International Journal of
Statistics and Applied Mathematics

ISSN: 2456-1452
Maths 2018; 3(1): 436-441
© 2018 Stats & Maths
www.mathsjournal.com
Received: 04-11-2017
Accepted: 09-12-2017

P Srivyshnavi
Senior Assistant Professor, Department of CSE, SPMVV Engineering College, Tirupati, Andhra Pradesh, India

SK Nafeez Umar
Assistant Professor, Department of Statistics and Mathematics, ANGR Agricultural University, Bapatla, Andhra Pradesh, India

G Mokesh Rayalu
Assistant Professor, Department of Mathematics, VIT, Vellore, Tamil Nadu, India

M Jahnavi
Research Scholar, Department of Business Management, SPMVV, Tirupati, Andhra Pradesh, India

M Venkataramaniah
Rtd. Professor, Professor, Department of Statistics, S.V University, Tirupati, Andhra Pradesh, India

P Balasiddamuni
Rtd. Professor, Professor, Department of Statistics, S.V University, Tirupati, Andhra Pradesh, India

Some advanced statistical tools for data analysis of research projects

P Srivyshnavi, SK Nafeez Umar, G Mokesh Rayalu, M Jahnavi, M Venkataramaniah and P Balasiddamuni

Abstract

Research is a method or process to discover truth based on practical or critical thinking. A plan to do research is a systematic manner is called Research Design. One of the important steps involved in planning Research design is 'Data Analysis of Projects' by using Statistical Tools or Techniques. Besides the basic statistical measures such as measures of central tendency and measures of variation, statistical inferential procedures will help to draw valid conclusions about the characteristics of the population on the basis of sample observations. The major areas of statistical inference are 'Theory of Estimation' and 'Theory of Testing the Hypothesis'.

Researchers in the fields of Applied Sciences frequently collect measurements on several variables. Generally, univariate statistical tools may not be suitable to analyze multivariate data. Some advanced multivariate statistical tools which are analogous to univariate statistical tools are essential to analyze multivariate data. Multivariate analysis is an inherently difficult subject to understand by applied research workers. More mathematics is required to derive multivariate statistical techniques for drawing inferences than in a univariate setting.

Modern computer software statistical packages such as SPSS version-21, SAS, QSB, RATS, R-software etc., readily provide the numerical results to rather complex multivariate statistical analysis.

In the present research article, some advanced statistical tools for data analysis of Research Projects have been proposed for the purpose of Research workers in applied sciences.

Keywords: Some advanced statistical, systematic manner, research design

1. Introduction

Research is a process for the discovery of truth, which is a method of practical thinking. Research project depicts a scheme for research under research project, a plan of action known as 'Research Design' may be proposed to do research in a systematic manner. The various steps involved in planning the Research Design are: Selection of Research Problem; Defining the variables; Objectives of the research study; Formulation of Hypotheses; Collection of sample data; Classification and Tabulation of Collected data; Analysis of data by using statistical techniques; and Interpretation about the research study based on results obtained from the analysis.

The main task of statistician is to apply some statistical tools to analyse data collected under research project.

In the present research article, an attempt has been made by discussing some advanced statistical techniques together with simplified formulae, which have a wide number of applications in the data analysis of various research projects in almost all the fields of science like Biology, Medicine, Agricultural Science, Psychology, Economics, Business, Technology and Criminology etc.

2. Some important basic statistical measures

(A) Measures of central tendency

The concentration of the given observations in the central part of the data or frequency distribution is called 'Central Tendency'.

Correspondence
G Mokesh Rayalu
Assistant Professor, Department of Mathematics, VIT, Vellore, Tamil Nadu, India

A measure which measures the concentration of the observations in the central part of the data is known as a ‘Measure of Central Tendency’ or an ‘Average’ or a ‘Measure of Location’ or ‘Central Value’,
Eg: Arithmetic Mean, Median, Mode, Geometric Mean, Harmonic Mean etc.

Arithmetic Mean ($\bar{X}$)
It is an ideal measure of central tendency. For unclassified data, it is defined as,

$$\overline{\mathrm{X}} = \frac{\sum \mathrm{x}}{\mathrm{n}} = \frac{\text{Sum of observations in the data}}{\text{Total number of observations}}$$

For a frequency distribution,
$\overline{\mathrm{X}} = \frac{\sum \mathrm{fx}}{\mathrm{N}}$, where $\sum fx$ = Sum of products of variables values (Mid values of classes in the case of continuous frequency distribution) and their corresponding frequencies.
N = Total frequency

(B) Measures of variation or dispersion
Measures of central tendency are generally inadequate to give us complete idea about the data. They must be supplemented by some other measures which are known as Measures of Dispersion. Dispersion means the scatteredness of the observations in the given data. A measure, which measures the scatteredness of the observations around the central value (Measure of central tendency), is called a Measure of dispersion.
Eg: Range, Quartile Deviation, Mean Deviation, Standard Deviation, Coefficient of Variation etc.

Standard Deviation (σ)
It is an ideal measure of dispersion. It is defined as the positive square root of their arithmetic mean of the squares of the deviations of the given observations from their arithmetic mean. It is denoted by σ.

For unclassified data, $\sigma = \sqrt{\frac{\sum(X-\overline{X})^2}{n}} = \sqrt{\frac{\sum X^2}{n} - (\bar{X})^2}$
Here, $\bar{X} = \frac{\sum X}{n}$
For frequency distribution $\sigma = \sqrt{\frac{\sum f(X-\overline{X})^2}{N}} = \sqrt{\frac{\sum fX^2}{N} - (\bar{X})^2}$
Here, $\bar{X} = \frac{\sum fX}{n}$
X’s are the values of the variable (mid values of classes in the case of continuous frequency distribution.
N is the total frequency.

Coefficient of Variation (C.V.)
It is a relative measure of dispersion based on $\bar{X}$ and σ. It is useful to compare the variations in two or more series of data. The data with smaller value of coefficient of variation is known as ‘Consistent Data’.
C.V. is defined as
C.V. = $\left(\frac{\sigma}{\bar{X}}\right) 100$.

(C) Measures of Skewness
Skewness means ‘Lack of symmetry’ or asymmetry in the data. A measure of skewness gives an idea about the shape of the frequency curve for the given data. For a symmetric data, the values of Arithmetic Mean, Median and Mode are all equal. Some important measures of skewness are given by:

Karl Pearson’s Coefficient of Skewness (S_k)

It is defined as $S_k = \frac{Mean-Mode}{Standard\ Deviation}$

or $S_k = \frac{3(Mean-Median)}{Standard\ Deviaiton}$ (When Mode is ill defined)

[Mean > Mode] or [Mean > Median] ⟹ Positive skewness in the data

[Mean < Mode] or [Mean < Median] ⟹ Negative skewness in the data

[Mean = Median = Mode] or $S_k = 0$ ⟹ Symmetry in the data
The value of S_k lies between (-3, 3)

Bowley’s coefficient of Skewness of Skewness (S_k)
It is defined as

$$S_k = \left[\frac{Q_3-Q_1+2Q_2}{Q_3-Q_1}\right]$$

Here, Q_1, Q_2 and Q_3 are first, second and third quartiles respectively. The value of S_B lies between [-1and +1].

(D) Measures of correlation between two variables
A bivariate distribution or data (data on two variables simultaneously), correlation analysis refers to the relationship between two or more variables. If the change in one variable results corresponding change in other variable, then one may say that there is correlation between two variables.
The frequently used measures of correlation are:
1. Scatter Diagram Method
2. Karl pearson’s Coefficient of Correlation (r)
3. Spearman’s Coefficient of Rank Correlation (ρ)
4. Kendall’s Coefficient of Concurrent Deviation (τ) Method

Karl Pearson’s coefficient of correlation between two variables X and Y is given by

$$r_{XY} = \frac{cov(X,Y)}{\sqrt{V(X)V(Y)}}$$
or
$$r_{XY} = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{\left[\sum X^2 - \frac{(\sum X)^2}{n}\right]\left[\sum Y^2 - \frac{(\sum Y)^2}{n}\right]}}$$

where $cov(X,Y) = \frac{1}{n}\left[\sum XY - \frac{(\sum X)(\sum Y)}{n}\right]$ = Covariance between X and Y

$V(X) = \frac{1}{n}\left[\sum X^2 - \frac{(\sum X)^2}{n}\right]$ = Variance of X

and $V(Y) = \frac{1}{n}\left[\sum Y^2 - \frac{(\sum Y)^2}{n}\right]$ = Variance of Y

The value of r always lies between 0 and 1.

3. Advanced statistical tools for data analysis
Application of advanced statistical tools for data analysis of research projects depends on the availability and reliability of statistical data. Some important types of statistical data can be exist in the literature are:
1. Univariate statistical data
2. Bivariate statistical data
3. Multivariate statistical data

4. Time series data
5. Cross section data
6. Pooling of Time series and Cross section data or Panel data

Some important advanced statistical tools to analyse different types of statistical data are given by

1. Simple Linear Regression Analysis
2. Multiple Linear Regression Analysis
3. Multiple and Partial Coefficients of Correlation
4. Large Sample and Small Sample tests of Significance: z, t, χ^2 and F tests
5. Analysis of Variance (ANOVA) technique
6. Nonlinear Regression Analysis: Exponential and Polynomial Regressions
7. Nonparametric Tests of significance
8. Multiple R^2 and $\bar{R}^2$ Criteria
9. Analysis of Covariance (ANCOVA) technique
10. Kruskal – Wallis Test for one way classified ranked data
11. Friedman's Test for two way classified ranked data
12. Bartlett's Test
13. Hotelling's T^2 – test
14. Analysis of Dispersion: Wilks Lambda Criterion
15. Multivariate ANOVA (MANOVA) technique
16. Multivariate ANCOVA (MANCOVA) technique
17. MANOVA technique for Ranked data
18. Stepwise Regression for selection of variables: Beta Weights and Variable Inflation Factor (VIF)
19. Principal Components Analysis
20. Path coefficients Analysis
21. Factor Analysis
22. Cluster Analysis
23. Discriminant Analysis: Mahalanobis D^2 – Test
24. Canonical Correlation Analysis
25. Generalized Regression Analysis
26. Sets of Linear Regression Analysis or Seemingly Unrelated Regression Equations (SURE) Analysis
27. Linear and Compound Growth Rates and their Tests of Significance
28. Chow Test for Structural Change
29. Tests for Autocorrelation: Durbin – Watson Test
30. First order / Second order / p^{th} order Autoregressive Analysis
31. First order / Second order / q^{th} order Moving Average Regression Analysis
32. Advanced Time Series Regression Analysis: Vector Autoregressive Analysis
33. Advanced Forecasting Techniques
34. Random Coefficients Regression (RCR) Analysis
35. Multifactor Conjoint Analysis
36. Partial Linear Regression Analysis
37. Univariate Logistic and Poisson Regression Analysis
38. Multiple Logistic and Multiple Poisson Regression Analysis
39. Tests for Manifold Contingency Tables involving several attributes
40. Operations Research Techniques
41. Experimental Designs: Complete Block and Incomplete Block experimental designs
42. Simulation Techniques
43. Neural Networks
44. Diagnostic Tests
45. Bootstrap Techniques
46. Tests for Normality
47. Circular Data Analysis
48. Statistical Production Function Analysis
49. Model Selection Criteria
50. Advanced Applied Regression Analysis etc.

4. Simple time series data analysis

(A) Computation of Linear Growth Rate and Its Test of Significance

Suppose that there exists a linear relationship between a study variable (Y) and a time variable (t) as

$y_i = a + bt_i$, i = 1, 2, ..., n

By using coded time variable X in the place of t, the linear model can be written as

$Y_i = a + bX_i$, i = 1, 2, ..., n

or simply, Y = a + bX.
By adding an error variable ∈, one may write statistical linear regression model as
Y = a + bX + ∈
where, Y = Dependent variable (study variable)
X = Independent variable (coded time variable)
and a, b are the parameters of the linear model.
The Least Squares estimates (Best estimates) of a and b are given by

$$\hat{b} = \left[\frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}}\right]$$

and $\hat{a} = \bar{Y} - \hat{b}\bar{X}$
Here, $\bar{X} = \frac{\sum X}{n}$, $\bar{Y} = \frac{\sum Y}{n}$ and n is the number of observations.
The estimated linear model is given by $\hat{Y} = \hat{a} + \hat{b}X$. This estimated model will be used for the prediction analysis.
An estimate of LGR is now given by

$$LGR = \left[\frac{\hat{b}}{\bar{Y}}\right] \times 100.$$

Test for significance of linear growth rate

To test for the significance of LGR, we use the following student's t-test statistic:

$$t = \frac{\hat{b}\sqrt{\sum X^2 - \frac{(\sum X)^2}{n}}}{\hat{\sigma}}$$

where $\hat{\sigma} = \sqrt{\frac{\left[\sum Y^2 - \frac{(\sum Y)^2}{n}\right] - \hat{b}\left[\sum XY - \frac{(\sum X)(\sum Y)}{n}\right]}{n-2}}$

One may compare the calculated value of $|t|$ with its critical value (table value) for (n-2) degrees of freedom at an appropriate level of significance and draw the inference accordingly.

(B) Computation of Compound Growth Rate and Its Test of Significance

Consider the formula for compound interest as

$$Y_n = Y_0\left(1 + \frac{r}{100}\right)^n.$$

One may replace Y_n by y; Y_0 by 'a', $\left(1 + \frac{r}{100}\right)$ by 'b'; and n by a coded time variable X and then write an exponential functional relationship between y and X as

$y = a\, b^X$

where y = Dependent variable (study variable) X = Independent variable (coded time variable) and a, b are the parameters of the exponential model.

Since, directly fitting of exponential model includes several difficulties, one may generally transform the model into a log-linear form and then it may be fit to the data.

Taking logarithms on both sides of the model, one may get

log y = log a + X log b

or Y = A + B X, which is a linear model.

Here, Y = log y, A = log a, B = log b

The Least Squares estimates of A and B are given by

$$\hat{B} = \left[\frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}}\right]$$

and $\hat{A} = \bar{Y} - \hat{B}\bar{X}$

Here, n is the number of observations.

The estimated linear model is given by $\hat{Y} = \hat{a} + \hat{b}X$.

An estimate of original parameter b is given by

$\hat{b} = Antilog(\hat{B})$.

Thus, $\left(1 + \frac{r}{100}\right) = \hat{b}$.

Hence, an estimate of CGR is now given by

$CGR = \hat{r} = [\hat{b} - 1]100$

Test of significance of CGR

To test for the significance of CGR, one may use the following students's t-test statistic:

$$t = \frac{\hat{B}\sqrt{\sum X^2 - \frac{(\sum X)^2}{n}}}{\hat{\sigma}}$$

where $\hat{\sigma} = \sqrt{\frac{\left[\sum Y^2 - \frac{(\sum Y)^2}{n}\right] - \hat{B}\left[\sum XY - \frac{(\sum X)(\sum Y)}{n}\right]}{n-2}}$

One may compare the calculated value of $|t|$ with its critical value for (n-2) degrees of freedom at an appropriate level of significance and draw the inference accordingly.

Remarks

1. Suppose $\widehat{b_1}$ and $\widehat{b_2}$ be the estimates of parameters of linear models of two time series data respectively. The linear growth rates of the two time series data can be compared by using the following student's t-test statistic as

$$t = \frac{\widehat{b_1} - \widehat{b_2}}{\sqrt{\left[S.E(\widehat{b_1})\right]^2 + \left[S.E(\widehat{b_2})\right]^2}}$$

where, S.E. denotes the standard error of estimate.

We compare $|t_{cal}|$ values with the critical value of t test statistic for ($n_1 + n_2 - 4$) degrees of freedom at a desired level of significance and draw the inference accordingly.

2. Suppose one may wish to investigate whether there is any change in the growth of production between war time and peace time periods (or any two different time periods). Such a change is referred to as a 'Structural Change in the growth'. It can be tested by using a test given in 'Applied Regression Analysis', known as Chow test for structural change.

5. Chow test for structural change in two time series data

Suppose that there are two regressions, representing observations in two countries for a cross section study in two different periods for time series study. One may ask whether the behaviour of two countries or in the two different time periods differs by testing the null hypothesis as

H$_0$: The regression coefficients in the two regression models are equal for the two samples.

Consider two general linear regression models in matrix notation for two samples of n_1 and n_2 observations respectively as

$$Y_1 = X_1\beta_1 + \epsilon_1 \qquad (5.1)$$
$$Y_2 = X_2\beta_2 + \epsilon_2 \qquad (5.2)$$

where Y_1 is ($n_1 \times 1$), Y_2 is ($n_2 \times 1$);

X_1 is ($n_1 \times k$), X_2 is ($n_2 \times k$); and

β_1 is ($k \times 1$), β_2 is ($k \times 1$) matrices.

The null hypothesis can be written as

H$_0$: $\beta_1 = \beta_2$ i.e., there is no structural change in two countries data or in two time series data.

Now, one may write,

(i) Unrestricted Linear Regression model in matrix notation

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}_{(n_1+n_2)\times 1} = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix}_{(n_1+n_2)\times 2k} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}_{(2k\times 1)} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix}_{(n_1+n_2)\times 1} \qquad (5.3)$$

By applying least squares estimation method, one may estimate the unrestricted linear regression model (5.3) and obtain the unrestricted least squares residual sum of squares of in matrix notation as $(e'e)_{UR}$.

Restricted Linear Regression model under H$_0$: $\beta_1 = \beta_2$ in matrix notation

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}_{(n_1+n_2)\times 1} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}_{(n_1+n_2)\times k} \beta_{(k\times 1)} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix}_{(n_1+n_2)\times 1} \qquad (5.4)$$

Again, applying least squares estimation method to estimate restricted linear regression model (5.4), one may obtain the restricted least squares residual sum of squares in matrix notation as $(e'e)_R$.

Now, the chow test statistic for testing the structural change under H$_0$ is given by

$$F = \frac{\left[(e'e)_R - (e'e)_{UR}\right]\big/ k}{(e'e)_{UR}\big/ ((n_1+n_2)-2k)} \sim F_{[k,(n_1+n_2)-2k]} \qquad (5.5)$$

One may compare the calculated value of F-test statistic with its critical value for (k, ($n_1 + n_2 - 2k$)) degrees of freedom at an appropriate level of significance and draw the inference accordingly.

6. Conclusions

The main aspect of Applied Statistician is the analysis of real – life data by applying various suitable statistical techniques in drawing valid conclusions about the characteristics of the data. Data analysis of Research Projects in India by using advanced statistical techniques gained momentum in the recent years. Applications of various statistical tools to analyse different research problems depend upon the availability and reliability of statistical data.

In the present study, an attempt has been made by giving some important advanced statistical tools to be used in practice for the data analysis of research projects. Further, simple statistical techniques for time series data analysis have been discussed in the study/

7. References

1. Aaker DA et al., Marketing Research", John Wiley & Sons, New York, 2001.
2. Draper NR, Smith H. Applied regression analysis, Second Edition, John Wiley & Sons, New York, 1981.
3. Johnson RA, Bhattacharya GK. Statistics: Principles and Methods, John Wiley & Sons, New York, 1996.
4. Johnson RA, Wichern DW. Applied Multivariate Statistical Analysis, Sixth Edition, Pearson Prentice Hall, New Jersey, 2007.
5. Johnston J, Dinardo J. Econometric Methods", Fourth Edition, McGraw Hill, New York, 1996.
6. Judge GG *et al.* The Theory and Practice of Econometrics", Second Edition, John Wiley & Sons, Inc, 1985.
7. Maddala GS. Introduction to Econometrics, Second Edition, MacMillan Publishing Company, New York, 1992.
8. Makridakis S et al. Forecasting: Methods and Applications", Second Edition, John Wiley & Sons, New York, 1983.
9. Montgomery DC. Design and Analysis of Experiments", Eighth Edition, John Wiley & Sons, Inc. 1983, ISBN: 978-1-118-14692-7.
10. Panneerselvam R. Research Methodology, Second Edition, PHI Learning Pvt. Limited, New Delhi, 2014, ISBN: 978-81-203-4946-9.
11. Rao CR. Advanced Statistical Methods in Biometric Research, John Wiley and Sons, Inc. 1952.
12. Rao CR. Linear Statistical Inference and Its Applications, Second Edition, Wiley, New York, 1973.
13. Rawlings JO, SG Pantula, Dickey DA. Applied Regression Analysis: A Research Tool, Springer - Verlag, New York, 1998.