BLOCK 3

RESEARCH AND STATISTICAL METHODS IN PUBLIC HEALTH

ignou
THE PEOPLE'S
UNIVERSITY

UNIT 8 RESEARCH METHODS AND STATISTICAL TOOLS*

Contents

Learning Objectives

After reading this Unit, you will be able to understand:

- The basic principles of research;
- The need for statistical tools in research;
- Design of protocols for sample studies and case studies;
- Statistical description of data; and
- Statistical tests of significance.

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

8.0 INTRODUCTION

The health of an individual is the primary determinant of quality of life. About 70% of the Indian population lives in rural areas. There is a great disparity in health care access between rural and urban areas. The burden of disease among people is an indicator of health care in any country. According to World Health Organization (WHO) 53.6% of total deaths in 1990 were due to communicable diseases including maternal, neonatal and nutritional diseases. In 2016 there were 61.8% of total deaths attributable to non-communicable diseases including cancer, heart diseases, diabetes etc. (Source: Health of the Nation's States: The India State-Level Disease Burden Initiative, 2017 published by the Indian Council of Medical Research, Public Health Foundation of India and the Institute for Health Metrics and Evaluation).

The quality of health care services at an affordable cost is a vital issue. The Central and State Governments as well as health insurance companies have come up with various schemes to meet the financial burden of health care services. However, a major role in health care delivery system is attributed to private sector which consists of 58% of the hospitals in the country, 29% of beds in the hospitals, and 81% of doctors according to a study done by Thayyil and Jeeja (2013).

The functioning of public health institutions largely depends on the statistical data available with them. For instance, the government obtains data on issues like percentage of houses without toilets, number of villages with safe drinking water, number of children to be vaccinated etc.

A lot of research is going on in the field of innovative health care with different objectives. Some are listed below:

Finding new methods of identifying factors causing the diseases;

Educating people on healthy lifestyles;

Estimating the number of deaths (mortality) due to specific diseases like cancers, kidney diseases, and heart diseases;

New and improved methods for clinical investigations (like X-Ray, Scans, MRI etc.), medication, patient care, follow-up etc.

All this needs a scientific approach and the research team usually comprises a statistician.

Role of Statistics: Statistics is a subject that deals with collection, organization and analysis of data. It helps in drawing inferences about population based on sample data. Research studies are broadly classified as follows:

a) Prospective studies in which outcomes are observed in response to interventions (known antecedents). They can be either observational studies or comparative studies.

b) Case Control studies which are retrospective studies in nature. The outcomes are known, and researcher investigates the possible causes for the outcome.

Every research study requires a well written protocol that specifies: a) aim and objective(s) of study, b) target group (cohort) to be addressed, c) duration of the study, d) sampling design, d) method of data collection and analysis, e) ethical issues, if any, f) budget estimates etc.

In the following section we shall understand some concepts related to sampling.

8.1 POPULATION AND SAMPLE

A population is the collection of all subjects (people, animals, plants etc.) related to the goal of the study and is also known as cohort or study group. A sample is a representative portion (subset) of population.

For instance, we may define a population as 'all women in a town below 30 years and suffering from anemia'. We do not know exactly; the characteristics of this population and the researcher aims at knowing them with the help of a sample study. The size of the population is usually large and if every member of the population is to be studied, it is called census or screening. This is, however, costly, time consuming and demands a large team of trained investigators for data collection. In some cases, screening carries no meaning as in the case of 'attempting to draw the total blood from a person to know the blood-sugar level'.

Sampling, on the other hand, is less costly and data collection can be done with few trained persons. The results obtained from the sample will be generalized to the population. This is known as inductive approach and the results are often considered as estimates of the unknown parameters of the target group.

A sampling design is a scheme (or a plan) according to which data will be organized in the study. It specifies the aspects such as, sampling frame (the complete list of population members), sample size, method of sampling, design of questionnaire, data entry and validation and reliability measures for data.

Sampling should be unbiased so that the investigator shall not influence the selection of respondents or the data collection process. Sampling methods are of two types viz., random sampling and non-random sampling.

Check Your Progress

1) Write a note on the role of statistics in health research.

.....

.....

.....

.....

2) Distinguish between population and sample with suitable examples.

.....

.....

.....

3) What is a sample design? What are its main components?

.....

.....

.....

.....

.....

We shall understand about the details of these methods as below:

8.2 RANDOM SAMPLING

In this method, the population members (units) are included in the sample by a random or lottery type mechanism. It prevents personal bias in recruiting the members into the study. It is the best way to produce unbiased conclusions. We have the following methods of random sampling.

8.2.1 Simple Random Sampling

In this method every unit of the population will have equal chance of getting selected into the sample. It is applicable when the population is homogeneous with respect to factors like age, gender, body mass, level of education etc. For instance, let there be 150 houses in a village. In order to select 30 houses, write 150 slips, each slip having the house number, put them in a box, shuffle the box and select one unit at a time, until 30 houses are selected (repetitions to be dropped).

8.2.2 Stratified Random Sampling

This method is used when the population units are heterogeneous like having differences in level of education, place of residence, socio economic status etc. Each group is called a *stratum* and sampling must cover all the *strata* (pleural of stratum) to avoid over representation of a few groups.

8.2.3 Systematic Sampling

This method is used when the population units are already arranged in a sequence' like the residential houses in a colony like 1, 2, 3, …,100. Systematic sampling starts with one member at random and selects successive houses with a fixed *gap* of houses.

8.2.4 Cluster Sampling

Cluster Sampling is quick and easy to administer. Suppose we wish to carry out a study on immunization. We may for instance select 20 villages, each village being a cluster. Within each cluster let us take 30 households at random so that 600 households will be covered by the study. Each cluster therefore contains heterogeneous members and hence represents the most characteristics of the population. This is one method recommended by the World Health Organization (WHO) for conducting surveys on immunization.

8.2.5 Multi Stage Sampling

This method is necessary for conducting a survey in a large area like a state or a big region within a state. Suppose we wish to study the prevalence of anemia in a region by visiting a fixed number of households. It is then convenient to first select in stage 1, a predetermined number of districts at random. In stage-2 we may select few Primary Health Centers (PHC) at random since each PHC covers some villages. In stage-3 we select a fixed number of villages at random and finally in each village a predetermined number of households may be selected at random. Thus, in this method, the sampling units change from stage to stage.

The method of sampling shall be specified in the study protocol along with the sample size.

8.3 NON-RANDOM SAMPLING

Non-random sampling is another method, where the researcher contacts purposefully a group of persons relevant to the study. Though this method does not support a scientific approach, still it is found useful to obtain quick results like a survey on a health care insurance. Such a survey is usually done only from those who have registered for the insurance and accessible to the researcher. However, the results of such studies can't be generalized to the target group (population).

Snowball sampling is one method to extract information from people having a special medical condition which carry stigma. Examples include survey on sex workers, drug abusers or HIV patients. It is difficult to know the complete population of such people and hence a random sample may not be feasible. If we are able to catch one person relevant to the study, he/she may serve as a guide to reach similar persons in the study area and all such persons form the sample.

8.4 CASE CONTROL STUDIES

In a case-control study the researcher investigates a set of patients called *cases* who are identified with a health condition. For instance, persons with hypothyroidism will be cases and the researcher attempt to identify the possible causes for it.

It is a common practice to select as cases only those patients who are newly diagnosed for the disease under study because it is easy to identify the exposure to conditions that might have caused the disease. The control subjects are usually taken as matched, in the sense that they have most factors (like age, gender, body weight) like those of cases except for the incidence of disease under study.

The pattern of factors like diabetes, obesity, age will then serve as possible causes often known as *biomarkers* for the disease. In a case-control study, the cases are compared with controls who are either normal subjects or those treated with a placebo. This helps in identifying factors, if any, which are dominant in cases but not in controls. Indrayan and Satyanarayana (2006) contain several practical examples on research designs.

In the following section we shall study the basic statistical methods of describing statistical data.

8.5 DESCRIPTIVE STATISTICS

Statistical data is usually expressed in three forms viz., tables, graphs and summary values. Data which is observed on a nominal or an ordinal scale are summarized by number (count) and per cent. Such a data is called *Categorical data*. If the data is measured on an interval scale, we summarize it using averages and some measures of variation.

Let us see a situation of categorical data and its summary in illustration 8.1 below.

Illustration 8.1 (Description of categorical data): The following data refers to the distribution of Leprosy patients according to the type of Leprosy and Gender.

Type	Number of Patients		
	Male	Female	Total
Tuberculoid	77	74	151
Lepromatous	35	33	68
Indeterminate	10	8	18
Borderline	7	5	12
Total	129	120	249

This data contains information on the type of leprosy and the number of cases gender-wise.

From the above tables it is clear that out of 249 patients, 151 (60.64%) have Tuberculoid. The pattern of disease is more or less similar between males and females.

The variable of interest is the type of leprosy which is given as 4 categories. The data on each category is the count (number of cases observed). There is another factor namely gender (male and female). The data is therefore two dimensional. Summarization of such data is usually done in terms of percentage. For instance, you can observe what percentages of males are exposed to Lepromatous? The answer is simply 35 out of 249 and or 14.05%.

You may observe what percent of female patients fall under borderline category. You should get 4.16%. We also observe that Tuberculoid is the dominant type of leprosy with 151 out of 249 which means 60.6%

Another way of describing such data is by using a bar chart as shown in Figure 8.1.

As an alternative you can describe the data by separate pie chart for male and female patients. Excel can be used to draw these charts.

Suppose you have data on Serum (blood) Creatinine measured in mg/dl. So, the data is continuous. The values are not necessarily whole numbers and fractional values are also allowed.

Similarly, the body mass index, fasting blood glucose and birth weight of a new born child are some examples of continuous variables. Such variables are described by averages instead of counts and percentages.

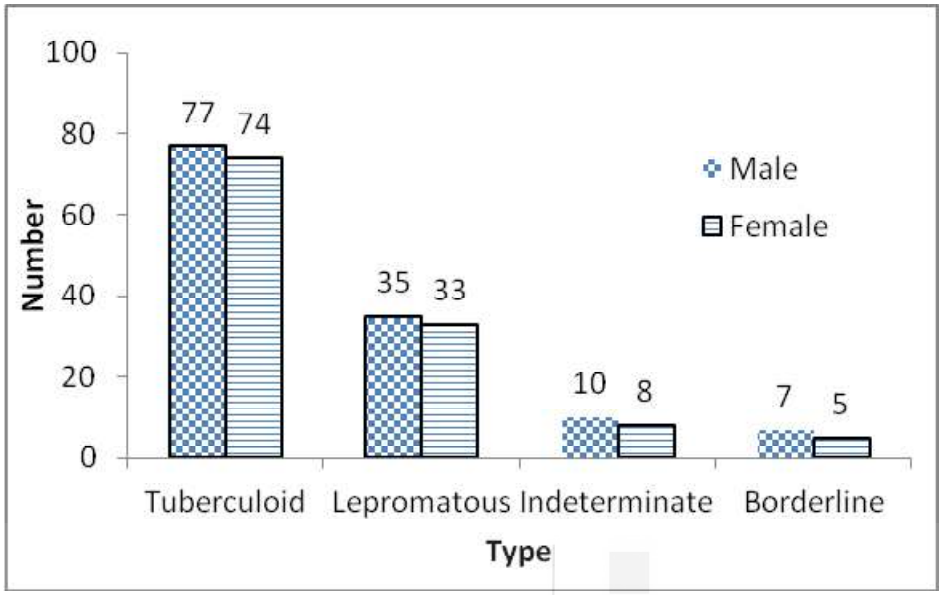


Fig. 8.1: Distribution of Patient by Type of Leprosy

8.6 MEASURES OF CENTRAL TENDENCY

Very often in a data, we find a tendency that most of the data is clustered around a central value which is known as the average. This is called Central Tendency and expressed in terms of some measures called averages. We shall discuss three commonly used averages.

8.6.1 Mean

The most commonly used average of the data is the Arithmetic Mean or simply the mean. It is simply the sum of all values divided by the number of values. If $x_1, x_2,\ldots,x_n$ are the n-values in the data, then the mean is given by

$$\overline{X} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Here is an illustration.

Illustration 8.2: The Creatinine levels (mg/dl) of 13 patients are given below.

0.70, 0.60, 0.20, 0.26, 0.40, 0.50, 0.35, 0.15, 0.30, 0.36, 0.60, 0.20 and 0.45

The sum of these 13 values is 5.07. So, you get mean = 5.07/13 = 0.39 mg/dl

The mean is an important measure and popularly used summary of data. It is based on all the data values and mean cannot be larger than the largest value. Mean also has a draw back in the sense that by including very large or very small values to the data, the mean gets drastically changed.

You may note that the mean (average) of the values {2, 3, 4, 5, 6} is 20/5 = 4. Suppose the last value is recorded as 36 by mistake; the mean goes to 50/5 = 10! Again, in the original data when 6 is written as 0 the mean goes down to 14/5 = 2.8 (the number of values is still 5 because 0 is also a value).

When the size of data is large say 100 or 300 values, one way of describing the data is by preparing a frequency table or table of counts. Here is an example.

Illustration 8.3: The Fluoride level (mg/L) observed in 200 samples collected from different parts of a district are given bellow.

Fluoride level	0.3 - 0.5	0.5 -0.7	0.7 -0.9	0.9-1.1	1.1-1.3	1.3-1.5	1.5-1.7
# samples	6	24	67	61	22	12	8

# indicates number or frequency.

We wish to know the average (mean) fluoride level.

Analysis: We have described the data by using *intervals* of fluoride level and counted the number of samples for which the value belongs to each interval. These intervals are also called *classes* or *bins.* In each interval, all the values less than the upper limit will be counted.

Now to find the mean of this data, we find the mid value of each interval which is the average of the upper and lower limits. For instance, the mid value of the interval *0.7-0.9* is (0.7+0.9)/2 = 0.8. Then the mean is found as follows (* indicates multiplication).

$$\{0.4*6 + 0.6*24 + 0.8*67 + 1.0*61 + 1.2*22 + 1.4*12 + 1.6x8\}/200$$

This gives $\bar{x}$ =187.4/200 = 0.94 mg/L. It means on an average the level of fluoride in the study area is 0.94 mg/L.

8.6.2 Median

Median is another average often used for ordinal data and also for data that contains extreme values. It is the middle value of the data when the data is arranged in ascending or descending order. In case of even number of data values there will be two middle values and we take their average as the median. It is largely used in life-testing and survival analysis. Median has the property that 50% of data values will be below the median and the other 50% will be above the median. We also call it 50^{th} Percentile.

The first quartile (Q_1) is a value that has approximately 25% of the data below it. The third quartile (Q_3) is a value that has approximately 75% of the data below it. By this rule Q_2 will be the median.

8.6.3 Mode

The value which occurs maximum number of times in a data is called the Mode. There can be single mode, two modes and sometimes multiple modes. There is no mode if the data is {2, 5, 8, 1, 4, 9, 6} since no value has got repeated.

8.7 MEASURES OF VARIABILITY

Variation is an inherent characteristic of measured values. It occurs due to several reasons some of which cannot be controlled. The spread of data values around a target is called dispersion or scatter. A stable or consistent data will have less

dispersion than an unstable data. The numerical measures of variation are called *dispersion measures* or *measures of spread.*

The following are some measures.

8.7.1 Range

It is the difference between the largest and the smallest values of the data. It is useful when the data is fairly stable like the height of an adult male. When the range is high, it means the data has high variation. This situation occurs when some *abnormal* values occur in the data. Sometimes range is specified as {min, max} but for comparing two or more data sets we have to use Range = {Max – Min}, which gives a single value. A range of 0.6 mg/L of fluoride level indicates less variation than a range of 1.1 mg/L.

8.7.2 Variance and Standard Deviation

Variance is a measure of spread of data values around the mean (M). If many values are away from the mean, we get high variance and if many are close to the mean, we get less variance. The population variance is denoted by σ^2 and given by the formula

$$\sigma^2 = \frac{\sum (x_i - M)^2}{N}$$

where N is the number of units in the population.

It is always positive but expressed in squared units. For instance, if height is measured in centimeters, the variance has to be expressed in 'centimeter square' which is difficult for comprehension.

In order to measure the variation in natural units we use the Standard Deviation (SD) which is the positive square root of variance given by

$$s = \sqrt{\frac{\sum (x_i - M)^2}{N}}$$

This sample standard deviation (s) calculated from a sample of size 'n' by using the following formula is the estimate of σ.

$$s = \sqrt{\frac{\sum (X_i - \overline{X})^2}{n}}$$

where $\overline{X}$ is the sample mean. In case of *small samples,* we use a different formula for SD given by $s = \sqrt{\frac{\sum (X_i - \overline{X})^2}{(n-1)}}$ Note the denominator is (n-1) here and it is useful for both small and large samples.

There are two other measures of dispersion which are used to describe the shape of the data distribution (or histogram). These are outlined below.

8.7.3 Skewness

It is a measure of lack of symmetry in the distribution. When the distribution has equal number of values below and above the central value (mean) we say the distribution is *symmetric*. A distribution with a long-left tail is said to be *left-skewed.* With the same logic, a *right-skewed* distribution will have long right tail. Such distributions are shown in Figure-8.2. below.

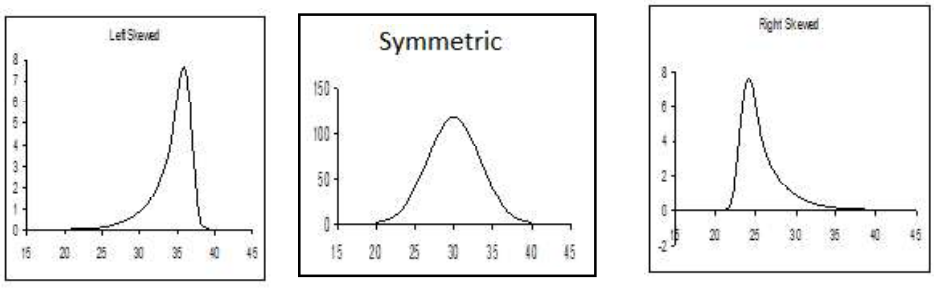


Fig. 8.2: Pattern (shape) symmetric and assymetric distributions

Skewness is measured by using Karl Pearson's coefficient

$S_k = \frac{3(\text{Mean} - \text{Median})}{\text{SD}}$ and this value can be positive, negative or zero. If the distribution happens to be symmetric then $S_k = 0$ because in that case Mean = Median.

8.7.4 Kurtosis

It is a measure of the *peakedness* in the shape of the distribution. Some distributions tend to have a high *peak* while some of them look *flat*. The distribution which is neither very tall nor very flat is known as *normal*. A distribution that is peaked higher than the normal is known as *Leptokurtic* distribution and the one that is peaked lower than the normal is known as *platykurtic*. For a well behaved data the value of Kurtosis will be 3. Several details of basic statistics in biology and health sciences can be found in Sundar Rao and Richard (2012).

In the following section we learn some basic ideas of statistical inference.

8.8 TESTS OF SIGNIFICANCE

Statistical methods not only help in describing data patterns and summarizing the data but also in drawing meaningful inferences about the unknown features of a population, based on sample data. This area of statistics is known as inferential statistics. The focus lies on two important areas;

a) Estimating the unknown parameters (like the prevalence of 'low birth weight')

b) Testing the truth of a hypothesis (belief) about the population characteristics using sample data.

Statistical tests of hypothesis are a part of statistical inference. A hypothesis is a numerical statement about the unknown parameters of a population. In literal

terms a hypothesis is a belief or a verifiable statement made by the researcher. In essence we wish to check whether the sample findings are an occurrence by chance (not significant) or can they be attributed to known factors? There is a subtle difference between the concept of tests of significance and test of hypothesis but in all practice, they convey the same.

One or more hypotheses are framed before taking up a study and the truth of these hypotheses is verified in the light of the sample data, as evidence, collected by the researcher. For obvious mathematical reasons we start with a hypothesis that there will be no effect or no phenomenon and check how much likely it is to be true. The answer appears only in terms of probability.

We need to know the following technical terms.

1) **Null hypothesis**: Denoted by H_0 this is a statement that mentions a *null effect* or absence of an effect. It is stated as a single value addressing the parameter of interest. Sometimes it is also stated as a hypothesis of *no difference*. Here are a couple of examples

   a) $H_{0:}$ The average knowledge scores before and after training remains the same.

   b) H_0: There is no stunting (lack of growth) in children in the study area.

   c) H_0: The prevalence of smoking in a given village is 30%.

2) **Alternative hypothesis**: When the null hypothesis is not supported by the data, we say it is *rejected* and we agree to *accept* another statement called *alternative hypothesis* denoted by H_1. Either the null or the alternative hypothesis will be true but not both in a given context.

   There are two ways of specifying the alternative hypothesis as below.

   a) **One sided alternative**: In this method we specify the direction of the result, if it is not zero. For instance, 'prevalence of smoking < 30%' is a one-sided alternative. We can also consider 'prevalence of smoking > 30%' as the alternative.

   b) **Two-sided alternative**: In this method we do not specify the direction of the result; it can be positive or negative. For instance, 'prevalence of smoking not equal to 30%' is a two sided alternative hypothesis.

3) **Type-I and Type-II errors**: Since the decision on the null hypothesis (H_0) is based on a sample, from the population, we are likely to reject H_0 even if it was really true (the sample might be poor evidence). This is a *false rejection* and called type-I error. Similarly, we may commit type-II error of *false acceptance*. Both these errors cannot be totally avoided but we can fix the *error rates* and develop a statistical procedure.

4) **Level of significance**: The maximum tolerable rate of false rejection is called the level of significance (LOS) denoted by the Greek letter α (alpha). By convention this value is taken as 5% though we sometimes take 1%. We write $\alpha = 0.05$ to mean that in 5% of instances the procedure may reject H_0 even when it is really true.

5) **Critical Value**: The test procedure, after using a formula, gives a value called *test value* obtained from the sample data. This value is compared

with a *critical value* or *threshold value* which is available in statistical tables. These critical values are based on the type of test and the value of α. When the test value exceeds the critical value, we reject H_0 at 5% LOS. It means that the null hypothesis is very unlikely to be true. For tests based on large samples (not less than 30), the critical value at 5% level for a two sided alternative is 1.96 and for one sided alternative, the critical value is 1.65.

6) **Power of the test**: This is the probability with which we are able to accept the alternative hypothesis when it is really true. This is denoted by $(1-\beta)$ where β denotes the *rate of false acceptance.* Thus, power is the *rate of true acceptance.* As a convention, researchers look at tests with 80% power or more. It means $\beta = 0.20$.

7) **p-value of a test**: It is an alternative approach to using critical values from tables. The p-value is the actual probability of type-I error calculated based on the sample data. This is compared with α. If p-value is less than α, we reject the null hypothesis and say that the findings are *significant*. As a rule, smaller p-value leads to statistical significance. If p-value exceeds a we say that 'the findings could also be due to chance'. The calculation of p-value is computer intensive procedure.

Check Your Progress

4) Explain various methods of random sampling.

.....

.....

.....

.....

.....

5) What are measures of central tendency? Explain about arithmetic mean.

.....

.....

.....

.....

.....

6) Write a short note on: a) null hypothesis and b) p-value.

.....

.....

.....

.....

.....

In the following section we shall discuss some statistical tests commonly used in public health studies.

8.8.1 Test for Proportions

The objective here is to compare the observed prevalence of a disease with a hypothetical prevalence. For instance, the researcher finds that the prevalence is 35% basing on a sample of say 250 individuals. It is hypothetically believed that the prevalence is 50% in the population. We wish to test whether the difference between the hypothetical and observed values of prevalence is significant. This is called one sample test for proportion because percentage and proportion convey the same meaning.

One sample test for proportion: H_0: $p = p_o$ (hypothetical value expressed as proportion) and H_1: $p \neq p_0$ (two sided alternative). The test value is computed by using the formula

$$Z = \frac{p - p_0}{\sqrt{\frac{P_0(1-p_0)}{n}}}$$

The numerator is the difference and the denominator is called *standard error.* This value can be either positive or negative but we ignore sign and read the test value. Taking $\alpha = 0.05$ the critical value is 1.96. If $Z > 1.96$ reject H_0 and conclude that the difference is significant (not an occurrence by chance).

Illustration 8.4: In a sample study on 150 school children, it is found that 26 students had hearing difficulty. Will this study support the statement (belief) that 20% of school children in general have hearing difficulty?

Solution: Here n = 150 and the sample proportion is p = 26/150 = 0.17 or 17%. The null hypothesis is H_0: $p = 0.20$ (p_0) and H_1: $p \neq 0.20$ (two tailed hypothesis). Now find

a) Difference = 0.17 – 0.20 = –0.03

b) Standard Error = $\sqrt{\frac{P_0(1-p_0)}{n}} = \sqrt{\frac{0.20 * 0.80}{150}} = 0.033$

c) Test value (Z) = 0.03/0.033 = 0.909

Taking $\alpha = 0.05$ the critical value is 1.96. Since $Z < 1.96$ we cannot reject the null hypothesis and hence the difference is not significant. We may accept the belief of the researcher.

Two sample test for proportions: We wish to test whether the difference between the proportions obtained from two independent groups is statistically significant.

If p_1 and p_2 denote the two proportions, we frame the null hypothesis as H_0: $p_1 = p_2$ (difference is zero) and H_1: $p_1 \neq p_2$ (two sided alternative). The test value is computed as

$$Z = \frac{p_1 - p_2}{\sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}}}$$

Where, $p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$ is called the combined proportion and q = (1-p). Taking $\alpha = 0.05$ we get the critical value as 1.96. If Z > 1.96 reject H_0 and consider that the difference is significant.

Illustration 8.5: In a community health study covering a cohort A it is found that 22 out of 120 pregnant women are found to be anemic and 41 out of 150 pregnant women in cohort B are found to be anemic. At 5% level of significance, can we consider that the cohort B has more anemic women than cohort A?

Solution: Here $n_1 = 120$ and $n_2 = 150$. The sample proportions are $p_1 = 22/120 = 0.18$ and p2 = 41/150 = 0.27. The null hypothesis is H_0: $p_1 = p_2$ (no difference) and H_1: $p_1 < p_2$ (one sided hypothesis). Now find

a) Difference = 0.18-0.27 = -0.09

b) Combined proportion $(p) = \frac{120*0.18+150*2.27}{120+150} = 0.17$

c) q = 1-0.17 = 0.83

d) Standard Error $= \sqrt{\left\{\frac{0.17*0.83}{120} + \frac{0.17*0.83}{150}\right\}} = \sqrt{0.0012 + 0.0009} = 0.046$

e) Test value (Z) = 0.09/0.046 = 1.96 (ignoring the negative sign)

Taking $\alpha = 0.05$ the critical value is 1.65. Since Z > 1.65 we cannot accept the null hypothesis and hence the difference is significant. We may accept the belief of the researcher that cohort B has more anemic women than cohort A.

8.8.2 Test for Means (t-test)

With these tests we can compare the observed sample mean of a characteristic with a hypothetical mean. We can also test for the significance of the difference between the means of two independent or dependent samples. It is assumed that the individual data values follow normal distribution.

- When the sample size is large and the null hypothesis is true, then test value follows normal distribution and the tests are known as Z-tests (proposed by R.A. Fisher)
- With small samples the normality assumption of the test value does not hold good. In this case we use a special distribution called Student's t-distribution (proposed by W. S. Gosset whose pen name was Student). These tests are known as t-tests.
- Interestingly, t-test can be applied for both small and large samples while Z-test cannot be applied on small samples.

Two sample t-test for means: This is a test for comparing the difference between the means of characteristic, observed in two *independent samples.* Suppose hemoglobin is measured from two independent samples of patients. One group is treated with a *high protein diet* and the other with *normal diet.* We wish to test whether the difference in the sample means is significant.

The null hypothesis is H_0: *The difference is zero*. The two-sided alternative could be taken as H_1: *The difference is not zero*. Assume that for the two groups, sample sizes are n_1 and n-$_2$ and the means are $\overline{X_1}$ and $\overline{X_2}$ respectively. Further let s_1 and s_2 be the standard deviations of values in the two groups respectively.

We have to find the combined SD denoted by $S = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$. The test value is calculated as $t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$. The critical value is found from tables of t-distribution with $n_1 + n_2 - 2$ degrees of freedom. If the calculated value of t, ignoring the sign, exceeds the critical value, we consider the difference as significant.

Sometimes we get a situation where a measurement is taken for each subject before and after giving an intervention (like treatment, training etc.). In such cases we have to use the *paired t-test.* This test contains only a single sample of n values but for each case there will be two observations, one *before* and another *after* the treatment. Such data is sometimes called as 'pre-post data'. We wish to test the significance of the difference between the pre and post means.

We end this unit with a note that *statistics is the science for understanding real life phenomena in the presence of uncertainty* and the *search for the truth* is the goal of analysis.

8.9 SUMMARY

In this unit we have learnt the following.

- Statistics helps in data collection and analysis of public health studies. The importance of statistics lies in summarizing vast volumes of data into meaningful summary like tables, charts and measures like averages.
- Sampling is the ideal method of data collection because total survey of a population is difficult, costly and time consuming. Several sampling methods like simple random sampling, stratified sampling and cluster sampling are available to avoid investigator's bias in selection.
- Statistical data is interpreted with the help of summary values like mean, variance and standard deviation. Since all the sample values are estimates of the population features (parameters) they are subject to sampling errors. For this reason we conduct tests of hypothesis to ensure that the errors in the inferences do not exceed what is promised (level of significance and power of test).

8.10 REFERENCES

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

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

Thayyil, J. & Jeeja, M. C. (2013). Issues of Creating a New Cadre of Doctors for Rural India. *International Journal of Medicine and Public Health*, *3*: 8-11.

8.11 ANSWERS TO CHECK YOUR PROGRESS

1) Statistics helps in drawing inferences about population based on sample data. For details refer section 8.0.

2) A *population* is the collection of all subjects (people, animals, plants etc.) related to the goal of the study and is also known as *cohort* or *study group.* A *sample* is a representative portion (sub set) of population. For details refer section 8.1.

3) A *sampling design* is a scheme (or a plan) according to which data will be organized in the study. It specifies the following aspects. (a) Sampling frame (b) Sample size (c) Method of sampling (d) Design of questionnaire (e) Data entry and validation (f) Reliability measures for data.

4) In random sampling, the population members (units) are included in the sample by a *random* or lottery type mechanism. Various methods of random sampling are: (a) Simple Random Sampling (b) Systematic Random Sampling (c) Stratified Random Sampling (e) Cluster Sampling Multi-Stage Sampling. For details refer section 8.2.

5) Different measures of central tendency are: (a) Mean (b) Median (c) Mode. The most commonly used average of the data is the Arithmetic Mean or simply the *mean*. It is simply the sum of all values divided by the number of values. For details refer section 8.6.

6) Null hypothesis is a statement that mentions a *null effect* or absence of an effect. It is denoted by $H_{0.}$ p-value is an alternative approach to using critical values from tables. The p-value is the actual probability of type-I error calculated basing on the sample data. For details refer section 8.8.

UNIT 9 DATA ANALYSIS*

Contents

Learning Objectives

After reading this Unit, you will be able to:

- Create a data file in Excel;
- Edit data editing in the Excel Sheet;
- Save, retrieve and export data to other software;
- Perform basic statistical analysis with Excel;
- Work with SPSS using Excel data; and
- Perform Regression analysis and ANOVA with SPSS.

9.0 INTRODUCTION

Microsoft (MS) Excel is a spreadsheet package designed to organize simple data sets, perform general calculations, create graphs and handle statistical analysis. This is a member of the MS-Office family that contains Word, Excel, Power Point and Access in a single suit. Excel offers a number of built-in programs, which run with simple mouse operations.

We can use Excel to store patient's data, case by case and store the values of a large number of parameters. We can make simple calculations like total, difference, percentage and even scientific calculations on any number of values. Several organizations use Excel as the platform to share data. Many web sites provide an option to download data into Excel format (like the statement of your bank account). Excel is also available in smart phones and tablets (WPS office is one such tool).

There are several software packages for statistical applications but many of them are expensive. Excel offers some tools to perform basic statistical analysis.

Let us start understanding how a data file can be created in Excel.

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

9.1 CREATING A DATA FILE IN EXCEL

When MS-Office is installed in the computer we find the Excel button either on the task bar or in the start menu. A click on this button opens Excel. A new workbook named Book1 will open and it normally contains 3 sheets named Sheet1, Sheet2 and Sheet3 and the active sheet is Sheet1. In Excel a data file is called a workbook. We refer to Excel 2010 for discussion.

The following are some features of an Excel worksheet.

a) Every sheet contains a 16384 columns labeled as A, B,…Z, AA, AB,…XFD.

b) The number of rows in a sheet is 1048576 numbered as 1,2,3.

c) Each cell of the sheet is identified with its cell address. For instance, A3 means the cell at the intersection of column A and row 3.

d) Each cell is used to enter different types of data like numbers, text, date, time, currency etc. The default status is called *General* for which no specific format is specified.

e) The first row is normally used to indicate the Column Heading of the data file. There should not be multiple headings, if the sheet is meant for analysis. Every column should have only one heading.

Remark: Some users continue to type the data sets one after another top to bottom, by providing a sub heading in between. It is not correct.

Consider the following illustration.

Illustration 9.1: A typical data file on the number of health camps conducted under NTR Vaidya Seva Scheme 2016-17 is shown in table 9.1. (**Source:** Socio Economic Survey 2016-17, Government of Andhra Pradesh). Let us create a data file in Excel.

Procedure: Here are the steps.

1) Data shall be entered in the cells using keyboard. After every entry in the cell the Enter Key should be pressed.

2) The data can be aligned properly by using the left, right and center alignment buttons appearing in the toolbars.

Table 9.1: Health Camps data (2016-17)

Sl. No.	District	Camps Conducted	Patient Screened	Out - patients	In - patients
1	Srikakulam	35	7682	8671	15279
2	Vizianagaram	35	9425	7277	14649
3	Visakhapatnam	35	8624	7030	21132
4	East Godavari	30	7030	29653	32117
5	West Godavari	35	5356	28062	22397
6	Krishna	0	0	18242	23753
7	Guntur	7	1739	31627	28937

8	Prakasam	56	12035	24022	19362
9	SPS Nellore	21	6658	15013	20608
10	Y.S.R.	17	3930	12372	16591
11	Kurnool	54	4265	6293	18483
12	Ananthapuramu	0	0	5211	15720
13	Chittoor	28	7528	10822	20358
	Total	353	74272	204295	269386

3) After entering the data, this file should be saved using the Save As option of the File menu. Let us save this file on the desktop as 'Health Camps Data'.

4) The File created in Excel will now look like the one shown in Figure 9.1.

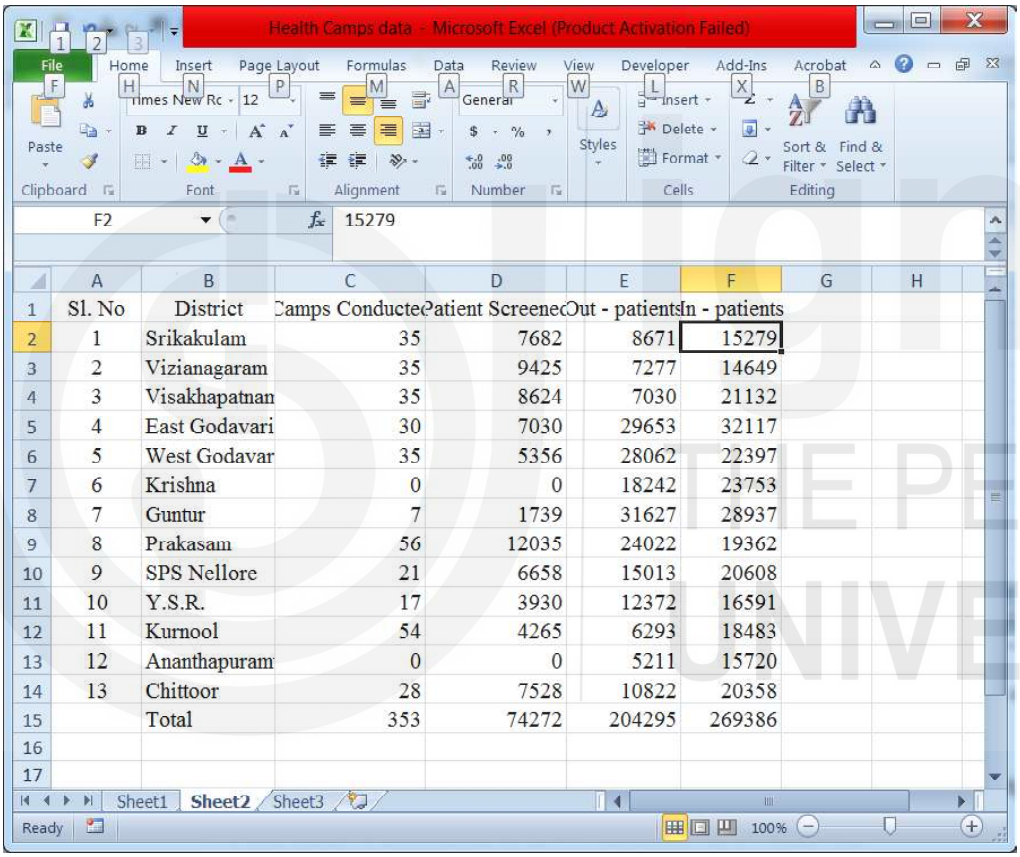


Fig. 9.1: Excel sheet for the Health Camps Data

We can create different data sets in different sheets of the same workbook or in different workbooks.

We can insert new work sheet by clicking on the button available to the right of sheet 3.

We can insert a maximum of 256 worksheets in a workbook.

9.2 EDITING FEATURES OF EXCEL SHEET

The Excel sheet has several features that make data entry very simple. Some of them are given below.

a) **Data selection**: A portion of data can be selected with mouse so that any changes in the selected area can be performed (like changing the letter size or font).

b) **Column width**: The width of a column is usually set by default as 8.43 points and it is called Standard. When the contents of the column exceed 8 or 9 characters, the rest goes hidden and not visible. We can select all the required columns with mouse and double click on the line between any two columns in the selected area.

c) **Freeze Panes**: When the data has more than 22 rows the headings disappear when we start scrolling down. Similar situation arises when the column heading is very wide. This gives trouble in editing data. To keep the first row (headings row) and the first column always visible, use Freeze Panes option from the main menu. As a rule, put mouse on the cell B2 and click freeze panes. This will create a dark like at B2 and the first row and first column will remain visible. If not required, use View → Unfreeze panes

d) **Sort/Filter**: This option helps in sorting the data in a column. When you can sort in the increasing or decreasing order all the other data elements in the rows get sorted automatically. The filter option helps to select records satisfying a particular condition like Gender = 'M' or Age < 35 years. The sort/filter button appears in the main menu.

e) **Cut, Copy and Paste**: A group of cells or a column or a complete row can be copied to another location in the sheet or to a different sheet using Cut, Copy and Paste operations. We can use short cuts Ctrl+C for **Copy** and Ctrl+V for **Paste**.

f) **Paste Special:** This is an important editing feature. We can copy a portion of data and paste it in another location without disturbing the back-end formulas. This is done by choosing 'values' option.

g) Exporting data to Word: The data or results obtained in Excel can be copied and pasted in a Word document or Power Point Presentation. Similarly, a data created the Word table, can also be copied and pasted in Excel.

It is also possible to create new entries like totals, sub totals, percentages etc., by writing a formula. We need not write a code; instead we can click on the cells that are involved in the formula. An interesting feature of Excel is that after finding the results (like totals), if we make any changes in the original data, the results will be automatically get updated.

Check Your Progress

1) Write about the structure of an Excel Workbook? How many sheets are found in a book?

.....

.....

.....

.....

2) Mention any four important editing features of an Excel Sheet.

.....

.....

.....

.....

.....

3) Write briefly about 'freeze panes' options in Excel.

.....

.....

.....

.....

.....

9.3 CREATING GRAPHS IN EXCEL

A variety of statistical graphs can be created with Excel. Commonly used charts include bar chart, column chart, pie chart and line chart. All these charts are available in some variants (style changes) and plotted with Excel by choosing from the options of insert menu.

Bar chart and **column chart** are often used to depict summary values like total, average or count in the case of categorical variables. Here is an example.

Age Group	Obese persons
< 30 yrs	21
31-40	42
>= 41 yrs	15

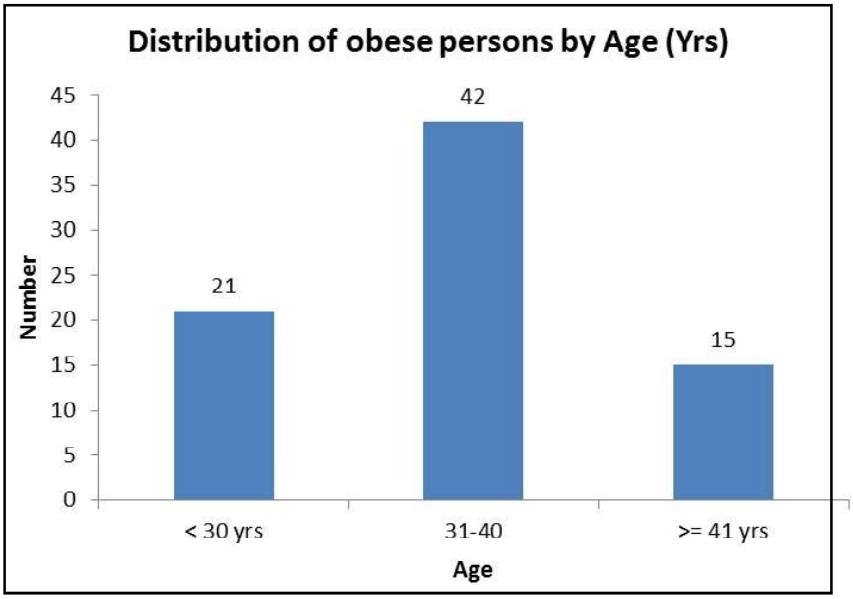


Fig. 9.2: Bar Chart

A **Pie chart** is used to display various components (like percentages) of a characteristic. The monthly percentage of blood donors according to blood group is better displayed by a pie chart than a bar chart. You may note that the percentages add to 100 as shown below.

Blood Group	Donors
O	56
A	12
B	10
AB	6

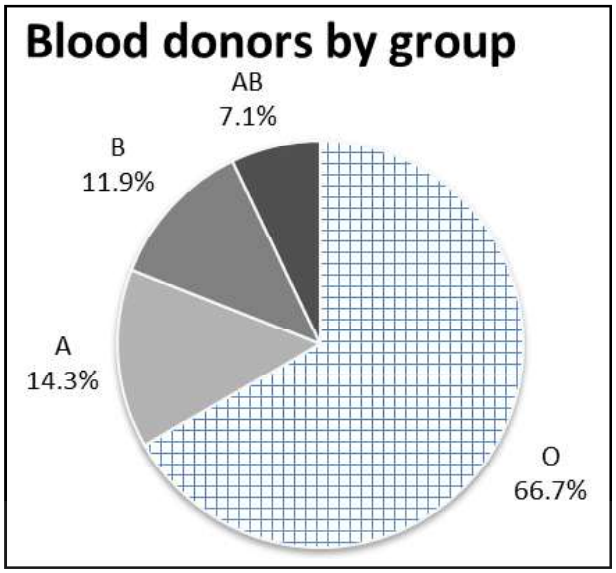


Fig. 9.3: Pie Chart

A **Line chart** is used to explain the trend (changing patterns over time) of a study variable. For instance, the growth in the number of minor surgeries performed in an area hospital is best displayed by a line chart as shown below.

Month	# of surgeries
Jan	21
Feb	29
Mar	27
Apr	28

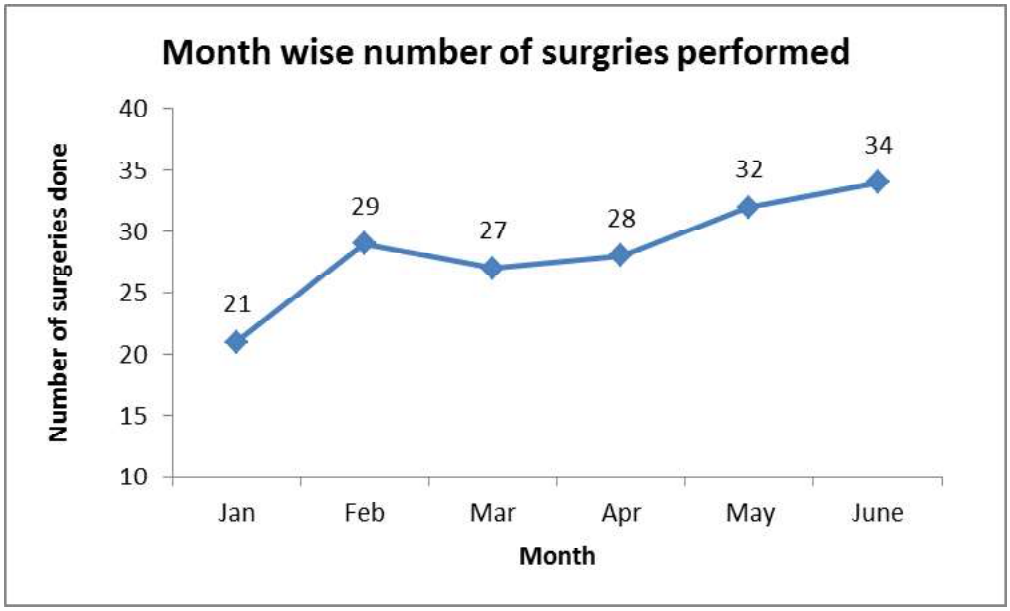


Fig. 9.4: Line Chart

Excel also helps in understanding relationship between two variables like birth weight and foot length of newborn babies. This is called scatter diagram as the one showing in the Illustration 10.3 and Figure 10.1 of Unit 10.

9.4 SIMPLE STATISTICAL ANALYSIS WITH EXCEL

Excel has a built-in statistical package to perform data analysis. This is called Analysis ToolPak given in the Data menu. It is a member of the Tools menu but usually not displayed in the menu. However, it can be activated through the Add-ins available in Excel options. Here is a partial list of features available in this tool for basic statistical analysis.

- ANOVA: Single factor
- ANOVA: Two-Factor with Replication
- Correlation
- Descriptive Statistics
- Histogram
- Random number generation
- Moving Average·Regression
- Sampling
- t-test
- Z-test

Each of these tools however needs some statistical knowledge of its application. It is also possible to use several mathematical, financial, logical and other functions.

Here are two illustrations to learn the utilities of Excel in performing simple statistical tests.

Illustration 9.2 (Two sample t-test): The following data is available on the Body Mass Index (BMI) observed in two independent groups A and B.

Group-A	20.43	22.51	18.99	20.49	23.12	25.63	18.08	20.63	22.55	22.43	22.77	23.23
Group-B	17.7	21.4	20.7	19.3	21	17.9	18.6	18.5	18.2	20.3		

We wish to test whether the difference in the mean BMI between the two groups is significant.

Solution: If we use the formula given in unit 8, we proceed as follows:

For the group A we have $n_1 = 12$, $\overline{X}_1 = 21.734$ and $s_1 = 2.078$. For the group B $n_2 = 10$, $\overline{X}_2 = 19.36$ and $s_2 = 1.378$. The combined SD will be $s = 1.78$.

The null hypothesis is H_0: Difference = 0 and H_1: Difference not zero (two-sided hypothesis). Now from the sample data we get

a) Difference = 21.734 -19.36 = 2.378

b) Standard Error = $1.797\sqrt{\left\{\frac{1}{12}+\frac{1}{10}\right\}}$ 1.797*0.4282 = 0.7694

c) Test value (t) = 2.378/0.7694 = 3.0907

The degrees of freedom are (12+10-2) = 20. Taking á = 0.05 the critical value from tables for 20 degrees of freedom is 2.08. Since t > 2.08 we cannot accept the null hypothesis and hence the difference is *significant at 5% level*.

Working with Excel: Alternatively, we can use the Analysis ToolPak of Excel in the Data menu. We find the option *t-test:* Two-Sample Assuming Equal Variances. The data of the two groups can be entered as two separate columns (or rows) with proper headings. The option window and the output are shown in figure 9.5 and the Excel output appears as shown in table 9.2).

Table 9.2: t-test output

t-Test: Two-Sample Assuming Equal Variances		
	Group-A	*Group-B*
Mean	21.738	19.360
Variance	4.319	1.898
Observations	12	10
Pooled Variance	3.229	
Hypothesized Mean Difference	0	
Df	20	
t Stat	**3.090**	
P(T<=t) one-tail	0.002	
t Critical one-tail	1.724	
P(T<=t) two-tail	**0.005**	
t Critical two-tail	2.085	

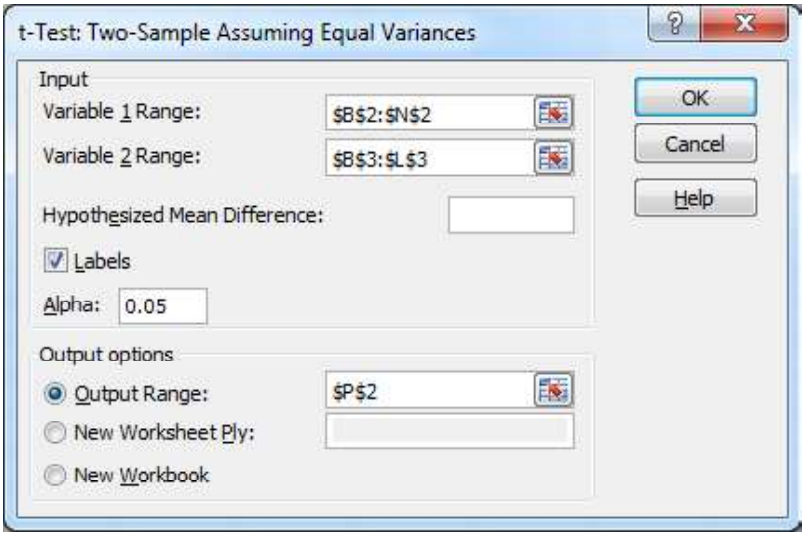


Figure 9.5: t-test options

The output shows the test value as 3.090 (t Stat) and this is compared with the critical value of 2.085 for the two tailed test. Since the test value is higher than the critical value, we reject the null hypothesis and conclude that the difference in the means is significant.

Instead, we can use the p-value of the test, shown in Figure 9.5 and table 9.2 as P (T<=t) two-tailed which is 0.005. Since this value is far less than 0.05 (5% level of significance) we reject null hypothesis and consider the difference as significant.

Here is another illustration on a different type of t-test called paired t-test.

Illustration-9.3 (Paired t-test): The following data refers to the Fasting Blood Sugar (FBS) before treatment and after 30 days for 15 persons. Test whether there is a significant decrease in the FBS levels.

Patient No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Before	176	142	148	130	196	177	162	194	152	156	140	173	192	148	180
After	154	122	127	110	171	157	141	164	136	133	128	148	157	130	166

We wish to test whether the change in the mean FBS is statistically significant.

Solution: Let us use the Analysis ToolPak of Excel and select the *t-test: Paired Two-Sample for Means*. The output shows the following results.

t-test: Paired Two Sample for Means		
	Before	*After*
Mean	164.4	142.9333
Variance	449.6857	330.3524
Observations	15	15
Pearson Correlation	0.9676	
Hypothesized Mean Difference	0	
Df	14	
t Stat	14.2325	
P(T<=t) one-tail	5.09E-10	
t Critical one-tail	1.7613	
P(T<=t) two-tail	1.02E-09	
t Critical two-tail	2.1448	

You will find that the p-value of two-tailed test is given as1.02E-09 in scientific notation. It is equal to 0.00000000102 which is very less than 0.05 and hence the change in FBS is statistically significant. When the p-value is so small like this, we simply report as $p < 0.001$ instead of displaying all the digits.

Another important utility of excel is a tool called Analysis of Variance (ANOVA) used for comparing the means of more than two independent groups.

Here is an illustration.

Illustration 9.4 (ANOVA): Suppose we wish to compare the B12 level (pictograms per milliliter or pg/ml) of anemic persons measured by three independent teams A, B and C. The following table shows an illustrative data.

A	101	92	97	102	115	98	125	101		
B	110	108	180	125	132	147				
C	313	181	252	173	345	197	241	223	250	257

We wish to compare the mean values among the three groups and test for statistical significance at 5% level of significance. The technique is called one-factor ANOVA or single-factor ANOVA because the total data with 24 values is divided or grouped according to a single factor called *team* with three levels A, B and C.

Procedure: In Excel, the data has to be entered in three separate columns or rows with suitable headings. Then open the Analysis ToolPak and select the tool *Anova Single Factor* and press OK. The option window shows the data links and the output appears as shown below.

ANOVA: Single Factor						
SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
A	8	831	103.875	116.125		
B	6	802	133.6667	724.2667		
C	10	2432	243.2	2988.178		
ANOVA						
Source of Variation	*SS*	*Df*	*MS*	*F*	*p-value*	*F crit*
Between Groups	96473.82	2	48236.91	32.33469	3.88E-07	3.4668
Within Groups	31327.81	21	1491.8			
Total	127801.6	23				

The above table is called ANOVA table and always appears in the format as shown above. The comparison between the means is done by a comparison of B12 variation between groups and within groups. The null hypothesis is that all the group means are equal. The F-value shown against 'Between Groups', denotes the test value. The corresponding p-value here is 0.00000039 which is far less than 0.05. Hence the mean B12 values differ significantly among the three groups.

We end the discussion on Excel with the note that learning by doing is the best way of gaining working knowledge in Excel. More details on the use of Excel for data analysis can be found from Sarma (2010) given in references.

9.5 CREATING DATA FILES IN SPSS

SPSS (Statistical Package for Social Sciences) is a software meant for performing statistical calculations and advanced analyses. Originally released 50 years ago by SPSS Inc., this software has undergone several improvements and found

popular among health researchers apart from survey scientists, marketing managers and social science researchers.

The present version is known as IBM SPSS after its takeover by the IBM Corporation in 2009 and the current versions are labeled as IBM SPSS and version 20 is used in this discussion.

Here are some features of SPSS for general statistical applications.

- We can open Excel data in SPSS.
- Several variables can be handled at a time during analysis. For instance, simple frequency tables or descriptive statistics like means and standard deviations for many variables can be computed in a single step.
- Two dimensional and multi-dimensional tables can be generated along with Chi-Square tests.
- Statistical tests like t-test, ANOVA can be performed for several variables at a time by simply selecting the variables and groups (to be compared).
- Advanced statistical features like multivariate analysis, repeated measures analysis, step wise linear regression can be handled easily.
- Continuous variables like body mass index can be coded into categories; existing categories can be re-categorized with simple menu options.
- We can create new variables by writing simple formula (not code) by mouse clicks from the option windows.
- In SPSS data setup, each row is called a *case*. Subset of data can be analyzed by selecting a portion of records (rows) either randomly or by a known rule (like Gender = "F" and Age < 40 yrs)
- Charts with complex conditions (like group wise histogram) are available for 2D and 3D visualization.

Above all there is no need to write any code to run a program. In fact, all operations are based on mouse clicks but still the back-end code can be saved as *syntax*.

An important utility of SPSS is the output. The output for most applications appears as formatted tables so that we can copy and paste the same in MS-Word or MS-Excel without retyping the contents of the output.

When there are several objects (like tables or charts) in the output, the complete output can be exported to other format like PDF or Word by using the 'Export' option in the File menu.

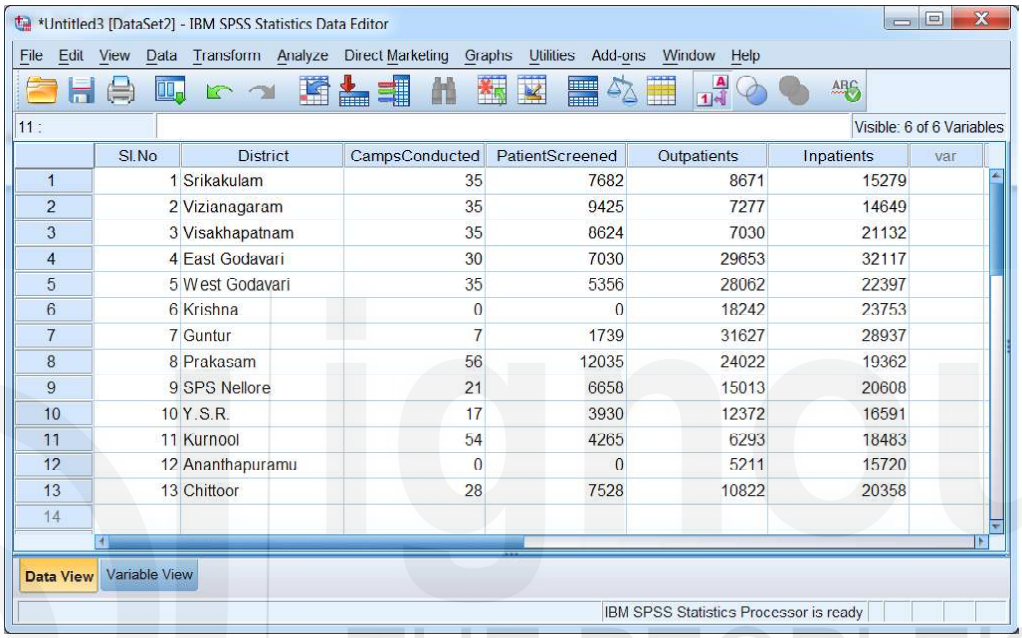
Let us recall the Health Camp data (Table 9.1) and open the Excel file in SPSS. The following are the steps:

1) Open SPSS and click the following options
2) File®Open®Choose type as Excel
3) Select the file from its location (here it is desktop). All the sheets will be available to view
4) Click on the sheet where the data is located
5) Press OK.

The excel data in SPSS window as shown in Figure 9.6.

Let us save the file on the desktop as *Health Camps* Data. The file will be saved with extension '.sav' while the same file in Excel had the extension '.xls'.

Each data in has two views viz., Data View and Variable View as shown in figure 9.6. In the 'Variable View' we find the names of the variables and their properties. It is important to note that SPSS asks to specify whether a variable is numeric or string. We should also specify the type of data (nominal, ordinal or scale). Some statistical tests depend on the type of data.



	Sl.No	District	CampsConducted	PatientScreened	Outpatients	Inpatients	var
1	1	Srikakulam	35	7682	8671	15279	
2	2	Vizianagaram	35	9425	7277	14649	
3	3	Visakhapatnam	35	8624	7030	21132	
4	4	East Godavari	30	7030	29653	32117	
5	5	West Godavari	35	5356	28062	22397	
6	6	Krishna	0	0	18242	23753	
7	7	Guntur	7	1739	31627	28937	
8	8	Prakasam	56	12035	24022	19362	
9	9	SPS Nellore	21	6658	15013	20608	
10	10	Y.S.R.	17	3930	12372	16591	
11	11	Kurnool	54	4265	6293	18483	
12	12	Ananthapuramu	0	0	5211	15720	
13	13	Chittoor	28	7528	10822	20358	
14							

Fig. 9.6: Excel Data Opened in SPSS

Let us find the summary statistics for Health Camps Data. To do this, choose the options Analyze®Descriptive Statistics®Frequencies. In the resulting menu we can select the variables for which descriptive statistics are required. We can optionally select *summary values* like mean, standard deviation, minimum, maximum etc. as output.

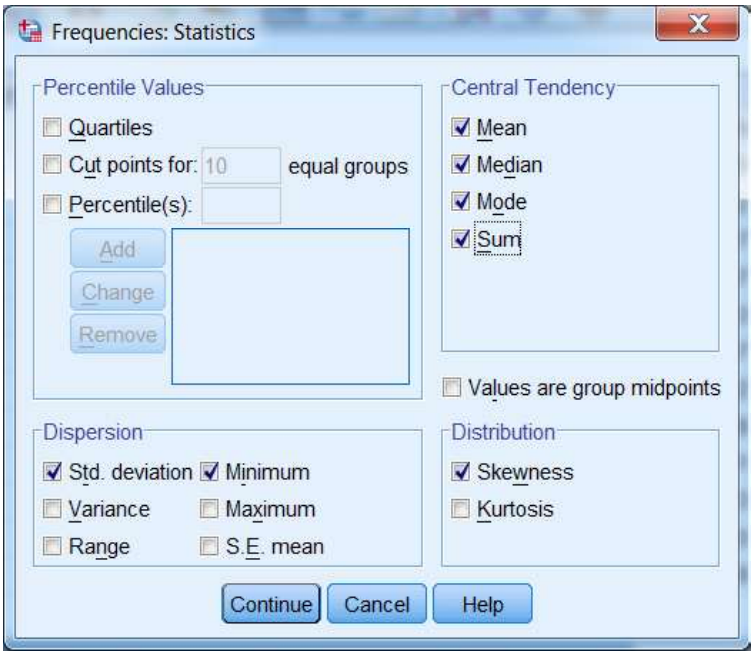


Fig. 9.7: Statistics Options

Table 9.3: SPSS output

Statistics				
		Patient Screened	**Out - patients**	**In - patients**
N	Valid	13	13	13
	Missing	0	0	0
Mean		5713.23	15715.00	20722.00
Median		6658.00	12372.00	20358.00
Mode		0	5211[a]	14649[a]
Std. Deviation		3634.234	9601.497	5204.917
Skewness		-.204	.607	1.028
Std. Error of Skewness		.616	.616	.616
Minimum		0	52111	4649
Sum		74272	204295	269386
a. Multiple modes exist. The smallest value is shown				

Figure 9.7 shows the options window. If we select three variables viz., patients screened, out-patients and in-patients the output will be like the one shown in table 9.3.

When a table is pasted into Word from SPSS, it can be formatted suitably with the options available in Word Tables. SPSS generates the frequency tables by counting the number of times an individual value has occurred. We do not directly get a table with class intervals. This is very useful to identify abnormal value in the data!

9.6 CROSS TABULATION AND CHI SQUARE TESTS

We can generate two dimensional tables from raw data using SPSS. For instance, we may like to count the number of cases of DM (yes/no) among male and female patients. For small data sets we can count with hand calculations but with thousands of records we need software and SPSS has a module called crosstabs to perform this operation. In addition to creating table of counts, we have options to perform a test of hypothesis between categorical characteristics.

Here is an Illustration.

Illustration 9.5: The following is a portion of illustrative data created in an Excel file containing the details of 30 patients in a health care study.

Table 9.4: Illustrative Data with 30 Cases and 12 Variables

Hosp_ID	AGE	SEX	BMI	DM	HTN	Toba-cco	Smo-king	Alcohol	SBP	DBP
1001	60	1	20.76	0	0	0	0	0	100	70
1002	72	2	21.4	0	0	0	0	0	100	60

1003	66	2	23.5	1	1	0	0	0	150	100
1004	46	1	21.1	1	1	1	1	0	120	80
1005	51	1	30.61	0	0	1	1	1	120	80
1006	60	1	21.78	0	0	0	1	1	130	80
1007	33	1	24.11	0	0	1	1	1	100	70
1008	63	1	20.2	0	1	1	1	0	110	70
1009	53	1	26.77	1	1	1	1	1	130	80
1010	42	1	19.15	0	0	1	1	1	110	80
1011	69	1	19.49	0	‘0	1	1	0	120	80
1012	47	1	25.28	0	1	1	1	1	170	100
1013	47	1	26.56	1	0	1	1	1	140	80
1014	67	2	24.8	1	0	0	0	0	140	90
1015	67	1	21.8	1	0	1	1	0	140	90
1016	60	1	22.86	1	0	0	0	0	160	90
1017	58	2	23.56	1	0	0	0	0	140	90
1018	67	2	20.81	1	1	0	0	0	140	90
1019	59	1	24.03	1	1	0	0	0	120	80
1020	45	2	25.68	0	1	0	0	0	120	80
1021	49	1	24.22	0	1	0	0	0	150	90
1022	65	1	18.73	1	1	1	1	1	160	90
1023	40	2	28.54	1	1	0	0	0	120	80
1024	41	1	26.22	0	1	0	0	0	150	70
1025	55	2	19.63	1	0	0	0	0	130	80
1026	68	2	25.97	1	1	0	0	0	140	90
1027	45	1	22.86	1	0	1	1	1	110	80
1028	60	1	20.76	0	0	0	0	0	110	70
1029	60	1	28.89	0	1	1	1	1	160	90
1030	60	1	18.31	0	1	1	1	0	140	80

The following are the codes used.

1) Sex (1 = Male, 2 = Female)
2) DM = Diabetes (1 = Yes, 0 = No)
3) HTN = Hypertension (1 = Yes, 0 = No)
4) Tobacco = Tobacco Chewing (1 = Yes, 0 = No)
5) Smoking (1 = Yes, 0 = No)
6) Alcohol (1 = Yes, 0 = No)

This file can be opened in SPSS and the codes are assigned as described above. The data file is shown in figure 9.8.

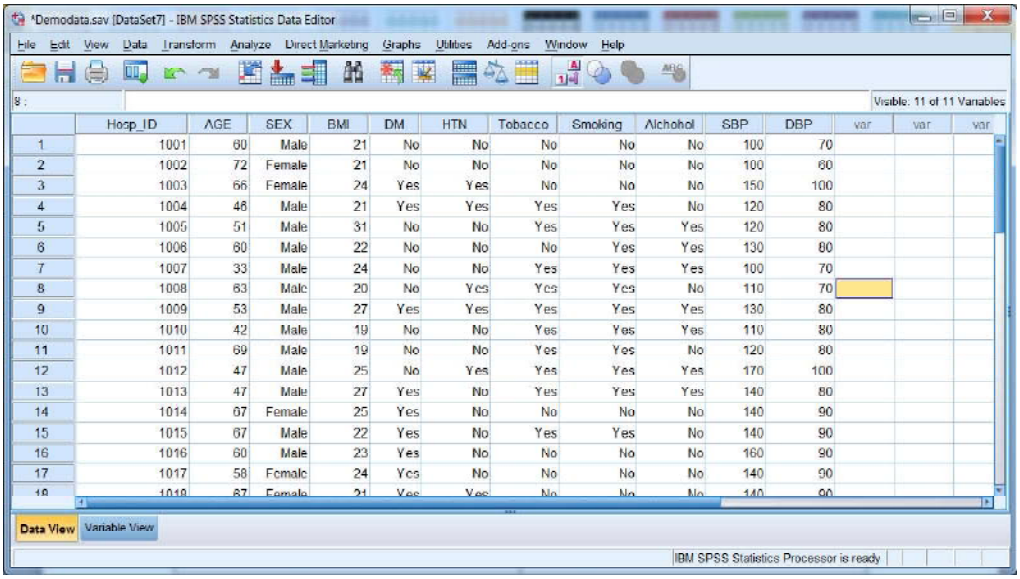


Fig. 9.8: Data File with Labels for the Codes

For variables like Sex, DM or HTN we have to assigning labels for the numerical codes. This can be done in the 'variable view' of the data file. There is a column titled 'values' and we have to click in this column corresponding to the variable say (DM). The labels are assigned as shown in Figure 9.9.

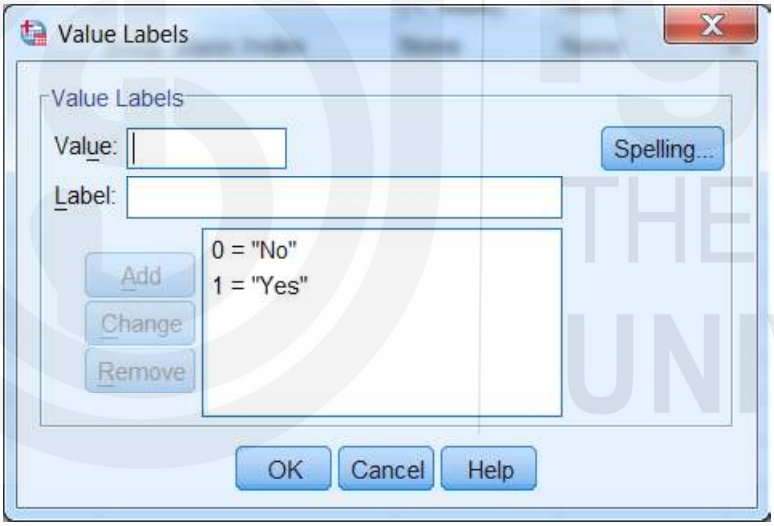


Fig. 9.9: Options for Assigning Labels for Numeric Codes

In the data view of the file, if we click on the View®Value Labels, the actual labels appear instead of numeric codes. Now let us answer the following queries. Some related details on coding and selection of cases can be Sarma (2010) given in references.

Check Your Progress

4) What are the statistical tools available in Analysis ToolPak of Excel? Write about t-test options.

.....

.....

.....

5) What is meant by cross tabulation in SPSS? Give an example.

.....

.....

.....

.....

.....

Query 9.1: Count of patients who smoke and also have DM. This is done by choosing Analyze®Descriptive Statistics®Crosstabs. In the resulting window send DM to rows and SMOKING to columns. Press OK. This gives the following two-way table of counts (number of patients in each category). You will find options to report percentage (click on Cells button).

Diabetes * Smoking Cross tabulation			
Count			
	Smoking		
Diabetes	**No**	**Yes**	**Total**
No	6	9	15
Yes	9	6	15
Total	15	15	30

Query 9.2: Is there any relationship between Tobacco chewing and alcohol consumption?

The relationship between tobacco and alcohol shall be expressed in terms of association because the two variables are qualitative (categorical). We can test the null hypothesis that tobacco and alcohol are independent. If this hypothesis is rejected, there will be reason to believe that they are not independent which means they are likely to be associated. Now the Chi Square test is an option in the crosstabs. If you click on 'statistics' button in crosstabs window, you will check for 'Chi Square'. This gives the following table and the Fisher's Exact Test shows p-value = 0.001 (marked as Exact Sig. (2-sided)). Since the p–value is smaller than 0.05 we may reject the null hypothesis of independence and conclude that tobacco and alcohol consumptions are associated with each.

Tobacco Chewing	**Alcohol**		**Total**
	No	**Yes**	
No	15	1	16
Yes	5	9	14
Total	20	10	30

Query 9.3: What are the mean and standard deviation of SBP and DBP among of male and female patients?

Since SBP and DBP are continuous (not categorical) we can find the mean and standard deviation. The options are Analyze®Compare Means®Means. Within

the options window select SBP and DBP into the 'dependent list'; sex into 'independent list' box. Leave the options as they are and press OK. This gives the following results.

SEX		Systolic Blood Pressure	Diastolic Blood Pressure
Male	Mean	130.95	80.95
	N	21	21
	Std. Deviation	21.425	8.309
Female	Mean	131.11	84.44
	N	9	9
	Std. Deviation	15.366	11.304
Total	Mean	131.00	82.00
	N	30	30
	Std. Deviation	19.538	9.248

The mean and standard deviation are given gender wise and for the overall data. The 'Total' indicates the 'overall' sample (irrespective of sex). You should not add the means to get this value.

Here is another illustration.

Illustration 9.6: Reconsider the data used in Illustration 9.3. Let us divide age into three groups say: i) de 50 years, ii) 51-60 years and iii) 61 and above. Then we can compare the mean of SBP or DBP age wise.

Approach: To categorize age, let us use the menu options Transform® Recode into Different Variables. In the options window choose Age and send it to the box shown in the middle.

Then assign a name to the new variable as 'Age code'. Then click 'change' and then click on 'old and new values'. To each category, assign a number like 1, 2, 3 and press 'Add' each time. These options are shown in Figure 9.8.

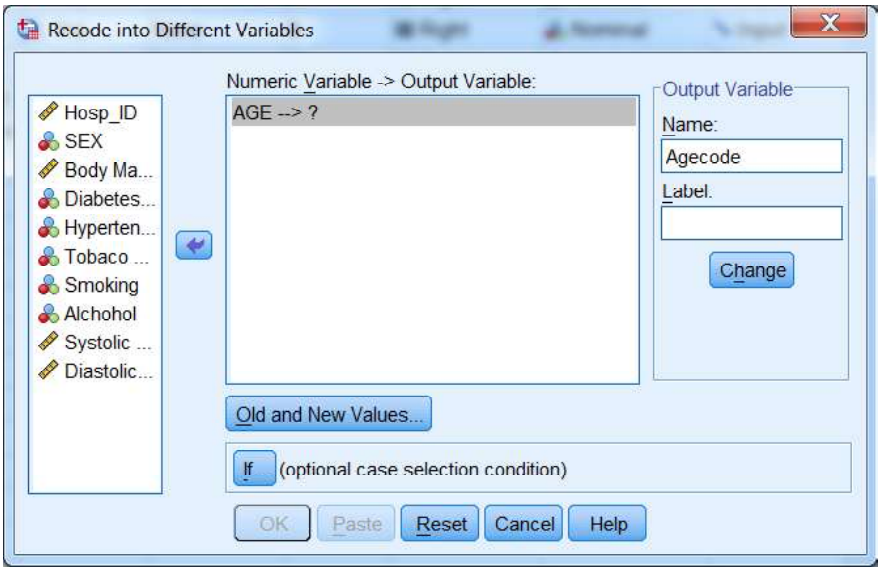


Figure 9.10 (a): Selection of Variables for Recoding

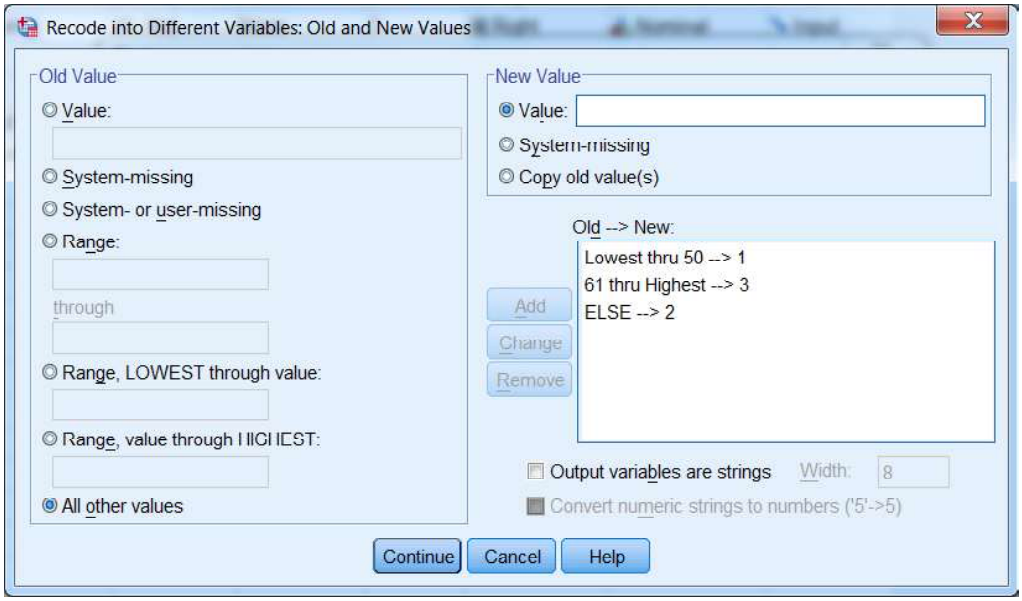


Fig. 9.10 (b): Conversion of Old Values into New Codes

Once done, you may click on OK button. This creates a new variable 'Age code' and a new column with the corresponding codes is generated. Then assign labels to the codes as (1 = 'd" 50 years', 2 = '51-60 years', 3 = '61 and above'.

This completes recoding process. We can find the mean values of SBP and DBP by using the options given in the answer for Q3. The group wise means are shown in Table 9.5.

Table 9.5: Descriptive Statistics of SBP and DBP by Age Group

Age	Systolic Blood Pressure			Diastolic Blood Pressure		
	Mean	N	Std. Deviation	Mean	N	Std. Deviation
<= 50 Years	129.00	10	22.336	81.00	10	8.756
51-60 Years	130.91	11	18.684	80.91	11	7.006
61 & Above	133.33	9	19.365	84.44	9	12.360
Total	131.00	30	19.538	82.00	30	9.248

Since we have attached labels for the age groups, we write the heading as Age and not Age code. The original table produced by SPSS is altered in the output itself, by using 'pivot options' (try by double click on the table in SPSS output).

We end the discussion of basics of Excel and SPSS with the observation that Excel serves something like entry level software to handle statistical data while SPSS helps in a broader way.

9.7 SUMMARY

- We have seen that Excel is popular tool among researchers to handle various types of research data. Data sets of moderate to large size can be created using Excel worksheets. There are several editing and computing facilities to effectively manage data as well as to make calculations. Statistical

diagrams like bar charts, pie charts and line charts can be effectively created using Excel.

- We can also perform certain statistical analyses like finding average, standard deviation, correlation and performing simple tests of hypothesis.
- We have also learnt that SPSS is exclusively meant for handling basic as well as advanced statistical analyses. We can open Excel data files in SPSS and carry out analysis in addition to data manipulation (like recoding, attaching labels etc.). In essence both Excel and SPSS are user-friendly to public health researchers.

9.8 REFERENCES

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

Socio Economic Survey (2016-17). Government of Andhra Pradesh.

9.9 ANSWERS TO CHECK YOUR PROGRESS

1) In Excel a data file is called a workbook. It normally contains three sheets. Every sheet contains a 16384 columns numbered as A, B,.. and 1048576 rows numbered as 1,2,3. For details refer section 9.1.

2) Four important editing features of an Excel sheet are: a) Data selection, b) Sort/Filter, c) Cut, Copy and Paste, d) Exporting data to Word. For details refer section 9.2.

3) In Excel, to keep the first row (headings row) and the first column always visible Freeze Panes option is used from the main menu. For details refer section 9.2.

4) A number of statistical tools are available in Analysis ToolPak of Excel. For e.g. ANOVA, Correlation, Descriptive Statistics, Histogram, Random Number Generation, Sampling, t-test, Z-test etc. For details refer section 9.4.

5) We can generate two dimensional tables from raw data using SPSS. For instance, we may like to count the number of cases of DM (yes/no) among male and female patients. For small data sets we can count with hand calculations but with thousands of records we need software and SPSS has a module called crosstabs to perform this operation. For details refer section 9.6.

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'

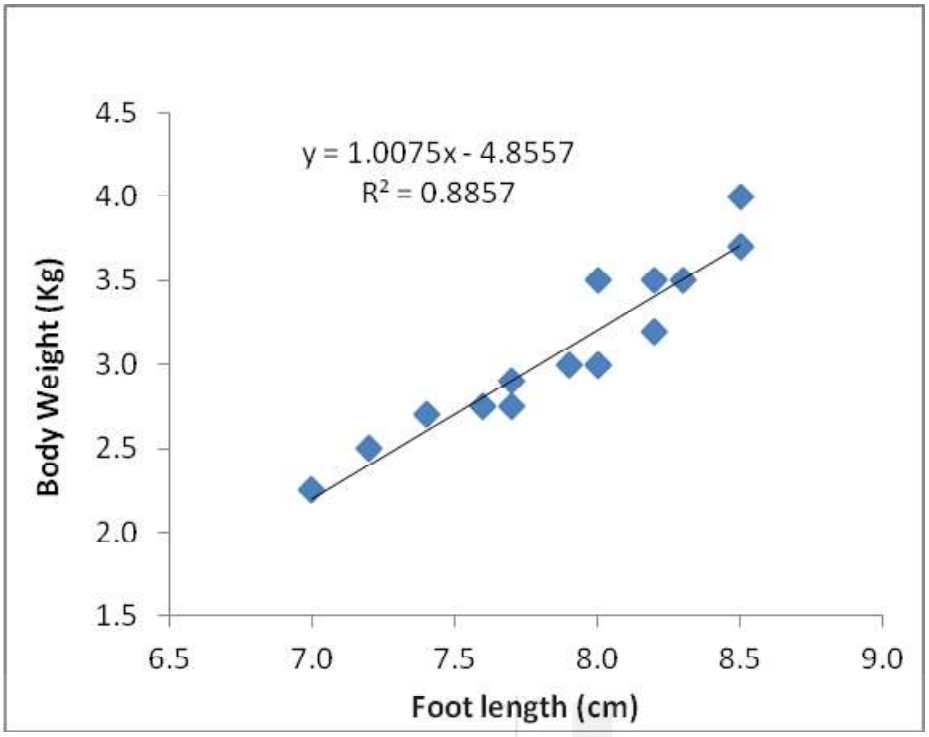


Fig. 10.1: Scatter Diagram and Linear Regression Equation

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.