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.

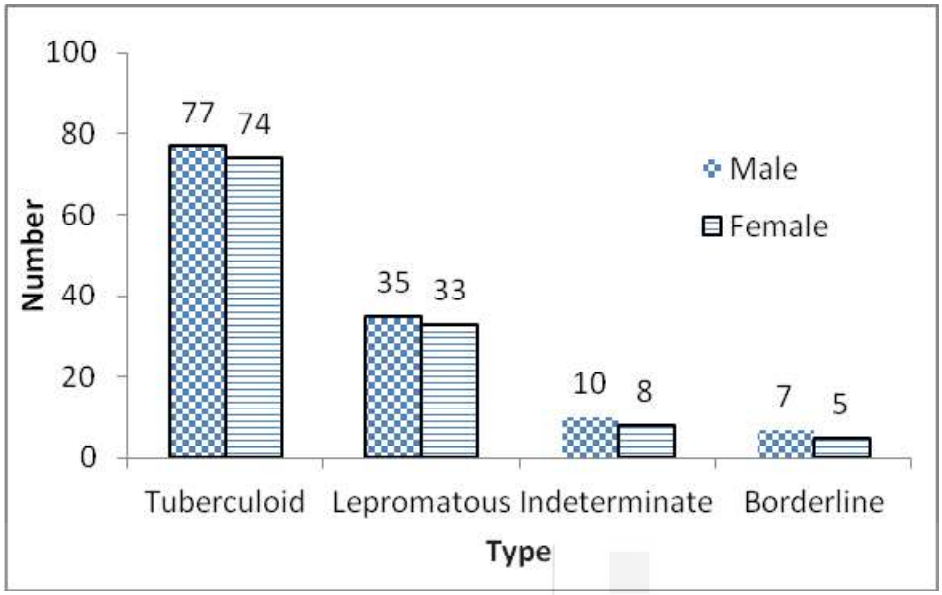

Fig. 8.1: Distribution of Patient by Type of Leprosy

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.

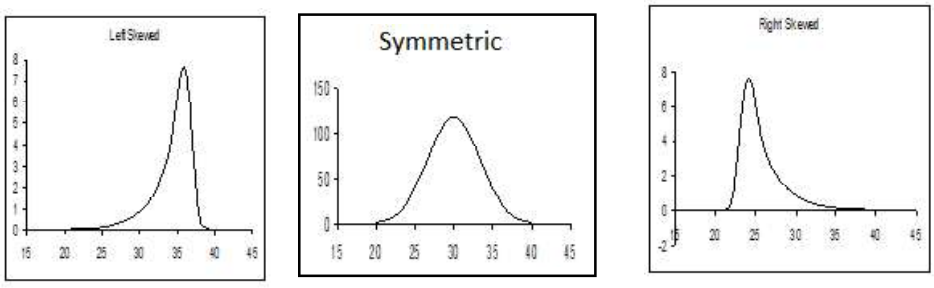

Fig. 8.2: Pattern (shape) symmetric and assymetric distributions

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.