NAHEP

Institutional Development Plan (IDP), SKUAST Jammu

Strengthening Institutional Capacities for Delivering Competent Skilled Professionals

Azadi Ka Amrit Mahotsav

NSD-2022

COMPENDIUM ON

BIG DATA ANALYSIS AND RESEARCH METHODS USING STATISTICAL SOFTWARES

Published by

IDP, SKUAST Jammu

Division of Statistics & Computer Science

Compiled & Edited by

Manish Kr. Sharma

M. Iqbal Jeelani Bhat

Sunali Mahajan

Technical Support

Sanveet Kour, Nishant Jasrotia

Sher-e-Kashmir University of Agricultural Sciences & Technology of Jammu

CHAPTER 1

SPECIAL LECTURE: ABOUT THE NATIONAL STATISTICS DAY

Vinod Kumar Gupta
Former National Professor (Statistics), ICAR
Email: vkgupta_1751@yahoo.co.in

In 2006, with an act of Parliament it was decided to observe 29 June as the National Statistics Day. Since then it is being celebrated every year. On National Statistics Day, India commemorates renowned statistician P.C. Mahalanobis and also aims at creating public awareness about the significance of statistics in daily life and in the process of planning and development. It also marks the birth anniversary of P.C. Mahalanobis.

Prasanta Chandra Mahalanobis

Born June 29, 1893, Calcutta [now Kolkata], India—died June 28, 1972, Calcutta), Indian statistician who devised the Mahalanobis distance and was instrumental in formulating India's strategy for industrialization in the Second Five-Year Plan (1956–61).

Born to an academically oriented family, Mahalanobis pursued his early education in Calcutta (now Kolkata). After graduating with honours in Physics from Presidency College, Calcutta, in 1912, he moved to England to study physics and mathematics at the University of Cambridge. Just before Mahalanobis left the university in 1915, he was introduced to statistics by one of his tutors. In England, Mahalanobis was introduced to the journal Biometrika. This interested him so much that he bought a complete set and took them to India. He discovered the utility of statistics to problems in meteorology and anthropology, beginning to work on problems on his journey back to India. When he returned to India, he accepted a temporary position teaching physics at Presidency College, and he became a professor of physics there in 1922. However, his interest in statistics had evolved into a serious academic pursuit, and he applied statistical methods to problems in anthropology, meteorology, and biology. On December 17, 1931, he established the Indian Statistical Institute in Calcutta.

Mahalanobis devised a measure of comparison between two data sets that is now known as the **Mahalanobis distance**. He was the founder of Indian Statistical Institute, Kolkata (1932) and introduced innovative techniques for conducting large-scale sample surveys and calculated acreages and crop yields by using the method of random sampling. He devised a statistical method called **fractile graphical analysis**, which could be used to compare the socioeconomic conditions of different groups of people. He also applied statistics to

economic planning for flood control. He also founded the first Indian Journal on Statistics - Sankhya in 1933.

With the objective of providing comprehensive socioeconomic statistics, Mahalanobis established the National Sample Survey in 1950 and also set up the Central Statistical Organization to coordinate statistical activities in India. He was also a member of the Planning Commission of India from 1955 to 1967. The Planning Commission's Second Five-Year Plan encouraged the development of heavy industry in India and relied on Mahalanobis's mathematical description of the Indian economy, which later became known as the Mahalanobis model.

Mahalanobis held several national and international portfolios. He served as the chairman of the United Nations Sub-Commission on Sampling from 1947 to 1951 and was appointed the honorary statistical adviser to the government of India in 1949. For his pioneering work, he was awarded the Padma Vibhushan, one of India's highest honours, by the Indian government in 1968. He was the Fellow of the Royal Statistical Society.

His famous books are (a) Experiments in Sampling in the Indian Statistical Institute; (b) Rabindranath Tagore's visit to Canada.

Theme of NSD

"Data for Sustainable Development"

The 17 UN Sustainable Development Goals give us a global plan for a sustainable future, both economically, environmentally and not least socially.

1. No Poverty
2. Zero Hunger
3. Good Health and Well-Being
4. Quality Education
5. Gender Equality
6. Clean Water and Sanitation
7. Affordable and Clean Energy
8. Decent Work and Economic Growth
9. Industry, Innovation and Infrastructure
10. Reduced Inequalities
11. Sustainable Cities and Communities
12. Responsible Consumption and Production

13. Climate Action
14. Life Below Water
15. Life on Land
16. Peace, Justice and Strong Institutions
17. Partnerships for the Goals

Cross-sectoral partnerships that recognise the **crucial links** between social and environmental issues are key to a better future. COVID-19 has presented unprecedented challenges, reversing decades of development and causing a deep global recession. Never has there been a more critical time for strengthening partnerships and securing the next ten years of collaboration for sustainable development. The international community must foster recognition of the urgent need to end human population growth as soon as is ethically possible, and promote greater investment in empowering solutions

Traces of Statistics

Mahabharata; Vana Prarva; Nala – Damyanti Akhyan

Nala and king Bhangasuri were moving in a chariot through a forest. Bhangasuri told Nala that if he can count how many fallen leaves and fruits are there, he (Bhangasuri) can tell the number of fruits and leaves on two strongest branches of Vibhitak tree. One above one hundred are the number of leaves and one fruit informed Nala after counting the fallen leaves and fruit. Bhangasuri avers 2095 fruits and five ten million leaves on the two strongest branches of the tree (actually it is 5 koti leaves and 1 koti is 10 million). Nala counts all night and is duly amazed by morning. Bhangasuri accepts his due "I of dice possess the science—and in numbers thus am skilled." said Bhangasuri. Vahuca replied; "That science—if to me thou wilt impart, In return, O king, receive thou—my surpassing skill in steeds." This indeed is a strong application of survey sampling.

SOME IMPORTANT JOURNALS

JRSS -

Around 1645, a group of scientists started regular meetings to establish a society for statisticians. William Petty was one of members in the first group of scientists. On 28 November 1660, twelve men held first meeting. The Royal Statistical Society was founded as the Statistical Society of London in 1834. Florence Nightingale was its first female member. The first part of the Journal of the Statistical Society of London was published in May 1838. On 31 January 1887, the Society was incorporated by Royal Charter and became the Royal

Statistical Society. In 1887 the name of the Journal of the Statistical Society of London was changed to the Journal of the Royal Statistical Society.

JASA -

A team of five people organized a Statistical Society called American Statistical Society. On February 5, 1840, it was renamed as the American Statistical Association (ASA) at its first annual meeting held in Boston. Richard Fletcher was elected as the first president (1839-1845). Lemuel Shattuck, the first secretary, was the true pioneer in founding the association. In 1888, the association started publishing Publications of the American Statistical Association, which resulted in increased national interest in the association and as a result, number of members increased to more than 500 in 1898 from 160 in 1889. The publication series introduced by Walker is now known as Journal of the American Statistical Association (JASA). JASA today is one of the most important journals in the field of statistical sciences.

International Statistical Institute (ISI), Netherlands

The International Statistical Congress was founded in 1853. In1885, it was renamed as The International Statistical Institute (ISI). It is a professional association of statisticians. The Institute's activities include publication of a variety of books and journals and holding an international conference every two years. The biennial convention of the ISI was commonly known as the ISI Session. Since 2011 it is known as the ISI World Statistics Congress. The permanent office of ISI is located in The Hague, in The Netherlands. ISI publishes the well renowned journal The International Statistical Review.

Biometrika –

The Journal Biometrika was established in1901. Karl Pearson was the editor for the first 35 years of its existence. Francis Galton and Walter Weldon were the other two who helped in establishing the Journal.

Annals of Mathematical Statistics -

Annals of Mathematical Statistics were a statistics journal published by the Institute of Mathematical Statistics from 1930 to 1972. In 1938, Samuel Wilks became editor-in- chief of the Annals and recruited a remarkable editorial board comprising of Sir R A Fisher, Jerzy Neyman, Harald Cramér, Harold Hotelling, Egon Sharpe Pearson, Georges Darmois, Allen T. Craig, William Edward Deming, Richard Edler von Mises, Henry Louis Rietz, and Walter A. Shewhart. It was superseded by the Annals of Statistics and the Annals of Probability in 1973.

Sankhya –

Sankhya an Indian Journal of Statistics was founded in 1933 by Late Professor Prasanta Chandra Mahalanobis and is being published since then by the Indian Statistical Institute. Prasanta Chandra Mahalanobis was the founder Editor-in-Chief of the Journal and remained so till his death.

JISAS –

Sir C.V. Raman, the then President of the Indian Science Academy, during 1945-46 felt that agricultural scientists and agricultural statisticians should knit themselves together for exchange of views and pooling up of experiences as Statistics had made the largest contributions in agriculture. So he advised to form a scientific society for the promotion of study and research in agricultural statistics which would provide an opportunity for publishing at one place their research papers and their contributions to AGRICULTURAL STATISTICS. Accordingly on 03 January 1947 the Indian Society of Agricultural Statistics was founded at a meeting of statisticians and other agricultural scientists who had gathered together in Delhi on the occasion of the 34th session of Indian Science Congress. Hon'ble Late Dr. Rajendra Prasad, the then Union Agricultural Minister, Government of India was its founder President. Hon'ble Dr. Rajendra Prasad continued to be the President of the Society even after becoming the President of the Republic of India. He remained President of the Society for 16 years since its formation and it was under his guidance that the society grew in its stature. The Society also started publishing a journal called "Journal of the Indian Society of Agricultural Statistics" right since the inception of the society.

Two D's of Statistics

(i) Data (Plural Sense)

By statistics we mean aggregates of facts affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to reasonable standard of accuracy, collected in a systematic manner for a predetermined purpose, and placed in relation to each other.

(ii) Discipline (Singular Sense)

Statistics refers to the body of technique or methodology, which has been developed for the collection, presentation and analysis of quantitative data and for the use of such data in decision making.

From the above two senses of statistics, modern definitions have emerged as given below:

STATISTICS is the practice or science of collecting and analysing numerical data in large quantities, especially for the purpose of inferring proportions in a whole from those in a representative sample.

STATISTICS is a science dealing with drawing conclusions under uncertainty.

STATISTICS is a science comprising of a body of scientific methods for making wise decisions in the face of uncertainty.

STATISTICS is a body of methods for obtaining and analyzing numerical data in order to make better decisions in an uncertain world.

So, from above definitions we find that science of statistics also includes the methods of collecting, organizing, presenting, analyzing and interpreting numerical facts and decisions are taken on their basis.

After analyzing the various definitions of statistics, we may define statistics as:

Statistics in the plural sense are numerical statements of facts capable of some meaningful analysis and interpretation, and in singular sense, it relates to the collection, classification, presentation and interpretation of numerical data.

STATISTICS is a way to generate information from data.

Definition has two components

(a) body of scientific methods which embrace uncertainty

(b) wise decision making

(a)→ fundamental (Basic) research

(b) → data generation – data analysis – data interpretation – knowledge – technology

Role of statistics important in data generation, data analysis and data interpretation.

In NARES, the focus of research in Statistics, in both the D's, is through

- Design of Experiments; Sample Surveys
- Modeling; Simulation; Econometrics; Forecasting.
- Statistical Genetics; Statistical Genomics; Bioinformatics.
- Applications of GIS + Remote Sensing + ANN + Expert Systems Decision Support System / Software Development.
- Big Data / Data Analytics / Data Science / Data Doctor

plane answers to complex questions

salt in statistics

(JM Sengupta, ISI)

- Communal riots in Delhi in 1947 immediately after India achieved Independence
- 'Large number' of minority community took refuge in Red Fort and a small number in Humayun Tomb
- Government responsible to feed these refugees
- Task entrusted to contractors and in absence of any knowledge about the number of refugees, government forced to accept and pay amounts quoted by the contractors for different commodities purchased by them to feed the refugees
- Government expenditure on this account seemed to be extremely high
- Suggested that statisticians may be asked to count the number of refugees inside the Red Fort
- Problem before experts to estimate the number of persons inside a given area without any prior information about the order of magnitude of the number, without having any opportunity to look at the concentrations of persons inside the area and without using any known sampling techniques for estimation or census methods
- Experts had, however, access to bills submitted by contractors to the government, which gave quantities of various commodities such as rice, pulses and salt purchased by them to feed refugees
- Let R, P, and S represent the quantities of rice, pulses and salt used per day to feed all the refugees
- From consumption surveys, per capita requirements of these commodities are known, say, r, p and s respectively
- Then R/r, P/p and S/s must provide parallel (equally valid) estimates of the same number of persons
- When these ratios were computed using values R, P and S quoted by the contractors it was found that S/s had the smallest value and R/r the largest value indicating that the quantity of rice, which is the most expensive commodity compared to salt, was probably exaggerated
- The estimate S/s was proposed by the statisticians for the number of refugees in the Red Fort

- The proposed method verified to provide a good approximation to the number of refugees in the Humayun tomb (the smaller of the two camps with only a relatively small number of refugees), which was independently ascertained
- The estimate provided by the statisticians useful to government in taking administrative decisions
- The method used is unconventional and ingenious, not to be found in any text book
- The idea behind it is statistical reasoning or quantitative thinking. Perhaps, it also involves an element of art

The future of statistics, or at least the next large chunk of future, will be preoccupied, I believe, with problems of large-scale inference raised in our revolutionary scientific environment. For example, how should one analyze 10, 000 related hypothesis tests or 100, 000 correlated estimates at the same time?

- Handling massive data, modeling complex systems and dealing with uncertainty are becoming priorities. All three are primary interests of the discipline of statistics. Never before has statistical knowledge been more important—nor as widely useful—to the scientific enterprise.
- Need for
- Universities to add new programmes in data science and statistics; produce skilled workers to meet the demand for statistical analysis and data mining skills.
- With the explosion of data, statisticians will be in high demand.
- Private (Freelance) Practicing Statisticians like Doctors, Lawyers, etc.

"We are surrounded by data, but starved for insights."

"Information is the oil of the 21st century, and analytics is the combustion engine."

CHAPTER 2

DATA VISUALIZATION IN R STUDIO

M .Iqbal Jeelani Bhat, Manish Kr. Sharma, S.E.H.Rizvi, Sunali Mahajan
Division of Statistics & Computer Science, FBSc
SKUAST- Jammu, Jammu
Email: jeelani.miqbal@gmail.com

Introduction

Data visualization is the graphical representation of information and data by using visual elements like charts, graphs, and maps which provides an accessible way to see and understand trends, outliers, and patterns in data. Our eyes are drawn to colors and patterns and we can quickly identify red from blue, square from circle, besides our culture is visual, including everything from art and advertisements to TV and movies. Data visualization is a form of visual art that grabs our interest and keeps our eyes on the message. When we see a chart, we quickly see trends and outliers. It's storytelling with a purpose and If you have ever stared at a massive spreadsheet of data and couldn't see a trend, you know how much more effective a visualization can be. Data visualization helps to tell stories by transforming data into a form easier to understand, highlighting the trends and outliers. A good visualization of data tells a story, removing the noise from data and highlighting the useful information. However, it's not simply as easy as just dressing up a graph to make it look better or slapping on the "info" part. Effective data visualization is a delicate balancing act between form and function. The plainest graph could be too boring to catch any notice or it make tell a powerful point and the most stunning visualization could utterly fail at conveying the right message or it could speak volumes. The data and the visuals need to work together, and there's an art to combining great analysis with great storytelling, where R software has played an incredible role.

R is basically a system for statistical analysis and graphics which is available freely on internet and can be downloaded on website http://cran-project.org. It was by R. Ihaka and R. Gentleman in 1995 and has has gained popularity in recent past in biological, social, behavioural sciences and has now become the most popular analytical tool in various multinational business industries. (R development core team, 2019). It has an R an IDE version also known as R studio which was developed by J.J Allaire 2011.This software can handle a huge amount of data sets and has a tremendous amount of flexibility for building

attractive visualization of large data sets through one of its incredible package ggplot2 developed by Hadley Wickham 2005. The concept behind ggplot2 divides plot into three different fundamental parts: **Plot = data + Aesthetics + Geometry**.

Components of GGplot2:

- **data** is a data frame
- **Aesthetics** is used to indicate x and y variables. It can also be used to control the **color**, the **size** or the **shape** of points, the height of bars, etc…..
- **Geometry** defines the type of graphics (**histogram**, **box plot**, **line plot**, **density plot**, **dot plot**, ….)

There are two major functions in **ggplot2** package: **qplot()** and **ggplot()** functions.

- **qplot()** stands for quick plot, which can be used to produce easily simple plots.
- **ggplot()** function is more flexible and robust than **qplot** for building a plot piece by piece.

The main function in the ggplot2 package is ggplot(), which can be used to initialize the plotting system with data and x/y variables.Before running ggplot2 we need to install R which can be can be downloaded freely from the Comprehensive R Archive Network (CRAN) webpage (http://cran.r-project.org/). For the sake of simplicity iris data set which is perhaps the best known in built data set available in any R package and contains 3 classes of 50 instances each, where each class refers to a type of iris plant, which can be recalled by executing following code `data(iris)`. This famous (Fisher's or Anderson's) iris data set gives the measurements in centimetres of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are *Iris setosa*, *versicolor*, and *virginica*

```
library(ggplot2)
library(GGally)
library(reshape)
library(corrplot)
library(PerformanceAnalytics)
data("iris")
head(iris)
vol <- ggplot(data=iris, aes(x = Sepal.Length))
vol + stat_density(aes(ymax = ..density.., ymin = -..density..,
```

```
                        fill = Species, color = Species),
                    geom = "ribbon", position = "identity") +
  facet_grid(. ~ Species) + coord_flip() + xlab("Sepal Length")

ggpairs(iris)
g1 <- ggpairs(data=iris, title="ggpairs plot Iris Data set",
              mapping=ggplot2::aes(colour = Species),
              lower=list(combo=wrap("facethist",binwidth=1)))
g1
mydata=iris[c(1:4)]
res<-cor(mydata)
corrplot(res, type = "lower")
pairs(data=res,
      ~ Sepal.Length + Sepal.Width + Petal.Length + Petal.Width)
chart.Correlation(res,
                  method="pearson",
                  histogram=TRUE,
                  pch=16)
col<-colorRampPalette(c("blue","white","red"))(20)
heatmap(x=res,col=col,symm=TRUE)
```

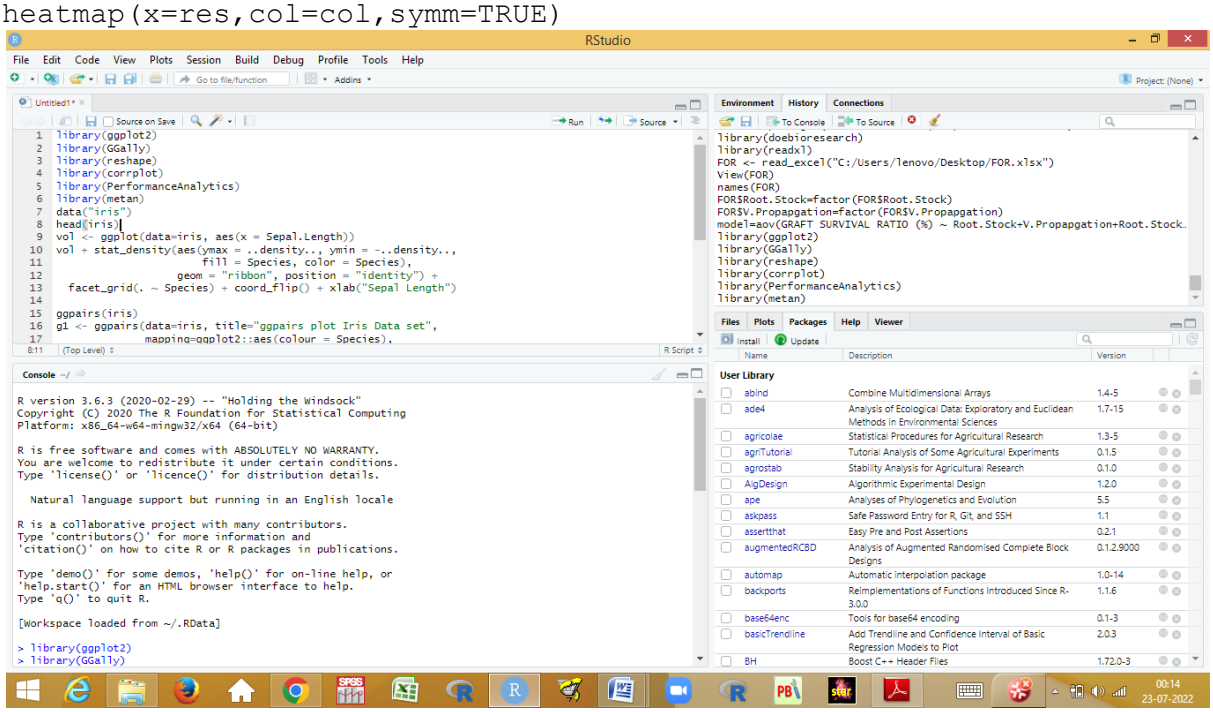
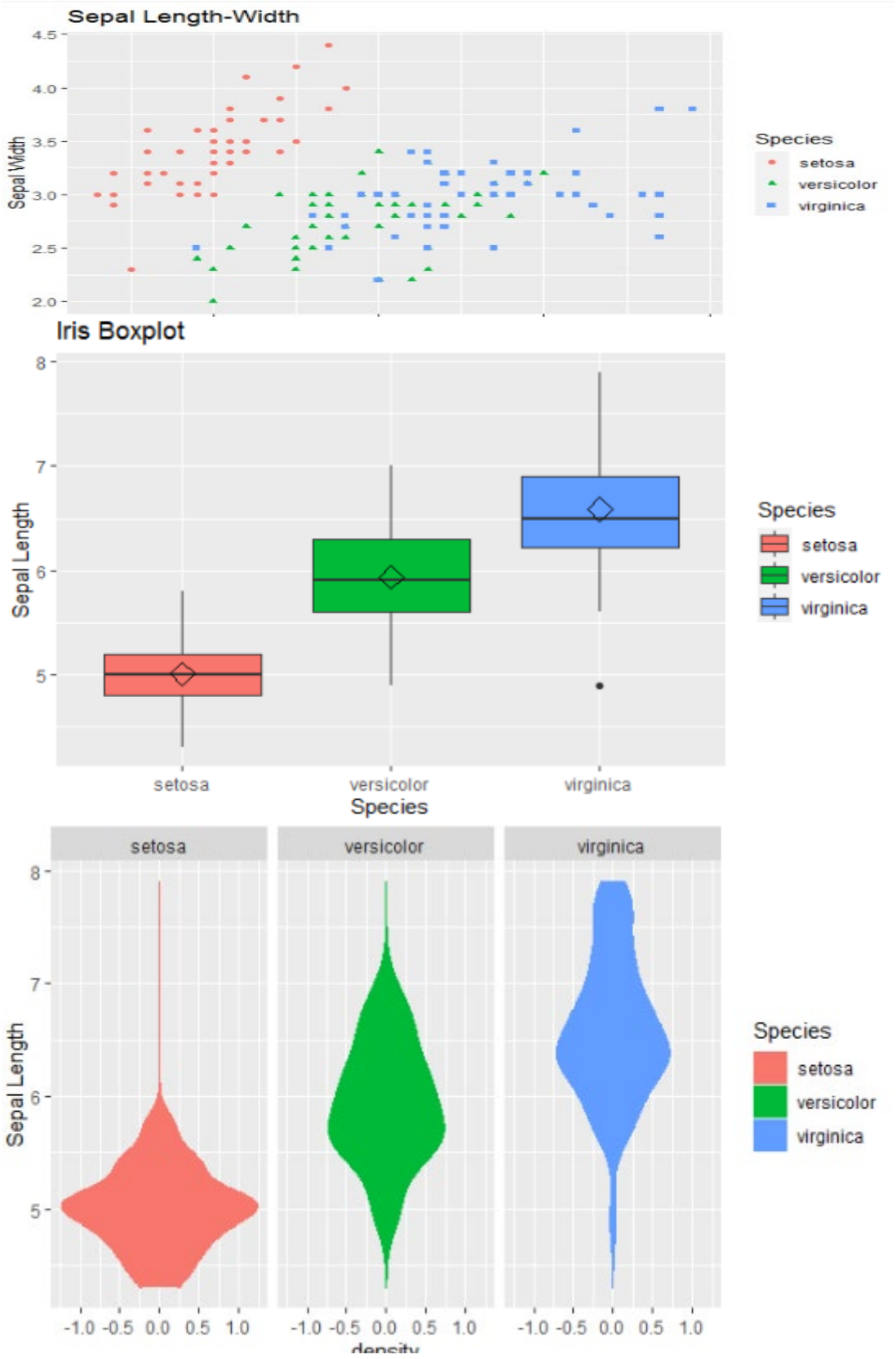


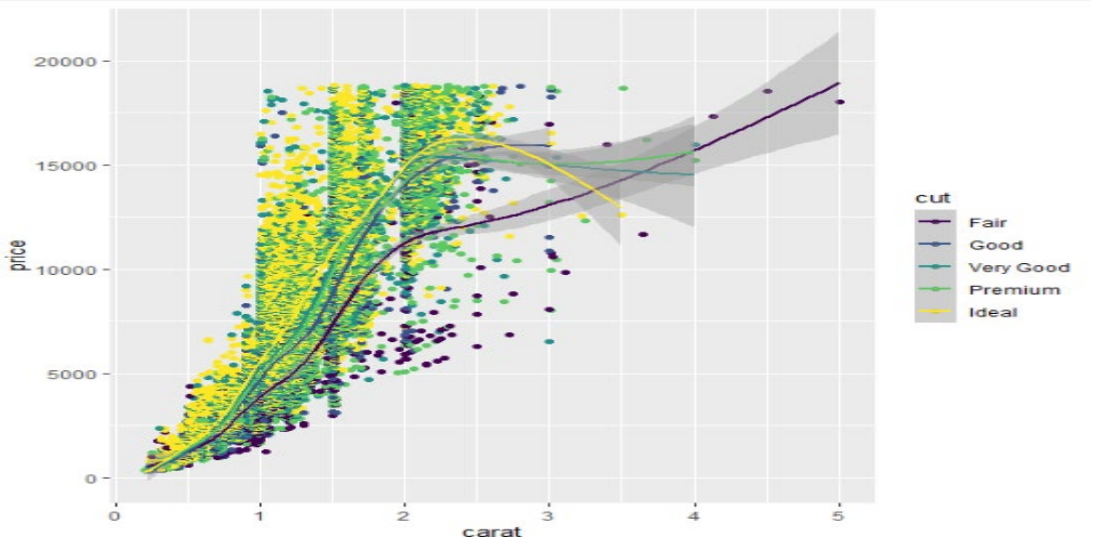
Fig.1 : Overview of R Studio Interface


Sepal Length-Width
Sepal Width
4.5
4.0
3.5
3.0
2.5
2.0
Species
setosa
versicolor
virginica
Iris Boxplot
Sepal Length
8
7
6
5
setosa
versicolor
virginica
Species
Species
setosa
versicolor
virginica
setosa
versicolor
virginica
Sepal Length
8
7
6
5
-1.0 -0.5 0.0 0.5 1.0
-1.0 -0.5 0.0 0.5 1.0
-1.0 -0.5 0.0 0.5 1.0
density
Species
setosa
versicolor
virginica

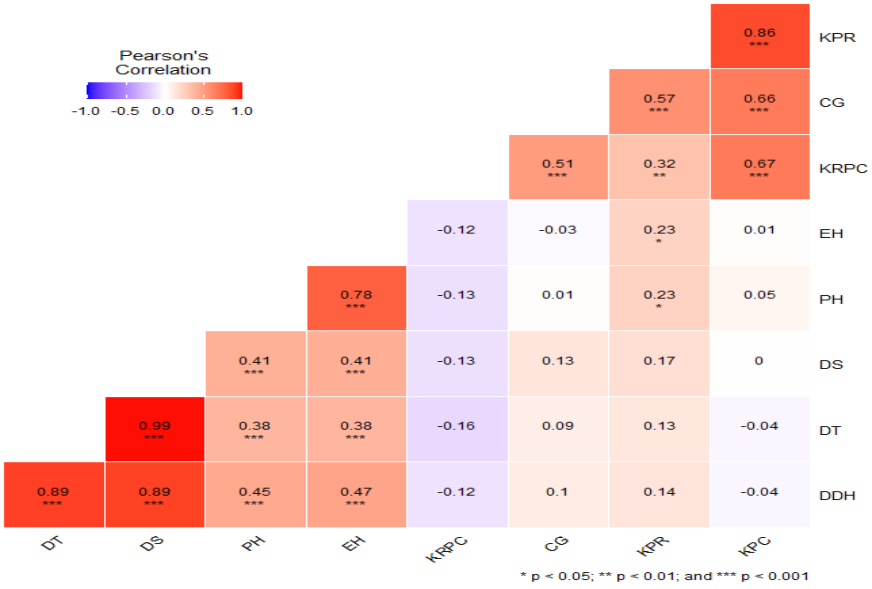

```
library(ggplot2)
data("diamonds")
#A dataset containing the prices and other attributes of almost
54,000 diamonds with 53940 rows and 10 variables:

ggplot(diamonds)  # if only the dataset is known.
ggplot(diamonds, aes(x=carat))  # if only X-axis is known. The Y-
axis can be specified in respective geoms.
```

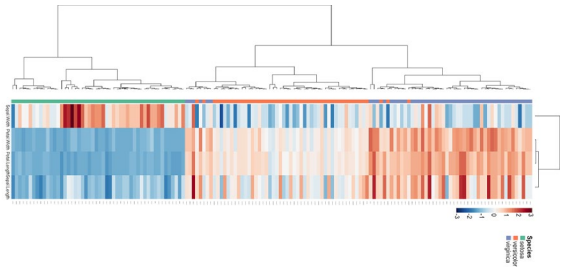
```
ggplot(diamonds, aes(x=carat, y=price))  # if both X and Y axes are fixed for all layers.
ggplot(diamonds, aes(x=carat, color=cut))  # Each category of the 'cut' variable will now have a distinct  color, once a geom is added.
ggplot(diamonds, aes(x=carat), color="steelblue")
ggplot(diamonds, aes(x=carat, y=price, color=cut)) + geom_point() + geom_smooth() # Adding scatterplot geom (layer1) and smoothing geom (layer2)
ggplot(diamonds) + geom_point(aes(x=carat, y=price, color=cut)) + geom_smooth(aes(x=carat, y=price, color=cut)) # Same as above but specifying the aesthetics inside the geoms.
```




```
library(metan)
all <- corr_coef(data)
plot(all)
```



```
#heat map
library(dplyr)
library(NMF)
library(RColorBrewer)
iris2 = iris # prep iris data for plotting
rownames(iris2) = make.names(iris2$Species, unique = T)
iris2 = iris2 %>% select(-Species) %>% as.matrix()
aheatmap(iris2, color = "-RdBu:50", scale = "col", breaks = 0,
annRow = iris["Species"], annColors = "Set2",
distfun = "pearson", treeheight=c(200, 50),
fontsize=13, cexCol=.7,
filename="heatmap.png", width=8, height=16)
```



- Conclusion

The increased popularity of big data and data analysis projects have made visualization more important than ever. Companies are increasingly using machine learning to gather massive amounts of data that can be difficult and slow to sort through, comprehend and explain. Visualization offers a means to speed this up and present information to business owners and stakeholders in ways they can understand, and R packages can facilitate alot due to its powerful graphical capabilities and can allow scientists and researchers to gain greater insight from their experimental data than ever before.

References :

- R Development Core Team.(2019) . R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria.
- R Development Core Team.2016. R: A Language and Environment for Statistical Computing. *The R Foundation for Statistical Computing*. Vienna, Austria. R version 3.2.4 (2016-03-10).*https://www.Rproject.org*.
- Fisher, R. A. (1936) The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**, Part II, 179–188.

CHAPTER 3
TESTING OF HYPOTHESIS AND VARIOUS TEST

Manish Sharma, S.E.H. Rizvi, Faizan Danish, Sunali Mahajan and Nishant Jasrotia
Division of Statistics and Computer Science, FBSc, SKUAST Jammu
Email to: manshstat@gmail.com

Introduction

One of the most important test within the branch of inferential statistics is the Student's t-test. The Student's t-test for two samples is used to test whether two groups (two populations) are different in terms of a quantitative variable, based on the comparison of two samples drawn from these two groups. In other words, a Student's t-test for two samples allows to determine whether the two populations from which your two samples are drawn are different (with the two samples being measured on a quantitative continuous variable). The reasoning behind this statistical test is that if your two samples are markedly different from each other, it can be assumed that the two populations from which the samples are drawn are different. On the contrary, if the two samples are rather similar, we cannot reject the hypothesis that the two populations are similar, so there is no sufficient evidence in the data at hand to conclude that the two populations from which the samples are drawn are different. Note that this statistical tool belongs to the branch of inferential statistics because conclusions drawn from the study of the samples are generalized to the population, even though we do not have the data on the entire population. To compare two samples, it is usual to compare a measure of central tendency computed for each sample. In the case of the Student's t-test, the mean is used to compare the two samples. However, in some cases, the mean is not appropriate to compare two samples so the median is used to compare them via the Wilcoxon test. The two tests (Student's t-test and Wilcoxon test) have the same final goal, that is, compare two samples in order to determine whether the two populations from which they were drawn are different or not. Note that the Student's t-test is more powerful than the Wilcoxon test (i.e., it more often detects a significant difference if there is a true difference, so a smaller difference can be detected with the Student's t-test) but the Student's t-test is sensitive to outliers and data asymmetry. Furthermore, within each of these two tests, several versions exist, with each version using different formulas to arrive at the final result. It is thus necessary to understand the difference between the two tests and which version to use in order to carry out the

appropriate analyses depending on the question and the data at hand. In particular, the test provides good results even when the population is not normal or the sample size is small, provided that the sample is reasonably symmetrically distributed about the sample mean. This can be determined by graphing the data. The following are indications of symmetry:

- The boxplot is relatively symmetrical; i.e. the median is in the center of the box and the whiskers extend equally in each direction
- The histogram looks symmetrical
- The mean is approximately equal to the median
- The coefficient of skewness is relatively small

The impact of non-normality is less for a two-tailed test than for a one-tailed test and for higher alpha values than for lower alpha values. The other assumption for the t-test is that we have a random sample. If, for example, we are interested in the mean cholesterol level of a population, then our sample must consist of the cholesterol levels of people chosen at random. We can't use the t-test for a sample consisting of cholesterol levels for the same person at different points in time.

Null and alternative hypothesis

Before diving into the computations of the Student's t-test by hand, let's recap the null and alternative hypotheses of this test:$H_0: \mu_1 = \mu_2$ vs $H_1: \mu_1 \neq \mu_2$, where μ_1 and μ_2 are the means of the two populations from which the samples were drawn. As mentioned in the introduction, although technically the Student's t-test is based on the comparison of the means of the two samples, the final goal of this test is actually to test the following hypotheses:H_0: the two populations are similar vs H_1: the two populations are different. This is in the general case where we simply want to determine whether the two populations are **different** or not (in terms of the dependent variable). In this sense, we have no prior belief about a particular population mean being larger or smaller than the other. This type of test is referred as a **two-sided** or bilateral test. If we have some prior beliefs about one population mean being larger or smaller than the other, the Student's t-test also allows to test the following hypotheses:$H_0: \mu_1 = \mu_2$ vs $H_1: \mu_1 > \mu_2$ Or $H_0: \mu_1 = \mu_2$ vs $H_1: \mu_1 < \mu_2$

In the first case, we want to test if the mean of the first population is significantly larger than the mean of the second, while in the latter case, we want to test if the mean of the first

population is significantly smaller than the mean of the second. This type of test is referred as a one-sided or unilateral test. Some authors argue that one-sided tests should not be used in practice for the simple reason that, if a researcher is so sure that the mean of one population is larger (smaller) than the mean of the other and would never be smaller (larger) than the other, why would she needs to test for significance at all? This a rather philosophical question and it is beyond the scope of this article. Interested readers are invited to see part of the discussion in Rowntree (2000).

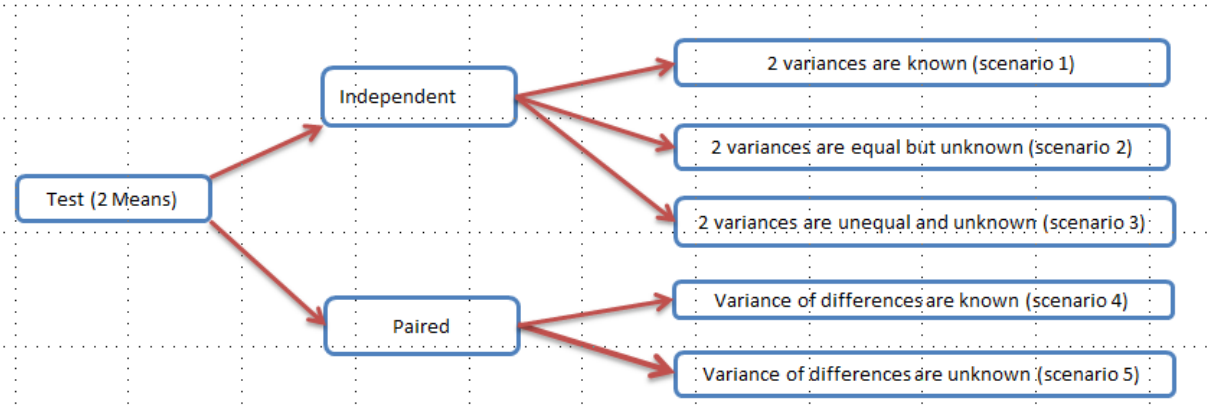
Hypothesis testing

In statistics, many statistical tests are in the form of hypothesis tests. Hypothesis tests are used to determine whether a certain belief can be deemed as true (plausible) or not, based on the data at hand (i.e., the sample(s)). Most hypothesis tests boil down to the following 4 steps:

1) State the null and alternative hypothesis.
2) Compute the test statistic, denoted t-stat. Formulas to compute the test statistic differ among the different versions of the Student's t-test but they have the same structure.
3) Find the critical value given the theoretical statistical distribution of the test, the parameters of the distribution and the significance level α.For a Student's t-test and its extended version, it is either the normal or the Student's t distribution (t denoting the Student distribution and z denoting the normal distribution).
4) Conclude by comparing the t-stat (found in step 2.) with the critical value (found in step. 3). If the t-stat lies in the rejection region (determined thanks to the critical value and the direction of the test), we reject the null hypothesis, otherwise we do not reject the null hypothesis. These two alternatives (reject or do not reject the null hypothesis) are the only two possible solutions, we never "accept" an hypothesis. It is also a good practice to always interpret the decision in the terms of the initial question.

Different versions of the Student's t-test

There are several versions of the Student's t-test for two samples, depending on whether the samples are independent or paired and depending on whether the variances of the populations are (un)equal and/or (un)known:



On the one hand, **independent** samples means that the two samples are collected on **different** experimental units or different individuals, for instance when we are working on women and men separately, or working on patients who have been randomly assigned to a control and a treatment group (and a patient belongs to only one group). On the other hand, we face **paired** samples when measurements are collected on the **same** experimental units, same individuals. This is often the case, for example in medical studies, when testing the efficiency of a treatment at two different times. The same patients are measured twice, before and after the treatment, and the dependency between the two samples must be taken into account in the computation of the test statistic by working on the **differences** of measurements for each subject. Paired samples are usually the result of measurements at two different times, but not exclusively. Suppose we want to test the difference in vision between the left and right eyes of 50 athletes. Although the measurements are not made at two different time (before-after), it is clear that both eyes are dependent within each subject. Therefore, the Student's t-test for paired samples should be used to account for the dependency between the two samples instead of the standard Student's t-test for independent samples. Another criteria for choosing the appropriate version of the Student's t-test is whether the variances of the populations (not the variances of the samples!) are known or unknown and equal or unequal. This criteria is rather straightforward, we either know the variances of the populations or we do not. The variances of the populations cannot be computed because if you can compute the variance of a population, it means you have the data for the whole population, then there is no need to do a hypothesis test anymore… So the variances of the populations are either given in the statement (use them in that case), or there is no information about these variances and in this case, it is assumed that the variances are unknown. In practice, the variances of the populations are most of the time unknown and the only thing to do in order to choose the appropriate version of the test is to check whether the

variances are equal or not. However, we still illustrate how to do all versions of this test by hand and in R in the next sections following the 4 steps of hypothesis testing.

Assumptions

As for many statistical tests, there are some assumptions that need to be met in order to be able to interpret the results. When one or several of them are not met, although it is technically possible to perform these tests, it would be incorrect to interpret the results or trust the conclusions. Below are the assumptions of the Student's t-test for two samples, how to test them and which other tests exist if an assumption is not met:

Variable type:

A Student's t-test requires a mix of one quantitative dependent variable (which corresponds to the measurements to which the question relates) and one qualitative independent variable (with exactly 2 levels which will determine the groups to compare).

Independence:

The data, collected from a representative and randomly selected portion of the population, should be independent between groups and within each group. The assumption of independence is most often verified based on the design of the experiment and on the good control of experimental conditions rather than via a formal test. If you are still unsure about independence based on the experiment design, ask yourself if one observation is related to another (if one observation has an impact on another) within each group or between the groups themselves. If not, it is most likely that you have independent samples. If observations between samples (forming the different groups to be compared) are dependent (for example, if two measurements have been collected on the same individuals as it is often the case in medical studies when measuring a metric (i) before and (ii) after a treatment), the paired version of the Student's t-test, called the Student's t-test for paired samples, should be preferred in order to take into account the dependency between the two groups to be compared.

Normality:

- ✓ With small samples (usually $n<30$), when the two samples are independent, observations in **both samples** should follow a normal distribution. When using the Student's t-test for paired samples, it is the difference between the observations of the two samples that should follow a normal distribution. The normality assumption can be tested visually thanks to a histogram and a QQ-plot, and/or formally via

a normality test such as the Shapiro-Wilk or Kolmogorov-Smirnov test. If, even after a transformation (e.g., logarithmic transformation, square root, etc.), your data still do not follow a normal distribution, the Wilcoxon test (wilcox.test(variable1 ~ variable2, data = dat in R) can be applied. This non-parametric test, robust to non normal distributions, compares the medians instead of the means in order to compare the two populations.

- ✓ With large samples (usually n≥30n≥30), **normality of the data is not required** (this is a common misconception!). By the central limit theorem, sample means of large samples are often well-approximated by a normal distribution even if the data are not normally distributed (Stevens 2013). It is therefore not required to test the normality assumption when the number of observations in each group/sample is large.

Equality of variances:

When the two samples are independent, the variances of the two groups should be equal in the populations (an assumption called **homogeneity of the variances**, or even sometimes referred as homoscedasticity, as opposed to heteroscedasticity if variances are different across groups). This assumption can be tested graphically (by comparing the dispersion in a boxplot or dotplot for instance), or more formally via the Levene's test (leveneTest(variable ~ group) from the {car} package) or via a F test (var.test(variable ~ group)). If the hypothesis of equal variances is rejected, another version of the Student's t-test can be used: the Welch test (t.test(variable ~ group, var.equal = FALSE)). Note that the Welch test does not require homogeneity of the variances, but the distributions should still follow a normal distribution in case of small sample sizes. If your distributions are not normally distributed or the variances are unequal, the Wilcoxon test should be used. This test does not require the assumptions of normality nor homoscedasticity of the variances.

Outliers:

An outlier is a value or an observation that is distant from the other observations. There should be no significant outliers in the two groups, or the conclusions of your t-test may be flawed. There are several methods to detect outliers in your data but in order to deal with them, it is your choice to either:

- ✓ use the non-parametric version (i.e., the Wilcoxon test)
- ✓ transform your data (logarithmic or Box-Cox transformation, among others)
- ✓ or remove them (be careful)

Concept of Variance

Variance is an important tool in the sciences including statistical science. In the Theory of Probability and statistics, variance is the expectation of the squared deviation of a random variable from its mean. Actually, it is measured to find out the degree to which the data in series are scattered around its average value. Variance is widely used in statistics, its use is ranging from descriptive statistics to statistical inference and testing of hypothesis. Relationship among variables under the said analysis, we use to examine the differences in the mean values of the dependent variable associated with the effect of the controlled independent variables, after taking into account the influence of the uncontrolled independent variables. We take the null hypothesis that there is no significant difference between the means of different populations. In its simplest form, analysis of variance must have a dependent variable that is metric (measured using an interval or ratio scale). There must also be one or more independent variables. The independent variables must be all categorical (non-metric). Categorical independent variables are also called factors. A particular combination of factor levels, or categories, is called a treatment. What type of analysis would be made for examining the variations depends upon the number of independent variables taken into account for the study purpose. One-way analysis of variance involves only one categorical variable, or a single factor. If two or more factors are involved, the analysis is termed n-way (eg. Two-Way, Three-Way etc.) F-tests are named after the name of Sir Ronald Fisher. The F-statistic is simply a ratio of two variances. Variance is the square of the standard deviation. For a common person, standard deviations are easier to understand than variances because they're in the same units as the data rather than squared units. F-statistics are based on the ratio of mean squares. The term "mean squares" may sound confusing but it is simply an estimate of population variance that accounts for the degrees of freedom (DF) used to calculate that estimate. F test can be defined as a test that uses the F-test statistic to check whether the variances of two samples (or populations) are equal to the same value. To conduct an F-test, the population should follow an F- distribution and the samples must be independent events. On conducting the hypothesis test, if the results of the f test are statistically significant then the null hypothesis can be rejected otherwise it cannot be rejected.

Assumptions

The assumptions for F-test for testing the variances of two populations are:

1. The populations from which the samples are drawn must be normally distributed.
2. The samples must be independent of each other.

The f test is used to check the equality of variances using hypothesis testing. The f test formula for different hypothesis tests is given as follows:

Left Tailed Test: Null Hypothesis: $H_0: \sigma_1^2 = \sigma_2^2$ vs **Alternative Hypothesis: $H_1: \sigma_1^2 < \sigma_2^2$**

Decision Criteria: If the f statistic < f critical value then reject the null hypothesis.

Right Tailed test: Null Hypothesis: $H_0: \sigma_1^2 = \sigma_2^2$ vs **Alternative Hypothesis: $H_1: \sigma_1^2 > \sigma_2^2$**

Decision Criteria: If the f test statistic > F- test critical value then reject the null hypothesis

Two Tailed test: Null Hypothesis: $H_0: \sigma_1^2 = \sigma_2^2$ vs **Alternative Hypothesis: $H_1: \sigma_1^2 \neq \sigma_2^2$**

Decision Criteria: If the F- test statistic > f test critical value then the null hypothesis is rejected

F Statistic

The F- test statistic or simply the F- statistic is a value that is compared with the critical value to check if the null hypothesis should be rejected or not. The F- test statistic formula for large samples:$F = \frac{\sigma_1^2}{\sigma_2^2}$, where σ_1^2 is the variance of the first population and σ_2^2 is the variance of the second population. And , F statistic for small samples: $F = \frac{s_1^2}{s_2^2}$, where s_1^2 is the variance of the first sample ands_2^2 is the variance of the second sample. The selection criteria for the σ_1^2 and σ_2^2 for an f statistic is given below:

- For a right-tailed and a two-tailed F-test, the variance with the greater value will be in the numerator. Thus, the sample corresponding to σ_1^2 will become the first sample. The smaller value variance will be the denominator and belongs to the second sample.
- For a left-tailed test, the smallest variance becomes the numerator (sample 1) and the highest variance goes in the denominator (sample 2).

F Test Critical Value

A critical value is a point that a test statistic is compared to in order to decide whether to reject or not to reject the null hypothesis. Graphically, the critical value divides a distribution into the acceptance and rejection regions. If the test statistic falls in the rejection region then the null hypothesis can be rejected otherwise it cannot be rejected. The steps to find the f test critical value at a specific alpha level (or significance level), α, are as follows:

- Find the degrees of freedom of the first sample. This is done by subtracting 1 from the first sample size. Thus, $x = n_1-1$
- Determine the degrees of freedom of the second sample by subtracting 1 from the sample size. This given $y = n_2-1$
- If it is a right-tailed test then α is the significance level. For a left-tailed test 1 - α is the alpha level. However, if it is a two-tailed test then the significance level is given by α / 2.
- The F table is used to find the critical value at the required alpha level.
- The intersection of the x column and the y row in the f table will give the f test critical value.

ANOVA F Test : The one-way ANOVA is an example of an F-test. ANOVA stands for analysis of variance. It is used to check the variability of group means and the associated variability in observations within that group. The F- test statistic is used to conduct the ANOVA test. The hypothesis is given as follows:

H_0: The means of all groups are equal.
H_1: The means of all groups are not equal.
Test Statistic: F = explained variance / unexplained variance
Decision rule: If F > F critical value then reject the null hypothesis.
To determine the critical value of an ANOVA f test the degrees of freedom are given by df_1 = K - 1 and df_1= N - K, where N is the overall sample size and K is the number of groups.

F Test vs T-Test

F test and t-test are different types of statistical tests used for hypothesis testing depending on the distribution followed by the population data. The table given below outlines the differences between the F test and the t-test.

F-test	**T-test**
An F test is a test statistic used to check the equality of variances of two populations	The T-test is used when the sample size is small (n < 30) and the population standard deviation is not known.
The data follows an F distribution	The data follows a Student t-distribution
The F test statistic is given as $F = \frac{\sigma_1^2}{\sigma_2^2}$,	The t-test statistic for 1 sample is given by $t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$, where $\bar{X}$ is the sample mean, μ is the population mean, s is the sample standard deviation and n is the sample size.
The f test is used for variances.	It is used for testing means.

Important Notes on F Test

- The f test is a statistical test that is conducted on an F distribution in order to check the equality of variances of two populations.
- The F-test formula for the test statistic is given by $F = \frac{\sigma_1^2}{\sigma_2^2}$.
- The f critical value is a cut-off value that is used to check whether the null hypothesis can be rejected or not.
- A one-way ANOVA is an example of an F-test that is used to check the variability of group means and the associated variability in the group observations.

Example 1: A research team wants to study the effects of a new drug on insomnia. 8 tests were conducted with a variance of 600 initially. After 7 months 6 tests were conducted with a variance of 400. At a significance level of 0.05 was there any improvement in the results after 7 months?

Solution: As the variance needs to be compared, the f test needs to be used.

Null Hypothesis: $\mathbf{H_0: s_1^2 = s_2^2}$

Alternative Hypothesis: $\mathbf{H_1: s_1^2 > s_2^2}$

$n_1 = 8, n_2 = 6$

$df_1 = 8 - 1 = 7$

$df_2 = 6 - 1 = 5$

$\mathbf{s_1^2} = 600, \mathbf{s_2^2} = 400$

The f test formula is given as follows:

$$F = \frac{s_1^2}{s_2^2} = 600 / 400$$

$$F = 1.5$$

Now from the F table the critical value F(0.05, 7, 5) = 4.88

The F-Distribution with $\alpha = 0.05$								
$\nu_2 \backslash \nu_1$	2	3	4	5	6	7	8	9
2	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
3	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18

As 1.5 < 4.88, thus, the null hypothesis cannot be rejected and there is not enough evidence to conclude that there was an improvement in insomnia after using the new drug. **Answer:** Fail to reject the null hypothesis.

Example 3: A toy manufacturer wants to get batteries for toys. A team collected 41 samples from supplier A and the variance was 110 hours. The team also collected 21 samples from supplier B with a variance of 65 hours. At a 0.05 alpha level determine if there is a difference in the variances.

Solution: This is an example of a two-tailed F test. Thus, the alpha level is 0.05 / 2 = 0.025

Null Hypothesis: $\mathbf{H_0: s_1^2 = s_2^2}$

Alternative Hypothesis: $\mathbf{H_1: s_1^2 \neq s_2^2}$

$$n_1 = 41, n_2 = 21$$

$$df_1 = 41 - 1 = 40$$

$$df_2 = 21 - 1 = 20$$

$$\mathbf{s_1^2} = 110, \mathbf{s_2^2} = 65$$

$$F = \frac{s_1^2}{s_2^2} = 110 / 65$$

$$F = 1.69$$

Using the F table F(0.025, 40, 20) = 2.287

As 1.69 < 2.287 thus, the null hypothesis cannot be rejected,

Answer: Fail to reject the null hypothesis.

Analysis of testing problem through SPSS

Professor Jessica Utts quoted "Decisions or predictions are often based on data—numbers in context. These decisions or predictions would be easy if the data always sent a clear message, but the message is often obscured by variability. Statistics provides tools for describing variability in data and for making informed decisions that take it into account". Now a days, Statistics has played a vital role in Veterinary, Agriculture, Medical and Allied Sciences starting from planning of the experiment to the analysis of data. The experimentation comprises of two major components *viz.,* designing the experiment (or the way of generating the data) and the analysis of data generated to draw meaningful and valid conclusions. It is through statistical packages only, the drawn results and designs could be developed by graphics, which is simpler to understand even by a layman.

In order to make our research more competitive, the modern statistical methodologies are required for the collection and analysis of data and the interpretation of results. A wide variety of proprietary software packages such as Statistical Analysis System (SAS), Statistical Package for Social Sciences (SPSS), SYSTAT, JMP, GENSTAT, GLIM, MINITAB, S-PLUS, STATISTICA etc. MS-Excels is also very useful to solve a wide type of problems. The free online available software packages are Libre Office cal, PSPP, R and Epilnfo. Here, it is not possible to discuss all in details. So, in this chapter the SPSS software package is discussed which, is a very comprehensive and widely available package for statistical analysis.

SPSS (originally standing for 'Statistical Package for the Social Sciences) but now '(Statistical Product and Service Solutions)' or '(Superior performing Statistical Software)' is one of the leading programmes for managing and analyzing social and scientific data. It was one of the earliest statistical package with version1 being released in 1968, well before the advent of computers. SPSS is now one of the most widely accepted statistical analysis package used worldwide. The capability of SPSS is truly amazing; it enables you to obtain statistics ranging from simple descriptive numbers to complex analyses of multivariate matrices. You can plot the data in histograms, scatter plots, and other ways. You can combine files, split files, and sort files. You can modify existing variables and create new ones. In

short, you can do just about anything you'd ever want with a set of data using this software package.

SPSS includes the following:

- Comprehensive statistical capabilities, including descriptive statistics, simulations and distributions, elementary inferential statistics, analysis of variance, regression, analysis of categorical data, multivariate analysis, non-parametric, control charts and time series analysis.
- A menu-driven interface providing easy access to SPSS statistical, graphical and data management capabilities.
- A data window permitting data entry, editing, and browsing in a spreadsheet like display.
- The ability to import and export data, including data from excel.

Feature of SPSS Software:

- One of the most powerful statistical packages that is also easy to use.
- Can use either with menus or syntax files
- It includes a full range of data management system and editing tools
- It provides in-depth statistical capabilities
- More powerful then MINITAB
- SPSS is a powerful and versatile tool that will enhance their learning experience.
- One of the most widely used statistical packages in academia
- It offers complete plotting, reporting and presentation features

There are a number of different types of windows of SPSS the most common in used are as

(i) **Data Editor Window** is the first window you encounter while clicking on the icon of the SPSS which, is used to define and enter your data and to perform statistical procedures.

(ii) **Output Window** displays the results of the statistical tests.

(iii) **Syntax Window** can be used to keep a record of the operations that you perform on your data. This window will be open automatically when someone clicks a **Paste** function of the dialogue box.

These windows can be saved. Basic information on these windows is summarized as.

Windows	**File Suffix**	**Function**
Editor	.sav	Used to define, enter, and edit your data and to run statistical tests
Output	.spo	Contains the results of the statistical procedures
Syntax	.sps	This window is activated when you click on the Paste function and records a record of the operations that are "pasted." Although this is beyond the goals of these lessons, you may want to know that SPSS commands can actually be run from this window.

The utility of this statistical software has been discussed with the help of numerical examples in the following sections which include:

E. **ONE SAMPLE T TEST**

Example: The length of snout of 12 fishes of a species in a tank was observed as 61, 62, 62, 63, 64, 65, 66, 66, 69, 69, 70, 71 mm. Is the mean length of snout of all fishes of the tank is 65mm?

SPSS Commands (Step by Step):

Go to Analyze: Click on compare means and then one sample T test ; Shift Y to test variable : Give test value equal to 65 : Click OK.

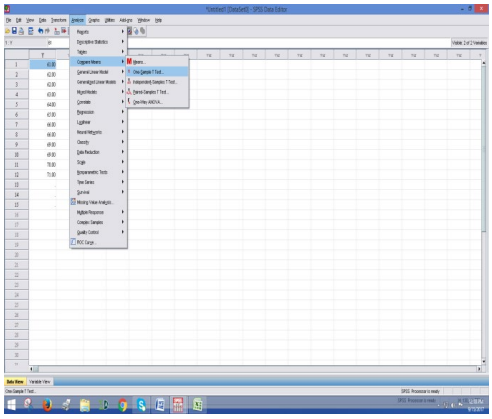


Fig. 11: Selection of command (Compare means one sample t test) from the menu

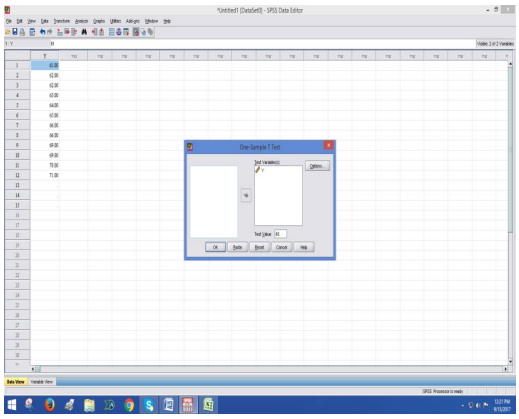


Fig. 12: Selection of command (Y to test variable) from the menu

Output:

One-Sample Test						
	Test Value = 65					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Y	.675	11	.513	.66667	-1.5064	2.8397

Interpretation: Here, the **p value** of t test is greater than 0.05 indicating that there is no change in the mean length of snout of all fishes.

F. PAIRED T TEST

Example: The weights of 15 pigs when brought in piggery and after six months are given below. Check whether the gain in weight is statistically significant or not.

weight when bought first	38	45	43	43	35	46	41	47	40	36	37	36	46	41	49
weight after six months	49	52	49	48	46	49	42	48	45	48	49	41	48	46	50

SPSS Commands (Step by Step):

1. Go to Analyze
2. Click on compare means and then paired t test.
3. Shift Y and X together to paired variable.
4. Click OK.

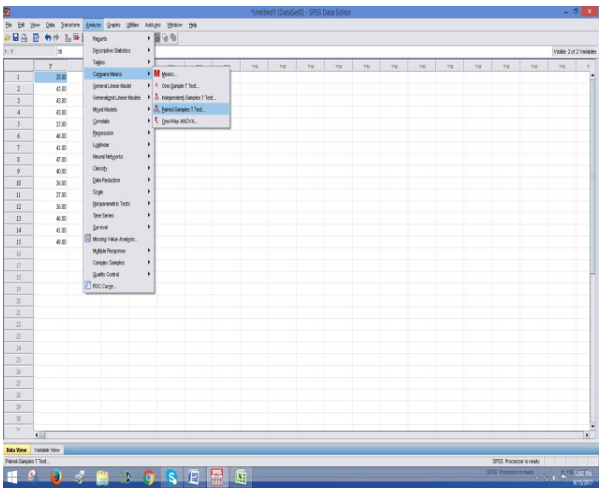


Fig. 13: Selection of command (compare means paired t test) from the menu

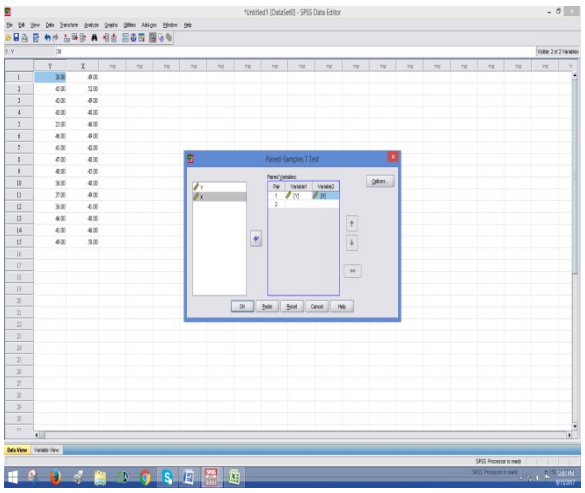


Fig. 14: Selection of command (Y and X to var.1 and var 2. resp.)

Output:

Paired Samples Test

	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
Pair 1 Y - X	-5.80000	4.02137	1.03831	-8.02696	-3.57304	-5.586	14	.000

Interpretation: Here, the p value is less than 0.05; therefore, null hypothesis can be rejected. Hence, it can be concluded that the gain in weight is statistically significant when the pigs are brought to piggery.

G. **INDEPENDENT SAMPLES T TEST**

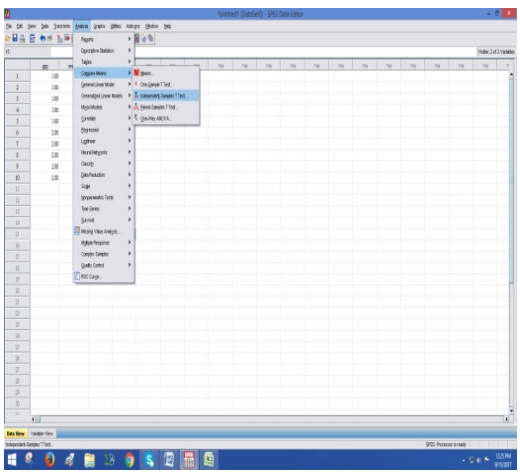
Example: Following data have been obtained when two groups of dogs were weighed following inhalation of a vapour.

Dog weight with inhalation of vapour(x_1)(kg)	Control group(x_2)(kg)
8.2	8.3
8.4	8.9
9.1	9.6
9.3	9.7
9.3	9.4

Test whether the weights of the two groups of dogs are same or not.

SPSS Commands (Step by Step):

Go to Analyze : Click on compare means and then independent sample t test ; Shift X_1 and X_2 together under test variable and coding variable to Group variable. Click OK.



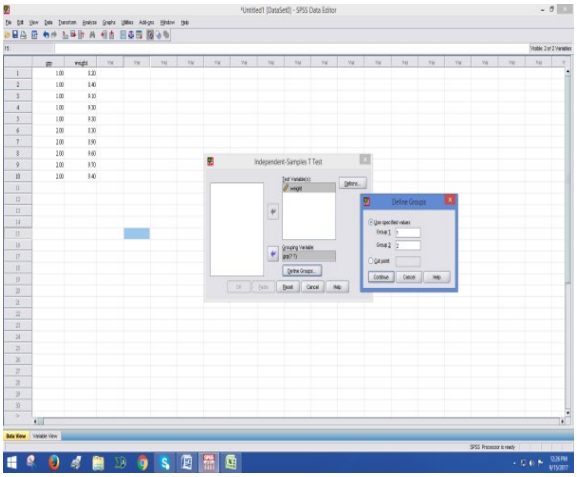


Fig. 15: Selection of command (compare means independent sample test) from the menu
Fig. 16: Selection of command (test variable and define groups.)

Output:

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
weight	Equal variances assumed	.014	.908	-.916	8	.386	-.32000	.34928	-1.12545	.48545
	Equal variances not assumed			-.916	7.913	.387	-.32000	.34928	-1.12700	.48700

Interpretation: Since, the p value of the F statistic is greater than 0.05 indicating that the two groups have equal variances. Therefore, we will apply the t-test with equal variances otherwise with unequal variances. Further, the p-value of the t-test is more than 0.05 so the test is insignificant means accept the null hypothesis. Thus on average, weights of the two groups of dogs are same.

H. CHI SQUARE

Example: RBCs count lac/mm^3 and Hb% g/100ml of 500 persons of test locality was recorded as follows. Is there any significant relation between RBCs count and Hb%?

RBCs count	HB%	
	Above Normal	Below Normal
Above Normal	85	75
Below Normal	165	175

SPSS Commands (Step by Step):

1. Go to data: Give Weight case by F and click OK : Go to Analyze : Click on Descriptive : Go to Crosstab : Shift R to Row , C to Column : Click on Statistic and select Chi square: Click to Continue and then OK.

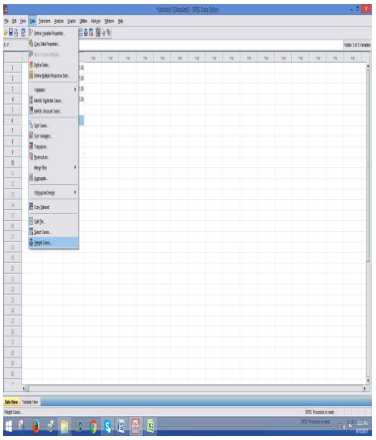


Fig.17 Insertion of the Data into the Data window

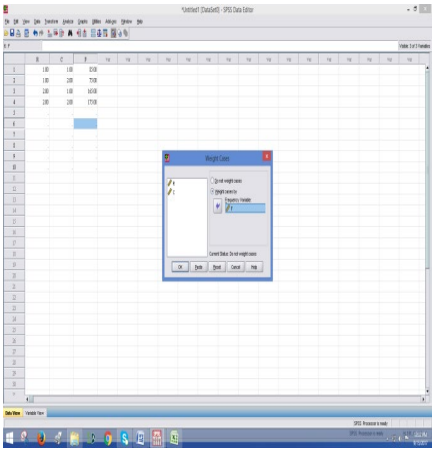


Fig.18 Assigning weight to the Variable F

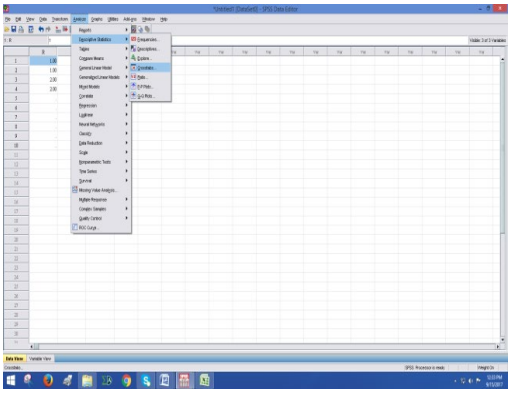


Fig.19 Selection of the command crosstab (Descriptive statistics)

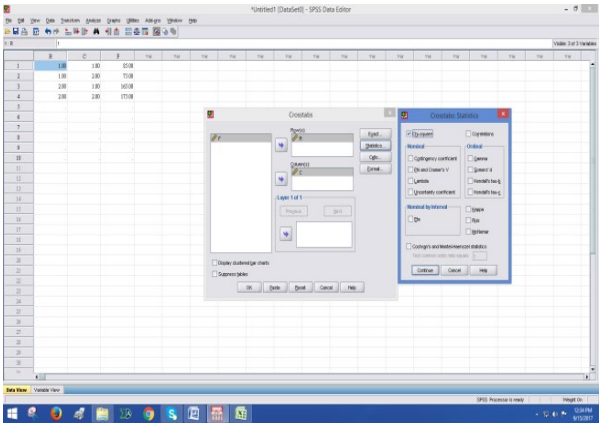


Fig.20 Selection of the variables and test chi square

Output:

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.919[a]	1	.338		
Continuity Correction[b]	.744	1	.388		
Likelihood Ratio	.920	1	.338		
Fisher's Exact Test				.388	.194
Linear-by-Linear Association	.917	1	.338		
N of Valid Cases[b]	500				

Interpretation: From the above table; it has been observed that the Pearson chi square value is significant as p- value is more than 0.05. Thus, it can be concluded that the two attributes are independent of each other.

References:

- Ajai S. Gaur & Sanjaya S. Gaur 2009. Statistical Methods for Practice and Research: a guide to Data analysis Using SPSS. Sage Publications Inc.
- L. S. Aiken and S.G. West .199). Multiple Regression: testing and Interpreting Interactions. Newbury Park, CA: Sage Publications
- Manish Kumar Sharma, Anil Bhat and M I J Bhat. 2017. Computer in Agriculture: Fundamentals and research. New India Publishing Agency.
- Ramez Elamseri, 2008. Fundamental of Database Systems, Pearson Education India.
- V.N. Amble. 1975. Statistical Methods in Animal Sciences. Indian Society of Agricultural Statistics, New Delhi.
- Rowntree, Derek. 2000. Statistics Without Tears.
- Stevens, James P. 2013. Intermediate Statistics: A Modern Approach. Routledge.

CHAPTER 4

LINEAR REGRESSION & LOGISTIC REGRESSION

Sunali Mahajan, Manish Sharma, M. Iqbal Jeelani Bhat and Shavi Gupta

"Statisticians are like artists, have the bad habit of falling in love with their models."
George Box

Regression analysis is the art and science of fitting straight lines to patterns of data. In linear regression model, the relationship between dependent variable and one or more independent variables can be obtained by fitting a linear equation/best fit line to observed data. A linear regression line equation is written in the form of: $Y=\beta o+\beta 1X$ where, X is the independent variable (plotted along the x-axis), Y is the dependent variable (plotted along the y-axis), β1 is the slope of the line, and βo is the intercept (the value of y when x = 0). Based on the number of input features, linear regression could be of two types:

Simple Linear Regression (SLR): In Simple Linear Regression (SLR), we will have a single input variable based on which we predict the output variable. Input variables can also be termed as Independent/predictor variables, and the output variable is called the dependent variable. The equation for SLR is $y=\beta o+\beta 1x+\epsilon$, where, y is the dependent variable, x is the predictor, βo, β1 are coefficients/parameters of the model, and Epsilon(ϵ) is a random variable called Error Term.

Multiple Linear Regression (MLR): In Multiple Linear Regression (MLR), we predict the output based on multiple inputs, also known as multivariable linear regression. In the case of multiple regression, there will be a set of independent variables that helps us to explain better or predict the dependent variable y. The multiple regression equation is given by $\mathbf{y} = \beta o + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_k x_k$ where; y is the dependent variable, x's are the predictors, β's are the coefficients/parameters of the model. Almost all real-world regression patterns include multiple predictors, and basic explanations of linear regression are often explained in terms of the multiple regression form.Least Square Regression Line or Linear Regression Line: The most popular method to fit a regression line in the XY plot is the method of least-squares. This process determines the best-fitting line for the noted data by reducing the sum of the squares of the vertical deviations from each data point to the line. If a point rests on the fitted line accurately, then its perpendicular deviation is 0. Because the variations are first squared, then added, their positive and negative values will not be cancelled. Linear regression

determines the straight line, called the least-squares regression line or LSRL, that best expresses the observations in a bivariate analysis of data set. Suppose Y is a dependent variable, and X is an independent variable, then the population regression line is given by: $Y = \beta o + \beta 1X$. If a random sample of observations is given, then the regression line is expressed by; $\hat{y} = bo + b1x$ where, b_0 is a constant, b_1 is the regression coefficient, x is the independent variable, and $\hat{y}$ is the predicted value of the dependent variable.

For the regression line where the regression parameters b_0 and b_1 are defined, the properties are given as:

- The line reduces the sum of squared differences between observed values and predicted values.
- The regression line passes through the mean of X and Y variable values
- The regression constant (b_0) is equal to y-intercept the linear regression
- The regression coefficient (b_1) is the slope of the regression line which is equal to the average change in the dependent variable (Y) for a unit change in the independent variable (X).

Regression Coefficient: In the linear regression line, we have seen the equation is given by;

$y = \beta o + \beta 1x$ where, βo is a constant and β1 is the regression coefficient. Now, the formula to find the value of the regression coefficients are $\beta o = b_0 = \bar{y} - b_1\bar{x}$; $\beta 1 = b_1 = \Sigma [(x_i - \bar{x})(y_i - \bar{y})] / \Sigma [(x_i - \bar{x})^2]$ where x_i and y_i are the observed data sets and $\bar{x}$ and $\bar{y}$ are the mean values. The analysis of Simple linear regression and Multiple linear regression can be done through SPSS:

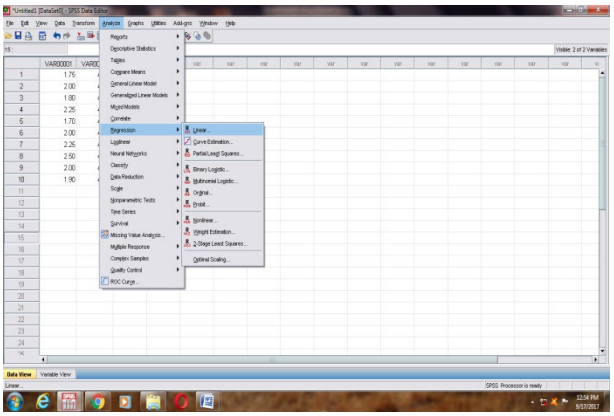
(i) REGRESSION ANALYSIS (LINEAR)

Example: Estimate the linear regression relating the influence of hen weights (Y) on feed intake (X) in a year.

Y	1.75	2.00	1.80	2.25	1.70	2.00	2.25	2.50	2.00	1.90
X	42.0	44.0	43.0	46.0	43.5	44.0	46.0	48.0	44.0	43.5

SPSS Commands (Step by Step):

1. Go to Analyze: Click on Regression and then Linear: Shift Y (hen weight) to dependent and X (feed intake) to independent list: Click OK to get the output.



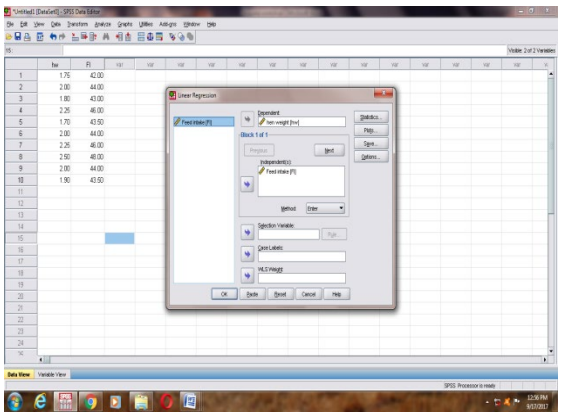


Fig. 1: Selection of command (Regression linear) from the menu. Fig. 2: Selection of input variable dependent and independent.

Output:

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.959[a]	.919	.909	.07643

ANOVA						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.529	1	.529	90.472	.000[a]
	Residual	.047	8	.006		
	Total	.575	9			

Coefficients						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-4.096	.643		-6.371	.000
	Feed intake	.138	.014	.959	9.512	.000

Interpretation: In case of simple linear regression the F value of ANOVA is 90.472 (p-value=0.000), which indicates the model is significant and the independent variable can be used to study the variable Y i.e the body weight of hen. Further, R^2 = 0.92 i.e. 92 percent variation of the Y can be explained by the variable X i.e Feed intake. The regression equation is Y= -4.096**+ 0.138X**, which implies that if X increases by one unit then Y i.e. the body weight of hen will be increased by 0.138 unit from the average level of Y.

(ii) Regression (Multiple)

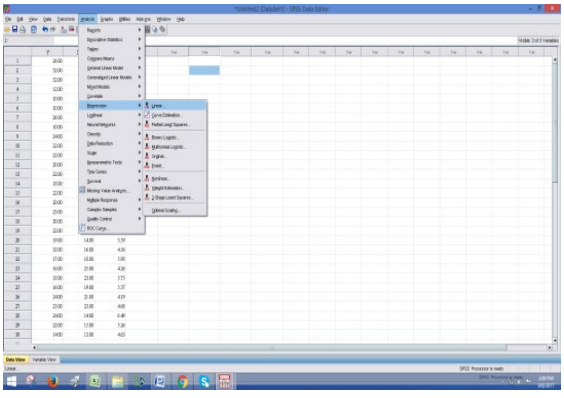
Example: Following is the data on keeping quality of milk samples(y), Methylene Blue Reduction time(x_1) and logarithm of standard plate count (x_2) obtained in a study of bacteriological quality of milk.

y	26.0	32.0	32.0	12.0	10.0	10.0	26.0	10.0	24.0	22.0	22.0	20.0	22.0	18.0	22.0
x_1	16.0	16.0	24.0	2.0	6.0	2.0	22.0	3.0	13.0	2.0	4.0	14.0	4.0	22.0	16.0
x_2	5.3	4.2	5.5	4.6	5.1	5.3	4.8	5.0	5.2	7.4	6.5	6.9	4.7	4.2	6.2
y	20.0	23.0	30.0	22.0	19.0	18.0	17.0	16.0	18.0	16.0	24.0	23.0	24.0	22.0	14.0
x_1	15.0	12.0	13.0	15.0	14.0	16.0	18.0	25.0	23.0	19.0	21.0	23.0	14.0	15.0	13.0
x_2	5.3	6.0	6.2	7.5	3.6	4.6	6.0	4.3	5.8	5.6	4.2	4.7	6.5	5.3	4.6

Fit the **Multiple Regression Analysis** y versus x_1 and x_2.

SPSS Commands are same as for Simple Linear regression:

Analyze > Regression > Linear > Y (move to Dependent variable box), x_1 & x_2 (move to Independent variable box) and OK.



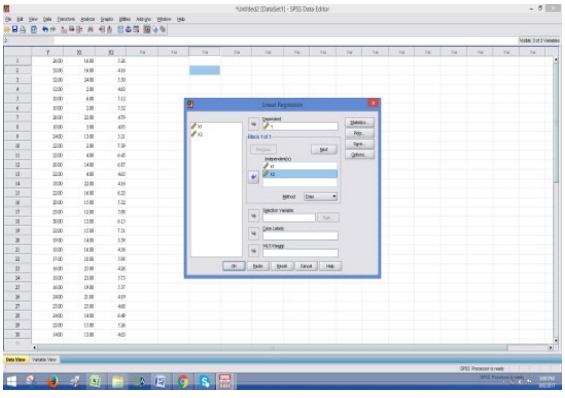


Fig. 3 Selection of the command (Regression linear) from the menu Fig.4 Selection of the input variables Y and X_1 & X_2 (independent)

Output:

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.471[a]	.222	.164	5.36252

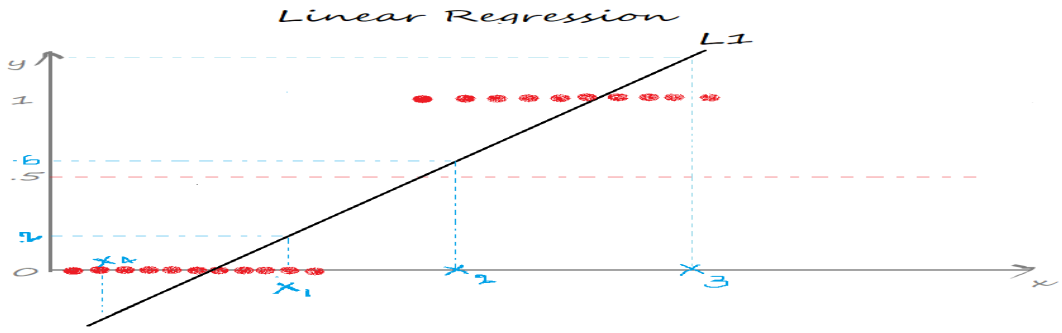
ANOVA						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	221.038	2	110.519	3.843	.034[a]
	Residual	776.429	27	28.757		
	Total	997.467	29			

Coefficients						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	6.909[ns]	6.467		1.068	.295
	X1	0.381*	.145	.459	2.630	.014
	X2	1.530[ns]	1.047	.255	1.461	.155

Interpretation: In case of multiple linear regression, the F value of ANOVA is 3.843 (p-value=0.034), which indicates the model is significant and the independent variables(X_1 & X_2) can be used to study the Y. Further, $R^2 = 0.222$ i.e. 22.2 percent variation of the Y can be explained by the variables X_1 & X_2. Further, the regression equation is $Y = 6.909^{ns} + 0.381X_1^{*} + 1.530X_2^{ns}$ .In this equation, the value $b_{yx1.x2} = 0.381$ means that on the average, the

quality of milk samples increases by 0.381 units if the Methylene blue reduction time is increased by one unit, eliminating the linear effect of the standard plate count.

Use of Linear Regression for classification: OLS(Ordinary Least Squares) is the algorithm to find the right coefficients for the minimum sum of squared errors. Now, let us try if we can use linear regression to solve a binary class classification problem. Assume we have a dataset that is linearly separable and has the output that is discrete in two classes (0, 1).



In Linear regression, we draw a straight line(the best fit line) L1 such that the sum of distances of all the data points to the line is minimal. The equation of the line L1 is y=mx+c, where m is the slope and c is the y-intercept. We define a threshold T = 0.5, above which the output belongs to class 1 and class 0 otherwise. y = mx+c , Threshold T = 0.5 ; y = $\begin{cases} 1, & mx + c \geq 0.5 \\ 0, & mx + c < 0.5 \end{cases}$

Case 1: the predicted value for x1 is ≈0.2 which is less than the threshold, so x1 belongs to class 0.

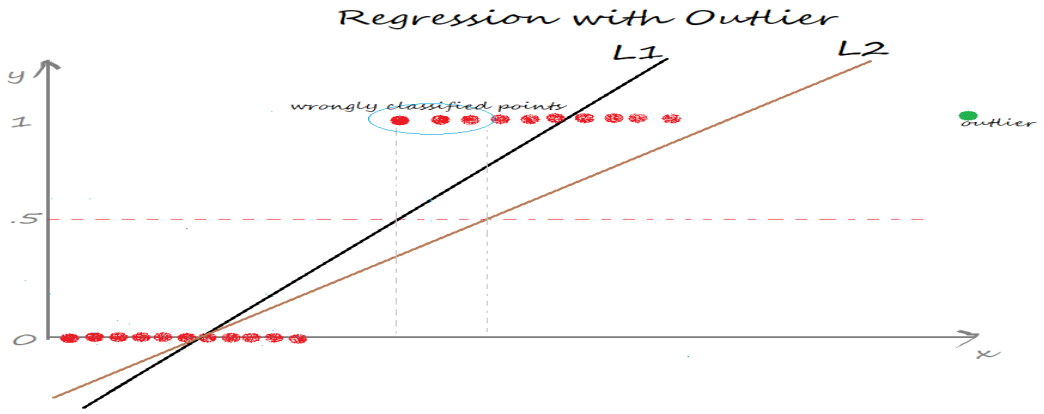
Case 2: the predicted value for the point x2 is ≈0.6 which is greater than the threshold, so x2 belongs to class 1.

Case 3: the predicted value for the point x3 is beyond 1.

Case 4: the predicted value for the point x4 is below 0.

The predicted values for the points x3, x4 exceed the range (0,1) which doesn't make sense because the probability values always lie between 0 and 1. And our output can have only two values either 0 or 1. Hence, this is a problem with the linear regression model.

Now, introduce an outlier and see what happens. The regression line gets deviated to keep the distance of all the data points to the line to be minimal.



L2 is the new best-fit line after the addition of an outlier. But the problem is, if we closely observe, some of the data points are wrongly classified. Certainly, it increases the error term. This again is a problem with the linear regression model. The two limitations of using a linear regression model for classification problems are:

- the predicted value may exceed the range (0,1)
- error rate increases if the data has outliers

There definitely is a need for Logistic regression here. The logistic regression equation is quite similar to the linear regression model. Logistic Regression can be used to model the probability of a certain class or event. It is used when the data is linearly separable and the outcome is binary or dichotomous in nature. That means Logistic regression is usually used for Binary classification problems.

Binary Classification refers to predicting the output variable that is discrete in two classes. A few examples of Binary classification are Yes/No, Pass/Fail, Win/Lose, Cancerous/Non-cancerous, etc. There are two types of logistic regression:

- **Simple Logistic Regression:** a single independent is used to predict the output
- **Multiple logistic regression:** multiple independent variables are used to predict the output

Although it is said Logistic regression is used for Binary Classification, but it can be extended to solve multiclass classification problems.

a) Multinomial Logistic Regression: The output variable is discrete in three or more classes with no natural ordering.

i. Food texture: Crunchy, Mushy, Crispy
ii. Hair colour: Blonde, Brown, Brunette, Red

b) Ordered Logistic Regression: Aka Ordinal regression model. The output variable is discrete in three or more classes with the ordering of the levels.

i. Customer Rating: extremely dislike, dislike, neutral, like, extremely like.
ii. Income level: low income, middle income, high income

Mathematically: Consider we have a model with one predictor "x" and one Bernoulli response variable "ŷ" and p is the probability of ŷ=1. The linear equation can be written as:

$p = b_0+b_1x$ -------- (1)

The right-hand side of the equation (b_0+b_1x) is a linear equation and can hold values that exceed the range (0,1). But we know probability will always be in the range of (0,1). To overcome that, we predict odds instead of probability. Odds are the ratio of the probability of an event occurring to the probability of an event not occurring i.e; Odds = p/(1-p). The equation 1 can be re-written as:

$p/(1-p) = b_0+b_1x$ --------(2)

Odds can only be a positive value, so to tackle the negative numbers, we predict the **logarithm of odds**.

Log of odds = ln(p/(1-p))

The equation 2 can be re-written as:

$\ln(p/(1-p)) = b_0+b_1x$ --------(3)

To recover p from equation 3, we apply exponential on both sides.

$\exp(\ln(p/(1-p))) = \exp(b_0+b_1x)$

$e^{\ln(p/(1-p))} = e^{(b0+b1x)}$

From the inverse rule of logarithms, algebraic manipulations and dividing numerator and denominator by $e^{(b0+b1x),}$ we get the equation for logistic model with one predictor as

$p = [e^{(b0+b1x)} / 1+ e^{(b0+b1x)}]$ or $[1 / (1 + e^{-(b0+b1x)})]$

Similarly, the equation for a logistic model with 'n' predictors is as below:

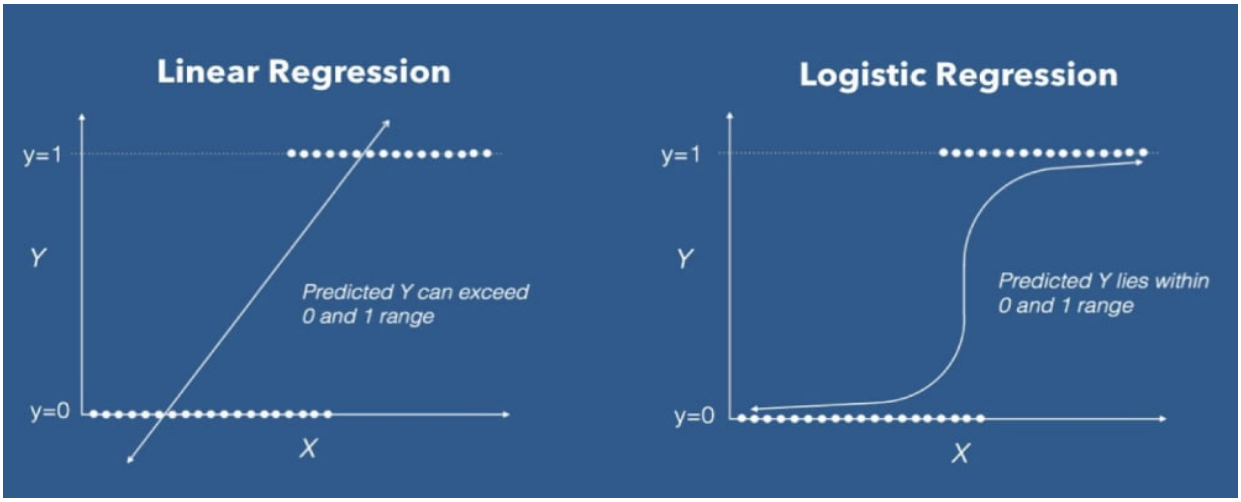
$p = 1/ (1 + e^{-(b0+b1x1+b2x2+b3x3+----+bn\ xn)})$

The right side part looks like the **sigmoid function**. It helps to squeeze the output to be in the range between 0 and 1. The sigmoid function is useful to map any predicted values of probabilities into another value between 0 and 1. In this chapter, we started with a linear equation and ended up with a logistic regression model with the help of a sigmoid function.

Linear model: $\hat{y} = \beta_0+\beta_1x$

Sigmoid function: $\varphi(z) = 1/(1+e^{-z})$

Logistic regression model: $\hat{y} = \varphi(b_0+b_1x) = 1/(1+e^{-(b0+b1x)})$



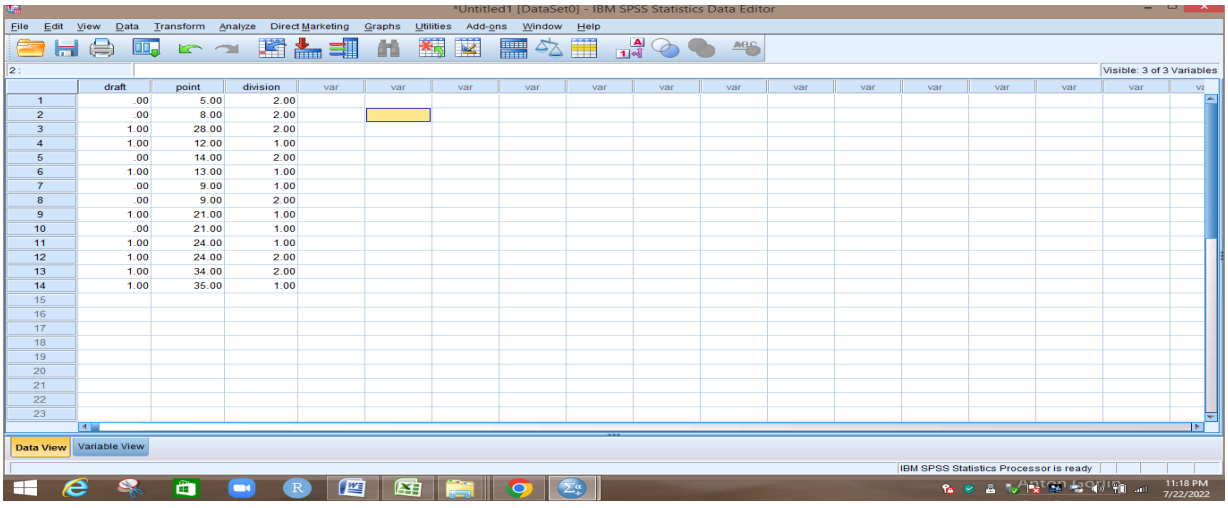
Example: Logistic Regression in SPSS

Use the following steps to perform logistic regression in SPSS for a dataset that shows whether or not college basketball players got drafted into the NBA (draft: 0 = no, 1 = yes) based on their average points per game and division level.

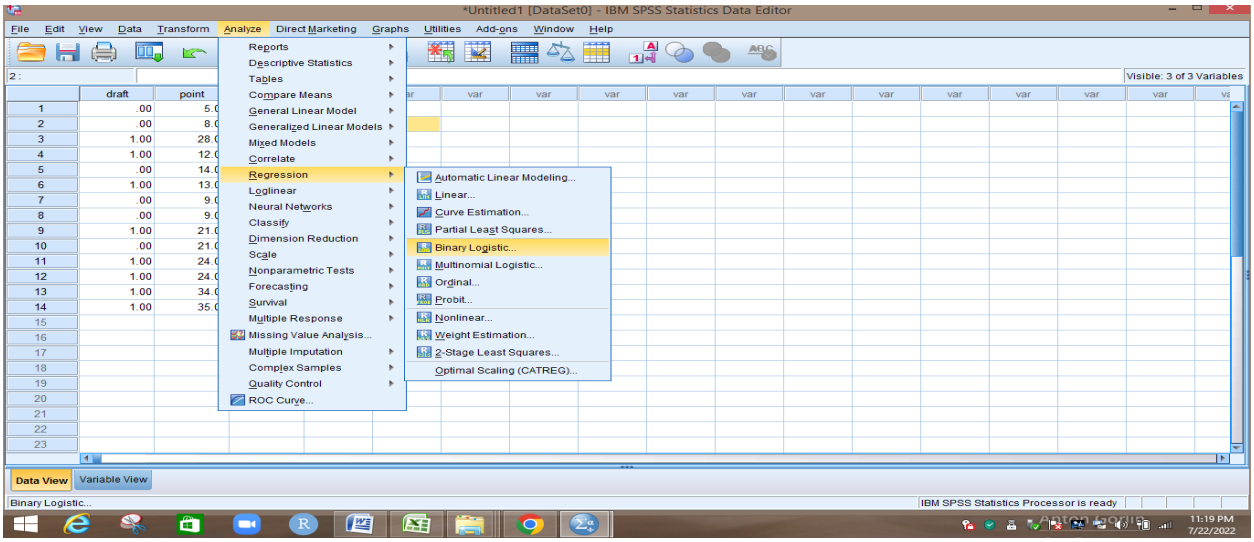
Draft	0	0	1	1	0	1	0	0	1	0	1	1	1	1
Point	5	8	28	12	14	13	9	9	21	21	24	24	34	35
Division	2	2	2	1	2	1	1	2	1	1	1	2	2	1

Steps:

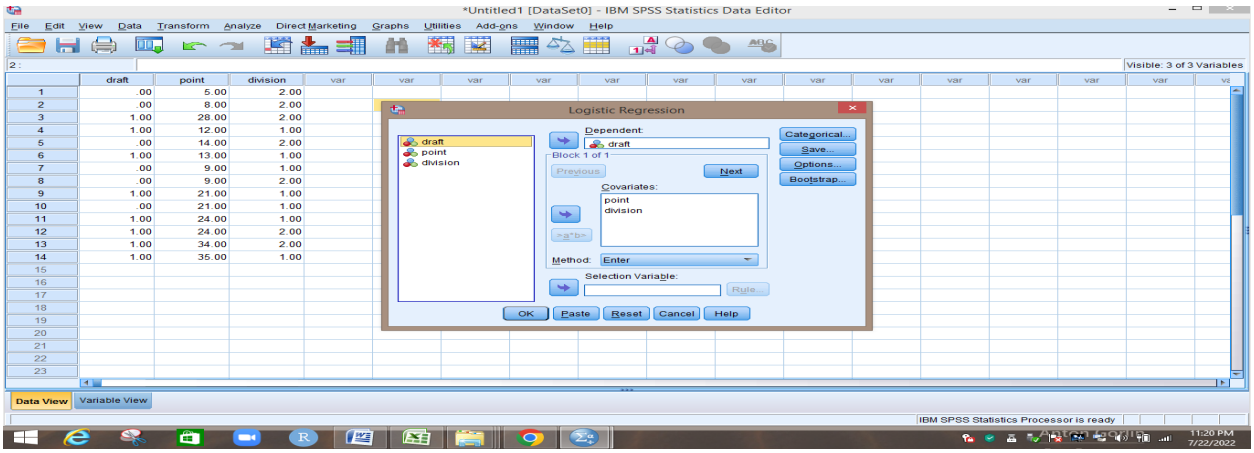
1) Enter the data in data view box of SPSS window.



2) Click on Analyse, then Regression, then Binary Logistic Regression.



The following dialog box will appear. Move draft in the Dependent box and the Independent variables Point and Division to Covariate box.



Click on Ok, you will get an output.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	8.256[a]	.540	.725

Classification Table[a]

Observed			Predicted		
			draft		Percentage Correct
			.00	1.00	
Step 1	draft	.00	5	1	83.3
		1.00	1	7	87.5
	Overall Percentage				85.7

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	point	.277	.134	4.243	1	.039	1.319
	division	-1.082	1.843	.345	1	.557	.339
	Constant	-3.152	3.408	.855	1	.355	.043

a. Variable(s) entered on step 1: point, division.

Interpretation: Logistic regression was performed to determine how points per game and division level affect a basketball player's probability of getting drafted. A total of 14 players were used in the analysis.

Model Summary: The most useful metric in this table is the Nagelkerke R Square, which tells us the percentage of the variation in the response variable that can be explained by the predictor variables. In this case, points and division are able to explain 72.5**%** of the variability in draft.

Classification Table: The most useful metric in this table is the Overall Percentage, which tells us the percentage of observations that the model was able to classify correctly. In this case, the logistic regression model was able to correctly predict the draft result of **85.7%** of players.

Variables in the Equation: Last table provides us with several useful metrics, including:

Wald: The Wald test statistic for each predictor variable, which is used to determine whether or not each predictor variable is statistically significant.

Sig: The p-value that corresponds to the Wald test statistic for each predictor variable. We see that the p-value for **points** is .039 and the p-value for **division** is .557.

Exp(B): The odds ratio for each predictor variable. This tells us the change in the odds of a player getting drafted associated with a one unit increase in a given predictor variable. For example, the odds of a player in division 2 getting drafted are just .339 of

the odds of a player in division 1 getting drafted. Similarly, each additional unit increase in points per game is associated with an increase of 1.319 in the odds of a player getting drafted.

We can then use the coefficients (the values in the column labeled B) to predict the probability that a given player will get drafted, using the following formula:

$$\text{Probability} = e^{-3.152 + .277(\text{points}) - 1.082(\text{division})} / (1+e^{-3.152 + .277(\text{points}) - 1.082(\text{division})})$$

For example, the probability that a player who averages 20 points per game and plays in division 1 gets drafted can be calculated as:

$$\text{Probability} = e^{-3.152 + .277(20) - 1.082(1)} / (1+e^{-3.152 + .277(20) - 1.082(1)}) = \mathbf{0.787}.$$

Since this probability is greater than 0.5, we would predict that this player would get drafted.

References:

Ajai S. Gaur & Sanjaya S. Gaur 2009. Statistical Methods for Practice and Research: a guide to Data analysis Using SPSS. Sage Publications Inc.

L. S. Aiken and S.G. West 1991). Multiple Regression: testing and Interpreting Interactions. Newbury Park, CA: Sage Publications

Berkson, J., (1944). Application Of The Logistic Function To Bio-assay. *Journal of the American Statistical Association*, **39** (227): 357–65.

Kalbfleisch, J. D., Prentice, R. L., (1980). The Statistical Analysis Of Failure Time Data. John Wiley, New York.

Pryanishnikov, I., Zigova, K., (2003). Multinomial Logit Models For The Austrian Labor Market. *Austrian Journal Of Statistics*, **32** (4): 267–282.

V.N. Amble. 1975. Statistical Methods in Animal Sciences. Indian Society of Agricultural Statistics, New Delhi.

CHAPTER 5

MULTICOLLINEARITY, HETEROSCEDASTICITY AND AUTOCORRELATION

S. E. H. RIZVI, MANISH KR. SHARMA & M.I.J. BHAT

Division of Statistics & Computer Science, Faculty of Basic Sciences, SKUAST-Jammu

A model is defined as the formal expression of the relationship that exists in the real world in mathematical terms and consequently allows us not only to explain observable facts but also to predict possibly unobserved events. Models are handy tools because they parallel the way we human think. Modelling can provide researchers a tool to make sound recommendations, to aid the conceptualization and sometimes to predict the consequences of an action that would be expensive, difficult or destructive to do with the real world. A model should provide information that is sufficiently precise and comprehensive to execute the intended purpose with its simplicity, is easily understood and helpful for drawing inferences. The best model is the one whose estimated prediction error is least. Mathematics provides us with ideas and tools that enable us to describe and understand the real-world studies in the construction, development and application of models. According to Draper and Smith (1998) model evaluation should try to reveal any errors and deficiencies in the model, in part, by establishing:

- Whether the equations used adequately represent the processes involved?
- If the equations have been combined correctly in the model?
- Whether the numerical constants obtained in fitting the model are the "best" estimates?
- Whether the model provides realistic predictions throughout the likely range of Application?
- If the model satisfies specified accuracy requirements?
- How sensitive model predictions are to errors in estimated coefficients and input Variables?

Regression analysis : Regression analysis includes any techniques for modelling and analyzing several variables, when the focus is on the relationship between a dependent variable (also called endogenous variable, predictand, explained variable or regressand) and one or more independent variable (also called exogenous variable, predictor explanatory variable or regresser). Most commonly, regression analysis estimates the conditional expectation of the dependent variable given the independent variable – that is, the average

value of the dependent variable when the independent variables are held fixed. Regression procedure can be classified according to number of variables involved in the form of functional relationship between the dependent variables and the independent variables. The procedure is simple if only two variables one dependent and one independent are involved and multiple otherwise.

Simple linear regression model is expressed the form: $Y = \alpha + \beta X + \varepsilon$

Multiple linear regression is expressed the form: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$

Where, Y indicates the dependent variable and X_i's denote p independent variables and ε is a stochastic disturbance or simply error term. Based on given observations, say n, these models are fitted through ordinary least square (OLS) method which provides the estimates of the parameters involved in the model.

Assumptions of Linear Regression Model:

In order to obtain the parameter certain assumptions are made about the behaviour of disturbance term in the regression model. For simple linear regression the basic assumptions are:

1. ε_i's are normally distributed
2. Zero mean i.e. $E(\varepsilon_i) = 0$, for every i
3. Constant variance or homoscedasticity i.e. $E(\varepsilon_i^2) = \sigma^2$, for every i
4. Non- auto correlation i.e. $E(\varepsilon_i \varepsilon_j) = 0$, for $i \neq j$
5. Each explanatory variable is non-stochastic with values fixed in repeated sampling.

In case of multiple linear regression model, two more points in addition to the assumption of simple linear regression model are to be satisfied which are given below :

6. The number of observations exceeds the number of coefficients to be estimated. In other words the rank of the matrix of observations on explanatory variables is the same as the number of explanatory variables.
7. No exact relationship should exist between any of the explanatory variables.

Violations of the Basic Assumptions of the Linear Regression Model & their Remedial Measures:

There are sufficient conditions for the least-squares estimator to possess desirable properties, in particular, the assumptions as given above imply that the parameter estimates will be unbiased, consistent and efficient in the class of linear unbiased estimators. Variation from the assumptions can sometimes be used as a measure of how far the model is from being useful. Many of these assumptions may be relaxed in more advanced treatments. Reports of

statistical analyses usually include analyses of tests on the sample data and methodology for the fit and usefulness of the model. There may be spatial trends and spatial autocorrelation in the variables that violates statistical assumptions of regression. Thus, we have to will look at what happens it these assumptions are violated in the real world and how they can be rectified? There are three major problems of violation of assumptions in regression models as follows:

1. Multicollinearity
2. Heteroscedasticity
3. Autocorrelation.

MULTICOLLINEARITY:

While dealing with multiple linear regression models, in many situations in practice it may be observed that the explanatory variables may not remain independent due to various reasons. In some cases, this relationship may be low and acceptable whereas in other situations it may be high. The situation where the explanatory variables are highly intercorrelated is referred to as **multicollinearity.** Multicollinearity does not also lessen the predictive or reliability of the regression model as a whole, it only affects the individual regressors. Note that, multicollinearity refers only to the linear relationships among the regressors, it does not rule out the nonlinear relationships among them. Broadly speaking there are two types of multicollinearity viz. **(i) Perfect multicollinearity (ii) High but imperfect multicollinearity.**

Detecting Multicollinearity: Exact or perfect multicollinearity is the case where one or more columns of X (the data matrix) are exact linear functions of one or more other columns of X otherwise it is of imperfect type of multicollinearity. Multicollinearity problems are, in general, very difficult to diagnose. Multicollinearity is a question of degree and not of kind. The detection of multicollinearity involves three aspects, namely, (i) Determining its presence (ii) Determining its severity, and (iii) Determining its form or location. Indicators of multicollinearity include:

a. A very high correlation among two or more variables which can be visualized through correlation matrix.
b. Very high correlations among two or more estimated coefficients.
c. High F- statistic, but low t-statistics for all coefficients.

d. Significant regression of one X variable on one or more other dependent variables.

Some multicollinearity is nearly always present, so formal tests are of little use. The relevant question is whether multicollinearity is serious enough to cause appreciable damage to the regression. No formal tests exist which adequately address this issue.

Variance Inflation Factor (VIF) is another way to indicate the presence of multicollinearity in the data. The VIF for the j^{th} explanatory variable is defined as VIF $=1/(1- R^2_j)$.This is the factor which is responsible for inflating the sampling variance. One or more large *VIF*s indicate the presence of multicollinearity in the data. In practice, usually, a *VIF* > 5 or 10 indicates that the associated regression coefficients are poorly estimated because of multicollinearity.

Dealing with Multicollinearity: Multicollinearity is among the most intractable of the problems that are badly faced by regression analysts. In severe cases, its treatment is rarely completely satisfactory. An important question arises about how to diagnose the presence of multicollinearity in the data on the basis of given sample information? Several diagnostic measures are available, and each of them is based on a particular approach. Various techniques have been proposed to deal with the problems resulting from the presence of multicollinearity in the data. Some are as follows:

1. Obtain more data: The harmful multicollinearity arises essentially because the rank of $X'X$ falls below p and $X'X$ is close to zero. Additional data may help in reducing the sampling variance of the estimates. The data need to be collected such that it helps in breaking up the multicollinearity in the data. It is always not possible to collect additional data for various reasons as follows.

2. Drop some variables that are collinear: In some studies researchers try to remedy multicollinearity by omitting one of more of the affected variables form the regression equation.

If possible, identify the variables which seem to cause multicollinearity. The process of omitting the variables may be carried out on the basis of some kind of ordering of explanatory variables, e.g., those variables can be deleted first which have smaller value of t -ratio. In such cases, one can get the estimators of the parameters of interest which have smaller mean squared errors than the variance of OLS estimator by dropping some variables.

3. Use of a priori information:

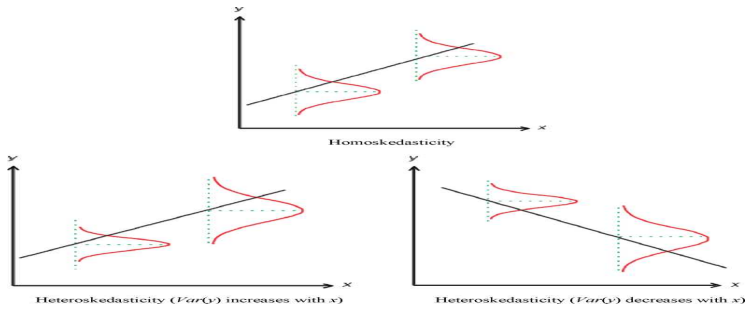
One may search for some relevant prior information about the regression coefficients. This may lead to the specification of estimates of some coefficients. The more general situation includes the specification of some exact linear restrictions and stochastic linear restrictions.

In addition, there are some other approaches like use of principal component and ridge regression.

HETEROSCEDASTICITY: In the regression model one of the assumptions states is regarding constant variance or homoscedasticity i.e. $E(\varepsilon_i^2) = \sigma^2$, for every i .In many situations, this assumption may be violated, and the variances may not remain the same. The disturbances whose variances are not constant across the observations are called **heteroscedastic disturbance.**

Detecting Heteroscedasticity: There are a number of procedures to test for the presence of heteroscedasticity. Residual plots as diagnostic tool for the presence of heteroscedasticity. Heteroskedasticity is suggested by "flaring" residual plots when the residual are graphed against Y, some of the X's or some other variable Z not even included in the regression.

Graphically, the following pictures depict homoskedasticity and heteroskedasticity.



Possible reasons for heteroscedasticity: There are various reasons due to which the heteroscedasticity is introduced in the data. Some of them are as follows:

1. The nature of the phenomenon under study may have an increasing or decreasing trend. For example, the variation in consumption pattern on food increases as income increases.

2. Sometimes the observations are in the form of averages, and this introduces the heteroscedasticity in the model. For example, it is easier to collect data on the expenditure on clothes for the whole family rather than on a particular family member.

Dealing with heteroscedasticity: The presence of heteroscedasticity affects the estimation and test of hypothesis. The heteroskedasticity can enter into the data due to various reasons. The tests for heteroscedasticity assume a specific nature of heteroscedasticity. Various tests are available some of them are listed below:

1. Goldfeld Quandt test
2. Glesjer test
3. Test based on Spearman's rank correlation coefficient
4. Park test

Goldfeld Quandt test: This test developed by Godfeld and Quandt divide the date set into two groups. This division may be based on discrete or qualitative difference between subsets of the observations. Alternatively the division may be based on rank ordering by one of the X's or some other variable z.

Glesjer test: Another test, developed by Glejser screens for heteroscedasticity by estimating regressions of the various form absolute values of residuals on X. The usual t and F-tests can be used to determine the significance of the regressions, thus giving a formal test for heteroscedasticity.

Spearman's rank correlation test: It d_t denotes the difference in the ranks assigned to two different characteristics of the i^{th} object or phenomenon and n is the number of objects or phenomenon ranked, then the well known Spearman's rank correlation coefficient is used to test the present heteroscedasticity

AUTOCORRELATION

One of the basic assumptions in the linear regression model is that the random error components or disturbances are identically and independently distributed. When the assumption of non- auto correlation i.e. $E(\varepsilon_i\varepsilon_j) = 0$, for $i \neq j$ is violated, then such problem is termed as the problem of autocorrelation.

Detecting Autocorrelation: There are number of methods to detect autocorrelation (also called serial correlation) problems. Will describe several of them.

1. Residual plots: Residual plots are one way to screen for serial correlation problems. Presence of serial correlation is judged on the basis of pattern of graph.

2. Draper and Smith's z-test for runs of + and - residuals:Draper and Smith suggest a nonparametric z-test of the probability that the observed number of "runs" of + and – residuals occur from an uncorrelated random process

Tests for autocorrelation:

Durbin Watson test:

The Durbin-Watson *(D-W)* test is used for testing the hypothesis of lack of first-order autocorrelation in the disturbance term. The null hypothesis is

$H_0 : \rho = 0$ vs. $H_1 : \rho \neq 0$

The *D-W* test statistic is

$$\boldsymbol{d} = \frac{\sum_1^n (\boldsymbol{e}_t - \boldsymbol{e}_{t-1})^2}{\sum_1^n (\boldsymbol{e}_t)^2}$$

As -l<r<l, so

if -l<r<0, then *2<d<4* and

if 0<r<l, then *0<d<2*. So *d* lies between 0 and 4.

Since *e* depends on *X,* so for different data sets, different values of *d* are obtained. So the sampling distribution of *d* depends on *X.* Consequently, exact critical values of *d* cannot be tabulated owing to their dependence on *X.* Durbin and Watson, therefore, obtained lower and upper values of *d* statistics..The test procedure is as follows:

$H_0 : \rho = 0$			
Nature of H_1	Reject H_0 when	Retain H_0 when	The test is inconclusive when
$H_1 : \rho > 0$	$d < d_L$	$d > d_U$	$d_L < d < d_U$
$H_1 : \rho < 0$	$d > (4\_d_L)$	$d < (4-d_u)$	$(4-d_U) < d < (4-d_L)$
$H_1 : \rho \neq 0$	$d < d_L$ or $d > (4\text{-}t/L)$	$d_v < d < (4-d_U)$	$d_L < d < d_U$ or $(4-d_U) < d < (4-d_L)$
Values of d_L and d_U are obtained from tables.			

Limitations of *D-W* test

If *d* falls in the inconclusive zone, then no conclusive inference can be drawn. This zone becomes fairly larger for low degrees of freedom. This test gives a satisfactory solution when values of x_t's change slowly, e.g., price, expenditure etc. The *D-W* test is not applicable when the intercept term is absent in the model. In such a case, one can use another critical value, say d_u in place of d_L .

Source of autocorrelation

Some of the possible reasons for the introduction of autocorrelation in the data are as follows:

1. Carryover of effect, at least in part, is an important source of autocorrelation. For example, the

monthly data on expenditure on household is influenced by the expenditure of preceding month. The autocorrelation is present in cross-section data as well as time-series data. In the cross-section data, the neighbouring units tend to be similar with respect to the characteristic under study. In time-series data, time is the factor that produces autocorrelation. Whenever some ordering of sampling units is present, the autocorrelation may arise.

2. Another source of autocorrelation is the effect of deletion of some variables. In regression modeling, it is not possible to include all the variables in the model. There can be various reasons for this, e.g., some variable may be qualitative, sometimes direct observations may not be available on the variable etc. The joint effect of such deleted variables gives rise to autocorrelation in the data.

3. The misspecification of the form of relationship can also introduce autocorrelation in the data. It is assumed that the form of relationship between study and explanatory variables is linear. If there are log or exponential terms present in the model so that the linearity of the model is questionable, then this also gives rise to autocorrelation in the data.

4. The difference between the observed and true values of the variable is called measurement error or errors–in-variable. The presence of measurement errors on the dependent variable may also introduce the autocorrelation in the data.

CHAPTER 6

NON PARAMETRCIC TESTS

V.K. Shivgotra
Department of Statistics, University of Jammu

Introduction

A **parametric** statistical test is one that makes assumptions about the parameters (defining properties) of the population distribution(s) from which one's data are drawn.

A **non-parametric** test is one that makes no such assumptions. In this strict sense, "non-parametric" is essentially a null category, since virtually all statistical tests assume one thing or another about the properties of the source population(s).

Which is more powerful?

Non-parametric statistical procedures are less powerful because they use less information in their calculation. For example, a parametric correlation uses information about the mean and deviation from the mean while a non-parametric correlation will use only the ordinal position of pairs of scores.

Parametric Assumptions

- The observations must be independent
- The observations must be drawn from normally distributed populations
- These populations must have the same variances
- The means of these normal and homoscedastic populations must be linear combinations of effects due to columns and/or rows

Nonparametric Assumptions

Certain assumptions are associated with most nonparametric statistical tests, but these are fewer and weaker than those of parametric tests. Nonparametric tests do not require that samples come from populations with normal distributions or have any other particular distributions. Consequently, nonparametric tests are called distribution-free tests.

Advantages of Nonparametric Methods

1. Nonparametric methods can be applied to a wide variety of situations because they do not have the more rigid requirements of the corresponding parametric methods. In particular, nonparametric methods do not require normally distributed populations.

2. Unlike parametric methods, nonparametric methods can often be applied to categorical data, such as the genders of survey respondents.

3. Nonparametric methods usually involve simpler computations than the corresponding parametric methods and are therefore easier to understand and apply.

Disadvantages of Nonparametric Methods

1. Nonparametric methods tend to waste information because exact numerical data are often reduced to a qualitative form.

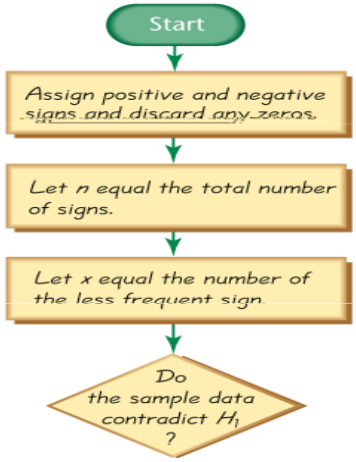
2. Nonparametric tests are not as efficient as parametric tests, so with a nonparametric test we generally need stronger evidence (such as a larger sample or greater differences) before we reject a null hypothesis.

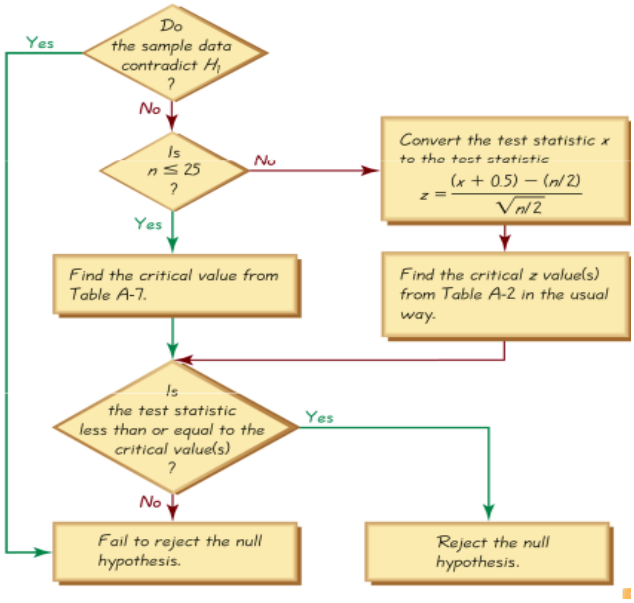
Sign Test: The main objective is to understand the sign test procedure, which involves converting data values to plus and minus signs, then testing for disproportionately more of either sign.

The sign test is a nonparametric (distribution free) test that uses plus and minus signs to test different claims, including:

1) Claims involving matched pairs of sample data;
2) Claims involving nominal data;
3) Claims about the median of a single population.

The basic idea underlying the sign test is to analyze the frequencies of the plus and minus signs to determine whether they are significantly different.





Requirements

1. The sample data have been randomly selected.

2. There is no requirement that the sample data come from a population with a particular distribution, as a normal distribution.

Notation for Sign Test

x = the number of times the less frequent sign occurs

n = the total number of positive and negative signs combined

Test Statistic

For $n \leq 25$: x (the number of times the less frequent sign occurs)

$$\text{For } n > 25;\ z = \frac{(x+0.5)-\frac{n}{2}}{\frac{\sqrt{n}}{2}}$$

Claims Involving Matched Pairs

When using the sign test with data that are matched pairs, we convert the raw data to plus and minus signs as follows:

1. Subtract each value of the second variable from the corresponding value of the first variable.

2. Record only the sign of the difference found in step 1.

Exclude ties: that is, any matched pairs in which both values are equal.

Example:

Yields of Corn from Different Seeds Use the data in Table 13-3 with a 0.05 significance level to test the claim that there is no difference between the yields from the regular and kiln -dried seed.

Table 13-3 Yields of Corn from Different Seeds

Regular	1903	1935	1910	2496	2108	1961	2060	1444	1612	1316	1511
Kiln dried	2009	1915	2011	2463	2180	1925	2122	1482	1542	1443	1535
Sign of difference	−	+	−	+	−	+	−	−	+	−	−

Use the data in Table 13-3 with a 0.05 significance level to test the claim that there is no difference between the yields from the regular and kiln -dried seed.

H0: The median of the differences is equal to 0.

H1: The median of the differences is not equal to 0. $\alpha = 0.05$ yields from the regular and kiln -dried seed. x = minimum (7, 4) = 4 (From Table 13-3, there are 7 negative signs and 4 positive signs.)
Critical value = 1 (From Table where n = 11 and $\alpha = 0.05$)

With a test statistic of x = 4 and a critical value of 1, we fail to reject the null hypothesis of no difference. There is not sufficient evidence to warrant rejection of the claim that the median of the differences is equal to 0.

Wilcoxon signed-rank test:

- The Wilcoxon signed-rank test is the nonparametric test equivalent to the dependent t-test.
- As the Wilcoxon signed-ranks test does **not assume normality** in the data, it can be used when this assumption has been violated and the use of the dependent t-test is inappropriate.
- It is used to compare two sets of scores that come from the same participants.
- This can occur when we wish to investigate any change in scores from one time point to another, or when individuals are subjected to more than one condition. *For example*, you could use a Wilcoxon signed-rank test to understand whether there was a difference in smokers' daily cigarette consumption before and after a 6 week hypnotherapy programme

(i.e., your dependent variable would be "daily cigarette consumption", and your two related groups would be the cigarette consumption values "before" and "after" the hypnotherapy programme).

Assumptions

- **A Wilcoxon signed-rank test is only appropriate where the following two assumptions are met:**
- **Assumption #1:** Your **dependent variable** should be measured at the **ordinal** or **interval/ratio level**. Examples of **ordinal variables** include Likert scales (e.g., a 7-point scale from strongly agree through to strongly disagree), amongst other ways of ranking categories (e.g., a 3-pont scale explaining how much a customer liked a product, ranging from "Not very much", to "It is OK", to "Yes, a lot"). Examples of **interval/ratio variables** include revision time (measured in hours), intelligence (measured using IQ score), exam performance (measured from 0 to 100), weight (measured in kg), and so forth.

- **Assumption #2:** Your independent variable should consist of two categorical, "related groups" or "matched pairs". "Related groups" indicates that the same subjects are present in both groups. The reason that it is possible to have the same subjects in each group is because each subject has been measured on two occasions on the same dependent variable. For example, you might have measured 10 individuals' performance in a spelling test (the dependent variable) before and after they underwent a new form of computerised teaching method to improve spelling. You would like to know if the computer training method to improve spelling. You would like to know if the computer training improved their spelling performance. The first related group consists of the subjects at the beginning (prior to) the computerised spelling training and the second related group consists of the same subjects, but now at the end of the computerised training. The Wilcoxon signed-rank test can also be used to compare different subjects, but this does not happen very often.

Example: A pain researcher is interested in finding methods to reduce lower back pain in individuals without having to use drugs. The researcher thinks that having acupuncture in the lower back might reduce back pain. To investigate this, the researcher recruits 25 participants to their study. At the beginning of the study, the researcher asks the participants to rate their back pain on a scale of 1 to 10, with 10 indicating the greatest level of pain. After 4 weeks of

twice weekly acupuncture the participants are asked again to indicate their level of back pain on a scale of 1 to 10, with 10 indicating the greatest level of pain. The researcher wishes to understand whether the participants' pain levels changed after they had undergone the acupuncture so they run a Wilcoxon signed-rank test.

Mann-Whitney U test: The Mann-Whitney U test is used to compare differences between two independentss groups when the dependent variable is either ordinal or interval/ratio, but not normally distributed. *For example*, you could use the Mann-Whitney U test to understand whether attitudes towards pay discrimination, where attitudes are measured on an ordinal scale, differ based on gender (i.e., your dependent variable would be "attitudes towards pay discrimination" and your independent variable would be "gender", which has two groups: "male" and "female"). *Alternately*, you could use the Mann-Whitney U test to understand whether salaries, measured using an interval scale, differed based on education level (i.e., your dependent variable would be "salary" and your independent variable would be "educational level", which has two groups: "high school" and "university").

The Mann-Whitney U test is the nonparametric alternative to the independent t-test.

Assumptions

When you choose to analyse your data using a Mann-Whitney U test, part of the process involves checking to make sure that the data you want to analyse can actually be analysed using a Mann-Whitney U test. You need to do this because it is only appropriate to use a Mann-Whitney U test if your data "passes" four assumptions that are required for a Mann-Whitney U test to give you a valid result.

Four assumptions:

Assumption #1: Your **dependent variable** should be measured at the **ordinal** or **interval/ratio level**. Examples of **ordinal variables** include Likert scales (e.g., a 7-point scale from strongly agree through to strongly disagree), amongst other ways of ranking categories (e.g., a 3-pont scale explaining how much a customer liked a product, ranging from "Not very much", to "It is OK", to "Yes, a lot"). Examples of **interval/ratio variables** include revision time (measured in hours), intelligence (measured using IQ score), exam performance (measured from 0 to 100), weight (measured in kg), and so forth.

Assumption #2: Your **independent variable** should consist of **two categorical**, **independent groups**. Example independent variables that meet this criterion include gender

(2 groups: male or female), employment status (2 groups: employed or unemployed), smoker (2 groups: yes or no), and so forth.

Assumption #3: You should have **independence of observations**, which means that there is no relationship between the observations in each group or between the groups themselves. For example, there must be different participants in each group with no participant being in more than one group. This is more of a study design issue than something you can test for, but it is an important assumption of the Mann-Whitney U test. If your study fails this assumption, you will need to use another statistical test instead of the Mann-Whitney U test (e.g., a Wilcoxon signed-rank test).

Assumption #4: A Mann-Whitney U test can be used when your two variables are **not normally distributed**. However, for a Mann-Whitney U test to be able to provide a valid result, **both distributions must be the same shape** (i.e., the distribution of scores for both categories of the independent variable must have the same shape).

You can check assumptions #4 using SPSS. Before doing this, you should make sure that your data meets assumptions #1, #2 and #3, although you don't need SPSS to do this. Just remember that if you do not check assumption #4, but your data ends up violating this assumption, the results you get when running a Mann-Whitney U test will not be valid.

Example:The following are the weight gains (in pound) of two random samples of young Indian fed on two different diets but otherwise kept under identical conditions:

Diet I:	16.3	10.1	10.7	13.5	14.9	11.8	14.3	10.2
	12.0	14.7	23.6	15.1	14.5	18.4	13.2	14.0
Diet II	21.3	23.8	15.4	19.6	12.0	13.9	18.8	19.2
	15.3	20.1	14.8	18.9	20.7	21.1	15.8	16.2

Use U test at 0.01 level of significance to test the null hypothesis that two population samples are identical against the alternative diet produces a greater gain in weight

The Kruskal-Wallis H-test : The Kruskal-Wallis H-test is a non-parametric statistical procedure for comparing more than two samples that are independent. The parametric equivalent to this test is the one-way analysis of variance (ANOVA). Yet ANOVA is used for normally distributed data, but Kruskal Wallis can be perform without the data being normally distributed.

The H-test is a generalization of the Mann-Whitney test, a test for knowing whether the two samples chosen are taken from the same population. The p value for both the Kruskal Wallis and the Mann- Whitney test are equal.

Assumptions

- **Assumption #1:** Your **dependent variable** should be measured at the **ordinal** or **continuous level** (i.e., **interval** or **ratio**). Examples of **ordinal variables** include Likert scales (e.g., a 7-point scale from "strongly agree" through to "strongly disagree"), amongst other ways of ranking categories (e.g., a 3-pont scale explaining how much a customer liked a product, ranging from "Not very much", to "It is OK", to "Yes, a lot"). Examples of **continuous variables** include revision time (measured in hours), intelligence (measured using IQ score), exam performance (measured from 0 to 100), weight (measured in kg), and so forth. You can learn more about ordinal and continuous variables in our article: Types of Variable.
- **Assumption #2:** Your **independent variable** should consist of **two or more categorical, independent groups**. Typically, a Kruskal-Wallis H test is used when you have **three or more** categorical, independent groups, but it can be used for just two groups (i.e., a Mann-Whitney U test is more commonly used for two groups). Example independent variables that meet this criterion include ethnicity (e.g., three groups: Caucasian, African American and Hispanic), physical activity level (e.g., four groups: sedentary, low, moderate and high), profession (e.g., five groups: surgeon, doctor, nurse, dentist, therapist), and so forth.
- **Assumption #3:** You should have **independence of observations**, which means that there is no relationship between the observations in each group or between the groups themselves. For example, there must be different participants in each group with no participant being in more than one group. This is more of a study design issue than something you can test for, but it is an important assumption of the Kruskal-Wallis H test. If your study fails this assumption, you will need to use another statistical test instead of the Kruskal-Wallis H test (e.g., a Friedman test). If you are unsure whether your study meets this assumption, you can use our Statistical Test Selector, which is part of our enhanced content.

- As the Kruskal-Wallis H test does not assume normality in the data and is much less sensitive to outliers, it can be used when these assumptions have been violated and the use of a one-way ANOVA is inappropriate. In addition, if your data is ordinal, a one-way ANOVA is inappropriate, but the Kruskal-Wallis H test is not. However, the Kruskal-Wallis H test does come with an additional data consideration, **Assumption #4**, which is discussed below:

- **Assumption #4:** In order to know how to interpret the results from a Kruskal-Wallis H test, you have to determine whether the **distributions** in each group (i.e., the distribution

of scores for each group of the independent variable) have the **same shape** (which also means the **same variability**).

Example: A random sample of three models of scooter were tested for the petrol mileage (no of kilometre per litre).Use KruskalWallies test to determine if the average mileage of the three models is same

Model-A	60	54	76	48	66	52	62	56
Model-B	62	58	52	48	70	79	86	90
Model-C	42	64	48	65	42	60	82	74

Friedman Test : The Friedman test is the non-parametric alternative to the one-way ANOVA with repeated measures. It is used to test for differences between groups when the dependent variable being measured is ordinal. It can also be used for continuous data that has violated the assumptions necessary to run the one-way ANOVA with repeated measures (e.g., data that has marked deviations from normality).

Assumptions

You need to do this because it is only appropriate to use a Friedman test if your data "passes" the following five assumptions:

Assumption #1:One group that is measured on **three or more different occasions**.

Assumption #2: Group is a random sample from the population.

Assumption #3: Your **dependent variable** should be measured at the **ordinal** or **interval/ratio level**. Examples of **ordinal variables** include Likert scales (e.g., a 7-point scale from strongly agree through to strongly disagree), amongst other ways of ranking categories (e.g., a 3-pont scale explaining how much a customer liked a product, ranging from "Not very much", to "It is OK", to "Yes, a lot"). Examples of **interval/ratio variables** include revision time (measured in hours), intelligence (measured using IQ score), exam performance (measured from 0 to 100), weight (measured in kg), and so forth. You can learn more about interval and ratio variables in our article.

Assumption #4: Samples do **NOT need to be normally distributed**.

The Friedman test procedure in SPSS will not test any of the assumptions that are required for this test. In most cases, this is because the assumptions are a methodological or study design issue, and not what SPSS is designed for. In the case of assessing the types of variable

you are using, SPSS will not provide you with any errors if you incorrectly label your variables as nominal

Example: A researcher wants to examine whether music has an effect on the perceived psychological effort required to perform an exercise session. The dependent variable is "perceived effort to perform exercise" and the independent variable is "music type", which consists of three categories: "no music", "classical music" and "dance music". To test whether music has an effect on the perceived psychological effort required to perform an exercise session, the researcher recruited 12 runners who each ran three times on a treadmill for 30 minutes. For consistency, the treadmill speed was the same for all three runs. In a random order, each subject ran: (a) listening to no music at all; (b) listening to classical music; and (c) listening to dance music. At the end of each run, subjects were asked to record how hard the running session felt on a scale of 1 to 10, with 1 being easy and 10 extremely hard. A Friedman test was then carried out to see if there were differences in perceived effort based on music type.

Runners	**Listening to no music**	**Listening to classical music**	**Listening to dance music**
1	8	8	7
2	9	6	8
3	6	8	6
4	4	9	4
5	9	8	9
6	4	7	8
7	7	9	4
8	8	8	7
9	9	6	8
10	9	6	4
11	7	8	4
12	3	9	9

CHAPTER 7

ANALYTICAL TOOLS USED IN PRODUCTION AND MARKETING OF AGRICULTURAL PRODUCE

Anil Bhat
Assistant Professor
Division of Agricultural Economics and ABM
Sher-e-Kashmir University of Agricultural Sciences and Technology of Jammu

Agricultural production is the use of cultivated plants or animals to produce products for sustaining or enhancing human life. Agricultural production has always involved the exploitation of resources such as soil, water, and energy. Increasing production to feed a growing world population while at the same time conserving resources for future generations has led to a search for 'sustainable' agricultural methods. Farm managers must take a long-term view when making decisions about which technologies to follow and what commodities to produce while still generating sufficient profits in the short run to earn a living. Farm managers must also be aware of possible trends in climatic conditions, and learn how to adapt their production methods accordingly (Edwards, 2014). ***Production*** is the process of combining various material inputs and immaterial inputs (plans, knowledge) in order to make something for consumption (output). Agricultural Production economics deals with the economic analysis of production of agricultural commodities for which various analytical procedures and tools used are explained below:

Costs and Return analysis: The information with respect to variable and fixed costs involved in the cultivation and production of agricultural crops, establishment of orchards as well as returns obtained thereby from the whole process falls under cost and return analysis. It is based on number of variables which needs to be measured and quantified.

Quantification of the variables: The various inputs used in the production of crops along with various costs and returns concepts are quantified as follows:

Human Labour: It included both hired and family labour. Human labour cost comprised of wages actually paid to the hired labour as also those paid to the labour obtained on contract for the whole year or part thereof and imputed value of labour put in by the family members working on the field. Existing casual labour wages for different operations are used to work out the total wage bill of labour employed per hectare of any crop.

Bullock labour: Hired bullock labour charges are considered for 8 hours a day, actually paid in the locality. Family bullock labour charges accounted equal to charges paid to the hired bullock labour.

Tractor charges: Tractor charges actually paid by the farmer are calculated by multiplying the area of the land with the market rate charged per unit of the land.

Manures and fertilizers: This item included the expenditure incurred on the purchase of chemical fertilizers and farmyard manure used for the production of fruits on the sample orchards. The farmyard manure used at the orchard was assessed at the prices prevailing in the study area. Similarly, the physical quantities of different fertilizers used were multiplied with the market price.

Plant protection: The various plant protection chemicals used per hectare for the crop will be assessed. Charges actually paid by the farmer are calculated by multiplying the quantity of the various pesticides, insecticides, fungicides and weedicides with the market rates charged per unit of each plant protection item used.

Seed/ Planting material: The market value of seeds/ planting material are considered.

Land revenue: The land revenue actually paid by the farmer to the government is considered.

Gross returns : Gross returns from the crop cultivated per hectare can be obtained by the value of the total produce harvested during the year. These are the post harvest market prices.

Interest on working capital : Interest on working capital can be calculated at the prevailing bank rate for short term loans.

Interest on fixed capital : Interest on fixed capital can be calculated at the prevailing bank rate for long term loans.

Production Function Analysis : In order to study the relationship between output and various inputs used, Cobb- Douglas production function is used. This function is used extensively in agricultural production function analysis. The functional form applies is given as under:

$$Y_t = \beta_0 \left(\prod_{i=1}^{n} X_i \beta i \right) u_t \qquad (i = 1,2,3, \ldots, n)$$

Where Y and X_i (i=1,2,3, ….., n) are the output and levels of inputs. The constant β_0 and β_i's (i-1,2,3, ….., n) represent the efficiency parameters and the production elasticities of the respective input variables for the given population at a particular period, t.

The fitted Cobb-Douglas production may be written with five input variables as:

$$Y = a_0 x_1^{b1} x_2^{b2} x_3^{b3} x_4^{b4} x_5^{b5}$$

On log transformation, the above function can be transformed to a linear form as:

Log y = log a_0 + b_1 log x_1 + b_2 log x_2 + b_3 log x_3 + b_4 log x_4 + b_5 log x_5

Or log y = $\mathbf{\log a_0 + b_i} \sum_{i=1}^{5} \log$ xi

Where

Y = Output in quintals as dependent variables

x_1 = Human labour in man days

x_2 = Manure and fertilizers (Rs)

x_3 = Expenditure on plant protection (Rs)

x_4 = Expenditure on irrigation (Rs)

x_5 = Expenditure on training and pruning (Rs)

a_0 = Constant

b's = Elasticities of production of respective resource categories

To examine the productivity of different inputs used in production of studied fruits, marginal value productivities of inputs were estimated at geometric mean levels of inputs. To calculate Marginal Value Productivity (MVP) of resource x_i, the following formula can be used.

$$MVP = \widehat{bi} \frac{GM(Y)}{GM(xi)}$$

Where,

MVP (x_i) = marginal value productivity of i_{th} resource

$\widehat{bi}$ = regression coefficient (estimated)

GM (Y) = geometric mean of output

GM (x_i) = geometric mean of inputs

For estimating the costs and returns

The cost and returns analysis can be worked out using CACP cost concepts like cost A_1, cost A_2, cost B_1, cost B_2, cost C_1, cost C_2 and cost C_3

Cost A_1 includes:

1. Wages of hired human labour
2. Value of hired bullock labour*
3. Value of owned bullock labour*
4. Value of owned machinery labour
5. Hired machinery charges
6. Value of seed (both farm produced and purchased)
7. Value of insecticides and pesticides

8. Value of manure (owned and purchased)
9. Value of fertilizer
10. Depreciation on implements and farm buildings
11. Irrigation charges
12. Land revenue, cesses and other taxes
13. Interest on working capital
14. Miscellaneous expenses

Cost A_2: Cost A_1 + rent paid for leased-in land

Cost B_1: Cost A_1 + interest on the value of owned fixed capital assets (excluding land).

Cost B_2: Cost B_1 + rental value of owned land + rent paid for leased in land

Cost C_1: Cost B1 + imputed value of family labour

Cost C_2: Cost B_2 + imputed value of family labour

Cost C_3: Cost C_2 + value of management input (10 % of Cost C_2)

Economic Viability : The economic viability is assessed using net present value (NPV), pay-back period, internal rate of return (IRR) and benefit–cost ratio (BCR).

Net present value : Net present value (NPV) of an investment is the discounted value of all cash inflows and cash outflow of the project during its life time. It can be computed as

$$NPV = \sum_{t=0}^{n} \{(B_t - C_t) / (1 + r)^t\}$$

Internal Rate of Return (IRR) : Internal rate of return is the rate of return at which the Net Present value of a stream of payments/ incomes is equal to zero.

$$IRR = \sum_{t=0}^{n} \{(B_t - C_t) / (1 + IRR)^t\} = 0$$

Benefit Cost Ratio (BCR) : The benefit cost ratio (BCR) of an investment is the ratio of the discounted value of all cash inflows to the discounted value of all cash outflows during the life of the project. It can be estimated as follows

$$BCR = \sum_{t=0}^{n} \{(B_t)/(1+r)^t / \sum_{t=0}^{n} [(C_t) / (1 + r)^t\}$$

Where,

B_t = gross returns in time t

C_t = variable cost in time t

r = rate of interest

t = time period (t = 0, 1, 2, …….., i, ……., 30)

Pay-back Period

The Pay-back period is defined as the length of time required to recover an initial investment through cash flows generated by the investment.

$$\text{Pay Back Period} = \frac{\text{Cost of investment}}{\text{Annual net cash flow}}$$

Methods of computation of Break-Even quantity

$$TR = TC$$

$$P \times Q = TFC + (AVC \times Q)$$

$$(P \times Q) - (AVC \times Q) = TFC$$

$$(P - AVC) \times Q = TFC$$

$$\mathbf{Q = TFC/ (P - AVC)}$$

Where, TR = Total revenue; TC = Total cost; P = Price of the product;
Q = Break-even output; TFC = Total fixed cost; and AVC =Average variable cost.

Value Added : It reflected the difference between price for which it sells its products and the cost incurred on the purchased inputs by a firm. This difference represented the value added by the productive activities of the firm.

Value added = Selling price of the product - Cost of the total inputs

Agricultural Marketing : The term agricultural marketing is composed of two words- agriculture and marketing. Agriculture, generally means growing and/or raising of crops and livestock while, marketing encompasses a series of activities involved in moving the goods from the point of production to point of consumption.

Marketing of crops : In planned economic development programme, exchange of goods play a very important role in maintaining equilibrium between production and consumption. The marketing of agricultural products is becoming more important as created by the new world trade order under the WTO agreements. The small and resource poor cultivators have to face big world players for marketing their produce. The prosperity of the cultivators thus depends not only on the increased rate of production, but also upon the method and efficiency with which they dispose of their produce to the great advantage. It assumes special significance in the marketing of perishable commodities like fruits and vegetables because

very small portion of it is consumed by the farm families, therefore, farmers have more marketable surplus.In specialized farming the producers who are in a position to adjust their production to the demand, reap the maximum benefit of the market. If cultivators are unable to adjust their production to the demand of the market, there can be no appreciable improvement in their condition even if the output is better in quality and larger in quantity. Inspite of the partial failure to adjust the production on the farm to the demand of the market, the progressive farmers realize the importance of the study of the market. The element of time is an important factor in marketing of agricultural produce in general and fruits in particular. The marketing possibility of the perishable commodities like fruits depends very largely on the rapidity with which they can be transported to the market. An efficient system of marketing, however, would require many aspects like less number of intermediaries, nominal commission, loading/ unloading charges, minimum marketing cost besides the development of means of transport. Efficient marketing should be such that the produce should reach the consumer in good state without damage, with less cost and within a short time after the produce is harvested.

Market structure : It is always imperative to study the activities of each agency taking part in marketing of agricultural commodities. The agencies that facilitated the flow of commodities till they reached the ultimate consumer are commission/ forwarding agents, wholesalers and retailers etc.

a) Commission agent/ forwarding agent: The function of the commission agent is to sell the produce of a producer without any risk of loss or cost in the study area. In lieu of his services, he charges certain percentage on the total sale value of the commodity. Many commission agents also performed the function of wholesalers and therefore obtained maximum profit out of trade.

b) **Wholesaler :** Wholesaler have got the key position in agricultural marketing and also play sometimes the role of commission agent as well as broker. Generally, they purchased the produce either from commission agents or directly from the producers in the market.

c) **Retailer :** The retailer is a marketing functionary, who caters the needs of the consumers by retailing, generally keep a small establishment and reap a maximum profit especially in fruit trade due to their higher share in the consumers' rupee. Mostly they buy the produce from wholesalers as well as producers early in the morning and sell it out during the remaining day and thus a vertically integrated markets benefits both the producers and

consumers. Some small producers themselves perform the function of retailing and receive better price of their produce in the nearby local markets.

Analysis of Marketing : The data collected is tabulated and analyzed for examining the marketing cost, margins, price spread and the marketing efficiency.

Marketing Margins, Costs and Loss : The post harvest loss at various stages of marketing is included either in the farmer's net margin or market intermediaries margin. The modified formulae can be used for separating the 'post harvest loss during marketing' at different stages of marketing as well as for estimating the producers' share, marketing margins and marketing loss.

a) Net Farmers Price : The net price received by the grower can be estimated as the difference in gross price received and sum of marketing costs and value loss during harvesting, grading, transit and marketing. Thus, the net farmer's price is expressed mathematically as follows:

$$NP_F = GP_F - \{C_F + (L_F \times GP_F)\} \text{ or}$$

$$NP_F = \{GP_F\} - \{C_F\} - \{L_F \times GP_F\} \quad (1)$$

Where NP_F is net price received by the farmers (Rs./kg),

GP_F is gross price received by the farmers or wholesale price to farmers (Rs./kg),

C_F is the cost incurred by the farmers during marketing (Rs./kg),

L_F is physical loss in produce from harvest till it reaches assembly market (per kg)

b) Marketing Margins : The margins of market intermediaries included profit and returns, which accrued to them for storage, the interest on capital and establishment after adjusting for the marketing loss due to handling. The general expression for estimating the margin for intermediaries is given below.

Intermediaries Margin = Gross price (sale price) – Price paid (cost price) – Cost of marketing – Loss in value during wholesaling

Net marketing margin of the wholesaler is given mathematically by

$$MM_w = GP_w - GP_F - C_w - (L_w \times GP_w) \text{ or}$$

$$MM_w = \{GP_w - GP_F\} - \{C_w\} - \{L_w \times GP_w\} \quad (2)$$

Where MM_w is net margin of the wholesaler (Rs./kg),

GP_w is wholesaler's gross price to retailers or purchase price of retailer (Rs./kg)

C_w is cost incurred by the wholesalers during marketing (Rs./kg),

L_w is physical loss in the produce at the wholesale level (per kg)

In the marketing chain, when more than one wholesaler is involved, i.e., primary wholesaler, secondary wholesaler, etc, then the total margin of the wholesaler is the sum of the margins of all wholesalers. Mathematically,

$$MM_w = MM_{w1} + \ldots\ldots\ldots + MM_{wi} + \ldots\ldots\ldots + MM_{wn}$$

Where MM_{wi} is the marketing margin of the i-th wholesaler.

Net marketing margin of retailer is given by:

$$MM_R = GP_R - GP_W - C_R - (L_R \times GP_R) \quad \text{or}$$

$$MM_R = \{GP_R - GP_W\} - \{C_R\} - \{L_R \times GP_R\} \qquad (3)$$

Where MM_R is net margin of the retailer (Rs./kg),

GP_R is price at the retail market or purchase price of the consumers (Rs./kg)

L_R is physical loss in the produce at the retail level (per kg),

C_R is the cost incurred by the retailers during marketing (Rs./kg).

The first bracketed term in equations (1), (2) and (3) indicates the gross return, while the second and third bracketed terms indicate respectively the cost and loss at different stages of marketing.

Thus, the total marketing margin of the market intermediaries (MM) is calculated as

$$MM = MM_W + MM_R \qquad (4)$$

Similarly, the total marketing cost (MC) incurred by the producer/ seller and by various intermediaries is calculated as

$$MC = C_F + C_W + C_R \qquad (5)$$

Total loss in the value of produce due to injury/ damage caused during handling of produce from the point of harvest till it reaches the consumers is estimated as

$$ML = \{L_F \times GP_F\} + \{L_W \times GP_W\} + \{L_R \times GP_R\} \qquad (6)$$

c) Marketing Efficiency

Most commonly used measures are conventional input to output marketing ratio, Shepherd's ratio of value (price) of goods marketed to the cost of marketing (Shephard, 1965) and Acharya's modified marketing efficiency formula (Acharya and Agarwal, 2001). However, all these measures do not explicitly mention the loss in the produce during the marketing process as a separates item in marketing. As reduction in loss itself is one of the efficiency parameters, there has been a need to incorporate this component explicitly in the existing marketing ratios to get correct measures of marketing efficiency while comparing alternate markets/ channels. 'Marketing loss' component was incorporated in the widely used

formula as given by Acharya and Agarwal (2001) and the modified marketing efficiency (ME) formula is given below.

$$ME = \frac{NP_F}{MM + MC + ML} \quad (7)$$

Where NP_F is net price received by the farmers (Rs./kg),

MM is the marketing margin,

MC is marketing cost,

ML is marketing loss.

Analysis of Seasonal Indices

$$S.I = \frac{PI}{MA(12)} \times 100$$

$$MA(12) = \frac{1}{12} \sum PI$$

Where

MA(12)= Twelve month moving average

PI = Market arrivals/ Price indices

S.I = Seasonal Indices for market arrivals/ prices

References

Acharya, S. S. and Aggarwal, N. L. 2001. *Agricultural Marketing in India*. Third edition, Oxford & IBH Publishing Company, New Delhi.

Bhat, A. 2011. Economic Analysis of Production and Marketing of Citrus in Jammu region of Jammu and Kashmir state. Ph.D (Agricultural Economics) Thesis. SKUAST-Jammu.

Edwards, W. and Duffy, P. 2014. Farm Management in Encyclopedia of Agriculture and Food Systems, 100-112.

CHAPTER 8

BASIC DESIGNS AND THEIR STATISTICAL ANALYSIS USING MINITAB

Bilal Ahmad Bhat and Manish Kumar Sharma

Division of Social Sciences, Faculty of Fisheries, SKUAST-K and Division of Statistics and Computer Science, Faculty of Basic Sciences, SKUAST-J, Main campus, Chatha Jammu

In decision making experimentation is an important component of any scientific investigation. Design of experiment means how to design an experiment in the sense that how the measurements or observations should be obtained to answer a query in a valid, efficient and inexpensive way. The designing of the experiment and the analysis of obtained data are inseparable. The data generated from any scientific experiment is valid, if the experiment is designed properly keeping in mind the question and proper analysis of data provides the valid statistical inferences. In case experiment is not well designed, the validity of the statistical inferences is questionable and may be invalid. A proper planning of an experiment essentially demands to identify factors that may cause variability. It is essential to understand first the basic terminologies used in the experimental design.

Experimental unit: For conducting an experiment, the experimental material is divided into smaller parts and each part isreferred to as an experimental unit. The experimental unit is randomly assigned to treatment is the experimental unit. The phrase "randomly assigned" is very important in this definition.

Experiment: A way of getting an answer to a question which the experimenter wants to know.

Treatment: Different objects or procedures which are to be compared in an experiment are called treatments.

Sampling unit: The object that is measured in an experiment is called the sampling unit. This may be different from the experimental unit**.**

Factor: A factor is a variable defining a categorization. A factor can be fixed or random in nature. A factor is termedas a fixed factor if all the levels of interest are included in the experiment. A factor is termed as a random factor if all the levels of interest are not included in the experiment and thosethat are can be considered to be randomly chosen from

all the levels of interest.

Replication: It is the repetition of the experimental situation by replicating the experimental unit.

Experimental error: The unexplained random part of the variation in any experiment is termed as experimental error. An estimate of experimental error can be obtained by replication.

Treatment design: A treatment design is the manner in which the levels of treatments are arranged in an experiment.

Example: Suppose some varieties of fish food is to be investigated on some fish species. The food is placed in the water tanks containing the fishes. The response is the increase in the weight of fish. The experimental unit in this case isthe tank, as the treatment is applied to the tank, not to the fish. Note that if the experimenter had taken the fish in hand and placed the food in the mouth of fish, then the fish would have been the experimental unit as long as each of the fish got an independent scoop of food.

Contrast and Analysis of Variance : The main procedure adopted for the analysis and interpretation of the data collected from an experiment is the analysis of variance procedure that essentially consists of partitioning the total variation in an experiment into components ascribable to different sources of variation due to the controlled factors and error.

The discussion below attempts to relate the method of analysis of variance to comparisons among treatment effects that in terms of symbols can be called *contrasts*.

Contrasts:

Suppose $y_1, y_2, y_3, \ldots, y_n$ denote n observations or any other quantities. The linear function $C = \sum_{i=1}^{n} l_i y_i$ where l_i's are given numbers such that $\sum_{i=1}^{n} l_i = 0$, is called a contrast of $y_i's$.

Suppose $y_1, y_2, y_3, \ldots, y_n$ be independent random variables with a common mean μ and variance σ^2, the expected value of the random variable, C is zero and the variance is $\sigma^2 \sum_{i=1}^{n} l_i^2$. In what follows we shall not distinguish between a contrast and its corresponding variable.

Sum of squares (s.s) of contrasts. The sum of square due to the contrast C is defined as

$\sigma^2 / \sigma^{-2}Var(C) = C^2 / \left(\sum_{i=1}^{n} l_i^2\right)$. It is known that this square has a $\sigma^2 \chi^2$ distribution. With one degrees of freedom when the yi's are normally distributed. Rhus the sum of squares due to two or more contrasts has also a distribution if the contrasts are independent.

It is to be remembered that multiplication of any contrast by a contrast does not change the contrast. The sum of squares due to a contrast as defined above is not evidently changed by such multiplication.

Orthogonal contrast. Two contrasts, $C_1 = \sum_{i=1}^{n} l_i y_i$ and $C_2 = \sum_{i=1}^{n} m_i y_i$ are said to be orthogonal if $\sum_{i=1}^{n} l_i m_i = 0$. This condition ensures that the covariance between C_1 and C_2 is zero when the observations are independent, because $Cov(C_1, C_2) = \sigma^2 \sum_{i=1}^{n} l_i m_i$.

When there are more than two contrasts, they are said to be mutually orthogonal if they are orthogonal pairwise. For example, in case of four observations y_1, y_2, y_3, y_4, we may write following three mutually orthogonal contrasts:

(i) $y_1+y_2-y_3-y_4$

(ii) $y_1-y_2-y_3+y_4$

(iii) $y_1-y_2+y_3-y_4$

The sum of squares due to a set of mutually orthogonal contrasts has a $\sigma^2 \chi^2$ distribution with as many degrees of freedom as the number of contrasts in the set.

Maximum number of orthogonal contrasts. Given a set of n observations y_1, y_2, y_3, …, y_n, the maximum number of mutually orthogonal contrasts among them is n-1. One way of writing such contrasts is to progressively introduce the values as under:

(i) $y_1 - y_2$

(ii) $y_1+y_2-2y_3$

.

.

.

(n) $y_1+y_2+y_3+ \ldots + y_{n-1} -(n-1)y_n$.

Another set of orthogonal contrasts for values of n is available in the Tables for Statisticians and Biometricians prepared by Fisher and Yates (1973) under the name of

orthogonal polynomials.

Design of experiment: One of the main objectives of designing an experiment is how to verify the hypothesis in an efficient and economical way. In the contest of the null hypothesis of equality of several means of normal populations having the same variances, the analysis of variance technique can be used. Note that such techniques are based on certain statistical assumptions. If these assumptions are violated, the outcome of the test of a hypothesis then may also be faulty and the analysis of data may be meaningless. So the main question is how to obtain the data such that the assumptions are met and the data is readily available for the application of tools like analysis of variance. The designing of such a mechanism to obtain such data is achieved by the design of the experiment. After obtaining the sufficient experimental unit, the treatments are allocated to the experimental units in a random fashion. Design of experiment provides a method by which the treatments are placed at random on the experimental units in such a way that the responses are estimated with theutmost precision possible.

Principles of experimental design: There are three basic principles of design which were developed by Sir Ronald A. Fisher.

There are three Basic principles

(i) Replication: the repetition of the treatments under investigation. Replication provides an efficient way of increasing the precision of an experiment. The precision increases with the increase in the number of observations. Replications are essential to obtain a valid estimate of the experimental error variance and are necessarily required to attach a probability statement with estimated treatment differences.

(ii) Randomization: The principle of randomization involves the allocation of treatment to experimental units at random to avoid any bias in the experiment resulting from the influence of some extraneous unknown factor that may affect the experiment. In the development of analysis of variance, we assume that the errors are random and independent. In turn, the observations also become random. The principle of randomization ensures this.

The random assignment of experimental units to treatments results in the following outcomes.

a) It eliminates systematic bias.
b) It is needed to obtain a representative sample from the population.
c) It helps in distributing the unknown variation due to confounded variables throughout the experiment and breaks the confounding influence.

Randomization forms a basis of a valid experiment but replication is also needed for the validity of the experiment. If the randomization process is such that every experimental unit has an equal chance of receiving each treatment, it is called complete randomization.

(iii) Local Control: is a device to maintain greater homogeneity of experimental units within a block of an experiment or as a whole. For example, in animal experiments, local control means the animals of the same breed, age, weight, lactation etc., be grouped together to constitute a block. Local control reduces experimental error and makes designs more efficient.

Complete and incomplete block designs: In most of the experiments, the available experimental units are grouped into blocks having more or less identical characteristics to remove the blocking effect from the experimental error. Such design is termed as block designs.

The number of experimental units in a block is called the block size.
If size of block = number of treatments and each treatment in each block is randomly allocated,

then it is a full replication and the design is called a complete block design.

In case, the number of treatments is so large that a full replication in each block makes it too heterogeneous with respect to the characteristic under study, then smaller but homogeneous blocks can be used. In such a case, the blocks do not contain a full replicate of the treatments. Experimental designs with blocks containing an incomplete replication of the treatments are called incomplete block designs.

Analysis of Variance (ANOVA) : ANOVA technique is a powerful statistical tool for test of significance. Suppose we want to assess the performance of students of various schools in a common examination. The mean score of each school will show a variation, also the scores of individual students within each school. It is at times difficult to tell at a glance whether the variation between schools are significant compared to variation within schools. It is for this

reason that techniques of "Analysis of Variance" hence been developed. The basic purpose of the analysis of variance is to test the homogeneity of several means. The term "Analysis of Variance" was introduced by Professor R.A.Fisher in 1920's. According to him Analysis of Variance (ANOVA) is the "Separation of variance ascribable to one group of causes from the variance ascribable to the other group". The total variation in any set of numerical data is due to a number of causes which may be classified as: Assignable Causes (ii) Chance Causes

Characteristics of ANOVA : ANOVA technique is not designed to test the equality of several population variances. Rather its objective is to test the equality of several population means or the homogeneity of several independent sample means.

Assumptions of ANOVA

1. Normality
   - Populations are Normally Distributed
2. Homogeneity of Variance
   - Populations have Equal Variances
3. Independence of Errors
   - Independent Random Samples are Drawn

One-Way and Two-Way Analysis of Variance : we use t-test in "Sampling and Tests of significance", to test the null hypothesis when we have means of only two samples. However, in situations where we have three or more samples to consider at a time an alternative method is needed for testing the hypothesis that all samples could likely be drawn from the same population. A One-Way ANOVA ("analysis of variance") compares the means of two or more independent groups in order to determine whether there is statistical evidence that the associated population means are significantly different. It is a parametric test. A two-way ANOVA instead compares multiple groups of two factors.

Some Simple designs we explain here

Completely Randomized Design (CRD) : This is the simplest type of design, in which the whole experimental material is divided into a number of experimental units depending upon the number of treatments and the number of replications for each. After that the treatments are allotted to the units entirely by chance. In case of field experiments the whole field is divided into a required number of equal plots and then the treatments are randomized in these plots.

If there are 5 treatments **A**,**B**,**C**,**D** and **E** and 4 replications to each, the number of plots will be 20 and each treatment will be allotted to 4 plots selected at random as 20 is a 2-digit

number, so 2-digit random number table will be consulted and a series of 20 random numbers will be taken excluding those which are greater than 20. Suppose the random numbers are 4,18,2,14,3,7,13,1,6,10,17,20,8,15, 11, 5, 9, 12, 16, 19. After this plots will be serially numbered and the treatment **A** will be allotted to the plots bearing serial numbers 4, 18, 2, 14 and so on for treatments **B**,**C**,**D** and **E.** In a similar way the randomization can be done for any number of treatments.

Statistical Model

The Statistical model in case of CRD is given by

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

where

Y_{ij} = jth individual unit in the ith treatment; μ = General mean; t_i = ith treatment effect and e_{ij} = error term.

Here source of variation are Between and Within treatments.

Statistical Analysis

It is analogous to that of analysis of variance in case of one-way classified data. The various steps (with usual notations) required are given as

$$Correction\ \text{Factor (C.F)} = \frac{(G.T)^2}{n} \quad \text{where} \quad G.T = \sum_{i=1}^{t}\sum_{j=1}^{r} y_{ij}$$

$$TSS = \sum_{i=1}^{t}\sum_{j=1}^{r} y_{ij}^2 - C.F$$

$$SST = \left[\frac{T_1^2}{r_1} + \frac{T_2^2}{r_2} + \ldots + \frac{T_t^2}{r_t}\right] - C.F \quad \text{(For unequal replications)}$$

$$SST = \left[\frac{T_1^2 + T_2^2 + \ldots + T_t^2}{r}\right] - C.F \text{ (For equal replications)}$$

$$SST = \frac{\sum_{i=1}^{t} T_i^2}{r} - C.F$$

SSE=TSS-SST

Analysis of variance Table for CRD

If we suppose the number of treatments to be ‘t’ and number of replications to be ‘r’ for every treatment, the total number of experimental units will be N = t x r. If the treatments have varying number of replications, r_1 , r_2 , …, r_t, then the total number of units (=N) will be given by

$$N = r_1 + r_2 + \ldots + r_t$$

Here the two independent sources of variation are (i) Between treatments and (ii) Within treatments. The ANOVA table is given as

Sources of Variation	d.f.	SS	MSS	F_{cal}
Between Treatments	t-1	SST	$SST/(t-1) = s_t^2$	s_t^2/ s_e^2
Within Treatments	N-t	SSE	$SSE/(N-t) = s_e^2$	
Total	N-1			

$F_{tab(t-1,\ N-t,\ 0.05)} = ?$

If $F_{cal} < F_{tab}$, accept H0 i.e., treatments do not differ significantly and in case $F_{cal} > F_{tab}$, reject H0 i.e., treatments differ significantly.

Pairwise Comparison

If $\overline{Y_{i.}} - \overline{Y_{j.}} \geq C.D$ in that case jth treatment differs and if $\overline{Y_{i.}} - \overline{Y_{j.}} < C.D$ in that case jth treatment does not differ.

Standard Errors

The standard error of difference between the two treatment means based on r_i and r_j replicates is estimated by the following relation

$$S.E._{diff} = \sqrt{Error\ \text{MS}\left(\frac{1}{r_i} + \frac{1}{r_j}\right)}$$

If the number of replications are equal i.e., $r_i = r_j = r$ then we have

$$S.E._{diff} = \sqrt{\frac{2\ Error\ \text{MS}}{r}}$$

$$Critical\ \text{Difference} = S.E._{diff} \times t_{5\%} \text{ for error d.f.}$$

Remark: (i) If C.V. (coefficient of variation) is below 15%, experiment is excellent i.e., experiment has high precision

(ii) If C.V. is below 15-20 %, experiment is moderate and

(iii) If C.V. is above 20 %, experiment is very poor and needs to be repeated.

Advantages of CRD

(i) In this design any number of treatments and replicates may be used. The number of replicates can also be varied at will from treatment to treatment.

(ii) The statistical analysis of data is very easy and it remains so even if number of replicates are not same for all treatments.

(iii) The method of analysis remains simple when the results from some units are missing/rejected.

(iv) The relative loss of information due to missing data is smaller than that with any other design.

(v) The design provides maximum number of degrees of freedom for the estimation of error as compared with other designs, for a given number of treatments and a given number of experimental units.

Disadvantages of CRD

The main objection to this design is on the grounds of accuracy. Since there is no restriction on the randomization of treatments, we cannot be sure about the fact. The units receiving one treatment are similar to those receiving the other treatment, and therefore the whole of variation among the units enters into experimental error. In field experiments, it has been replaced by RBD.

Applications of CRD

(i) where the experimental material is limited in quantity and is homogenous

(ii) where it is expected that some of the units will be destroyed or will fail to respond.

(iii) In smaller experiments, where increased accuracy from the alternate design is not sufficient to exceed in importance the loss of error degrees of freedom.

(iv) This design is frequently used in laboratory experiments.

Example 1. (Treatment with equal number of replications)

The following table gives the yield in pounds per plot of 5 varieties of wheat after being applied to each of 4 plots, completely randomized.

Varieties	**Yield in lbs**			
A	**08**	**08**	**06**	**10**
B	**10**	**12**	**13**	**09**
C	**18**	**17**	**13**	**16**
D	**12**	**10**	**15**	**11**
E	**8**	**11**	**09**	**08**

Analyse the data.

Solution: We set up null hypothesis that the treatments do not differ significantly.

Varieties	Yield in lbs				Total	Mean
A	**08**	**08**	**06**	**10**	**32**	**8**
B	**10**	**12**	**13**	**09**	**44**	**11**
C	**18**	**17**	**13**	**16**	**64**	**16**
D	**12**	**10**	**15**	**11**	**48**	**12**
E	**8**	**11**	**09**	**08**	**36**	**9**
Total					**224**	

Correction Factor (C.F)= $224^2/20 = 2508.8$

TSS = $(8^2 + 8^2 + \ldots + 9^2 + 8^2) - C.F = 207.2$

SS for varieties = $(32^2 + 44^2 + 64^2 + 48^2 + 36^2)/4 - C.F = 155.2$

SS for error = TSS – SS for varieties = 207.2 – 155.2 = 52.0

The ANOVA table is given as

Source of Variation	d.f.	SS	MSS	F	F_{tab} 5%	1%
Between Varieties	4	155.2	38.80	11.18	3.06	8.25
Within Varieties (error)	15	52.0	3.47			
Total	19	207.2				

Since calculated value of F is greater than tabulated value of F at 1% level of significance, we reject null hypothesis H0 and conclude that treatments differ significantly. To compare varieties we compute CD, given by

$$Critical\ \text{Difference} = S.E._{diff} \times t_{5\%} \text{ for } 15 \text{ d.f.}$$

i.e., $$Critical\ \text{Difference} = \sqrt{\frac{2 \times 3.47}{4}} \times 2.131.$$

Therefore, CD= 2.81.

The varieties can be compared by setting them in decreasing order of their yields

Varieties:	C	D	B	E	A
Mean yields/plot:	16	12	11	9	8

The varieties which do not differ significantly have been underlined by a bar.

Example 2: (Treatments with unequal number of replications)

The following table gives the yields in pounds per plot, of 5 varieties of wheat after being applied to 4, 3, 2, 4 and 3 plots respectively. Completely randomized

Varieties	Yield in lbs				Total	Mean
A	08	08	06	10	32	8.0
B	10	09	08	-	27	9.0
C	08	10	-	-	18	9.0
D	07	10	09	08	34	8.5
E	12	08	10	-	30	10.0
Total					141	

Analyse the data.

Solution:

Correction Factor (C.F)= $141^2/16 = 1242.5625$

TSS = $(8^2 + 8^2 + \ldots + 8^2 + 10^2) - C.F = 32.4375$

SS for varieties = $(32^2/4 + 27^2/3 + 18^2/2 + 34^2/4 + 30^2/3) - C.F = 7.4375$

SS for error = TSS – SS for varieties = 32.4375 – 7.4375 =25.0

The ANOVA table is given as

Source of Variation	d.f.	SS	MSS	F	F_{tab} 5%	1%
Between Varieties	4	7.4375	1.8594	0.81	3.36	5.41
Within Varieties (error)	11	25.0	2.2727			
Total	15	32.4375				

Since calculated value of F is less than tabulated value of F at 5% level of significance, we accept null hypothesis H0 and conclude that treatments do not differ significantly. Here treatments do not differ significantly so we do not compute CD.

The Randomized Complete Block design (RCBD)

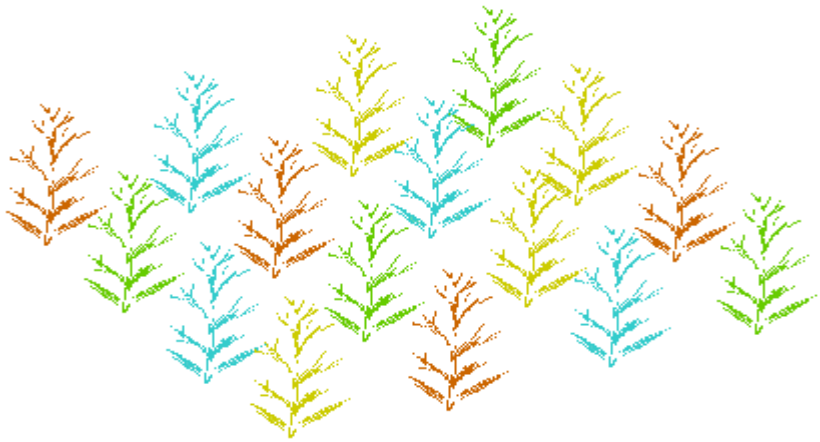
The RCBD or RBD is the standard design for agricultural experiments. The field or orchard is divided into units to account for any variation in the field. Treatments are then assigned at random to the subjects in the blocks-once in each block.

Field marks:

- Treatments are assigned at random within blocks of adjacent subjects, each treatment once per block.
- The number of blocks is the number of replications.
- Any treatment can be adjacent to any other treatment, but not to the same treatment within the block.
- Used to control variation in an experiment by accounting for spatial effects.

Sample layout:

Different colors represent different treatments; each horizontal row represents a block. There are 4 blocks (I-IV) and 4 treatments (A-D) in this example.



Block I A B C D

Block II D A B C

Block III B D C A

Block IV C A B D

Statistical Model

The Statistical model in case of RBD is given by

$$Y_{ij} = \mu + t_i + b_j + e_{ij}$$

where

Y_{ij} = yield of experimental unit from ith treatment and jth block; μ = General mean; t_i = ith treatment effect; bj=effect due to jth block or replicate and e_{ij} = error effect due to random

component assumed to be independently normally distributed with mean zero and variance σ_e^2 i.e., e_{ij} are i.i.d. N(0, σ_e^2).

ANOVA table format:

Source of variation	Degrees of freedom	Sums of squares (SSQ)	Mean square (MS)	F
Blocks (*B*)	b-1	SSQ_B	$SSQ_B/(b-1)$	MS_B/MS_E
Treatments (*Tr*)	t-1	SSQ_{Tr}	$SSQ_{Tr}/(t-1)$	MS_{Tr}/MS_E
Error (*E*)	(t-1)*(b-1)	SSQ_E	$SSQ_E/((t-1)*(b-1))$	
Total (*Tot*)	t*b-1	SSQ_{Tot}		
[a]where t=number of treatments and b=number of blocks or replications.				

Advantages of the RCBD

1. Generally more precise than the CRD.
2. No restriction on the number of treatments or replicates.
3. Some treatments may be replicated more times than others.
4. Missing plots are easily estimated.
5. Whole treatments or entire replicates may be deleted from the analysis.
6. If experimental error is heterogeneous, valid comparisons can still be made.

Disadvantages of the RCBD

1. Error df is smaller than that for the CRD (problem with a small number of treatments).
2. If there is a large variation between experimental units within a block, a large error term may result (this may be due to too many treatments).
3. If there are missing data, a RCBD experiment may be less efficient than a CRD

NOTE: The most important item to consider when choosing a design is the uniformity of the experimental units.

Example: Plan and yield of paddy strains (kg/plot) in a RBD experiment is shown below:

Block I	Block II	Block III	Block IV
A(12)	B(4)	B(7)	F(8)
E(14)	C(6)	C(9)	A(18)
C(11)	E(11)	D(9)	C(10)
D(7)	A(16)	E(15)	B(6)

B(5)	D(8)	F(12)	D(8)
F(10)	F(9)	A(14)	E(12)

Analyse the data.

Solution: We set up the null hypothesis that blocks also treatments do not differ significantly.

The given data can be arranged in the following two way classification

Treatment	Replications or Blocks				Treatment Total
	I	II	III	IV	
A	12	16	14	18	60
B	5	4	7	6	22
C	11	6	9	10	36
D	7	8	9	8	32
E	14	11	15	12	52
F	10	9	12	8	39
Block Total	59	54	66	62	241

Grand Total = 241, replication r = 4; number of treatments (t) = 6, rt = N = 24 plot observations

Correction Factor (C.F)= $241^2/24 = 2420$

Total SS = $(12^2 + 5^2 + \ldots + 12^2 + 8^2) - C.F = 2717 - 2420 = 297$

Block or Replication SS = $[(59^2 + 54^2 + 66^2 + 62^2)/6] - C.F = 2432 - 2420 = 12$

Treatment or Variety SS = $[(60^2 + 22^2 + 36^2 + 32^2 + 52^2 + 39^2)/4] - C.F = 2657 - 2420 = 237$

Error SS = Total SS – Block SS – Variety SS = 297 – 12 – 237 = 48

ANOVA Table for RBD

Sources of Variation	d.f.	SS	MSS	F_{cal}	Table F 5%	Table F 1%
Blocks	r-1=3	12	4			
Varieties	t-1=5	237	47	14.6^{**}		
Error	(r-1)(t-1)=15	48	3.2		2.90	4.56
Total	rt-1=23	297				

** Significant at 1 % level of significance.

Since the calculated value of F is greater than the tabulated value of F for 5 and 15 d.f at 1 % level of significance, the conclusion is that the varieties differ significantly at 1 % level or the varietal differences are highly significant.

The critical difference is given by (for treatments)

$$Critical\ \text{Difference} = S.E._{diff} \times t_{5\%} \text{ for } 15 \text{ d.f.}$$

i.e., $$Critical\ \text{Difference} = \sqrt{\frac{2 \times 3.2}{4}} \times 2.131.$$

Therefore, C.D= 2.69.

Arrange the varieties in order of preference as

Varieties:	A	E	F	C	D	B
Mean yield (Kg/plot):	15.0	13.0	9.8	9.0	8.0	5.5

It is clear that varieties A and E are superior to B,D, C, F; while D,C,F are on par (on par varieties are underlined).

Latin Square Design

In RBD whole of the experimental area is divided into relatively homogenous groups (blocks) and treatments are allocated at random to units within each block, i.e., randomization was subjected to one restriction, i.e., within blocks. But in field experimentation it may happen that experimental area (field) exhibit fertility in strips e.g., cultivation might result in alternative strips of high or low fertility. RBD will be effective if the blocks happen to be parallel to these strips and would be extremely inefficient if the blocks are across the strips. Initially fertility gradient is seldom known. A useful method of eliminating fertility variations consists in an experimental layout which will control variation in two perpendicular directions. Such a layout is a Latin Square Design (LSD).

Layout of Design

In this design the number of treatments is equal to the number of replications. Thus in case of m treatments, there have to be $m \times m = m^2$ experimental units. The whole of experimental area is divided into m^2 experimental units (plots) arranged in a square so that each row as well as each column contains m units (plots). The m treatments are then allocated at random to these rows and columns in such a way that every treatment occurs once in each row and

column. LSD and is extensively used in agricultural experiments, e.g., if we are interested in studying the effects of m types of fertilizers on the yield of a certain variety of wheat, it is customary to conduct the experiments on a square field with m^2-plots of equal area and to associate treatments with different fertilizers and row and column effects with variations in fertility of soil. Obviously, there can be many arrangements for an $m \times m$ LSD and a particular layout in an experiment must be determined randomly, e.g., with four treatments A,B,C and D, one typical arrangement of 4×4 LSD is given below:

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

Fisher and Yates have tabulated Latin Square upto 12×12 .

Linear Statistical Model

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + e_{ijk}$$

where Y_{ij} = jth individual unit in the ith treatment; μ = constant mean effect;α_i, β_j and γ_k are the effects due to the ith row, jth column and kth treatments respectively and e_{ijk} = error term.

ANOVA Table for $m \times m$ LSD (with usual notations)

Sources of Variation	d.f.	SS	MSS	F_{cal}
Rows	m-1	S_R^2	$S_R^2/(m-1) = s_r^2$	$F_R = s_r^2/ s_e^2$
Columns	m-1	S_C^2	$S_C^2/(m-1) = s_c^2$	$F_C = s_c^2/ s_e^2$
Treatments	m-1	S_T^2	$S_T^2/(m-1) = s_t^2$	$F_T = s_t^2/ s_e^2$
Error	(m-1)(m-2)	S_E^2	$S_E^2/(m-1) = s_e^2$	
Total	$m^2 - 1$			

Here we set up the null hypothesis

for row effects, H_α, : $\alpha_1 = \alpha_2 = \ldots = \alpha_m = 0$,

for column effects, H_β, : $\beta_1 = \beta_2 = \ldots = \beta_m = 0$ and

for treatment effects, H_γ, : $\gamma_1 = \gamma_2 = \ldots = \gamma_m = 0$.

If $F_R > F_\alpha$ (i.e., tabulated value of F for (m-1), (m-1)(m-2) d.f) we reject H_α otherwise we may we may accept H_α. Similarly, we can test H_β and H_γ.

Example: Test the effect of 5 varieties A,B,C,D and E of a crop in a Latin Square Design. Plan and yield (kg/plot) are as given below:

Column ↓	Rows →					
	I	II	III	IV	V	Row Total
I	B(6)	A(11)	E(8)	D(6)	C(5)	36
II	A(9)	D(9)	C(4)	E(14)	B(10)	46
III	C(3)	B(8)	D(7)	A(12)	E(8)	38
IV	E(10)	C(5)	A(10)	B(7)	D(10)	42
V	D(8)	E(15)	B(9)	C(3)	A(8)	43
Column Total	36	48	38	42	41	205

Analyse the data.

Solution: First we set up the null hypothesis that rows, columns and treatment effects are non-significant

Yield (kg/plot) as per treatments

A	B	C	D	E
9	6	3	8	10
11	8	5	9	15
10	9	4	7	8
12	7	3	6	14
8	10	5	10	8
Total = 50	40	20	40	55
Mean = 10	8	4	8	11

Now, we compute

Correction Factor (C.F)= $205^2/25 = 1681$

TSS = $(9^2 + 6^2 + \ldots + 10^2 + 8^2) - C.F = 1903-1681=222$

Row SS = $[(36^2 + 46^2 + 38^2 + 42^2 + 43^2)/5] - C.F = 1694 - 1681 = 13$

Column SS = $[(36^2 + 48^2 + 38^2 + 42^2 + 41^2)/5] - C.F = 1698 - 1681 = 17$

Treatment SS = $[(50^2 + 40^2 + 20^2 + 40^2 + 55^2)/5] - C.F = 1825 - 1681 = 144$

Error SS = Total SS – Row SS – Column SS – Treatment SS = 222 – 13 – 17 -144 = 48

ANOVA Table for 5×5 LSD

Sources of	d.f.	SS	MSS	F_{cal}	Table F

Variation					5%	1%
Rows	4	13	3.25	0.80		
Columns	4	17	4.25	1.06		
Treatments	4	144	36.00	9.00	3.26	5.41
Error	12	48	4.00			
Total	24	222				

Comparing the F ratio for treatments, with the table value of F (for 4 and 12 d.f), it is found that differences in varietal means are highly significant. Further, rows and columns effect are non-significant, indicating that fertility variation in either direction is not very much. In this case a simple RBD experiment would have been more efficient. The critical difference is given by (for treatments)

$$Critical\ \text{Difference} = S.E._{diff} \times t_{5\%} \text{ for } 12 \text{ d.f.}$$

i.e., $$Critical\ \text{Difference} = \sqrt{\frac{2 \times 4}{5}} \times 3.055.$$

Therefore, C.D=3.85

Factorial Experiments: A factorial experiment in experimental design is used to study two or more factors, each with multiple discrete possible values or "levels". Several factors affect simultaneously the characteristic under study in factorial experiments and the experimenter is interested in the main effects and the interaction effects among different factors. A factorial experiment is necessary when interactions may be present to avoid misleading conclusions.It is suggested reader must go through literature available on design of experiments and allied fields (Das and Giri, 1986; Montgomery, 2017)

Missing plot techniques:

It happens many time in conducting the experiments that some observation are missed. This may happen dueto several reasons. For example, in a clinical trial, suppose the readings of blood pressure are to be recorded after three days of giving the medicine to the patients. Suppose the medicine is given to 20 patients and one of the patients doesn't turn up for providing the blood pressure reading. Similarly, in an agriculturalexperiment, the seeds are sown and yields are to be recorded after few months. Suppose some cattle destroy the crop of any plot or the crop of any plot is destroyed due to storm, insects etc. In such cases, one option is to

- somehow estimate the missing value on the basis of available data,
- replace it back in the data and make the data set complete.

Now conduct the statistical analysis on the basis of completed data set as if no value was missing by making necessary adjustments in the statistical tools to be applied. Such an area comes under the purview of "missing data models" and a lot of development has taken place. One can read in detail about these in books like Little and Rubin, D.B., 2002 and Schafer, 1997.

Analysis of Statistical Designs Using MINITAB Software

(a) Completely Randomized Design (CRD)

Example: Following are the yields obtained in kgs of three varieties WL-711, WG-357 and 1562 of

wheat sown in 14 plots.

WL-711:	12	13	14	13	
WG-357	10	9	10	9	9
1562:	13	14	13	12	14

Is there any significant difference in the production of three varieties? Also calculate critical difference.

Computer Program:

```
MTB > Name c1 ='WL-711'
MTB > Name c2 ='WG-357'
MTB > Name c3='1562'
MTB > Set c1
DATA> 12, 13, 14, 13
DATA> End
MTB > Set c2
DATA> 10, 9, 10, 9, 9
DATA> End
MTB > Set c3
DATA> 13, 14, 13, 12, 14
DATA> End
MTB > Aovoneway c1, c2, c3
```

The data should be entered in the following manner(Menu commands)

C1	C2
Yield	Level
12	1
13	1
14	1
13	1
10	2
9	2
10	2
9	2
9	2
13	3
14	3
13	3
12	3
14	3

Menu commands

Stat→ ANOVA→ One-way → Yield → button [puts Yield under variables list Response] → Level →

button[puts Level under variable list Factor] → Comparisions→ (Tick) Tukey's family error rate→ OK

Result:

One-way ANOVA: WL-711, WG-357, 1562

Analysis of Variance

Source	DF	SS	MS	F	P
Factor	2	44.357	22.179	40.66	0.000
Error	11	6.000	0.545		
Total	13	50.357			

Individual 95% CIs For Mean
Based on Pooled StDev

Level	N	Mean	StDev	---+---------+---------+---------+---
WL-711	4	13.000	0.816	(-----*----)

```
WG-357     5     9.400     0.548      (----*----)
1562       5    13.200     0.837                          (----*----)
                                      ---+---------+---------+---------+---
Pooled StDev =    0.739                9.0      10.5      12.0      13.5
Program for Critical Difference
MTB > Invcdf 0.975;
SUBC> t  11.
Inverse Cumulative Distribution Function
Student's t distribution with 11 DF
P( X<= x )         x
      0.9750       2.2010
MTB > Let k1 = 2.2010*sqrt(2*0.545/3)
MTB > Print k1
Data Display
K1    1.32670 (Critical Difference)
```

Interpretation: From ANOVA table we obtain that the varieties differ highly significantly (P-value = 0.000) as regards to yield of wheat. A further comparison among varieties can be made through CD value.

Multiple comparisons of means allow us to examine which means are different and to estimate by how much they are different. We can assess the statistical significance of differences between means using a set of confidence intervals, a set of hypothesis tests or both. The confidence intervals allow us to assess the practical significance of differences among means, in addition to statistical significance. As usual, the null hypothesis of no difference between means is rejected if and only if zero is not contained in the confidence interval.

Which multiple comparison method should I use with One-Way ANOVA?

The selection of the appropriate multiple comparison method depends on the inference that we want. It is inefficient to use the Tukey all-pairwise approach when Dunnett or MCB is suitable, because the Tukey confidence intervals will be wider and the hypothesis tests less powerful for a particular family error rate. For the same reasons, MCB is superior to Dunnett if you want to eliminate factor levels that are not the best and to identify those that are best or

close to the best. The choice of Tukey versus Fisher's LSD methods depends on which error rate, family or individual, we want to specify.

The characteristics and advantages of each method are summarized in the following table:

Method	**Normal Data**	**Strength**	**Comparison with a Control**	**Pairwise Comparison**
Tukey	Yes	Most powerful test when doing all pairwise comparisons.	No	Yes
Dunnett	Yes	Most powerful test when comparing to a control.	Yes	No
Hsu's MCB method	Yes	The most powerful test when you compare the group with the highest or lowest mean to the other groups.	No	Yes
Games-Howell	Yes	Used when you do not assume equal variances.	No	Yes

Duncan's new *multiple range test* (MRT): In statistics, *Duncan's* new *multiple range test* (MRT) is a *multiple comparison* procedure developed by David B. *Duncan* in 1955 which makes use of a least significant difference value for each pair of treatment means. LSD test is usually not suitable, when the total number of treatment is large and in such cases, DMRT is useful. Here treatment means are arranged in order and the least significant value for comparing two treatment means situated at distance n in an ordered set and α level of significance calculated by the formula

$D_{\alpha,n} = s_{\bar{x}} \times (Duncan's\ significant\ range\ value$ for α level, n distance and ν error d.f.) We can note the Duncan's significant ranges from Biometrics, 11, 1-42, 1955 by D.B.Duncan. A different Duncan's least significant value is obtained and compared with the difference in the two treatment means for each pair situated at different distances. The two treatments are taken to differ significantly in their effect if the difference is greater than or equal to

Duncan's least significant value $D_{\alpha,n}$. In a two way difference table all possible pairs off treatment means can conveniently be compared by taking the treatment means in ascending order along rows and descending order along columns. In this two way table all differences are covered by the upper diagonal and diagonal elements. The biggest advantage of this test lies in the fact that it allows the experimenter to commit fewer Type II error and more Type I error than lsd and Newman-Kuel's test. Duncan showed that the level of significance for k-treatments changes according to the following formula, $\alpha_k = 1- (1 – \alpha)^{k-1}$

DMRT involves the computation of numerical boundaries, that allow for the classification of the difference between any 2 treatment means as significant or non-significant.

REMARK

One-Way ANOVA also offers Fisher's LSD method for individual confidence intervals.
Fisher's is not a multiple comparison method, but instead contrasts the individual confidence intervals for the pairwise differences between means using an individual error rate. Fisher's LSD method inflates the family error rate, which is displayed in the output.
Which multiple comparison method should I use with Fit General Linear Model or Fit Mixed Effects Model?
After we use Fit General Linear Model or Fit Mixed Effects Model, use the corresponding analysis to obtain multiple comparisons of means:
Stat > ANOVA > General Linear Model > Comparisons
Stat > ANOVA > Mixed Effects Model > Comparisons
We must make the following choices when using multiple comparisons:
Pairwise comparisons or comparisons with a control
The method of comparison

Pairwise comparisons or comparison with a control

Choose **Pairwise** in the **Options** sub-dialog box when we do not have a control level and we want to compare all combinations of means.

Choose **With a Control** to compare the level means to the mean of a control group. When this method is suitable, it is inefficient to use pairwise comparisons because pairwise confidence intervals are wider and the hypothesis tests are less powerful for a given confidence level.

The multiple comparison method

Choose the comparison procedure based on the group means that we want to compare, the type of confidence level that we want to specify, and how conservative we want the results to be. "Conservative" in this context indicates that the true confidence level is likely to be greater than the confidence level that is displayed.

Except for Fisher's method, the multiple comparison methods have protection against false positives built-in. By protecting against false positives with multiple comparisons, the intervals are wider than if there were no protection.

Some characteristics of the multiple comparison methods are summarized below:

Comparison method	**Properties**	**Confidence level that you specify**
Tukey	All pairwise comparisons only, not conservative	Simultaneous
Fisher	No protection against false positives due to multiple comparisons	Individual
Dunnett	Comparison to a control only, not conservative	Simultaneous
Bonferroni	Most conservative	Simultaneous
Sidak	Conservative, but slightly less than Bonferroni	Simultaneous

What if the p-value from the ANOVA table conflicts with the multiple comparisons output?

The p-value in the ANOVA table and the multiple comparison results are based on different methodologies and can occasionally produce contradictory results. For example, it is possible that the ANOVA p-value can indicate that there are no differences between the means while the multiple comparisons output indicates that some means that are different. In this case, you can generally trust the multiple comparisons output.

We do not need to rely on a significant p-value in the ANOVA table to reduce the chance of detecting a difference that doesn't exist. This protection is already incorporated in the Tukey, Dunnett, and MCB tests (and Fisher's test when the means are equal).

(b) Randomized Block Design (RBD)

Example: For the data (given below) work out ANOVA for RBD alongwith least significant difference

Treatments	Blocks					
	1	2	3	4	5	6
A	42.4	34.3	24.1	36.5	54.2	49.1
B	33.2	33.3	5.0	26.2	30.2	28.6
C	8.5	21.9	6.2	16.0	13.5	15.4
D	16.6	19.3	16.6	2.1	11.1	11.2

Computer Program:

MTB > Name c1 = 'Yield'

MTB > Name c2 = 'Blocks'

MTB > Name c3 = 'Treats'

MTB > Set c1

DATA> 42.4, 34.3, 24.1, 36.5, 54.2, 49.1, 33.2, 33.3, 5.0, 26.2, 30.2, 28.6, 8.5,

DATA>21.9, 6.2, 16.0, 13.5, 15.4, 16.6, 19.3, 16.6, 2.1, 11.1, 11.2

DATA> End

MTB > Set c2

DATA> 1, 2, 3, 4, 5, 6, 1, 2, 3, 4, 5, 6, 1, 2, 3, 4, 5, 6, 1, 2, 3, 4, 5, 6

DATA> End

MTB > Set C3

DATA> 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4

DATA> End

MTB > ANOVA c1 = c2 c3;

SUBC> MEANS c3.

The data should be entered in the following manner(Menu commands)

C1	C2	C3
Yield	Blocks	Treats
42.4	1	1
34.3	2	1
24.1	3	1
36.5	4	1
54.2	5	1
49.1	6	1
33.2	1	2
33.3	2	2

5.0	3	2
26.2	4	2
30.2	5	2
28.6	6	2
8.5	1	3
21.9	2	3
6.2	3	3
16.0	4	3
13.5	5	3
15.4	6	3
16.6	1	4
19.3	2	4
16.6	3	4
2.1	4	4
11.1	5	4
11.2	6	4

Menu commands

Stat→ ANOVA→ One-way → Yield → button [puts Yield under variables list Response] →Level → button[puts Level under variable list Factor] → Comparisions→ (Tick) Tukey's family error rate→ OK

Result:

ANOVA: Yield versus Blocks, Treats

Factor Type Levels Values
Blocks fixed 6 1 2 3 4 5 6
Treats fixed 4 1 2 3 4

Analysis of Variance for Yield

Source	DF	SS	MS	F	P
Blocks	5	632.41	126.48	2.17	0.113
Treats	3	2965.23	988.41	16.93	0.000
Error	15	875.56	58.37		
Total	23	4473.20			

Treatment Means

Treats	N	Yield
1	6	40.100
2	6	26.083
3	6	13.583
4	6	12.817

Program for Critical Difference

```
MTB > Invcdf 0.975;
SUBC> t 15.
Inverse Cumulative Distribution Function
Student's t distribution with 15 DF
P( X<= x )         x
   0.9750       2.1314
MTB > Let k1 = 2.1314*sqrt(2*58.37/6)
MTB > Print k1
```

Data Display
K1 9.40154 (critical difference)
Interpretation:

From ANOVA table we obtain that the treatments differ highly significantly (P-value = 0.000). A further comparison among treatments can be made through CD value i.e. 9.40154.

(c) Latin Square Design (LSD)

Example: The following are the results of the Latin Square experiment on the effect of four manorial treatments A, B, C, D on the yield of sugarcane. Perform the ANOVA and calculate the critical difference for the treatment mean yields.

Row ↓ Column →	**I**	**II**	**III**	**IV**
I	**D (29.1)**	**B (18.9)**	**C (29.4)**	**A (5.7)**
II	**C (16.4)**	**A (10.2)**	**D (21.2)**	**B (19.1)**
III	**A (5.4)**	**D (38.8)**	**B (24.0)**	**C (37.0)**
IV	**B (24.9)**	**C (41.7)**	**A (9.5)**	**D (28.9)**

Computer Program:

```
MTB > set c1
DATA>
29.1,18.9,29.4,5.7,16.4,10.2,21.2,19.1,5.4,38.8,24.0,37.0,24.9,41.7,9.5,2
8.9
DATA> end
MTB > set c2
DATA> 1,1,1,1,2,2,2,2,3,3,3,3,4,4,4,4
DATA> end
MTB > set c3
DATA> 1,2,3,4,1,2,3,4,1,2,3,4,1,2,3,4
DATA> end
MTB > set c4
DATA> 4,2,3,1,3,1,4,2,1,4,2,3,2,3,1,4
DATA> end
MTB > name c1='yield'
MTB > name c2='Rows'
MTB > name c3='Columns'
MTB > name c4='Treatments'
MTB > ancova c1=c2 c3 c4;
SUBC> means c4.
```

Result (Output of MINITAB)

ANCOVA: yield versus Rows, Columns, Treatments
Factor Levels Values

```
Rows        4    1    2    3    4
Columns     4    1    2    3    4
Treatmen    4    1    2    3    4
Analysis of Variance for yield
Source      DF       SS        MS       F       P
Rows         3     259.31     86.44    3.32   0.099
Columns      3     155.27     51.76    1.99   0.218
Treatmen     3    1372.12    457.37   17.55   0.002
Error        6     156.37     26.06
Total       15    1943.08
```

Treatment Means

```
Treatment   N    yield
1           4     7.700
2           4    21.725
3           4    31.125
4           4    29.500
```

```
Program for Critical Difference
MTB > invcdf 0.975;
SUBC> t 6.
Inverse Cumulative Distribution Function
Student's t distribution with 6 DF
P( X<= x )         x
   0.9750       2.4469
MTB > let k1=2.4469*sqrt(2*26.06/4)
MTB > print k1
Data Display
K1    8.83260 (Critical Difference)
```

Interpretation:

From ANOVA table we obtain that the treatments differ highly significantly (P-value = 0.000). A further comparison among treatments can be made through CD value i.e. 8.83260.

(d) Transformation of data: In MINITAB for making logarithmic, square root and arcsine transformations, one can use the Calc → Calculator. It is followed by storing the result in a variable

by entering a target column in in MINITAB worksheet. Then define the functions that are to be used

for transformation of data in the Expression SubDialog Box. We use LOGT(Column number or

variable name to be transformed) for logarithmic transformation and click OK. The transformed data

will appear in the target column. SQRT (Column number or variable to be transformed), we use for

square root transformation in the Expression Subdialog Box and for Arcsine transformation, we use

Expression ASIN (sqrt of the column number in which data is given/100)*180*7/22). The

multiplication of 180*7/22 is done to convert the data from radians to degrees. In case the data lies

between 0 and 1, then do not divide by 100.

Testing Normality: Stat→Basic Statistics→ Normality →In the Dialog Box. Select the stored residual as variable in the Variable list and select one of the three tests viz Anderson-Darling, Ryan-Joinder and Kolmogrov-Smirnov tests and Click OK.

Test for homogeneity of Variance: Stat→ANOVA→ Test for Equality of Variance → In the Dialog Box. Chose the stored residual in the Response Box and Treatment in the Factors Box and then chose the confidence level and click OK.

Example: The data obtained on percentage of unsalable ears of corn is given below:

	Blocks					
Treatments	I	II	III	IV	V	VI
A	42.4	34.3	24.1	39.5	55.5	49.1
B	33.3	33.3	5.0	26.3	30.2	28.6
C	8.5	21.9	6.2	16.0	13.5	15.4
D	16.6	19.3	16.6	2.1	11.1	11.1

Transform the data by using Arc Sin transformation for proportions.

Computer Program:

```
MTB > Name c1 = 'Response'
MTB > Name c2 = 'Transformed data'
MTB > Set c1
DATA> 42.4, 34.3, 24.1, 39.5, 55.5, 49.1, 33.3, 33.3, 5.0, 26.3, 30.2, 28.6, 8.5,
DATA>21.9, 6.2, 16.0, 13.5, 15.4, 16.6, 19.3, 16.6, 2.1
DATA> 11.1, 11.1
DATA> End
MTB > Let c2 = asin(sqrt(c1/100))*57.296
MTB > End
MTB > Print c2
```

Result

Data Display

Transformed data

40.6287	35.8499	29.4010	38.9390	48.1578	44.4845	35.2443
35.2443	12.9210	30.8530	33.3360	32.3298	16.9507	27.9030
14.4183	23.5783	21.5569	23.1058	24.0436	26.0605	24.0436
8.3323	19.4612	19.4612				

References

D.C. Montgomery (2017). *Design and Analysis of Experiments,* 9th edition, Wiley, Hoboken NJ

Giri, N. C. and Das, M. N. (1986). Design and Analysis of Experiments (2nd Edition). *New Age*

International Publishers – New Delhi, pp.1-500.

Little, R.J.A. and Rubin, D.B. (2002). Statistical Analysis with Missing Data, 2nd edition, New York:

John Wiley.

Schafer, J.L. (1997). Analysis of Incomplete Multivariate Data. Chapman & Hall, London etc.

Snedecor, G.W. and *Cochran*, W.G. (*1989*). *Statistical Methods*. 8th Edition, Iowa State University Press.

CHAPTER 9

TIME SERIES MODEL FOR FORECASTING

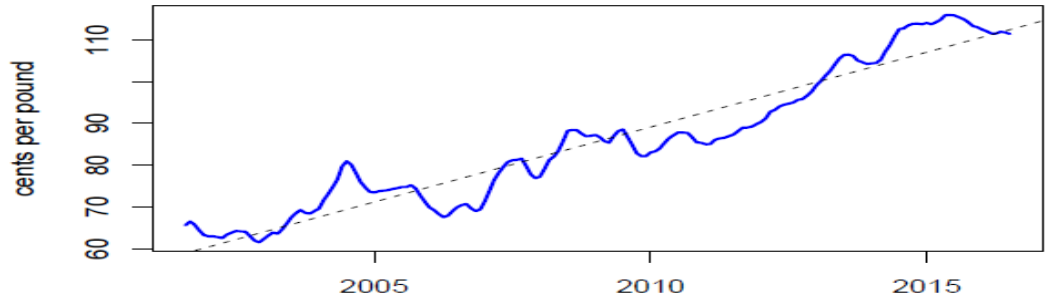
Sunali Mahajan, Manish Sharma, S.E.H. Rizvi and Nishant Jasrotia

Introduction:

John Naisbitt **said** "The most reliable way to forecast the future is to try to understand the present". **The aim of forecasting time series data is to understand how the sequence of observations will continue in the future.** A time series data is the data on the variable which is collected at regular intervals and in a chronological order.

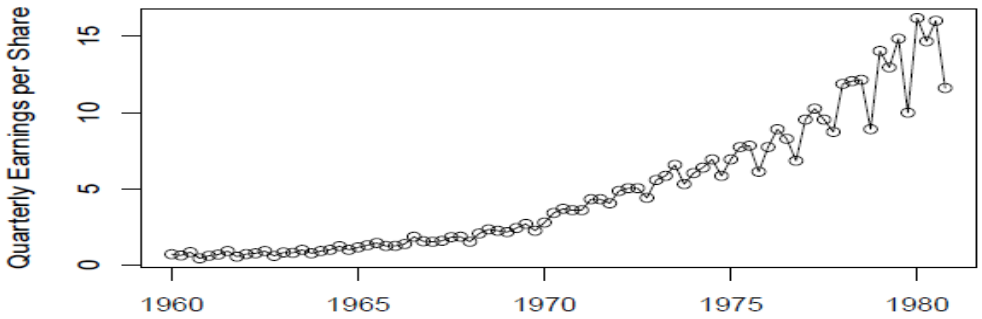
Time Series: A time series is a sequential set of data points, measured typically over successive times. Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. In general, a time series is affected by four components, i.e.; trend, seasonal, cyclical and irregular components.

1. **Trend:** The general tendency of a time series to increase, decrease or stagnate over a long period of time.



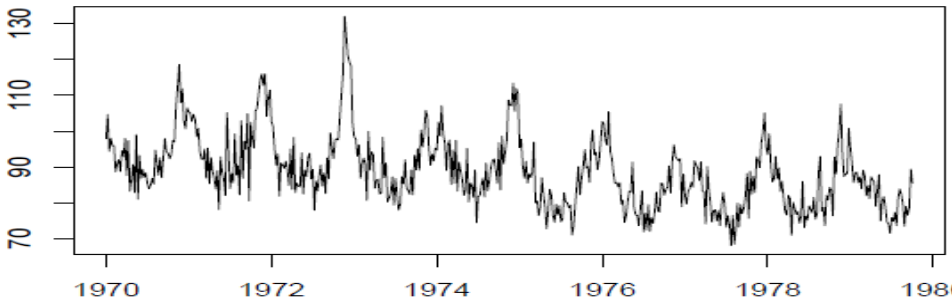
The price of chicken: monthly whole bird spot price, Georgia docks, US cents per pound, August 2001 to July 2016, with fitted linear trend line.

2. **Seasonal variation:** This component explains fluctuations within a year during the season, usually caused by climate and weather conditions, customs, traditional habits, etc.



Johnson & Johnson quarterly earnings per share, 84 quarters, 1960-I to 1980-IV.

3. **Cyclical variation:** This component describes the medium-term changes caused by circumstances, which repeat in cycles. The duration of a cycle extends over longer period of time.



Average weekly cardiovascular mortality in Los Angeles County. There are 508 six-day smoothed averages obtained by filtering daily values over the 10 year period 1970-1979.

4. **Irregular variation:** Irregular or random series are caused by unpredictable influences, which are not regular and also do not repeat in a particular pattern. These variations are caused by incidences such as war, strike, earthquake, flood, revolution, etc. There is no defined statistical technique for measuring random fluctuations in a time series.

Combination of Four Components: Considering the effects of these four components, two different types of models are generally used for a time series.

- **Additive Model:** $Y(t) = T(t) + S(t) + C(t) + I(t)$

Assumption: These four components are independent of each other.

- **Multiplicative Model:** $Y(t) = T(t) * S(t) * C(t) * I(t)$

Assumption: These four components of a time series are not necessarily independent and they can affect one another.

ARIMA Modeling:

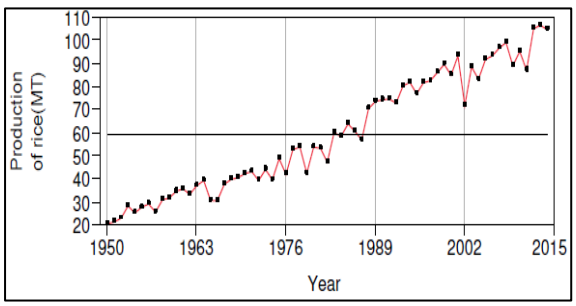
A time series is a set of measurements, y_t, taken over equally-spaced time periods (e.g., daily stock prices, monthly sales, quarterly GDP etc.). The three approaches for modeling y_t:

I. Fit y_t as a function of time (OLS regression)
II. Fit y_t as a function of its past values (Autoregressive)
III. Fit y_t as a function of random noise (Moving-Average)

When we focus on II and III approach, then the model is called ARMA models or ARIMA models when data is differenced--it brings AR and MA processes together!

ARIMA stands for Auto Regressive Integrated Moving Average (ARIMA), which is used for evaluating the future values through Box Jenkins methodology (1976). The Box-Jenkins procedure is concerned with fitting an ARIMA model to a given set of data. The main objective in fitting ARIMA model is to identify the stochastic process of the time series and predict the future values accurately. These methods have also been useful in many types of situations which involve the building of models for discrete time series and dynamic systems. However, this method is not good for lead times or for seasonal series with a large random component (Granger and Newbold, 1973). Originally ARIMA models have been studied extensively by George Box and Gwilym Jenkins during 1968 and their names have frequently been used synonymously with general ARIMA process applied to time series analysis, forecasting and control. The main stages in setting up a Box- Jenkins forecasting model are identification, estimating the parameters, diagnostic checking and forecasting. These can be understood from the study entitled "ARIMA Modelling for Forecasting of Rice Production: A case study of India" which have been carried out by Mahajan et.al. (2020);

i. **Identification of the characteristics (p, d, q) for the model:** The foremost step in the process of modeling is to check for the stationarity of the series by using appropriate tests like Augmented Dickey-Fuller (Dickey and Fuller, 1981 and Fuller, 1996), 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 through appropriate differencing because an ARIMA model is applicable only to a stationary time series.



Lag	AutoCorr	-.8-.6-.4-.2 0 .2 .4 .6 .8	Ljung-Box Q	p-Value	Lag	Partial	-.8-.6-.4-.2 0 .2 .4 .6 .8
0	1.0000		.	.	0	1.0000	
1	0.9230		57.9753	<.0001*	1	0.9230	
2	0.8802		111.532	<.0001*	2	0.1906	
3	0.8258		159.437	<.0001*	3	-0.0540	
4	0.7977		204.872	<.0001*	4	0.1324	
5	0.7668		247.552	<.0001*	5	0.0359	
6	0.7282		286.693	<.0001*	6	-0.0778	
7	0.6907		322.511	<.0001*	7	-0.0121	
8	0.6390		353.709	<.0001*	8	-0.1106	
9	0.5943		381.177	<.0001*	9	-0.0398	
10	0.5473		404.898	<.0001*	10	-0.0307	
11	0.5151		426.297	<.0001*	11	0.0447	
12	0.4723		444.629	<.0001*	12	-0.0589	
13	0.4534		461.845	<.0001*	13	0.1293	
14	0.4059		475.914	<.0001*	14	-0.1332	
15	0.3704		487.866	<.0001*	15	-0.0182	
16	0.3188		496.898	<.0001*	16	-0.0864	

Figure 1: Behaviour of the annual production of rice in India

Figure 2: Correlogram of annual production of rice in India

Table 1: ADF statistic for production of rice crop in India

Augmented Dickey-Fuller (ADF)	**Test-Statistic (-6.05)**	**P-value (0.98)**
Mean	59.36	
SD	25.46	

ii. **Estimation:** On the basis of identification of the parameters (p, d, q) the series is subjected to fitting of the appropriate ARIMA (p, d, q) model. The procedure for fitting the model involves transforming the series through appropriate differencing, if non-stationary, and then subjecting the differenced series to fitting. Choice of parameters is on the basis of significant ACFs and PACFs**.**

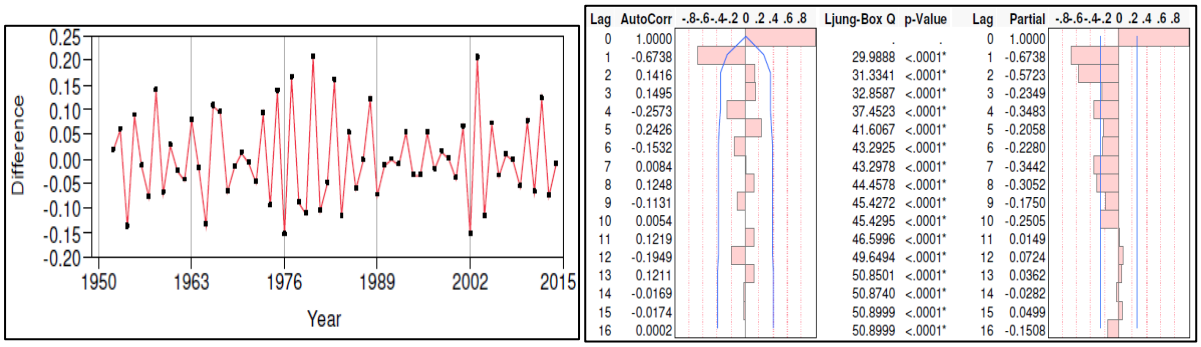


Figure 3: Behaviour of data after taking logarithm and differencing of order 2

Figure 4: Correlogram of data after taking logarithm and differencing of order 2

Table 2: ADF statistic for production of rice crop after taking logarithm and differencing of order 2

Augmented Dickey-Fuller (ADF)	**Test-Statistic (-17.41)**	**P-value (<0.0001)**
Mean	0.0110	
SD	0.0501	

Table 3: Different models for annual production of rice in India

Model	**Intercept**	**Significance of parameters/model**	**AIC**	**SBC**	**R^2**
ARIMA(0 2 2)	Yes	Significant	-216.40	-209.97	0.95
ARIMA(1 2 2)	Yes	Non-significant	-214.44	-205.87	0.95
ARIMA(2 2 2)	Yes	Non-significant	-212.45	-201.73	0.95
ARIMA(2 2 1)	No	Significant	-210.82	-204.39	0.94
ARIMA(2 0 0)	Yes	Significant	-205.21	-198.69	0.89

iii. **Diagnostics:** It has been done through Akaike Information Criteria (Akaike, 1979), Schwarz-Bayesian Information Criteria (Schwarz, 1978) and R2. AIC can be written as

$$AIC = -2 \log L + 2n,$$

where, L is the likelihood function and n is the number of hyper parameters estimated from the model. As an alternative to AIC, sometimes SBIC is also used which is given by

$$SBIC = -2\log L + n \log T,$$

where, T is total number of observations. Lower the value of these statistics better is the fitted model.

Table 4: Parameter estimates of ARIMA (0 2 2) for annual production of rice in India

Term	Lag	Estimate	Std error	t ratio	p-value	MAPE	-2logliklehood	Constant estimate
MA1	1	1.7833	0.1112	16.04	<.0001	1.85	-222.39	-0.0002
MA2	2	-0.7834	0.1053	-7.44	<.0001			
Intercept	0	-0.0002	0.0001	-2.27	0.0268			

Once the appropriate ARIMA model has been fitted, one can examine the goodness of fit by means of plotting the ACF and PACF of residuals of the fitted model. If most of the sample autocorrelation coefficients of the residuals are within the limits ± 1.96 / N; where, N is the number of observations upon which the model is based then the residuals are white noises indicating that the model is a good fit.

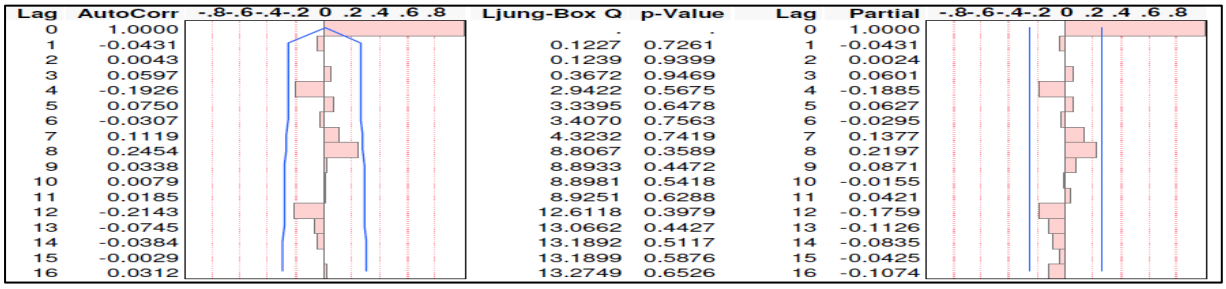


Figure 5: ACF and PACF plots of ARIMA (0 2 2) for the production of rice in India

The model that satisfies all the diagnostic checks is consider for forecasting.

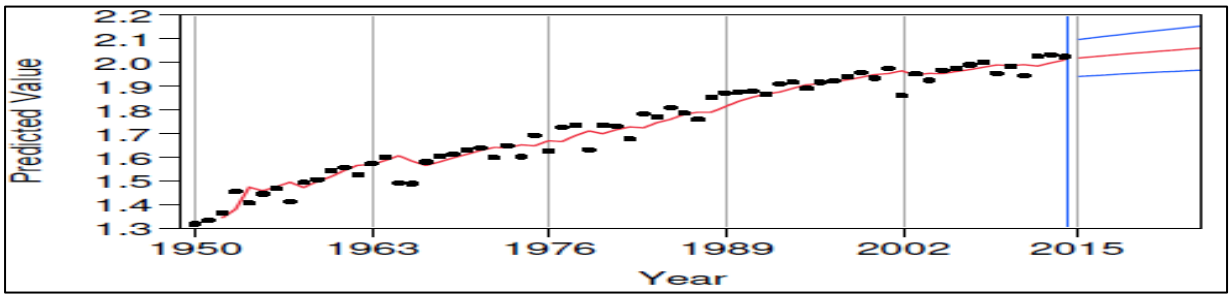


Figure 6: Forecasting graph of ARIMA (0 2 2) for the production of rice in India

There are several softwares which are good to follow Box-Jenkins methodology like SAS, Eviews, JMP, R etc.,

Among them JMP have been used in the analysis of above Example with the following steps:

1. Transform (e.g., differencing, log) the data to make it stationary by differencing, taking log etc.;

   - Time series is stationary if it has fixed mean and a constant variance.

2. Examine the sample autocorrelations and partial autocorrelations, and make an initial guess of small values of p and q on the basis of spikes of ACF and PACF plots;
3. Fit an ARIMA (p d q) model (p, d and q are the orders for AR, differencing and MA);
4. Perform the diagnostic checks to confirm that the model is consistent with the data;
   - Repeat 3 and 4 step to fit alternative models and compare them.
5. After comparison, the selected ARIMA (p d q) model is used for forecasting.

References:

Mahajan, S., Sharma, M., Peshin, R., Arya, V.M. and Kumar, B. (2018). Production performance of maize in India: An ARIMA Approach. *Journal of SAFE Agriculture*; **2**(1):24-27.

Nath, B., Dhakre, D.S. and Bhattacharya D. (2019). Forecasting wheat production in India: An ARIMA modelling approach**.** *Journal of Pharmacognosy and Phytochemistry*; **8**(1): 2158-2165.

Prabakaran, K. and Sivapragasam, C. (2014). Forecasting areas and production of rice in India using ARIMA model; *International Journal of Farm Sciences* **4**(1): 99-106.

Sharma, M., Jasrotia, N., Kumar, B., Bhat, A. and Mahajan, S. (2018). Modeling of Monthly Arrival of Rohu Fish using ARIMA in Jammu Region of J&K State. *Journal of Animal Research;* **8**(2): 259-262.

Mahajan, S., Sharma, M. and Gupta, A. (2020). ARIMA Modelling for Forecasting of Rice Production: A case study of India. *Agriculture Science Digest*; **40**(4): 404-407.

CHAPTER 10

PRINCIPAL COMPONENT ANALYSIS – WITH HANDS ON COMPUTER

Sunil Kumar
Assistant Professor, Department of Statistics, University of Jammu, Jammu
*sunilbhougal06@gmail.com

INTRODUCTION

Principal component analysis (PCA) is the oldest and best known technique of multivariate data analysis. It was first coined by Pearson (1901), and developed independently by Hotelling (1933). Like many other multivariate methods, it was not widely accepted nor used until the advent of electronic computers, but it is now well embedded in nearly every statistical software packages. **PCA** is one of the most frequently used multivariate data analysis methods that lets you investigate multidimensional datasets with quantitative variables. It is widely used in biostatistics, marketing, sociology, agriculture and many other fields.It is a **projection** method as it projects observations from a p-dimensional space with p variables to a k-dimensional space (where $k < p$) so as to conserve the maximum amount of information (information is measured here through the total variance of the dataset) from the initial dimensions. PCA **dimensions** are also called **axes** or **Factors**. If the information associated with the first 2 or 3 axes represents a sufficient percentage of the total variability of the scatter plot, the observations could be represented on a 2 or 3-dimensional chart, thus making interpretation much easier.

PCA can thus be considered as a **Data Mining** method as it allows to easily extract information from large datasets. There are several uses for it, including:

- The study and visualization of the correlations between variables to hopefully be able to limit the number of variables to be measured afterwards;
- Obtaining non-correlated factors which are linear combinations of the initial variables so as to use these factors in modelling methods such as linear regression, logistic regression or discriminant analysis.
- Visualizing observations in a 2- or 3-dimensional space in order to identify uniform or atypical groups of observations.

PCA has been called one of the most important results from applied linear algebra and perhaps its most common use is as the first step in trying to analyze large data sets. In general

terms, PCA uses a vector space transform to reduce the dimensionality of large data sets. Using mathematical projection, the original data set, which may have involved many variables, can often be interpreted in just a few variables (i.e. the principal components). The central idea of PCA is to reduce the dimensionality of a data set in which there are a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. This reduction is achieved by transforming to a new set of variables, the principal components, which are uncorrelated, and which are ordered so that the first few retain most of the variation present in all of the original variables. Computation of the principal components reduces to the solution of an eigenvalue-eigenvector problem for a positive-semi-definite symmetric matrix. Thus, the definition and computation of principal components are straightforward but, as will be seen, this apparently simple technique has a wide variety of different applications, as well as a number of different derivations. The central idea of PCA is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. This is achieved by transforming to a new set of variables, the principal components (PCs), which are uncorrelated, and which are ordered so that the first few retain most of the variation present in all of the original variables.

GOALS

The goals of PCA are to
1. extract the most important information from the data;
2. compress the size of the data by keeping only important information;
3. simplify the description of the data set; and
4. Analyze the structure of the observations and the variables.
5. Compress the data, by reducing the number of dimensions, without much loss of information.
6. This technique used in image compression

In order to analysis the data by Principal Component Analysis we have to be thorough in statistics and matrix algebra. So, we discuss on Statistics which looks at distribution measurements, how the data is spread out and also on Matrix Algebra by calculating eigenvectors and eigenvalues which is the fundamental principle to determine PCA.

Standard Deviation: Standard Deviation of a set of observations of a series is the positive square root of the arithmetic mean of the squares of all the deviations from the arithmetic mean.

Covariance: Covariance is a measure of the relationship between two random variables and to what extent, they change together.

Eigenvectors and Eigenvalues: Eigenvectors and eigenvalues are numbers and vectors associated to square matrices. Together they provide the eigen-decomposition of a matrix, which analyze the structure of a matrix. Even though the eigen-decomposition does not exist for all square matrices, it has a particularly simple expression for matrices such as correlation, covariance, or cross-product matrices. The eigen-decomposition of this type of matrices is important because it is used to find the maximum (or minimum) of functions involving these matrices. Specifically, PCA is obtained from the eigen-decomposition of a covariance or a correlation matrix.

EXAMPLE

Analysis is discussed by taking a simple example from Mishra et al. (2017). The steps are as follows:

Step 1: Get Some Data

In this simple example, we are going to use our own made-up data set. It's only got 2 dimensions, and the reason why we have chosen this is so that we can provide plots of the data to show what the PCA analysis is doing at each step. The data used is given in table 1

Table 1: Original data on the left, data with the subtracted mean taken on the right

Variable X	Variable Y	Deviation from mean for X	Deviation from mean for Y
2.5	2.4	0.69	0.49
0.5	0.7	-1.31	-1.21
2.2	2.9	0.39	0.99
1.9	2.2	0.09	0.29
3.1	3.0	1.29	1.09
2.3	2.7	0.49	0.79
2	1.6	0.19	-0.31
1	1.1	-0.81	-0.81
1.5	1.6	-0.31	-0.31
1.1	0.9	-0.71	-1.01

Step 2: For PCA to work properly, mean of the entire data was subtracted from each of the data dimensions. The mean subtracted is the average across each dimension. So, all the x values have x'(the mean of the x values of all the data points) subtracted, and all the y values have y' subtracted from them. This produces a data set whose mean is zero.

Step 3: Calculate the Covariance Matrix

This is done in exactly the same way as we have discussed previously. Since the data is 2 dimensional, the covariance matrix will be 2×2. There is no complicacy here, so the result will be

$$Cov = \begin{bmatrix} 0.616555556 & 0.615444444 \\ 0.615444444 & 0.716555558 \end{bmatrix}$$

As the non-diagonal elements in this covariance matrix are positive, it would be expected that both the x and y variable increase together.

Step 4: Calculate the Eigenvectors and Eigenvalues of the Covariance Matrix

Since the covariance matrix is square, the eigenvectors and eigenvalues for the matrix can be calculated. This step is most important, as it helps to get useful information about the data. By this we can determine the eigenvectors and eigenvalues are given below

$$Eigenvalue = \begin{bmatrix} 0.0490833989 \\ 1.28402771 \end{bmatrix}$$

$$Eigenvector = \begin{bmatrix} -0.735178656 & -0.677873399 \\ 0.677873399 & -0.735178656 \end{bmatrix}$$

It is important to notice that these eigenvectors are both unit eigenvectors as it is very important for PCA. Also, in most of the statistical software packages develop a unit eigenvectors.

Step 5: Choosing Components and Forming a Feature Vector

Here is the notion where data compression and reduced dimensionality comes into play. If we look at the eigenvectors and eigenvalues from the previous section, we will notice that the eigenvalues are quite different values. In fact, it turns out that the eigenvector with the highest eigenvalue is the principle component of the data set. The eigenvector with the largest eigenvalue was the one that pointed down the middle of the data. It is the most significant relationship between the data dimensions. In general, once eigenvectors are found from the covariance matrix, the next step is to order them by eigenvalue, highest to lowest. This gives you the components in order of significance. After that components with lesser significance were ignored. If the eigenvalues are small, the information will be much more accurate. If some the components were leaves out, the final data set will have fewer dimensions than the original. To be precise, if we originally have dimensions in our data, we would have calculated eigenvectors and eigenvalues, and then we choose only the first eigenvectors, then the final data set has only dimensions. What needs to be done now is we need to form a feature vector, which is just a fancy name for a matrix of vectors. This is constructed by taking the eigenvectors that we want to keep from the list of eigenvectors, and forming a matrix with these eigenvectors in the columns.

$$Feature\ Vector = (eig1, eig2, eig3 \dots eign)$$

Given with the example set of data, and the fact that here will be 2 eigenvectors, that means there will be two choices. A feature vector can be form with either of the eigenvectors

$$\begin{bmatrix} 0.677873399 & -0.735178656 \\ -0.735178656 & -0.677873399 \end{bmatrix}$$

Or, the smaller one can be chosen to leave out, less significant component and only have a single column vector

$$\begin{bmatrix} 0.677873399 \\ -0.735178656 \end{bmatrix}$$

The result of each of this step is used to develop a new data set in the next step.

Step 6: Deriving the New Data Set

Once the components (eigenvectors) chosen to keep in the data and a feature vector have been formed, then the transpose of the vector was calculated and multiply it on the left over of the original data set, transposed. Final Data = Row Feature Vector × Row Data Adjust Where, Row Feature Vector is the matrix with the eigenvectors in the columns transposed so that the eigenvectors are now in the rows, with the most significant eigenvector at the top, and Row Data Adjust is the mean-adjusted data transposed, i.e. the data items are in each column, with each row holding a separate dimension. The transpose of the feature vector and the data were calculated first. Then the final data set, with data items in columns, and dimensions along rows were determined. It will give the original data solely in terms of the vectors that was taken into consideration. The original data set had two axes, x and y, so the new data set was in terms of them. It is possible to express data in terms of any two axes that depends on the statistician. If these axes are perpendicular, then the expression is the most efficient. That's why eigenvectors are always perpendicular to each other. Here the original data have been changed in terms of the axes x and y to two newly developed eigenvectors. So, the new data set has reduced dimensionality, in terms of the vectors that satisfy the original data the most and leaving the unimportant eigenvectors. The new eigenvectors of the taken example are given below\

Table 2: The table of data by applying the PCA analysis using both eigenvectors

Variable X	Variable Y
-0.827970186	-0.175115307
1.77758033	0.142857227
-0.992197494	0.384374989
-0.274210416	0.130417207
-1.67580142	-0.209498461
-0.912949103	0.175282444
0.0991094375	-0.349824698
1.14457216	0.0464172582
0.438046137	0.0177646297
1.22382056	-0.162675287

To show this on the data, the final transformation with each of the possible feature vectors has to be done. The transpose of each of the result was carried out, to bring the data back to the nice table-like format. By this it is understandable that no information had been lost in this decomposition and there is a strong correlation newly developed data with the original one. The other transformation can be made by taking only the eigenvector with the largest eigenvalue. The table of data resulting from that is given below

Table 3: The data after transforming using only the most significant eigenvector

Transformed Data (Single eigenvector) for Variable X
-0.827970186
1.77758033
-0.992197494
-0.274210416
-1.67580142
-0.912949103
0.0991094375
1.14457216
0.438046137
1.22382056

As expected, it only has a single dimension. If we compare this data set with the one resulting from using both eigenvectors, we will notice that this data set is exactly the first column of the other. So, if we were to plot this data, it would be 1-dimensional, and would be points on a line in exactly the x positions of the points in the plot in Table 2. We have effectively thrown away the whole other axis, which is the other eigenvector. Basically we have transformed our data so that is expressed in terms of the patterns between them, where the patterns are the lines that most closely describe the relationships between the data. This is helpful because we have now classified our data point as a combination of the contributions from each of those lines. Initially we had the simple x and y axes. This is fine, but the x and y values of each data point don't really tell us exactly how that point relates to the rest of the data. Now, the values of the data points tell us exactly where (i.e. above/below) the trend lines the data point sits. In the case of the transformation using both eigenvectors, we have simply altered the data so that it is in terms of those eigenvectors instead of the usual axes (Table 3). But the single-eigenvector decomposition has removed the contribution due to the smaller eigenvector and left us with data that is only in terms of the other.

Geometrical Interpretation

• PCA projects the data along the directions where the data varies the most.

- These directions are determined by the eigenvectors of the covariance matrix corresponding to the largest eigenvalues.
- The magnitude of the eigenvalues corresponds to the variance of the data along the eigenvector directions.

CONCLUSION: One benefit of PCA is that we can examine the variances associated with the principle components. Often one finds that large variances associated with the first k < m principal components, and then a precipitous drop up. One can conclude that most interesting dynamics occur only in the first k dimensions. Both the strength and weakness of PCA is that it is a non-parametric analysis. There are no parameters to tweak and no coefficients to adjust based on user experience the answer are unique and independent of the user. This same strength can also be viewed as a weakness. If one knows a priori some features of the dynamics of a system, then it makes sense to incorporate these assumptions into a parametric algorithm or an algorithm with selected parameters. Thus, the appropriate parametric algorithm is to first convert the data to the appropriately centered polar coordinates and then compute PCA. Performing PCA is quite simple in practice. Organize a data set as an m×n matrix, where m is the number of measurement types and n is the number of trials. Subtract of the mean for each measurement type or row xi. Calculate the SVD or the eigenvectors of the co-variance. It was found that there were many interesting applications of PCA, out of which in day today life knowingly or unknowingly multivariate data analysis and image compression are being used alternatively.

References

Hotelling H. 1933. Analysis of a complex of statistical variables into principal components. J Educ Psychol. 25: 417-441.

Pearson K. 1901. On lines and planes of closest fit to systems of points in space. Philos Mag A. 6: 559-572.

Mishra, S., Sarkar, U., Taraphder, S., Datta, S., Swain, D., Saikhom, R., Panda, S. and Laishram, M. Multivariate Statistical Data Analysis – Principal Component Analysis (PCA). International Journal of Livestock Research, 7(5) (**2017**): 60-78.

CHAPTER 11

FACTOR ANALYSIS AND ITS APPLICATIONS

Jeevan Jyoti
University of Jammu

Factor analysis is a general name denoting a class of procedures primarily used for data reduction and summarisation. In marketing research, there may be a large number of variables, most of which are correlated and which must be reduced to a manageable level. Relationships among sets of many interrelated variables are examined and represented in terms of a few underlying factors. For example, the image of a fashion brand may be measured by asking participants to evaluate competing fashion brands on a series of items on a semantic differential scale or a Likert scale. These item evaluations may then be analysed to determine the **factors** underlying the image of a fashion brand.

In analysis of variance, multiple regression and discriminant analysis, one variable is considered the dependent or criterion variable and the others are considered independent or predictor variables. But no such distinction is made in factor analysis. Rather, factor analysis is an **interdependence technique** in that an entire set of interdependent relationshipsis examined.

Factor analysis is used in the following circumstances:

1 *To identify underlying dimensions, or factors, that explain the correlations among a set of variables.* For example, a set of lifestyle statements may be used to measure the psychographic profiles of consumers. These statements may then be factor analysed to identify the underlying psychographic factors, as illustrated in the opening example.

2 *To identify a new, smaller set of uncorrelated variables to replace the original set of correlated variables in subsequent multivariate analysis (regression or discriminant analysis).*

3. *To identify a smaller set of salient variables from a larger set for use in subsequent multivariate analysis.*

Factor analysis has numerous applications in marketing research.

For example:

• It can be used in market segmentation for identifying the underlying variables on which to group the customers. New car buyers might be grouped based on the relative emphasis they place on economy, convenience, performance, comfort and luxury. This might result

in five segments: economy seekers, convenience seekers, performance seekers, comfort seekers and luxury seekers.

• In product research, factor analysis can be employed to determine the brand attributes that influence consumer choice. Toothpaste brands might be evaluated in terms of protection against cavities, whiteness of teeth, taste, fresh breath and price.

• In advertising studies, factor analysis can be used to understand the media consumption habits of the target market. The users of frozen foods may be heavy viewers of horror films, play a lot of electronic games and listen to rock music.

• In pricing studies, factor analysis can be used to identify the characteristics of price-sensitive consumers. For example, these consumers might be methodical, economy minded and home centred.

Statistics Associated with Factor Analysis

The key statistics associated with factor analysis are as follows:

• *Bartlett's test of sphericity.* This is a test statistic used to examine the hypothesis that the variables are uncorrelated in the population. In other words, the population correlation matrix is an identity matrix; each variable correlates perfectly with itself ($r = 1$) but has no correlation with the other variables ($r = 0$).

• *Communality.* This is the amount of variance a variable shares with all the other variables being considered. It is also the proportion of variance explained by the common factors.

Correlation matrix. A correlation matrix is a lower triangular matrix showing the simple correlations, r, between all possible pairs of variables included in the analysis. The diagonal elements, which are all one, are usually omitted.

• *Eigenvalue.* The eigenvalue represents the total variance explained by each factor.

• *Factor loadings.* These are simple correlations between the variables and the factors.

• *Factor loading plot.* A factor loading plot is a plot of the original variables using the factor loadings as coordinates.

• *Factor matrix.* A factor matrix contains the factor loadings of all the variables on all the factors extracted.

• *Factor scores.* These are composite scores estimated for each participant on the derived factors.

• *Factor scores coefficient matrix.* This matrix contains the weights, or factor score coefficients,

used to combine the standardised variables to obtain factor scores.

• *Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy.* The KMO measure of sampling adequacy is an index used to examine the appropriateness of factor analysis. High values (between 0.5 and 1.0) indicate that factor analysis is appropriate. Values below 0.5 imply that factor analysis may not be appropriate.

• *Percentage of variance.* This is the percentage of the total variance attributed to each factor.

• *Residuals.* These are the differences between the observed correlations, as given in the input correlation matrix, and the reproduced correlations, as estimated from the factor matrix.

• *Scree plot.* A scree plot is a plot of the eigenvalues against the number of factors in order of extraction.

Factor Analysis Process

EFA is often recommended when researchers have no hypotheses about the nature of the underlying factor structure of their measure. Exploratory factor analysis has five basic decision points: (1) decide the number of factors, (2) choosing an extraction method, (3) choosing a rotation method (4) Factor Interpretation and (5)Model Interpretation

DECIDING THE NUMBER OF FACTORS The most common approach to deciding the number of factors is to generate a scree plot. The scree plot is a two dimensional graph with factors on the x-axis and eigenvalues on the y-axis. Eigenvalues are produced by a process called principal components analysis (PCA) and represent the variance accounted for by each underlying factor. They are not represented by percentages but scores that total to the number of items. A 12-item scale will theoretically have 12 possible underlying factors, each factor will have an eigenvalue that indicates the amount of variation in the items accounted for by each factor. If a the first factor has an eigenvalue of 3.0, it accounts for 25% of the variance (3/12=.25). The total of all the eigenvalues will be 12 if there are 12 items, so some factors will have smaller eigenvalues.

From the scree plot you can see that the first couple of factors account for most of the variance, then the remaining factors all have small eigenvalues. The term “scree” is taken from the word for the rubble at the bottom of a mountain. A researcher might examine this plot and decide there are 2 underlying factors and the remainder of factors are just “scree” or error variation. So, this approach to selecting the number of factors involves a certain amount

of subjective judgment. Another approach is called the Kaiser-Guttman rule and simply states that the number of factors are equal to the number of factors with eigenvalues greater than 1.0. I tend to recommend the scree plot approach because the Kaiser-Guttman approach seems to produce many factors.

FACTOR EXTRACTION Once the number of factors are decided the researcher runs another factor analysis to get the loadings for each of the factors. To do this, one has to decided which mathematical solution to use to find the loadings. There are about five basic extraction methods (1) PCA, which is the default in most packages. PCA assumes there is no measurement error and is considered not to be a true exploratory factor analysis; (2) maximum likelihood (a.k.a. canonical factoring); (3) alpha factoring, (4) image factoring, (5) principal axis factoring with iterated communalities (a.k.a. least squares). Without getting into the details of each of these, I think the best evidence supports the use of principal axis factoring and maximum likelihood approaches. I typically use the former. Gorsuch (1989) recommends the latter if only a few iterations are performed (not really possible in most packages). Snook and Gorsuch (1989) show that PCA can give poor estimates of the population loadings in small samples. With larger samples, most approaches will have similar results. The extraction method will produce factor loadings for every item on every extracted factor. Researchers hope their results will show what is called simple structure, with most items having a large loading on one factor but small loadings on other factors.

FACTOR ROTATION Once an initial solution is obtained, the loadings are rotated. Rotation is a way of maximizing high loadings and minimizing low loadings so that the simplest possible structure is achieved. There are two basic types of rotation: orthogonal and oblique. Orthogonal means the factors are assumed to be uncorrelated with one another. This is the default setting in all statistical packages but is rarely a logical assumption about factors in the social sciences. Not all researchers using EFA realize that orthogonal rotations imply the assumption that they probably would not consciously make. Oblique rotation derives factor loadings based on the assumption that the factors are correlated, and this is probably most likely the case for most measures. So, oblique rotation gives the correlation between the factors in addition to the loadings. Here are some common algorithms for orthogonal and oblique rotation.

- Orthogonal rotation: varimax, quartamax, equamax.
- Oblique rotation: oblimin, promax, direct quartimin

INTERPRETION OF FACTORS: Interpretation is facilitated by identifying the variables that have large loadings on the same factor. That factor can then be interpreted in terms of the variables that load high on it. Another useful aid in interpretation is to plot the variables, using the factor loadings as coordinates. Variables at the end of an axis are those that have high loadings on only that factor and hence describe the factor. Variables near the origin have small loadings on both the factors. Variables that are not near any of the axes are related to both the factors. If a factor cannot be clearly defined in terms of the original variables, it should be labelled as an undefined or a general factor.

DETERMINE THE MODEL FIT: The final step in factor analysis involves the determination of model fit. A basic assumption underlying factor analysis is that the observed correlation between variables can be attributed to common factors. Hence, the correlations between the variables can be deduced or reproduced from the estimated correlations between the variables and the factors. The differences between the observed correlations (as given in the input correlation matrix) and the reproduced correlations (as estimated from the factor matrix) can be examined to determine model fit. These differences are called *residuals*. If there are many large residuals, the factor model does not provide a good fit to the data and the model should be reconsidered.

References for Further Reading:

- Fabrigar, Leandre R.; Wegener, Duane T.; MacCallum, Robert C.; Strahan, Erin J. (1 January 1999). "Evaluating the use of exploratory factor analysis in psychological research" (PDF). *Psychological Methods*. **4** (3): 272–299.
- Faris, H., Gaterell, M., & Hutchinson, D. (2022). Investigating underlying factors of collaboration for construction projects in emerging economies using exploratory factor analysis. *International journal of construction management*, *22*(3), 514-526.
- Hair, J.F., Jr., R.E. Anderson, and R.L. Tatham. 1987. Multivariate data analysis. 2nd ed. Macmillan Publishing Company, New York, NY.
- Norris, Megan; Lecavalier, Luc (17 July 2009). "Evaluating the Use of Exploratory Factor Analysis in Developmental Disability Psychological Research". *Journal of Autism and Developmental Disorders*. **40** (1): 8–20.
- Stapleton, C. D. (1997). Basic Concepts in Exploratory Factor Analysis (EFA) as a Tool To Evaluate Score Validity: A Right-Brained Approach.

CHAPTER 12

K-MEAN CLUSTERING AND DECISION TREE USING SOFTWARES

Manish Sharma, M.I.J. Bhat, Sunali Mahajan and Arshid Bhat
Division of Statistics and CS, FBSc
Email:manshstat@gmail.com

In this chapter Univariate analysis, multivariate analysis and its techniques are discussed. Moreover, the various types of cluster analysis are hierarchical (Agglomerative clustering and Divisive) and non-hierarchical clustering with detailed description about K mean clustering algorithm is summarized using cluster analysis, its applications and objectives.

Introduction:

Univariate analysis is the simplest form of analyzing data. “Uni” means “one” so in Univariate analysis your data has only one variable. The Univariate data doesn’t deal with causes or relationships (unlike regression) and its major purpose is to describe, summarizes and finds patterns in the data. Sometimes the Univariate analysis can yield misleading results and in those cases multivariate analysis is more appropriate. Multivariate Analysis uses statistical techniques which allow us to focus and analyze more than two statistical variables at once. It is a collection of methods used when several measurements are made on an object in different samples. The measurements are referred to as variables and the objects are called units. It helps in summarizing data and reducing the chances of spurious results. Multivariate statistics is a subdivision of statistics encompassing the simultaneous observation and analysis of more than one outcome variable. Multivariate statistics concerns understanding the different aims and background of each of the different forms of multivariate analysis, and how they relate to each other. Multivariate analysis (MVA) is based on the principles of multivariate statistics. Typically, MVA is used to address the situations where multiple measurements are made on each experimental unit and the relations among these measurements and their structures are important. This can be complicated by the desire to include physics-based analysis to calculate the effects of variables for a hierarchical "system-of-systems". There are two main multivariate analysis methods (i) Dependence Analysis; it is used in predicting the dependency among variables like Multiple Regressions and Discriminant Analysis. And Interdependence Analysis; it is used in analyzing the relationships among variables or objects where none of them are dependent. like Factor Analysis, Cluster Analysis, PCA. Examples of Multivariate Data: Techniques of Multivariate

Analysis are Multiple Regression Analysis, Principal Component, Analysis, Cluster Analysis, Factor Analysis and Discriminant Analysis. Factor analysis is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors. For example, it is possible that variations in six observed variables mainly reflect the variations in two unobserved (underlying) variables. Factor analysis searches for such joint variations in response to unobserved latent variables. The observed variables are modelled as linear combinations of the potential factors plus "error" terms, hence factor analysis can be thought of as a special case of errors-in-variables models. Discriminant analysis is statistical technique used to classify observations into non-overlapping groups, based on scores on one or more quantitative predictor variables. For example, a doctor could perform a discriminant analysis to identify patients at high or low risk for stroke. The analysis might classify patients into high- or low-risk groups based on personal attributes (e.g., cholesterol level, body mass) and/or lifestyle behaviours (e.g., minutes of exercise per week, packs of cigarettes per day). Cluster analysis foundations rely on one of the most fundamental, simple and very often unnoticed ways (or methods) of understanding and learning, which is grouping "objects" into "similar" groups. Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). One very popular application of cluster analysis in business is market segmentation. Consider, customers are grouped into distinct clusters or market segments and each segment is targeted with different marketing mixes such as different promotional messages, different products, different prices, and different distribution channels. Other examples of clustering may be a clustering of products into different sub-groups based on attributes like price-elasticity, genres, etc. In this way, clustering compresses the entire data into a reduced set of sub-groups. So, clustering is a data reduction technique. The Main objective of clustering (i) Discover structures and patterns in high-dimensional data (ii) Group data with similar patterns together and it reduces the complexity and facilitates interpretation. The cluster analysis techniques are of main two types:

(i) **Hierarchical clustering:** Hierarchical clustering is basically an unsupervised clustering technique which involves creating clusters in a predefined order. The clusters are ordered in a top to bottom manner. In this type of clustering, similar clusters are grouped together and are arranged in a hierarchical manner. It can be further divided into two types namely agglomerative hierarchical clustering and

Divisive hierarchical clustering. In this clustering, we link the pairs of clusters all the data objects are there in the hierarchy.

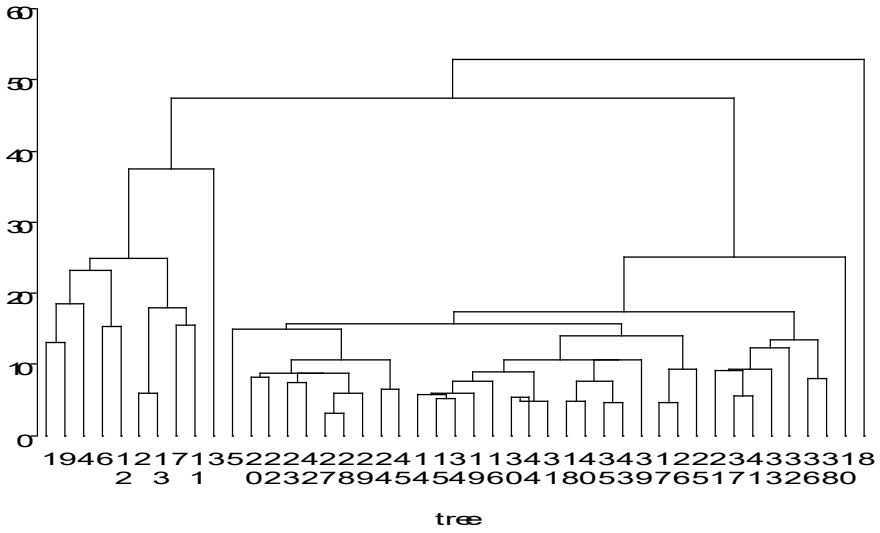


Fig.1: The graphically representation of various steps or stages of the clustering process is known as dendogram.

(ii) **Non-hierarchical clustering:** Non Hierarchical Clustering involves formation of new clusters by merging or splitting the clusters. It does not follow a tree like structure like hierarchical clustering. This technique groups the data in order to maximize or minimize some evaluation criteria. K means clustering is an effective way of non-hierarchical clustering. In this method the partitions are made such that non-overlapping groups having no hierarchical relationships between themselves.

The main Difference between Hierarchical Clustering and Non Hierarchical Clustering:

Hierarchical clustering	**Non Hierarchical clustering**
Hierarchical Clustering involves creating clusters in a predefined order from top to bottom	Non Hierarchical Clustering involves formation of new clusters by merging or splitting the clusters instead of following a hierarchical order.
It is considered less reliable than Non Hierarchical Clustering.	It is comparatively more reliable than Hierarchical Clustering.
It having less weight.	It having more weight

It is very problematic to apply this technique when we have data with high level of error.	It can work better then Hierarchical clustering even when error is there.
It is comparatively easier to read and understand.	The clusters are difficult to read and understand as compared to Hierarchical clustering.
It is relatively unstable techniques	It is a relatively stable technique.

Hierarchical clustering can be divided into two main types: Agglomerative and Divisive

Agglomerative clustering: It's also known as AGNES (Agglomerative Nesting). It works in a bottom-up manner. That is, each object is initially considered as a single-element cluster (leaf). At each step of the algorithm, the two clusters that are the most similar are combined into a new bigger cluster (nodes). This procedure is iterated until all points are member of just one single big cluster (root). The result is a tree which can be plotted as a dendrogram. It is good at identifying small clusters.

Divisive hierarchical clustering: It's also known as DIANA (Devise Analysis) and it works in a top-down manner. The algorithm is an inverse order of AGNES. It begins with the root, in which all objects are included in a single cluster. At each step of iteration, the most heterogeneous cluster is divided into two. The process is iterated until all objects are in their own cluster. Divisive hierarchical clustering is good at identifying large clusters. The most common methods of cluster agglomeration are: (i) Maximum or complete linkage clustering-It computes all pairwise dissimilarities between the elements in cluster 1 and the elements in cluster 2 and considers the largest value (i.e., maximum value) of these dissimilarities as the distance between the two clusters. It tends to produce more compact clusters. (ii) Minimum or single linkage clustering-It computes all pairwise dissimilarities between the elements in cluster 1 and the elements in cluster 2 and considers the smallest of these dissimilarities as a linkage criterion. It tends to produce long, "loose" clusters. Mean or average linkage clustering: It computes all pairwise dissimilarities between the elements in cluster 1 and the elements in cluster 2, and considers the average of these dissimilarities as the distance between the two clusters. (iii) Centroid linkage clustering:-It computes the dissimilarity between the centroid for cluster 1 (a mean vector of length p variables) and the centroid for cluster 2. (iv) Ward's minimum variance method-It minimizes the total within-cluster variance. At each step the pair of clusters with minimum between-cluster distance is merged.

Suppose values of p-variables be observed from n sample objects. The value of i-th object [i=1,2,…n] is denoted by a vector $Xi = [X_{i1}, X_{i2}, \ldots..X_{ip}]'$. d_{ij} be the distance between i-th

and j-th object, where $d_{ij} = \left[\sum_{k=1}^{p}\left|X_{ik} - X_{jk}\right|^{r}\right]^{1/r}$ is a special form of Minkowski metric. If r = 2, Euclidean distance $d_{ij} = \left[\sum_{k=1}^{p}\left|X_{ik} - X_{jk}\right|^{2}\right]^{1/2}$.If r=1, City block metric $d_{ij} = \sum_{k=1}^{p}\left|X_{ik} - X_{jk}\right|$ .Squared Euclidean distance $d_{ij} = \sum_{k=1}^{p}\left(X_{ik} - X_{jk}\right)^{2}$. Chebychev Distance Metric $d_{ij} = \max_{k}\left|X_{ik} - X_{jk}\right|$.

The different steps involved in the agglomerative hierarchical clustering for grouping N objects are

Step 1:Start with N clusters, each containing a single entity and an N x N symmetric matrix of distances (or similarity) D = $\{d_{ik}\}$

Step2: Search the distance matrix for the nearest (most similar) pair of clusters. Let the distance between "most similar" clusters U and V be d_{UV} .

Step3: Distance between two *p*-dimensional observations is

$\mathbf{x} = [x_1, x_2, \ldots, x_p]'$ **and** $\mathbf{y} = [y_1, y_2, \ldots, y_p]'$

$$d(\mathbf{x},\mathbf{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_p - y_p)^2} = \sqrt{(\mathbf{x}-\mathbf{y})'(\mathbf{x}-\mathbf{y})}$$

Methods of non-Hierarchical clustering:

K-Means is a non-hierarchical cluster analysis method that begins by determining the number of clusters desired. After the number of clusters is known, then the cluster process is carried out without following the hierarchical process. It is widely used in various fields because it is simple and easy to implement. K-Means clustering is very suitable for large data sizes because it has a higher speed than the hierarchical method. The results of clustering may depend on the order of data observations. The K-Means algorithm is a distance-based clustering method that divides data into a number of clusters and only works on numeric attributes.The most popular non-hierarchical procedure is the K-means methods. Here K-means describes an algorithm that assigns each item to the cluster having the nearest centroid (mean). In its simplest version the process is composed of three steps. (i) Partition the items into K-initial clusters. (ii) Proceed through the list of items, assigning an item to the cluster whose centroid (mean) is nearest. (Distance is usually computed using Euclidian distance). Recalculate the centroid for the cluster receiving the new item and for the cluster losing the item. (iii) Repeat step (ii) until no more reassignments take place Rather than starting with a partition of all items into K preliminary groups instep (i) one can specify K initial centroids

(seed points) and then proceed to step (ii). The final assignment of items to clusters will be to some extent dependent upon the initial partition or the initial selection of seed points. Experience suggests that most major changes in assignment occur with the first reallocation step. The K mean clustering is the algorithm to cluster n objects based on attributes into k partitions. Where k < n .It is similar to the expectation-maximization algorithm for mixtures of Gaussians in that they both attempt to find the centres of natural clusters in the data. It assumes that that the object attributes form a vector. It is the first thing that practitioners apply when solving clustering tasks to get an idea of the structure of the dataset. An algorithm for partitioning (or clustering) N data points into K disjoint subsets S_j containing data points so as to minimize the sum of squares criterion $J = \sum_{j=1}^{K} \sum_{n \in S_j} |x_n - \mu_j|^2$ Where x_n is a vector representing the n^{th} data points and u_j is the geometric centroid of the data points in S_j. Simply speaking k-means clustering is an algorithm to classify or to group the objects based on attributes or features into k number of group. K is positive integer number. The group is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid.

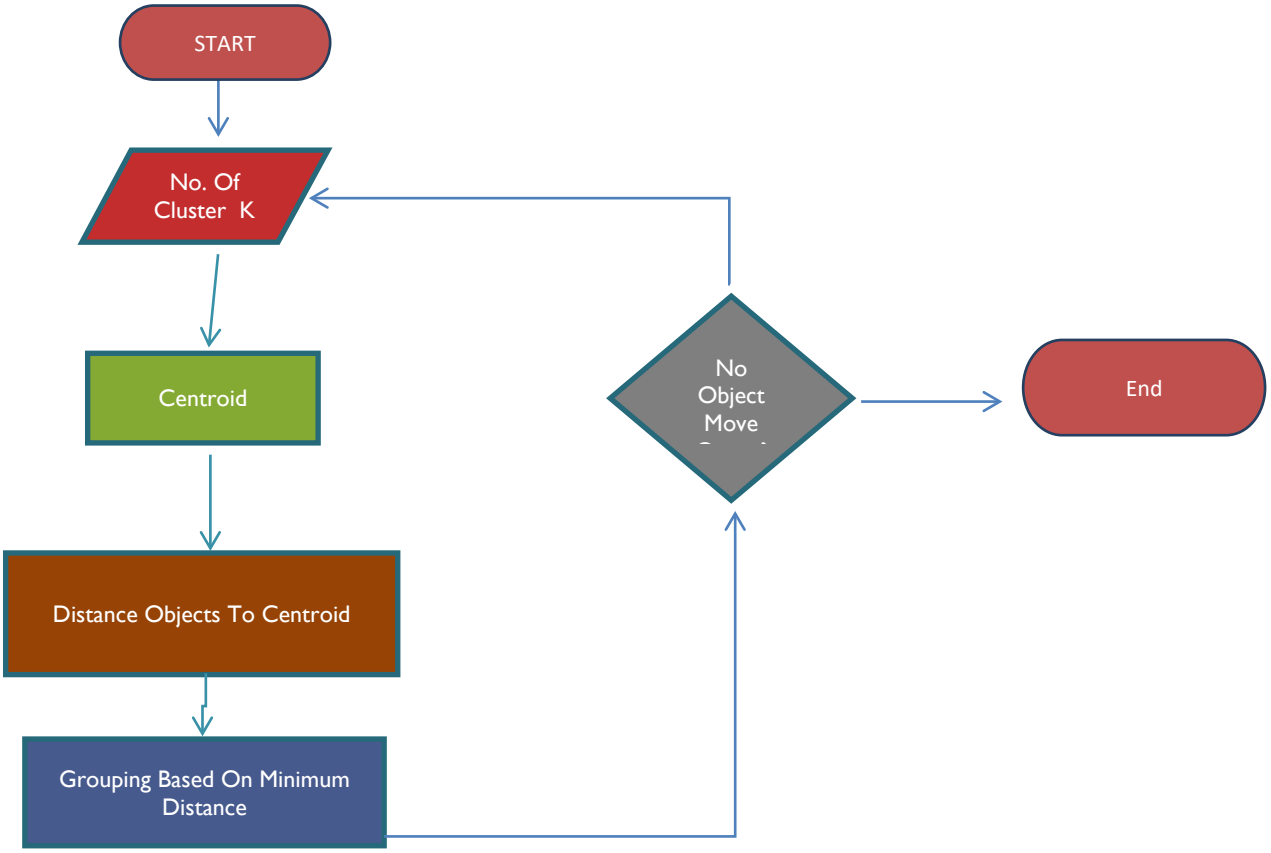


Figure: Flow chart for K-mean clustering.

The Main steps of Algorithm are as

Step-01: Choose the number of clusters K.

Step-02: Randomly select any K data points as cluster centres. Select cluster centers in such a way that they are as farther as possible from each other.

Step-03: Calculate the distance between each data point and each cluster centre. The distance may be calculated either by using given distance function or by using Euclidean distances formula.

Step-04: Assign each data point to some cluster. A data point is assigned to that cluster whose centre is nearest to that data point.

Step-05: Re-compute the centre of newly formed clusters. The centre of a cluster is computed by taking mean of all the data points contained in that cluster.

Step-06: Keep repeating the procedure from Step-03 to Step-05 until any of the following stopping criteria is met. Centre of newly formed clusters do not change, data points remain present in the same cluster, and maximum number of iterations are reached.

Problem-01: Cluster the following eight points (with (x, y) representing locations) into three clusters:

A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A7(1, 2), A8(4, 9)

Let the Initial cluster centers are: A1(2, 10), A4(5, 8) and A7(1, 2).

NOTE: The distance function between two points a = (x1, y1) and b = (x2, y2) is defined as-

$P(a, b) = |x2 – x1| + |y2 – y1|$

Use K-Means Algorithm to find the three cluster centers after the second iteration.

Solution- Iteration-01:

(i) Calculating Distance Between A1(2, 10) and C1(2, 10)-

$P(A1, C1) = |x2 – x1| + |y2 – y1|$

$= |2 – 2| + |10 – 10| = 0$

(ii) Calculating Distance Between A1(2, 10) and C2(5, 8)-

$P(A1, C2) = |x2 – x1| + |y2 – y1|$

$= |5 – 2| + |8 – 10| = 3 + 2 = 5$

(iii) Calculating Distance Between A1(2, 10) and C3(1, 2)-

$P(A1, C3) = |x2 – x1| + |y2 – y1|$

$= |1 – 2| + |2 – 10| = 1 + 8 = 9$

Given Points	Distance from center (2, 10) of Cluster-01	Distance from center (5, 8) of Cluster-02	Distance from center (1, 2) of Cluster-03	Point belongs to Cluster
A1(2, 10)	0	5	9	C1
A2(2, 5)	5	6	4	C3
A3(8, 4)	12	7	9	C2
A4(5, 8)	5	0	10	C2
A5(7, 5)	10	5	9	C2
A6(6, 4)	10	5	7	C2
A7(1, 2)	9	10	0	C3
A8(4, 9)	3	2	10	C2

New clusters are-

Cluster-01: A1 (2, 10)

Cluster-02:A3 (8, 4); A4 (5, 8); A5 (7, 5); A6 (6, 4); A8 (4, 9)

Cluster-03: A2 (2, 5) and A7 (1, 2)

The new cluster center is computed by taking mean of all the points contained in that cluster.

Cluster-01: There is only one point A1 (2, 10) in Cluster-01.So, cluster center remains the same.

Cluster-02: Center of Cluster-02

= ((8 + 5 + 7 + 6 + 4)/5, (4 + 8 + 5 + 4 + 9)/5) = (6, 6)

Cluster-03: Center of Cluster-03

= ((2 + 1)/2, (5 + 2)/2) = (1.5, 3.5).This is completion of Iteration-01.

Iteration-02: We calculate the distance of each point from each of the center of the three clusters. The distance is calculated by using the given distance function.

Given Points	Distance from center (2, 10) of Cluster-01	Distance from center (6, 6) of Cluster-02	Distance from center (1.5, 3.5) of Cluster-03	Point belongs to Cluster

A1(2, 10)	0	8	7	C1
A2(2, 5)	5	5	2	C3
A3(8, 4)	12	4	7	C2
A4(5, 8)	5	3	8	C2
A5(7, 5)	10	2	7	C2
A6(6, 4)	10	2	5	C2
A7(1, 2)	9	9	2	C3
A8(4, 9)	3	5	8	C1

New clusters are- **Cluster-01:** A1(2, 10) and A8(4, 9)

Cluster-02: A3(8, 4); A4(5, 8); A5(7, 5) and A6(6, 4)

Cluster-03: A2(2, 5) and A7(1, 2)

Now, we re-compute the new cluster clusters. The new cluster center is computed by taking mean of all the points contained in that cluster.

For Cluster-01: Center of Cluster-01

= ((2 + 4)/2, (10 + 9)/2)= (3, 9.5)

For Cluster-02: Center of Cluster-02

= ((8 + 5 + 7 + 6)/4, (4 + 8 + 5 + 4)/4)

= (6.5, 5.25)

For Cluster-03: Center of Cluster-03

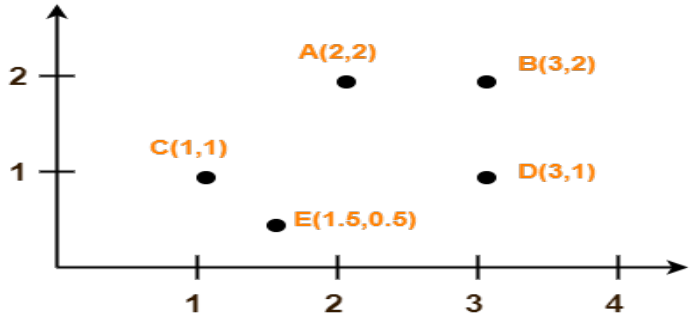
= ((2 + 1)/2, (5 + 2)/2)

= (1.5, 3.5). This is completion of Iteration-02.

After second iteration, the center of the three clusters are-

C1(3, 9.5) C2(6.5, 5.25) and C3(1.5, 3.5)

Problem-02: Use K-Means Algorithm to create two clusters-



Solution : Assume A(2, 2) and C(1, 1) are centers of the two clusters.

Iteration-01: Calculating Distance Between A(2, 2) and C1(2, 2)-

$P(A, C1) = \text{sqrt}\ [\ (x2 - x1)^2 + (y2 - y1)^2\ ]$

$= \text{sqrt}\ [\ (2 - 2)^2 + (2 - 2)^2\ ]$

$= \text{sqrt}\ [\ 0 + 0\ ] = 0$

Calculating Distance Between A(2, 2) and C2(1, 1)-

$P(A, C2) = \text{sqrt}\ [\ (x2 - x1)^2 + (y2 - y1)^2\ ]$

$= \text{sqrt}\ [\ (1 - 2)^2 + (1 - 2)^2\ ]$

$= \text{sqrt}\ [\ 1 + 1\ ] = \text{sqrt}\ [\ 2\ ] = 1.41$

In the similar manner, we calculate the distance of other points from each of the center of the two clusters.

Given Points	Distance from center (2, 2) of Cluster-01	Distance from center (1, 1) of Cluster-02	Point belongs to Cluster
A(2, 2)	0	1.41	C1
B(3, 2)	1	2.24	C1
C(1, 1)	1.41	0	C2
D(3, 1)	1.41	2	C1
E(1.5, 0.5)	1.58	0.71	C2

Cluster-01: A(2, 2); B(3, 2); E(1.5, 0.5) and D(3, 1)

Cluster-02: C(1, 1) and E(1.5, 0.5)

We re-compute the new cluster clusters. The new cluster center is computed by taking mean of all the points contained in that cluster.

For Cluster-01: Center of Cluster-01

= ((2 + 3 + 3)/3, (2 + 2 + 1)/3)

= (2.67, 1.67)

For Cluster-02: Center of Cluster-02

= ((1 + 1.5)/2, (1 + 0.5)/2)

= (1.25, 0.75)

This is completion of Iteration-01.Next, we go to iteration-02, iteration-03 and so on until the centres do not change anymore. The **strength** of K- algorithm is a classic algorithm for solving clustering problems because its algorithm is relatively simple and fast. The K-Means algorithm gives relatively good results in convex clusters. The K-Means algorithm has a fairly high accuracy of object size, so this algorithm is relatively more scalable and efficient for processing large amounts of data. The K-Means algorithm has no effect on the order of objects. The **Weakness** of K means clustering when the numbers of data are not so many, initial grouping will determine the cluster significantly. The number of cluster, k, must be determined beforehand. Its advantage is that it doesn't yield the same result with each run, since the resulting clusters depend on the initial random assignment. We never know the real cluster, using the same data, because it is imputed in a different order it may produce different cluster if the number of data is few. It is sensitive to initial condition. Different initial condition may produce different result of cluster. The algorithm may be trapped in the local optimum. The **Applications of K mean clustering** is relatively efficient and fast. It computes result at **O (tkn**), where n is the number of objects or points, k is the number of clusters and t us the number of iterations. K-means clustering can be applied to machine learning or data mining. Used on acoustic data in speech understanding to convert wave forms into one of the k categories Also used for choosing palettes on old fashioned graphical devices and image quantization. Thus the **K**-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K mean clustering is the fastest and most efficient algorithm to categorize data points into groups even when very little information is available about data. The steps involved in the analysis of cluster through SAS and SPSS are as

Programme for Cluster analysis through SAS

Data initial;

Input Food Calories Protein Fat Calcium Iron ;

Cards;

proc cluster simple noeign method = centroid RMSSTD RSQURE

nonorm out = tree;

Id food;

var Calories Protein Fat Calcium Iron;

run;

proc tree data = tree out= clus3 nclusters=**3**;

Id food;

Copy Calories Protein Fat Calcium Iron;

proc sort;

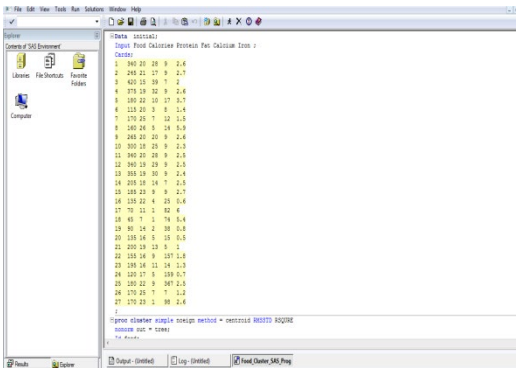
by cluster;

proc print;

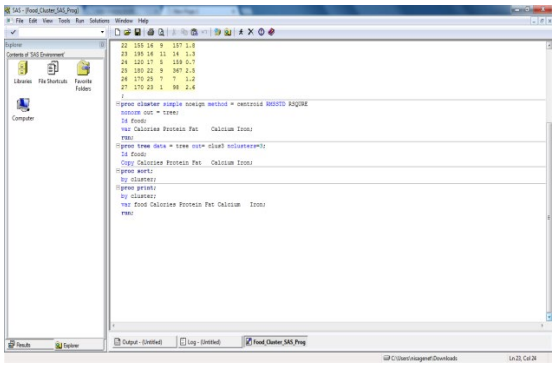
by cluster;

var food Calories Protein Fat Calcium Iron;

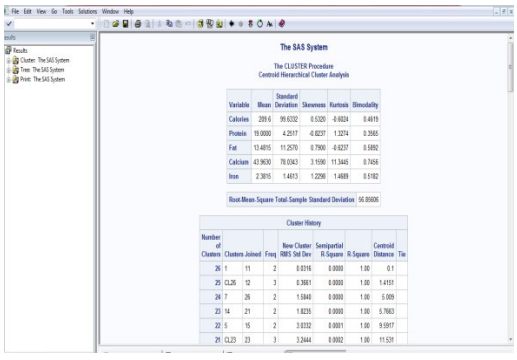
run;



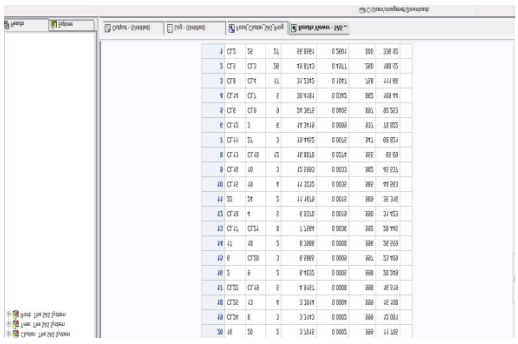
Step1: The input of data in the command window



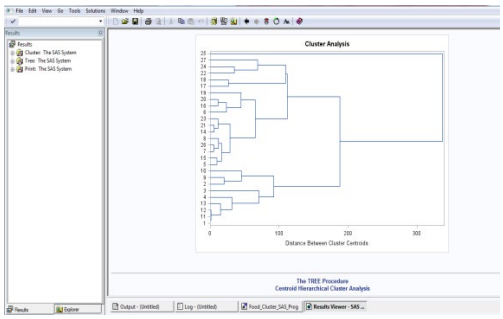
Step 2: The commands in the SAS window



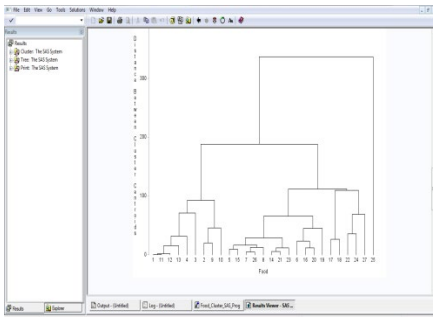
Step 3(a): The output through SAS.



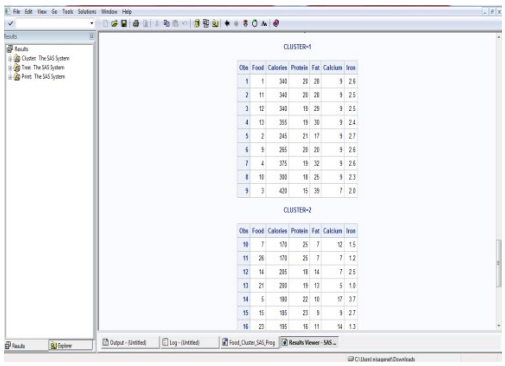
Step 3(b): The output through SAS.



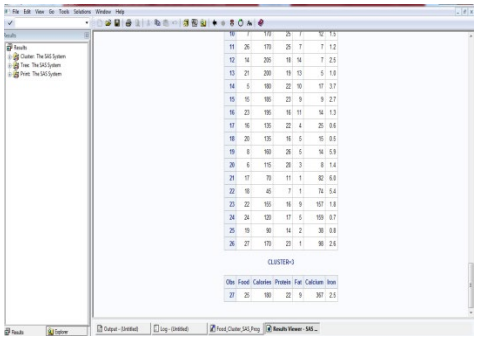
Step 3(c): The output through SAS.



Step 3(d): The output through SAS.



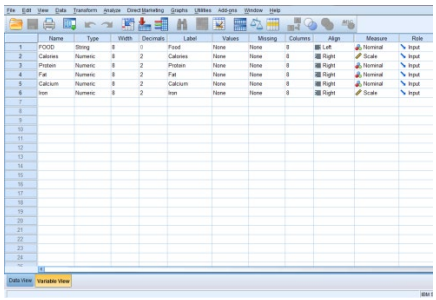
Step 3(e): The output through SAS.



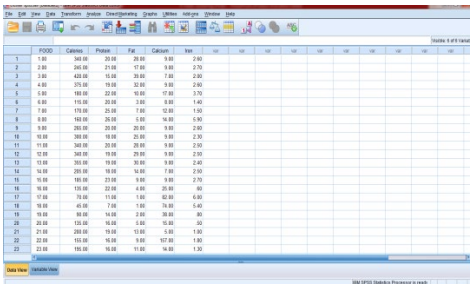
Step 3(f): The output through SAS.

Fig. 2: The screenshots of steps of cluster analysis output through SAS.

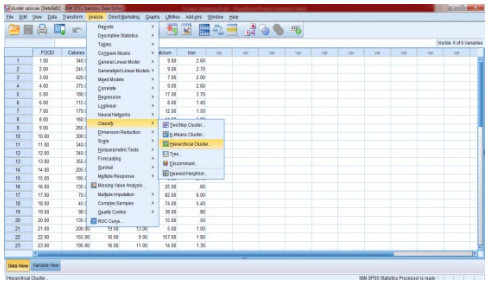
Cluster analysis through SPSS



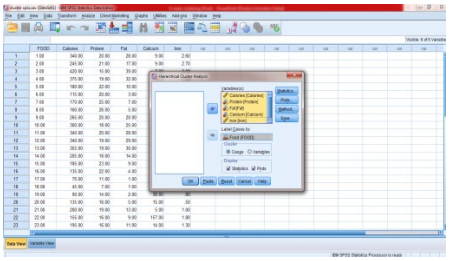
Step1: The variable input in variable view window



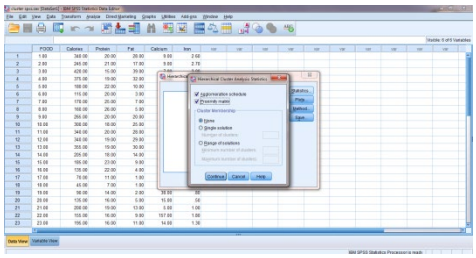
Step 2: The input of the data



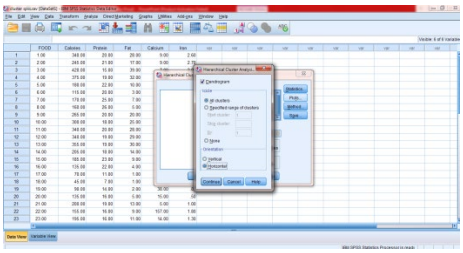
Step3: The selection of the command from Drop Drag menu.



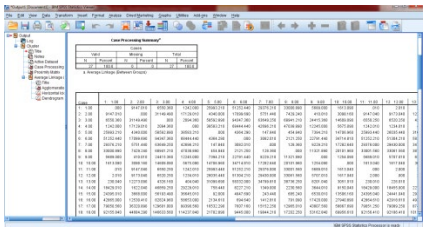
Step 4: Transferring of the variables in dialogue box



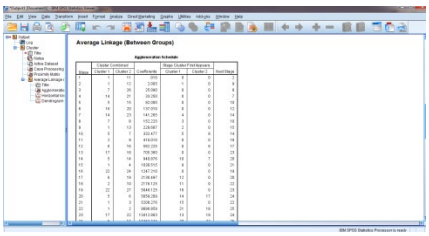
Step5: The selection of the command from sub dialogue box.



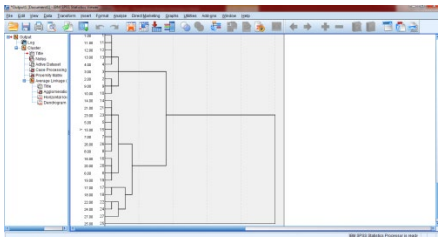
Step 6: selection of the commands in sub dialogue box



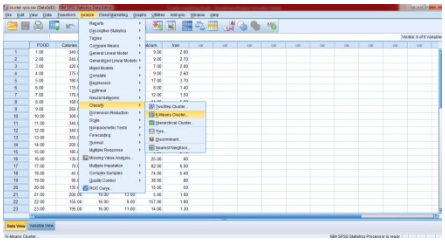
Step7: The output with matrix.



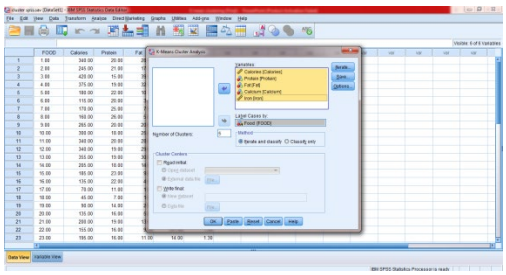
Step 8: The output related to merger of clusters.



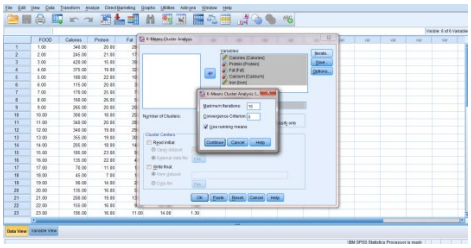
Step9: The dendogram ouput.



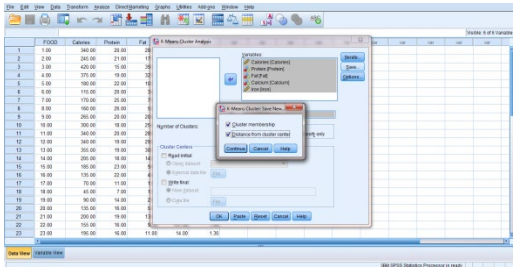
Step 10: The selection of command related K-Mean



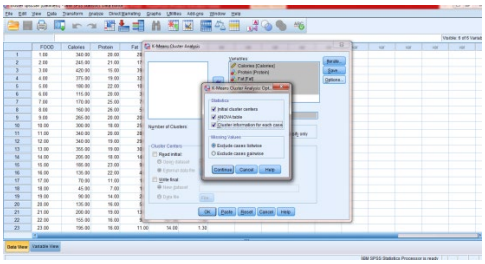
Step 11: The selection of the command from dialogue box.



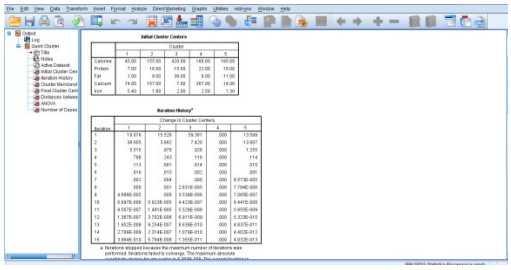
Step 12: Selection of the commands in sub dialogue box



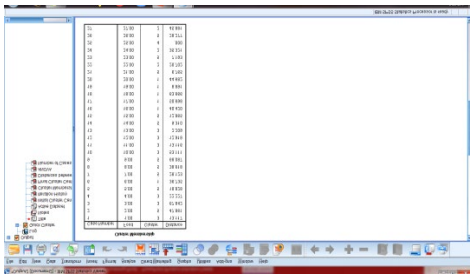
Step 13: The selection of the command from dialogue box.



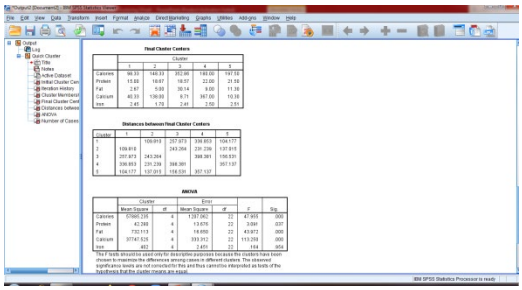
Step 14: Selection of the commands in sub dialogue box



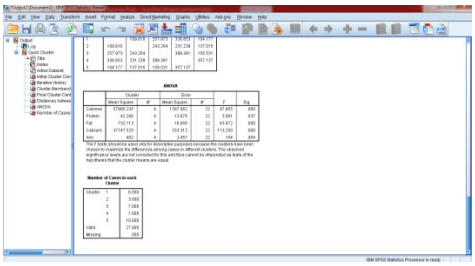
Step15: The output with matrix.



Step 16: The output related to merger of clusters.



Step17: The output.



Step 18: The output related information of clusters.

Fig. 3: The screenshots of steps of cluster analysis output through SPSS.

Reference:

- Agarwal, N., Agarwal. K., (2012). An Improved K-Means Clustering Algorithm for Data Mining, Lambert Academic Publishing.
- Bagadi. R.C., (2020). Computation of Optimal Number of Clusters in the K-Means Clustering Algorithm Independently published.
- Nandal. P., (2021). Optimizing Web Search Results for Image. K-means Clustering Algorithm, Grin Verlag.
- Patel. S., (2019). K-means Clustering Algorithm: Implementation and Critical Analysis, Scholars' Press.
- Wong. M.A., (2018). Asymptotic Properties of K-Means Clustering Algorithm as a Density Estimation Procedure, Forgotten Books.
- Wu. J., (2012). Advances in K-means Clustering, Berlin-Heidelberg, Springer.

CHAPTER 13

STATISTICAL AND BIOMETRICAL TECHNIQUES IN AGRICULTURE

BUPESH KUMAR
Associate Prof. Division of Plant Breeding and Genetics, FoA, SKUAST-Jammu, Main Campus, Chatha-180009

Branch of genetics which utilizes various statistical concepts and procedures to study the biological problems is known as biometrical genetics or biometrics. Biometrics and statistics have become integral in deciphering the mechanism involved in the transmission of traits from one generation to the other in plant breeding experiments. Successful breeding for crop improvement depends on scientific field experiments but, the moment such biological populations are subjected to field experimentation, some external non-heritable agencies come into play thereby, masking real differences among the populations. Two kinds of errors have been encountered under field experimentation viz., 1. Controllable and (ii) non-controllable. Further, data recorded in an experiment is so voluminous that it is difficult to arrive at meaningful conclusion with respect to genetic information. In addition genotype x environment interaction under field experiments is also to be taken into consideration for authentic information/data. Controllable errors can be scaled down by adopting appropriate field designs but the non-controllable or the chance errors have to be minimized by applying appropriate statistical methods. Therefore, good knowledge of statistical methods and their appropriate application helps in minimizing all possible sources of error. Now a days number of statistical and biometrical techniques are available with a plant breeder to employ them suitably for his crop improvement programmes. However, choice 'of an appropriate design/model or a statistical/biometrical technique is perhaps the most delicate and critical job because a wrong choice may lead to wastage of time and resources.

Statistical methods are utilized to obtain, assemble, classify and to interpret 'voluminous quantitative data that might have been influenced by many external factors. These methods are basically concerned with measuring and accounting for random variation departure from true value, so as to detect, identify and determine whether observed differences among varieties/treatments are real or not due to various kinds of errors. Various

parameters and mating designs which are commonly used in plant breeding experiments are detailed as follows:

❖ **Measures of variation**

✓ **Range:** The lower and higher value of character determines its range, which is expressed as follows:

$$\text{Range} = \text{highest value} - \text{lowest value}$$

✓ **Mean:** The mean is calculated by the following formula:

$$\text{Mean} = \frac{\sum X_i}{N}$$

Where,

$\sum X_i$ = Summation of all the observation

N = Total number of observations

✓ **Coefficient of variation:** It is a standardized, unitless measure which allows to compare variability between disparate groups and characteristics. It is also known as the relative standard deviation (RSD).

$$CV = \frac{\text{Standard deviation}}{\text{Mean}}$$

✓ **Genetic parameters**: The genotypic and phenotypic variances can be estimated by using the formula given by Cochran and Cox (1957).

$$\text{Genotypic variance } (\sigma^2_g) = \frac{\text{MSS due to genotypes} - \text{MSS due to error}}{\text{Number of replications}}$$

$$\text{Phenotypic variance } (\sigma^2_p) = \text{Genotypic variance } (\sigma^2_g) + \text{Error variance } (\sigma^2_e)$$

✓ **Genotypic and phenotypic co-efficient of variation**: GCV and PCV can be estimated using the formula of Burton and Devane(1953).

$$\text{Genotypic coefficient of variation (GCV)} = \frac{\text{Genotypic standard deviation}}{\text{Mean}} \times 100$$

$$\text{Phenotypic coefficient of variation (PCV)} = \frac{\text{Phenotypic standard deviation}}{\text{Mean}} \times 100$$

✓ **Heritability**

Heritability in the broad sense refers to the proportion of genotypic variance to the total observed variance in the total population. Heritability (h^2) in the broad sense can be calculated according to the formula given by Allard (1960).

$$h^2(bs) = \frac{\sigma^2 g}{\sigma^2 p}$$

Where,

$h^2(bs)$ = heritability in broad sense

$\sigma^2 g$ = genotypic variance

$\sigma^2 p$ = phenotypic variance $(\sigma^2 g) + (\sigma^2 e)$

$\sigma^2 e$ = environmental variance

Estimates of heritability can be categorized as:

Low : 0-30%.

Moderate : 30-60 %

High : 60% and above

✓ **Genetic advance**

Genetic advance for each trait can be estimated by using the formula of Johnson *et al.* (1955).

$$GA = h^2 bs \times \sigma_p \times K$$

Where,

GA : expected genetic gain

h^2bs: heritability estimate in broad sense

σ_p : phenotypic standard deviation of the trait

K : selection differential, the value of which is 2.06 at 5% selection intensity

Further, the genetic advance as per cent of can be computed by using the following formula:

$$\text{Genetic advance as percent of mean} = \frac{GA}{\text{Grand mean}} \times 100$$

Genetic advance as per cent mean can be categorized as

Low : 0 - 10%

Moderate : 10.1 - 20%

High : >20.1%

❖ **Biometrical techniques in Agriculture**

Statistical procedures that are used in the study of biometrical genetics are known as biometrical techniques and these techniques are exploited in plant breeding as follows:

ASSESSMENT OF POLYGENIC VARIATION

✓ **METEROGLYPH ANALYSIS**

It is a semigraphic method for assessing the pattern of morphological variation in a large number of germplasm lines taken at a time. It was developed by Andreson (1957) to investigate the pattern of morphological variation in crop species. Salient characteristics of this technique are:

- ❖ Analysis is based on first order statistics and, therefore, results are statistically more reliable and robust.
- ❖ Analysis is possible from both replicated and non replicated data.
- ❖ Variability pattern is depicted by glyph on the graph.

✓ **D^2 STATISTICS**

In plant breeding, genetic diversity plays an important role. Genetic diversity arises due to geographical separation or due to genetic barriers to cross ability. This has been observed in fescue, maize, alfalfa, cotton and several other crops. **D^2 statistics** technique was developed by P.C. Mahalanobis in 1928. Rao (1952) suggested the application of this technique for the assessment of genetic diversity in plant breeding. This is one of the potent techniques of measuring genetic divergence. Now this technique is extensively used in plant breeding and genetics for the study of genetic divergence in the various breeding material. Salient characteristics of this technique are:

- ❖ It is a numerical approach which is based on second order statistics.
- ❖ Analysis is more difficult than metroglyph analysis.
- ❖ Analysis is possible from replicated data only.
- ❖ Genetic diversity is depicted by cluster diagram.

DETERMINATION OF YIELD COMPONENTS

✓ **CORRELATION**

Correlation (r) is a statistical measure which is used to find out the degree and direction of relationship between two or more variables. A + value of r indicates that the changes of two variables are in the same direction, i.e., high values of one variable are associated with high values of other and *vice -versa.* When r is negative, the movements are in opposite directions, *i.e.,* high values of one variable are associated with low values of other. The salient features of correlation coefficient are.

- ❖ It is independent of the unit of measurement.
- ❖ Its value lies between – 1 and 1.
- ❖ It measures the degree and direction of association between two or more variable.

Correlation coefficient analysis measures the mutual relationship between various plant characters and determines the component characters on which selection can be based for genetic improvement in yield.

- ✓ **PATH ANALYSIS**

Path coefficient analysis is a standardized partial regression coefficient which measures the direct and indirect contribution of various independent traits on a dependent trait. It was originally developed by Wright in 1921, but it was first used for plant selection by Dewey and Lu in 1959. Path analysis reveals whether the association of these traits with yield is due to their direct effect on yield or is a consequence of their indirect effects via other component characters. Salient features of path coefficient analysis are:

- ❖ It measures the cause of association between two variables.
- ❖ Analysis is based on all possible simple correlations among various characters.
- ❖ Provides information about direct and indirect effects of independent variable on dependent variable.
- ❖ Helps in determining yield contributing characters and thus is useful in indirect selection.

- ✓ **DISCRIMINANT FUNCTION ANALYSIS**

In this technique the desirable genotypes are discriminated from the undesirable ones based on the combination of various characters. The use of this technique for plant selection was first proposed by Smith in 1936. Salient features of this technique are:

- ❖ Measures the efficiency of various character combinations in selection. Selection index leads to simultaneous manipulation of several traits for genetic improvement of economic yield.
- ❖ Provides information on yield components and thus aids indirect selection for genetic improvement of yield.

- ❖ Analysis involves variances and co variances

ASSESSMENT OF GENE ACTION

- ✓ **DIALLEL ANALYSIS**

Diallel cross refers to all possible crosses among the n lines and the analysis of such a set of crosses is known as diallel analysis.

Plan of Crossing for a Diallel Design

Parents	1	2	3	4	5
1	*	X	X	X	X
2	+	*	X	X	X
3	+	+	*	X	X
4	+	+	+	*	X
5	+	+	+	+	*

Where, X, +, and* = direct crosses, reciprocals, and parents respectively

TYPES OF DIALLEL CROSS: Depending upon number of crosses to be made, diallel cross is of two types

FULL DIALLEL: In a full diallel, each parent is used as male and female for each mating. Salient features of full diallel are given below.

- ❖ Total number of single crosses in a full diallel is equal to p(p-1), where p is the number of parents used.
- ❖ Full diallel is used when (a) reciprocal differences are significant, and (b) parents do not have male sterility or self incompatibility
- ❖ Full diallel permits estimation of maternal effects.
- ❖ Each parent is used as male and female in the mating.

There are two methods for evaluation of full diallel cross *i.e.,* with parents and without parents

- ✓ **Full dialllel with parents (F_1 s, Reciprocals and Parents):** It includes both way crosses and parents.
- ✓ **Full Diallel without Parents (F_1s, and Reciprocal):** It includes all possible single crosses made among p parents in both the direction *i.e.,* direct and reciprocal crosses.

HALF DIALLEL: In this design all possible crosses among the selected parents are made in one direction only *i.e.,* direct crosses. Salient features of half diallel are briefly presented below:

- ❖ Each parent is used either as male or as female in the mating.
- ❖ Number of single crosses required is equal to p(p-1), where p is the number of parents used.
- ❖ It is used when reciprocal differences are not significant.
- ❖ It can be used when parents have male sterility or self incompatibility.

It can be evaluated in two ways *i.e.,* with parents and without parents as given below.

- ✓ **Half diallel with parents (F_1s and parents):** It includes one way crosses and parents.
- ✓ **Half Diallel without Parents (F_1s Only):** It includes all possible single crosses made in the one direction only.
- ✓ **PARTIAL DIALLEL CROSS**

Partial diallel is a modified form of a diallel cross in which only a part of all possible crosses made among n parents is utilized for evaluation and biometrical analysis. In other words, partial diallel is a genetic design and method of analysis which utilizes only a part of all possible crosses from a diallel mating. It is sometimes termed as fractional diallel. The concept of partial mating design was developed was developed by Kempthorne in 1957 and was further elaborated by Kempthorne and Curnow in 1961, and others. Salient features of partial diallel cross are:

- ❖ Partial diallel utilizes only a part of all possible crosses from a diallel cross.
- ❖ In partial diallel, each parent is crossed to some of the other parents, but not all.
- ❖ In partial diallel, total number of crosses is equal to ns/2, where n and s are number of parents and sample crosses.
- ❖ Analysis is based on the both first and second order statistics. Moreover, s are number of parents and sample crosses.
- ❖ Partial diallel provides information about gca and sca variances and gca effects and D and H components. It does not provide information about sca effects.
- ❖ Results obtained from partial diallel have lesser precision than those of diallel analysis.
- ❖ Partial diallel permits evaluation of more inbred lines at a time than be a complete diallel cross, though with certain loss of precision.

- ✓ **LINE X TESTER ANALYSIS:** It is a modified form of top cross, in case of top cross only one tester is used, while in case of line x tester cross several tester are used. In this the new inbred lines are crossed to a common parent and performance of cross combinations is ascertained. The concept of line x tester analysis was developed by Kempthorne in 1957.

Plan of Crossing for Line x Tester Cross Design

Female Parents	Male parents		
	m1	m2	m3
f1	X	X	X
f2	X	X	X
f3	X	X	X
f4	X	X	X
f5	X	X	X

The crosses obtained by mating each line to each tester are evaluated in a standard statistical design with required number of replications. Usually only crosses are evaluated. However plant breeders are always interested to study the heterosis of crosses produced by them. Therefore parents are also sometimes included along with the crosses and both are evaluated in a same experiment. The biometrical observations are taken on all the replications and are used for statistical analysis.

- ✓ **BIPARENTAL CROSS**

In biparental cross two pure lines or strains having contrasting performance are selected as parents. The selected pure lines are crossed to produce F_1. The F_1 plants are self pollinated to obtain F_2 seeds which are used to raise F_2 plants. Large F_2 population is grown and several single plants are randomly selected for crossing. Based on the fashion of crossing between selected F_2 plants, there are three mating designs of biparental cross. These are commonly known as North Carolina Design I, II and III. These designs provide estimates for the two most important genetic parameters namely additive genetic variance and variance due to dominance.

NCD I: A individual is randomly selected and used as male. A set of 4 randomly selected plants are used as females and mated to the above male. Thus a set of 4 full sib families are produced. This is denoted as a male group. Similarly a large number of male groups are produced. No female is used for any second mating. Four male groups (16 female groups) form a set. Several such sets are made for evaluation.

NCD II: In contrast to NCD I, both parental and maternal half sibs are produced in this design. From the base population of F_2 m males and f females are randomly selected and each male is crossed to each of females. Thus m x f crosses will form a set. Many such sets can be made for evaluation. For each set, male and female plants are selected afresh. The evaluation of crosses and statistical analysis is same as in NCD I.

NCD III: The materials for estimation of genetic parameters are produced by back crossing randomly selected F_2 individuals (n) (using as male) to each of the original parents or imbeds (*l*) (used as females). A male group in this design thus consists of two female progenies and a set contains only 2n full sib progenies. Many such sets can be made for evaluation.

- ✓ **TRIPPLE TEST CROSS**

The most efficient analysis of randomly mating populations is possible when two contrasting inbred lines are available using the extension of design NCD III. This was first devised by Kearsy and Jinks (1968). This allows for the test of one of the assumptions that there is no non-allelic interaction. Thus the epistatic component can be worked out. In this design selected F_2 individuals are crossed not only with two inbred parents but also to their F_1. Therefore, there are 3n families where n is the number of F_2 plants used for crossing. The evaluation of crosses and statistical analysis is same as NCD I.

VARIETAL ADAPTION

- ✓ **STABILITY MODELS**

Stability reflects the suitability of a variety for general cultivation over a wide range of environments. Estimation of phenotypic stability, which involves regression analysis has proved to be valuable technique for assessing the response of various genotypes under changing environment conditions. An evaluation of genotype-environment interactions gives an idea of the buffering capacity of the population under study. The low magnitude of genotype environmental interactions indicates consistent performance of a population over variable environments. In other words it shows high buffering ability of the population. Stability analysis is done from the data of replicated trials conducted over several locations or for several years on the same location or both. The stability analysis consists of following steps.

1. Location or environment wise analysis of variance (ANOVA)
2. Pooled analysis of variance for all the location/environment (POOLED ANOVA)

In pooled analysis, if GXE interaction is found significant, the stability analysis can be carried out using any one of the model.

- ❖ Finlay and Wilkinson model (1963)
- ❖ Eberhart and Russell model (1966)
- ❖ Perkins and Jinks model (1968)
- ❖ Freeman and Perkins model (1971)

- ✓ **Additive Main Effects and Multiplicative Interaction (AMMI) model**

Williams (1952) and Pike and Silverberg (1952) invented the AMMI model. Being a combination of ANOVA and PCA, and being true to its name, the additive Main effects and multiplicative interaction (AMMI) model is an additive and multiplicative model. Thus AMMI model is a hybrid model, the result is the least square analysis, which with further graphical representation of the numerical results (biplot anlysis) often allows a straight forward interpretation of the underlying causes of G X E interaction.

AMMI is applicable and useful for data (i) structured in a two-way factorial, with at least three rows and three coloumns replicated or not, (ii) containing one kind of data quantitative rather than categorical and (iii) fitted by the AMMI model reasonably well as ordinarily happens when main effects and interaction are both significant. The third requirement is fulfilled when all the three factors viz., genotype, environment and GX E interaction are significant in the pooled analysis of variance table. First the additive part of AMMI model uses ordinary ANOVA. ANOVA leaves a non-additive residual, the interaction. Second the multiplicative part of the AMMI model uses PCA to decompose the interaction into PCA axes 1 to N, and a residual ρ_{ge} remains if not all axes are used. These PCA scores are termed as interaction PCA scores or IPCA scores. Each IPCA axis is assigned G+E-1-2n as degrees of freedom where n is the number of axis. The member of the AMMI family with 1 IPCA while relegating all higher axes to the residual is denoted AMMI 1., while AMMI 2 retains 2IPCA axes and so on. If IPCA MS are significant and residual Ms is non-significant, the steps may be continues for development of biplot. The AMMI biplot is developed using genotype and environment mean (on X axis) and the respective IPCA values (on Y axis).

HETEROSIS AND INBREEDING DEPRESSION

- ✓ **Relative heterosis:** The superiority of F_1 hybrid over the mid parental value (ie., mean value or average of two parents involved in the cross) is known as mid parent heterosis, which is estimated as follows:

  Relative heterosis percent = (F_1-MP)/MP * 100
- ✓ **Heterobeltiosis;** The superiority of F_1 hybrid over the better parent or superior parent out of two parents involved in the cross is referred to as heterobeltiosis, which is estimated as follows:

  Heterobeltiosis=(F_1-BP)/BP * 100
- ✓ **Standard Heterosis:** The superiority of F_1 hybrid over the standard commercial variety/ hybrid is known as standard heterosis. The term useful heterosis was used by

Meredith and Bridge (1972). It is also called as economic heterosis. This type of heterosis is of direct practical value in plant breeding. It is estimated as follows:

Standard heterosis= (F_1-SV)/SV *100

- **Inbreeding depression:**

The inbreeding depression is estimated when both F_1 and F_2 populations of the same cross are available.

Inbreeding depression= (F_1-F_2)/F_1 * 100

The following inferences can be drawn from the estimates of heterosis and inbreeding depression.

- ✓ If high heterosis is followed by high inbreeding depression, it indicates the presence of non-additive gene action.
- ✓ If heterosis is followed by low inbreeding depression, it indicates presence of additive gene action
- ✓ The heterosis will be high when some alleles are fixed in one parent and other alleles in the other parent (dispersion of alleles).
- ✓ The genes with lack of dominance will not exhibit heterosis in F_1 but may show increase in performance in F_2 (low inbreeding depression) due to fixation of genes i.e., additive gene action.

CHAPTER 14

QUALITATIVE AND QUANTITATIVE TECHNIQUES IN RESEARCH

R.K. Salgotra, Faizan Danish* and Manish Sharma
School of Biotechnology, Faculty of Biotechnology, SKUAST-Jammu
*Department of Mathematics, School of Advanced Sciences, VIT-AP University, Vijayawada, Amaravati, Andhra Pradesh
Division of Statistics & CS, Faculty of Basic Sciences, SKUAST-Jammu

Quantitative Research

A process of inquiry based on testing a theory composed of variables, measured with numbers, and analyzed using statistical techniques. The goal of quantitative methods is to determine whether the predictive generalization of a theory hold true

Assumptions underlying quantitative methods

Reality is objective, "out there" and independent of the researcher, therefore reality is something that can be studied objectively. The researcher should remain distant and independent of what is being researched. Research is based primarily on deductive forms of logic, and theories and hypotheses are tested in a cause-effect order. The goal is to develop generalization that contributes to theory that enable the researcher to predict, explain, and understand a phenomenon.

Three general types of quantitative methods:

1. Experiments: True experiments are characterized by random assignment of subjects to experimental conditions and the use of experimental controls.

2. Quasi-Experiments : Quasi-experimental studies share almost all the features of experimental designs except that they involve non-randomized assignment of subjects to experimental conditions.

3. Surveys: Surveys include cross-sectional and longitudinal studies using questionnaires or interviews for data collection with the intent of estimating the characteristics of a large population of interest based on a smaller sample from that population.

Qualitative Research: A process of building a complex and holistic picture of the phenomenon of interest, conducted in a natural setting. Researcher who use qualitative methods seek a deeper truth. They aim to “study things in their natural setting", attempting to make sense of, or interpret, phenomena in terms of meanings people bring to them. The goal of qualitative research is to develop an understanding of a social or human problem from multiple perspectives

Assumptions underlying Qualitative methods: Multiple realities exist in any given situation. Researcher interacts with those he/she studies and actively works to minimize the distance between the researcher and those being researched. Researcher explicitly recognizes and acknowledges the value- laden nature of the research .Research is context-bound. Research is based on inductive forms of logic; categories of interest emerge mainly from informants (subject). The goal is to uncover and discover patterns of theories that help explain a phenomenon of interest. Determination of accuracy involves verifying the information with informants or triangulation among different sources of information.

Three general types of qualitative methods:

1. Case Studies: In a case study the researcher explores a single entity or phenomenon ('the case') bounded by time and activity (e.g., a program, event, institution, or social group) and collects detailed information through a variety of data collection procedures over a sustained period of time. The case study is a descriptive record of an individual's experiences and/or behaviors kept by an outside observer.

2. Ethnographic Studies: □ In ethnographic research the researcher studies an intact cultural group in a natural setting over a specific period of time. A cultural group can be any group of individuals who share a common social experience, location, or other social characteristic of interest -- this could range from an ethnographic study of rape victims in crisis shelters, to children in foster care, to a study of a cultural group in Africa.

3. Phenomenological Studies: In a phenomenological study, human experiences are examined through the detailed description of the people being studied -- the goal is to understand the 'lived experience' of the individuals being studied. This approach involves researching a small group of people intensively over a long period of time.

Qualitative vs quantitative research: Qualitative research is concerned with finding the answer to questions which begin with why?how? In what way? Quantitative research is concerned with questions about: how much? How many? How often? To what extent? Quantitative research collects numerical data in order to explain, predict and or control phenomena of interest. Data analysis is mainly statistical. It is categorized with descriptive research correlational research, causal-comparative research and experimental research. The result of research is a number, or a series of numbers, presented in tables, graphs or other forms of statistics. When conducting qualitative research, the researcher collects data consisting mostly of words, pictures, observations of events, etc. These may eventually be categorized in some way, and possibly quantified. Collects narrative data to gain insights into

phenomena of interest. Data analysis includes the coding of the data and production of a verbal synthesis. It offers a lot of freedom in terms of what to study. However, analyzing such data can be very time consuming, and may be influenced by researcher bias.

Quantitative Research: Involves the numerical representation and manipulation of observations for the purpose of describing and explaining the phenomenon that those observations reflect. It is argued that quantification allows for more precision in analysis and ease in summarizing data and making inferences.

Qualitative research: Involves the examination and interpretation of observations for the purpose of discovering underlying meanings and patterns of relationship. It provides much richer, in depth data, which often provide insights into subtle nuances that quantitative approaches might miss. Very useful for exploratory research and in the early stages of theory development.Some key qualitative and quantitative research methods

Quantitative, Randomized clinical trial, Cohort , Case-Control study, Qualitative Participant observation, Case study, Structural observation, Content analysis of documents performance data and Focus groups

Advantages and disadvantages of qualitative and quantitative research

Over the years, debate and arguments have been going on with regard to the appropriateness of qualitative or quantitative research approaches in conducting social research. Robson (2002) noted that there has been a paradigm war between constructivists and positivists. But the two methods are incompatible in the sense that each has its own unique ways of gathering and analyzing data. The two methods are tools used to achieve the same goal with different techniques and procedures, despite the fact that they have different strengths and logic (Paul, 2007; Maxwell, 2004; Maxwell and Loomis, 2002). Both research approaches fall on a research continuum (Creswell, 2009 and Johnson and Christensen, 2012). It is interesting to note that in the research approaches, whether qualitative or quantitative method, the key words “explaining phenomena” is used irrespective of the approach (Muijs, 2004). All the definitions, criticisms, arguments and counter arguments made by authors about the research approaches border only on the methods of data collection, analysis and summary of the results. The fact is that neither constructivists nor positivists have claimed that their instruments are more reliable and valid than the other, thus showing that they are meant to achieve the same goal. It is worth knowing that since qualitative and quantitative research approaches are based on divergent theories and assumptions, one should be more

advantageous than the other and vice versa, depending on the nature of research and data collection methods.

Advantages of Qualitative Research Approach

Berg and Howard (2012) characterise qualitative research as meanings, a concept, a definition, metaphors, symbols and a description of things. This definition clearly show that qualitative research contains all necessary instruments that can evoke recall which aids problem-solving. Qualitative data instruments such as observation, open-ended questions, in-depth interview (audio or video), and field notes are used to collect data from participants in their natural settings. The methods employed in data collection give full description of the research with respect to the participants involved. The participants' observation and focused group nature of qualitative research approach create wider understanding of behaviour. Hence, qualitative research approach provides abundant data about real life people and situations (De Vaus, 2014; Leedy and Ormrod, 2014). Secondly, the system through which data are retrieved in qualitative research approach is regarded as being unique. The reliance on the collection of non-numerical primary data such as words and pictures by the researcher who serves as an instrument himself makes qualitative research well-suited for providing factual and descriptive information (Johnson and Christensen, 2012). Thirdly, in this research approach, theory emerges from data. Different authors use different words or phrases such as: 'investigative, do-it-yourself and bottom-up' to explain the originality and independent nature of the qualitative research approach (Maxwell, 2013; Shank and Brown, 2007; Johnson and Christensen, 2012). The emergent of theory from data allows the researcher to construct and reconstruct theories where necessary, based on the data he generates, instead of testing data generated elsewhere by other researchers. Expressions and experiences of the participants are easily understood even when there are little or no information about them (Leedy and Ormrod, 2014). Moreover, a qualitative research approach views human thought and behaviour in a social context and covers a wide range of phenomena in order to understand and appreciate them thoroughly. Human behaviours, which include interaction, thought, reasoning, composition, and norms, are studied holistically due to in-depth examination of phenomena. The close relationship that exists between the researcher and the participants in this approach makes it easy for the participant to contribute to shaping the research. This however account for significant understanding of experiences as its participants understand themselves and also understand experience as unified (Sherman and Webb, 1990; Lichtman, 2013).

Disadvantages of Qualitative research Approach

Despite the usefulness of a qualitative research approach for conducting research in problem-solving instruction in secondary school science education curriculum, there are still some criticisms about the efficacy of the approach. The problems associated with using qualitative research approach in problem-solving instruction for secondary school science education are highlighted below. Christensen and Johnson (2012) found that qualitative researchers view the social world as being dynamic and not static. In view of this, they limit their findings to the particular group of people being studied instead of generalizing (De Vaus, 2014). In studying problem-solving instruction in secondary school science education, the research approach is presumably deemed to have covered a large proportion of the study group. Perhaps qualitative approach could have been good method for the study if its finding are reflective of a wider population (Shank and Brown, 2007). However, replicability is another problem associated with a qualitative research approach. Critics of this approach argue that the constructivist has abandoned the scientific methods and procedures of enquiry and investigation (Cohen, 2011). The users of the approach are said to write fictions because they have no means of verifying their true statements. Since the approach is characterized by feelings and personal reports, it is believed that the approach cannot give reliable and consistent data when compared to using quantifiable figures (Atkins and Wallac, 2012). As well, the subjective method employed by the qualitative approach users may be wrong, inaccurate and misleading, as suggested by Bernstein (1974) in Cohen and Morrison (2011). The authors' criticism was based on ontological and epistemological paradigms, that is, how the researchers understand and negotiate the situation. Researchers impose their meaning and understanding of a situation to a given time and place to other people. Denzin and Lincoln (2005) stated that constructivists' approach is a multidisciplinary field, therefore their research is only exploratory. Finally, non-use of numbers by qualitative researchers makes it difficult and impossible to simplify findings and observations. Qualitative researchers believe that the social world (phenomena and experiences) has many dimensions, hence explanations are based on the interpretations of the researcher (Leedy and Ormrod, 2014 ; De Vaus, 2014). In view of this, proper explanation cannot be given because the result depends on the explanation of the researcher at that time of which different researcher may give a different explanation. So, the research cannot be repeated by another researcher at another place and still get the same results (Williams and May, 1998).

Advantages of Quantitative Research Approach

The first advantage of this research approach is the use of statistical data as a tool for saving time and resources. (Bryman, 2001) argue that quantitative research approach is the research that places emphasis on numbers and figures in the collection and analysis of data. Imperatively, quantitative research approach can be seen as being scientific in nature. The use of statistical data for the research descriptions and analysis reduces the time and effort which the researcher would have invested in describing his result. Data (numbers, percentages and measurable figures) can be calculated and conducted by a computer through the use of a statistical package for social science (SPSS) (Gorard, 2001; Connolly, 2007) which save lot of energy and resources. Secondly, the use of scientific methods for data collection and analysis make generalization possible with this type of approach. Interaction made with one group can be generalized. Similarity, the interpretation of research findings need not be seen as a mere coincidence (Williams and May 1998). The study of problem-solving instruction in secondary school science education within one particular area or zone can be reflective of the wider society in terms of samples, contents and patterns (Shank and Brown, 2007; Cohen and Morrison, 2011). However, replicability is another benefit derivable from the use of this research approach. Since the research approach basically relies on hypotheses testing, the researcher need not to do intelligent guesswork, rather he would follow clear guidelines and objectives (Lichtman, 2013). The research study using this type of research tool is conducted in a general or public fashion because of its clear objective and guidelines , and can therefore be repeated at any other time or place and still get the same results (Shank and Brown, 2007). Moreover, this research approach gives room for the use of control and study groups. Using control groups, the researcher might decide to split the participants into groups giving them the same teaching, but using different teaching methods, bearing in mind the factors that he is studying. At the end of the study teaching, the groups can be gathered and the researcher can then test the problem-solving ability of the students and be able to access the teaching method that best impacts the problem-solving abilities amongst the students. (Johnson and Christensen, 2012). Finally, Denscombe (1998) describe quantitative research as “researcher detachment” research approach. When looking at the “researcher detachment”, it may be seen as strength of quantitative research approach from one angle, yet from another angle it may seen as its weakness. The issue of researcher being bias with either his data collection or data analysis will be highly eliminated when the researcher is not in direct contact with the participants, that is, he collects his data through either telephone, internet or even pencil-

paper questionnaire. There is full control for alternatives such as interpretations, explanations, and conclusions. In other words, the objectivity of the researcher will not be compromised. Secondly, this may perhaps guarantee respondent anonymity (Muijs, 2004; Litchman, 2006; Bryman, 2012; Creswell, 2009).

Disadvantages of Quantitative Research Approach: Researcher detachment from the participants is also a weakness within the quantitative research approach. Researcher detachment means that he is an "observer" or an "outside looking in". With this type of researcher/participant relationship, it will extremely difficult to get the in-depth study of the phenomena within its natural settings. He will neither understand the group or individuals working with him nor will he appreciate them (Shank and Brown, 2007; Berg, 2007; Christensen and Johnson, 2012). In studying problem- solving instructions for science education in secondary schools, the researcher need not be an observer nor detach himself from the participants. It is dehumanizing as well as undermining life and mind (Cohen, 2011). The experiences gathered may not be that of the participants mind and opinion (Berg and Howard, 2012). Quality and quantity are very important in any educational research since research is an instrument of change. Those two words cannot be neglected when explaining phenomena (Dabbs, 1982 cited in Berg and Howard, 2012). In the quantitative research approach, the participants have no room to contribute to the study. The researcher is at the "driver's seat" (Bryman, 2001). The linear and non-flexibility nature of a quantitative approach demands that the researcher follow a certain order. He starts by setting the research question and hypotheses, conducts a literature review, collects data, analyses the data and summarizes the result (Litchman, 2006; Creswell, 2009). For educational studies such as problem-solving instruction for secondary school science students, the researcher may decide to observe the teaching methods first and see how the method affects students. Following his initial observation, he may repeat the visit for another observation, if necessary, before planning the main research. Input made by the participants can help form researchers' point of orientation. This process is not possible within a quantitative research approach wherein its liturgical order of study does not support several ways of knowing. This is predicated through the use of variables to search for the meanings instead of patterns, as argued by Shank and Brown (2007). Researcher decides the orientation of the research even if participants have a significant point to make or not. A quantitative research approach is characterized as being structured with predetermined variables, hypotheses and design (Denscombe, 1998; Bryman, 2012; Creswell, 2009; Christensen and Johnson, 2012). As a result of using predetermined

working strategies, the approach does not require or encourage imaginative, critical and creative thinking (De Vaus, 1996). Any data collected is geared towards supporting or rejecting the predetermined paradigms. This, however, shows that the tool is effective for studying what is already known instead of assisting in unravelling the unknown and revamping the known. Perhaps, findings from the studies with this tool may lead to propounding laws and facts that can stand on their own regardless of it being true or not (Shank and Brown, 2007). When considering the existence of social differences in the society and schools in particular, a quantitative research approach is not well "suited to examine the complex and dynamic contexts of public education in its forms, sites and variations" (Denzin and Lincoln, 2005). But are there true experiments in educational research? Certainly there is no true experiment in educational research (Gorard, 2001).

References:

Atkins, L. & Wallac, S. (2012). Qualitative Research in Education. SAGE Publication.
BERG, B. L. (2007). Qualitative Research Methods for the Social Sciences. (6th ed). USA: Pearson Educational Inc.
Berg, B. L. & Howard, L. (2012). Qualitative Research Methods for the Social Sciences. (8th ed). USA: Pearson Educational Inc.
Bryman, A. (2001). Social Research Methods. New York: Oxford University Press. Bryman, A. (2008). Social Research Methods. (3rd ed). New York: Oxford University Press.
Bryman, A. (2012). Social Research Methods. 4th edition. New York: Oxford University Press.
Cohen, L., Manion, L. & Morrison, K. (2011). Research Methods in Education. (7th ed). London: Routledge.
Connolly, P. (2007). Qualitative Data Analysis in Education: A critical introduction using SPSS. London: Routledge.
Creswell, J. W. (2009). Research Design Qualitative, Quantitative and Mixed Methods Approach. (3rd ed).London: SAGE Publication.
De Vaus, D. A. (1996). Surveys in Social Research. (4th ed). Australia: UCL Press. De Vaus, D. A. (2014). Surveys in Social Research. (6th ed). Australia: UCL Press.
Denzin, N. K. & Lincoln,Y. S. (2005). The SAGE Handbook of Qualitative Research. (3rd ed). California: SAGE Publication.
Denscombe, M. (1998). The Good Research for Small –Scale Social Research Project. Philadelphia: Open University Press.
Gorard, S. (2001). Quantitative Methods in Educational Research: The role of numbers made easy. London: The Tower Building.
Johnson, B. & Christensen, L. (2012). Educational Research, Qualitative, Quantitative and Mixed Approach. (4th ed). California: SAGE Publication.
Leedy, P. & Ormrod, J. E. (2014). Practical Research Planning and Design. (10th ed). Edinburgh: Pearson Educational Inc.
Lichtman, M. (2006). Qualitative Research in Education: A User's Guide. London: SAGE Publication.
Lichtman, M. (2013). Qualitative Research in Education: A User's Guide. (3rd ed). USA: SAGE Publication.

Maxwell, J. A. & Loomis, D. (2002). Mixed Method Design: An alternative approach in Tashakkor, A. & Teddlie, C. eds: Handbook of Mixed Methods in Social and Behavioural Research, pp241-271. Thousand Oaks, C A: SGAE Publication.

Maxwell, J. A. (2004). Causal Explanation, Qualitative Research, and Scientific Inquiry in Education.Educational Research, 33(2), 3-11.

Maxwell, J. A. (2013). Qualitative Research Design: An Interactive Approach. (3rd ed). London: SAGE Publication.

May, T. & Williams, M. (1998). Knowing The Social World. Buckingham: Open University Press.

Robson, C. (2002). Real World Research: A Resource for Social Scientists and Practitioner-Researchers. (2nd ed). USA: Blackwell Publishing.

Shank, G. & Brown, L. (2007). Exploring Educational Research Literacy. New York: Routledge.

Sherman, R. R. & Webb, R. B. (1990). Qualitative Research in Education: Focus and Methods. London: Falmer Press.

CHAPTER 15

AMMI MODEL AND BI PLOTS

Imran Khan, SKUAST-Kashmir
Email: ik400123@gmail.com

Given multivariate data, there are a number of available methods for multivariate analysis. Gauch (1982) reviews three basic options: direct gradient analysis, classification, and ordination. Depending on the data and research purposes, one approach or some particular combination of approaches may be most appropriate. Direct gradient analysis relates variables of interest to measured and presumably causal factors. Classification places similar entities together in classes, and may additionally arrange the classes in a hierarchy. Ordination seeks an effective low-dimensional summary of high-dimensional multivariate data.The Additive Main effects and Multiplicative Interaction (AMMI) model combines regular analysis of variance (ANOVA) for additive main effects with principal components analysis (PCA) for multiplicative structure within the interaction (that is, within the residual from ANOVA). AMMI is also sometimes called Abiplot@ analysis, even though this term was actually intended to refer to a graph or plot containing two kinds of points or entities (Gabriel 1971; Bradu and Gabriel 1978). It is an appropriate ordination analysis for most yield trials.AMMI is effective for several purposes: (1) understanding genotype-environment interaction, including identifying mega-environments, (2) improving the accuracy of yield estimates, which increases the probability of successfully selecting genotypes with the highest yields, (3) imputing missing data, and (4) increasing the flexibility and efficiency of experimental designs (Gauch 1992, Gauch and Zobel 1996a). Ultimately, these advantages imply larger selection gains in breeding research and more reliable recommendations in agronomy research. AMMI is ordinarily the statistical method of choice when main effects and interaction are both important.

The necessary statistical background for AMMI models was available in 1918 after development of the two components: PCA by Pearson (1901) and ANOVA by Fisher (1918). AMMI models were not developed, however, until 1952 by Pike and Silverberg (1952) and Williams (1952). AMMI consists, quite simply, of fitting an additive ANOVA model in the usual manner (producing a grand mean, row means, and column means), and then for the interaction (that is, the non-additive residual from this additive model) fitting a multiplicative PCA model. By using all the PCA axes an exhaustive model would result,

Afitting@ all the data perfectly. However, the usual research purpose is rather to use only one to a few PCA axes to summarize the patterns in the interaction, leaving a residual. The three traditional approaches of agronomists and breeders for analyzing yield trial data are ANOVA (additive model), PCA (multiplicative model), and linear regression (Tukey 1949; Finlay and Wilkinson 1963; Wright 1971). These approaches are largely subsumed and integrated by AMMI models, which have proven to be exceptionally effective. The advantages from using AMMI models increase with dataset size and noise level.

AMMI offers new and exciting possibilities for effective research strategies. Most specifically, AMMI produces adjusted means that often have a predictive accuracy equivalent to unadjusted means based on much more data. AMMI analysis can often improve accuracy as much as would doubling or tripling the data collection effort. Although it takes time to learn AMMI and to run AMMI analyses, this alternative for improving accuracy can be remarkably cost effective relative to collecting expensive additional data.

AMMI Model Equation

Consider yield data Y_{ger} for G genotypes in E environments with R replications (where R may equal 1). Each genotype and environment combination, or more generically each row and column combination, is termed a treatment, and its average over replications is denoted by Y_{ge}. In other words, the experiment has a two-way factorial design, with each treatment specified by a genotype and environment combination. The AMMI model equation is:

$$Y_{ger} = \mu + \alpha_g + \beta_e + \Sigma_n \lambda_n \gamma_{gn} \delta_{en} + \rho_{ge} + \varepsilon_{ger}$$

where Y_{ger} is the yield of genotype *g* in environment *e* for replicate *r*,
μ is the grand mean,
α_g is the genotype *g* mean deviation (genotype mean minus grand mean),
β_e is the environment *e* mean deviation,
λ_n is the singular value for IPCA axis *n*,
γ_{gn} is the genotype *g* eigenvector value for IPCA axis *n*,
δ_{en} is the environment *e* eigenvector value for IPCA axis *n*,
ρ_{ge} is the residual, and
ε_{ger} is the error.

Note that $\Sigma\alpha_g = \Sigma\beta_e = 0$. The eigenvector values for each interaction PCA (IPCA) axis are scaled to unit vectors so that $\Sigma\gamma_g^2 = \Sigma\delta_e^2 = 0$. The eigenvalue for a given IPCA axis is the sum of squares (SS) accounted for by that axis, and it equals λ^2 or the square of the singular value λ. The sum of the eigenvalues $\Sigma_n\lambda_n^2$ for N axes, plus the residual SS of $R\Sigma_g\Sigma_e\rho_{ge}^2$ for a reduced model, equals the genotype-environment (GE) interaction SS. Given G

genotypes and E environments, there are GE□1 df for treatments, G□1 df for genotypes, E□1 df for environments, and (G□1)(E□1) df for interaction.

A convenient scaling for tabulating the multiplicative part of the AMMI model results from expressing genotype scores as $\lambda^{0.5}\gamma_g$ and environment scores as $\lambda^{0.5}\delta_e$. Then multiplication of a genotype score by an environment score gives the estimated interaction directly, without need of a further multiplication by λ. The units for μ, α, β, λ, and ρ are exactly the same units of yield as for the data Y, but γ and δ are dimensionless. Therefore, the genotype scores $\lambda^{0.5}\gamma_g$ and environment scores $\lambda^{0.5}\delta_e$ are in the units of the square root of yield, and hence the product of a genotype score times an environment score is in the units of yield as required.

If scaling of the additive effects as means is preferred to deviations as in the above AMMI model equation, then use the means $\mu+\alpha_g$ and $\mu+\beta_e$ instead of the deviations α_g and β_e, and modify the model equation to subtract rather than add the grand mean μ

The AMMI model can have up to min(G□1, E□1) axes. Using all axes results in the full model, which has as many df as the data and accordingly fits the data exactly, eliminating the residual term. However, ordinarily a reduced AMMI model is chosen, summing the IPCA axes over *n*=1, N axes, where N is smaller than the full model, and hence the model does leave a residual. It is useful to denote the AMMI family of models as AMMI0 for the AMMI model with no IPCA axis (namely, the ANOVA model), AMMI1 for 1 IPCA axis, AMMI2 for 2, and so on.

Yield estimates for a reduced AMMI model are obtained by deleting the residual term from the above equation. Also, since these estimates are for each genotype *g* in each environment *e*, the error term and the subscript for replications *r* are also deleted. And for the AMMI1 model, there is only one IPCA 1 axis, so the subscript for axis number *n* can also be deleted.

Biplot: A biplot is a scatter plot that graphically displays both the row factors and the column factors of a two-way data. The concept of biplots were first developed by Gabriel, K.R.(1971). Since then the biplot has been used in data visualization and pattern analysis in various research fields, from psychology to economics, to agronomy. This technique has extensively been used in the analysis of multi-environment trials. For generating a biplot, we take the matrix representing the effects of two factors. This matrix is then subjected to singular value decomposition.

CHAPTER 16

BIG DATA ANALYSIS USING STATISTICAL SOFTWARES IN ANIMAL SCIENCES

Dibyendu Chakraborty*, Dhirendra Kumar, Simran Singh, Mandeep Singh Azad, M. Iqbal Jeelani Bhat[#] and Manish K. Sharma[#]
Division of Animal Genetics & Breeding
FVSc&AH, SKUAST-Jammu, R. S. Pura, Jammu-181102
[#] Division of Statistics & Computer Science, SKUAST-Jammu, Chatha-180009
dibyendu_vet40@yahoo.co.in

Data recording and data analysis is very important aspects of any scientific activity. If we consider about the success of G. J. Mendel, we can see he succeeded only due to proper maintenance of data and interpretation of data with statistical tools. So, when it is about Animal Sciences, the improvement of animals in term of production and health is only possible when the proper data are maintained and data is analyzed statistically. With the advanced in technology animal farmers predict and prevent diseases like mastitis, lameness, postpartum diseases, African swine flu, Coccidiosis etc. (Neethirajan, 2020). The Now we have entered in the era of advanced computer programming and statistical softwares. So, the results in Animal Sciences can be interpreted in terms of big data. The term big data refers to extremely large sets of information which require specialized computational tools to enable their analysis and exploitation. With the advancement of molecular and breeding techniques the big data on animal health, production, reproduction and overall welfare can be analyzed and decisions can be made for empowering farmers to improve welfare, production and sustainability.

The artificial intelligence is also helping to manage the big data recoding such as monitoring precision livestock health data; developing wearable sensors (such as accelerometers and temperature sensors) for animals which can detect lameness and other health and welfare parameters; and an internet of things for livestock farming. Big data analysis in animal science helps to in precision animal agriculture including management, production, welfare, sustainability, health surveillance, and environmental footprint. The big data applications are helpful also in smart livestock farming like Biometric sensing and GPS tracking (Sonka, 2014); Breeding, monitoring (Cole *et al.*, 2012); Milk robots (Gobart, 2012) and livestock movements (Faulkner and Cebul, 2014;Wamba and Wicks, 2010). There are many statistical softwares are used for big data analysis in animal sciences. Some of them have been discussed here.

What is Big Data?

According to Gartner, the definition of Big Data –

"*Big data*" is high-volume, velocity, and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making."

Big Data refers to complex and large data sets that have to be processed and analyzed to uncover valuable information that can benefit organizations. There are some basic However, there are certain basic system of belief of Big Data that makes it more simpler to define Big Data:

- A massive amount of data that keeps on growing exponentially with time.
- It is so voluminous that it cannot be processed or analyzed using conventional data processing techniques.
- It includes data mining, data storage, data analysis, data sharing, and data visualization.
- The term is an all-comprehensive one including data, data frameworks, along with the tools and techniques used to process and analyze the data.

Why is Big Data Important?

The importance of big data does not revolve around how much data collected but it implies that how the collected data are utilized. The importance of Big Data are-

1. Cost Savings
2. Time Reductions
3. Understanding the present scenario
4. Solving the basic problems related to the research
5. Big Data Analytics to solve Researcher's curiosity
6. Big Data Analytics as a Driver of Innovations and Product Development

Big Data Analysis in Animal Sciences

There are many softwares are used for big data analysis. A survey results figured that R, SAS, and SPSS the most used platforms and 47%, 32%, and 32% of respondents were using R, SAS, and SPSS, respectively (Fig. 1). Some of the Animal Breeding softwares are handling the big data regarding the Animal Breeding. These are DFREML, WOMBAT, BLUPF90 FAMILY OF PROGRAMS etc.

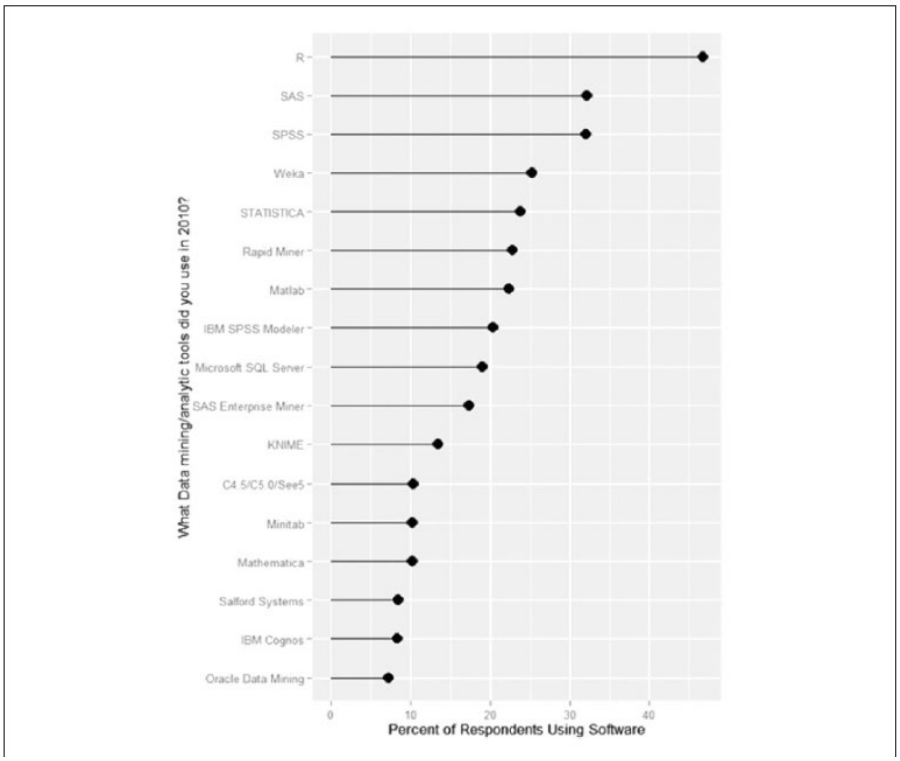


Fig. 1: Rever analytics survey results of analytic tools (*Source:* Muenchen, 2014).

SPSS

SPSS, originally termed Statistical Package for the Social Sciences, was released in 1968 as a software designed for the social sciences. Since then, IBM has replaced SPSS Inc. as the owner, and the software has expanded its user base past this one area. The software's former acronym has been replaced with Statistical Product and Service Solutions to reflect the greater diversity of its clients. Arguably, it still remains the leading statistical analysis software package for the social sciences. In fact, SPSS does look similar to typical spreadsheet applications like Excel, and its ease of use is very comparable to Excel as well.

There are many differences between Excel and SPSS that suit SPSS to better handle statistical methods. For one, "SPSS was designed specifically for statistical processing of large amount of data at an enterprise level," while spreadsheets are broadly applicable to many different tasks outside of statistical computing (Robbins, 2012). An advantage of this specialized design is that SPSS "keeps calculated statistics and graphs separate from the raw data but still easily accessible" (Robbins, 2012). SPSS software furthermore has a much more convenient platform for performing statistical tests. For instance, performing a one-sample *t* test in Excel requires some independent calculations by the user, whereas with SPSS, the user only needs to

"select a variable and supply the value to compare with [the] sample" and click "Ok" (Robbins, 2012). Another advantage of SPSS is that it links numerically coded data to its original meaning (Robbins, 2012). With most data being electronically stored in numerical fashion, this feature of SPSS is highly valuable

SAS

SAS (Statistical Analysis System) is a commercial statistical package that was developed during the 1960s and 1970s at North Carolina State University as part of an agricultural research project. Its usage has grown exponentially since then. Nowadays, 91 of the top 100 companies on the 2013 Fortune Global 500 list use the software (SAS, 2014). SAS's Analytics Pro bundle comes with three of the most popular SAS products: Base SAS, SAS/STAT, and SAS/GRAPH. The corporation's Visual Data Discovery collection includes SAS Enterprise Guide (SAS's only point–click interface) and JMP software to make discovery and exploratory analysis easier.

With either of these toolsets, programmers can perform a number of statistical tests. The Institute for Digital Research and Education website outlines a multitude of statistical tests and their corresponding SAS codes. The list includes 32 tests that come from statistical categories such as regression, factor analysis, discriminant analysis, ANOVA, nonparametric tests, and correlation.

SAS has built-in, functional packages for many specific industries, including health care, banking, insurance, law enforcement, communications, retail, casinos, utilities, sports, and more.

R

R is a free, open-source statistical software. Colleagues at the University of Auckland in New Zealand, Robert Gentleman and Ross Ihaka, created the software in 1993 because they mutually saw a need for a better software environment for their classes. R has certainly outgrown its origins, now boasting more than 2 million users according to an R Community website (Revolution Analytics, 2014).

R is a comprehensive statistical analysis toolkit. It can perform any statistical analysis desired, but users must either write the code or access the code from someone who has already written it. As stated on its website, people have already designed many standard data analysis tools "from accessing data in various formats, to data manipulation (transforms, merges, aggregations, etc.), to traditional and modern statistical models (regression, ANOVA, GLM, tree models, etc.

The key feature of R that differentiates it from other statistical softwares is its acceptance of customization. On one hand, the aforementioned software have "data-in-data-out black-box procedures".

DFREML

It is the derivative-free REML (Smith and Graser, 1986) developed by Meyer (1988). This program was extensively used and became gold standard in the 1990s till 2005. It used likelihood ratio test to compare the significance of the variance components used in the model. It supported 10 models that also included complex random regression.

WOMBAT

It is a slightly less flexible but free alternative to ASReml, written by Karen Meyer (2006). It can be downloaded via http://agbu.une.edu.au/~kmeyer/wombat.html. 32-bit and 64-bit Linux executables is capable of analysing complex models and large amounts of data, as well as a less powerful and efficient Windows version are available. As is the case for ASReml, users will require a simple text editor to generate the input files. A manual covering all the basics as well as a number of more advanced options can be downloaded at http://agbu.une.edu.au/~kmeyer/download.php?file=WombatManual.pdf. Additional information related to running WOMBAT and REML estimation in general can also be found at the WOMBAT Wiki (http://agbu.une.edu.au/~kmeyer/dokuwiki/doku.php).

BLUPF90 FAMILY OF PROGRAMS

The BLUPF90 family of programs is a collection of software in Fortran 90/95 for mixed model computations in animal breeding (http://nce.ads.uga.edu/wiki/doku.php?id=application_programs). The objective of the software is to be as simple as with a matrix package and as efficient as in a programming language. The programs support mixed models with multiple-correlated effects, multiple animal models and dominance.The programs can do data conditioning, estimate variances using several methods, calculate BLUP for very large data sets, calculate approximate accuracy, and use SNP information for improved accuracy of breeding values + for genome-wide association studies (GWAS). The programs have been designed with 3 goals in mind: 1. Flexibility to support a large set of models found in animal breeding applications. 2. Simplicity of software to minimize errors and facilitate modifications. 3. Efficiency at the algorithmic level. User manual can be

found at (http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=blupf90_all4.pdf).

Conclusion

The technology driven discoveries led to generate the big data in animal sciences. The data are generated by fully automated, high-throughput data recording or phenotyping platforms, including digital images, sensor and sound data, unmanned systems, and information obtained from real-time noninvasive computer vision. The big data analysis subsequently helps in taking more precise decisions. However, the big data also creates complexity in analysis and understanding in Animal Sciences. The potential in "big data" analysis have not been adequately explored in the animal science. In near future it may be fully utilized and can be applied to solve pressing problems in animal sciences.

References:

Cole JB, Newman S, Foertter F, Aguilar I and Coffey M. 2012. Breeding and Genetics Symposium: really Big Data: processing and analysis of very large data sets. *J. Anim. Sci.*, *90* (**2012**): 723–733.

Faulkner A. and Cebul K. Agriculture Gets Smart: The Rise of Data and Robotics, Cleantech Agriculture Report. Cleantech Group. (**2014**).

Grobart S. Dairy industry in era of Big Data: new gadgets help farmers monitor cows and analyze their milk. (**2012**).

Meyer K. DFREML. A set of programs to estimate variance components under an individual animal model. *J. Dairy Sci.*, *69*(Suppl. 2) (**1988**):33.

Meyer K. WOMBAT – Digging deep for quantitative genetic analyses by restricted maximum likelihood. *Proc. 8th World Congr. Genet. Appl. Livest. Prod., Communication*, *27* (**2006**): 4.

Morota, G. Ricardo VV, e, Silva FF, Koyama M and Fernando SC. BIG DATA ANALYTICS AND PRECISION ANIMAL AGRICULTURE SYMPOSIUM: Machine learning and data mining advance predictive big data analysis in precision animal agriculture. *J. Anim. Sci., 96*(4) (**2018**):1540-1550. doi: 10.1093/JAS/SKY014.

Muenchen RA. (**2014**). The popularity of data analysis software. Retrieved from http://r4stats.com/articles/popularity/

Neethirajan S. The role of sensors, big data and machine learning in modern animal farming. *Sensing and Bio-Sensing Research*, *29* (**2020**): 100367. https://doi.org/10.1016/j.sbsr.2020.100367.

Revolution Analytics. (**2014**). *What is R?* Retrieved from http://www.inside-r.org/what-is-r.

Robbins, S. (**2012**). *How does SPSS differ from a typical spreadsheet application*. Retrieved from https://publish.illinois.edu/commonsknowledge/2012/06/07/how-does-spss-differ-froma-typical-spreadsheet-application.

SAS. (**2014**). *About SAS*. Retrieved from http://www.sas.com/en_us/company-information.html.

Sonka, S. Big Data and the Ag sector: more than lots of numbers. *International Food and Agribusiness Management Review, 17*(**2014**): 1.

Wamba SF and Wicks A. RFID deployment and use in the dairy value chain: applications, current issues and future research directions. *Technology and Society* (ISTAS), International Symposium on IEEE, (**2010**), pp. 172–179.

Why use R. (**2014**). Retrieved from http://www.inside-r.org/whyuse-r.

CHAPTER 17

ROLE OF ARTIFICIAL INTELLIGENCE IN AGRICULTURE

Pawanesh Abrol[1], Palak Mahajan[2], Namrata Phonsa[3] and Parul Sharma[4]
[1 2 3 4] Department of Computer Science and IT, University of Jammu, J&K, India.
[1]pawanesh.abrol@gmail.com, [2]palak.mahajan18@gmail.com
[3]namrataphonsa@gmail.com, [4]sharmaparul1620@gmail.com

Plants existence is essential for human continued existence. They are directly or indirectly reliant on plants for food, shelter, and clothing. With the growing human population now, it becomes more important to protect them and keep their health good. There are very few experts who can recognize the species and assess the underlying health of the plants. This is where machine learning comes into fold. With its ability to do the classification, categorization, regression, decision making etc. it has become inevitable in nearly all research areas. Pertaining to plants, it has shown good results in its species recognition, disease recognition, health status, growth monitoring, yield prediction etc. Many user-friendly portable applications have been created for such purposes, but they all lack accuracy and wide coverage of species [1].

Artificial intelligence (AI) is the branch of computer science that deals with the simulation of human intelligence. Every vendor across the world wants to incorporate their product with AI for its more usability and acceptability. Deep learning is an advanced form of machine learning that works with huge amounts of data and comprises neural networks in its architecture. It has become a primary learning paradigm/tool of AI for its exceptional results in many areas. Convolutional Neural Network (CNN) is the most popular form of deep learning that deals with classifications, segmentation, object detection etc. Training a CNN model would require a large amount of image data which may take days to complete. Once the training phase is over, the CNN model having the weights adjusted in all the layers can be saved to be used later for similar types of problems. Artificial Intelligence is a discipline that deals with creating intelligent application programs and machines to do what the human cognitive mind does concerning vision, speech and reasoning. There are primarily two ways of programming by which computers learn to mimic humans' capabilities, these are: rule based and self-learned. In the former approach, pre-defined rules and logics enable the machines to act accordingly and learn. These are knowledge driven AI programs, whereas, in later systems a machine learns itself to extract out useful information from the data given to it. This way a computer formulates the appropriate combinations of rules to bring out most

accurate results is called machine learning [2]. The most popular form of machine learning is through neural networks whose design is inspired from the working of the human brain. The image in figure 1 shows a diagram of a biological neuron marked to describe one artificial neuron's function.

Neuron receives input signals from other neurons via its dendrites and passes output signals through its axon. Axon then branches out to join other neurons. In the figure above, the x_0 is the input signal; this signal is multiplied ($w_0 x_0$) with the weight variable (w_0). The influence is determined by summing the signal input and weight ($\sum_i w_i x_i + b$) which is then calculated by the activation function f, if it is above a certain threshold the neuron fires.

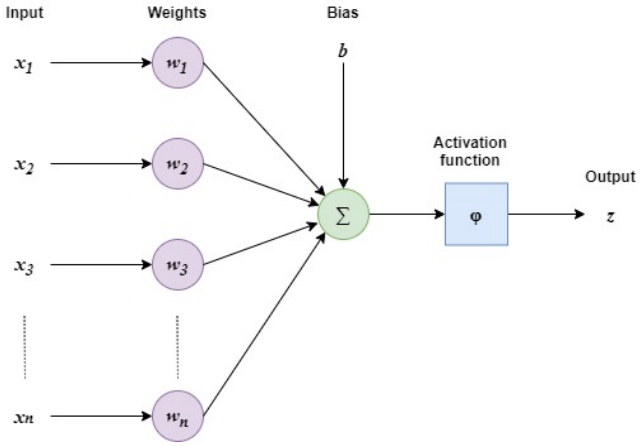


Figure 1. A n-input ANN model with w_i weights and b bias where φ represents the activation function that gives an output z [12].

Machine Learning (ML) is a subset of AI which enables the computer system to learn from the data given in the beginning and enhance its performance from experience with limited human interventions. It is helpful especially when the data is huge and humans are incapable to go through the details and find out the knowledge from its arrangement. Learning can be supervised, unsupervised or reinforced depending on the type of the application in which it will be used. Supervised learning models such as support vector machines (SVM), k nearest neighbor (kNN), naive-bayes (NB), decision tree, convolution neural network (CNN), etc. trains the machine in accordance to the labeled output. During the learning period, output is obtained after every iteration, which is compared with labeled output and depending upon the grade of similarity or dissimilarity, the weights are adjusted, and some bias value is introduced in the network. The weights and the bias are continuously altered until the desired output is achieved. Unsupervised learning models such as k-means, hierarchical clustering, and CNN etc. cluster output values based on their similarity or dissimilarity measures to each other. In reinforcement learning, there are no predefined classes, but it is trained to learn from

its experience. Action is taken in a particular situation to maximize reward. Based on the type of reward, it has two types, that is, negative and positive. Positive reinforcement is when the network output is in the desired direction whereas in negative reinforcement the output is not in the desired direction. This work focuses on the capacity of AI in the field of agriculture [3]. The role of AI in agriculture is comprehended from several research papers. AI is such a vast subject that it almost covers each problem that deals with the following:

- Finding patterns, trends, and associations'
- Implement plans
- Learn with experience and perform better
- Historical data-based knowledge discovery
- Assist in fact-based decisions
- Find inefficiencies.

There are some major areas where it has been and still used prevalently in the agricultural sector are:

Crop Management

Crop yield prediction: Machine learning has been beneficial for increasing crop productivity by evaluating yield estimations, yield mapping, and matching supply with demand [4]. Disease detection: Detection and identification of crop disease is one of the most important research projects works that concerns agriculture. It deals with automatic disease identification in the crops [5]. Weed Detection: Weeds presence in crops hinders its growth and productivity. Moreover, its presence is difficult to detect from crops. Apart from detection, species can also be identified as done in work [6]. Crop quality: The quality assessment of crops helps in improving the yield and reduces its loss by correct classification and estimation of crop quality [7]. Species recognition: This is yet another important application of machine learning in agriculture which machine learning techniques help in automatic classification and identification of crop or plant species. [8].

Livestock management

Animal wellbeing: Machine learning helps in observing animals for diseases and health monitoring. Animal behavior can be closely observed for their well-being as suggested by [9]. Livestock production: Animal produce like eggs, milk, yarn etc. can be observed for its quality and quantity assessment [10].

Other areas in agriculture where machine learning technologies has started to set strong hold are water and soil management for improved irrigations, evapotranspiration and various soil properties.

Case Study: In one of our research works [11], multiple component or input-based CNN is implemented to classify 10 similar looking citrus species. For input to CNN images of four different organs of the citrus plant namely leaves, fruits, flowers and entire plants were used. The dataset was self-created by capturing these images in natural environment from citrus orchard in SKUAST-Jammu. For leaves, fruits, flowers and entire plant 50, 30, 10 and 50 images per species were collected respectively. Some samples are shown in figure 2.

Figure 2: Samples of images of a. leaves, b. fruits, c. flowers and d. entire plant of Kinnow, Narangi, Grapefruit and Sweet orange citrus species (from left to right).

The methodology that was adopted is depicted in figure 3. The performance of the CNN model is evaluated for citrus plant species classification. The first experiment has been conducted using leaves, fruits, and entire plant organs of citrus plants separately as input for the convolution layer as shown in figure 2. Python 3.7 is used for transfer learning of Inceptionv3 CNN model with sequential API of Keras library to construct the single input-based model.

Result: First single organ was used to classify these species except flowers (because of a smaller number of its samples) and it was found out that leaves were best in categorizing citrus species followed by fruits. Next the combinations of two organs were evaluated and then their combinations with flower input were tested and it was found that there was a drop in accuracy. The combination of three inputs of leaves, fruits and entire plants gave the best final classification accuracy of 91.4%. Multiple input-based CNN was able to classify 10 similar-looking citrus plants.

Conclusion of case study: The three main observations were made on the basis of results obtained. First, even though citrus species closely resemble each other, the results obtained, when its multiple organs are combined for CNN, are comparable to a similar type of work conducted earlier. Second, in the case of citrus plants, leaves give better classification accuracy than other organs. Third, citrus flowers closely resemble such an extent that they cause misclassifications and thus decrease final classification accuracy. In general, the final classification accuracy can further be improved by increasing training dataset size, using more other components of the plant like stem, branch, etc., and using deeper CNN models like Inceptionv3.

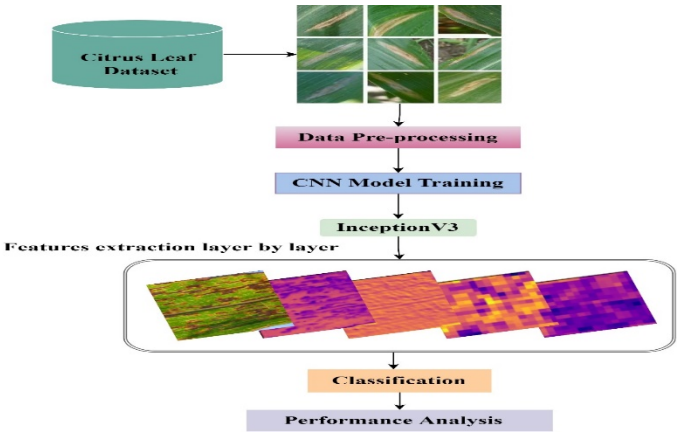
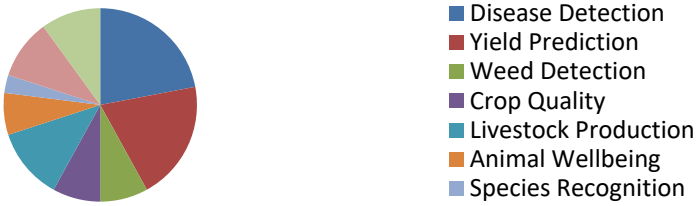


Figure 3: Block Diagram for plant leaf classification.

Conclusion:

Based on the studies executed in the contemporary research articles, it was observed that about 22% of the work has been done on disease detection, 20% on yield prediction, 12% on livestock production, 10% on water and soil management, 8% each on weed detection and crop quality, 7% on animal wellbeing, 3% on species recognition and rest on others like weather, fertilizers, smart field vehicles etc.

By applying AI techniques in various agricultural sectors, the ongoing research is giving improved insights and endorsing the decisions and actions with the aim to increase the quality and the quantity of the production. This will give a way for better knowledge-based agricultural techniques which will eventually result in better bio-product quality.



References

[1]Singh, A. K; Ganapathysubramanian, B; Sarkar, S; and Singh, A.; Deep Learning for Plant Stress Phenotyping: Trends and Future Perspectives. Trends in plant science, vol. *23*, no.10, 2018, pp. 883–898.

[2]LeCun, Y; Bengio, Y; and Hinton, G.; Deep learning. Nature, 521, no. 7553, 2015, pp. 436-444.

[3]Kamilaris, A; and Prenafeta-Boldú, F.X.; Deep learning in agriculture: A survey. Computers and Electronics in Agriculture, no. 147, 2018, pp. 70-90.

[4]Amatya, S.; Karkee, M.; Gongal, A.; Zhang, Q.; Whiting, M.D. Detection of cherry tree branches with full foliage in planar architecture for automated sweet-cherry harvesting. Biosyst. Eng., no. 146, 2015, pp. 3–15.

[5]Pantazi, X.E.; Tamouridou, A.A.; Alexandridis, T.K.; Lagopodi, A.L.; Kontouris, G.; Moshou, D.; Detection of Silybum marianum infection with Microbotryumsilybum using VNIR field spectroscopy. Comput. Electron. Agric., no. 137, 2017, pp. 130–137.

[6]Pantazi, X.-E.; Moshou, D.; Bravo, C.; Active learning system for weed species recognition based on hyperspectral sensing. Biosyst. Eng., 146, 2016, pp. 193–202.

[7]Zhang, M.; Li, C.; Yang, F.; Classification of foreign matter embedded inside cotton lint using short wave infrared (SWIR) hyperspectral transmittance imaging. Comput. Electron. Agric., no.139, 2017, pp. 75–90.

[8]Grinblat, G.L.; Uzal, L.C.; Larese, M.G.; Granitto, P.M. Deep learning for plant identification using vein morphological patterns. Comput. Electron. Agric., 127, 2016, pp. 418–424.

[9] Pegorini, V.; Karam, L.Z.; Pitta, C.S.R.; Cardoso, R.; da Silva, J.C.C.; Kalinowski, H.J.; Ribeiro, R.; Bertotti, F.L.; Assmann, T.S. In vivo pattern classification of ingestive behavior in ruminants using FBG sensors and machine learning. Sensors, no. 15, 2015, pp. 28456–28471.

[10]Morales, I.R.; Cebrián, D.R.; Fernandez-Blanco, E.; Sierra, A.P.; Early warning in egg production curves from commercial hens: A SVM approach. Comput. Electron. Agric., no. 121, 2016, pp. 169–179.

[11]Sharma, P.; Abrol, P.; Analysis of multi component-based CNN for similar citrus species classification, Machine Learning & Cognitive Science: A Walkthrough. Studies in Computational Intelligence, no 1027, 2022.

[12]Mahajan, P.; Abrol P.; Lehana P.K. Scene based Classification of Aerial Images using Convolution Neural Networks. Journal of Scientific Industrial Research. 79, 2020, pp. 1087-94.

CHAPTER 18

NEURAL NETWORK MODELLING IN R STUDIO

M .Iqbal Jeelani Bhat, Manish Kr. Sharma, S.E.H.Rizvi, Sunali Mahajan
Division of Statistics & Computer Science, FBSc
SKUAST- Jammu, Jammu
jeelani.miqbal@gmail.com

Introduction

The Artificial Neural Network (ANN) is one of the well-known prognostic methods used to find a solution when other statistical methods are not applicable. The advantages of this tool, such as the ability to learn from examples, fault tolerance, operation in a real-time environment, and forecasting non-linear data, all make it a widely used statistical tool. Moreover, ANN accurately fits in the nonlinear variables, which is an advantage compared to multivariate linear analysis based on linear variables. An inspiration for ANN was the human brain and biological neurons. The basic element of this structure is the perceptron. This is a mathematical equivalent of a neuron, which transfers electrical signals represented as numerical values. Artificial neurons are arranged in layers: input—taking the input data, hidden and output—producing a result. Each node connects with every neuron in the next layer. However, there are no connections among neurons in the same layer. .

Input layers: Layers that take inputs based on existing data
Hidden layers: Layers that use back propagation to optimize the weights of the input variables in order to improve the predictive power of the model
Output layers: Output of predictions based on the data from the input and hidden layers

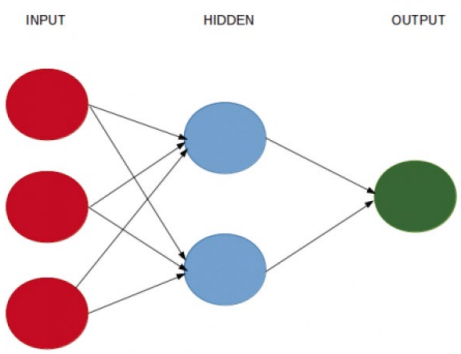
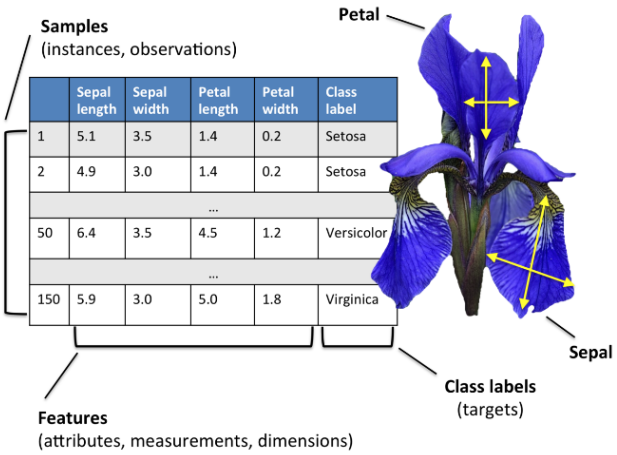


Fig.1 Basic Structure

The ANN learning process is based on adjusting weighted connections between nodes until the most efficient solution of a problem has been obtained. Moreover, providing both an input and output in the network allows for calculation of an error based on its target output and present output. This can be used for corrections of the network by updating its weights and to achieve optimal results . Artificial neural networks (ANN) or connectionist systems are computing systems that are inspired by, but not identical to, biological neural networks that constitute animal brains. In this chapter iris data which is an inbuilt data available in R software has been analyzed through Neural Network modelling The **data set** consists of 50 samples from each of three species of **Iris** (**Iris** setosa, **Iris** virginica and **Iris** versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters.



	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

Fig.2: This famous (Fisher's) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are *Iris setosa*, *versicolor*, and *virginica*.Analysis of iris data through Neural Network modelling in R is done with the help of below mentioned steps

- **Install Libraries**

```
install.packages("neuralnet")
install.packages("NeuralNetTools")
install.packages("ggplot2")
install.packages("GGally")
install.packages("caret")
```

- **Import Libraries**

```
library("neuralnet")
```

```
library("NeuralNetTools")
library("ggplot2")
library("GGally")
library("caret")
```

- **Load Iris:** The Iris data set exist inside the R, so, you can import conform the code below.

```
data(iris)
```

- **Exploratory Data Analysis**

In statistics, exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods.

- Correlation Analysis

```
ggplot <- function(...)
ggplot <- function(...)
  ggplot2::ggplot(...) +
  scale_color_brewer(palette="Purples") +
  scale_fill_brewer(palette="Purples")
unlockBinding("ggplot",parent.env(asNamespace("GGally")))
assign("ggplot",ggplot,parent.env(asNamespace("GGally")))
graph_corr <- ggpairs(iris, mapping = aes(color = Species),
                      columns = c('Sepal.Length',
                                  'Sepal.Width',
                                  'Petal.Length',
                                  'Petal.Width',
                                  'Species'),
                      columnLabels = c('Sepal.Length',
                                       'Sepal.Width',
                                       'Petal.Length',
                                       'Petal.Width',
                                       'Species'))
graph_corr <- graph_corr + theme_minimal()
graph_corr
```

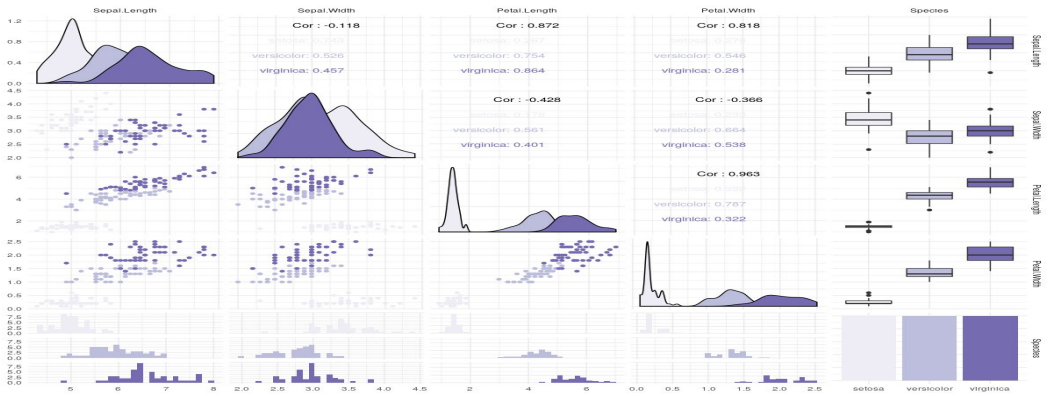


Fig.3: ggpairs plots which is combination of scatter, density, correlation and box plots

- **Normalization and Transform Data**

One of the most important procedures when forming a neural network is data normalization. This involves adjusting the data to a common scale so as to accurately compare predicted and actual values. Failure to normalize the data will typically result in the prediction value remaining the same across all observations, regardless of the input values. It can be done in the following ways in R :

- Scale the data frame automatically using the *scale* function in R
- Transform the data using a *max-min normalization* technique

For this example, Max-Min Normalization function is used.

```
norm.fun = function(x){(x - min(x))/(max(x) - min(x))}
Apply the function in the dataset to normalize data.
df_iris = iris[,c("Sepal.Length","Sepal.Width",
                  "Petal.Length","Petal.Width" )]
df_iris = as.data.frame(apply(df_iris, 2, norm.fun))
df_iris$Species = iris$Species
df_iris$setosa <- df_iris$Species=="setosa"
df_iris$virginica <- df_iris$Species == "virginica"
df_iris$versicolor <- df_iris$Species == "versicolor"
```

Data validation : Now we need split data between training and test data set. In code below, the training sample size represent 75% of data set total, and 25% represent the test data set.

```
## 75% of the sample size
smp_size <- floor(0.75 * nrow(df_iris))
## set the seed to make your partition reproducible
set.seed(123)
train_ind <- sample(seq_len(nrow(df_iris)), size = smp_size)
training.set <- df_iris[train_ind, ]
test.set <- df_iris[-train_ind, ]
```

Fitting of Model: The neural network model contain 4 variables in input layer, 2 hidden layers containing in each layer 10 neurons. The number of repetitions for the neural network's training is equal 5. The activation function used is **logistic**, and the function that is used for the calculation of the error is **ce (cross-entropy).**

```
model = as.formula("Species ~
                            Sepal.Length +
                            Sepal.Width +
                            Petal.Length +
                            Petal.Width")
iris.net <- neuralnet(model,
                      data=training.set,
                      hidden=c(10,10),
                      rep = 5,
                      act.fct = "logistic",
                      err.fct = "ce",
                      linear.output = F,
                      lifesign = "minimal",
                      stepmax = 1000000,
```

```
                    threshold = 0.001)
```

- **Visualization NN plot**

Visualize the neural network architecture with the code below.

```
plotnet(iris.net,
        alpha.val = 0.8,
        circle_col = list('purple', 'white', 'white'),
        bord_col = 'black')
```

Output Neural Network Model Visualization

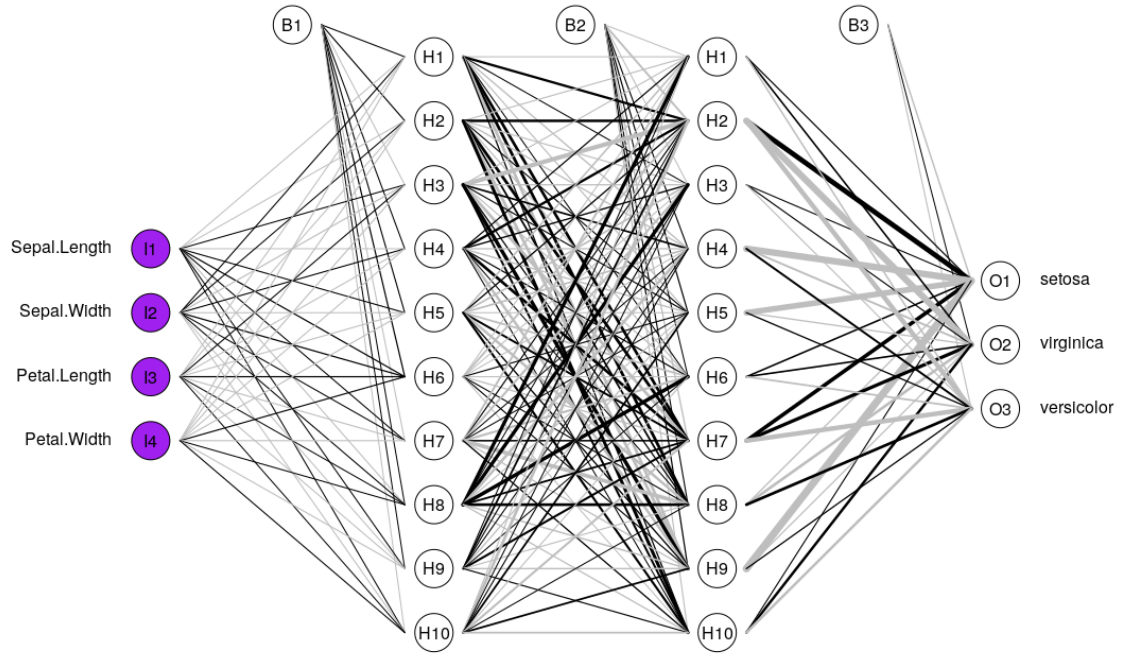


Fig.4: Neural Network Plot

- **Prediction**

Use the test data set as input to the Neural Network model to predict Iris classes.

```
iris.prediction <- compute(iris.net, test.set)
idx <- apply(iris.prediction$net.result, 1, which.max)
predicted <- as.factor(c('setosa', 'versicolor',
'virginica')[idx])
```

Model evaluation:

Performance of an algorithm, typically a supervised learning one is done by an error matrix known as confusion matrix. Each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class (or vice versa).

```
confusionMatrix(predicted, test.set$Species)
Output Confusion Matrix
Confusion Matrix and Statistics
            Reference
Prediction    setosa versicolor virginica
  setosa          11          0         0
  versicolor       0         13         1
  virginica        0          0        13

Overall Statistics

               Accuracy : 0.9737
                 95% CI : (0.8619, 0.9993)
    No Information Rate : 0.3684
    P-Value [Acc > NIR] : 2.196e-15

                  Kappa : 0.9604

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: setosa Class: versicolor Class: virginica
Sensitivity                 1.0000            1.0000           0.9286
Specificity                 1.0000            0.9600           1.0000
Pos Pred Value              1.0000            0.9286           1.0000
Neg Pred Value              1.0000            1.0000           0.9600
Prevalence                  0.2895            0.3421           0.3684
Detection Rate              0.2895            0.3421           0.3421
Detection Prevalence        0.2895            0.3684           0.3421
Balanced Accuracy           1.0000            0.9800           0.9643
```

According to the confusion matrix output applied in the test set, we had **97% accuracy** in the developed neural network model.

Conclusion: Predictive performance of statistical models should be tested by applying appropriate validation technique before final execution in scientific work. Artificial neural network (ANN) modelling technique should be given more emphasis than the conventional modelling approaches as ANN modelling does not require statistical assumptions and can overcome the problem of over fitting, Non-linearity.

References :

- R Development Core Team.(2019) . R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria.
- Rencher,A.C., and R Development Core Team .2004.. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria.ISBN 3-900051-00-3.
- R Development Core Team.2016. R: A Language and Environment for Statistical Computing. *The R Foundation for Statistical Computing*. Vienna, Austria. R version 3.2.4 (2016-03-10).*https://www.Rproject.org*.
- Fisher, R. A. (1936) The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**, Part II, 179–188.