UNIT 12 SAMPLING PROCEDURE*

Structure

* Prof. C G Naidu (retd.), School of Vocational Studies, IGNOU

12.0 OBJECTIVES

On completion of this Unit, you should be able to

- explain different methods of drawing a sample;
- describe different types of samples;
- use random number tables to draw a sample; and
- determine the sample size.

12.1 INTRODUCTION

In Unit 2 of this course, you have learned about different types of data, namely, primary data and secondary data. In that Unit, we also discussed the use of different survey techniques like face-to-face interview, telephone survey, postal survey, internet survey, etc. in collecting primary data. In Unit 15, you also have learned the meaning of sampling, advantages of sampling, and sampling error.

In statistics, we often rely on a sample (that is, a small subset of a larger set of data) to draw inferences about the population (that is, the larger set of data). For example, you are interested to know the voting behaviour of Delhi people in the next election. Who will you ask? Naturally, it is not possible for you to ask every single Delhi voter how he or she is likely to vote. Instead, you may querry a relatively small number of Delhi voters and draw inferences about entire Delhi from their responses. In this case total voters of Delhi constitute the population and the voters actually querried constitute your sample.

Ideally, the characteristics of a sample should reflect the characteristics of the population from which it is drawn. In such cases, the inferences drawn from a sample are probably applicable to the entire population.

In this unit you will learn how to draw a sample under different population characteristics and how to determine the sample size.

12.2 SAMPLING PROCESS

In conducting a sample survey, the sampling process determines which sampling units will be included in the survey. Sampling makes data process more manageable and affordable. It enables the population characteristics to be inferred with minimal errors on the basis of the sample. The sampling process includes defining the target population from which we draw the sample, identification of the sampling frame, selection of the sampling method, and selection of the sampling units.

1) Survey Objectives: A sample survey begins with the specification of the objectives. We should have a clear and un-ambiguous idea of the objectives

of the survey, because all other steps – target population, sampling frame, sampling procedure, etc. – are designed according to survey objectives.

2) Questionnaire Design: Keeping the objectives of the survey in view we are required to design a questionnaire. We have already learnt the major steps involved in designing of a questionnaire in Unit 1 of this course. In addition to the questionnaire we need to develop training documents for the investigators, particularly when the sample survey is conducted at a larger scale involving a number of investigators.

3) Defining the Target Population: To draw the samples from a population we must know the target population about which conclusions are to be drawn. The target population is also referred to as the universe. The target population is the group about which we wish to generalize or make inferences from the sample. For example, you want to conduct a sample survey on the family planning methods used by eligible couples in Delhi. Here, all those couples in Delhi in the reproductive age group form the target population.

4) Identifying Sampling Frame: The sampling frame is a list of cases from the target population. The sampling frame is the actual operational definition of the target population. In our earlier example of eligible couples in Delhi using family planning methods, all the people in the reproductive age group form the sampling frame. Many times we may not be able to list all the cases in the target population for some reason or other. For example, we want to list the people for a survey based on telephone directory. In this case, certainly those people who do not have telephone numbers in the directory will be excluded from the listing. This type of error is called sampling frame error.

5) Selecting Sampling Procedure: Once the sampling frame is identified, we select appropriate sampling procedure to select the sample for the survey. We will discuss various sampling procedures in detail in the next section of this Unit.

6) Selecting the Sampling Units: Sampling units are those cases from the sampling frame which are included in the sample by using appropriate sampling procedure. Essentially, a sampling unit is the case on which data is collected. For example, you may decide to take 1000 sampling units from the sampling frame (consisting of all the reproductive age group people in Delhi) for your sample survey.

7) Survey data Processing: After selection of sampling units the next step is data collection and processing. We need to check the incomplete questionnaires and edit or cross-check the responses wherever there is a doubt. Data entry and tabulation follows.

8) Analysis of Data: The next step in the sequence is analysis of data. Keeping in view our requirements we analyse the data by using various statistical tools.

9) Publication and Dissemination of Results: On the basis of data analysis we prepare technical and research reports. Finally, the socio-economic results of the survey and their implications are discussed in seminars.

12.3 TYPES OF SAMPLING

The method of selecting a sample from a given population is called *sampling*. Basically there are two types of sampling, viz., probability sampling and non-probability sampling. In probability sampling the sampling units are selected according to some chance mechanism or probability of selection. On the other hand, non-probability sampling is based on judgement or discretion of the person making a choice. Thus in non-probability sampling certain units may be selected because of convenience or they serve a purpose or the researcher feels that these units are representative of the population. No random selection on the basis of chance mechanism is involved here.

12.3.1 Probability Sampling

It is also called random sampling. It is a procedure in which every member of the population has a chance or probability of being selected in the sample. It is in this probabilistic sense that the sample is random. The word 'random' does not mean that the sample is obtained in a haphazard manner without following any rule.

Random sampling is based on the well-established principles of probability theory. There are quite a few variants of the random sampling, viz., simple random sampling, systematic random sampling and stratified random sampling. We discuss these types below.

a) Simple Random Sampling

If there is not much variation in the characteristics of the members of a population, we can follow the method of simple random sampling. In this method, we consider the population in its entirety as a homogeneous group and follow the principle of random sampling to choose the members for the sample.

There are two variants of simple random sampling, viz., simple random sampling with replacement (SRSWR) and simple random sampling without replacement (SRSWOR). This difference pertains to the way the sample units are selected. According to the procedure of simple random sampling with replacement (SRSWR), we draw one unit from the population, note down its features and put it back to the whole lot in the sense that the unit again becomes eligible for selection. In this way, the total number of units in the population always remains the same. In other words, the composition of the population remains unchanged, and each member of the population has the *same chance* or probability of being

selected in the sample. In fact, if N is the size of the population, this probability is $\frac{1}{N}$. On the other hand, in the case of simple random sampling without replacement, the unit once selected is not returned to the population in the sense that it becomes ineligible for selection again. As a result, after each draw, the composition of the population changes. Therefore, the probability of any particular unit being selected also changes.

b) Systematic Random Sampling

In this variant of random sampling, only the first unit of the sample is selected at random from the population. The subsequent units are then selected by following some definite rule. For example, suppose, we have to choose a sample of agricultural plots. In systematic random sampling, we begin with selecting one plot *at random* and then every j^{th} plot may be selected.

c) Stratified Random Sampling

Stratified random sampling is the appropriate method if the population under consideration consists of heterogeneous units. Here, first we divide the population into certain homogeneous groups or strata from each stratum. Secondly, some units are selected by simple random sampling. Thirdly, after selecting the units from each stratum, they are mixed together to obtain the final sample.

Let us consider an example. Suppose, we want to estimate the per capita income of Delhi by a sample survey. It is common knowledge that Delhi is characterised by rich localities, middle class localities and poor localities in terms of the income groups of the people living in these localities. Now, each of these different localities can constitute a stratum from which some people may be selected by adopting simple random sampling procedure.

d) Multi-Stage Random Sampling

Let us consider a situation where we want to obtain information from a sample of households in a large city, say, Delhi. Sometimes, it may not be possible to directly take a sample of households because a list of all the households may not be easily obtained. In such a situation, one may resort to take samples in various stages. Generally, the city is divided into certain geographical areas for administrative purposes. These areas may be termed as city blocks. So in the first stage, some of such blocks may be selected by random sampling. In the next stage, from each of the selected blocks in the first stage, some households may be selected again by the principle of random sampling. In this way, ultimately a sample of households from a large city may be obtained. The above-mentioned example is the case of a two-stage random sampling. However, if the nature of the inquiry so demands, the method of sampling can be extended to more than two stages.

12.3.2 Non-Probability Sampling

We have considered the method of random sampling and some of its variants above. It should be clear that the basic objective of the principle of random sampling is to eliminate or at least minimise the effect of the subjective bias of the investigator in the selection of the population sample. But for certain purposes, there is a need for using discretion. For example, suppose a teacher has to choose 4 participants from a class of 30 students in a debate competition. Here, the teacher may select the top 4 debaters on the basis of her own conscious judgement about the top debaters in the class. This is an example of purposive sampling. In this method, the purpose of the sample guides the choice of certain members or units of the population.

12.3.3 Mixed Sampling

In mixed sampling, we have some features of both non-probability sampling and random sampling. Suppose, an institute has to send 5 students for managerial training in a company during the summer vacation. Initially, it may shortlist about 20 students who are considered to be suitable for the training by applying its own discretion. Then from these 20 students, 5 students may finally be selected by random sampling.

We will discuss the process of drawing various types of samples later in this Unit.

12.4 SELECTION OF A SIMPLE RANDOM SAMPLE

Simple random sampling (SRS) is the basic sampling procedure where we select a sample from the population. In this procedure each unit is chosen entirely through chance mechanism and each unit of the population has an equal chance of being included in the sample. Every possible sample of a given size has the same chance of selection. That is, each unit of the population is equally likely to be chosen at any stage in the sampling process and selection of one unit should in no way influence the selection of another unit in the population.

Simple random sampling should be used with a homogeneous population. That is, all the units in a population should possess the same attributes that we are interested in measuring. The characteristics of homogeneity may include age, sex, income, social status, geographical region, etc.

There are two most commonly used methods to extract a simple random sample. The first is lottery method and the second is a random numbers selection method. Irrespective of which method we decide to use, every element in the sampling frame should be assigned an identifying number.

12.4.1 Lottery Method

The simplest method of selecting a random sample is drawing lottery. In this method, the unit identification numbers are placed in a container and mixed together. Finally, someone draws out numbers at random from the container until

the desired sample size is attained. Suppose, we want to select *n* sample units out of N population units. We assign the numbers 1 to N; one number to each of the population unit; and write these numbers on N slips. The slips are made as homogeneous as possible in shape, size, colour, etc. These slips are then put in a container and thoroughly shuffled. Finally, *n* slips are drawn one by one. The *n* units corresponding to numbers on slips drawn, will constitute a random sample.

Example 12.1

Let us assume that you are doing some research with a bank branch that wishes to assess customers views on the quality of service in the bank branch. You are asked to select 100 customers as the sample using simple random sampling procedure with lottery method. First, you have to get the sampling frame organized. For this, you will go through the bank records to identify the account holders. You then assign the serial numbers to all the account holders. Suppose there are 1000 account holders and you want to draw 100/1000=10% sample. You could print the serial numbers, tear them into separate strips, put the strips in a container, mix them up real good, close your eyes and pull out the first 100 strips.

The disadvantage of the lottery method is that it would be tedious and the quality of the sample would depend on how thoroughly you mixed them up and how randomly you picked them. Also, as the population size increases, it becomes more and more difficult to draw samples using lottery method.

12.4.2 Random Numbers Table Method

The random numbers are a collection of digits generated through a probability mechanism. The random numbers have the following properties:

The probability that each digit (0,1,2,3,4,5,6,7,8 or 9) will appear at any place is the same, that is 1/10.

The occurrence of any two digits in any two places is independent of each other.

In this method each unit in the population is assigned a unique number in a sequence. To draw the sample we use a table of random numbers. You can find the random number table (RNT), among other places, in Fisher and Yates (1963): Statistical *Tables for Biological, Agricultural and Medical Research*. An example of a random numbers table is shown in Table 12.1.

Table 12.1: Random Numbers Table

39634 62349 74088 65564 16379 19713 39153 69459 17986 24537
14595 35050 40469 27478 44526 67331 93365 54526 22356 93208
30734 71571 83722 79712 25775 65178 07763 82928 31131 30196

64628 89126 91254 24090 25752 03091 39411 73146 06089 15630
42831 95113 43511 42082 15140 34733 68076 18292 69486 80468
80583 70361 41047 26792 78466 03395 17635 09697 82447 31405
00209 90404 99457 72570 42194 49043 24330 14939 09865 45906
05409 20830 01911 60767 55248 79253 12317 84120 77772 50103
95836 22530 91785 80210 34361 52228 33869 94332 83868 61672
65358 70469 87149 89509 72176 18103 55169 79954 72002 20582
72249 04037 36192 40221 14918 53437 60571 40995 55006 10694
41692 40581 93050 48734 34652 41577 04631 49184 39295 81776
61885 50796 96822 82002 07973 52925 75467 86013 98072 91942
48917 48129 48624 48248 91465 54898 61220 18721 67387 66575
88378 84299 12193 03785 49314 39761 99132 28775 45276 91816
77800 25734 09801 92087 02955 12872 89848 48579 06028 13827
24028 03405 01178 06316 81916 40170 53665 87202 88638 47121
86558 84750 43994 01760 96205 27937 45416 71964 52261 30781
78545 49201 05329 14182 10971 90472 44682 39304 19819 55799
14969 64623 82780 35686 30941 14622 04126 25498 95452 63937
58697 31973 06303 94202 62287 56164 79157 98375 24558 99241
38449 46438 91579 01907 72146 05764 22400 94490 49833 09258
62134 87244 73348 80114 78490 64735 31010 66975 28652 36166
72749 13347 65030 26128 49067 27904 49953 74674 94617 13317
81638 36566 42709 33717 59943 12027 46547 61303 46699 76243
46574 79670 10342 89543 75030 23428 29541 32501 89422 87474
11873 57196 32209 67663 07990 12288 59245 83638 23642 61715
13862 72778 09949 23096 01791 19472 14634 31690 36602 62943
08312 27886 82321 28666 72998 22514 51054 22940 31842 54245
11071 44430 94664 91294 35163 05494 32882 23904 41340 61185

82509 11842 86963 50307 07510 32545 90717 46856 86079 13769
07426 67341 80314 58910 93948 85738 69444 09370 58194 28207
57696 25592 91221 95386 15857 84645 89659 80535 93233 82798
08074 89810 48521 90740 02687 83117 74920 25954 99629 78978
20128 53721 01518 40699 20849 04710 38989 91322 56057 58573
00190 27157 83208 79446 92987 61357 38752 55424 94518 45205
23798 55425 32454 34611 39605 39981 74691 40836 30812 38563
85306 57995 68222 39055 43890 36956 84861 63624 04961 55439
99719 36036 74274 53901 34643 06157 89500 57514 93977 42403
95970 81452 48873 00784 58347 40269 11880 43395 28249 38743
56651 91460 92462 98566 72062 18556 55052 47614 80044 60015
71499 80220 35750 67337 47556 55272 55249 79100 34014 17037
66660 78443 47545 70736 65419 77489 70831 73237 14970 23129
35483 84563 79956 88618 54619 24853 59783 47537 88822 47227
09262 25041 57862 19203 86103 02800 23198 70639 43757 52064

Source: Adapted from Table of Random Numbers at http://www.mrs.umn.edu/~sungurea/introstat/public/instruction/ranbox/randomnumbersII.html

The above random numbers table contains 450 (5 digit) random numbers.

12.4.3 Steps in the Use of RNT

We need to follow the following steps while selecting a SRS by using RNT.

1. Determine the population size (N).
2. Determine the sample size (n).
3. List all the units of the population. Assign the numbers in a serial order. Suppose there are 100 units in the population, assign the serial numbers from 00 to 99.
4. Determine the starting point of selecting the sample by picking up a page from the random number tables and dropping your finger on a number in the page blindly.

5. Choose the direction in which you want to read the numbers (say from left to right or right to left or top to bottom or bottom to top).
6. Suppose you are looking for two digit numbers (00 to 99) you may not get these numbers by direct reading from the tables since they are 5 digit numbers (see Table 12.1). You can either look at the last two digits or first two digits of the numbers. For example, if the 5 digit number you have chosen is 54245 (that is, the number in the 29th row and 10th column of the random number table given at Table 12.1). Then, the two digit number will be 45 if you chose the last two digits of the number.
7. Look only at the numbers assigned to each population unit. If the number represents one of the unit of the population it becomes part of the sample. Suppose you want to select 10 sample units, the other numbers you will be choosing are 71(11071), 30(44430), 64(94664),94(91294),63(35163),82(32883),04(23904), 40(41340), 85(61185). Observe that you have omitted 94(05494) since you have already chosen this number.
8. Once a number is chosen, do not select it again.
9. If you reach the end point of the table before obtaining the required sample, pick another starting point in the random number table and select the remaining units for the sample.

Example 12.2

Suppose you have to select 100 account holders as a sample out of total 1000 account holders in the population using random numbers table.

628	126	254	090	752	091	411	146	089	630
831	113	511	082	140	733	076	292	486	468
583	361	047	792	466	395	635	697	447	405
209	404	457	570	194	043	330	939	865	906
409	830	911	767	248	253	317	120	772	103
836	530	785	210	228	869	332	868	672	358
469	149	509	176	169	954	002	582	249	037
192	221	918	437	571	995	006	694	692	581
050	734	652	577	631	184	295	776	885	796
822	973	822	467	013	072	942	917	129	624

Here, first you assign each account holder a number from 000 to 999. To draw a sample of 100 account holders, you need to find 100 three digit numbers in the range 000 to 999. Pick up any row or column in the random numbers table given in Table 12.1. Suppose you have selected the fourth row and first column as starting point to draw the sample, the first digit number is 628(64628) if you chose last 3 digits as the number for your purpose. Here, you read the last 3 digits

of the number. If the number is within the range (000 to 999) include the number in the sample. Otherwise skip the number and read the next number in some identified direction. If a number is already selected omit it. In this example since you have started with fourth row and first column and moving from left to right direction the following 100 numbers are selected for the sample.

If the number of units in the population is very large, neither of the above two methods is feasible. These days by using a computer we can select a random sample in a much easier way. There are many computer program which can generate a series of random numbers if we have the units of the population listed in a computer.

We will explain one way of selecting a sample using computer generated random numbers. In our example, let us assume we can copy and paste the list of account holders into a column in an EXCEL spreadsheet. Then, in the column right to it we paste the function =RAND() which is EXCEL's way of putting a random number between 0 and 1 in the cells. Then, all we have to do is take the first 100 names in the sorted list. The entire process takes a minute if we are familiar with using EXCEL program in computer.

12.4.4 Advantages of SRS

1) In simple random sample we assure population units to be homogeneous and thus do not require additional information on the characteristics of the population.
2) Using simple random sampling we can select an unbiased sample. This is because, in SRS the chance of including each unit of population in the sample is equal. The bias due to human preferences is completely eliminated.
3) Through estimation of sampling error we can assess the accuracy of the results.
4) If the population size is not too large, simple random sampling is a simple and easily implementable sampling procedure.

12.4.5 Limitations of SRS

1) The greatest limitation of simple random sampling is that if the population size is too large then we need to spend a lot of time in listing the units of the population.
2) The simple random sampling procedure will be efficient only when we have a homogeneous population. Suppose we have a population with characteristics such as gender, age, social status, etc. Then, we need a larger sample size to accommodate a representative sample with all those characteristics of the population units. A better way to tackle this issue is to use stratified sampling procedure which you will learn later in this unit.

12.5 SELECTION OF SYSTEMATIC RANDOM SAMPLE

The systematic random sampling procedure is somewhat similar to the simple random sampling procedure. In this sampling procedure, we select a starting point at random and then systematically select the sample units from the population units at a specified sampling interval.

The starting point and the sampling interval are based on the required sample size. The sampling interval will be represented as K. The selection of a sample using systematic random sampling procedure is very simple. Suppose the population consists of *n* units and you have decided to select a sample of n units using systematic random sampling procedure. Follow the following steps.

1. Number the units in the population from 1 to N (suppose you have 1000 units).
2. Decide the sample size *n* you need (suppose you want to select 100 units).
3. Determine the sampling interval by dividing the population by the sample size. K = N/*n* = the interval size (here K=1000/100 = 10).
4. Select a unit at random from the first K units (1 to K) (suppose you have selected unit number 5 as your first sample unit).
5. Then select the subsequent sample units by adding K to the previous unit (the subsequent samples will be 15 (5+10), 25 (15+10), …, 995 (985+10).

Example 12.3

From a population consisting of 500 units draw a sample of 60 units using systematic random sampling procedure.

To use systematic random sampling, the first thing we need to do is listing of the population units in a random order by giving numbers from 1 to 500. The sampling interval is K=500/60 = 8.3 or say 8. Now we select the first sample unit at random from the first 8 population units. Suppose the first unit selected is 5. The subsequent sample units selected are: 13, 21, 29, 37.....477. Therefore, following are the population units selected in the sample.

5	13	21	29	37	45	53	61	69	77
85	93	101	109	117	125	133	141	149	157
165	173	181	189	197	205	213	221	229	237
245	253	261	269	277	285	293	301	309	317
325	333	341	349	357	365	373	381	389	397
405	413	429	429	437	445	453	461	469	477

Thus, in the systematic random sampling procedure, the first sample unit is selected at random and this sample unit in turn determines the subsequent sample units to be selected. However, it is essential that the units in the population are randomly ordered. In certain cases we prefer using systematic random sampling procedure to simple random sampling procedure because it is easier to select sample units. For example, if you want to find out the yield of coconut trees in a field, select a tree at random, other trees are automatically selected at a gap equivalent to sampling interval.

12.5.1 Advantages of Systematic Random Sampling

a) The main advantage of using systematic random sample is that the time taken and work involved in systematic random sampling is less than simple random sampling procedure. It is frequently used in exit polls on voting behaviour and obtaining the opinions and views of consumers in marketing research.

b) The other advantage of systematic random sampling procedure is that this method can be used even when no formal list of the population units is available. For example, suppose if we are interested in knowing the opinion of consumers on improving the services offered by a bank, we may simply choose every k^{th} account holder visiting a bank branch, provided that we know how many account holders are there : (1) For example, there are 2000 account holders in the population and we want to have 200 account holders as sample size. Then, K=2000/200=10) and we select every 10^{th} account holder visiting the bank.

12.5.2 Disadvantages of Systematic Random Sampling

a) The main disadvantage of systematic random sampling procedure is that if there is a periodicity in the occurrence of units of a population, the use of systematic random sampling procedure gives a highly unrepresentative sample. For example, suppose you are interested in obtaining the views/opinions of consumers of a store in your locality. You may arrange all the consumers of the store according to their date of visit and start selecting a sample of customers who visit the store on 1^{st} of every month. You know that the 1^{st} day of every month cannot be representative of the whole month.

b) The other disadvantage of systematic random sampling procedure is that every unit of the population does not have an equal chance of being selected. Rather the selection of population units in the sample depends on the initial unit of selection. Regardless of how we select the first unit of the sample, subsequent units are automatically determined. This lacks complete randomness.

12.6 SELECTION OF STRATIFIED RANDOM SAMPLE

In some cases the population may not be homogenous, that is, all the units may not be equal with respect to the characteristic we intent to survey. The characteristics of the population under study may be male/female, rural/urban, literate/illiterate, high income/low income groups, etc. In situations where these units vary widely, the simple random sampling procedure or the systematic random sampling procedure will not provide us with a representative sample. In such situations by using stratified random sampling we can obtain a representative sample.

In stratified sampling, we divide the population into different strata in such a way that units are homogenous within each stratum. Moreover, each stratum is different. Suppose we want to stratify the population on the basis of gender distribution then we list the population units separately according to males or females. Subsequently, we decide the sample size to be drawn from each stratum. There are two approaches to decide the sample size from each stratum. These are: (a) proportional stratified sample, and (b) disproportional stratified sample. We will discuss these two procedures below.

12.6.1 Proportional Stratified Sample

When we take a sample from a population with several strata, we are required to take samples from each stratum. Such sample could be in proportion of the stratum population size to the total population size. Suppose we divide the population (N) into K non-overlapping strata N_1, N_2, N_3..... N_K such that $N_1+N_2+N_3+.....+N_K = N$. We decide to draw a sample of the size n. Then the sample proportions of different strata are given by:

$$\frac{n_1}{N_1} = \frac{n_2}{N_2} = \frac{n_3}{N_3} = ... = \frac{n_{12}}{N_K}$$

Example 12.4

Suppose we want to draw a sample of 200 units from a population consisting of 1000 units. The population is heterogeneous in nature in terms of high income or low income and rural or urban. The strata population sizes are given as follows:

High income - urban = 200
Low income - urban = 400
High income- rural = 100
Low income-rural = 300

To have a representative sample each stratum in the sample should represent the corresponding stratum in the population. For this we should take a different sample size from each stratum depending upon the stratum size. The deciding factor in each of the stratum is same as the proportion of total sample to the

population. In our example, to have a sample of 200 units, the proportion of sample to the population in each stratum is

$$\frac{n}{N} = \frac{200}{1000} = 0.2$$

Observe that we are considering the same proportion for each stratum. Then the sample from each stratum will be as follows:

Stratum Category	Stratum Population size(N_i)	Sample to population proportion	Stratum sample size
(1)	(2)	(3)	(4)=(2)× (3)
High income - urban	200	0.2	40
Low income -urban	400	0.2	80
High income-rural	100	0.2	20
Low income-rural	300	0.2	60
Total	1000	0.2	200

There are several advantages of stratified sampling over simple random sampling. The stratified sampling ensures sample representation of not only the entire population, but also each stratum. This is important where the stratum size is small. Moreover, stratified sampling generally has more statistical precision than simple random sampling.

12.6.2 Disproportional Stratified Sampling

In proportional stratified sampling, we assumed that each stratum in the population is homogeneous. Consequently, we expect that variability within stratum is lower than the variability for the population as a whole. On the other hand, if the variability within each stratum is not small then we use disproportional stratified sampling. In disproportional stratified sampling, the strata allocation is based on size and variability (that is, the standard deviation of the characteristic under study). In this procedure a larger sample is drawn from the stratum having higher variability. This procedure is also sometimes called double weighing scheme and provides the most efficient sample and most precise/reliable estimates for a given sample size. The only requirement is that we should have knowledge/estimate of the standard deviation of the characteristic under study within each stratum.

Follow the steps given below for using disproportional stratified sampling.

1) Divide the population into strata based on the chosen characteristic (example, Rural/Urban, Male/Female, etc.)

2) The number of units taken from each stratum is directly proportional to the relative size of the stratum and standard deviation σ_i of the characteristic under consideration. Suppose, if σ_1, σ_2, σ_3, …σ_k are the standard deviations of k strata and P_1, P_2, P_3, …P_k are the stratum proportions to the total population, and n $(= n_1 + n_2 + ... + n_k)$ is the sample size required. Then the stratum sample size using disproportional stratified sampling procedure is

$$n_i = \frac{P_i \times \sigma_i \times n}{\sum P_i \sigma_i}$$

3) Choose the sample from each stratum using either simple random sampling or systematic random sampling.

Let us go back to Example 12.4, where we have divided the population into 4 strata. We observe that there are small number of households in high income strata and large number of households in low income strata. Assume that the variance of income among higher income groups is higher than the variance among the lower income groups. Therefore, in order to avoid under-representation of higher income groups in the sample, a disproportional sample is taken in each stratum. That means, if the variability within the stratum is higher, we must have larger sample size of that stratum to increase the precision of the estimates. Similarly, if the variability within the stratum is lower, we must have smaller sample size of that stratum. That is, higher the stratum variance larger the stratum sample size and lower the stratum variance smaller the sample size. This is in addition to the fact that larger stratum size requires a larger sample size.

Example 12.5

Consider Example 12.4 again. Suppose the stratum variances are given as follows:

Stratum	Variance ($\sigma^{2)}$
High income urban	6.5
Low income urban	2.5
High income rural	4.5
Low income rural	2.0

Use the disproportional stratified sampling procedure to choose a sample of size 200 from the four strata.

For this example the disproportional sample size for each stratum is given below:

Stratum	Stratum population	Stratum population proportion (P_i)	Stratum Variance (σ_i^2)	Stratum standard deviation ($\sigma_i = \sqrt{\sigma_i^2}$)	$P_i \times \sigma_i$	Sample size $\frac{P_i \times \sigma_i \times n}{\sum P_i \sigma_i}$
High income-urban	200	0.20	6.5	2.5	0.50	56
Low income-urban	400	0.40	2.5	1.6	0.64	72
High income-rural	100	0.10	4.5	2.1	0.21	24
Low income-rural	300	0.30	2.0	1.4	0.42	47
Total	1000				1.77	200

12.6.3 Advantages of Stratified Sampling

a) In stratified random sampling the sample is drawn from each stratum of the population. Therefore, the stratified random sampling procedure is more representative.

b) The stratified random sampling procedure is more precise than simple random sampling. Therefore, to a great extent this procedure avoids sample selection bias.

c) As we have seen in simple random sampling and systematic random sampling procedures, when there is heterogeneity of population we need to have a large sample size to have a fairly representative sample. However, in stratified random sampling this objective can be achieved with a smaller sample size. This saves a lot of time, money and other resources for data collection.

12.6.4 Disadvantages of Stratified Sampling

a) The main disadvantage of stratified random sampling procedure is that we need a detailed knowledge of the distribution of the characteristics in the population. If we cannot accurately identify the homogeneous groups, it is better to use simple random sample since improper stratification can lead to serious error.

b) The other disadvantage of stratified random sampling is that we need to prepare a list of population units for each stratum separately. As the list of population units may not be readily available for each characteristic, the preparation of lists may be a very difficult task.

12.7 SELECTION OF A CLUSTER SAMPLE

Very often population units are spread over a vast geographical area. In that case collection of data through simple random sampling requires a lot of time, money

and manpower as we have to cover the entire geographical area for collecting data on the selected units. Imagine taking a sample of respondents spread all over Uttar Pradesh in order to conduct personal interviews. Using simple random sample, the respondents will be spread all over the state and you have to travel and spend lot of money meeting the respondents. In such situation cluster sampling will be much useful.

The basic principles of cluster sampling are:

i) The differences or variability within a cluster should be as large as possible. As far as possible the variability within each cluster should be the same as that of population.

ii) The variability between clusters should be as small as possible. Once the clusters are selected, all the units in the selected clusters are included in the sample for obtaining data.

In cluster sampling we divide the population into groups called clusters. Then we select a sample of clusters using a simple random sampling.

The population units in each of the clusters are assumed to be as heterogeneous as those in the total population. That is, each cluster itself is a representative of the population.

12.7.1 Steps in Cluster Sampling

In cluster sampling, we follow the steps given below:

1) Divide the population into a number of clusters.
2) Determine the number of clusters needed for your sample.
3) Randomly select the sample of clusters.
4) Survey all units within the sampled clusters.

Suppose the division of clusters is based on the geographical boundaries of the population, then it is called *area sampling*. You have observed that in the case of cluster sampling the clusters are selected using random sampling method. Subsequently all the population units within each sampled cluster are included in the sample. Suppose instead of including all the population units within each selected cluster you chose to include only a sample of units within each cluster. Then you can clearly understand that there are two stages.

In the first stage you select the clusters and in the second stage you select the sample units within each sampled clusters. This sampling procedure is called *two-stage sampling.* Here, the clusters are called primary units and the units within the sampled clusters are called secondary units.

Example 12. 6

Suppose we are interested in finding the options of ATM customers of a Bank in Uttar Pradesh state. We can divide the state into say 30 clusters (may be we can

consider district as a unit and include one or two districts in one cluster). Here, we assume that each of these clusters will represent the opinions of the ATM customers of Uttar Pradesh as a whole. We then select a sample of clusters and obtain the opinion of all the ATM customers in each of the cluster.

12.7.2 Advantages of Cluster Sampling

a) The main advantage of cluster sampling is that it takes less travel time and related data collection costs.

b) Since the researcher need not cover all the clusters and only sample of clusters are covered, it is a more practical method which facilitates fieldwork.

12.7.3 Disadvantages of Cluster Sampling

a) In cluster sampling we assume that each cluster represents the heterogeneity of the population units of all clusters. However, this assumption may not be true in many cases, because often the tendency is that the units in the clusters are more homogeneous than the units of the entire population. That means it is difficult to form heterogeneous clusters.

b) The cluster sampling has a lower sampling efficiency for a given sample size than random sampling and stratified sampling. This method is cost ffective not statistically efficient.

12.8 MULTISTAGE SAMPLING

We have seen in cluster sampling that when we select a sample instead of covering all the units from each cluster, we call it two-stage sampling. The multistage sampling is as extension of two-stage sampling.

The four methods we have covered so far, namely, (a) simple random sampling, (b) systematic random sampling, (c) stratified sampling, and (d) cluster sampling are the simplest probability (or random) sampling procedures. However, in real-life, we use sampling methods that are more complex than the above four methods. The basic principle in multistage sampling is that we can combine these simple methods in a variety of useful ways to address our sampling needs. We call it multistage sampling when we combine two or more of the above sampling methods.

Example 12.7

Consider the case of interviewing school students in Haryana in order to grade the schools according to socio-economic background of the parents. For this problem, in the first stage we need to apply cluster sampling. We divide the state of Haryana into a number of clusters, say districts. Then we select a sample of districts (clusters) using simple random sampling method. In the second stage we

divide the schools using stratified sampling. Here the strata may be government schools, government-aided schools, central schools, and public schools. We select a sample of schools in each stratum using either a simple random sampling or a systematic random sampling. In the third stage we again use simple random sampling and select a sample of classes in each sampled school for face-to-face interviews with the students. In the fourth stage of sampling we consider selecting a sample of students from each sampled class using simple random sampling or systematic random sampling.

In multi stage sampling it is possible to consider as many stages as necessary to achieve a representative sample. In each stage a suitable method of sampling is used. Each stage results in a reduction of the sample size.

Advantages

a) Multistage sampling procedure provides cost gains by reducing the data collection costs.

b) Multistage sampling is more flexible and allows us to use different sampling procedures in different stages of sampling.

c) If the population is spread over a very wide geographical area, multistage sampling is the most appropriate sampling method.

Disadvantages

If the sampling units selected at different stages are not representative, multistage sampling becomes less precise and less efficient.

Check Your Progress 1

1) Which of the following is a procedure of selecting samples from a population?

   a) Random sampling

   b) Non-random sampling

   c) Stratified sampling

   d) All the above

2) Suppose you are applying a stratified random sampling procedure on a population. How do you make your sample selection?

   a) Select at random an equal number of units from each stratum

   b) Draw equal number of units from each stratum and weigh the results

   c) Select the sample at random from each stratum proportional to the population

   d) b. and c. both

   e) a. and c. both

3) State whether the following statements are 'True' or 'False'.

   a) A sampling procedure that selects units from a population at uniform intervals is called simple random sampling.

   b) A sampling procedure that divides the population into well-defined groups from which random samples are drawn is known as stratified sampling.

4) A population is made up of groups that have wide variation between groups but little variation within each group. In this situation the appropriate type of sampling procedure to use is

   a) Cluster sampling

   b) Systematic sampling

   c) Stratified sampling

   d) Multistage sampling

12.9 NON-PROBABILITY SAMPLING PROCEDURES

There are different types of non-probability sampling such as

1. Convenience Sampling
2. Judgment Sampling
3. Quota Sampling
4. Snowball Sampling

We discuss the procedure of drawing a non-probability sampling below.

12.9.1 Convenience Sampling

This is one of the most commonly used methods of non-probability sampling. In this method the researcher's convenience forms the basis for selection of the sample. Especially for a exploratory research there is a pressing need for data. In such situations the selection of sampling units is left to the interviewer. The population units are included in the sample simply because they are in the right place at the right time. This method is often used during preliminary research efforts to get a gross estimate of the results, without incurring the cost or time required to select a sample. For example, during the budget session or when the price of a product is increased or a new government is formed, convenience samples are used by the researchers/journalists to reflect public opinion. Convenience samples are extensively used in marketing research.

The advantage of convenience sampling is that it is less expensive and less time-consuming. The limitations of convenience sampling are: (a) it involves sample

selection bias, and (b) it does not provide a representative sample of the population and therefore we cannot generalise the results.

12.9.2 Judgment Sampling

This is another commonly used non-probability sampling procedure. This procedure is often referred to as *purposive sampling*. In this procedure the researcher selects the sample based on his/her judgment. The researcher believes that the selected sample elements are representative of the population. For example, the calculation of consumer price index is based on judgment sampling. Here the sample consists of a basket of consumer items and other goods and services which are expected to reflect a representative sample.

The prices of these items are collected from selected cities that are viewed as typical cities with demographic profiles matching the national profile.

The advantage of judgment sampling is that it is low cost, convenient and quick. The disadvantage is that it does not allow direct generalisations to population. The quality of the sample depends upon the judgment of the researcher.

12.9.3 Quota Sampling

In this procedure the population is divided into groups based on some characteristics such as gender, age, education, religion, income group, etc. A quota of units from each group is determined. The quota may be either proportional or non-proportional. The proportional quota sampling is based on the proportion of each characteristic in the population so that the proportion in the sample represents the population proportion. For example, if you know that there are 80% of the households whose income is below say Rs.100000 per annum and 20% households have income above Rs.100000 per annum in a city. You want to take a sample of size 100 households. Then you include 80 households from below Rs.100000 income and 20 households from above Rs.100000 income. The objective here is to meet the proportional quota of sampling from each characteristic in the population.

The non-proportional quota sampling is a bit less restrictive. In this procedure, you specify the minimum number of sampled units from each group. You are not concerned with having proportions in the population. For instance, in the above example you may simply interview 50 households from each income group instead of 80% and 20%. The interviewer is instructed to fill the quota for each group based on convenience or judgment. The very purpose of quota sampling is that various groups in the population are represented to the extent the investigator desires.

Do not confuse the quota sampling with stratified sampling that you have learned earlier. In stratified sampling you select random samples from each stratum or group whereas in quota sampling the interviewer has a fixed quota. For example, in a city there are five market centres. A company wants to assess the demand for its new product and sends 5 investigators to assess the demand by interviewing

50 prospective customers from each market. It is left to the investigator whom he/she will interview at each market centre. If the product is targeted to women, this way you cannot elicit the information among various groups of women customers like housewives or employed women or young or old. In this sampling you are simply fixing a quota for each investigator.

The quota sampling has the advantage over others if the sample meets the characteristics of the population that you are looking into. In addition, the cost and time involved in collecting the data are greatly reduced. However, there are many disadvantages as well. In quota sampling, the samples are selected according to the convenience of the investigator instead of selecting random samples. Therefore, the selected samples may be biased.

If there are a large number of characteristics on the basis of which the quotas are fixed, then it becomes very difficult to fix the quotas/sub-quotas for each group/sub-group. Also the investigators have the tendency to collect information only from those who are willing to provide information and avoid unwilling respondents.

12.9.4 Snowball Sampling

In snowball sampling, we begin by identifying someone who meets the criteria for inclusion in our study. We then ask him/her to recommend others who also meets the criteria. Although this method would hardly lead to representative samples, there are times when it may be the best method available. Snowball sampling is especially useful when we are trying to reach populations that are inaccessible or hard to find. For example, if we are studying the homeless, we are not likely to find good lists of homeless people within a specific geographical area. However, if we go to that area and identify one or two, we may find that they know very well who the other homeless people in their vicinity are and how we can find them.

12.10 DETERMINATION OF THE SAMPLE SIZE

The use of appropriate sampling procedure is necessary for a representative sample. However, this condition is not sufficient. In addition to the above, we should determine the sample size. The question of how large a sample should be is a difficult one. Sample size can be determined by various considerations. The following are the some of the considerations in determining the sample size:

Sampling error

Number of comparisons to be made

Response rates

Funds available

Sampling Error: In Unit 13 you will learn that smaller samples have greater sampling error than large samples. On the other hand, larger samples have larger non-sampling errors than smaller samples. The sampling error is a number that

describes the precision of an estimate of the sample. It is usually expressed as a margin of error associated with a statistical level of confidence. For example, for a prime minister preferential poll you may say that the incumbent is favored by 65% of votes, with a margin of error (precision) of plus or minus 5 percentage points at a 95% confidence level. This means that if the same surveys were conducted with 100 different samples of voters, 95 of the surveys would be expected to show the incumbent favoured by between 60% and 70% of the voters (65% ± 5%). Remember as you increase the precision level of your results you need larger sample size.

Number of Comparisons to Make: Sometimes we may be interested in making comparisons of two or more groups (strata) in the sample. For example, we may want to make the comparison between male and female respondents or between urban and rural respondents. Or we may want to compare the results for 4 geographical regions of the country say north, south, west and east. Then we need an adequate sample size in each region or stratum of the population. Therefore, the heterogeneity of population characteristics plays a significant role in deciding the sample size.

Response Rates: In mail surveys, we know that all those questionnaires mailed to the respondents may not reach us back after filling the questionnaires. As per the experiences on mail survey, the response rate ranges between 10% and 50%. Then, if you are expecting a 20% response rate, for example, you will have to mail 5 times the number of sample size required.

Funds Available: The funds available may influence the sample size. If the funds available for the study are limited then you may not be able to spend more than a certain amount of the total money available with you on collecting the data.

It is even more difficult to decide the sample size, when you use the non-probability sampling procedures. This is because there are no definite rules to be followed in non-probability sampling procedures. It all depends upon on what you want to know, the purpose of inquiry, what will be useful, what will have credibility and what can be done with available time and resources. In purposive sampling, the sample should be judged on the basis of purpose. In non-probability sampling procedures, the validity, meaningfulness, and insights generated have more to do with the information-richness of the sample units selected rather than the sample size.

Some Formulae to Determine the Sample Size

Technical considerations suggest that the required sample size is a function of the precision of the estimates you wish to achieve, the variance of the population and the confidence level you wish to use. If you want more precision and confidence level then you may need larger sample size. The more frequently used confidence levels are 95% and 99%. And the more frequently used precision levels are 1% and 5%. There are different formulae used to determine the sample size

depending upon various considerations discussed above. In this section we will discuss three of them.

If we wish to report the results as percentages (proportions) of the sample responding, we use the following formula:

$$n_i = \frac{P_i(1-P_i)}{\frac{A^2}{Z^2} + \frac{P_i(1-P_i)}{N_i}}$$

Where, n_i = sample size of the i^{th} attribute required

P_i = estimated proportion of the population possessing i^{th} attribute of interest (for example, proportion of males, females, urban, rural, etc.)

A = precision required (0.01, 0.05 etc.)

Z = standardized value indicating the confidence level (Z=1.96 at 95% confidence level and Z=2.58 at 99% confidence level)

N_i = population size of the i^{th} attribute (known or estimated)

Example 12.8: A population consists 80% rural and 20% urban people. Given that the population size is 50000, determine the sample size required. Assume that the desired precision and confidence levels are 1% and 99% respectively.

In this example,

P_1 = proportion of rural people = 0.80

P_2 = proportion of urban people = 0.20

N_1 = rural population size = $50000 \times 0.80 = 40000$

N_2 = urban population size = $50000 \times 0.20 = 10000$

A = 0.01

Z = 2.58 (at 99% confidence level)

The required sample size is

$$n_1 = \text{rural sample} = \frac{P_1(1-P_1)}{\frac{A^2}{Z^2} + \frac{P_1(1-P_1)}{N_1}}$$

$$= \frac{0.80(1-0.80)}{\frac{0.01^2}{2.58^2} + \frac{0.80(1-0.80)}{40000}}$$

$$= \frac{0.80(0.20)}{\frac{0.0001}{6.6564} + \frac{0.80(0.20)}{40000}}$$

$$= \frac{0.16}{0.000019 + \frac{0.16}{40000}}$$

$$= \frac{0.16}{0.000019 + 0.000004}$$

$$= \frac{0.16}{0.000023} = 8410.8 \text{ or say } 8411$$

n_2 = urban sample = $\dfrac{P_2(1-P_2)}{\dfrac{A^2}{Z^2} + \dfrac{P_2(1-P_2)}{N_2}}$

$$= \frac{0.20(1-0.20)}{\dfrac{0.01^2}{2.58^2} + \dfrac{0.20(1-0.20)}{10000}}$$

$$= \frac{0.20(0.80)}{\dfrac{0.0001}{6.6564} + \dfrac{0.20(0.80)}{10000}}$$

$$= \frac{0.16}{0.000019 + \dfrac{0.16}{10000}}$$

$$= \frac{0.16}{0.000019 + 0.000016}$$

$$= \frac{0.16}{0.000035} = 4568.4 \text{ or say } 4568$$

Therefore we need to have a sample of size 8411+4568 = 12979 units.

If we wish to report the results as means (averages) of the sample responding, we use the following formula:

$$n_i = \frac{P_i^2}{\dfrac{A^2}{Z^2} + \dfrac{P_i^2}{N_i}}$$

Where, n_i = sample size of the i^{th} attribute required

P_i = estimated standard deviation of the i^{th} attribute of interest (for example, average income of high income group, low income group etc.)

A = precision required (0.01 or 0.05 as the case may be)

Z = standardized value indicating the confidence level (Z=1.96 at 95% confidence level and Z=2.58 at 99% confidence level)

N_i = population size of the i^{th} attribute (known or estimated)

Example 12.9: It is planned to conduct a study to know the average income of households. Given that the standard deviation of households is 2.5 and the population size is 10000, determine the sample size required. Assume that the desired precision and confidence levels are 5% and 95% respectively.

In this example,

P_1 = standard deviation of income = 2.5

N_1 = number of households = 10000

A = 0.05

Z = 1.96 (at 95% confidence level)

The required sample size is

$$n_1 = \frac{P_1^2}{\frac{A^2}{Z^2} + \frac{P_1^2}{N_1}}$$

$$= \frac{2.5^2}{\frac{0.05^2}{1.96^2} + \frac{2.5^2}{10000}}$$

$$= \frac{6.25}{\frac{0.0025}{3.8416} + \frac{6.25}{10000}}$$

$$= \frac{6.25}{0.000651 + 0.000625}$$

$$= \frac{6.25}{0.001276} = 4898$$

If we wish to report the results in a variety of ways or we have the difficulty in estimating the proportion or standard deviation of the attribute of interest, we use the following formula:

$$n = \frac{0.25}{\frac{A^2}{Z^2} + \frac{0.25}{N}}$$

Where, n = sample size required

A = precision required (0.01 or 0.05 as the case may be)

Z = standardized value indicating the confidence level (Z=1.96 at 95% confidence level and Z=2.56 at 99% confidence level)

N = population size (known or estimated)

Example 12.10: Given that the population size is 10000, determine the sample size required when desired precision and confidence levels are 5% and 99% respectively.

In this example,

N = 10000

A = 0.05

Z = 2.58 (at 99% confidence level)

The required sample size is

$$n = \frac{0.25}{\frac{0.05^2}{2.58^2} + \frac{0.25}{10000}}$$

$$n = \frac{0.25}{\frac{0.0025}{6.6564} + \frac{0.25}{10000}}$$

$$n = \frac{0.25}{0.0003756 + 0.000025} = \frac{0.25}{0.000401} = 624$$

Check Your Progress 2

1) Say whether the following statements are 'True' or 'False'.

   a) When the units included in the sample are based on judgment of the investigator, the sampling is said to be random.

   b) With increasing sample size the sampling error decreases.

   c) Convenience sampling has the disadvantage that it may not be representative sample.

2) One of the major disadvantage of judgment sampling is

   a) The procedure is very cumbersome

   b) The sample selection depends on the individual judgment of the investigator

   c) It gives small sample size

   d) It is very expensive

12.11 LET US SUM UP

The most commonly used probability sampling procedure is the simple random sampling which allows a chance to all population units to be included in the sample. The sample units are chosen using random number tables. A systematic random sample uses the first sample unit at random as a starting point and the subsequent sample units are chosen systematically. A stratified sample guarantees inclusion of units from each stratum. A cluster sample involves complete enumeration of one or more randomly selected clusters.

The non-probability sampling procedures include convenience sampling, judgment sampling, quota sampling and snowball sampling. These sampling procedures are not independent from sampling bias but still popular in some situations particularly marketing research.

A number of factors decide the sample size. It may be the number of groups in the population, the heterogeneity of population, funds and time available, etc.

Using a sample saves a lot of money, time and manpower. If a suitable sampling procedure is used in selecting units, appropriate sample size is selected and necessary precautions are taken to reduce sampling errors, then a sample should yield a valid and reliable information about the population.

12.12 ANSWERS/HINTS TO CHECK YOUR PROGRESS EXCERCISES

Check Your Progress 1

1) d

2) c

3) a) False

   b) True

4) c

Check Your Progress 2

1) a) False

   b) True

   c) True

2) b

UNIT 13 STATISTICAL ESTIMATION*

Structure

13.0 OBJECTIVES

After going through this Unit you will be in a position to

- distinguish between sampling error and non-sampling error;
- explain the concept of sampling distribution;
- explain the concept of standard error;
- explain the concept of estimation;
- distinguish between point estimation and interval estimation;
- estimate confidence interval for a parameter; and
- explain the concept of confidence level.

* Prof. Kaustuva Barik, Indira Gandhi National Open University, New Delhi

13.1 INTRODUCTION

Many times due to certain constraints such as inadequate funds or manpower or time we are not in a position to survey all the units in a population. In such situations we take resort to sampling, that is, we survey only a part of the population. On the basis of the information contained in the sample we try to draw conclusions about the population. This process is called statistical inference. We must emphasize that statistical inference is widely applied in economics as well as in many other fields such as sociology, psychology, political science, medicine, etc. For example, before election process starts or just before declaration of election results many newspapers and television channels conduct exit polls. The purpose is to predict election results before the actual results are declared. At that point of time, it is not possible for the surveyors to ask all the voters about their preferences for electoral candidates – the time is too short, resources are scarce, manpower is not available, and a complete census before election defeats the very purpose of election!

In the above example the surveyor actually does not know the result, which is the outcome of votes cast by all the voters. Here all the voters taken together comprise the population. The surveyor has collected data from a representative sample of the population, not all the voters. On the basis of the information contained in the sample, (s)he is making forecast about the entire population.

In this unit we deal with the concept of statistical inference and methods of statistical estimation. Parameter, as you know, is a function of population units while statistic is a function of sampling units. There could be a number of *parameters* and corresponding *statistics*. However, in order to keep our presentation simple, we will confine ourselves mostly to arithmetic mean.

13.2 STATISTICAL BACKGROUND

In the previous two units we have discussed two important aspects: theoretical probability distributions, and sampling techniques. These two aspects form the basis of statistical inference.

In Unit 10 we explained the concept of a random variable. We learnt that X is a random variable if it assumes values $x_1, x_2, ..., x_n$ with corresponding probabilities $p_1, p_2, ..., p_n$ attached to it. Here the probability of occurrence of x_1 is p_1, the probability of occurrence of x_2 is p_2, and so on. If the values $x_1, x_2, ..., x_n$ are discrete we call X a discrete random variable and find out the probability for isolated values of X. On the other hand, if X is a continuous random variable we can find out the probability of X within certain range such that $P(a \le X \le b) = p_1$.

In Units 10 and 11 we discussed theoretical discrete probability distributions (such as binomial and Poisson) and continuous probability distributions (such as normal and t). We learnt that if the range of X increases infinitely then these

probability distributions approach normal distribution. Thus normal distribution is a limiting case of these probability distributions and is considered as a sort of ideal among probability distributions.

The normal distribution is defined by two parameters: mean (μ) and standard deviation (σ). If the probabilities associated with a random variable are distributed according to normal distribution (that means, if X follows normal distribution), we can find out the probability of $P(a \le X \le b) = p_1$ by using the equation for its probability distribution function.

A problem encountered here is that μ and σ can take any values and finding out corresponding probability is time consuming. This problem is tackled by subtracting μ from the normal variable and dividing it by σ. This way we obtain the 'standard normal variate', $z = \frac{x-\mu}{\sigma}$, which has mean = 0 and standard deviation = 1. By plotting the probabilities for different values of z on a graph paper we obtain 'standard normal curve' which is symmetrical and area under the curve is = 1. Remember that in the case of standard normal curve we measure $z = \frac{x-\mu}{\sigma}$ on the x-axis and probability of occurrence of z, that is $p(z)$, on the y-axis. Thus if we consider a particular segment of the normal curve (bounded by two values of z, say, z_1 and z_2) the area under the curve gives its probability. Remember that normal curve is different from the frequency curve given in Unit 2 of this course. *You should note that the area under the normal curve does not give frequencies; it gives probabilities.*

13.3 CERTAIN CONCEPTS

We explain below some of the concepts frequently used in sampling theory.

13.3.1 Parameter

In a statistical inquiry, our interest lies in one or more characteristics of the population. A measure of such a characteristic is called a *parameter*. For example, we may be interested in the mean income of the people of some region for a particular year. We may also like to know the standard deviation of these incomes of the people. Here, both mean and standard deviation are parameters.

Parameters are conventionally denoted by Greek alphabets. For example, the population mean is usually denoted by μ and population standard deviation is usually denoted by σ.

It is important to note that the value of a parameter is computed from all the population observations. Thus, the parameter 'mean income' is calculated from all the income figures of different individuals that constitute the population. Similarly, for the calculation of the parameter 'correlation coefficient of heights and weights', we require the values of all the pairs of heights and weights in a population.

Thus, we can define a *parameter as a function of the population values.* If θ is a parameter that we want to obtain from the population values $X_1, X_2, X_3, \ldots\ldots, X_N$ then

$$\theta = f(X_1, X_2, X_3, \ldots\ldots, X_N)$$

13.3.2 Statistic

While discussing the census and the sample survey, we have seen that due to various constraints, sometimes it is difficult to obtain information about the whole population. In other words, it may not be always possible to compute a population parameter. In such situations, we try to get some idea about the parameter from the information obtained from a sample drawn from the population. This sample information is summarised in the form of a *statistic*. For example, sample mean or sample median or sample mode is called a statistic. Thus, a statistic is calculated from the values of the units that are included in the sample. So, a *statistic can be defined as a function of the sample values*. Conventionally, a statistic is denoted by an English alphabet. For example, the sample mean may be denoted by $\bar{x}$ and the sample standard deviation may be denoted by *s*. If *T* is a statistic that we want to obtain from the sample values $x_1, x_2, \ldots., x_n$, then

$$T = f(x_1, x_2, \ldots., x_n)$$

13.3.3 Estimator and Estimate

The basic purpose of a statistic is to estimate some population parameter. The procedure followed or the formula used to compute a statistic is called an *estimator* and the value of a statistic so computed is known as an *estimate*.

If we use the formula $\bar{x} = \frac{x_1+x_2+\cdots+x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$ for calculating a statistic, then this formula is an estimator. Next, if we use this formula and get $\bar{x} = 10$, then this '10' is an estimate.

13.3.4 Sampling Distribution

As a sample is much smaller than the parent population, many samples can be selected from the same population. Since an 'estimate' of a parameter depends upon the sample values, and these values may change from one sample to another, there can be different estimates or values of a statistic for the same parameter. This variation in values is called *sampling fluctuation*. Suppose, a number of samples, each of size *n*, are drawn from a population of size N and for each sample, the value of the statistic is computed. If the number of samples is large, these values can be arranged in the form of a relative frequency distribution. When the number of samples tends to infinity, the resultant relative frequency distribution of the values of a statistic is called the *sampling distribution* of the given statistic.

Suppose, we are interested in estimating the population mean (which is a parameter), denoted by μ. A random sample of size n is drawn from this population (of size N). The sample mean $\bar{x} = \frac{1}{n}\sum x_i$ is a statistic corresponding to the population mean μ. We should note that $\bar{x}$ is a random variable as its value changes from one sample to another in a probabilistic manner.

Example 13.1

Consider a population consisting of the following 5 units: 2, 4, 6, 8, and 10. Suppose, a sample of size 2 is to be selected from it by the method of simple random sampling without replacement. Obtain the sampling distribution of the sample mean and its standard error.

Solution: The number of samples that can be selected without replacement

$$= {}^{N}c_{n} = {}^{5}c_{2} = \frac{5!}{2!(5-2)!} = \frac{5!}{2!\,3!} = \frac{5\times 4}{2\times 1} = \frac{20}{2} = 10.$$

The possible samples along with the corresponding sample means $\left(\bar{x}\right)$ are presented in Table 13.1.

Table 13.1: Possible Samples and Sample Means

Sample	Sample Mean $\left(\bar{x}\right)$
(2, 4)	3
(2, 6)	4
(2, 8)	5
(2, 10)	6
(4, 6)	5
(4, 8)	6
(4, 10)	7
(6, 8)	7
(6, 10)	8
(8, 10)	9

Now, we can have a frequency distribution of the sample means:

Table 13.2:Frequency Distribution of Sample Means

Sample Mean $\left(\bar{x}\right)$	Frequency (f)
3	1
4	1
5	2
6	2
7	2
8	1
9	1

From the frequency distribution given in Table 13.2, we can present the probability distribution of the sample mean as given in Table 13.3.

Table 13.3: Sampling Distribution of Sample Means

Sample Mean $(\bar{x})$	Probability $\left(\frac{f}{\Sigma f}\right)$
3	$\frac{1}{10}$
4	$\frac{1}{10}$
5	$\frac{2}{10}$
6	$\frac{2}{10}$
7	$\frac{2}{10}$
8	$\frac{1}{10}$
9	$\frac{1}{10}$

We note here that $\sum f$, which, from the frequency distribution of the sample mean presented earlier, is equal to 10. In Table 13.3, we have used the relative frequency for the calculation of the probabilities.

13.3.5 Standard Error of a Statistic

In the previous Section we learnt that we can draw a number of samples depending upon the population and sample sizes. From each sample we get a different value for the statistic we are looking for. These values can be arranged in the form of a probability distribution, which is called the sampling distribution of the concerned statistic. The statistic is also similar to a random variable since a probability is attached to each value it takes. In Table 13.3 we have presented the statistic along with its probability.

We have learnt in Unit 10 that mathematical expectation of a random variable is equal to its arithmetic mean. Let us find out the mathematical expectation and standard deviation of the sampling distribution.

We notice two important properties of the sampling distribution.

1) The expectation of the sampling distribution of the statistic is equal to the population parameter. Thus if we have the sampling distribution of

sample means ($\bar{x}$), then its expected value is equal to population mean (μ). Symbolically, E($\bar{x}$) = μ.

2) The standard deviation of the sampling distribution is called 'standard error' of the concerned statistic. Thus if we have sampling distribution of sample means, then its standard deviation is called the 'standard error of sample means'. Thus standard error indicates the spread of the sample means away from the population mean. In Unit 14 we would see that standard error is used for hypothesis testing and statistical estimation.

Example 13.2

Find out the standard error of the sampling distribution given in Table 13.3.

Solution: We know that standard error of the sample mean is standard deviation of the sampling distribution. Thus,

$$\sigma_{\bar{x}} = \sqrt{E\left(\bar{x}\right)^2 - \left[E\left(\bar{x}\right)\right]^2}$$

Now,

$$E\left(\bar{x}\right) = 3\times\frac{1}{10} + 4\times\frac{1}{10} + 5\times\frac{2}{10} + 6\times\frac{2}{10} + 7\times\frac{2}{10} + 8\times\frac{1}{10} + 9\times\frac{1}{10} = \frac{60}{10} =$$

and

$$E\left(\bar{x}\right)^2 = 9\times\frac{1}{10} + 16\times\frac{1}{10} + 25\times\frac{2}{10} + 36\times\frac{2}{10} + 49\times\frac{2}{10} + 64\times\frac{1}{10} + 81\times\frac{1}{10} = \frac{390}{10}$$

$= 39.$

$$\therefore \sqrt{E\left(\bar{x}\right)^2 - \left[E\left(\bar{x}\right)\right]^2} = \sqrt{39-36} = \sqrt{3} = 1.73.$$

Thus, the standard error of the sample mean in this case is 1.73.

Now a question may be shaping up in your mind.

Do we have to draw all possible samples to find out standard error? In Example 13.2 above we first noted down all the possible samples, arranged these in a relative frequency distribution form and thereafter calculated the standard deviation. In Example 13.2 the population size and sample size were quite small, and thus the task was manageable. But, can you imagine what would happen when we have much larger population and sample sizes? It is too difficult and cumbersome a task. In fact the entire advantages of sampling disappear if we start selecting all possible samples!

Secondly, is it possible to fit a theoretical probability distribution to the sampling distribution? In fact, the *Central Limit Theorem* says that, "if samples of size *n* are drawn from any population, the sample means are approximately normally distributed for large values of *n*". Thus whatever be the distribution of the population, the sampling distribution of $\bar{x}$ is approximately normal for large enough sample sizes.

If the population is normal, then sampling distribution of $\bar{x}$ is normal for any sample size. If population is approximately normally distributed than sampling distribution of $\bar{x}$ is approximately normal even for small sample size. Moreover, even if population is *not* normally distributed, sampling distribution of $\bar{x}$ is approximately normal for large sample sizes.

Thirdly, what is the relationship between standard deviation of the population from which the sample is drawn and the standard error of $\bar{x}$? Obviously, the spread of $\bar{x}$ will be less than the spread of the population units. The standard error of $\bar{x}$ is given by

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ where $\sigma_{\bar{x}}$ is standard error of $\bar{x}$ and σ is standard deviation of the original population.

Thus standard error is always smaller in value than standard deviation of the population, because standard error is equal to the standard deviation of the population divided by square root of the sample size.

The above is true for simple random sampling with replacement. When sampling is without replacement in that case we have to make some finite population correction and standard error is given by $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \times \frac{N-n}{N-1}$.

When the ratio $\frac{n}{N}$ is very small both the procedures give almost similar results. But when sample size is not negligible compared to population size the correction factor needs to be applied.

How do we interpret the standard error? As mentioned earlier it shows the spread of the statistic. Thus, if standard error is smaller then there is a greater probability that the estimate is closer to the concerned parameter.

Check Your Progress 1

1) Define the following concepts:

   a) Population

   b) Sample

   c) Parameter

   d) Statistic

   ……

   ……

   ……

   ……

   ……

2) Distinguish between the following:
   a) Estimator and Estimate
   b) Census and Sample Survey

   ……

   ……

   ……

   ……

   ……

3) Explain the following concepts:
   a) Sampling Distribution
   b) Standard Error

   ……

   ……

   ……

   ……

   ……

4) Given a population: 2, 4, 6. Assume that a sample of size 2 is selected from this population by the method of random sampling without replacement.
   a) Present the sampling distribution of the sample mean.
   b) Compute the standard error.

   ……

   ……

   ……

   ……

   ……

13.4 NON-SAMPLING AND SAMPLING ERRORS

The basic purpose of sampling is to draw inferences about the population on the basis of the sample. For example, we have to find out the per capita income of a village. Due to shortage of time, money and personnel we do not undertake a complete census and opt for a sample survey. In this case it is very likely that the per capita income obtained from the sample is not equal to the actual per capita income of the village. This discrepancy could arise because of two reasons:

Since we are collecting data from only a part of the population (i.e., the sample selected by us), sample mean (per capita income in this case) is not equal to population mean. If at all both are equal, it is a rare coincidence! If we take sample mean as population mean we are committing an error called sampling error.

A second source of error could arise because of wrong reporting or recording or tabulation or processing of data. This type of error is termed non-sampling error. Remember that non-sampling error, as its name suggests, has nothing to do with our sampling process. Wrong reporting or recording or processing of data can take place in a sample survey also.

We explain the sources of these errors below.

13.4.1 Non-Sampling Error

Various sources of non-sampling error are given below.

(i) Error due to measurement

It is a well-known fact that precise measurement of any magnitude is not possible. If some individuals, for example, are asked to measure the length of a particular piece of cloth independently up to, say, two decimal points; we can be quite sure that their answers will not be the same. In fact, the measuring instrument itself may not have the same degree of accuracy.

In the context of sampling the respondents of an inquiry, for example, may not be able to provide the accurate data about their incomes. This may not be a problem with individuals earning fixed incomes in the form of wages and salaries. However, self-employed persons may not be able to do so.

(ii) Error due to non-response

Sometimes the required data are collected by mailing questionnaires to the respondents. Many of such respondents may return the questionnaires with incomplete answers or may not return them at all. This kind of an attitude may be due to the following reasons:

a) The respondents are too casual to fill up the answers to the questions asked.

b) They are not in a position to understand the questions.

c) They may not like to disclose the information that has been sought.

We should note that the error due to non-response may also arise because of the possibility of the questionnaire being lost in transit.

If the data are collected through personal interviews, some of the reasons for the error due to non-response pointed out above may not arise. However, in that case this error may arise because some of the individuals:

a) may not like to give the information, or

b) may not simply be available even after repeated visits.

(iii) Error in recording

This type of error may arise at the stage when the investigator records the answers or even at the tabulation stage. A major reason for such error is the carelessness on the part of the investigator.

(iv) Error due to inherent bias of the investigator

Every individual suffers from personal prejudices and biases. Despite the provision of the best possible training to the investigators, their personal biases may come into play when they interpret the questions to be put to the respondents or record the answers to these questions.

In complete enumeration the extent of non-sampling error tends to be significantly large because, generally, a large number of individuals are involved in the data collection process. We try to minimise this error through:

i) a careful planning of the survey,

ii) providing proper training to the investigators,

iii) making the questionnaire simple.

However, we would like to emphasize that complete enumeration is always prone to large non-sampling errors.

13.4.2 Sampling Error

By now it should be clear that in the sampling method also, non-sampling error may be committed. It is almost impossible to make the data absolutely free of such errors. However, since the number of respondents in a sample survey is much smaller than in census, the non-sampling error is generally less pronounced in the sampling method. Besides the non-sampling errors, there is sampling error in a sample survey. Sampling error is the absolute difference between the parameter and the corresponding statistic, that is, $|T-\theta|$.

Sampling error is not due to any lapse on the part of the respondent or the investigator or some such reason. It arises because of the very nature of the procedure. It can never be completely eliminated. However, we have well developed sampling theories with the help of which the effect of the sampling error can be minimised.

13.5 DESIRABLE PROERTIES OF AN ESTIMATOR

Suppose, θ is an unknown population *parameter* that we are interested in. We may want to estimate θ on the basis of a random sample drawn from the population. For this purpose we may use a statistic T (which is a function of the sample values). Here T is an *estimator* of θ and the value of T that is obtained from the given sample is an *estimate* of θ. In fact, the value is known as a *point estimate* in the sense that it is one particular value of the estimator (see Unit 14 for details).

Earlier, we have discussed the concepts of sampling and non-sampling errors. We recapitulate here that the absolute difference (ignoring the sign) between a sample statistic and the population parameter, i.e., $|T-\theta|$ measures the extent of the sampling error. We may note here that an estimator is essentially a formula

for computing an estimate of the population parameter and there can be several potential estimators (alternative formulae) that may be used for this purpose. So, there should be some desirable properties on the basis of which we can select a particular estimator for estimating the population parameter. A very simple requirement for T to be a good estimator of θ is that the difference $|T-\theta|$ should be as small as possible. Various approaches have been suggested to ensure this.

13.5.1 Unbiasedness

We have already noted that the value of a statistic varies from sample to sample due to sampling fluctuation. Although the individual values of a statistic may be different from the unknown population parameter, on an average, the value of a statistic should be equal to the population parameter. In other words, the sampling distribution of T should have a central tendency towards θ. This is known as the property of unbiasedness of an estimator. It means that although an individual value of a given estimator may be higher or lower than the unknown value of the population parameter, there is no bias on the part of the estimator to have values that are always greater or smaller than the unknown population parameter. If we accept that mean (here, expectation) is a proper measure for central tendency, then T is an *unbiased estimator* for θ

if $E(T)=\theta$.

13.5.2 Minimum-Variance

It is also desirable that the average spread of all the possible values of an unbiased estimator around the population parameter is as small as possible. It will reduce the chance of an estimate being far away from the parameter. If we accept that variance is a proper measure for average spread (dispersion), we want that among all the unbiased estimators, T should have the smallest variance. Symbolically, $V(T)\le V(T')$ where, V stands for variance and T' is any other unbiased estimator.

An estimator T, which is unbiased and among all the unbiased estimator has the minimum variance, is known as a *minimum-variance unbiased estimator*. Let us consider an example. Suppose, we have a random sample of size n from a given population of size N. In this case, the sample mean is given by $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$ where x_i is the i^{th} member of the sample. It can be proved that it is an unbiased estimator of the population mean μ. Symbolically

$$E(\bar{x})=\mu$$

However, it can be shown that the sample variance defined as $s^2=\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2$ is not an unbiased estimator of the population variance σ^2. Symbolically,

$$E\left(s^2\right)\neq\sigma^2$$

On the contrary, if we define the sample variance as $s'^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$, then s'^2 is an unbiased estimator of $\sigma'^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$

Suppose further that the sample values are not only random but also independent (random sample with replacement) and the underlying population is normal. It can be shown that the sample mean $\bar{x}$ is not only an unbiased estimator of the population mean μ but also it has the minimum variance among all the unbiased estimators of μ.

13.5.3 Consistency and Efficiency

Another approach may be to suggest that the estimator T should approximate the unknown population parameter θ as the sample size n increases. Since T itself is a random variable, we may express this requirement in probabilistic or stochastic terms as the statistic T should converge to the parameter θ stochastically (i.e., in probability) as $n \to \infty$. A statistic T with this property is called a *consistent* estimator of θ.

In real life, a large number of consistent estimators of the same parameter θ have often been found. In such a situation, obviously, some additional criterion is needed to choose among these consistent estimators. One such criterion may be to demand that not only T should converge stochastically to θ but also it should do so quite rapidly. Without going into the details, we may mention here that some times an estimator assumes the form of a normal distribution when the sample size n increases indefinitely. Such estimators are called *asymptotically normal*. If we focus on consistent estimators that are asymptotically normal, the rapidity of their convergence is indicated by their respective asymptotic variances. In fact, the convergence is the fastest for the estimator that has the *lowest asymptotic variance.* Such kind of an estimator is known as an *efficient estimator* among all the asymptotically normal consistent estimators of a population parameter.

Check Your Progress 2

1) What are the sources of non-sampling error?

……

……

……

……

……

……

……

2) Define the following concepts:

a) Unbiasedness estimator

b) Minimum variance estimator

c) Consistent estimator

……

……

……

……

…….....….....

3) State whether the following statements are true or false.

a) Normal distribution is a limiting case of binomial distribution.

b) Standard deviation of sampling distribution of a statistic is termed as standard error.

c) Poisson distribution is an example of continuous distribution.

d) Statistical estimation is a part of statistical inference.

……

……

……

……

13.6 CONCEPT OF STATISTICAL INFERENCE

As mentioned earlier, statistical inference deals with the methods of drawing conclusions about the population characteristics on the basis of information contained in a sample drawn from the population. Remember that population mean is not known to us, but we know the sample mean. In statistical inference we would be interested in answering two types of questions. First, what would be the value of the population mean? The answer lies in making an informed guess about the population mean. This aspect of statistical inference is called 'estimation'. The second question pertains to certain assertion made about the population mean. Suppose a manufacturer of electric bulbs claims that the mean life of electric bulbs is equal to 2000 hours. On the basis of the sample information, can we say that the assertion is not correct? This aspect of statistical inference is called hypothesis testing.

Thus statistical inference has two aspects: estimation and hypothesis testing. We will discuss about statistical estimation in this unit while testing of hypothesis will be taken up for discussion in the next unit.

Fig. 13.1 below summarises different aspects of statistical inference. A crucial factor before us is whether we know population variance or not. Of course when we do not know the population mean, how do we know the population variance? We begin with the case where population variance is known, because it will help us in explaining the concepts. Later on we will take up the more realistic case of unknown population variance.

Estimation could be of two types: point estimation and interval estimation. In point estimation we estimate the value of the population parameter as a single point. On the other hand, in the case of interval estimation we estimate lower and upper bounds around sample mean within which population mean is likely to remain.

The assertion or claim made about the population mean would be in the form of a null hypothesis and its counterpart, alternative hypothesis. We will explain these concepts and the methods of testing of hypothesis in the next unit.

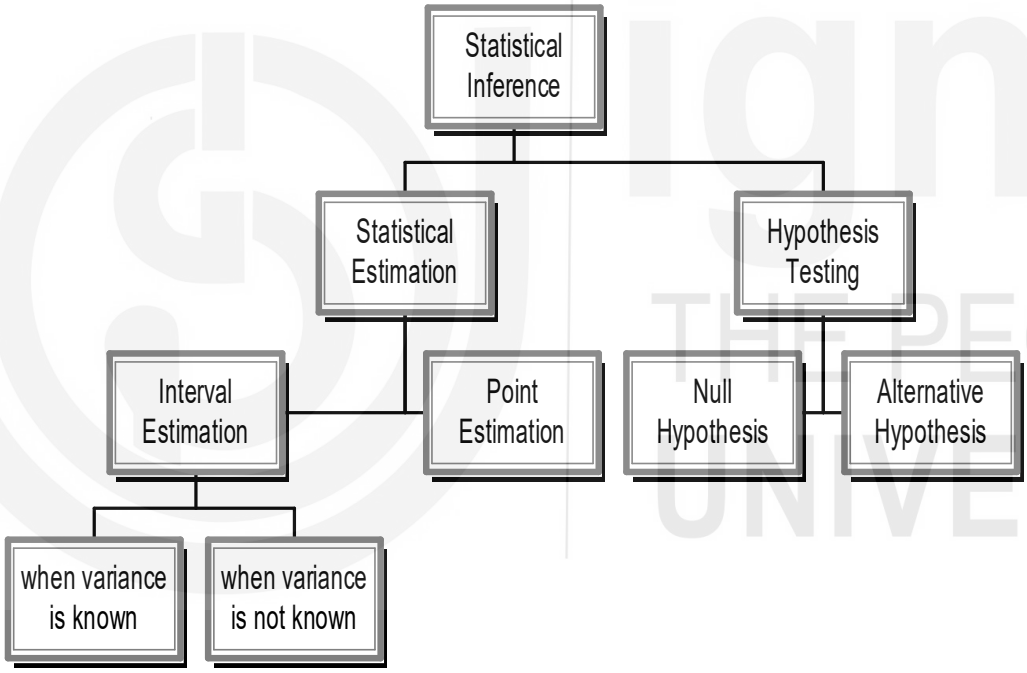


Fig. 13.1: Statistical Inference

13.7 POINT ESTIMATION

As mentioned earlier we do not know the parameter value and want to guess it by using sample information. Obviously the best guess will be the value of the sample statistic.

For example, if we do not know the population mean the best guess would be the sample mean. Here in this case we use a single value or point as 'estimate' of the parameter.

Recall that estimator is the formula and estimate is the particular value obtained by using the formula. For example, if we use sample mean for estimation of population mean, then $\frac{1}{n}\sum_i x_i$ is the estimator. Suppose I collect data on a sample, and put the sampling units to this formula and obtain a particular value for sample mean, say 120. Then 120 is an estimate of population mean. It is possible that you draw another sample from the same population, use the formula for sample mean, that is $\frac{1}{n}\sum_i x_i$, and obtain a different value, say 123. Here both 120 and 123 are estimates of population mean. But in both the cases the estimator is the same, which is $\frac{1}{n}\sum_i x_i$. Remember that the term statistic, which is used to mean a function of sample values, is a synonym for estimator.

There may be situations when you would find more than one potential estimator (alternative formulae) for a parameter. In order to choose the best among these estimators, we need to follow certain criteria. Based on these criteria an estimator should fulfill certain desirable properties. There are quite a few desirable properties for an estimator, but the most important is its unbiasedness.

Unbiasedness means that an estimate may be higher or lower than the unknown value of the parameter. But the expected value of the estimate should be equal to the parameter. For example, sample mean ($\bar{x}$) may fluctuate from sample to sample but on an average it would be equal to population mean. In other words, $E(\bar{x}) = \mu$.

However, $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is not an unbiased estimator of the population variance $\sigma^2 = \frac{1}{N}\sum(X_i - \bar{X})^2$. In fact, if we define $s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$, then s^2 is an unbiased estimator of σ^2. Usually a sample is less dispersed than the population from which it is drawn. Therefore, there is a tendency for the sample standard deviation *s* to be little less than population standard deviation σ. In order to rectify this condition we artificially inflate *s* by dividing by a smaller number (*n-1)*, instead of *n*.

The point estimate is very important for testing of hypothesis, as we will see in Unit 14.

13.8 INTERVAL ESTIMATION

We have seen above that in point estimation, we estimate the parameter by a single value, usually the corresponding sample statistic. The point estimate may not be realistic in the sense that the parameter value may not exactly be equal to it.

An alternative procedure is to give an interval, which would hold the parameter with certain probability. Here we specify a lower limit and an upper limit within which the parameter value is likely to remain. Also we specify the probability of

the parameter remaining in the interval. We call the *interval* as 'confidence interval' and the *probability* of the parameter remaining within this interval as 'confidence level' or 'confidence coefficient'.

13.8.1 Confidence Interval

Let us take an example. Suppose you are asked to estimate the average income of people in Raigarh district of Chhatisgarh state. You collected data from a sample of 500 households and found the average income (say, $\bar{x}$) to be Rs. 18250 per annum. This sample average may not be equal to the actual average income of Raigarh district of Chhatisgarh (μ) because of sampling error. Thus we are not sure whether average income of the above district is Rs. 18250 or not. On the other hand, it will be more sensible if we say that average income of Raigarh district of Chhatisgarh is between Rs. 17900 and Rs. 18600 per annum. Also we may specify that the probability that average income will remain within these limits is 95 per cent. Thus our confidence interval in this case is Rs. 17900-18600 and the confidence level or confidence coefficient is 95 per cent.

Here a question may be shaping up in your mind, 'How do we find out the confidence interval and confidence coefficient?' Let us begin with confidence coefficient. We know that the sampling distribution of $\bar{x}$ for large samples is normally distributed with mean μ and standard error $\frac{\sigma}{\sqrt{n}}$, where n is the size of the sample. By transforming the sampling distribution ($z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$) we obtain *standard normal variate*, which has zero mean and unit variance. The standard normal curve is symmetrical and therefore, the area under the curve for $0 \le z \le \infty$ is 0.5 which is presented in the form of a table (See Appendix Table A.1 given at the end of the book). Let us assume that we want our confidence coefficient to be 95 per cent (that is, 0.95). Thus we should find out a range for z which will cover 0.95 area of the standard normal curve. Since distribution of z is symmetrical, 0.475 area should remain to the right and 0.475 area should remain to the left of $z = 0$. If look into normal area table (see Appendix Table A.1) we find that 0.475 area is covered when z=1.96. Thus the probability that z ranges between –1.96 to 1.96 is 0.95. From this information let us work out backward and find the range within which μ will remain.

We find that

$$P(-1.96 \le z \le 1.96) = 0.95 \qquad \ldots(13.1)$$

$$or\ P\left(-1.96 \le \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \le 1.96\right) = 0.95$$

$$or\ P\left(-1.96\frac{\sigma}{\sqrt{n}} \le \bar{x} - \mu \le 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$or\ P\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95 \quad \ldots(13.2)$$

Let us interpret the above. Recall that each sample would provide us with a different value of $\bar{x}$. Accordingly, the confidence interval would be different. In each case the confidence interval would contain the unknown parameter or it would not. Equation (13.2) means that if a large number of random samples, each of size *n*, are drawn from the given population and if for each such sample the interval $\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \le \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$ is determined, then in about 95% of the cases, the interval will include the population mean μ.

The confidence coefficient is denoted by $(1-\alpha)$ where α is the level of significance (we will discuss the concept of 'level of significance' in Unit 14). Confidence coefficient could take any value. We can ask for a confidence level of say 81 per cent or 97 per cent depending upon how precise our conclusions should be. However, conventionally two confidence levels are frequently used, namely, 95 per cent and 99 per cent. Of course at times we take 90 per cent confidence level , though not frequently.

Let us find out the confidence interval when confidence coefficient $(1-\alpha)$ = 0.99. In this case 0.495 area should remain on either side of the standard normal curve. If we look into the normal area table (Table A.1) we find that 0.495 area is covered when z = 2.58.

Thus

$$P(-2.58. \le z \le 2.58) = 0.99 \quad \ldots(13.3)$$

By rearranging the terms in the above we find that

$$P\left(\bar{x} - 2.58\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + 2.58\frac{\sigma}{\sqrt{n}}\right) = 0.99 \quad \ldots(13.4)$$

Equation (13.4) implies that 99 per cent confidence interval for μ is given by $\bar{x} \pm 2.58\frac{\sigma}{\sqrt{n}}$.

By looking into the normal area table you can work out the confidence interval for confidence coefficient of 0.90 and find that

$$P\left(\bar{x} - 1.65\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + 1.65\frac{\sigma}{\sqrt{n}}\right) = 0.90 \quad \ldots(13.5)$$

We observe from (13.2), (13.4) and (13.5) that as the interval widens, the chance for the interval holding a population parameter (in this case μ) increases.

13.8.2 Confidence Limits

The two limits of the confidence interval are called *confidence limits*. For example, for 95 per cent confidence level we have the lower confidence limit as $\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}$ and upper confidence limit as $\bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$. The confidence coefficient

can be interpreted as the confidence or trust that we place in these limits for actually holding μ.

Example 13.3

A paper company wants to estimate the average time required for a new machine to produce a ream of paper. A random sample of 36 reams shows an average production time of 1.5 minutes per ream of paper. The population standard deviation is known to be 0.30 minute. Construct an interval estimate with 95% confidence limits.

Solution: The information given is

$$\bar{x} = 1.5,\ \sigma = 0.30 \text{ and } n = 36$$

Since n = 36 (> 30) , we can take the sample as a large sample and accordingly $\bar{x}$ is normally distributed with mean μ and standard error $\frac{\sigma}{\sqrt{n}}$. Now, the standard error is

$$\frac{\sigma}{\sqrt{n}} = \frac{0.30}{\sqrt{36}} = 0.05$$

The 95% confidence interval is given by

$$\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$$

$$or\ 1.5 - 1.96 \cdot 0.05 \le \mu \le 1.5 + 1.96 \cdot 0.05$$

$$i.e.,\ 1.402 \le \mu \le 1.598$$

Thus with 95% confidence, we can state that the average production time for the new machine will be between 1.402 minutes and 1.598 minutes. Here, 1.402 is the 'lower confidence limit' and 1.598 is the 'upper confidence limit'.

13.8.3 Confidence Interval for Unknown Variance

We estimated confidence interval for population mean on the assumption that population variance is known. It is a bit unrealistic that we do not know population mean (we want to estimate it) but know population variance. A realistic case would be the assumption that both population mean and variance are unknown. On the basis of sample mean and variance we want to find out confidence interval for population mean.

Since the population standard deviation (σ) is not known we use the sample standard deviation (s) in its place.

However, in such a case the sampling distribution of $\bar{x}$ is not normal, rather it follows student's t distribution.

The standard error of the sample means would be

$$\frac{s}{\sqrt{n}}.$$

Like the standard normal variate, z, the *t*-distribution has a mean of zero, is symmetrical about mean and ranges between $-\infty$ to ∞. But its variance is greater than 1. Actually its variance changes according to degrees of freedom. However, when $n > 30$ the *t*-distribution has a variance very close to 1 and thus resembles *z*-distribution.

The t-statistic, like the *z*-statistic, is calculated as

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

By looking into the area table for *t*-distribution (see Appendix Table A.3) we find the probability values for the confidence level that we require. The degrees of freedom is $(n-1)$. Thus the confidence interval would be

$$\bar{x} - t \cdot \frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t \cdot \frac{s}{\sqrt{n}} \qquad \ldots(13.6)$$

Example 13.4

The mean weight (in kilogram) of 20 children are found to be 15 with a standard deviation of 4. On the basis of the above information estimate 95 per cent confidence interval for mean weight of the population from which the sample is drawn. Assume that population is normally distributed.

Solution: Since population is normal and sample size is small we apply *t*-distribution for estimation of confidence interval. Since $n = 20$ we have degrees of freedom (d.f.) = 19. We move down the first column of Appendix Table A.3 till we reach the row corresponding to d.f.= 19. Since we need 95 per cent confidence interval we should leave 0.025 area on each side of $t = 0$ as we did in the previous Section. Thus for 19 degrees of freedom and $\alpha = 0.025$ we find that *t*-value is 2.093.

Hence the confidence interval is

$$15 - 2.093 \times \frac{4}{\sqrt{20}} \le \mu \le 15 + 2.093 \times \frac{4}{\sqrt{20}}$$

or $\quad 15 - 1.87 \le \mu \le 15 + 1.87$

or $\quad 13.13 \le \mu \le 16.87$

Similarly, you can find out confidence intervals for different sample sizes and confidence coefficients.

Let us summarise the rules for application of z or *t* statistic for estimation of confidence interval.

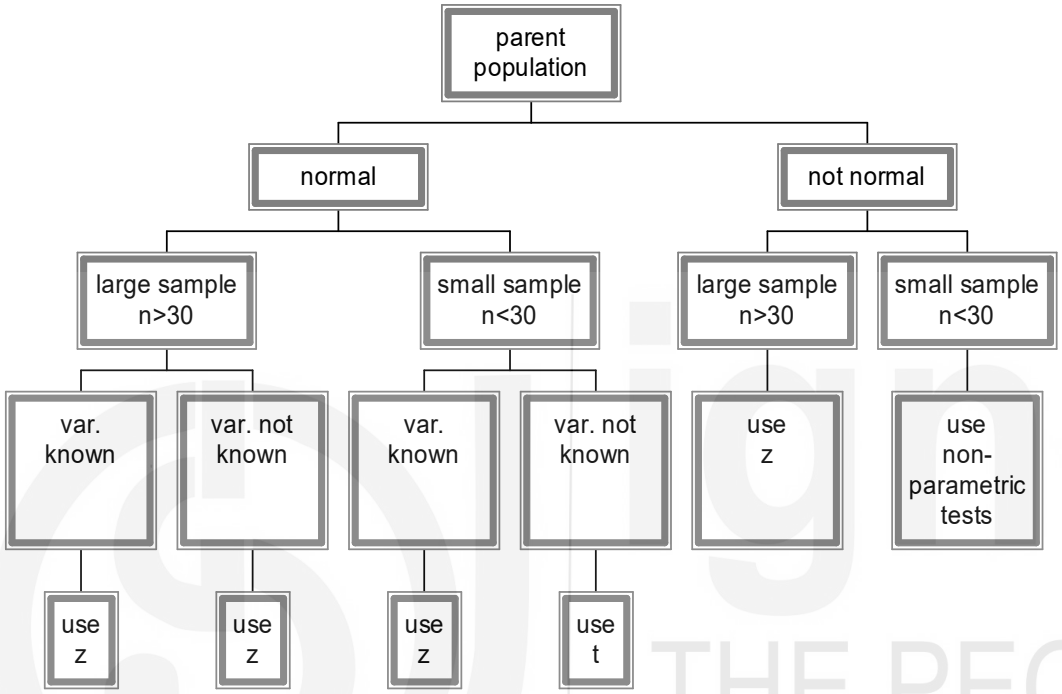
1. If sample size is large (n>30) apply z-statistic -- it does not matter whether (i) parent population is normal or not, and (ii) variance is known or not.
2. If sample size is small ($n \le 30$) check whether (i) parent population is normal, and (ii) variance is known.

a) If parent population is not normal apply nonparametric tests.

b) If parent population is normal and variance is known apply z.

c) If parent population is normal and variance is not known apply t.

In Fig. 13.2 we present the above in the form of a chart.



1.1 Fig. 13.2: Selection of Proper Test Statistic

2 Check Your Progress 3

1. A sample of 50 employees were asked to provide the distance commuted by them to reach office. If sample mean was found to be 4.5 km. Find 95 percent confidence interval for the population. Assume that population is normally distributed with a variance of 0.36.

   ……

   ……

   ……

2. For a sample of 25 students in school the mean height was found to be 95 cm. with a standard deviation of 4 cm. Find the 99 percent confidence interval.

   ……

   ……

   ……

3. State whether the following statements are True or False.

a) When parent population is not normal and sample size is small we use t-distribution to estimate confidence interval.

b) The range of *t*-distribution is 0 to infinity.

c) When confidence level is 90 per cent, level of significance is 10 per cent.

……

……

13.9 LET US SUM UP

Drawing conclusions about a population on the basis of sample information is called statistical inference. Here we have basically two things to do: estimation and hypothesis testing. In this unit we took up the first issue while the send one will be discussed in the next unit.

An estimate of an unknown parameter could be either a point or an interval. Sample mean is usually taken as a point estimate of population mean. On the other hand, in interval estimation we construct two limits (upper and lower) around the sample mean. We can say with stipulated level of confidence that the population mean, which we do not know, is likely to remain within the confidence interval. In order to construct confidence interval we need to know the population variance or its estimate. When we know population variance, we apply normal distribution to construct the confidence interval. In cases where population variance is not known, we use student's *t* for the above purpose. Remember that when sample size is large (n>30) *t*-distribution approximates normal distribution. Thus for large samples, even if population variance is not known, we can use normal distribution to confidence interval on the basis of sample mean and sample variance.

13.10 ANSWERS/HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

Go through Section 13.3 and define these terms in one or two sentences each.

1) Read the text and distinguish between these terms in a few sentences each.

2) Go through Section 13.3 and explain.

3) a) Identify all possible samples; calculate sample mean, arrange the sample means in the form of a frequency distribution; find out probability of occurrence of each sample mean.

b) Explain what the term means. Use the formula $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. Your answer should 0.94.

Check Your Progress 2

1) Go through Section 13.4 and answer.
2) Read Section 13.5 and answer.
3) a) True b) True c) False d) True

Check Your Progress 3

1. Since it is large sample we apply z-statistic. The confidence interval is

   $4.40 \leq \mu \leq 4.60$

2. Since it is small sample and population variance is not given we apply *t*-statistics with degrees of freedom 24. The tabulated value of *t* at 99

   percent confidence level is 2.49. The confidence interval is
   $93.01 \leq \mu \leq 96.99$.

3. a) False b) False c) True

UNIT 14 TESTING OF HYPOTHESIS*

Structure

14.0 OBJECTIVES

After going through this Unit you will be in a position to

- explain the concepts of null hypothesis and alternative hypothesis;
- identify critical region based on the level of significance;
- distinguish between Type I and Type II errors;
- test for hypothesis concerning population mean on the basis of one sample; and
- test for the difference in sample means obtained from two samples.

14.1 INTRODUCTION

In the previous Unit we learnt about the estimation of confidence interval for population mean on the basis of sample data. In the present Unit we will look into another aspect of statistical inference, that is, hypothesis testing. Hypothesis is a statement or assertion or claim about the population parameter. For example, suppose we have a hitch that the per capita income of Chhatisgarh state is Rs. 20000 per annum. We can be sure about the truth in the above statement if we undertake a complete census of households in the state.

* Prof. Kaustuva Barik, Indira Gandhi National Open University, New Delhi.

This implies we collect data on the income of all the households in Chhatisgarh and calculate the per capita income of the state. However, because of constraints such as time, money and manpower may restrict us to go for a sample survey and reach a conclusion about the statement on the basis of sample information. The procedure followed in the above is the subject matter of hypothesis testing.

Hypothesis testing is applied widely in various fields and to various situations. For example, suppose the effectiveness of a new drug in curing tuberculosis needs to be tested. Obviously all the patients suffering from tuberculosis need not be administered with the new drug to see its effectiveness. What we need is a representative sample and test whether the new drug is more effective than existing drugs. As another example let us take the case of planner who asserts that the crude birth rate is the same in the states Bihar and Rajasthan. In this case it may not be possible on our part to go for a census survey of all the births that have taken place in Bihar and Rajasthan during the last year and calculate the crude birth rate. Instead a sample survey is undertaken and the assertion made by the planner is put to test.

In hypothesis testing we try to answer questions of the following types: Is the sample under consideration is drawn from a particular population? Is the difference between two samples significant enough so that they cannot belong to the same population?

14.2 FORMULATION OF A HYPOTHESIS

A hypothesis is a tentative statement about a characteristic of a population. It could be an assertion or a claim also. For example, official records for recent years show that female literacy in Orissa is 51 per cent. Here a statement or a claim about the rate of female literacy is being made. Thus it could be considered as a hypothesis.

In hypothesis testing there are four important components: i) null hypothesis, ii) alternative hypothesis, iii) test statistic, and iv) interpretation of results. We discuss each of these below.

Usually statistical hypotheses are denoted by the alphabet *H*. There are two types of hypothesis: null hypothesis and alternative hypothesis. A null hypothesis is the statement that we consider to be true about the population and put to test by a test statistic. Usually we denote null hypothesis by H_0. In the example on female literacy in Orissa our null hypothesis is

$$H_0 : \mu = 0.51 \qquad \ldots(14.1)$$

where μ is the parameter, in this case female literacy in Orissa.

There is a possibility that the null hypothesis that we intend to test is not true and female literacy is not equal to 51 per cent. Thus there is a need for an alternative hypothesis which holds true in case the null hypothesis is not true.

We denote alternative hypothesis by the symbol H_A and formulate it as follows:

$$H_A : \mu \neq 0.51 \qquad \ldots(14.2)$$

We have to keep in mind that null hypothesis and alternative hypothesis are mutually exclusive, that is, both cannot be true simultaneously. Secondly, both H_0 and H_A exhaust all possible options regarding the parameter, that is, there cannot be a third possibility. For example, in the case of female literacy in Orissa, there are two possibilities - literacy rate is 51 per cent or it is not 51 per cent; a third possibility is not there.

It is a rare coincidence that sample mean ($\bar{x}$) is equal to population mean (μ). In most cases we find a difference between $\bar{x}$ and μ. Is the difference because of sampling fluctuation or is there a genuine difference between the sample and the population? In order to answer this question we need a test statistic to test the difference between the two. The result that we obtain by using the test statistic needs to be interpreted and a decision needs to be taken regarding the acceptance or rejection of the null hypothesis.

The development of test statistic for hypothesis testing and interpretation of results requires elaboration. Before discussing further on these two steps we present another concept -- critical or rejection region.

14.3 REJECTION REGION AND TYPE OF ERRORS

The underlying idea behind hypothesis testing and interval estimation (discussed in the previous Unit) is the same. Recall from Unit 18 that a confidence interval is built around sample mean with certain confidence level. A confidence level of 95 per cent implies that in 95 per cent cases the population mean would remain in the confidence interval estimated from the sample mean. It is implicit that in 5 per cent cases the population mean will not remain within the confidence interval. Note that when the population mean does not remain within the confidence interval we should reject the null hypothesis.

14.3.1 Rejection Region for Large Samples

Let us explain the concept of critical region for large samples. Later on we will extend the concept to small samples.

As you already know from previous Units, sampling distribution of sample mean ($\bar{x}$) follows normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$. Thus, $\bar{x}$ can be transformed into a standard normal variable, z, so that it fol clows normal distribution with mean 0 and standard deviation 1.

In notations, $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ and $z \sim N(0,1)$.

Recall from Unit 11 that area under the standard normal curve gives the probability for different range of values assumed by z. These probabilities can also be presented in a table (see Appendix Table A1).

Let us look into the standard normal curve presented in Fig. 14.1, where the x-axis represents the variable z and the y-axis represents the probability of z, that is $p(z)$. We should note the following points.

- When sample mean is equal to population mean (that is, $\bar{x} = \mu$) we find that z = 0. When $\bar{x} > \mu$ we find that z is positive. On the other hand, when $\bar{x} < \mu$ we find that z is negative.
- Note that we are concerned with the difference between $\bar{x}$ and μ. Therefore, negative or positive sign of z does not matter much.
- Higher the difference between $\bar{x}$ and μ, higher is the absolute value of z. Thus z-value measures the discrepancy between $\bar{x}$ and μ, and therefore can be used as a test statistic.
- We should find out a critical value of z beyond which the difference between $\bar{x}$ and μ is significant.
- If the absolute value of z is less than the critical value we should not reject the null hypothesis.
- If the absolute value of z exceeds the critical value we should reject the null hypothesis and accept the alternative hypothesis.

  Thus in the case of large samples the absolute value of z can be considered as test statistic for hypothesis testing such that

$$z = \frac{|\bar{x} - \mu|}{\sigma/\sqrt{n}} \qquad \ldots(14.3)$$

Let us explain the concept of critical region through the standard normal curve given in Fig. 14.1 below. When we have a confidence coefficient of 95 percent, the area covered under the standard normal curve is 95 per cent. Thus 95 per cent area under the curve is bounded by $-1.96 \le z \le 1.96$. The remaining 5 per cent area is covered by $z \le -1.96$ and $z \ge 1.96$. Thus 2.5 per cent of area on both sides of the standard normal curve constitute the rejection region. This area is shown in Fig. 14.1. If the sample mean falls in the rejection region we reject the null hypothesis.

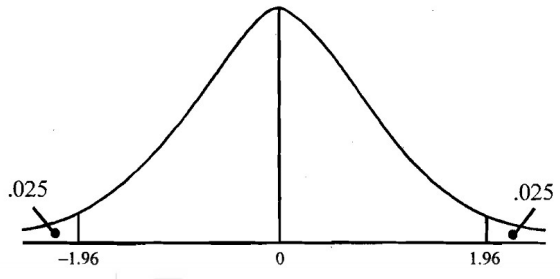


Fig. 14.1: Critical Regions

14.3.2 One-tail and Two-tail Tests

In Fig. 14.1 we have shown the rejection region on both sides of the standard normal curve. However, in many cases we may place the rejection region on one side (either left or right) of the standard normal curve.

Remember that if α is the level of significance, then for a two-tail test $\frac{\alpha}{2}$ area is placed on both sides of the standard normal curve. But if it is a one-tail test then α area is placed on one-side of the standard normal curve. Thus the critical value for one-tail and two tail test differ.

The selection of one-tail or two-tail test depends upon the formulation of the alternative hypothesis. When the alternative hypothesis is of the type $H_A : \bar{x} \neq \mu$ we have a two-tail test, because $\bar{x}$ could be either greater than or less than μ. On the other hand, if alternative hypothesis is of the type $H_A : \bar{x} < \mu$, then entire rejection is on the left hand side of the standard normal curve. Similarly, if the alternative hypothesis is of the type $H_A : \bar{x} > \mu$, then the entire rejection is on the right hand side of the standard normal curve.

The critical values for z depend upon the level of significance. In Table 14.1 these critical values for certain specified levels of significance (α) are given for the tests to be conducted under the assumption of normal distribution. The values are given for both two-tail and one-tail tests.

Table 14.1: Critical Values for z-statistic

Significance Level (α)	0.10	0.05	0.01	0.005
two-tail test	1.65	1.96	2.58	2.81
one-tail test	1.28	1.65	2.33	2.58

Note: The table is derived from Appendix Table A1.

14.3.3 Type I and Type II Errors

In hypothesis testing we reject or do not reject a hypothesis with certain degree of confidence. As you know, a confidence coefficient of 0.95 implies that in 95 out of 100 samples the parameter remains within the acceptance region and in 5 per cent cases the parameter remains in the rejection region. Thus in 5 per cent cases the sample is drawn from the population but sample mean is too far away from the population mean. In such cases the sample belongs to the population but our test procedure rejects it. Obviously we commit an error such that H_0 is true but gets rejected. This is called 'Type I error'. Similarly there could be situations when the H_0 is not true, but on the basis of sample information we do not reject it. Such an error in decision making is termed 'Type II error' (see Table 14.2).

Note that Type I error specifies how much error we are in a position to tolerate. Type I error is equal to the level of significance, and is denoted by α. Remember that confidence coefficient is equal to $1-\alpha$.

Table 14.2: Type of Errors

	H_0 **true**	H_0 **not true**
Reject H_0	Type I Error	Correct decision
Do not reject H_0	Correct decision	Type II Error

14.3.4 Rejection Region for Small Samples

Let us go back to Fig. 13.2 in Unit 13 where we have given certain criteria for use of proper test statistic in interval estimation. We see that in the case of small samples ($n \leq 30$), if population standard deviation is known we apply *z*-statistic for hypothesis testing. On the other hand, if population standard deviation is not known we apply *t*-statistic. The same criteria apply to hypothesis testing also.

In the case of small samples if population standard deviation is known the test statistic is

$$z = \frac{|\bar{x} - \mu|}{\sigma/\sqrt{n}} \quad \ldots(14.4)$$

On the other hand, if population standard deviation is not known the test statistic is

$$t = \frac{|\bar{x} - \mu|}{s/\sqrt{n}} \quad \ldots(14.5)$$

In the case of *t*-distribution, however, the area under the curve (which implies probability) changes according to degrees of freedom. Thus while finding the critical value of *t* we should take into account the degrees of freedom. When sample size is *n*, degrees of freedom is $n-1$. Thus we should remember two things while finding critical value of *t*. These are: i) significance level, and ii) degrees of freedom.

Check Your Progress 1

1. Distinguish between the following:

   a) Null hypothesis and Alternative hypothesis

   b) One-tail and Two-tail tests

   c) Confidence level and Level of significance

   d) Type I and Type II errors

……

……

……

……

……

2. Suppose a sample of 100 students has mean age of 12.5 years. Show through diagram the rejection region at 5 per cent level of significance to test the hypothesis that the sample has a mean age greater the population mean. Assume that population mean and standard deviation are 10 years and 2 years respectively.

……

……

……

……

……

14.4 TESTING OF HYPOTHESIS FOR A SINGLE SAMPLE

In many situations we are asked to judge whether a sample is significantly different from a given population. For example, let us assume that we surveyed a sample of 400 households of Raigarh district of Chhatisgarh state and calculated the per capita income of these households. Subsequently, our task is to test the hypothesis that per capita income calculated from the sample is not different from the per capita income of the district.

In the above example we can have two different situations: i) population (in this case all the households of the district) variance is known, ii) population variance is not known to us. We explain the steps to be followed in each case below.

14.4.1 Population Variance is Known

Let us consider the case that we know the per capita income of Raigarh district of Chhatisgarh as well as its variance. Suppose the data available in official records show that per capita income of Raigarh district is Rs. 10000 and standard deviation of per capita income is Rs. 1500. However, we did a sample survey of 400 households and found that their per capita income is Rs. 10500. Do we accept the data provided in official records?

In this case μ = Rs. 10000

σ = Rs. 1500

$\bar{x}$ = Rs. 10500

n = 400

From the central limit theorem we know that when sample size is large, sample is approximately normally distributed. This is true even in cases where the parent population is normally distributed. Thus this example is appropriate for application of normal distribution.

Our null hypothesis in this case is

$$H_0 : \bar{x} = \mu$$

The null hypothesis suggests that sample mean is equal to population mean. In other words, per capita income obtained from the sample is the same as the data provided in official records.

Our alternative hypothesis is

$$H_A : \bar{x} \neq \mu$$

Suppose we do not have any reason to say that per capita income obtained from the sample ($\bar{x}$) is greater than or small than the per capita income available in official records. Thus our alternative hypothesis is that $\bar{x}$ could be on either side of μ. Therefore, we should go for two-tail test so that rejection region is on both sides of the standard normal curve and the test statistic is

$$z = \frac{|\bar{x} - \mu|}{\sigma/\sqrt{n}}. \qquad \ldots(14.5)$$

By substituting values in the above we obtain

$$z = \frac{|10500 - 10000|}{1500/\sqrt{400}} = \frac{500}{500/20} = \frac{500}{75} = 6.67$$

Recall the standard normal curve and the area for different values of z (See Table 15.1 in Unit 15). We notice that when z = 1.96 the area covered under standard normal curve is 0.4750. Therefore, the level of significance is 5 per cent. Similarly when z = 2.58 the area covered under standard normal curve is 0.4950. therefore, the level of significance is 1 per cent.

In the above case since z = 6.67, the sample lies in the critical region and we reject the hypothesis. Thus the per capita income obtained from the sample is significantly different from the per capita income provided in official records.

The steps you should follow are:

1. Specify the null hypothesis.
2. Find out whether it requires one-tail or two-tail test. Accordingly identify your critical region. This will help in specification of alternative hypothesis.
3. Apply sample values to *z*-statistic given at (14.5).
4. Find out from *z*-table the critical value according to level of significance.
5. If you obtain a value lower than the critical value do not reject the null hypothesis.
6. If you obtain a value greater than the critical value reject the null hypothesis and accept the alternative hypothesis

Example 14.1

Suppose the voltage generated by certain brand of battery is normally distributed. A random sample of 100 such batteries was tested and found to have a mean voltage of 1.4 volts. At 0.01 level of significance, does this indicate that these batteries have a general average voltage, that is different from 1.5 volts? Assume that population standard deviation is 0.21 volts.

Here, $H_0: \mu = 1.5$

Since average voltage of the sample can be different from average voltage of the population if it is either less than or more than 1.5 volts, our rejection region is on both sides of the normal curve. Thus it is a case of two-tail test and alternative hypothesis is

$H_1: \mu \neq 1.5$

Since the population standard deviation σ is known, the test statistic is

$$z = \frac{|\bar{x} - \mu|}{\frac{\sigma}{\sqrt{n}}} = \frac{|1.4 - 1.5|}{\frac{0.21}{\sqrt{100}}} = 4.8$$

From the table for the area under the standard normal curve, we find that the critical value at the 1 per cent level significance is 2.58. Since the observed value of z is greater than 2.58 we reject the null hypothesis at 1% level and accept the alternative hypothesis that the average life of batteries is different from 1.5 volts.

14.4.2 Population Variance not Known

The assumption that population standard deviation (σ) is known to us is unrealistic, as we do not know population mean itself. When σ is unknown we have to estimate it by sample standard deviation (s). In such situations there are two possibilities depending upon the sample size. If the sample size is large ($n > 30$) we apply z-statistic, that is,

$$z = \frac{|\bar{x} - \mu|}{s/\sqrt{n}} \qquad \ldots(14.6)$$

In case the sample size is small ($n \leq 30$) we apply t-statistic with $n - 1$ degrees of freedom. The test statistic is

$$t = \frac{|\bar{x} - \mu|}{s/\sqrt{n}} \qquad \ldots(14.7)$$

The steps you should follow are:

1. Specify the null hypothesis.
2. Find out whether it requires one-tail or two-tail test. Accordingly identify the rejection region in the standard normal curve. This will help in specification of alternative hypothesis.

3. Check whether sample size is large ($n > 30$) or small ($n \leq 30$).
4. In case $n > 30$, apply *z*-statistic given at (14.6).
5. Find out from z-table the critical value according to level of significance (α).
6. In case $n \leq 30$, apply *t*-statistic given at (14.7).
7. Find out from *t*-table (given in Table 15.3 in Block 5) the critical value for $n-1$ degrees of freedom and level of significance (α).
8. If you obtain a value lower than the critical value do not reject the null hypothesis.
9. If you obtain a value greater than the critical value reject the null hypothesis and accept the alternative hypothesis

Example 14.2

A tablet is supposed to contain on an average 10 mg. of aspirin. A random sample of 100 tablets show a mean aspirin content of 10.2 mg. with a standard deviation of 1.4 mg. Can you conclude at the 0.05 level of significance that the mean aspirin content is indeed 10 mg.?

Here, the null hypothesis is $H_0: \mu = 10$

The rejection region is on both sides of 10 mg. Thus it requires a two-tail test and $H_A : \mu \neq 10$.

Also, the sample mean is $\bar{x} = 10.2$ and the sample size n = 100. Since population standard deviation is not known we estimate it by sample standard deviation s and our test statistic is $z = \dfrac{|\bar{x} - \mu|}{s/\sqrt{n}}$. By applying relevant values from the sample we obtain

$$z = \frac{|\bar{x} - \mu|}{\frac{s}{\sqrt{n}}} = \frac{|10.2 - 10|}{\frac{1.4}{\sqrt{100}}} = 1.43$$

At 5 per cent level of significance the critical value of z is 1.96. since the z value that we have obtained is less than 1.96, we do not reject the null hypothesis. Therefore the mean level of aspirin is 10 mg.

Example 14.3

The population of Haripura district has a mean life expectancy of 60 years. Certain health care measures are undertaken in the district. Subsequently, a random sample of 25 persons shows an average life expectancy of 60.5 years with a standard deviation of 2 years. Can we conclude at the 0.05 level of significance that the average life expectancy in the district has indeed gone up?

Here, $H_0: \mu = 60$

We have to test for an increase in life expectancy. Thus it is a case of one-tail test and the rejection region will be on the right-hand tail of the standard normal curve.

Hence our alternative hypothesis is

$H_1: \mu > 60$

Here population standard deviation σ is not known and we estimate it by the sample standard deviation s. Here the sample size is small hence we have to apply t-statistic given at (14.7).

$$t = \frac{|\bar{x} - \mu|}{\frac{s}{\sqrt{n}}} = \frac{|60.5 - 60|}{\frac{2}{\sqrt{25}}} = 1.25$$

Since sample size is 25, degrees of freedom is 25 – 1 = 24. From the t-table we find that for 24 degrees of freedom, 5 per cent level of significance, and one-tail test the t-value is 1.71.

Since t-value obtained above is less than the critical value we do not reject the hypothesis. Therefore, life expectancy has not changed Therefore, we accept the alternative hypothesis that life expectancy for the district has not changed after the health care measures.

Check Your Progress 2

1. A report claimed that in the 'School Leaving Examination', the average marks scored in Mathematics was 78 with a standard deviation of 16. However, a random sample of 37 students showed an average of 84 marks in Mathematics. In the light of this evidence, can we conclude that actually the average was more than 78? Use 0.05 level of significance.

   ……

   ……

   ……

   …….....

2. A passenger car company claims that average fuel efficiency of cars is 35 kms per litre of petrol. A random sample of 50 cars shows an average of 32 kms per litre with a standard deviation of 1.2 kms. Does this evidence falsify the claim of the passenger car company at 0.01 level of significance?

   ……

   ……

   ……

   ……

   …….....

3. A random sample of 200 tins of coconut oil gave an average weight of 4.95 kg per tin with a standard deviation of 0.21 kg. Do we accept the hypothesis of net weight of 5 kg per tin at 0.01 level of significance?

   ……

   ……

   ……

   ……

   …….....

4. According to a report, the national average annual income of the government employees during a recent year was Rs. 24,632 with a standard deviation of Rs. 1827. A random sample of 49 government employees during the same year showed an average annual income of Rs. 25,415. On the evidence of this sample, at 0.05 level of significance, Can we conclude that the national average annual income of government employees was indeed Rs. 24,632?

   ……

   ……

   ……

   ……

   …….....

14.5 TESTS FOR DIFFERENCE BETWEEN TWO SAMPLES

Many times we need to test for the difference between two samples. The objective may be to ascertain whether both samples are drawn from the same population or to check whether a particular characteristic is the same in two populations. For example, we formulate a hypothesis that the production per worker in plant A is the same as the production per worker in plant B. We discuss below the procedure for testing of such a hypothesis.

Here again we deal with two different situations: whether variance of both the populations are known. Another consideration is sample size: large or small.

The null hypothesis is the statement that population means of both the populations are the same. In notations

$$H_0 : \mu_1 = \mu_2 \qquad \ldots(14.8)$$

The alternative hypothesis is the statement that both the population means are different. In notations

$$H_A : \mu_1 \neq \mu_2 \qquad \ldots(14.9)$$

14.5.1 Population Variance is Known

When standard deviations (positive square root of variance) of both the populations are known we apply z statistic specified as follows:

$$z = \frac{\left|(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)\right|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad \ldots(14.10)$$

In (14.10) above, subscript 1 refers to the first sample and subscript 2 refers to the second sample. By applying relevant data in (14.10) we obtain the observed value of z and compare it with the critical value for specified level of significance.

Example 14.4

A bank wants to find out the average savings of its customers in Delhi and Kolkata. A sample of 250 accounts in Delhi shows an average savings of Rs. 22500 while a sample of 200 accounts in Kolkata shows an average savings of Rs. 21500. it is known that standard deviation of savings in Delhi is Rs. 150 and that in Kolkata is Rs. 200. Can we conclude at 1 per cent level of significance that banking pattern of customers in Delhi and Kolkata is the same?

In this case the null hypothesis is $H_0 : \mu_1 = \mu_2$

and the alternative hypothesis is $H_A : \mu_1 \neq \mu_2$

We are provided with the information that

$\bar{x}_1$ = Rs. 22500 $\quad\quad$ σ_1 = Rs. 150

$\bar{x}_2$ = Rs. 22400 $\quad\quad$ σ_2 = Rs. 200

n_1 = 250 $\quad\quad$ n_2 = 200

Since σ_1 and σ_2 are known we apply z-test.

The test statistic is $z = \frac{\left|(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)\right|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$.

By applying the information provided above we obtain

$$z = \frac{\left|22500 - 22400\right|}{\sqrt{\frac{150^2}{250} + \frac{200^2}{200}}} = \frac{100}{\sqrt{90 + 200}} = 5.87$$

We find that at 1 per cent level of significance the critical value obtained from Table 14.2 is 2.58.

Since the observed value of *t* is greater than the critical value of *t* the null hypothesis is rejected and the alternative hypothesis is accepted. Thus the banking pattern of customers in Delhi and Kolkata are different.

14.5.2 Population Variance is not Known

When population standard variance (σ^2) is not known we estimate it by sample standard variance (s^2). If both samples are large in size (n>30) then we apply z statistic as follows:

$$z = \frac{\left|(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)\right|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

…(14.11)

On the other hand, if samples are small in size ($n \leq 30$) then we apply t-statistic as follows:

$$t = \frac{\left|(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)\right|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \qquad \ldots(14.12)$$

Degrees freedom for t-test = $(n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$

Example 14.5

A mathematics teacher wants to compare the performance of Class X students in two sections. She administers the same set of questions to 25 students in Section A and 20 students in Section B. She finds that Section A students have a mean score of 78 marks with standard deviation of 4 marks while Section B students have a mean score of 75 marks with standard deviation of 5 marks. Is the performance of students in both Sections different at 1 per cent level of significance?

In this case the null hypothesis is $H_0 : \mu_1 = \mu_2$

and the alternative hypothesis is $H_A : \mu_1 \neq \mu_2$

We are provided with the information that

$\bar{x}_1 = 78$ $\quad$ $s_1 = 4$

$\bar{x}_2 = 75$ $\quad$ $s_2 = 5$

$n_1 = 25$ $\quad$ $n_2 = 20$

Since σ_1 and σ_2 are not known and sample sizes are small we apply t-test.

$$t = \frac{\left|(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)\right|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{\left|78 - 75\right|}{\sqrt{\frac{4^2}{25} + \frac{5^2}{20}}} = \frac{3}{1.37} = 2.18$$

The degrees of freedom in this case is 25+20-2 = 43.

We find out from Table 15.3 that at 1 per cent level of significance the t-value for 43 degrees of freedom is 2.69.

Since the critical value of *t* is less than the observed value of *t* we reject the null hypothesis and accept the alternative hypothesis. Therefore, students in Section A and Section B are different with respect to their performance in mathematics.

Check Your Progress 3

1. You are given the following information.

   $n_1=50$ $n_2=50$

   $\bar{x}_1=52.3$ $\bar{x}_2=52.3$

   $\sigma_1 = 6.1$ $\sigma_2 = 6.1$

   Test the following hypothesis.

   $H_o : \mu_1 = \mu_2$

   $H_A : \mu_1 \neq \mu_2$

   $\alpha = 0.05$

   ……

   ……

   ……

   ……...

2. From two normal populations two samples are drawn. The following information is obtained.

   $n_1=15$ $n_2=10$

   $\bar{x}_1=140$ $\bar{x}_2=150$

   $s_1 = 10$ $s_2 = 15$

   Test the hypothesis at 1% level of significance that there is no difference between both the populations.

   ……

   ……

   ……

   ……...

3. Suppose that samples of size $n_1=20$ and $n_2=15$ are drawn from two normal populations. The sample statistics are as follows:

   $\bar{x}_1=110$ $\bar{x}_2=125$

   $s_1^2 = 225$ $s_2^2 = 150$

   Can we conclude at the 5% level of significance that $\mu_1 < \mu_2$?

   ……

   ……

   ……

   ……

   ……..

14.6 LET US SUM UP

In the present Unit we discussed about the methods of testing a hypothesis and drawing conclusions about the population. Hypothesis is a statement about the population parameter. In order to test a hypothesis we formulate test statistic on the basis of the information available to us. In this Unit we considered two situations: i) description of a single sample, and ii) comparison between two samples.

Construction of the test statistic depends on the knowledge about population variance and sample size. When population variance is known to us or the sample size is large we apply normal distribution and use *z* statistic to test the hypothesis. On the other hand, when we do not know the population variance and sample size is small we construct the test statistic on the basis of *t* distribution. Remember that for large samples *t* distribute on approximates normal distribution and therefore we can use *z* statistic.

14.11 ANSWERS/HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1. Go through Sections 18.2 and 18.3, and answer.
2. It is large sample and σ is unknown. It requires one-tail test. Thus rejection region is to the right hand tail of standard normal curve. Accordingly draw the diagram.

Check Your Progress 2

1. Since it is large sample with known variance, we apply z-statistic. Since alternative hypothesis is $\mu > 78$, we apply one-tail test. The observed value of z is 2.28 and critical value of z at 5% level of significance is 1.65. Since the observed value is greater than the critical value we reject the null hypothesis. Therefore, we conclude that the average marks was more than 78.
2. It is a large sample with unknown variance. It requires two-tail test. The observed value of z is 17.68 and critical value of z at 1% level of significance is 2.58. Since the observed value is greater than the critical value, the null hypothesis is rejected.
3. It is a large sample with unknown variance. Requires two-tail test with z-statistic. Observed value of z is 3.37. Null hypothesis is rejected.
4. Since it is large sample with known standard deviation, we apply z-statistic. Requires two-tail test. Observed value of z is 3.00. critical value of z at 5% level of significance is 2.58. null hypothesis is rejected. Therefore, the national average of annual income of government employees was different from Rs. 24632.

Check Your Progress 3

1. The samples sizes are large and population standard deviations are known. Hence we apply z-statistics and observed value of z is 2.58. Since the alternative hypothesis is $\mu_1 \neq \mu_2$, we have a two tail test at $\sigma = 0.05$ the critical value of z is 1.96. Null hypothesis is rejected.
2. The sample sizes are small and σ is not known. Hence we apply t-statistic and observed value of *t* is 0.61. The hypothesis are H_o: $\mu_1 = \mu_2$ against H_A: $\mu_1 \neq \mu_2$. Thus, it requires a two-tail test. For 23 d.f., at 1% level of significance, the critical value of *t* is 2.50. H_o is not rejected.
3. The sample sizes are small and σ is not known, *t*-statistic is applied. Observed value of *t* is 0.72. H_o is $\mu_1 = \mu_2$ and H_A is $\mu_1 < \mu_2$. Hence one-tail test is required. Thus critical value of t at 33 d.f. for 5% level of significance is 2.00. H_o is not rejected.

UNIT 15: CHI-SQUARED TEST FOR NOMINAL DATA*

Structure

15.0 OBJECTIVES

After going through this Unit you will be in a position to

- present nominal data in the form of contingency table;
- explain the chi-squared test statistic; and
- apply chi-squared test to contingency tables.

15.1 INTRODUCTION

In the previous two Units we discussed the procedures of drawing conclusions about population parameters on the basis of sample information. In many cases, particularly for nominal variables, we do not have parameters. Here a variable (or attribute) assumes values pertaining to a finite number of categories and we can count the number of observations in each category. Drawing inferences in the case of nominal or categorical data is the subject matter of the present Unit.

The method of hypothesis testing discussed in the previous Unit requires certain assumptions about the population from which the sample is drawn. For example, application of *t*-test for small samples requires that the parent population is normally distributed. Similarly the hypothesis is formulated by specifying a particular value for the parameter. Hence these tests are called parametric tests.

When it is not possible to make any assumption about the value of a parameter the test procedure described in the previous Unit fails. In situations where the population does not follow normal distribution or where it is not possible to specify the parameter value, we use non-parametric tests. There are quite a few non-parametric tests depending upon our need. However, we confine ourselves to a common procedure, that is, chi-squared (pronounced as *kai*-squared) test. In Unit 11 we have defined the chi-squared distribution and its important features.

* Prof. Kaustuva Barik, Indira Gandhi National Open University, New Delhi

You should recall that the chi-squired distribution is a squared standard normal variable, given by $z = \frac{X-\mu}{\sigma}$, where μ is the mean and σ is the standard deviation of the population. Thus chi-squared statistic takes non-negative values only. Further, chi-squared statistic has a skewed distribution (right-skewed). The extent of skewness depends on its degree of freedom. As degrees of freedom increase, its skewness decreases (that is, it becomes more symmetric). The chi-squired test is extensively used in econometrics; primarily for comparison of variance. In many circumstances we do not know the population variance and estimate it from a sample. The estimated value of variance is compared with theoretical distributions on the basis of chi-squared test. Thus, the chi-squared statistic is used in building confidence intervals also. However, in this Unit we will limit ourselves to the use of chi-squared statistic for test of independence between variables.

15.2 CONTINGENCY TABLE

Contingency table is a rectangular table in which observations from the population are classified according to two characteristics. It is also called a two-way table, which is discussed in Unit 7. Recall that qualitative data can be arranged into categories and presented in the form of a two-way table.

In order to explain the application of chi-squared test let us take a concrete example. In Table 15.1 we present data where Occupation of Father is a nominal variable and Number of Children is a numerical variable. We divide occupation into five categories – i) unemployed, ii) unskilled labour, iii) skilled labour, iv) self-employed, and v) professional. Similarly we divide families into five categories according the number of children – i) no child, ii) one child, iii) two children, iv) three children, and v) more than three children. For a sample of 650 families the data obtained is presented in Table 15.1.

Table 15.1: Observed Frequency on Occupation and Number of Children

Number of Children	*Occupation of Father*					*Total*
	Unemployed	*Unskilled Labour*	*Skilled Labour*	*Self-Employed*	*Professional*	
	(1)	(2)	(3)	(4)	(5)	
0	10	15	10	12	11	58
1	35	25	17	18	25	120
2	22	33	45	40	43	183
3	11	40	48	58	30	187
≥ 4	11	33	30	19	9	102
Total	89	146	150	147	118	650

Table 15.1 is a called contingency table, because we are trying to find whether the number of children is contingent upon the occupation of the father.

Our purpose is test for possible relationship between the number of children and the occupation of father. Thus the null hypothesis is specified as

H_0: Number of children and occupation of father are independent

against the alternative hypothesis

H_A: Number of children and occupation of father are dependent

In Table 15.1 we have presented the observed frequency for each cell in the table. What should be the expected frequency when there is no relationship between the variables under consideration? We will answer this question below.

15.3 EXPECTED FREQUENCIES

As mentioned above, expected frequency is calculated under the assumption that there is no relationship between the number of children and the occupation of father. For each cell in Table 15.1 the expected frequency is obtained by multiplying the sample size *n* by the cell probability. In order to calculate the cell probability we first find out the marginal frequencies for each row and column. As you know from Unit 7 'row marginal' are given by the row totals. Similarly, 'column marginal' are given by column totals.

For the rows, we can find out the 'marginal row probability'. For row 1 the marginal row probability, $p(r_1)$, is given by

$$p(r_1) = \frac{58}{650} = 0.09 \qquad \ldots(15.1)$$

Marginal row probabilities for other rows are

$$p(r_2) = \frac{120}{650} = 0.18 \qquad p(r_3) = \frac{183}{650} = 0.28$$

$$p(r_4) = \frac{187}{650} = 0.29 \qquad p(r_5) = \frac{102}{650} = 0.16$$

Similarly for column 1 the marginal column probability, $p(c_1)$, is given by

$$p(c_1) = \frac{89}{650} = 0.14 \qquad \ldots(15.2)$$

Marginal column probabilities for other columns are

$$p(c_2) = \frac{146}{650} = 0.22 \qquad p(c_3) = \frac{150}{650} = 0.23$$

$$p(c_4) = \frac{147}{650} = 0.23 \qquad p(c_5) = \frac{118}{650} = 0.18$$

Recall from Unit 10 that if events A and B are independent then the probability of joint occurrence of A and B is given by

$$p(A \cap B) = p(A).p(B)$$

Therefore, if we assume the null hypothesis to be true, then cell probability for the first cell (c_1, r_1) will be

$$p(r_1 \cap c_1) = p(r_1).p(c_1) = \frac{58}{650} \times \frac{89}{650} = 0.0892 \times 0.1369 = 0.0122 \quad \ldots(15.3)$$

Hence, expected frequency for the first cell will be

$$E_{11} = n.p(r_1 \cap c_1) = 650 \times 0.0122 = 7.94 \quad \ldots(15.4)$$

In general terms we can say that the expected frequency of cell ij is

$$E_{ij} = \frac{(\text{Row } i \text{ total}) \ (\text{Column } j \text{ total})}{\text{Sample size}} \quad \ldots(15.5)$$

By applying (15.5) we calculate the expected cell frequency for each cell and prepare a table as given in Table 15.2.

Table 15.2: Calculation of Expected Frequency for Each Cell

Number of Children		*Occupation of Father*					*Total*
		Unemployed	*Unskilled Labour*	*Skilled Labour*	*Self-Employed*	*Professional*	
		c_1	c_2	c_3	c_4	c_5	
0	r_1	7.94	13.03	13.38	13.12	10.53	58.00
1	r_2	16.43	26.95	27.69	27.14	21.78	120.00
2	r_3	25.06	41.10	42.23	41.39	33.22	183.00
3	r_4	25.60	42.00	43.15	42.29	33.95	187.00
≥ 4	r_5	13.97	22.91	23.54	23.07	18.52	102.00
Total		89.00	146.00	150.00	147.00	118.00	650.00

The next step is to compare the observed frequency with the expected frequency.

15.4 CHI-SQUARED TEST STATISTIC

In order to compare the observed frequency with the expected frequency we construct the chi-squared statistic, which is given by

$$\chi^2 = \Sigma \frac{(O_i - E_i)^2}{E_i} \qquad \ldots(15.6)$$

where O refers to observed frequency and E refers to expected frequency.

The chi-squared statistic has degrees of freedom $(r-1)(c-1)$. For example, if there are 3 rows and 4 columns, then degrees of freedom is $(3-1)(4-1)=6$.

Let us summarise the steps to be followed in chi-squared test. These are:

1. specify the null and alternative hypotheses
2. calculate the expected frequency for each cell by using (15.5)
3. calculate the observed value of χ^2 statistic by using (15.6)
4. determine the degrees of freedom according to the formula $(r-1)(c-1)$
5. check the level of significance (α) required
6. from Table (15.3) given in Unit 15, Block 5 find out the critical value of χ^2 for α and relevant degrees of freedom
7. compare the observed value of χ^2 with the critical value of χ^2
8. if the observed value is less than the critical value, then do not reject H_0
9. if the observed value is greater than the critical value, then reject H_0 and accept H_A

For the data given in Table 15.1 let us find out the observed value of χ^2.

Table 15.3: $\frac{(O_i - E_i)^2}{E_i}$ for each Cell

Number of Children		*Occupation*					*Total*
		Unemployed	*Unskilled Labour*	*Skilled Labour*	*Self-Employed*	*Professional*	
		c_1	c_2	c_3	c_4	c_5	
0	r_1	0.53	0.30	0.86	0.10	0.02	1.80
1	r_2	20.99	0.14	4.13	3.08	0.47	28.81
2	r_3	0.37	1.60	0.18	0.05	2.88	5.08
3	r_4	8.33	0.10	0.54	5.84	0.46	15.26
≥ 4	r_5	0.63	4.44	1.77	0.72	4.89	12.46
Total		30.85	6.58	7.48	9.77	8.72	63.41

Since there are 5 rows and 5 columns, the degrees of freedom is $(5-1)(5-1)=16$.

For 16 d.f. the critical value of χ^2 at 5 per cent level of significance (see Table A2 in the Appendix) is 26.30. We find from Table 15.3 we find that the observed value of χ^2 is 63.41. Since the observed value is greater than the critical value we reject the null hypothesis and accept the alternative hypothesis. Therefore, we conclude that the variables 'number of children' and 'occupation of father' are dependent.

Check Your Progress 2

1. Explain the following concepts.
   a) marginal frequency
   b) cell probability
   c) expected frequency
   d) critical value of χ^2

   .....
   .....
   .....
   .....
   .....

2. There are three brands (orange, cola and lemon) of soft drinks produced by a company. A survey of 160 persons in two states (one form north- Punjab and one from south- Tamil Nadu) provides the following information.

	orange	cola	lemon
Punjab	33	26	31
Tamil Nadu	17	24	29

   Test the hypothesis that there is no preference for particular brand of soft drink in both the states ($\alpha = 0.05$).

   .....
   .....
   .....
   .....
   .....

15.5 LET US SUM UP

In the case of qualitative data we cannot have parametric values. Therefore, hypothesis testing on the basis of *z*-statistic or *t*-statistic cannot be performed. Chi-squared test is applied to such situations. Chi-squared test is a non-parametric test, where no assumption about population is required. There are various types of non-parametric tests beside the chi-squared test. Moreover, chi-squared test can be applied to many situations besides contingency table.

In contingency table we test the null hypothesis that variables under consideration are independent of each other against the alternative hypothesis that variables are related. Here we compare expected frequency with observed frequency and construct the chi-squared statistic. If the observed value of chi-squared exceeds the expected value of chi-squared we reject the null hypothesis.

15.6 ANSWERS/HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1. Go through the text and explain these terms.
2. The expected frequencies are

	orange	cola	lemon
Punjab	28.13	28.13	33.75
Tamil Nadu	21.88	21.88	26.25

The observed value of chi-squared statistic is 2.98. Degrees of freedom are 2. The critical value of chi-squared at 5 per cent level of significance at 2 degrees of freedom is 5.99. Hence null hypothesis is not rejected and soft drink consumption is independent of the region.

APPENDIX TABLES

Table A1: Normal Area Table

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

Table A2: Critical Values of Chi-squared Distribution

df\area	0.1	0.05	0.025	0.01	0.005
1	2.706	3.841	5.024	6.635	7.879
2	4.605	5.991	7.378	9.210	10.597
3	6.251	7.815	9.348	11.345	12.838
4	7.779	9.488	11.143	13.277	14.860
5	9.236	11.071	12.833	15.086	16.750
6	10.645	12.592	14.449	16.812	18.548
7	12.017	14.067	16.013	18.475	20.278
8	13.362	15.507	17.535	20.090	21.955
9	14.684	16.919	19.023	21.666	23.589
10	15.987	18.307	20.483	23.209	25.188
11	17.275	19.675	21.920	24.725	26.757
12	18.549	21.026	23.337	26.217	28.300
13	19.812	22.362	24.736	27.688	29.819
14	21.064	23.685	26.119	29.141	31.319
15	22.307	24.996	27.488	30.578	32.801
16	23.542	26.296	28.845	32.000	34.267
17	24.769	27.587	30.191	33.409	35.718
18	25.989	28.869	31.526	34.805	37.156
19	27.204	30.144	32.852	36.191	38.582
20	28.412	31.410	34.170	37.566	39.997
21	29.615	32.671	35.479	38.932	41.401
22	30.813	33.924	36.781	40.289	42.796
23	32.007	35.172	38.076	41.638	44.181
24	33.196	36.415	39.364	42.980	45.559
25	34.382	37.652	40.646	44.314	46.928
26	35.563	38.885	41.923	45.642	48.290
27	36.741	40.113	43.195	46.963	49.645
28	37.916	41.337	44.461	48.278	50.993
29	39.087	42.557	45.722	49.588	52.336
30	40.256	43.773	46.979	50.892	53.672

Table A3: Critical Values of *t* Distribution

Df\p	0.25	0.10	0.05	0.025	0.01	0.005
1	1.0000	3.0777	6.3138	12.7062	31.8205	63.6567
2	0.8165	1.8856	2.9200	4.3027	6.9646	9.9248
3	0.7649	1.6377	2.3534	3.1825	4.5407	5.8409
4	0.7407	1.5332	2.1318	2.7765	3.7470	4.6041
5	0.7267	1.4759	2.0150	2.5706	3.3649	4.0321
6	0.7176	1.4398	1.9432	2.4469	3.1427	3.7074
7	0.7111	1.4149	1.8946	2.3646	2.9980	3.4995
8	0.7064	1.3968	1.8595	2.3060	2.8965	3.3554
9	0.7027	1.3830	1.8331	2.2622	2.8214	3.2498
10	0.6998	1.3722	1.8125	2.2281	2.7638	3.1693
11	0.6974	1.3634	1.7959	2.2010	2.7181	3.1058
12	0.6955	1.3562	1.7823	2.1788	2.6810	3.0545
13	0.6938	1.3502	1.7709	2.1604	2.6503	3.0123
14	0.6924	1.3450	1.7613	2.1448	2.6245	2.9768
15	0.6912	1.3406	1.7531	2.1315	2.6025	2.9467
16	0.6901	1.3368	1.7459	2.1199	2.5835	2.9208
17	0.6892	1.3334	1.7396	2.1098	2.5669	2.8982
18	0.6884	1.3304	1.7341	2.1009	2.5524	2.8784
19	0.6876	1.3277	1.7291	2.0930	2.5395	2.8609
20	0.6870	1.3253	1.7247	2.0860	2.5280	2.8453
20	0.6870	1.3253	1.7247	2.0860	2.5280	2.8453
21	0.6864	1.3232	1.7207	2.0796	2.5177	2.8314
22	0.6858	1.3212	1.7171	2.0739	2.5083	2.8188
23	0.6853	1.3195	1.7139	2.0687	2.4999	2.8073
24	0.6849	1.3178	1.7109	2.0639	2.4922	2.7969
25	0.6844	1.3163	1.7081	2.0595	2.4851	2.7874
26	0.6840	1.3150	1.7056	2.0555	2.4786	2.7787
27	0.6837	1.3137	1.7033	2.0518	2.4727	2.7707
28	0.6834	1.3125	1.7011	2.0484	2.4671	2.7633
29	0.6830	1.3114	1.6991	2.0452	2.4620	2.7564
30	0.6828	1.3104	1.6973	2.0423	2.4573	2.7500
inf	0.6745	1.2816	1.6449	1.9600	2.3264	2.5758

Table A4: Critical Values of *F* Distribution

(5% level of significance)

df2/df1	1	2	3	4	5	6	7	8	9	10
1	161.448	199.500	215.707	224.583	230.162	233.986	236.768	238.883	240.543	241.882
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385	19.396
3	10.128	9.552	9.277	9.117	9.014	8.941	8.887	8.845	8.812	8.786
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964
5	6.608	5.786	5.410	5.192	5.050	4.950	4.876	4.818	4.773	4.735
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637
8	5.318	4.459	4.066	3.838	3.688	3.581	3.501	3.438	3.388	3.347
9	5.117	4.257	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137
10	4.965	4.103	3.708	3.478	3.326	3.217	3.136	3.072	3.020	2.978
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602
15	4.543	3.682	3.287	3.056	2.901	2.791	2.707	2.641	2.588	2.544
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538	2.494
17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548	2.494	2.450
18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456	2.412
19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477	2.423	2.378
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348
21	4.325	3.467	3.073	2.840	2.685	2.573	2.488	2.421	2.366	2.321
22	4.301	3.443	3.049	2.817	2.661	2.549	2.464	2.397	2.342	2.297
23	4.279	3.422	3.028	2.796	2.640	2.528	2.442	2.375	2.320	2.275
24	4.260	3.403	3.009	2.776	2.621	2.508	2.423	2.355	2.300	2.255
25	4.242	3.385	2.991	2.759	2.603	2.490	2.405	2.337	2.282	2.237
26	4.225	3.369	2.975	2.743	2.587	2.474	2.388	2.321	2.266	2.220
27	4.210	3.354	2.960	2.728	2.572	2.459	2.373	2.305	2.250	2.204
28	4.196	3.340	2.947	2.714	2.558	2.445	2.359	2.291	2.236	2.190
29	4.183	3.328	2.934	2.701	2.545	2.432	2.346	2.278	2.223	2.177
30	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266	2.211	2.165
40	4.085	3.232	2.839	2.606	2.450	2.336	2.249	2.180	2.124	2.077
60	4.001	3.150	2.758	2.525	2.368	2.254	2.167	2.097	2.040	1.993
120	3.920	3.072	2.680	2.447	2.290	2.175	2.087	2.016	1.959	1.911
inf	3.842	2.996	2.605	2.372	2.214	2.099	2.010	1.938	1.880	1.831

Table A4: Critical Values of *F* Distribution (Contd.)

(5% level of significance)

df2/df1	12	15	20	24	30	40	60	120	INF
1	243.906	245.950	248.013	249.052	250.095	251.143	252.196	253.253	254.314
2	19.413	19.429	19.446	19.454	19.462	19.471	19.479	19.487	19.496
3	8.745	8.703	8.660	8.639	8.617	8.594	8.572	8.549	8.526
4	5.912	5.858	5.803	5.774	5.746	5.717	5.688	5.658	5.628
5	4.678	4.619	4.558	4.527	4.496	4.464	4.431	4.399	4.365
6	4.000	3.938	3.874	3.842	3.808	3.774	3.740	3.705	3.669
7	3.575	3.511	3.445	3.411	3.376	3.340	3.304	3.267	3.230
8	3.284	3.218	3.150	3.115	3.079	3.043	3.005	2.967	2.928
9	3.073	3.006	2.937	2.901	2.864	2.826	2.787	2.748	2.707
10	2.913	2.845	2.774	2.737	2.700	2.661	2.621	2.580	2.538
11	2.788	2.719	2.646	2.609	2.571	2.531	2.490	2.448	2.405
12	2.687	2.617	2.544	2.506	2.466	2.426	2.384	2.341	2.296
13	2.604	2.533	2.459	2.420	2.380	2.339	2.297	2.252	2.206
14	2.534	2.463	2.388	2.349	2.308	2.266	2.223	2.178	2.131
15	2.475	2.403	2.328	2.288	2.247	2.204	2.160	2.114	2.066
16	2.425	2.352	2.276	2.235	2.194	2.151	2.106	2.059	2.010
17	2.381	2.308	2.230	2.190	2.148	2.104	2.058	2.011	1.960
18	2.342	2.269	2.191	2.150	2.107	2.063	2.017	1.968	1.917
19	2.308	2.234	2.156	2.114	2.071	2.026	1.980	1.930	1.878
20	2.278	2.203	2.124	2.083	2.039	1.994	1.946	1.896	1.843
21	2.250	2.176	2.096	2.054	2.010	1.965	1.917	1.866	1.812
22	2.226	2.151	2.071	2.028	1.984	1.938	1.889	1.838	1.783
23	2.204	2.128	2.048	2.005	1.961	1.914	1.865	1.813	1.757
24	2.183	2.108	2.027	1.984	1.939	1.892	1.842	1.790	1.733
25	2.165	2.089	2.008	1.964	1.919	1.872	1.822	1.768	1.711
26	2.148	2.072	1.990	1.946	1.901	1.853	1.803	1.749	1.691
27	2.132	2.056	1.974	1.930	1.884	1.836	1.785	1.731	1.672
28	2.118	2.041	1.959	1.915	1.869	1.820	1.769	1.714	1.654
29	2.105	2.028	1.945	1.901	1.854	1.806	1.754	1.698	1.638
30	2.092	2.015	1.932	1.887	1.841	1.792	1.740	1.684	1.622
40	2.004	1.925	1.839	1.793	1.744	1.693	1.637	1.577	1.509
60	1.917	1.836	1.748	1.700	1.649	1.594	1.534	1.467	1.389
120	1.834	1.751	1.659	1.608	1.554	1.495	1.429	1.352	1.254
inf	1.752	1.666	1.571	1.517	1.459	1.394	1.318	1.221	1.000

Table A4: Critical Values of *F* Distribution (contd.)

(1% level of significance)

df2/df1	1	2	3	4	5	6	7	8	9	10
1	4052.181	4999.500	5403.352	5624.583	5763.650	5858.986	5928.356	5981.070	6022.473	6055.847
2	98.503	99.000	99.166	99.249	99.299	99.333	99.356	99.374	99.388	99.399
3	34.116	30.817	29.457	28.710	28.237	27.911	27.672	27.489	27.345	27.229
4	21.198	18.000	16.694	15.977	15.522	15.207	14.976	14.799	14.659	14.546
5	16.258	13.274	12.060	11.392	10.967	10.672	10.456	10.289	10.158	10.051
6	13.745	10.925	9.780	9.148	8.746	8.466	8.260	8.102	7.976	7.874
7	12.246	9.547	8.451	7.847	7.460	7.191	6.993	6.840	6.719	6.620
8	11.259	8.649	7.591	7.006	6.632	6.371	6.178	6.029	5.911	5.814
9	10.561	8.022	6.992	6.422	6.057	5.802	5.613	5.467	5.351	5.257
10	10.044	7.559	6.552	5.994	5.636	5.386	5.200	5.057	4.942	4.849
11	9.646	7.206	6.217	5.668	5.316	5.069	4.886	4.744	4.632	4.539
12	9.330	6.927	5.953	5.412	5.064	4.821	4.640	4.499	4.388	4.296
13	9.074	6.701	5.739	5.205	4.862	4.620	4.441	4.302	4.191	4.100
14	8.862	6.515	5.564	5.035	4.695	4.456	4.278	4.140	4.030	3.939
15	8.683	6.359	5.417	4.893	4.556	4.318	4.142	4.004	3.895	3.805
16	8.531	6.226	5.292	4.773	4.437	4.202	4.026	3.890	3.780	3.691
17	8.400	6.112	5.185	4.669	4.336	4.102	3.927	3.791	3.682	3.593
18	8.285	6.013	5.092	4.579	4.248	4.015	3.841	3.705	3.597	3.508
19	8.185	5.926	5.010	4.500	4.171	3.939	3.765	3.631	3.523	3.434
20	8.096	5.849	4.938	4.431	4.103	3.871	3.699	3.564	3.457	3.368
21	8.017	5.780	4.874	4.369	4.042	3.812	3.640	3.506	3.398	3.310
22	7.945	5.719	4.817	4.313	3.988	3.758	3.587	3.453	3.346	3.258
23	7.881	5.664	4.765	4.264	3.939	3.710	3.539	3.406	3.299	3.211
24	7.823	5.614	4.718	4.218	3.895	3.667	3.496	3.363	3.256	3.168
25	7.770	5.568	4.675	4.177	3.855	3.627	3.457	3.324	3.217	3.129
26	7.721	5.526	4.637	4.140	3.818	3.591	3.421	3.288	3.182	3.094
27	7.677	5.488	4.601	4.106	3.785	3.558	3.388	3.256	3.149	3.062
28	7.636	5.453	4.568	4.074	3.754	3.528	3.358	3.226	3.120	3.032
29	7.598	5.420	4.538	4.045	3.725	3.499	3.330	3.198	3.092	3.005
30	7.562	5.390	4.510	4.018	3.699	3.473	3.304	3.173	3.067	2.979
40	7.314	5.179	4.313	3.828	3.514	3.291	3.124	2.993	2.888	2.801
60	7.077	4.977	4.126	3.649	3.339	3.119	2.953	2.823	2.718	2.632
120	6.851	4.787	3.949	3.480	3.174	2.956	2.792	2.663	2.559	2.472
inf	6.635	4.605	3.782	3.319	3.017	2.802	2.639	2.511	2.407	2.321

Table A4: Critical Values of *F* Distribution (contd.)

(1% level of significance)

df2/df1	12	15	20	24	30	40	60	120	INF
1	6106.321	6157.285	6208.730	6234.631	6260.649	6286.782	6313.030	6339.391	6365.864
2	99.416	99.433	99.449	99.458	99.466	99.474	99.482	99.491	99.499
3	27.052	26.872	26.690	26.598	26.505	26.411	26.316	26.221	26.125
4	14.374	14.198	14.020	13.929	13.838	13.745	13.652	13.558	13.463
5	9.888	9.722	9.553	9.466	9.379	9.291	9.202	9.112	9.020
6	7.718	7.559	7.396	7.313	7.229	7.143	7.057	6.969	6.880
7	6.469	6.314	6.155	6.074	5.992	5.908	5.824	5.737	5.650
8	5.667	5.515	5.359	5.279	5.198	5.116	5.032	4.946	4.859
9	5.111	4.962	4.808	4.729	4.649	4.567	4.483	4.398	4.311
10	4.706	4.558	4.405	4.327	4.247	4.165	4.082	3.996	3.909
11	4.397	4.251	4.099	4.021	3.941	3.860	3.776	3.690	3.602
12	4.155	4.010	3.858	3.780	3.701	3.619	3.535	3.449	3.361
13	3.960	3.815	3.665	3.587	3.507	3.425	3.341	3.255	3.165
14	3.800	3.656	3.505	3.427	3.348	3.266	3.181	3.094	3.004
15	3.666	3.522	3.372	3.294	3.214	3.132	3.047	2.959	2.868
16	3.553	3.409	3.259	3.181	3.101	3.018	2.933	2.845	2.753
17	3.455	3.312	3.162	3.084	3.003	2.920	2.835	2.746	2.653
18	3.371	3.227	3.077	2.999	2.919	2.835	2.749	2.660	2.566
19	3.297	3.153	3.003	2.925	2.844	2.761	2.674	2.584	2.489
20	3.231	3.088	2.938	2.859	2.778	2.695	2.608	2.517	2.421
21	3.173	3.030	2.880	2.801	2.720	2.636	2.548	2.457	2.360
22	3.121	2.978	2.827	2.749	2.667	2.583	2.495	2.403	2.305
23	3.074	2.931	2.781	2.702	2.620	2.535	2.447	2.354	2.256
24	3.032	2.889	2.738	2.659	2.577	2.492	2.403	2.310	2.211
25	2.993	2.850	2.699	2.620	2.538	2.453	2.364	2.270	2.169
26	2.958	2.815	2.664	2.585	2.503	2.417	2.327	2.233	2.131
27	2.926	2.783	2.632	2.552	2.470	2.384	2.294	2.198	2.097
28	2.896	2.753	2.602	2.522	2.440	2.354	2.263	2.167	2.064
29	2.868	2.726	2.574	2.495	2.412	2.325	2.234	2.138	2.034
30	2.843	2.700	2.549	2.469	2.386	2.299	2.208	2.111	2.006
40	2.665	2.522	2.369	2.288	2.203	2.114	2.019	1.917	1.805
60	2.496	2.352	2.198	2.115	2.028	1.936	1.836	1.726	1.601
120	2.336	2.192	2.035	1.950	1.860	1.763	1.656	1.533	1.381
inf	2.185	2.039	1.878	1.791	1.696	1.592	1.473	1.325	1.000

GLOSSARY

Array : An array is an arrangement of data in ascending or descending order. It is also called a simple array.

Area diagrams : These are *two dimensional* diagrams. Here both the height and the base of the diagram are important. That is why they are known as area diagrams. They can be either rectangles, or squares or circles.

Arithmetic Mean : Sum of observed values of a set divided by the number of observations in the set is called a mean or an average.

Actuarial Science : Actuarial Science is concerned with the application of mathematical and statistical methods to finance and insurance, particularly where this relates to the assessment of risks in the long term. In actuarial science we compute the insurance risks and premiums.

Bar diagram : It is often defined as a set of thick lines corresponding to various values of the variable. It is different from histogram where width of the rectangle is important.

Base Year : Preferably a normal year, in terms of variable concerned, base year index is invariably taken as 100. Current year index is expressed as a percentage of base year index.

Chain Index : Current Year's index is expressed as a percentage of previous index.

Cyclical Variations : Oscillatory movements of a time series where the period of oscillation, called cycle, is more than a year.

Cohort : A group of people sharing a common demographic experience who are observed through time. For example, the birth cohort of 2003 is the people born in that year.

Conditional Probability : **:** If A and B are not mutually exclusive events then the probability of B given that A has already occurred is known as the conditional probability of B given that A has occurred and is denoted by $P(B/A)$.

Complementary Event : If A is an event, then the non-occurrence of the event A, denoted by $\overline{A}$ is called complement to the event A. The sum of probabilities of any event and its complement is always equal to 1.

Continuous Probability Distribution : It is the probability distribution for a continuous random variable.

Continuous Random Variable : It is a random variable that can take all values in a certain interval.

Coefficient of Determination : It is given as r^2, i.e., the square of the correlation coefficient. It shows the percentage variation in the dependent variable y explained by the independent variable x.

Coefficient of Kurtosis : It is a measure of the relative peakedness of the top of a frequency curve.

Coefficient of Variation : It is a relative measure of dispersion which is independent of the units of measurement. As opposed to this Standard Deviation is an absolute measure of dispersion.

Condensation of data : It is a process of classifying and arranging complex and unorganised mass of data to make them fit for comparison and analysis.

Confidence Level : It gives the percentage (probability) of samples where the population mean would remain within the confidence interval around the sample mean. If α is the significance level the confidence level is $(1-\alpha)$.

Contingency Table : A two-way table to present bivariate data. It is called contingency table because we try to find whether one variable is contingent upon the other variable

Contingency Table : A two-way table to present bivariate data. It is called contingency table because we try to find whether one variable is contingent upon the other variable.

Continuous frequency distribution : A continuous frequency distribution in formed where the variable can take any value between two numbers like height and weight, income, temperatures, etc.

Continuous Probability Distribution : It is the probability distribution for a continuous random variable.

Continuous Random Variable : It is a random variable that can take all the values in a given interval.

Cumulative frequency distribution : It is obtained by successive totaling of the simple frequencies of a discrete or continuous distributions. This totaling can be done either from above (giving "less-than" cumulative frequency distribution) or from below (giving "more-than" cumulative frequency distribution).

Caption : It is a part of a table and labels data presented in the column of a table. It is also called *box head*. It may contain one or more than one column head.

Census Data : The data obtained by observing all the items of population.

Chi-square Distribution : It is an asymmetric distribution where the range of variation for the random variable is from zero to infinity. For fairly large degrees of freedom, it approaches the normal distribution.

Chi-square Variable : A random variable that follows the chi-square distribution.

Class and class limits : It is a decided group of magnitudes having two ends called class limits or *class boundaries*.

Class range : Also called *class interval* is the difference of two limits of a class. It is equal to upper limit minus lower limit. It is also called *class width*.

Cluster Sampling : It is a sampling procedure where the entire population is divided into groups called clusters and then a random number of clusters are selected. All observations in the selected clusters are included in the sampling.

Convenience Sampling : It refers to the method of obtaining a sample that is most conveniently available to the researcher.

Data Point : It is an observation from an individual or item.

Data Set : It is the collection of all data points.

Degrees of Freedom : It refers to the number of pieces of independent information that are required to compute some characteristic of a given set of observations.

Discrete frequency distribution : A discrete distribution or discrete series is formed where the variable can take only discrete values like 1,2,3,..... Number of children in a family, number of students in a university, etc. are examples of discrete variable.

Discrete Probability Distribution : It is the probability distribution for a discrete random variable.

Discrete Random Variable : It is a random variable that either assumes a finite number of values or an infinite sequence (like 1, 2, 3, …).

Estimate : It is the particular value that can be obtained from an estimator.

Estimation : It is the method of prediction about parameter values on the basis of sample statistics.

Estimator : It is specific functional form of a statistic or the formula involved in its calculation. Generally, the two terms, statistic and estimator, are used interchangeably.

Exclusive type class interval : A class interval which includes all observations that are greater than or equal to the lower limit but less than the upper limit.

Expected Frequency : It is the expected cell frequency under the assumption that both the variables are independent.

***F* Distribution** : It is an asymmetric distribution that is skewed to the right. For fairly large degrees of freedom, it approaches the normal distribution

***F* Variable** : A random variable that follows the *F* distribution.

Frequency array : It is an array or series formed by writing various possible values of the variable along with their respective frequencies.

Frequency curve : It is a smoothened graph of a frequency distribution obtained from frequency polygon through free hand tracing in such a way that the area under both of them is approximately the same.

Frequency Distribution : The arrangement of data in the form of frequency distribution that describes the basic pattern which the data assumes in the mass.

Frequency polygon : It is a broken line graph to represent a frequency distribution and can be obtained either from a histogram or directly from the distribution.

Gaussian Distribution : It is the other name for the normal distribution.

Geometric Mean : It is the mean of *n* values of a variable computed as the nth root of their product.

Harmonic Mean : It is the inverse of the arithmetic mean of the reciprocals of the observations of a set

Histogram : It is a set of adjacent rectangles presented vertically with areas proportional to the frequencies.

Historigram : The line graph of a time series is called historigram (For example, steel production since 1950).

Hypothesis : It is an assertion or statement about a population.

Inclusive type class interval : A class interval in which all observations lying between and including the class limits are included.

Independent : Two events *A* and *B* are said to be mutually independent if

Events		the occurrence of *B* does not depend upon the occurrence of *A* and vice versa.
Index Number	:	A pure number, expressed as a percentage to base year value. Index Number measures the relative changes over time in the variable concerned (price, quantity sales or say, exports) of a group of commodities. This is a special type of weighted average of prices (or any other attribute) of the commodities or items included.
Investigator	:	The person responsible for the collection of information from respondents.
Irregular Movement	:	The random movement of time series which is not
Infant mortality rate	:	The number of deaths of infants below one year old per 1000 live births in a given year.
Judgment Sampling	:	In this sampling procedure the selection of sample is based on the researcher's judgment about some appropriate characteristic required of the sample units.
Life expectancy	:	The average number of additional years a person could expect to live if the current mortality trends continue for the rest of that person's life. Frequently we use life expectancy at birth.
Line graph	:	It is the locus of different points obtained with the combinations of X and Y coordinates measured on X-axis and Y-axis respectively.
Migration	:	The movement of people across a specified boundary for the purpose of establishing a new or semi permanent residence.
Mid-year population	:	It is the average of end-year estimates. For example, the mid-year population of 2020 will be the average of the population as on 31st December 2019 and 31st December 2020.
Multistage Sampling	:	The sample selection is done in a number of stages.
Main body of the table	:	It is certainly the most important part of the table and contains numerical information about which a hint is already made clear by the title. It is also called *field of the table*.
Mathematical Expectation	:	The mathematical expectation or the expected value of a random variable is the sum of the products of the values of the random variable and the corresponding probabilities.
Mean Deviation	:	It is the arithmetic mean of absolute deviations (i.e., the differences) from mean or median or mode.
Median	:	In a set of observations, it is the value of the middlemost item when they are arranged in order of magnitude.

Method of Least Squares : When a polynomial function is fitted to the time series, the method of least squares requires that the parameters of the function should be so chosen as to make the sum of squares of the deviations between actual observations and empirical values to be minimum.

Mid-point : Also called mid-value, it is the average value of two class limits. It falls just in the middle of a class.

Mode : In a set of observations, it is the value which occurs with maximum frequency.

Moment of order *r* : It is defiend as the arithmetic mean of the r^{th} power of deviations of observations.

Moving Average Method : Taking a suitable period of moving average, (say 3 years), the moving averages are calculated as series of mean values of three (in this case) consecutive years. The average obtained is entered against the middle year.

Nominal Variable : Such a variable takes qualitative values and do not have any ordering relationships among them. For example, gender which assumes two qualitative values, male and female has no ordering in 'male' and 'female' status. A nominal variable is also called an attribute.

Normal Distribution : The best known of all the theoretical probability distributions. It traces out a bell-shaped symmetric probability curve.

Normal Equations : A set of simultaneous equations derived in the application of the least squares method, for example in regression analysis. They are used to estimate the parameters of the model.

Normal Variable : A random variable that follows the normal distribution.

Number Ogive : It is the graph of cumulative frequency. Graph of "less-than" cumulative frequencies gives "less-than" ogive and that of "more-than" gives "more-than" ogive.

Natural Increase : The surplus of births over deaths in a population in a given period of time.

Open-end class : A class in which one of the limits is not specified.

Parameter : It is a measure of some characteristic of the population.

Pictographs : Here the data are presented in the form of pictures.

Pie diagram : It is a circle sub-divided into components to present proportion of different constituent parts of a total. It is also called pie chart.

Population : It is the entire collection of units of a specified type in a given place and at a particular point of time.

Posterior Probabilities : The revised probabilities of various events are known as posterior probabilities. This revision is made on the basis of the occurrence or non-occurrence of certain events by using Bayes' Theorem.

Price Relative : In the construction of an index number price relative for a commodity is the ratio of the current year price to base year price of that commodity.

Primary Data : Data obtained by observing the items or individuals under the ambit of a problem under consideration.

***a priori* Probabilities** : The probabilities assigned to various events on the basis of the classical definition or statistical definition or in a subjective manner are priori probabilities.

Probability Density Function : It is a function of a continuous random variable. However, like the probability mass function, it cannot directly give the probability for a specified value of the random variable. Here, we can only find the probability of the random variable lying in an interval.

Probability Distribution : It is a statement about the possible values of a random variable along with their probabilities.

Probability Mass Function : It is a function that gives the probability for a specified value of a discrete random variable.

Probability : It is a relative measure for the degree of certainty (and implicitly that of non-chance) associated with an event. For an event A, the probability.

Problem of Estimation : We may be interested in some feature of the population that is *completely* unknown to us and we want to make some intelligent guess about it on the basis of a random sample drawn from the population. This problem of statistical inference is known as the problem of estimation.

Quantity Index : The variable considered is the quantity of commodities.

Questionnaire or Schedule : It is a list of questions that are relevant to the inquiry at hand.

Quota Sampling : In this sampling procedure the samples are selected on the basis of some parameters such as age, gender, geographical region, education, income, religion, etc.

Random Sampling : It is a procedure where every member of the population has a definite chance or probability of being selected in the sample. It is also called probability sampling.

Range : It is the difference between the largest and the smallest observations of a given set or data.

Regression	:	It is a statistical measure of the average relationship between two or more variables in terms of the original units of the data.
Relative frequency distribution	:	It is frequency distribution where the frequency of each value is expressed as a *fraction* or a *percentage* of the total number of observations.
Respondent	:	The person who supplies the information.
Rate of Natural increase	:	The rate at which a population increases in a given year because of surplus of births over deaths expressed as per 1000 of the population. This excludes migration.
Simple Random Sampling	:	It is the basic sampling procedure when we select samples using lottery method or using random number tables.
Snowball Sampling	:	Snowball sampling relies on referrals from initial sampling units to generate additional sampling units.
Stratified Sampling	:	In this sampling procedure the population is divided into groups called strata and then the samples are selected from each stratum using a random sampling method.
Systematic Sampling	:	A sampling procedure in which units are selected from the population at uniform interval that is measured in time, order or space.
Sample Data	:	The data obtained by observing only those items which are included in the sample.
Sample	:	It is a sub-set of the population. It can be drawn from the population in a scientific manner by applying the rules of probability so that personal bias is eliminated. Many samples can be drawn from a population and there are many methods of drawing a sample.
Sampling Distribution	:	It is the relative frequency or probability distribution of the values of a statistic when the number of samples tends to infinity.
Sampling Distribution	:	It refers to the probability distribution of a statistic.
Sampling Error	:	The absolute difference between population parameter and relevant sample statistic.
Sampling Fluctuation	:	It is the variation in the values of a statistic computed from different samples.
Seasonal Variation	:	Periodical movement where the period is not longer than one year.

Secondary Data : Data obtained from the already collected data of some agency.

Secular Trend : The smooth, regular and long-term movement of a time series over a period of time. Trend may be upward or rising, downward or declining or it may remain more or less constant over time.

Significance Level : There may be certain samples where population mean would not remain within the confidence interval around sample mean. The percentage (probability) of such cases is called significance level. It is usually denoted by α. When $\alpha = 0.05$ (that is, 5 percent) we can say that in 5 per cent cases we are likely to reach an incorrect decision or commit Type I error. Level of significance could be at any level but it is usually taken at 5 percent or 1 percent level.

Simple and sub-divided bar diagram : In the case of simple bar diagram only one variable can be presented. A sub-divided bar diagram is used to show various components of a phenomenon.

Simple Random Sampling : This is a sampling procedure, in which, each member of the population has the *same chance* of being selected in the sample

Skewness : Departure from symmetry is skewness.

Standard Deviation : It is positive square root of the variance.

Standard Error : It is the standard deviation of the sampling distribution of a statistic.

Standard Normal Variate : A normal variable with mean 0 and standard deviation equal to 1.

Statistic : It is a function of the values of the units that are included in the sample. The basic purpose of a statistic is to estimate some population parameter.

Statistical Inference : It is the process of concluding about an unknown population from a known sample drawn from it.

Statistical Survey : It is a method for the collection of data by observing all or a sample of items under the ambit of a given problem.

Statistical Unit : It is a characteristic or a set of characteristics of an item that are observed to collect data.

Stub : Also a part of a table, it consists of stub head and stub entries. Each stub entry labels a given data placed in the rows of the table. Both *stub head* and *stub entries* appear on the left-hand column of a table. They describe the row heads.

Student's-*t* Distribution	:	It is a symmetric distribution about the value zero. The range of variation for the student's-*t* random variable is $-\infty$ to $+\infty$. It is, however, flatter than the normal distribution curve. For fairly large degrees of freedom, it approaches the normal distribution.
Tabulation	:	It is a systematic presentation of data in rows and columns.
Test of Hypothesis	:	Testing the validity of a hypothesis on the basis of collected data. The probability of success, p, in a binomial distribution is very small and the number of trials, n, is so large that the expectation, $\mu = np$ is finite.
Theoretical Distribution	:	It is a probability distribution that is generated by specifying the conditions of a random experiment. Some examples of probability distributions are the binomial distribution, the Poisson distribution, the normal distribution, etc.
Variance	:	If $E(x)$ is the mathematical expectation of a random variable x, the variance of x is defined as $\mathrm{E}\,[\,x\text{-}\mathrm{E}(x)]^2$.
Variance	:	It is the arithmetic mean of squares of deviations of observations from their arithmetic mean.
Volume diagrams	:	These are *three dimensional* diagrams. In their construction length, width and height are used. They consist of boxes, cubes, blocks, spheres and cylinders.

SOME USEFUL BOOKS

Devore, Jay L., 2010, *Probability and Statistics for Engineers*, Cengage Learning

Freund, John E., 1992, *Mathematical Statistics*, Prentice Hall

Larsen, Richard J. and Morris M. Marx, 2011, An Introduction to Mathematical Statistics and its Applications, Prentice Hall

Cochran, William G., 2007, Sampling Techniques, John Wiley