

A Study the new Multivariate Anova and Ancova Techniques by Existing Statistical Techniques in Biostatistics

Jagmohan Singh Dhakar^{1*}, Dr. Sudesh Kumar²

¹ Research Scholar of Sunrise University

² Associate Professor, Sunrise University, Alwar, Rajasthan

Abstract- *Biometrics is a discipline of statistics that involves the use of various computational and scientific approaches to biological research challenges. Biostatisticians collect data on a variety of variables in biological and medical research investigations on a regular basis. Multivariate statistical analysis concerns in understanding different aims and background of each of the different forms of multivariate regression models and how they relate to each other. Any data study including the description of the connection between a response variable and one or more explanatory factors has used regression methods. There exists sufficient literature on applications of Multivariate Statistical tools in Biostatistics. The research work develops some new multivariate ANOVA and ANCOVA techniques for biostatistics involving multiple covariates for different classified data with multiple observations besides the proposed tests for testing sequential contingencies across different groups of various manifold classifications.*

Keywords - *Biometrics, ANOVA and ANCOVA techniques, Multivariate Statistical Techniques, Regression Methods*

-----X-----

INTRODUCTION

Biostatistics is largely concerned with Mathematicians & Statisticians who work in the biological sciences. Biometrical procedures are used frequently by biologists & physicians. Biometrics is a discipline of statistics that involves the use of various computational and scientific approaches to biological research challenges. Biostatisticians collect data on a variety of variables in biological and medical research investigations on a regular basis. These multivariate data are described & analysed using multivariate statistical techniques. Multivariate statistical approaches can be thought of as a generalisation of univariate methods. Researchers in the biological & medical sciences must use multivariate statistical tools to analyse correlations amongst multiple variables, which are intrinsically challenging to apply. Multivariate analysis is naturally challenging for investigators in the biological and medical sciences to comprehend because of the interactions between multiple variables. In order to make inferences using multivariate statistical approaches, more mathematics is required than in a univariate environment. Modern computer software packages, including SPSS version 20, SAS, RATS, SYSSTAT, R-Software, and others, can quickly produce numerical data for multivariate statistical analysis. It's worth noting that many multivariate

statistical approaches are built on the "Multivariate Normal Distribution," which is a basic probability distribution.

Multivariate Statistical Techniques: Categorization & Selection

To make the analysis of complicated data sets easier, multivariate statistical approaches are applied (okluk et al., 2010). Data sets with several independent and dependent variables can be studied using these strategies. Mertler and Vannatta (2005) point out that the simultaneous study of all variables' correlations is a significant advantage. As a result, using univariate statistical approaches to evaluate these associations simultaneously is not possible. Scientific research is far too complicated to be clarified by a single factor. While many aspects influence the problematic while answering a research topic, and the problem to be solved should be assessed in light of these numerous factors. As a result of the limitations of univariate statistics, multivariate statistical studies were developed. As a result, research yields more objective and consistent results, as the presumed limits in univariate statistics are removed. The most significant disadvantage of univariate analyses is that they maintain numerous elements under experimental

control in the event and only assess the influence of one item at a time. Except for some controlled trials, no restrictive or feature can be asserted in multivariate statistical analysis. Since the investigation of many attributes is concerned, multivariate statistical analyses require at least two variables. More and more researchers are turning to multivariate approaches such as structural equation modelling, canonical correlation, multiple regression, conjoint analysis, multivariate discriminant analysis and linear probability models, as well as multiple factor analyses (both explanatory and confirmatory) and cluster analyses, multidimensional scaling analysis, and correspondence analyses.

Analytical methods based on the classification of measured & latent variables as causal & relational variables are called SEMs (correlation-based). SEM is based on the study of hypothesis testing in relation to theoretically built structural models. The foundation of these structural models is a network of causal relationships between variables. Causal links are described using regression equations. Causal equations can be made easier to grasp through the use of schematic representations. When used to the social sciences, behavioural sciences, educational sciences, business, marketing, or health sciences, SEM is a statistical method that relies on a causal and relational characterization of variables that can be observed and those that cannot be (Raykov&Marcoulides, 2000). SEM's broad use now is largely due to the fact that direct and indirect effects between observable & unobserved variables may be examined in a single model (Bryne, 2010). Due to its ability to account for both visible and unobservable effects in one model, SEM has become extremely popular in recent years. It's known as multiple regression analysis (SEM). As a member of the linear model family, SEM modelling can describe complex systems with simultaneous and linked linkages, and it may model interactions amongst non-observable variables. SEM focuses on the relationships between variables and their causes. Thus, it's widely utilized in the social & behavioural sciences (Pang, 1996). When a model is employed to analyse a hypothetical situation, it can generate fictional or relevant data called SEM. There are a lot of models based on real or speculative ideas (Raykov&Marcoulides, 2000, p. 6-7). Research environments are explained and defined by these concepts. SEM is one of a kind because it enables for extremely accurate modelling of measurement error. The SEM can be used to test a theory once it has been created about a circumstance. It's known as validation in SEM applications. Structural models are used in the same way in constructed validity. Measurement methods used in these applications are examined to determine how much of an unobservable variable is recorded (Raykov&Marcoulides, 2000).

Manova & Mancova

MANOVA is a method for simultaneously revealing the associations between two or more metric dependent variables & large number of category independent factors. This is a more advanced variant of a one-variable variance analysis. After the experiment, ANCOVA can be utilized in the final part of the MANOVA to lessen the impact of the uncontrolled metric independent variable on the dependent variable. This is the same as decreasing the influence of the third variable on the bivariate correlation Data sets with two or more variables with a normal distribution & common variables can be utilised to test hypotheses (Ünlükaplan, 2008).

Multiple-variable regression analysis: In multiple regression analysis, there is one dependent variable & many independent variables that are expected to influence this dependent variable. To put it another way, multiple regression is an enhanced version of simple linear regression (Alpar, 2001, p. 132). Changes in independent variables have an effect on dependent variables through the use of multiple regression. This is an excellent strategy for determining the amount of a dependent variable. For illustration, a examiner could evaluation a company's revenues by using advertising spending, the number of salespeople, & number of branches (independent variables) (Yener, 2007).

Probability models based on linear models: The linear probability model incorporates multiple regression and multiple discriminant analysis. Using a high number of independent variables, this method, like multiple regression, estimates a dependent variable. This method differs from multiple regression and is similar to discriminant analysis because the dependent variable is not metric. This approach is comparable to multiple regression, with the exception of the distinction noted. Multiple regression can be used to analyse the nonmetric scale of the dependent variable. In contrast to discriminant analysis, this analysis could use both metric & nonmetric independent variables. If you have more than one dependant variable in your model, you should use discriminant analysis (Tatldil, 1992).

LITERATURE REVIEW

Attila Csala et al. (2019) The state-of-the-art multivariate statistical approaches for high-dimensional multisetomics data analysis are covered in this chapter. Recent biotechnological advancements have enabled large-scale measurement of diverse biomolecular data spread over many omics domains, such as genotypic and phenotypic data. A new research path is to use an integrated method to study different data sources in order to better model and understand the underlying biology of complicated illness states. This chapter provides an overview of some recent advances

in multivariate techniques for high-dimensional omics data analysis, with a focus on two well-known multivariate methods, canonical correlation analysis (CCA) and redundancy analysis (RDA). In the realm of omics data analysis, penalised variants of CCA are common, and there has been new work on multisetpenalised RDA that is relevant to multisetomics data. These methods are discussed in terms of how they address the statistical issues that come with high-dimensional multisetomics data processing and how they contribute to our understanding of the human condition in terms of health and disease.

David Núñez-Alonso, et al. (2019) From 2010 to 2017, 22 monitoring stations in Madrid city and province were used to report on the distribution of pollutants in the city and province. Air pollution data was interpreted and modelled using statistical methods. The data includes yearly average nitrogen oxide, ozone, and particle matter (PM10) concentrations gathered in Madrid and its suburbs, which is one of Europe's largest metropolitan areas and whose air quality has not been adequately researched. A map of the distribution of these contaminants was created to demonstrate the relationship between them as well as the region's population. Correlation analysis, PCA, and cluster analysis (CA) were used in a multivariate analysis to establish a correlation between different contaminants. The findings allowed separate monitoring stations to be classified based on each of the four pollutants, exposing information about their sources and methods, displaying their spatial distribution, and monitoring their levels according to the legislation's average yearly restrictions. The conclusion generated from the multivariate analysis indicating NO₂ levels surpassing the yearly limit in the centre, south, and east of the Madrid province was also corroborated by the development of contour maps using the geostatistical approach of ordinary kriging.

Avijit Hazra et al. (2017) Multivariate analysis is a statistical approach that examines three or more variables in connection to the subjects under inquiry at the same time in order to discover or clarify links between them. Dependence techniques, which look at the correlation between one or more dependent variables and their independent predictors, & interdependence techniques, which don't make that distinction and treat all variables equally in their search for underlying relationships, are two types of these techniques. A situation in which a single numerical dependent variable is to be predicted from many numerical independent variables is modelled by multiple linear regression. While the outcome variable is dichotomous, logistic regression is utilised. The log-linear technique can be used to evaluate cross-tabulations with more than two variables because it models count data. An expansion of ANOVA, analysis of

covariance incorporates an extra independent variable of interest, the covariate, into the study. It attempts to determine if a difference exists after "controlling" for the effects of a covariate on the numerical dependent variable of interest. When many numerical dependent variables must be included in the study, MANOVA is a multivariate extension of ANOVA. Psychometrics, social sciences, & market research are the most common uses of interdependence methods. Exploratory factor analysis and principal component analysis are two similar techniques that aim to extract a smaller number of composite factors or components from a larger number of metric variables that are linearly connected to the original variables. Cluster analysis seeks to find reasonably homogeneous groupings called clusters in a large number of examples without any prior knowledge of the groups.

Dr. Sateesh Kumar Ojha et al. (2016) If variables are not correctly examined, we often get misled conclusions in research. All of the latent and observable variables must be correctly understood in order for management decisions to be relevant and effective in various functional areas of management. The purpose of this work is to look into the usage of various multivariate tools for analysing in management research, whether they are applied or basic. Data comes from both original and secondary sources. The first step is observing various research articles published in the proceedings of various conferences. The secondary section contains a variety of multivariate analysis-related papers. The investigation uncovered the reasons behind the lack of use of such research techniques. According to the preliminary findings, the majority of studies do not make extensive use of such analytical methods. The main cause for not applying proper design is carelessness in design while addressing the design aspect.

OBJECTIVES

1. To study the various multivariate statistical tools for Biostatistics existing in the literature
2. To study the univariate & multivariate ANOVA and ANCOVA techniques for different types of data

HYPOTHESIS

1. There will be no significance the various multivariate statistical tools for Biostatistics existing in the literature
2. There will be no significance difference between the univariate & multivariate ANOVA and ANCOVA techniques for different types of data

RESEARCH METHODOLOGY

The major purpose of this research is to create some new multivariate statistical procedures by changing existing biostatistical statistical techniques. Many researchers have used univariate statistical techniques in biostatistics, but just a few have used multivariate statistical methods at the fundamental level. The current study makes an attempt to analyse numerous research difficulties in the biological and medical sciences by suggesting some advanced multivariate statistical techniques based on the Multivariate Analysis of Covariance (MANCOVA) technique. The goal of this work is to use the generalised weighted least squares approach to estimate the parameters of a multivariate logistic regression model.

DATA ANALYSIS AND INTERPRETATION

- **Multivariate Multiple Regression Analysis**

Let's look at an example of how to model the relationship between m responses (Y_1, Y_2, \dots, Y_m) and a set of predictor variables (z_1, z_2, \dots, z_r) that are linked to each other. It is assumed that each response would follow its own regression model, hence

$$\left. \begin{aligned} Y_1 &= \beta_{01} + \beta_{11}z_1 + \dots + \beta_{r1}z_r + \epsilon_1 \\ Y_2 &= \beta_{02} + \beta_{12}z_1 + \dots + \beta_{r2}z_r + \epsilon_2 \\ &\vdots \\ Y_m &= \beta_{0m} + \beta_{1m}z_1 + \dots + \beta_{rm}z_r + \epsilon_m \end{aligned} \right\}$$

The error vector $\epsilon = \epsilon_1, \dots, \epsilon_m$ has $E(\epsilon) = 0$ and $\text{var}(\epsilon) = \Sigma$. There may be a correlation between the error terms associated with various replies. If you want to use the classical linear regression model's nomenclature, you'll need to use $[z_1, z_2, \dots, z_r]$ to express the values of the predictor variables, or $Y_j = Y_{j1}, \dots, Y_{jm}$ for the responses and $\epsilon_j = \epsilon_{j1}, \dots, \epsilon_{jm}$ for the errors. The design matrix is represented in matrix notation.

$$Z_{(n \times (r+1))} = \begin{bmatrix} Z_{10} & Z_{11} & \dots & Z_{1r} \\ Z_{20} & Z_{21} & \dots & Z_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{n0} & Z_{n1} & \dots & Z_{nr} \end{bmatrix}$$

is exactly the identical as the single-response regression model. There are multivariate analogues for the other matrix quantities as well Set

$$Y_{(n \times m)} = \begin{bmatrix} Y_{11} & Y_{12} & \dots & Y_{1m} \\ Y_{21} & Y_{22} & \dots & Y_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n1} & Y_{n2} & \dots & Y_{nm} \end{bmatrix} = [Y_{(1)} \quad Y_{(2)} \quad \dots \quad Y_{(m)}]$$

$$\beta_{((r+1) \times m)} = \begin{bmatrix} \beta_{01} & \beta_{02} & \dots & \beta_{0m} \\ \beta_{11} & \beta_{12} & \dots & \beta_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{r1} & \beta_{r2} & \dots & \beta_{rm} \end{bmatrix} = [\beta_{(1)} \quad \beta_{(2)} \quad \dots \quad \beta_{(m)}]$$

$$\epsilon_{(n \times m)} = \begin{bmatrix} \epsilon_{11} & \epsilon_{12} & \dots & \epsilon_{1m} \\ \epsilon_{21} & \epsilon_{22} & \dots & \epsilon_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \epsilon_{n1} & \epsilon_{n2} & \dots & \epsilon_{nm} \end{bmatrix} = [\epsilon_{(1)} \quad \epsilon_{(2)} \quad \dots \quad \epsilon_{(m)}]$$

$$= \begin{bmatrix} \epsilon_{(1)}^T \\ \vdots \\ \epsilon_{(2)}^T \\ \vdots \\ \epsilon_{(m)}^T \end{bmatrix}$$

The multivariate linear regression model

$$Y_{(n \times m)} = Z_{(n \times (r+1))} \beta_{((r+1) \times m)} + \epsilon_{(n \times m)}$$

with $E(\epsilon_{(i)}) = 0$ and $\text{cov}(\epsilon_{(i)}, \epsilon_{(k)}) = \sigma_{ik}I$; $i, k = 1, 2, \dots, m$

The covariance matrix of the m observations on the j th trail $\Sigma = \sigma_{ik}$, but the observations from the various trails are unconnected. The design matrix Z has j th rows $[z_{j0}, z_{j1}, \dots, z_{jr}]$ and has unknown parameters β and σ_{ik} . The i th response, $Y(i)$, follows a linear regression model, to put it simply.

$$Y_{(i)} = Z \beta_{(i)} + \epsilon_{(i)}; \quad i = 1, 2, \dots, m$$

With $\text{cov}(\epsilon_i) = \sigma_{\mu i}$. However, the errors for different responses on the same trial can be correlated.

Given the outcomes Y and the values of the predictor variables Z with full columnrank, one can determine the least squares estimates $\hat{\beta}_i$ exclusively from the observations $Y(i)$ on the i th response. In conformity with the single response solution, one may take

Likelihood Ratio Tests for Regression Parameters

The hypothesis that the responses do not based on $Z_{q+1}, Z_{q+2}, \dots, Z_r$ becomes

$$H_0: \beta_{(2)} = 0 \text{ where } \beta = \begin{bmatrix} \beta_{(1)} \\ \dots \\ \beta_{(2)} \end{bmatrix}$$

Setting $Z = \begin{bmatrix} Z_1 & \dots & Z_2 \end{bmatrix}$ one can write the general model as

$$E(Y) = Z\beta = [Z_1 \quad \dots \quad Z_2] \begin{bmatrix} \beta_{(1)} \\ \dots \\ \beta_{(2)} \end{bmatrix} = Z_1\beta_{(1)} + Z_2\beta_{(2)}$$

Under $H_0: \beta_{(2)} = 0$, $Y = Z_1\beta_{(1)} + \epsilon$ and the likelihood ratio test of H_0 is based on the quantities involved in the extra sum of squares and cross products

$$\begin{aligned} &= (Y - Z_1\hat{\beta}_{(1)})^T (Y - Z_1\hat{\beta}_{(1)}) - (Y - Z\hat{\beta})^T (Y - Z\hat{\beta}) \\ &= n(\hat{\Sigma}_1 - \hat{\Sigma}) \end{aligned}$$

where $\hat{\beta}_{(1)} = (Z_1^T Z_1)^{-1} Z_1^T Y$ and $\hat{\Sigma}_1 = n^{-1} (Y - Z_1\hat{\beta}_{(1)})^T (Y - Z_1\hat{\beta}_{(1)})$

The likelihood ratio, taken in terms of generalized variances:

$$\Lambda = \frac{\max_{\beta, \Sigma} L(\beta, \Sigma)}{\max_{\beta, \Sigma} L(\beta, \Sigma)} = \frac{L(\hat{\beta}_{(1)}, \hat{\Sigma}_1)}{L(\hat{\beta}, \hat{\Sigma})} = \left(\frac{|\hat{\Sigma}|}{|\hat{\Sigma}_1|} \right)^{\frac{n}{2}}$$

Equivalently, Wilks' lambda statistic

$$\Lambda^n = \frac{|\hat{\Sigma}|}{|\hat{\Sigma}_1|}$$

Could be utilized.

Result: Let the multivariate multiple regression model hold with Z for full rank $r+1$ and $(r+1) + m \leq n$. Let the errors E be normally distributed. Under $H_0: \beta_{(2)} = 0$, $n\hat{\Sigma}$ is distributed as $W_{p, n-r-1}(\Sigma)$ independently of $n(\hat{\Sigma}_1 - \hat{\Sigma})$ which, in turn, is distributed as $W_{p, r-q}(\Sigma)$. The likelihood ratio test of H_0 is equivalent to rejecting H_0 for large values of

$$-2 \ln \Lambda = -n \ln \left(\frac{|\hat{\Sigma}|}{|\hat{\Sigma}_1|} \right) = -n \ln \frac{|n\hat{\Sigma}|}{|n\hat{\Sigma} + n(\hat{\Sigma}_1 - \hat{\Sigma})|}$$

For large n , the modified statistic

$$-\left(n - r - 1 - \frac{1}{2}(m - r + q + 1) \right) \ln \left(\frac{|\hat{\Sigma}|}{|\hat{\Sigma}_1|} \right)$$

Chi-square distribution with degrees of freedom of $(n - r - q)$ is an approximation of this distribution.

Multiple Variable Regression Predictions

If the model $Y = Z\beta + \epsilon$ has been fitted & examined for inadequacies, it will have normal errors. When a model is accurate, it can be used for forecasting.

Predicting the mean responses to fixed values of the predictor variables is a challenge. The distribution theory could be used to draw conclusions about the median responses. As a result, it is clear that

$\hat{\beta}^T Z_0$ is distributed as $N_m(\beta^T Z_0, Z_0^T (Z^T Z)^{-1} Z_0 \Sigma)$

and $n\hat{\Sigma}$ is independently distributed as $W_{n-r-1}(\Sigma)$.

The unknown value of the regression function at is

$Z_0 \beta$. So, from the T^2 -statistic, one can write

$$T^2 = \left(\frac{\hat{\beta}^T Z_0 - \beta^T Z_0}{\sqrt{Z_0^T (Z^T Z)^{-1} Z_0}} \right)^T \left(\frac{n}{n-r-1} \hat{\Sigma} \right)^T \left(\frac{\hat{\beta}^T Z_0 - \beta^T Z_0}{\sqrt{Z_0^T (Z^T Z)^{-1} Z_0}} \right)$$

As $100(1-\alpha)\%$ confidence ellipsoid for $\beta^T Z_0$ is provided by the inequality

$$(\hat{\beta}^T Z_0 - \beta^T Z_0)^T \left(\frac{n}{n-r-1} \hat{\Sigma} \right)^{-1} (\hat{\beta}^T Z_0 - \beta^T Z_0) \leq Z_0^T (Z^T Z)^{-1} Z_0 \left[\left(\frac{m(n-r-1)}{n-r-m} \right) F_{m, n-r-m}(\alpha) \right]$$

If $F_{m, n-r-m}(\alpha)$ is the upper $(100\alpha)\%$ percentile of an F-distribution with m & $n-r-m$ degrees of freedom

The $100(1-\alpha)\%$ simultaneous confidence intervals for $E(Y_i) = Z_0^T \beta_{(i)}$ are

$$Z_0^T \hat{\beta}_{(i)} \pm \sqrt{\left(\frac{m(n-r-1)}{n-r-m} \right) F_{m, n-r-m}(\alpha)} \sqrt{Z_0^T (Z^T Z)^{-1} Z_0 \left(\frac{n}{n-r-1} \hat{\sigma}_{ii} \right)},$$

where $\hat{\beta}_{(i)}$ is the i^{th} column of $\hat{\beta}$ and $\hat{\sigma}_{ii}$ is the i^{th} diagonal

The second prediction problem is concerned with forecasting new responses

$Y_0 = \beta^T Z_0 + \epsilon_0$ at Z_0 . Here ϵ_0 is independent of ϵ . Now,

$Y_0 - \hat{\beta}^T Z_0 = (\beta - \hat{\beta})^T Z_0 + \epsilon_0$ is distributed as $N_m(0, (1 + Z_0^T (Z^T Z)^{-1} Z_0) \Sigma)$

independently of $n\hat{\Sigma}$, so that 100(1-a)% prediction ellipsoid for Y_0 becomes

$$(y_0 - \hat{\beta}'z_0)' \left(\frac{n}{n-r-1} \hat{\Sigma} \right)^{-1} (y_0 - \hat{\beta}'z_0) \leq (1 + z_0'(Z'Z)^{-1}z_0) \left[\left(\frac{m(n-r-1)}{n-r-m} \right) F_{m,n-r-m}(\alpha) \right]$$

The 100(1-a)% simultaneous prediction intervals for the individual responses Y_{0i} are

$$z_0' \hat{\beta}_{(i)} \pm \sqrt{\left(\frac{m(n-r-1)}{n-r-m} \right) F_{m,n-r-m}(\alpha)} \sqrt{1 + z_0'(Z'Z)^{-1}z_0} \left(\frac{n}{n-r-1} \hat{\sigma}_{ii} \right), i=1,2,\dots,m$$

where $\hat{\beta}_{(i)}$, $\hat{\sigma}_{ii}$, and $F_{m,n-r-m}(\alpha)$ are the same quantities appearing. A comparison reveals that the intervals for predicting actual response variable values are larger than those for predicting predicted values. There is a random error ϵ , which is shown by the additional width.

• **Canonical Correlation Analysis**

Analyzing the canonical correlation between two sets of variables aims to find and quantify the relationships between them. To illustrate how arithmetic speed & power relate to reading speed & power, H. Hotelling (1935, 1936) first proposed the concept.

When doing a canonical correlation analysis, researchers look at the correlation between a linear combination of variables in one set and another set of variables. Before moving on, it is important to figure out which pair of linear combinations has the greatest correlation in terms of correlation. One can then discover the linear combination that is most correlated with the initial pair of linear combinations. It is the canonical variables with their canonical correlations that are referred to as the canonical terms.

• **Anova Technique for three way Classified Data**

When a data is classified into different levels 1, m and n levels of three factors A, B and C respectively in 1 x m x n table such that each cell may contain either one or more observations, then that forms a three way classified data. Based on the number of observations in the cells, one may discuss the following two types of ANOVA for three way classified data (i) ANOVA for three way classified data with single observation per cell (ii) ANOVA for three way classified data with multiple but equal number of observations per cell.

• **Anova Technique for two way Classified Data**

When a data is classified into different levels r and s levels of two factors A & B respectively in r x s table such that each cell may contain either one or more

observations, then that forms a two way classified data. Based on the number of observations in the cells, one can have the following four types of ANOVA for two way classified data.

- ANOVA for two way classified data with single observation per cell
- ANOVA for two way classified data with multiple but equal number of observations per cell
- ANOVA for two way classified data with unequal number of observations in the cells with no interaction
- ANOVA for two way classified data with unequal number of observations in the cells with interaction (very rare)
- **Anova Technique for one way Classified Ranked Data**

An alternative approach to the F test analysis of variance that does not rely on the assumption of normality may be used in cases where the normality assumption is incorrect. Kruskal & Wallis came up with this method (1952). Only two populations can be compared using the Wilcoxon rank sum test. It is a generalization of the Wilcoxon test, the Kruskal-Wallis test. If the k treatments are all identical, the Kruskal-Wallis test is performed to determine whether or not there are any observations that are statistically significantly larger than the rest of the samples. The Kruskal-Wallis test can be thought of as a measure of treatment quality because of the procedure's focus on looking for variations in means. By ranks, it is a one-way ANOVA. Non-parametric Kruskal-Wallis testing is an alternative to standard variance analysis.

CONCLUSION

The literature on Biostatistics focuses only on practical applications of ANOVA and ANCOVA techniques and entirely ignores theoretical aspects of MANOVA and MANCOVA techniques. The statistical examination of correlations between many variables, especially three or more, is the focus of multivariate statistical approaches. In the present study, the various Multivariate statistical tools in Biostatistics. Besides basic univariate ANOVA and ANCOVA techniques for one way, two way and three way classified data, the Multivariate ANOVA (MANOVA) and Multivariate ANCOVA (MANCOVA) techniques have been provided in a systematic manner. The present research study has brought out some new multivariate statistical techniques for Biostatistics by developing theoretical aspects of various ANCOVA and MANCOVA techniques for different types of data involving single and multiple covariate, besides a comparison test for sequential connection across various groups of manifold contingency tables involving several attributes.

REFERENCES

1. Bryne, B. (2010). Structural Equation Modelling with AMOS: Basic Concept, Applications and Programming, Second Edition, New Jersey, USA: Lawrence Erlbaum Associates Publisher.
2. Buhlmann, P. and Yu, B. (2003). Boosting with the L2 loss: regression and classification, *Journal of the American Statistical Association*, **98**, 324–339.
3. Buja, A. and Swayne, D.F. (2002). Visualization methodology for multidimensional scaling, *Journal of Classification*, **19**, 7–43.
4. Buja, A., Swayne, D.F., Littman, M.L., Dean, N., Hofmann, H., and Chen, L. (2008). Data visualization with multidimensional scaling, *Journal of Computational and Graphical Statistics*, **17**, 444–472.
5. Buntine, W.L. and Weigend, A.S. (1991). Bayesian back-propagation, *Complex Systems*, **5**, 603–643.
6. Burdick, R.K. and Graybill, F.A. (1992), "Confidence intervals on variances components", Marcel Decker, New York.
7. Burges, C.J.C. (1998). A tutorial on support vector machines for pattern recognition, *Data Mining and Knowledge Discovery*, **2**, 121–167.
8. Bursac, Z., Gauss, H.C., Williams, D.K. and Homer, D.W. (2008), "Purposeful Selection of Variables in Logistic Regression", *Sources Code for Biology and Medicine*, 16,3-17.
9. Burt, C. (1950). The factorial analysis of qualitative data, *British Journal of Psychology, Statistics Section* **3**, 166-185.
10. Calinski, T. and Kageyama, S. (2000), "Block design: A Randomization approach", Vol. 1: Analysis, New York : Springer.
11. Canary, J. (2013), "A Comparison of Three Goodness of Fit Tests for the Logistic Regression Model", Unpublished Doctoral Dissertation, University of Tasmania, Hobart Tasmania, Australia.
12. Chen, C.C. (2001). "Extended Rank Analysis of Covariance as a Most Efficient Matched Analysis Considering Trend Information", *Biometric Journal*, Vo1.43, Issue.7, pp. 895-907.

Corresponding Author

Jagmohan Singh Dhakar*

Research Scholar of Sunrise University