4.30	13.20	9.10	15.57
24	5.10	14.50	10.10	22.56
25	1.50	2.50	1.20	0.17
26	1.90	3.80	2.50	1.56
27	2.40	3.90	1.70	0.58
28	3.80	7.50	5.50	4.47
29	3.60	9.30	6.50	5.94
30	3.30	7.00	3.90	4.63
31	5.60	13.70	9.20	15.26
32	4.50	9.90	4.70	12.11
33	5.20	11.20	7.70	12.57
34	4.60	11.70	9.70	16.21
35	5.00	18.40	10.80	21.15
36	4.60	12.80	8.60	14.44

**R-codes:**

**Step 1:** Create and set the working directory.

```
setwd("../file location")
```

**Step 2:** Save the data in a text file and data reading. For example name the data file as: "data.txt"

```
dat <- read.table(file="data.txt", header=T, row.names = 1, sep="\t")
```

**Step 3:** PCA

```
dat.pca <- prcomp(dat, center = TRUE, scale. = TRUE)
```

**Step 4:** Results summary(dat.pca)

Importance of components:

	PC1	PC2	PC3	PC4
Standard deviation	1.9008	0.45588	0.37934	0.18753
Proportion of Variance	0.9033	0.05196	0.03597	0.00879
Cumulative Proportion	0.9033	0.95523	0.99121	1.00000

Step 5: Representation through a lower ortho-dimensional space.

Rotation (n x k) = (4 x 4):

	PC1	PC2	PC3	PC4
Height	0.4832289	0.8509266	0.1995658	-0.050865416
Collar diam.	0.5127805	-0.1463091	-0.4324909	0.727049786
DBH	0.5104606	-0.2202849	-0.4712747	-0.684693116
Crown diam.	0.4929258	-0.4538637	0.7423108	0.002579604

**Interpretation:**

The proportion of total variation accounted for by the first principal component is 0.903 and the first two components account for a proportion of .9552. Hence, in further analysis, the first or first two principal components PCA1 and PCA2 could replace four variables by sacrificing negligible information about the total variation in the system. The scores of principal components can be obtained by substituting the values of variables in equations of PCAs. For the data under study, the first two principal component scores for first observation i.e. for tree1 can be worked out as

$$PC1 \text{ score} = .483 \times 4.00 + .512 \times 10.50 + .510 \times 6.90 + .49 \times 15.13$$

$$PC2 \text{ score} = .851 \times 4.00 + -.146 \times 10.50 + -.220 \times 6.90 - .4533 \times 15.13$$

Similarly for all other trees the first two principal components scores can be worked out. Thus the whole data with four variables can be converted to a new data set with two principal components.

**Factor Analysis**

We have demonstrated the FA with the following data example in R.

**Example 2:** Consider a hypothetical data on six characters with 15 observations as below:

Obs.	X1	X2	X3	X4	X5	X6
1	609.40	164.99	61.11	15.77	449.89	318.38
2	1960.90	4.30	54.74	33.47	37.14	1.43
3	1846.20	72.92	64.28	36.09	927.87	79.43
4	1002.70	211.76	49.15	42.60	1198.60	280.20
5	2801.10	59.43	82.32	4.40	329.55	108.05
6	1060.00	156.00	69.97	14.07	318.33	229.67
7	512.80	642.81	68.59	8.74	497.83	865.83
8	919.40	18.50	77.13	7.63	403.25	142.80
9	450.40	13.90	54.46	3.48	124.42	25.42

## Multivariate Data Analysis Using R

10	1449.90	129.93	67.04	20.04	530.53	210.44
11	2153.40	96.49	90.92	12.72	881.04	84.66
12	1237.85	147.97	64.38	37.43	643.96	165.81
13	744.90	95.75	77.27	25.03	551.82	154.47
14	1320.90	29.11	68.87	28.54	344.87	63.15
15	1846.20	21.40	63.31	33.87	261.71	12.48

**R code:** Repeat the Steps 1 and 2.

```
dat.fa <- factanal(dat, factors = n) ###choose 'n'
```

**Results:**

Uniquenesses:

```
  x1    x2    x3    x4    x5    x6
0.775 0.030 0.741 0.005 0.738 0.005
```

Loadings:

```
      Factor1 Factor2
x1 -0.474
x2  0.980
x3          -0.509
x4 -0.249  0.966
x5  0.200  0.471
x6  0.997
```

```
      Factor1 Factor2
SS loadings  2.282  1.424
Proportion var 0.380  0.237
Cumulative var 0.380  0.618
```

Null hypothesis: 2 factors are sufficient

Chi square statistic: 6.62 (4 degrees of freedom)

p-value: 0.158

**Interpretation:** Before we interpret the results of the FA, recall the basic idea behind it. FA creates linear combinations of factors to abstract the variable's underlying communality. To the extent that the variables have an underlying communality, fewer factors capture most of the variance in the dataset. This allows us to aggregate a large number of observable variables in a model to represent an underlying concept, making it easier to understand the data. The variability in our data, is given by  $\Sigma$ , and its estimate  $\hat{\Sigma}$  is composed of the variability explained by the factors (linear combination of the factors (communality)) and part of the variability cannot be explained by a linear combination of the factors (uniqueness).

From the above FA it is evident that two factors are sufficient as the test is not significant. Variables X1, X2, and X6 define *factor 1* (high loadings on factor 1, small or negligible loadings on factor 2), variables X3, X4, and X5 define *factor 2* (high loadings on factor 2, small or negligible loadings on factor 1).

### Cluster Analysis and Discriminant Analysis

Rudimentary, exploratory procedures are often quite helpful in understanding the complex nature of multivariate relationship. Searching the data for a structure of "natural" grouping is an important exploratory technique. The most important techniques for data classification are: Cluster analysis and Discriminant analysis.

Although both cluster and discriminant analyses classify objects into different categories, discriminant analysis requires one to know group membership for the cases (*i.e.*, prior class information) used to decide the classification rule whereas in cluster analysis group membership for all cases is unknown. In addition to membership, the number of groups is also generally unknown. In cluster analysis the units within cluster are similar but different between clusters. The grouping is done on the basis of some criterion like similarities measures etc. Thus in the case of cluster analysis the inputs are similarity measures or the data from which these can be computed.

### ***Cluster Analysis***

Cluster analysis is a technique used for combining observations into groups such that:

- (a) Each group is homogeneous or compact with respect to certain characteristics *i.e.*, observations in each group are similar to each other.
- (b) Each group should be different from other groups with respect to the characteristics *i.e.*, observations of one group should be different from the observations of other groups.

There are various mathematical methods which help to sort objects in to a group of similar objects called a Cluster. Cluster analysis is used in diversified research fields. In biology, cluster analysis is used to identify diseases and their stages. For example by examining patients who are diagnosed as depressed, one finds that there are several distinct sub-groups of patients with different types of depression. In marketing cluster analysis is used to identify persons with similar buying habits. By examining their characteristics it becomes possible to plan future strategies more efficiently.

**Example 3:** We will use the data given in Example 2 for cluster analysis using R.

**Steps 1, 2:** Follow the Steps 1 and 2 mentioned in PCA.

**Step 3:** Normalize the data (Sometimes normalization is essential for cluster analysis)

```
means <- apply(dat1,2,mean)
sds <- apply(dat1,2,sd)
nor <- scale(dat1,center=means,scale=sds)
```

**Step 4:** Calculate the distance matrix

```
distance = dist(nor)
```

**Step 5: Select the clustering method**

- a. Default method (Hierarchical agglomerative clustering)

```
mydata.hclust = hclust(distance)
plot(mydata.hclust)
```

Cluster membership:

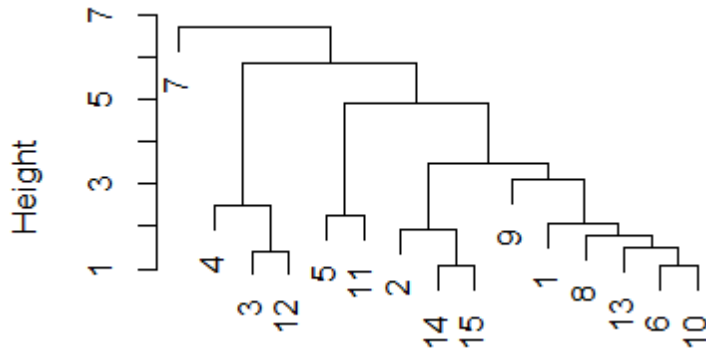
```
member = cutree(mydata.hclust, n) ###Select n (number of clusters, say n=3)
```

```
[1] 1 1 2 2 1 1 3 1 1 1 1 2 1 1 1
```

Table: member

```
1 2 3
11 3 1
```

Dendrogram:



a. **Average linkage method**

```
mydata.hclust = hclust(distance, method = "average")
plot(mydata.hclust)
```

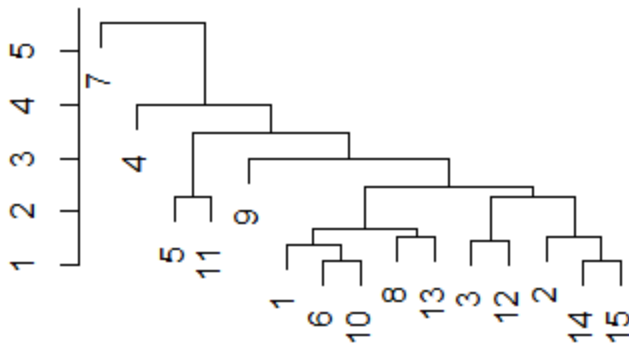
Cluster membership:

```
member = cutree(mydata.hclust, n) ###select n (number of clusters,
say n=3)
[1] 1 1 1 2 1 1 3 1 1 1 1 1 1 1 1
```

Table: member

	1	2	3
11	3	1	

Dendrogram:



b. **K-means clustering**

```
set.seed(123) ##set seed for results reproducibility
kc<-kmeans(nor, 3)
print(kc)
```

K-means clustering with 3 clusters of sizes 5, 3, 7

Cluster members:

```
[1] 3 1 2 2 1 3 3 3 3 3 1 2 3 1 1
```

Cluster means:

	x1	x2	x3	x4	x5	x6
1	1.01492019	-0.5231903	0.40335875	0.07683345	-0.4159255	-0.6089339
2	0.05085677	0.1264368	-0.75013641	1.30454261	1.3631827	-0.0362353
3	-0.74673875	0.3195201	0.03337364	-0.61397073	-0.2871315	0.4504822

Within cluster sum of squares by cluster:

```
[1] 17.946535 4.653991 29.628320 (between_ss / total_ss = 37.8 %)
```



**REFERENCES**

- Chatfield, C. and Collins, A.J. (1990). Introduction to multivariate analysis. *Chapman and Hall publications*.
- Cheng (1997). Applications of GIS and Multivariate Statistical Analysis in Planning Water Conservation Protected Forest -- An Example of the Experimental Forest of National Taiwan University. *Taiwan Journal of Forestry*.
- Johnson, R.A. and Wichern, D.W. (1996). Applied multivariate statistical analysis. *Prentice-Hall of India Private Limited*.
- Salam and Naguchi. (1998). Factors influencing the loss of forest cover in bangladesh: An analysis from socioeconomic and demographic perspectives, *Journal of Forest Research*, **3**, 145-150.
- Sharma, S. (1996). Applied Multivariate Techniques, *John Wiley & Sons*, New York.

---

# MACHINE LEARNING

## (Implementing Support Vector Machine and Random Forest in R)

---

**Prabina Kumar Meher**

*ICAR-Indian Agricultural Statistics Research Institute*

*Library Avenue, New Delhi - 110 012*

[prabina.meher@icar.gov.in](mailto:prabina.meher@icar.gov.in)

---

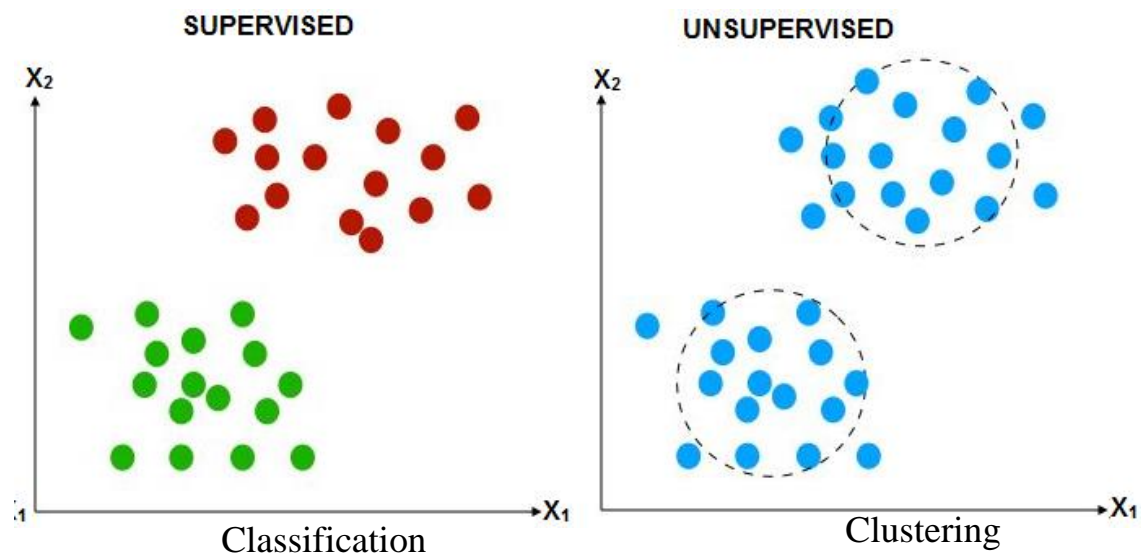
Machine learning (ML) is a branch of Artificial Intelligence. In ML, computer-based algorithms are developed to make the system learn the complex pattern from the data and based on the learned pattern predictions are done for the new individuals. Prediction can be broadly categorised into two classes

1. Predicting the label of the observations (disease, no-disease)
2. Predicting the values, continuous or discrete (yield)

The first type of prediction is classed classification and the second type is known as regression. Further, ML can be broadly categorised into two classes i.e., supervised learning and unsupervised learning. These two learning techniques are mainly differs based on the input-output relation. In other words, in the supervised learning algorithm the labels (output) are attached to each observation that we want to predict (classification or regression), whereas in the case of unsupervised learning no labels are attached to the observations and here our aim is to mainly grouping of the observations.

Target (Y)	Predictors			
	X <sub>1</sub>	X <sub>2</sub>	...	X <sub>p</sub>
y <sub>1</sub>	x <sub>11</sub>	x <sub>12</sub>	...	x <sub>1p</sub>
y <sub>2</sub>	x <sub>21</sub>	x <sub>22</sub>	...	x <sub>2p</sub>
y <sub>3</sub>	x <sub>31</sub>	x <sub>32</sub>	...	x <sub>3p</sub>
...	...	...	...	...
y <sub>n</sub>	x <sub>n1</sub>	x <sub>n2</sub>	x <sub>n3</sub>	x <sub>np</sub>

Predictors			
X <sub>1</sub>	X <sub>2</sub>	...	X <sub>p</sub>
x <sub>11</sub>	x <sub>12</sub>	...	x <sub>1p</sub>
x <sub>21</sub>	x <sub>22</sub>	...	x <sub>2p</sub>
x <sub>31</sub>	x <sub>32</sub>	...	x <sub>3p</sub>
...	...	...	...
x <sub>n1</sub>	x <sub>n2</sub>	x <sub>n3</sub>	x <sub>np</sub>



**Source:** The images have been taken from Google

In this lecture note, our focus is only on the supervised machine learning algorithms. Here, we will discuss two commonly used supervised machine learning algorithms that are support vector machine (SVM) and random forest (RF).

### Measuring accuracy

Different types of performance metrics are utilized for measuring the accuracy in classification and regression problem.

#### *Classification accuracy metrics*

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + TN + FP}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{recall}}{\text{Precision} + \text{Recall}}$$

TP: true positive

TN: true negative

FP: false positive

FN: false negative

#### *Regression accuracy metrics*

$$\text{Mean Absolute Error (MAE)} = \frac{1}{n} \sum_{i=1}^n |y - \hat{y}|$$

$$\text{Mean Square Error (MSE)} = \frac{1}{n} \sum_{i=1}^n (y - \hat{y})^2$$

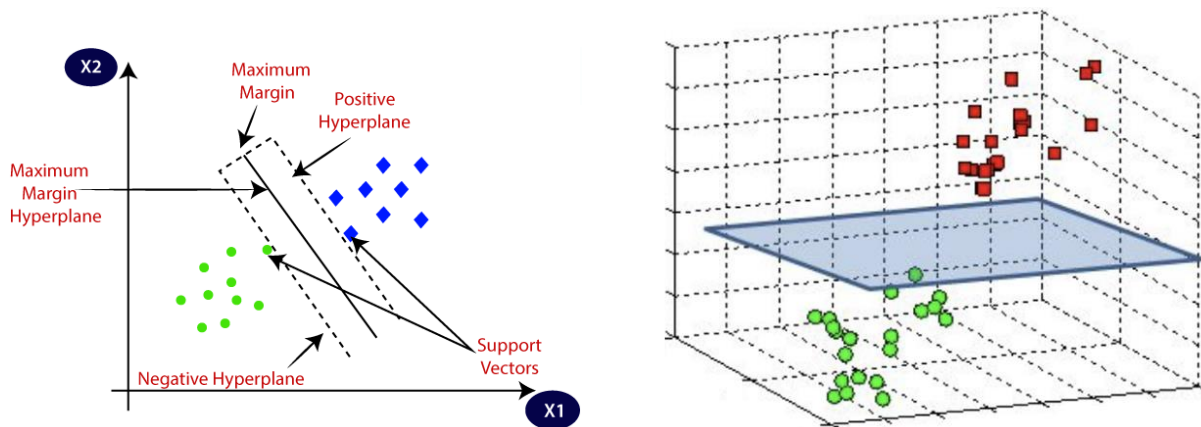
$$\text{Mean Percentage Error (MPE)} = \frac{1}{n} \sum_{i=1}^n \frac{(y - \hat{y})}{y} \times 100$$

$$\text{Mean Absolute Percentage Error (MAPE)} = \frac{1}{n} \sum_{i=1}^n \left| \frac{y - \hat{y}}{y} \right| \times 100$$

In the following subsections, we will discuss the R-code to implement the SVM, RF and ANN. For this, the user needs to install the R from <https://cran.r-project.org/bin/windows/base/>.

## Support Vector Machine

Support vector machine (SVM) is a supervised machine learning algorithm which can be used for both classification and regression. In case of classification, the predictive ability of SVM is largely dependent upon the type of kernel function used that maps the input data set to a high-dimensional feature space, where the observations belong to different classes are linearly separable by the optimal separating hyper-plane. In case of regression, the best line of fit is searched and this is nothing but the hyper plane with maximum number of points.



**Source:** The images have been taken from Google images

## Implementing SVM classification in R

Install the R-package “e1071” to implement the SVM.

```
install.packages(caTools)
install.packages("e1071")
library(e1071)
library(caTools)
```

Read the data from the directory, where the data has been saved. Here, we will use the inbuilt dataset available in the R i.e. *iris* dataset. This is a benchmark dataset which comprises 150 observations of three different flower species i.e. setosa, versicolor and virginica with 50 observations for each type. There are four variables (predictors) such as sepal length, sepal width, petal length and petal width. This data can be loaded in to R console by simply typing `iris`.

```
> iris
   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1           5.1         3.5         1.4         0.2   setosa
2           4.9         3.0         1.4         0.2   setosa
3           4.7         3.2         1.3         0.2   setosa
4           4.6         3.1         1.5         0.2   setosa
5           5.0         3.6         1.4         0.2   setosa
```

Since the response is different species (labels) of flowers, this is a classification problem. For classification or regression, training and test datasets are required. Training set contains the

observations with the respective labels whereas the labels are predicted for the observations of the test set. So, first we need to bifurcate the dataset in to training and test sets. The percentage of dataset to be used for training and testing depends upon the user. However, for better fitting of the model the training dataset should be large enough. The following command can be used for this purpose.

```
library(caTools)
set.seed(123)
dataset <- iris
part <- sample.split(dataset$Species, SplitRatio = 0.70)
train_set <- subset(dataset, part == TRUE)
test_set <- subset(dataset, part == FALSE)

#Check the number of observations of training and test set
table(train_set$Species)
table(test_set$Species)
```

Fitting of the support vector machine classification model using the training dataset

```
# Fitting SVM model to the Training set

library(e1071)
set.seed(123)
svm_class <- svm(formula = Species ~ .,
                 data = train_set,
                 type = 'C-classification',
                 kernel = 'radial')
```

One can see the detail by printing the model.

```
> svm_class

Call:
svm(formula = Species ~ ., data = train_set, type = "C-classification",
    kernel = "radial")

Parameters:
  SVM-Type:  C-classification
  SVM-Kernel: radial
    cost:  1

Number of Support Vectors:  40
```

In the model fitting, the values of the other parameters are kept default. However, the user must tune the respective parameter(s) of the kernel function to maximize the classification accuracy. Here, we have used the radial basis kernel (RBF) kernel function, but there is a choice to select the other kernels as well i.e. polynomial, linear and sigmoid. The classification accuracy varies according to the kernel functions. The support vectors are the observations that lie closely to the hyper-plane as well as influence the position and direction of the hyper-plane. Therefore, the numbers of support vectors are always less than the total numbers of training observations. Maximum number of observations appearing closely to the hyper-plane represents better fitting of the model. After fitting the model, the next part is to

predict the labels of the test set (for classification) by using the trained model. Let us print and check the test set

```
> test_set
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
2             4.9         3.0         1.4         0.2   setosa
4             4.6         3.1         1.5         0.2   setosa
5             5.0         3.6         1.4         0.2   setosa
8             5.0         3.4         1.5         0.2   setosa
11            5.4         3.7         1.5         0.2   setosa
16            5.7         4.4         1.5         0.4   setosa
```

So, one can see that the labels (Species) are there in the test dataset. But, in reality, there wouldn't have labels and we need to predict. Thus, let us remove the labels from the test dataset and predict the labels using the model.

```
test_set1 <- test_set[-5]
```

```
> test_set
  Sepal.Length Sepal.Width Petal.Length Petal.Width
2             4.9         3.0         1.4         0.2
4             4.6         3.1         1.5         0.2
5             5.0         3.6         1.4         0.2
8             5.0         3.4         1.5         0.2
11            5.4         3.7         1.5         0.2
```

```
#Predicting the Test set results
y_pred = predict(svm_class, newdata = test_set1)
y_pred
```

```
> y_pred
  2      4      5      8     11     16
setosa setosa setosa setosa setosa setosa
21     24     26     31     32     34
setosa setosa setosa setosa setosa setosa
50     53     58     59     65     67
setosa versicolor versicolor versicolor versicolor versicolor
```

Here, we have predicted the labels of the test set. One can predict the probability with which these labels are predicted and for this a separate argument need to be passed while training the model as follows.

```
#Training of the model with probability option
svm_class <- svm(formula = Species ~ .,
                 data = train_set,
                 type = 'C-classification',
                 kernel = 'radial', probability=TRUE)

#Prediction fro the test set with probability option
pred_prob <- predict(svm_class, newdata = test_set1, probability=TRUE)
```

```
> pred_prob
      setosa  versicolor  virginica
2  0.95757442 0.027914287 0.014511291
4  0.96313332 0.022506155 0.014360530
5  0.97177768 0.015992024 0.012230293
8  0.96951296 0.017704090 0.012782952
11 0.96913348 0.018803440 0.012063083
16 0.87234295 0.072760478 0.054896568
```

The next step is to compute the confusion matrix that comprises number of correctly classified and mis-classified observations. First of all, we know the labels of the test set that was used for prediction and we called this label as observed labels and the labels obtained through prediction is called the predicted labels.

```
observed <- test_set$Species
predicted <- y_pred
#Creating confusion matrix
#install.packages("caret")
library(caret)
conmat <- confusionMatrix(data=predicted, reference = observed)
```

```
> conmat
Confusion Matrix and Statistics

              Reference
Prediction   setosa versicolor virginica
setosa       15          0          0
versicolor   0          12         1
virginica    0           3         14

Overall Statistics

                Accuracy : 0.9111
                95% CI   : (0.7878, 0.9752)
    No Information Rate : 0.3333
    P-Value [Acc > NIR] : 8.467e-16

                Kappa   : 0.8667

    McNemar's Test P-Value : NA

Statistics by Class:

              Class: setosa Class: versicolor Class: virginica
Sensitivity      1.0000      0.8000      0.9333
Specificity      1.0000      0.9667      0.9000
Pos Pred Value   1.0000      0.9231      0.8235
Neg Pred Value   1.0000      0.9062      0.9643
Prevalence       0.3333      0.3333      0.3333
Detection Rate   0.3333      0.2667      0.3111
Detection Prevalence 0.3333      0.2889      0.3778
Balanced Accuracy 1.0000      0.8833      0.9167
```

Sensitivity is the proportion of correctly predicted positive instances. Specificity is the proportion of correctly predicted negative instances. Positive predicted value is the ratio of the number of correctly predicted positive instances to the total number of predicted

positives, also known as precision. Similarly, negative predicted value is the ratio of the number of correctly predicted negative instances to the total number of predicted negatives. Detection prevalence is defined as the number of predicted positive events (both true positive and false positive) divided by the total number of predictions. The detection rate is the proportion of correctly predicted instances out of total number of instances. The balanced accuracy is the average of the sensitivity and specificity. There are also other performance metrics that can be computed for evaluating the classification performance of machine learning algorithms.

### Implementing SVM regression in R

Based on the similar principle of SVM classification, support vector regression (SVR) is useful for the numerical dependent variable rather than the categorical response. SVR is a non-parametric technique that does not depend upon the underlying distribution of both independent and dependent variables, unlike simple linear regression. It is based on the principle of maximum margin that allows viewing SVR as a convex optimization problem. The penalty parameter (cost) can also be incorporated to avoid over-fitting of the SVR model. Let us discuss how to fit the SVR using R-package “e1071” using a sample dataset. Here, we will use the Los Angeles ozone pollution data, 1976 available in the R-package “mlbench”.

```
library(e1071)
library(mlbench)
library(caret)
library(MLmetrics)
data(Ozone)
```

This dataset contain 366 observations on 13 variables, each observation is one day-basis

V1	Month (1 = January, ..., 12 = December)
V2	Day of month (1, 2, ...,31)
V3	Day of week ( 1 = Monday, ..., 7 = Sunday)
V4 (y)	Daily maximum one-hour-average ozone reading
V5	500 millibar pressure height (m) measured at Vandenberg AFB
V6	Wind speed (mph) at Los Angeles International Airport (LAX)
V7	Humidity (%) at LAX
V8	Temperature (degrees F) measured at Sandburg, CA
V9	Temperature (degrees F) measured at El Monte, CA
V10	Inversion base height (feet) at LAX
V11	Pressure gradient (mm Hg) from LAX to Daggett, CA
V12	Inversion base temperature (degrees F) at LAX
V13	Visibility (miles) measured at LAX

The variables V5, V7, V8, V9, V10, V11 and V12 contain “NA” values and hence the corresponding observations are removed and the resultant dataset comprised of 203 observations on 13 variables.

```
#Removing NA values
dat <- na.omit(Ozone)
rownames(dat)<- as.numeric(1: nrow(dat))
head(dat)
```



```
> head(dat)
  V1 V2 V3 V4   V5 V6 V7 V8   V9  V10 V11  V12 V13
1  1  5  1  5 5760  3 51 54 45.32 1450  25 57.02  60
2  1  6  2  6 5720  4 69 35 49.64 1568  15 53.78  60
3  1  7  3  4 5790  6 19 45 46.40 2631 -33 54.14 100
4  1  8  4  4 5790  3 25 55 52.70  554 -28 64.76 250
5  1  9  5  6 5700  3 73 41 48.02 2083  23 52.52 120
6  1 12  1  6 5720  3 44 51 54.32  111  9 63.14 150
```

So, here our objective is to predict the response i.e., the daily maximum one-hour-average ozone reading (y). First, we will prepare the training and test dataset.

```
set.seed(123)
index <- createDataPartition(dat$V4, p = .7, list = FALSE)
train <- dat[index, ]
test <- dat[-index, ]
```

```
> head(train)
  V1 V2 V3 V4   V5 V6 V7 V8   V9  V10 V11  V12 V13
3  1  7  3  4 5790  6 19 45 46.40 2631 -33 54.14 100
4  1  8  4  4 5790  3 25 55 52.70  554 -28 64.76 250
5  1  9  5  6 5700  3 73 41 48.02 2083  23 52.52 120
6  1 12  1  6 5720  3 44 51 54.32  111  9 63.14 150
8  1 14  3  4 5780  6 19 54 56.12 5000 -44 56.30 200
9  1 15  4  4 5830  3 19 58 62.24 1249 -53 75.74 250
> head(test)
  V1 V2 V3 V4   V5 V6 V7 V8   V9  V10 V11  V12 V13
1  1  5  1  5 5760  3 51 54 45.32 1450  25 57.02  60
2  1  6  2  6 5720  4 69 35 49.64 1568  15 53.78  60
7  1 13  2  5 5760  6 33 51 57.56  492 -44 64.58  40
17 1 30  5 11 5790  3 28 63 57.38  793 -15 65.84 120
20 2  4  3  2 5590  3 76 36 37.40 5000  70 37.94 100
26 2 13  5  6 5700  4 86 55 49.28 2398  21 53.78 200
```

The training and test data sets are ready. Now, we will fit the SVM regression model using the training dataset with default parameter setting. As mentioned earlier, one can choose any one of the kernel function out of four kernels i.e., 'linear', 'polynomial', 'radial basis' and 'sigmoid' for training and predicting. Here, we will use the “radial” kernel function which is the default kernel parameter.

```
#fitting of the SVM regression model
svr_model = svm(train$V4~., data=train)
summary(svr_model)
```

## Machine Learning

```
Call:
svm(formula = train$V4 ~ ., data = train)
```

```
Parameters:
  SVM-Type:  eps-regression
  SVM-Kernel: radial
    cost:    1
   gamma:   0.01754386
  epsilon:  0.1
```

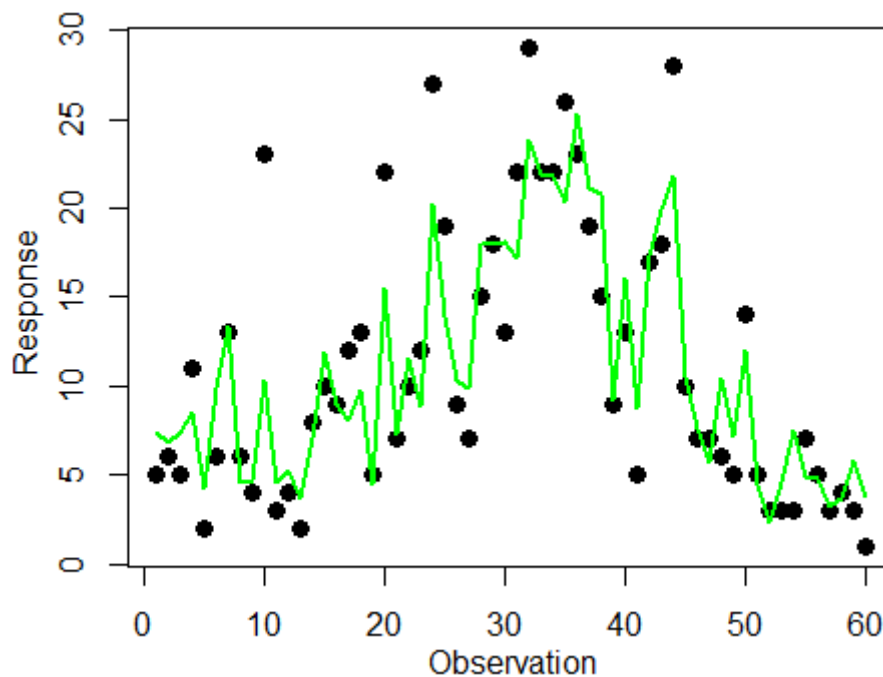
```
Number of Support Vectors: 121
```

Next, we will predict for the test set and plot the predicted observation values with the real values.

```
#Prediction for the test set
pred_svr <- predict(svr_model, test[, -4])
print(pred_svr)
```

```
> print(pred_svr)
      1      2      7     17     20     26
7.401885 6.873475 7.417459 8.526543 4.230745 9.856104
     28     29     30     34     36     38
13.310310 4.513160 4.687134 10.321620 4.556551 5.226023
     42     43     50     53     54     60
3.634022 7.074853 11.854427 8.968529 8.094056 9.696712
```

```
#plotting
x <- 1:length(test$V4)
plot(x, test$V4, pch=16, col="black", cex=1.3, xlab="Observation",
      ylab="Response")
lines(x, pred_svr, lwd="2", col="green")
```



Now, we will evaluate the performance (prediction accuracy) of the model with different metrics such as mean square error (MSE), mean absolute error (MAE), root mean square error (RMSE), R-squared and mean absolute percentage error (MAPE).

```
#Performance metrics
mse <- MSE(y_true=test$V4, y_pred=pred_svr)
mae <- MAE(y_true=test$V4, y_pred=pred_svr)
mape <- MAPE(y_true=test$V4, y_pred=pred_svr)
rmse <- RMSE(y_true=test$V4, y_pred=pred_svr)
Rsqr <- R2_Score(y_true=test$V4, y_pred=pred_svr)
Accuracy <- data.frame(MSE=mse, MAE=mae, MAPE=mape, RMSE=rmse, R2=Rsqr)
Accuracy
```

```
Accuracy
      MSE      MAE      MAPE      RMSE      R2
11.21423 2.471234 0.332201 3.348766 0.8041752
```

In SVR, the prediction accuracy can be improved by optimising the hyper parameter of the kernel functions. It is also advised that user should compare the accuracy of different kernel function to have a better idea about the comparative accuracy.

### Random Forest

Classification and regression trees (CART) work on the principle of information gain at each node of the tree. In other words, splitting at that node happens where information gain is maximum. This process is repeated until all the nodes are exhausted or there is no further information gain. The CARTs have very low predictive power and often referred as weak learners. RF algorithm is based on the ensemble concept of such weak learners.

Random Forest (RF) is a supervised machine learning algorithm that can be used for both classification and regression problems. RF is an ensemble learning method comprises several tree-based classifier, where each classifier (classification tree) is constructed on a bootstrap resample of the training dataset. This method is robust to noise, problem of overfitting and can handle large dataset. In each bootstrap sample, a classification tree is constructed and the observations abstain from taking part in the classifier construction are used as test set for that tree classifier. On-an-average, each classifier in RF is built on  $2/3^{\text{rd}}$  of the training data and tested on the  $1/3^{\text{rd}}$  Out-of-Bag sample. These OOB samples are the source of data for measuring the prediction error of RF. More clearly, the error for each classifier in RF is measured based on its OOB samples (called as OOB error) and these OOB errors are averaged over all the decision trees to compute the OOB error of the forest. As far as classification of test instance is concerned, each classifier of RF votes each test instances to one of the pre-defined classes and the test instance is predicted by the label of winning class. With regard to the RF regression, the final prediction is made by taking the average of

the predictions made by each individual decision tree in the forest. RF is also non-parametric in nature which does not depend upon the underlying assumption about the statistical distribution of the dependent and independent variables.

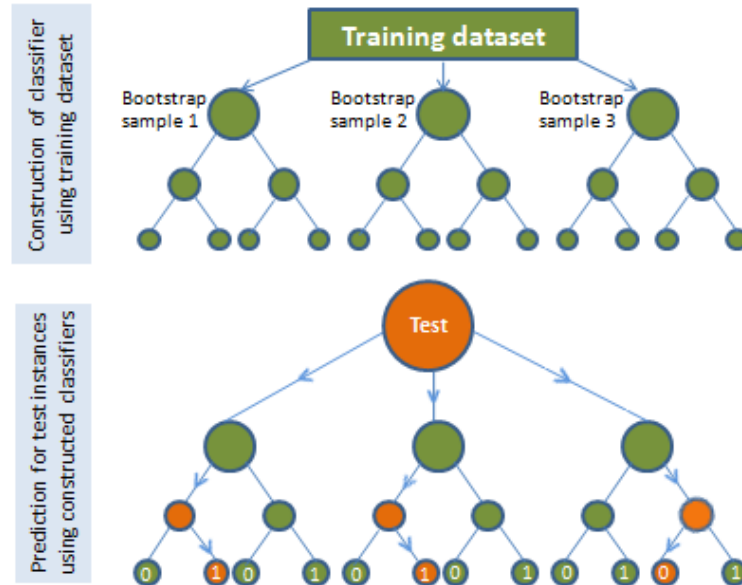


Image Source: Meher et al. (2019) *BMC Genetics* 20 (1), 1-13

## Implementing RF classification in R

To implement the RF classification, first we need to install the *randomForest* R-package.

```
install.packages(randomForest)
library(randomForest)
```

Here, we will use the inbuilt dataset available in the R i.e. iris dataset. This is a benchmark dataset which comprises 150 observations of three different plant species i.e. setosa, versicolor and virginica with 50 observations for each type. There are four variables (predictors) such as sepal length, sepal width, petal length and petal width. This data can be loaded into the R console by simply typing `iris`.

```
#Load the dataset
dat <- iris
```

```
> iris
   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1           5.1           3.5           1.4           0.2  setosa
2           4.9           3.0           1.4           0.2  setosa
3           4.7           3.2           1.3           0.2  setosa
4           4.6           3.1           1.5           0.2  setosa
5           5.0           3.6           1.4           0.2  setosa
```

For classification or regression, training and test datasets are required. Training set contains the observations with the respective labels whereas the labels are predicted for the observations of the test set. We will split the dataset into train and validation set in the ratio 70:30. The percentage of dataset to be used for training and validation depends upon the user.

```
# Split into Train and Validation sets
# Training Set : Validation Set = 70 : 30 (random)
set.seed(100)
train <- sample(nrow(dat), 0.7*nrow(dat), replace = FALSE)
Train_Set <- dat[train,]
Valid_Set <- dat[-train,]
summary(Train_Set)
summary(Valid_Set)

> summary(Train_Set)
  Sepal.Length   Sepal.Width   Petal.Length   Petal.Width
Min.   :4.300   Min.   :2.200   Min.   :1.000   Min.   :0.100
1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
Median :5.800   Median :3.000   Median :4.300   Median :1.300
Mean   :5.811   Mean   :3.054   Mean   :3.702   Mean   :1.197
3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
  Species
setosa   :36
versicolor:34
virginica :35

> summary(Valid_Set)
  Sepal.Length   Sepal.Width   Petal.Length   Petal.Width
Min.   :4.400   Min.   :2.000   Min.   :1.300   Min.   :0.100
1st Qu.:5.300   1st Qu.:2.800   1st Qu.:1.500   1st Qu.:0.200
Median :5.900   Median :3.000   Median :4.400   Median :1.400
Mean   :5.918   Mean   :3.064   Mean   :3.889   Mean   :1.204
3rd Qu.:6.500   3rd Qu.:3.300   3rd Qu.:5.500   3rd Qu.:1.800
Max.   :7.700   Max.   :4.200   Max.   :6.700   Max.   :2.500
  Species
setosa   :14
versicolor:16
virginica :15
```

Now, we will train the RF model with default parameter setting. There are mainly two parameters to be tuned in RF model i.e., the number of trees to grow (*ntree*) and the number of variables randomly sampled at each split (*mtry*). The *ntree* value should not be set too small. It should be ensured that every observation gets predicted at least few times. The default *mtry* value for is square root of the number of variable and one-third of the number of predictors for the regression problem. The default value of *ntree* is 500.

```
# Create a Random Forest model with default parameters
model_RF <- randomForest(Species ~ ., data = Train_Set)
model_RF
```

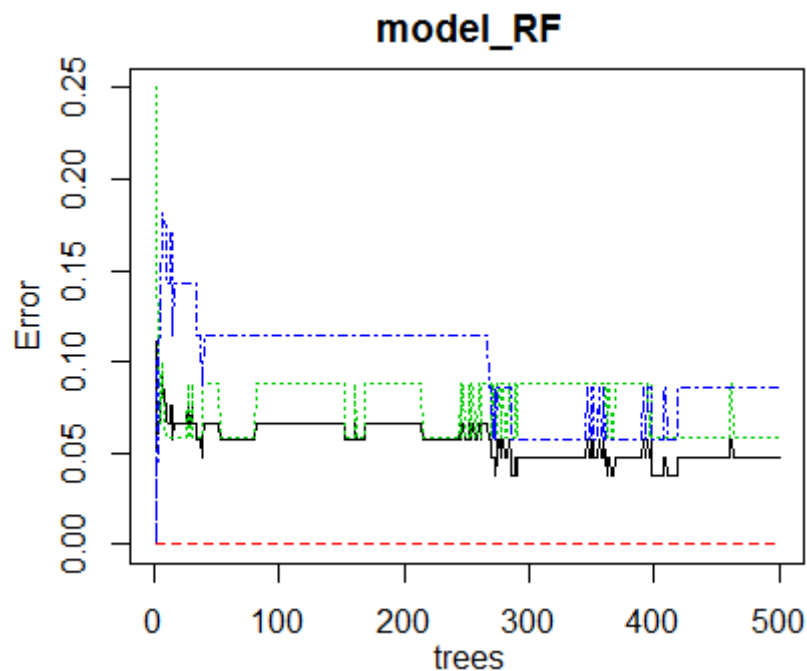
## Machine Learning

```
> model_RF

Call:
  randomForest(formula = Species ~ ., data = Train_Set)
  Type of random forest: classification
  Number of trees: 500
No. of variables tried at each split: 2

  OOB estimate of error rate: 4.76%
Confusion matrix:
      setosa versicolor virginica class.error
setosa      36         0         0 0.00000000
versicolor  0         32         2 0.05882353
virginica   0          3        32 0.08571429
```

Here, we can see that the numbers of classification trees are 500 and the number of variables tried at each split is 2 which is square root of 4. The overall OOB error is 4.76%. The misclassification errors for setose, versicolor and virginica are 0, 0.05 and 0.08 respectively. One can also plot the RF model and visualize the error rate with respect to the number of trees.



Now we will predict for the training set as well as for the validation set.

```
# Predicting on train set
pred_Train <- predict(model_RF, Train_Set[,-5], type = "class")
# Checking classification accuracy
table(pred_Train, Train_Set$Species)
```

```
> table(pred_Train, Train_Set$Species)
```

```
pred_Train  setosa versicolor virginica
setosa      36         0          0
versicolor  0         34          0
virginica   0         0          35
```

```
# Predicting on Validation set
pred_Valid <- predict(model_RF, Valid_Set[,-5], type = "class")
# Checking classification accuracy
table(pred_Valid, Valid_Set$Species)
```

```
> table(pred_Valid, Valid_Set$Species)
```

```
pred_Valid  setosa versicolor virginica
setosa      14         0          0
versicolor  0         14          0
virginica   0         2          15
```

While predicting the training set, we can see that all the instances of the three species are correctly classified. This may be due to the fact that in this case both training and test sets are same. On the other hand, all the test instances of the setosa and versicolor are correctly predicted, whereas 2 observations of the virginica are mis-classified into the versicolor. So, the accuracy of the test will always be either equal or less than that of the training set. The next step is to compute the confusion matrix and the performance metric thereafter. First of all, we know the labels of the test set that was used for prediction and we called this label as observed labels and the labels obtained through prediction is called the predicted labels.

```
observed <- Valid_Set$Species
predicted <- pred_Valid
#Creating confusion matrix
#install.packages("caret")
library(caret)
conmat <- confusionMatrix(data=predicted, reference = observed)
conmat
```

```
> conmat
Confusion Matrix and Statistics

          Reference
Prediction setosa versicolor virginica
setosa      14          0          0
versicolor  0          14          0
virginica   0           2         15

Overall Statistics

          Accuracy : 0.9556
          95% CI   : (0.8485, 0.9946)
No Information Rate : 0.3556
P-Value [Acc > NIR] : < 2.2e-16

          Kappa   : 0.9333

McNemar's Test P-Value : NA

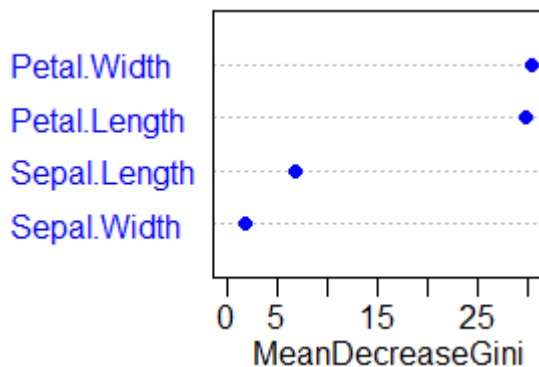
Statistics by Class:
```

```
          Class: setosa Class: versicolor Class: virginica
Sensitivity          1.0000          0.8750          1.0000
Specificity          1.0000          1.0000          0.9333
Pos Pred Value       1.0000          1.0000          0.8824
Neg Pred Value       1.0000          0.9355          1.0000
Prevalence           0.3111          0.3556          0.3333
Detection Rate       0.3111          0.3111          0.3333
Detection Prevalence 0.3111          0.3111          0.3778
Balanced Accuracy    1.0000          0.9375          0.9667
```

We can also use check importance of each variables. The mean decrease in accuracy for each of the variables can be computed and plotted by using the following function.

```
importance(model_RF) #computation
varImpPlot(model_RF) #plotting
```

```
> importance(model_RF)
          MeanDecreaseGini
Sepal.Length      6.923897
Sepal.Width       1.894913
Petal.Length     29.887463
Petal.Width      30.505017
```





## Implementing Random Forest regression in R

Like simple linear regression, RF regression is based on the concept of dependent and independent variables. In RF regression, the ensemble learning technique is employed which combines the results from several tree-based learners. RF regression is more accurate and powerful as compared to several other regression methods. It also performs well on the dataset that have features with non-linear relationship. However, to avoid the over-fitting, sufficient number of trees should be constructed.

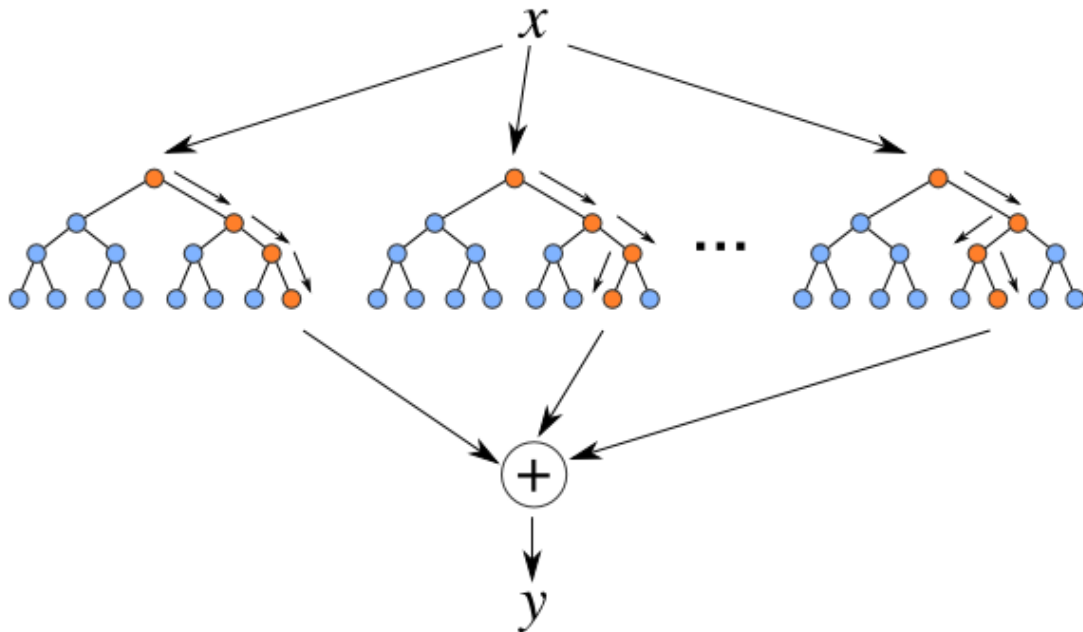


Image source: <https://levelup.gitconnected.com/random-forest-regression-209c0f354c84>

Now, we will discuss different steps for performing the RF regression using the R-software. So, first we need to install the *randomForest* R-package in R-console.

```
#package installation
install.packages("randomForest")
library(randomForest)
```

Here, we will use the same dataset that has been used for implementing support vector regression in R. In other words, we will use the Los Angeles ozone pollution data, 1976 available in the R-package “mlbench”. More details about the dataset can be found in the subsection “Implementing support vector regression in R”. Here also, we will use the dataset after removing NA values.

## Machine Learning

```
#Loading the dataset
library(mlbench)
data(Ozone)

#Removing NA values
dat <- na.omit(Ozone)
rownames(dat) <- as.numeric(1:nrow(dat))
head(dat)
```

```
> head(dat)
  V1 V2 V3 V4   V5 V6 V7 V8   V9  V10 V11  V12 V13
1  1  5  1  5 5760  3 51 54 45.32 1450  25 57.02  60
2  1  6  2  6 5720  4 69 35 49.64 1568  15 53.78  60
3  1  7  3  4 5790  6 19 45 46.40 2631 -33 54.14 100
4  1  8  4  4 5790  3 25 55 52.70  554 -28 64.76 250
5  1  9  5  6 5700  3 73 41 48.02 2083  23 52.52 120
6  1 12  1  6 5720  3 44 51 54.32  111  9 63.14 150
```

So, here our objective is to predict the response i.e., the daily maximum one-hour-average ozone reading ( $y$ ). First, we will first prepare the training and test dataset. The splitting of dataset into training and test set is important in the sense that the training set contains both the response and predictors from which the model learns off. The test set then tests the model's predictions based on the learned model from the training set.

```
set.seed(123)
index <- createDataPartition(dat$V4, p = .7, list = FALSE)
train <- dat[index, ]
test <- dat[-index, ]
```

```
> head(train)
  V1 V2 V3 V4   V5 V6 V7 V8   V9  V10 V11  V12 V13
3  1  7  3  4 5790  6 19 45 46.40 2631 -33 54.14 100
4  1  8  4  4 5790  3 25 55 52.70  554 -28 64.76 250
5  1  9  5  6 5700  3 73 41 48.02 2083  23 52.52 120
6  1 12  1  6 5720  3 44 51 54.32  111  9 63.14 150
8  1 14  3  4 5780  6 19 54 56.12 5000 -44 56.30 200
9  1 15  4  4 5830  3 19 58 62.24 1249 -53 75.74 250

> head(test)
  V1 V2 V3 V4   V5 V6 V7 V8   V9  V10 V11  V12 V13
1  1  5  1  5 5760  3 51 54 45.32 1450  25 57.02  60
2  1  6  2  6 5720  4 69 35 49.64 1568  15 53.78  60
7  1 13  2  5 5760  6 33 51 57.56  492 -44 64.58  40
17 1 30  5 11 5790  3 28 63 57.38  793 -15 65.84 120
20 2  4  3  2 5590  3 76 36 37.40 5000  70 37.94 100
26 2 13  5  6 5700  4 86 55 49.28 2398  21 53.78 200
```

The training and test data sets are ready. Now, we will fit the RF regression model using the training dataset with default values of  $mtry$  and  $ntree$  parameters. As mentioned earlier, one can optimize the  $mtry$  and  $ntree$  for getting maximum accuracy. Here, we used  $mtry=4$  (one-third of the number of predictors) and  $ntree=500$  (default values).

```
#fitting of the RF regression model
RF_model = randomForest(train$V4~., data=train)
summary(RF_model)
```

## Machine Learning

```
> RF_model
```

```
Call:
```

```
randomForest(formula = train$V4 ~ ., data = train)
  Type of random forest: regression
  Number of trees: 500
```

```
No. of variables tried at each split: 4
```

```
Mean of squared residuals: 24.82561
  % Var explained: 64.86
```

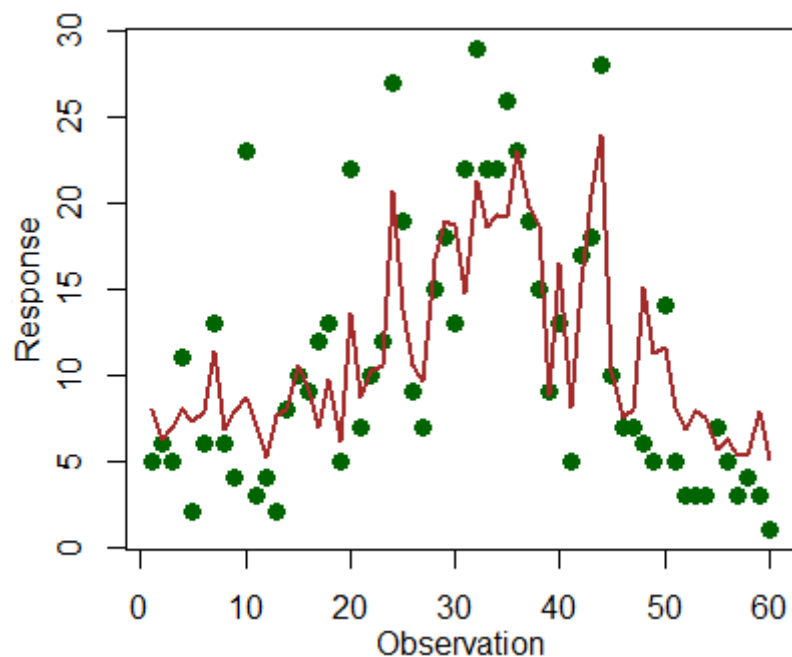
Next, we will predict for the test set and plot the predicted observation values with the real values.

```
#Prediction for the test set
pred_RF <- predict(RF_model, test[, -4])
print(pred_RF)
```

```
> print(pred_RF)
```

1	2	7	17	20	26
7.960000	6.265333	6.999900	7.999733	7.343100	7.855833
30	34	36	38	42	43
7.900767	8.737433	7.077767	5.209833	7.661100	7.970033
54	60	61	68	70	71
6.946400	9.707567	6.120767	13.533100	8.640700	10.195667
88	90	91	100	103	107
13.788700	10.557533	9.675800	16.693967	18.934533	18.772033

```
#plotting
x <- 1:length(test$V4)
plot(x, test$V4, pch=16, col="darkgreen", cex=1.3, xlab="Observation",
      ylab="Response")
lines(x, pred_RF, lwd="2", col="brown")
```



## Machine Learning

Now, we will evaluate the performance (prediction accuracy) of the model with different metrics such as mean square error (MSE), mean absolute error (MAE), root mean square error (RMSE), R-squared and mean absolute percentage error (MAPE).

```
#Performance metrics
library(MLmetrics)
mse <- MSE(y_true=test$V4, y_pred=pred_svr)
mae <- MAE(y_true=test$V4, y_pred=pred_svr)
mape <- MAPE(y_true=test$V4, y_pred=pred_svr)
rmse <- RMSE(y_true=test$V4, y_pred=pred_svr)
Rsqr <- R2_Score(y_true=test$V4, y_pred=pred_svr)
Accuracy <- data.frame(MSE=mse, MAE=mae, MAPE=mape, RMSE=rmse, R2=Rsqr)
Accuracy
```

```
Accuracy
      MSE      MAE      MAPE      RMSE      R2
17.35143 3.188692 0.5369437 4.165505 0.6970064
```

In RF regression, the prediction accuracy depends upon the type dataset used. Also, the accuracy can be improved by optimizing the parameters. It is also advised that user should compare the accuracy of different regression methods to have a better idea about the comparative accuracy.

### References

- Breiman, L., 2001. Random forests. *Machine learning*, 45(1), 5-32.
- Liaw, A. and Wiener, M., 2002. Classification and regression by randomForest. *R news*, 2(3),18-22.
- Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.C., Lin, C.C. and Meyer, M.D., 2019. Package ‘e1071’. *The R Journal*.
- Vapnik, V., Guyon, I. and Hastie, T., 1995. Support vector machines. *Machine Learning*, 20(3), 273-297.

# Faculty Members

Name of Faculty	E-mail ID	Contact number
<b>Dr. Rajender Parsad</b> Director, ICAR-IASRI	director.iasri@icar.gov.in	011-25841479
<b>Dr. Ajit</b> Course Coordinator	ajit@icar.gov.in	9415092880
<b>Dr. Ranjit Kumar Paul</b> Course Coordinator	ranjit.paul@icar.gov.in	8287778896
<b>Dr. Soumen Pal</b> Course Coordinator	soumen.pal@icar.gov.in	9654670940
<b>Mr. Upendra Kumar Pradhan</b>	upendra.pradhan@icar.gov.in	7807176593
<b>Dr. Samarendra Das</b>	samarendra.das@icar.gov.in	9861345735
<b>Dr. Prabina Kumar Meher</b>	prabina.meher@icar.gov.in	9310714631
<b>Dr. Himadri Shekhar Roy</b>	himadri.roy@icar.gov.in	9013846158
<b>Mr. Prakash Kumar</b>	prakash.kumar@icar.gov.in	8800877135
<b>Dr. Md. Yeasin</b>	md.yeasin@icar.gov.in	8926261427
<b>Dr. Ankur Biswas</b>	ankur.biswas@icar.gov.in	9968000281
<b>Dr. Susheel Kumar Sarkar</b>	susheel.sarkar@icar.gov.in	8368096196

## ICAR-Indian Agricultural Statistics Research Institute (IASRI)



Library Avenue, Pusa  
New Delhi-110012

Visit us at : <https://iasri.icar.gov.in/>



<https://iasri.icar.gov.in>

भा.कृ.अ.प.—भारतीय कृषि सांख्यिकी अनुसंधान संस्थान  
ICAR- Indian Agricultural Statistics Research Institute  
An ISO 9001:2008 Certified Institute

Home About Us Divisions Academics Publications Services Notifications Downloads Training Events Gallery Contact हिन्दी

**INFORMATION HUB**

- KVK Portal
- Krishi Portal
- NABG
- NAHEP Component-2
- PG School, IARI-MS
- Agridaksh
- Mushroom Agridaksh
- Expert System (Wheat)
- Expert System (Seed Spices)
- ePlatform (Seed Spices)
- Agricultural Research Data Book

**STATISTICAL HUB**

**LATEST EVENTS**

- "Data Analytics as Career Option" on 12th August 2021 from 6.00 PM onwards on occasion of Azadi Ka Amrit Mahostav.
- Virtual Classroom and Agri-Diksha (Web Education Channel) Inaugurated

**Machine Learning for Climate Change Analysis**  
by **Dr. Pabitra Mitra**  
Professor (Computer Science and Engineering) & Head, Centre for Campus Indian Institute of Technology, Kharagpur  
on **18th September 2021 at 12:00 PM**  
Join us at : [shorturl.at/osB58](https://shorturl.at/osB58)  
Meeting ID : 979 7899 8991 | Passcode : 469993

In Commemoration Of India's 75 Years Of Independence, ICAR-IASRI Is Organizing The Webinar On "Machine Learning For Climate Change Analysis" By Dr Pabitra Mitra On 18th September 2021 At 12.00 PM Onwards

About Us

**Thank You!**

**Dear participants**

Hope this manual lecture notes may be of use in your future research and teaching endeavour

**Team-IASRI**