Development and Application of Chi-Square in R: A Case Study on Indian Computer Industries

Sandeep Kaur

Abstract—Statistical, presentation tools and data management make R a perfect integrated package for research in the field of computer industries. In this paper we describes the various steps which helps to import the data into R, to prepare that data for analysis using nineteen different states calculations with the help of chi square test. This paper outline the data analysis techniques of chi square test in R with side-by-side compare the data with SPSS statistical software package. In this paper we show the various result of the R and the SPSS and in the both cases the hypothesis is null.

Index Terms—SPSS, SAS, Splus, R, programming language, statistical software.

I. INTRODUCTION

The R language is a project designed to create a free, open source language which can be used as a replacement for the Splus language, originally developed as the S language at AT&T Bell Labs, and currently marketed by Insightful Corporation of Seattle, Washington. While R is not 100% compatible with Splus (it is often described as a language “which bears a passing resemblance to S”), many Splus programs will run under R with no alterations. Accordingly, much of the existing documentation for Splus can still be useful under R, and many authors of code for either language are careful to make sure that their code will be suitable for both languages. Broadly speaking, R is a core programming language in which one writes statistical software packages and SPSS is a statistical software package that can be programmed. R is a dialect of the S language, which in turn is a language with semantics for the most part derived from LISP with numeric array processing operations very much like those in APL. The SPSS programming language is more procedural and resembles a macro assembler. In terms of current feature sets, given the Comprehensive R Archive Network (CRAN) library of packages, there are few statistical processes one can't get done in either R or SPSS. SPSS has more marketing and business intelligence type processes built in than you'll find in CRAN, but CRAN has more bioinformatics, finance and spatial processes. The final differentiator is cost. The basic version of R is open source, as are most of the packages in CRAN. The only cost is the bandwidth to download R and whatever packages you need. There is a commercial version of R from Revolution Analytics that's more highly tuned and has support and Training included. SPSS is not open source, and is sold as a base package and add-on packages for various optional processes. Describing how vectors work in R, it is helpful to distinguish two ideas of vectors in order to set the correct expectations [12].

II. VARIOUS PROGRAMMING AND STATISTICAL LANGUAGES USED

A. SPSS [8]

SPSS is among the most widely used programs for statistical analysis in social science. It is used by market researchers, health researchers, survey companies, government, education researchers, marketing organizations and others. The many features of SPSS are accessible via pull-down menus or can be programmed with a proprietary 4GL command syntax language. Command syntax programming has the benefits of reproducibility, simplifying repetitive tasks, and handling complex data manipulations and analyses.

Statistics included in the base software:
1. Descriptive statistics: cross tabulation, frequencies, Descriptive, Explore, Descriptive Ratio Statistics

B. R (programming language) [11]

R is a language and environment for statistical computing and graphics. It is GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers R and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R. Provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering .................) and graphical. Techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity. One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. Great care has been taken over the defaults for the minor design choices in graphics, but the user retains full control. R is available as Free Software under the terms of the free

All Rights Reserved © 2012 IJARCSEE

32
III. APPLICATION OF CHI-SQUARE TEST [9]

A chi-square test is a statistical test commonly used for testing independence and goodness of fit. Testing independence determines whether two or more observations across two populations are dependent on each other (that is, whether one variable helps to estimate the other). Testing for goodness of fit determines if an observed frequency distribution matches a theoretical frequency distribution. In both cases the equation to calculate the chi-square statistic is

\[ \chi^2 = \sum \frac{(O - E)^2}{E} \]

where \( O \) equals the observed frequency and \( E \) the expected frequency. The results of a chi-square test, along with the degrees of freedom, are used with a previously calculated table of chi-square distributions to find a \( p \)-value. The \( p \)-value can then be used to determine the significance of the test.

IV. PURPOSED ALGORITHM

1. Prepare the Excel sheet to calculate the average growth for all states.
2. Read the data from the Excel sheet in R as (.csv) file.
3. Fill the total values into the array totalc=array(1:21)
4. Prepared a choices dialogue box for 19 states so that any one state chosen from the all states and also add the average growth choice so that should be compare with the state.
5. To find the chi square test \( \chi^2 = \frac{(O - E)^2}{E} \) denote the values to the observed and expected from equation 1.
6. Print the all values with print("\%+15.10f",c).
7. Print the results as print(sprintf("Chi-Square Answer = %+15.10f",c)).
8. Find out the degree of freedom \( n-1 \).
9. The significant result has been investigated for R hypothesis with the help of critical value table degree of freedom 15 on 0.05 is 24.996.
10. The chi square result is -43.6696390449. This is rejected because it is greater than the 0.05 significance value and in the SPSS the value is asymp.sig (2-signed) is 2.400E2 a P value called Pearson value. It has been evaluated from the both results that the hypothesis (H0) is rejected in both cases. If the results came less than the 0.05 the hypothesis (H0) is null hypothesis means the values of DELHI is dependent on the average growth.

V. IMPLEMENTATION AND RESULTS FORMULATION

A. Implementation in R

The Figure 1 shows the results i.e. the observed value denoted by O, Expected value denoted by E, Observed – Expected and the (Observed-Expected) 2/ Expected. It also shows the degree of freedom and the critical value which is taking from the chi square table. The result depicts that the hypothesis (H0) is accepted because it is greater than the 0.05 significance value.

B. Implementation in SPSS

In the SPSS we take the Pearson chi-square value 240.0 as a result, in the results of the SPSS we didn’t see the critical value chart because it take the p value. called Pearson value as a result. The cases are 16 which are shown as an N of valid cases. Table 1 and 2 shows the results using SPSS statistical package.
VI. CONCLUSION

The comparative investigation of different statistical software named SPSS and statistical and programming language R. The significant result has been investigated for R hypothesis with the help of critical value table degree of freedom 15 on 0.05 is 24.996. The chi square result is 97.5628475009. This is rejected because it is greater than the 0.05 significance value and in the SPSS the value is asymp.sig (2-sided) is 2.400E-02 a P value called Pearson value. It has been evaluated from the both results that the hypothesis (H0) is rejected in both cases. If the results came less than the 0.05 the hypothesis (H0) is null hypothesis means the values of Andhra Pradesh is dependent on the average growth.

ACKNOWLEDGMENT

I am thankful to Ministry of Statistics and Programme Implementation, Government of India who provided me the data of different states. I wish to express my gratitude and heartfelt thanks to our Sukhwinder Kaur, Assistant Professor, Department of Management, Lovely Professional university for her timely advice and support in carrying out my work.

REFERENCES

Sandeep Kaur was born in Amritsar. She has done her MCA from Guru Jambheswar University of Science and Technology Hisar and is currently pursuing Masters of Technology in CSE from Lovely Professional University (LPU) Phagwara, India. Her research areas of interests include software engineering, Statistical Programming, R Programming, Security, Databases, and Mobile Ad hoc networks.