UNIT 10 ADVANCED STATISTICS*

Contents

Learning Objectives

After reading this unit, you will be able to:

- Compute and interpret various measures of association;
- Perform Chi Square test;
- Work with linear regression; and
- Perform Analysis of Variance and interpret the findings.

10.0 INTRODUCTION

Very often in medical and public health research, we come across *categorical variables* which are also known as *qualitative factors*. Data on such factors is not a measurement but it like an option to choose from a discrete list of possible values. For instance, the sanitation level in a village may be expressed as *poor, moderate* and *good* which may be numerically coded as 1, 2, 3 respectively. This type of data is called *ordinal data* because the order or options has a meaning. A higher value indicates a better situation. Sometimes categories are just nominal in the sense the numeric value does not indicate any order. For instance, the type of leprosy under 4 categories may be coded as 1,2,3,4 and the code 4 may indicate a better status when compared with code 1.

For categorical data we cannot use measures like mean and standard deviation. Instead it will be expressed *count of cases* and percent (or proportion).

When two categorical variables are to be compared, we summarize the data in the form of a two-way table of counts, known as *contingency table* or *cross tabulation.*

For instance, the gender wise distribution of the prevalence of *cataract* in a study area may be presented as a 2 x 2 table shown in Table 10.1.

* Prof. K.V. S Sarma (retd.), Department of Statistics, Sri Venkateswara University, Tirupati

Table 10.1: Cross tabulation of gender versus cataract

Gender	Cataract		Total
	Yes	No	
Male	73	54	127
Female	35	38	73
Total	108	92	200

There are two factors in this context; one is gender (male or female) and the other is presence of cataract (yes or no). We wish to know whether prevalence of cataract has any association with gender or cataract is independent of gender.

This type of contingency tables can be easily prepared on large data by using the tools of Excel and SPSS. In Excel we use the option 'Pivot table' from the insert menu.

10.1 CHI SQUARE TEST OF ASSOCIATION

Association is the term used to indicate the relationship between two qualitative factors which are also known as *attributes.* The 4 values given in Table 10.1 may be identified in general as shown below so that we can work out a formula to perform further analysis.

Gender	Cataract		Total
	Yes	No	
Male	*a*	*b*	(a+b)
Female	*c*	*d*	(c+d)
Total	(a+c)	(b+d)	N

If we wish to know whether presence of cataract has any association (relationship) with gender we perform a statistical test of significance called *Chi-Square test of independence*.

The null hypothesis is H_0: The two attributes are independent and let us take $\alpha = 0.05$. We calculate a test value denoted by the Greek letter χ^2 as follows.

$$\chi^2 = \frac{N(ad - dc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

If the calculated test value exceeds the critical value (to be read from statistical tables), we reject H_0 and conclude that prevalence of cataract has some association with gender. The critical value is however linked to the degrees of freedom given by (# rows-1) x (# columns-1). In a 2x2 table we get (2-1) x (2-1) = 1. From tables of Chi Square values, we get the critical value as 3.84, at 5% level of significance with 1 degree of freedom. Further details on the Chi Square test can be read from Sundara Rao and Richard (2012).

Here is an illustration.

Illustration 10.1: Consider the contingency table given in Table-10.1 which is reproduced as below with *a, b, c, d* marked in brackets.

Gender	Cataract		Total
	Yes	No	
Male	73 (a)	54 (b)	127
Female	35 (c)	38 (d)	73
Total	108	92	200

Using Chi Square test, check whether there is any association between gender and presence of cataract.

Solution: From the above table we observe the following.

1) (a+b) =108, (c+d)= 92, (a+c)= 127 and (b+d)= 73 and N = 200

2) Chi-Square = $\frac{200*(2774-1890)^2}{(108)(92)\ (127)(73)} = 1.697$

3) The critical value from statistical tables at 5% level of significance with 1 degree of freedom is 3.84.

Since the calculated Chi-Square is much smaller than the critical value, we conclude that gender has no association (relation) with the presence of cataract. It means the presence of cataract is likely to be independent of gender.

Chi Square test for bigger tables: Sometimes we get contingency tables of size larger than a 2 × 2 table. If one factor has 3 levels and another has 4 levels, we get a 3 × 4 table. There will be 3 rows and 4 columns.

The calculation of Chi Square value for such tables has a general formula in which we find the *expected frequency* (E) for each cell and compare them with the *observed frequency* (O) of the cell. A cell is a part of the table at the intersection of a row and a column. A cell contains the observed data.

The expected frequency is found from the contingency table as

E = (Row total * Column total)/Grand total.

The Chi Square value is computed with the formula $\chi^2 = \sum \frac{(0-E)^2}{E}$ If there are k cells in the table you will get

$$\chi^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \ldots + \frac{(O_k - E_k)^2}{E_k}$$

The null hypothesis is again H_0: The two factors are independent (there is no association).

The degrees of freedom for a 3 x 4 table will be (3-1) x (4-1) = 2 x 3 = 6 and the critical value will be read for these degrees of freedom.

Here is an illustration with 3 rows and 3 columns.

Illustration 10.2: A batch of 124 health workers was given field training on using a new devise for a screening test. The distribution of respondents according to the performance score (max = 100) and experience (in years) is given below.

Experience (years)	Performance score		
	< 40	41 - 60	Above 60
Up to 5	19	11	8
6 – 10	10	15	18
Above 10	6	17	20

We wish to test whether the performance score has any association with experience of the health worker.

Solution: We can use an online calculator to perform this test and obtain the following results. For instance, if we use www. socscistatistics.com/tests/ chisquare2/Default2.aspx, we get the intermediate calculations. This is a convenient method instead of the hand-calculation method.

	< 40	41 - 60	Above 60	Row Total
Up to 5	19 (10.73) [6.38]	11 (13.18) [0.36]	8 (14.10) [2.64]	38
6 – 10	10(12.14) [0.38]	15(14.91) [0.00]	18(15.95) [0.26]	43
Above 10	6(12.14) [3.10]	17(14.91) [0.29]	20(15.95) [1.03]	43
Column Total	35	43	46	124

Each cell of the table contains three entries as follows:

i) Expected frequency (E) of a cell shown in ordinary bracket.

ii) Chi Square value for the cell and the figure shown in square bracket. Adding the Chi Square values of cells give the test value as 14.442. The p-value of the test is p = 0.006 which is less than 0.05. Hence there is a significant association between score and experience.

Remark: The Chi Square test measures the significance of the association between two categorical variables.

10.2 ASSOCIATION RELATED MEASURES

In some epidemiological studies we come across measures like *relative risk* and *odds ratio* both of which are based on 2 x 2 table of counts as done in Chi Square tests. The proportion of people of a population, having disease among all those who are exposed to a condition, is called the risk of disease.

Relative Risk

Suppose in a study group of 300 males, we found 156 smokers and out of them 85 had a heart disease. The risk of heart disease due to smoking in this group

will be 85/156 = 0.54 or 54%. Among 146 non-smokers suppose 28 persons had a heart disease. Then the risk of disease for non-smokers is 28/146 = 0.19 or 19%. This information can be arranged as a 2 x 2 table as shown below.

Smoking	Disease		Total
	Present	Absent	
Yes	85	71	156
No	28	118	146
Total	113	189	300

The *relative risk* (RR) of a disease is calculated as follows.

RR = *Risk of disease in the exposed group* / *Risk of disease in the unexposed group*

In this case we get RR = 0.54/0.19 = 2.84. It means that smokers have 2.84 times more risk of a heart disease when compared to non-smokers.

RR is also called *risk rate* and popularly used to compare risks or incidence of various health conditions across time periods or geographical locations. You may wish to know the RR of dengue fever during September to December of the year when compared to the previous year.

Odds Ratio (OR)

The *odds ratio* (OR) is another measure of association used in the context of case-control studies. A group of individuals who are known to have the disease (called *cases*) are chosen for study and another group of comparable individuals without disease (called *controls*) are chosen for comparison.

When the data is presented as in Table 10.1 the OR is defined as $OR = \frac{a*d}{b*c}$

where * denotes multiplication.

In the case of data on smoking and heart disease we get OR = (85*118)/(28*71) = 5.04. This value is different from that of RR because the incidence of heart disease is common with smokers.

The OR value will be similar to that of RR when the disease is *uncommon* or *rare occurrence*. Further OR can be calculated from a case-control data while RR cannot be done. More information on RR and OR can be found in Indrayan and Satyanarayana (2006).

The strength of association two categorical between variables is understood with the help of some measures discussed below.

Yule's Coefficient (Y): This is a measure of association between two categorical variables proposed by Udny Yule in 1912. The value of Y lies between -1 and +1 and calculated from a 2x2 table using the formula $Y = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}$ and Y lies between -1 and +1. A value of +1 indicates perfect positive association and a value of -1 indicates perfect negative association.

Pearson's Phi coefficient (φ): This is a measure based on Chi Square value of a 2x2 contingency table and calculated using the formula $\varphi = \sqrt{\chi^2 / n}$ where n denotes the total of all values in the table. The value of φ lies between -1 and +1.

Cramer's V: This is another measure of association applicable for contingency tables having more than 2 rows or columns or both. The value of V lies between 0 and 1 such that 0 indicates no association and 1 indicates complete association.

The formula is $V = \sqrt{\frac{\chi^2 / n}{\min(k-1, r-1)}}$ where k denotes the number of columns and r denotes the number of rows.

Check Your Progress

1) What is meant by categorical variables? Write down the difference between ordinal and nominal variables.

.....

.....

.....

.....

.....

2) Write a short note on Relative Risk and Odds Ratio.

.....

.....

.....

.....

.....

10.3 LINEAR REGRESSION

Simple Linear Regression is a statistical technique used to explain the cause and effect relationship between two variables. It is a mathematical model (formula) which is derived from sample data collected as pairs of observations, on the predictor variable (X) and response variable (Y). It is assumed that the relationship is linear, to mean that the change in Y occurs at a constant rate with one unit change in X.

Regression analysis is commonly used to predict the response by reading the values of the predictors. Such predictors are called *prognostic factors* in some health studies. They can be either continuous or categorical. When more than one predictor is used to explain the response, we call the regression as *multiple linear regression*. There is another branch of regression analysis called non-linear regression which is beyond the scope of the present unit.

Regression analysis is closely related to the concept of *correlation coefficient* which measures the strength of linear relationship between two measured quantities like age and body weight. It is denoted by *r* and calculated from a sample of size n, by using Pearson's formula given below.

$$r = \frac{\sum XY - n\overline{XY}}{\sqrt{\left(\sum X^2 - n\overline{X}^2\right)}\sqrt{\left(\sum Y^2 - n\overline{Y}^2\right)}}$$

This correlation coefficient is also called product moment correlation. Modern computers help in quick calculation or *r*. The following are some points of interest to the practitioner.

1) The value of *r* lies between –1 and +1
2) *r* = 0 means no linear relationship
3) A smaller value of *r* indicates weak relationship
4) By using scatter diagram, the nature of relationship can be observed.

Simple Linear Regression

The simple linear regression model is given by Y = a + bX + e where 'a' is a constant called the intercept or baseline value and 'b' is called the regression coefficient. The term 'e' is called *random error component* and it represents the role of unexplained factors that might influence Y apart from X. The values of 'a' and 'b' are estimated from sample data by using a technique called *method of least squares*. Here is an illustration.

Illustration 10.3: The birth weight (Y) in kg of 15 newborn babies is measured along with the length of the foot (X) in cm.

S. No	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
X	7.7	7.9	8.3	8.0	8.2	8.5	7.4	8.0	8.5	7.2	7.0	8.0	7.7	8.2	7.6
Y	2.8	3.0	3.5	3.0	3.2	4.0	2.7	3.5	3.7	2.5	2.3	3.0	2.9	3.5	2.8

We wish to use foot length as a predictor to determine the birth weight (without a weighing machine) by using a linear regression model.

The scatter diagram drawn in Excel is shown in Figure 10.1. It can be seen that there is a linear and positive relationship between the foot length and birth weight of a baby (See Fig. 10.1).

The regression model is fitted with the following steps:

1) Right click on any dot on the scatter chart
2) Select the option 'Add trend line'
3) Select the option 'Display equation on chart'
4) Select the option 'Display R-squared value' on chart'
5) Click 'OK'

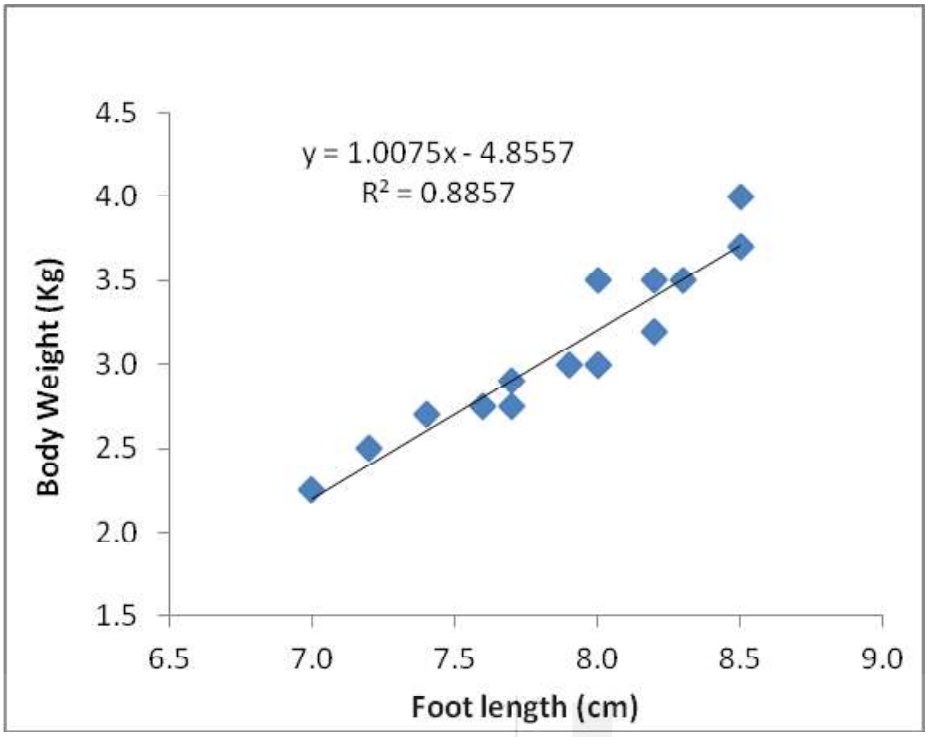

Fig. 10.1: Scatter Diagram and Linear Regression Equation

With these options we get the regression model displayed on the chart along with a measure for goodness of fit, called R-squared value. Let us understand the results:

a) The regression model is Y = –4.856 + 1.008* X.

b) The coefficients are a = –4.856 and b = 1.008

c) For an increase of foot length by one centimeter, the body weight increases marginally by 1.008 kg. Suppose a baby has foot length of 8 cm. Then the predicted body weight will be 3.20 kg for such babies.

d) The value of $R^2 = 0.8857$ which indicates that 88.57% of body weight can be explained by foot length by using this model.

e) The correlation coefficient between birth weight and foot length is simply the square root of 0.8857. Hence $r = 0.9411$ which is very high indicating that foot length is a good predictor of birth weight.

f) The *sign of correlation coefficient* is the same as that of the regression coefficient 'b' (+1.008) and hence the correlation is positive in this case.

While working with hand calculations we use the following formulas.

$$b = \frac{\sum\left(X-\overline{X}\right)\left(Y-\overline{Y}\right)}{\sum\left(X-Y\right)^2} \text{ and}$$

$a = \overline{Y} - b\overline{X}$ where $\overline{X}$ and $\overline{Y}$ denote the mean of X and Y respectively.

Multiple Linear Regression

The multiple linear regression is an extension of the simple regression. It relates Y with more than one explanatory variables $X_1, X_2, \ldots, X_k$ by using the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k K_k + e$$

where $\beta_1, \beta_2, \ldots, \beta_k$ denote the regression coefficients (weights) of the explanatory variables and β_0 is a constant. The value of β_0 represents the average of Y when all the X variables are set to zero. The term 'e' represents the random error component. The coefficients $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ are estimated from sample data by the method of least squares. The procedure involves complex calculations and a good method is to use software like Excel or SPSS. Once we estimate the coefficients, we say that the model is *fitted* to the data.

The goodness of the fitted model is judged by a measure called R^2 which lies between 0 and 1. Higher value indicates better fit. Since the fitting of the model is based on sample data, it is necessary to test for the statistical significance of (i) each regression coefficient (by Student's t-test) and (ii) the R^2 value (by F-test). By convention p-value < 0.05 indicates significance. Sarma (2010) discussed the details on multiple regression can be using SPSS.

Here is an illustration.

Illustration 10.4: The following data refers to the Quality of Life (QOL) measured on a 100 point scale of patients who have undergone a heart surgery. These patients were advised physio-therapy. The QOL is found to be dependent on age, gender, duration of physiotherapy (days) and the duration of hospital stay (days). The variables are coded as X_1 = Age (years), X_2 = Gender (Male = 1, Female = 2), X_3 = Duration of physiotherapy (days), X_4 = Duration of hospital stay (days) and Y = QOL score (max = 100). Sample data with 25 records is given below.

Table 10.2: Quality of Life (QOL) data on 25 patients

S. No	X_1	X_2	X_3	X_4	Y
1	21	1	12	5	65
2	25	1	10	8	58
3	26	2	6	5	59
4	26	2	10	3	63
5	26	1	12	4	64
6	27	1	6	4	61
7	28	2	11	3	65
8	28	1	10	2	67
9	29	1	11	8	54
10	30	1	12	6	62
11	31	1	8	7	60
12	31	1	6	6	59
13	32	2	13	5	65
14	32	2	7	7	57
15	32	1	14	2	65

16	34	2	10	4	65
17	34	2	10	3	63
18	35	2	5	8	54
19	35	1	8	5	63
20	36	1	15	5	66
21	36	2	9	7	56
22	36	2	15	4	67
23	38	2	10	3	68
24	39	1	8	8	61
25	40	1	11	3	63

We wish to build a multiple linear regression model relating Y to X_1, X_2, $X_3$and X_4.

Analysis

Using SPSS with options Analyze ® Regression ®Linear we get options window to select the dependent variable and other independent variables into the appropriate input boxes. Let us choose the 'method' as 'stepwise' and press OK. This gives the output which is summarized as below.

The model has $R^2 = 0.75$ which means 75% of the behavior of Y can be explained by the model.

1) The F-test (given in ANOVA table) shows high significance ($p < 0.001$) which means the goodness of the model is 'not an occurrence by chance'.
2) The regression coefficients (weights of explanatory variables) are as follows.

Table 10.3: Output of Multiple Linear Regression

Variable	***Coefficients***	***Standard Error***	***t Stat***	***P-value***
Intercept	62.202	3.944	15.77	<0.001
Age	0.075	0.093	0.800	0.433
Gender	–0.506	0.919	–0.551	0.5871
Duration of Physiotherapy	0.500	0.177	2.811	0.011*
Duration of Hospital Stay	–1.365	0.252	–5.398	< 0.001*

* Regression coefficient is statistically significant.

It can be seen that duration of physiotherapy and duration of Hospital Stay have a significant effect on the QOL. (p-value marked with *). The intercept (constant component of the model) is also significant but we are most of the time interested in the significance of predictor variables.

We end this section with the observation that linear regression is a tool useful to propose statistical models to explain the impact of explanatory variables on study outcomes. More details on handling regression analysis with SPSS can be found in Sarma (2010).

10.4 ANALYSIS OF VARIANCE (ANOVA)

The Analysis of Variance (ANOVA) is a statistical tool used to compare the mean values of a single continuous variable (Y) among three or more independent groups. The grouping variable is called a *factor* like the dose of age group, socio economic status etc., which is categorical with a few levels. ANOVA helps testing whether the group means differ significantly. It can be considered as an extension of the two-sample t-test. If there is only one factor affecting Y, we use *One-way ANOVA* but in general we can have more than one factor and we can test the significance due to all the factors and their combinations. Suppose we are measuring the response Y (like the Body Mass Index) of a group of persons receiving a treatment for weight reduction under three methods viz., A (Food Control), B (Exercise) and C (Both food control and exercise). We wish to check whether the mean of Y remains the same in all the three groups. Since the data is classified according to only one factor, it is called one-way classified data.

Let the means of Y in the groups A, B and C be denoted by μ_1, μ_2 and μ_3 respectively. Then we wish to test the null hypothesis H_0: $\mu_1 = \mu_2 = \mu_3$. The alternative hypothesis H_1: At least two means are not equal. It is assumed that the variance of Y remains the same in each group. It means that the response is consistent in all the groups and not contains extremely high or low values.

Then the null hypothesis is tested by using a ratio called F-ratio given as

$$F = \frac{\text{Mean Sum of Squares (MSS) due to the factor}}{\text{Residual MSS}}$$

The F-ratio is the test value. When the three means differ by a big gap we a high F-value. If the p-value is less than 0.05 we can reject the null hypothesis and consider that the factor has significant effect on Y; else accept the null hypothesis. When the F-ratio is significant, we infer that the difference in group means is not an occurrence *by chance*.

When the null hypothesis is rejected, we have to identify at which levels of the factor, the means happened to be different. This is done by using a *multiple comparison test* or pairwise comparison test. There are several such tests like Duncan's Multiple Range Test (DMRT), Least Significant Difference (LSD) test or Scheffe's test. The calculations for ANOVA can be done either with MS-Excel or with SPSS.

Consider the following illustration.

Illustration 10.5: A researcher has measured the Body Mass Index (BMI) of patients after undergoing a hormone therapy. The age of the patients was classified into three age groups as: i) < 30 years, ii) 31- 40 years and iii) 41 and above coded as 1,2,3 respectively. The data also contains another response variable called BMD (Bone Mineral Density) as shown in Table 10.4. We wish to check whether the mean BMI remains among the age groups.

Analysis

The calculations of ANOVA can be performed easily with Excel or SPSS. We illustrate this with SPSS. The following are the options.

1) Open the SPSS data file
2) Choose Analyze ® Compare Means ® One way ANOVA.
3) Select BMI into the 'dependent variable' box
4) Select Age Group into the 'Factor' window
5) Click on the *options* tab and select *descriptive statistics*
6) Press OK.

Table 10.4: BMI and BMD data

S. No	BMI	Age group	BMD
1	21.5	2	0.933
2	22.0	2	0.889
3	22.8	2	0.937
4	22.7	3	0.874
5	23.1	2	0.953
6	22.9	2	0.671
7	23.1	3	0.914
8	18.3	1	0.883
9	22.9	2	0.749
10	18.0	1	0.875
11	22.1	3	0.715
12	23.8	3	0.932
13	23.5	2	0.800
14	23.8	3	0.699
15	22.1	3	0.677
16	20.8	2	0.813
17	18.0	1	0.851
18	19.2	1	0.888
19	17.8	1	0.875
20	20.1	1	0.773

We will first report the mean and standard deviation of BMI in each group as shown below. (The SPSS output actually shows standard error and 95% confidence intervals in addition to the mean and standard deviation).

Age group	N	Mean	Std. Deviation
< 30	6	18.56	0.900
31 – 40	8	22.43	0.916
41 & above	6	22.93	0.771
Total	20	21.42	2.099

The ANOVA in its standard format appears as shown Table 10.5. For writing the report we can present the F value and the corresponding p-value (indicated by Sig. in SPSS output). We understand the following components.

1) There are two sources of variation viz., 'between groups' (due to age group) and 'within groups' (indicating the random and uncontrolled factors that might have influenced the mean BMI). The total variation in BMI is the sum of variation: a) due to age group (known factor) and b) due to unknown factors.

Table 10.5: One-way ANOVA table

Source of variation	Sum of squares	d. f.	Mean Square	F	Sig.
Between Groups	70.872	2	35.436	46.679	0.0001
Within Groups	12.905	17	0.759		
Total	83.778	19			

2) The degrees of freedom (*df*) is an indicator of the denominator to be used in Mean Square. The *df* indicates the number of independent observations (means) and the general formula is *df* = (k-1) if there are k-observations in hand. That is why the total *df* is (20-1) = 19. Since there are 3 groups we get (3-1) = 2 *df*. Finally, the *df* for within groups component is 19 (by subtraction).

3) The sum of squares and mean sum of squares are intermediate calculations, driving us to find the estimated variance due to age group. The F-value is called the F-ratio or *variance ratio* The heading 'sig.' indicates the p-value of the F-ratio. Since the p-value < 0.05 we reject the null hypothesis and conclude that the mean BMI among the three age groups differ significantly.

4) In a classical approach, one uses the critical value of F-ratio obtained from statistical tables. In this case the F-critical value for (2,17) degrees of freedom at 5% level of significance is 3.59. Since the obtained value is more than this, we reject null hypothesis.

Pairwise comparison of group means (like 1 versus 2, 1 versus 3 and 2 versus 3) is done with Duncan's test as a *post-hoc* procedure shown below. We should not use the two-sample t-test here!

The Duncan's test makes use of the Mean Sum of Squares calculated in the ANOVA table. There are 2 subsets into which the three means are classified.

Duncan's test for comparing the *mean BMI* among age groups			
Age group	**N**	**Subset for alpha = 0.05***	
		1	2
< 30	6	18.56	—-
31 – 40	8	—-	22.43
41 & above	6	—-	22.93
Sig.		1.000	0.318

* Means for groups in homogeneous subsets are displayed.

Those means which belong to the same subset are considered as homogenous (see the p-value), in the sense, they do not differ significantly. The mean BMI in

the age groups '31-30' and '41 and above' have no significant difference (p = 0.318) but both of them differ from the mean BMI of '<30' age group. This completes the one way ANOVA.

Check Your Progress

3) What is Multiple Linear Regression? How is it done with SPSS?

.....

.....

.....

.....

.....

4) What is ANOVA test? How is it performed in SPSS?

.....

.....

.....

.....

.....

Remark

a) It is also a practice to display the means by a bar chart, but SPSS shows by line chart.

b) Instead of Age group, if we use actual age (years) as *factor* in the SPSS options, we get an unpleasant output! Only categorical variables (on a nominal or ordinal scale) shall be used. This should not be done.

c) The ANOVA used here is called *univariate* ANOVA because only one response variable is considered among several groups. If two or more responses are studied at a time, we call it a *profile* we have to use an advance tool called *Multivariate ANOVA* or MANOVA.

d) ANOVA with more than one factor is done by SPSS using the *general linear model* available in the *Analyze option* in SPSS.

We end this discussion with the observation that ANOVA is a method statistical inference and needs careful interpretation. Reporting only the p-value is not enough. We have to comment on how the mean values differ among the groups.

10.5 SUMMARY

- We have learnt that measurement of association between categorical variables is studied by Chi Square test which is based on a contingency table. Some standard measures include Yule's Y, Pearson's Phi and the Cramer's V statistic. In the case of quantitative data (measured on an interval scale) we use Pearson's Correlation Coefficient.

- We have also seen that regression analysis, unlike correlation analysis measures the form of relationship between variables. Stepwise regression is a recommended method to establish a functional regression model.

- Further we have understood the principle of ANOVA meant for comparing the mean values of a characteristic among three or more groups of data sets. It is based on F-test and the computations can be carried out with SPSS. The analysis is completed only when a multiple comparison test (like Duncan's test) is performed.

10.6 REFERENCES

1) Indrayan, A., & Satyanarayana, L. (2006). *Biostatistics for Medical, Nursing and Pharmacy Students*. New Delhi: Prentice Hall of India.

2) Rao, P. S. & Richard, J. (2012). *Introduction to Biostatistics and Research Methods,* 5th edition, Prentice Hall of India.

3) Sarma, K.V.S. (2010). Statistics Made Simple Do it yourself on PC, 2nd Edition. New Delhi:Prentice Hall of India. New Delhi.

10.7 ANSWERS TO CHECK YOUR PROGRESS

1) Categorical variables are also known as *qualitative factors*. Data on such factors is not a measurement but it like an option to choose from a discrete list of possible values. For details refer section 10.0.

2) In some epidemiological studies we come across measures like *relative risk* and *odds ratio* both of which are based on 2 x 2 table of counts. The proportion of people of a population, having disease among all those who are exposed to a condition, is called the risk of disease whereas *odds ratio* (OR) is another measure of association used in the context of case-control studies. For details refer section 10.2.

3) The multiple linear regression is an extension of the simple regression. It relates Y with more than one explanatory variables. For details refer section 10.3.

4) The Analysis of Variance (ANOVA) is a statistical tool used to compare the mean values of a single continuous variable (Y) among three or more independent groups. The grouping variable is called a *factor* like the dose of age group, socio economic status etc., which is categorical with a few levels. For details refer section 10.4.

SUGGESTED READINGS

BLOCK 1: ESSENTIALS IN EPIDEMIOLOGY AND PUBLIC HEALTH

Beaglehole, R. & Bonita, R. (1997) *Public Health at the Crossroads*. Australia: Cambridge University Press.

Blumenthal, D. S. &Ruttenber, A. J. (1995). *Introduction to Environmental Health.* Second Edition. New York: Springer.

Last, John M. (1998). *Public Health and Human Ecology.* London: Prentice Hall.

Schneider, Mary- Jane. (2006). *Introduction to Public Health.* London: Jones and Bartlett.

Turnock, B. (1994). *Public Health.* Boston: Jones and Bartlett.

Park, K. (2007). *Park's Text Book of Preventive and Social Medicine*. Jabalpur: BanarsidasBhanot Publishers.

Grover, A.& Singh R.B. (2019). *Urban Health and Wellbeing.* Japan*: Springer*

Mahajan, M.C., Gupta B.K. (2013). *Textbook of Preventive and Social Medicine*. 4th edition, Revised by R.N Roy and I. Saha, Jaypee Brothers Medical Publishers Ltd.

https://mohfw.gov.in/

https://www.undp.org/content/undp/en/home/sustainable-development-goals.html

BLOCK 2: PSYCHOLOGICAL, BEHAVIOURAL, AND SOCIAL ISSUES IN PUBLIC HEALTH AND MANAGEMENT

Gupta Monica (2016). *Public Health in India: An Overview*. Working Paper Series 3787. World Health Organisation.

Glanz, K.,Rimer, B. & Viswanath, K. (Ed).(2008). *Health Behaviour and Health Education Theory, Research and Practice.* San Fransisco: Wiley Imprint.

Jenkins, David. (2003). *Building Better Health : A Handbook of Behavioural Change.*Washington DC: Pan American Health Organisation.

Kawachi, I., &Wamala, S. (Eds.). (2006). *Globalization and Health*. Oxford University Press.

Kleinman, A. &Benson, P. (2006) *Anthropology in the Clinic* PLoS Medicine(10): e294.

Kleinman, A. (2004). *Culture and Psychiatric Diagnosis and Treatment*:*The Trimbos Lecture.* Harvard University.

Park, K. (2007). *Park's Text Book of Preventive and Social Medicine*. Jabalpur: BanarsidasBhanot Publishers.

Rachel Davis, Rona Campbell, Zoe Hildon, Lorna Hobbs & Susan Michie (2015).

Theories of behaviour and behaviour change across the social and behavioural sciences: *a scoping review.* Health Psychology Review, 9:3, 323-344

World Health Organisation(2008). Geneva: *Commission on Social Determinants of Health Report.*

BLOCK 3: RESEARCH AND STATISTICAL METHODS IN PUBLIC HEALTH

Indrayan A &L.Satyanarayana (2006). *Biostatistics for Medical, Nursing and Pharmacy Students,*Delhi: PHI Learning Pvt.Ltd.

Microsoft Excel (2010) Step by Step (eBook) Web resources: https://www.spss-tutorials.com/basics/

Sabine Landau &Brian S. E. (2004). *A Handbook of Statistical Analyses using SPSS.*USA: Chapman & Hall/CRC Press LLC.

Sundar Lal & Vikas (2018), Public Health Management – Principles and Practice, Delhi: CBS Publishers and Distributors Pvt. Ltd.

Suresh K Sharma (2014). *Nursing Research and Statistics* (2nd Edition).Gurugram: Elsevier RELX India Private Limited.

Wayne W Daniel (2014). *Biostatistics: A Foundation for Analysis in the Health Sciences*. Wiley Series in Probability and Statistics.