ORIGINAL RESEARCH PAPER

Biological Science

STATISTICAL ANALYSIS OF CLINICAL STUDIES THROUGH SPSS AND R

KEY WORDS: SPSS, R-programming, Regression, Chi-square test, ANOVA

Aijaz Ahmad — Research Scholar, University Department of Statistics and Computer Applications, T.M. Bhagalpur University, Bhagalpur

Md Shahid Iqbal * — Research Scholar, University Department of Statistics and Computer Applications, T.M. Bhagalpur University, Bhagalpur *Corresponding Author

N. Ahmad — Professor and Head, University Department of Statistics and Computer Applications, T.M. Bhagalpur University, Bhagalpur

ABSTRACT

Field of Medicine is an empirical science because of heavy use of mathematical and computational models. Biostatisticians join the link between Medical science and Mathematics. In modern medicine, to make sense of data collected to decide if the treatment would be suitable or to find variable which are responsible for the disease, we require predictive models and statistical methods. There are varieties of predictive models and statistical methods available in literature. However, Chi-square test, analysis of variance (ANOVA), and regression models are powerful analytic tools, yield valid statistical inferences. This paper explores the uses of Chi-square test, analysis of variance (ANOVA), and regression analysis in clinical studies. We present the computational analysis for real clinical data using statistical package for the social sciences (SPSS) and R programming in simple steps. Further, we discuss the computational and graphical outputs of SPSS and R in details. This work may help medical researchers to identify the appropriate test for the analysis.

1. INTRODUCTION

Bio-statistical points of view are now taking place in medical books, research journals and pharmaceutical literature as a medical tool. Variation is inevitable aspect of life and uncertainties are profound. Thus, statistical methods are indeed to measure the magnitude of uncertainties and to minimize their impact on decisions making (Armstrong et al., 2011). Trends are established among variations and the trend yields results within clinical tolerance. Statistical medical practices are established to delineate and minimize the role of uncertainties and thus increase the efficiency of medical decisions (Chow et al., 2013).

Predictive models using historical data help to find most likelihood of disease for diagnosis and prognosis. They are helpful for clinical decisions making, allowing for care to be customized to each individual. Many statistical methods for data analysis are available in the literature, such as "Linear regression", "Logistic regression", "Multivariate Regression", "Cluster analysis", "Analysis of variance (ANOVA)", "Chi-squared test", "Factor analysis", "Time series", "Experimental design", "Bayesian theory Naïve Bayes classifier" (Spicer et al., 1987; Dunn and Everitt, 1995; Everitt, 2002; Collett, 2003; Armstrong & Hilton, 2004; Armstrong et al., 2010; 2011).

Statistical programming involves doing computations to aid in statistical analysis. For example, data must be summarized and displayed. Models must be fit to data, and the results displayed. These tasks can be done in number of different computer applications: Microsoft Excel, SPSS, SAS, S-PLUS, R, MINITAB, etc.

In this paper we discuss the procedure of chi-square test, analysis of variance and regression analysis through two Statistical Packages (SPSS and R). The paper follows with preliminary of SPSS and R, and discusses three statistical methods with illustration.

2. Preliminary of SPSS and R

A. SPSS: It stands for Statistical Package for the Social Sciences and is a comprehensive system for analyzing data (SPSS Inc., 2001). **The package** is used for **managing and analysis of real** data.

- SPSS package consists of a set of software tools for data entry, data management, statistical analysis and presentation.
- SPSS can take data from almost any type of file and use them to generate tabulated reports, charts, and plots of distributions and trends, descriptive statistics, and complex statistical analysis.

To start SPSS from the Windows Taskbar choose Start ➤ All Programs ➤ SPSS. The SPSS Data Editor offers facility like spreadsheet, which is divided into rows and columns. This data editor provides two views of the data (a) Data view: Displays the actual data values. (b) Variable view: Displays variable definition information, including defined variable and value labels, data type etc.

Entering and Editing Data: The easiest way to enter data in SPSS is to type it directly into the Data Editor window.

You enter all the data of a variable in a column of the Data Editor. To enter the name of the variable click the Variable View tab of the Data Editor. A window will appear in which the variables are numbered 1, 2, and so on. Type the name of variable that you want in the column headed Name.

Import a Data File: If your data are already in a Excel file, you could open it from the File Menu as follows:

- Choose File ➤ Open ➤ Data. The dialog box will appear.
- Choose the appropriate location of the file.
- In Files of type select Excel (*.xls, *.xlsx, *.xlsm).
- Select your required file and click on Open. The dialog box Opening Excel Data Source will appear. Check the boxes for which you are looking for.
- Click OK.

B. R-Programming: It is a free, open source, and cross-platform programming language that is well suited for statistical analyses and can be installed using URL https://www.R-project.org/ or search CRAN (Comprehensive R Archive Network) into Google (R Core Team, 2018). R is being used more and more in educational, academic, and commercial settings. A few advantages of working with R as a student, teacher, or researcher include:

- R functions return limited output. This helps prevent students from sorting through a lot of output they may not understand, and in essence requires the user to know what

output they're asking R to produce.
- Since all functions are open source, the user has access to see how pre-defined functions are written.
- There are powerful packages written for specific type of analyses.

3. Chi-square Test

This test is used to assess the significance of the difference between two categorical variables by comparing the observed and expected frequencies using the chi-square test. The test statistic is given as: Chi-square = $\sum \frac{(O-E)^2}{E}$. It is based on the following assumptions:

- Independence: Each participant should participate only once and should not influence others.
- Expected Frequency: In 2×2 design, all expected frequencies should be at least five. In larger design, no more than 20% of expected cell frequency should be lower than five.

Example: A study of Hodgkin's disease classified 538 numbers of patients by histological type (LP = lymphocyte predominance, NS = nodular sclerosis, MC = mixed cellularity, LD = lymphocyte depletion) and response to treatment like- positive, partial, and none are given in the following contingency table (Hand et al., 1994):

Histological Type	**Disease**		
	Positive	**Partial**	**None**
LP	74	18	12
NS	68	16	12
MC	154	54	58
LD	18	10	44

We analyze the data to identify the evidence that a patient's response to treatment for Hodgkin's disease varies by histological type.

SPSS: Steps for chi-square test
- Enter all the frequency data (Response) into first column and name the column *Count*.
- In second column (grouping variable indicating whether the response is from positive or partial or none by disease), enter code 1 if the value is positive, enter 2 if the value is partial, and enter 3 if the value is none. Name the column *Disease*.
- In third column, enter codes to identify the Histological Type. Enter the code 1, 2, 3, and 4 for LP, NS, MC, and LD, respectively. Name the column *Type*.
- Click on *Data* > *Weight Cases…* A dialog box will open.
- Select *Weight cases by,* double click on *Count* to move frequency data in the box and then click *OK*.
- Choose *Analyze* > *Descriptive statistics* > *Crosstab….*
- In *For rows:* Enter the variable *Type* containing the categories that define the rows of the table.
- In *For columns:* Enter the variable *Disease* containing the categories that define the columns of the table.
- Under Exact: You could select the exact test if the expected cell frequency is less than 5.
- Under Statistics: Select Chi-square value.
- Under *Cell:* You can select expected frequencies, percentage, etc.
- Click *OK.*

Table 1. SPSS Output of Observed and Expected frequencies.

Type * Disease Cross tabulation						
			Disease			Total
			1	2	3	
Type	1	Count	74	18	12	104
		Expected Count	60.7	18.9	24.4	104.0
	2	Count	68	16	12	96
		Expected Count	56.0	17.5	22.5	96.0
	3	Count	154	54	58	266
		Expected Count	155.2	48.5	62.3	266.0
	4	Count	18	10	44	72
		Expected Count	42.0	13.1	16.9	72.0
Total		Count	314	98	126	538
		Expected Count	314.0	98.0	126.0	538.0

Table 2. SPSS Output of Chi-square Test.

Chi-Square Tests			
	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	75.890a	6	.000
Likelihood Ratio	68.295	6	.000
Linear by Linear Association	46.267	1	.000
N of Valid Cases	538		
a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 13.12.			

R-Programming: If the data are summarized in a table (x) or matrix (x), then the R function for the test is: *chisq.test(x).* If the data are not summarized but stored in two variables x and y, then the R function is: *chisq.test(x,y)* or *chisq.test(table(x,y))*

```
> count<-matrix (c(74,68,154,18,18,16,54,10,12,12,58,44),
nrow=4)
> count
      [,1] [,2] [,3]
[1,]   74   18   12
[2,]   68   16   12
[3,]  154   54   58
[4,]   18   10   44

> chisq.test(count)

Pearson's Chi-squared test
data: count

X-squared = 75.89, df = 6, p-value = 2.517e-14
```

Table 1 and Table 2 shows the output of SPSS and output R with steps are also presented for chi-square test. We have the value of test statistic, x^2=75.89 and p-value = 000 (approx.) which is less than α=0.05. Hence we reject the null hypothesis and conclude that the Hodgkin's disease is highly significant for histological type.

4. Analysis of Variance

In this section, we discuss the analysis of variance (ANOVA) to test whether there is any significance difference between three or more sample means. ANOVA has following assumptions:

- Sample data are continuous.
- Groups are independent and should be approximately normally distributed.
- The variability in each group is same.
- Statistical model for the data is $y_{ij}=\mu+T_i+e_{ij}$.

Example: Steady-state hemoglobin levels were measured on a total of patients 41 with 3 types of sickle cell disease. The types of sickle cell disease are HB SS, HB ST (HB S/-thalassemia), and HB SC. The results are shown below (Hand et al., 1994):

SS	7.2 9.1	7.7 9.1	8.0 9.1	8.1 9.8	8.3 10.1	8.4 10.3	8.4	8.5	8.6	8.7
ST	8.1	9.2	10.0	10.4	10.6	10.9	11.1	11.9	12.0	12.1
SC	10.7 12.6	11.3 13.3	11.5 13.3	11.6 13.8	11.7 13.9	11.8	12.0	12.1	12.3	12.6

We analyze the data with the hypothesis that the three types of sickle cell disease have the same hemoglobin levels at 5% significance level.

SPSS: Steps for ANOVA

- Enter data (hemoglobin) into the Data Editor in a single column and enter grouping variable 1, 2… in another column to indicate to which group each growth data belongs. Go to Variable View and name the columns as *hemoglobin* and *disease* and then return to Data Editor.
- Use Explore and test for normality.
- Select Analyze > Compare Means > One-Way ANOVA…. A One-Way ANOVA dialog box will open.
- Move the variable *hemoglobin* to Dependent List box and move the variables *disease* to Factor box.
- In Options one can select Homogeneity of variance test, which calculates the Levene statistic to test for the equality of group variances.
- Click [OK].

Table 3. SPSS output of analysis of variance for Hemoglobin

ANOVA					
Hemoglobin					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	99.889	2	49.945	49.999	.000
Within Groups	37.959	38	.999		
Total	137.848	40			

Table 4. SPSS output of Tukey's test for Multiple Comparisons

Multiple Comparisons						
Hemoglobin Tukey HSD						
(I) Disease	(J) Disease	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Limit	Upper Limit
1	2	-1.91750*	.40289	.000	-2.9001	-.9349
	3	-3.58750*	.35920	.000	-4.4635	-2.7115
2	1	1.91750*	.40289	.000	.9349	2.9001
	3	-1.67000*	.40803	.001	-2.6651	-.6749
3	1	3.58750*	.35920	.000	2.7115	4.4635
	2	1.67000*	.40803	.001	.6749	2.6651
*. The mean difference is significant at the 0.05 level.						

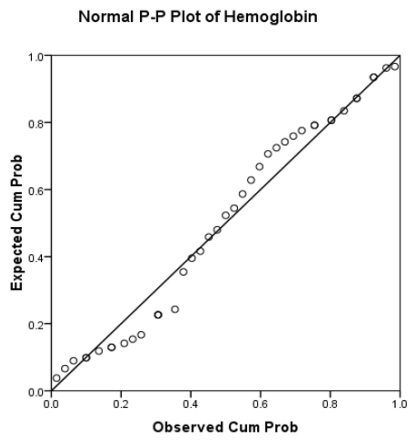

Figure 1. Normal probability plot

R-Programming: The R function for ANOVA is *oneway.test ()or aov()*.

```
> hemoglobin< -c(7.2, 7.7, 8.0, 8.1, 8.3, 8.4, 8.4, 8.5, 8.6, 8.7, 9.1,
9.1,9.1,9.8,10.1,10.3,8.1,9.2,10.0,10.4,10.6,10.9,11.1,11.9,12.0,
12.1,10.7,11.3,11.5,11.6,11.7,11.8,12.0,12.1,12.3,12.6,12.6,13.
3,13.3,13.8,13.9)

> disease<-rep(c("s1","t1","c1"),c(16,10,15))
> HB<-data.frame(hemoglobin,disease)
> HB
   hemoglobin disease
1         7.2      s1
2         7.7      s1
3         8.0      s1
4         8.1      s1
5         8.3      s1
6         8.4      s1
7         8.4      s1
8         8.5      s1
9         8.6      s1
10        8.7      s1
11        9.1      s1
12        9.1      s1
13        9.1      s1
14        9.8      s1
15       10.1      s1
16       10.3      s1
17        8.1      t2
18        9.2      t2
19       10.0      t2
20       10.4      t2
21       10.6      t2
22       10.9      t2
23       11.1      t2
24       11.9      t2
25       12.0      t2
26       12.1      t2
27       10.7      c3
28       11.3      c3
29       11.5      c3
30       11.6      c3
31       11.7      c3
32       11.8      c3
33       12.0      c3
34       12.1      c3
35       12.3      c3
36       12.6      c3
37       12.6      c3
38       13.3      c3
39       13.3      c3
40       13.8      c3
41       13.9      c3

> attach(HB)

The following objects are masked _by_ .GlobalEnv:
    disease, hemoglobin

> names(HB)
[1] "hemoglobin" "disease"

> Analysis<-aov(hemoglobin~disease)

> summary(Analysis)
             df  Sum Sq  Mean Sq  F value  P(>F)
disease       2  99.89   49.94    50       2.28 e-11 ***
Residuals    38  37.96   1.00

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> boxplot (hemoglobin~disease, names=c("SS","ST","SC"))
```

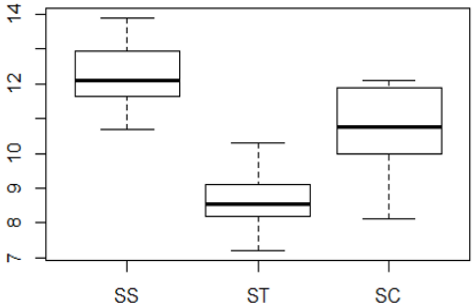

Figure 2. Box plots of hemoglobin levels

Table 3 and Table 4 shows the output of SPSS and output of *R* with steps are also presented for ANOVA test. We have the value of test statistic F=50 and p-value = 000 (approx.) is less than α=0.05, we reject H_0 (null hypothesis) and conclude that there is highly significant different between the average hemoglobin levels with different sickle cell disease. Table 4 presents Tukey multiple comparisons to check the pair-wise difference. It shows that all the pairs of treatments are significantly different. Normal probability plot presented in Figure 1 reveal that the error distribution of hemoglobin level is approximately normal since the points are cluster around straight line. Figure 2 presents box-plot for hemoglobin levels at each type of sickle cell disease. It indicates that the average hemoglobin level is different due to difference sickle cell disease.

5. Regression Analysis

In this section, we assess the linear relationship between dependent (or response or outcome) variable and independent (or predictor) variable. The regression analysis has the assumptions:

- Errors or residuals (e) have mean zero and variance constant (which can be assessed graphically using a scatter plot of residuals and predicted values).
- Residuals are normally distributed (which can be assessed graphically using a Normal Q-Q Plot of residuals).
- Errors are uncorrelated random variables with mean zero (Autocorrelation). One can use Durbin-Watson statistics, which should be within -2 to +2).
- Outcome variable is linearly related with predictors (Scatter plot or Lack-of-fit test can be used for testing linearity).
- Simple linear regression model is $y_i=\beta_0+\beta_1 x_i+e_i$

Example: The body mass index (BMI) and the systolic blood pressure of 6 people were measured to study a cardiovascular disease. The result are shown below

BMI (x)	Pressure (y):
26	170
23	150
27	160
28	175
24	155
25	150

We analyze the data with the hypothesis that whether a high BMI relates to a high blood pressure.

SPSS: Steps for Regression Analysis

- Select Analyze > Regression > Linear.... A linear regression dialog box will open.
- Move the variable *Pressure* to Dependent box and move *BMI* to Independent(s) box.
- Make sure that Enter is selected in the Method dialog box.
- Click on Statistics... and check the boxes Estimates, Confidence Interval (CI) and Model fit and then click Continue.
- If you wanted to compute predicted and residual values, click on Save and then select Unstandardized in the both Predicted Values and Residuals boxes. The required values will be saved in the additional columns of the Data Editor.
- If you want, click on Plots and select Normal probability plot for standardized Residual plot.
- If you want to check for linearity and constant variance, then plot the standardized residuals against the standardized predicted values by choosing moving *ZRESID to the box Y and *ZPRED to the box X.
- Click [OK].

Table 5. Regression analysis of systolic blood pressure

ANOVA[b]						
Model	Sum of Sq	df		F	Sig.	
1	Regression	365.714	1	365.714	7.938	.048a
	Residual	184.286	4	46.071		
	Total	550.000	5			

a. Predictors: (Constant), BMI
b. Dependent Variable: Pressure

Table 6. Point and Interval estimates of parameters

Coefficients[a]								
Model		Un-standardized Coefficients		Standardized Coefficients	t	Sig.	95% CI for B	
		B	Std. Error	Beta			Lower Limit	Upper limit
1	(Constant)	43.429	41.468		1.047	.354	-71.704	158.561
	BMI	4.571	1.623	.815	2.817	.048	.067	9.076

Normal P-P Plot of Regression Standardized Residual

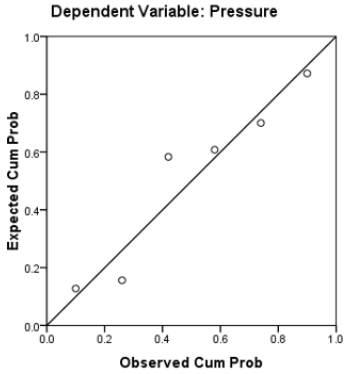

Figure 3. Normal probability plot of systolic blood pressure

R-Programming: The R function lm(), summary(), confint(), summary.aov() are used to obtain the regression coefficients, their summary, confidence interval of coefficients, and ANOVA table, respectively.

```
> BMI<-c(26,23,27,28,24,25)
> SBP<-c(170,150,160,175,155,150)

> model<-lm(formula=SBP~BMI)
> summary(model)
```

Call:
lm(formula = SBP ~ BMI)

Residuals:

1	2	3	4	5	6
7.714	1.429	-6.857	3.571	1.857	-7.714

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	43.429	41.468	1.047	0.354
BMI	4.571	1.623	2.817	0.048 *

Sign if. codes: 0'***'0.001'**'0.01'*'0.05'.'0.1' '1

Residual standard error: 6.788 on 4 degrees of freedom
Multiple R-squared: 0.6649, Adjusted R-squared: 0.5812

F-statistic: 7.938 on 1 and 4 DF, p-value: 0.04795

```
> plot(model)
```

```
> confint(model)
                2.5 %       97.5 %
(Intercept) -71.70392268 158.561066
BMI           0.06652078   9.076336

> summary.aov(model)
             df  Sum Sq  Mean Sq  F value  P(>F)
BMI          1   365.7   365.7    7.938    0.048 *
Residuals    4   184.3    46.1
—
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

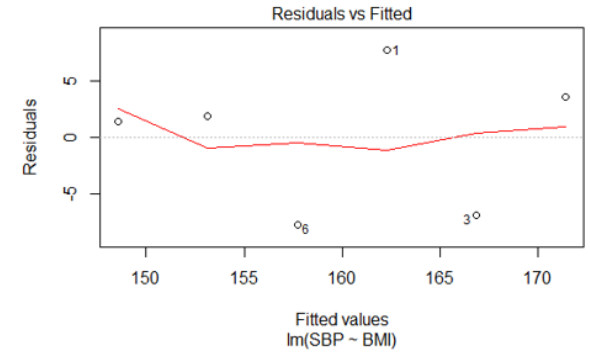

Figure 4. Residual vs fitted plot

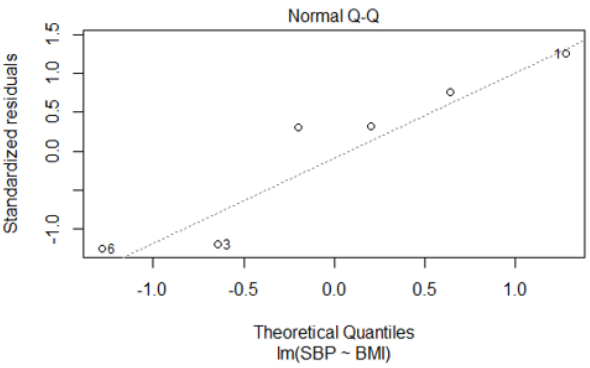

Figure 5. Normal Q-Q plot of Residuals

Table 5 and Table 6 shows the output of SPSS and output of R with steps are also presented for regression analysis. We have the value of test statistic, =7.938 , and *p-value* = 0.048 and is less than α=0.05, we reject the null hypothesis ie. H_0:β_1=0 and conclude that H_1:$\beta_1 \neq 0$, that is, there may be linear relationship between *x* and *y*. Table 6 and output of R (labeled coefficients) presents parameters estimate along with 95% confidence interval, and t value and p-value to test hypothesis for intercept (β_0) and slope (β_1) equals zero. The value of R-squared = 0.665 shows that 66.5% of the systolic BP is explained by the BMI. The fitted regression line is y=43.429+4.571x. Figure 3 and 5 presents the normal P-P & Q-Q plots. Since the residuals cluster approximately around straight line, we conclude that the error distribution is approximately normal. The residuals are also plotted against the fitted value in Figure 4. It shows no serious evidence of model inadequacies and shows no variance increasing.

CONCLUSION

In this paper we have presented the statistical analysis using chi-square test, ANOVA, and linear regression for clinical data through SPSS and R programming. We have also presented the basic and essential understanding of proposed statistical methods and steps wise procedure for analysis through SPSS and R. The experimental results and graphical presentations are discussed. The analysis discussed in this paper may be helpful for researchers to use a suitable statistical method that gives meaningful results in clinical studies.

REFERENCES

[1.] Altman, D. G. (1991). Practical Statistics for Medical Research. Chapman & Hall, London.
[2.] Armstrong, R.A., Davies, L.N., Dunne, M.C.M., & Gilmartin, B. (2011). Statistical guidelines for clinical studies of human vision. Ophthalmic Physiol Opt, 31, 123–136.
[3.] Armstrong, R.A., Eperjesi, F., & Gilmartin, B. (2010). The use of correlation and regression methods in optometry. Clin Exp Optom, 88, 81–88.
[4.] Armstrong, R.A. & Hilton, A. (2004). The use of analysis of variance (ANOVA) in applied microbiology. The Microbiologist, 5, 18–21.
[5.] Collett, D. (2003). Modelling Survival Data in Medical Research (2nd ed). Boca Raton, Chapman & Hall/CRC, FL.
[6.] Dunn, G. and Everitt, B. S. (1995). Clinical Biostatistics: An Introduction to Evidence-Based Medicine. Arnold, London.
[7.] Everitt, B. S. (2002) Modern Medical Statistics: A Practical Guide. Arnold, London.
[8.] Hand, D.J., Daly, E., Lunn, A.D., McConway, K.J., and Ostrowki, E. (1994). A Handbook of Small Data Sets. Chapman & Hall, London.
[9.] Spicer, C. C., Laurence, G. J., and Southall, D. P. (1987). Statistical analysis of heart rates and subsequent victims of sudden infant death syndrome. Statistics in Medicine, 6, 159–166.
[10.] SPSS Inc. (2001). SPSS Base 11.0 for Windows User's Guide: Englewood Cliffs, Prentice Hall, NJ.
[11.] R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.