Sampling Techniques and Sampling Distributions

[Sampling versus complete enumeration; Random and non-random sampling; Different types of random sampling; Sample statistic and population parameter; Practical methods of drawing a random sample; Sampling distributions - standard error; sampling distribution of the sample mean and the sample proportion; Sampling from normal and non-normal populations; Central Limit Theorem; Four basic distributions: Standard normal distribution; Chi-square distribution; t-distribution; F-distribution]

4.1 Introduction

Statisticians are often concerned in studying some characteristics of a larger whole of which every member cannot be observed for some practical reasons. For example, in order to estimate the average height of adult males of Calcutta, the investigator cannot afford the time or expenses required to measure the height of each adult males of Calcutta, or, a farmer, who wants to know what proportion of his potato crop is diseased, cannot afford time and expense to examine each and every potato. Likewise, in order to estimate the total production of a crop in a given season, it is out of question to harvest and weigh the produce from all the fields growing the crop which constitute the population. In such cases what an investigator can best do is to select a limited number of individuals and collect information from the selected individuals. ***Sampling theory*** consists of selecting a small representative part from the larger whole under study. The larger whole as mentioned above is known as ***population***. The small part which is selected from the larger whole in order to have some idea about the larger whole is known as *sample*. ***Population*** is the entire group of objects or individuals about which information is wanted. ***Sample*** is a representative part drawn from the *population* under study in order to have some idea or to take some decision about the population. For example, in opinion poll, a relatively small number of persons are interviewed, and their opinions on some current issues are gathered in order to get an idea about the opinion of the whole community, or, from a large lot of manufactured products a relatively small number of items are inspected in order to make decision about the quality (defective or non-defective) of the whole lot.

A population consists of units or elements and with each population

unit there associates some numerical value of one or more variables, or, category of one or more attributes. Size of the population is denoted by N and the size of the sample is denoted by n. Exactly what constitutes a population varies from one case to another depending on the objective of the survey. While studying a population, we are usually interested in one or more unknown characteristics of the population. These characteristics are known as ***parameters***. We imagine a population as an existing collection of real units, each of which could be identified and located, inspected, interviewed, observed or measured. When it is not possible from practical point of view to measure or observe the entire population, we draw a sample from it. The purpose of studying sampling theory is to evolve sampling procedures so that maximum efficiency (or information) can be achieved for a given cost or to achieve a fixed level of efficiency at a minimum cost. The procedure to observe the entire population is called ***complete enumeration*** or ***census***.

It is to be noted that the sample drawn from a population should be a *representative of the population* in the sense that the sample will reveal the main characteristics of the population. Otherwise, the decisions made or idea formed about the population on the basis of the sample will be wrong.

4.2 Advantages of Sampling over Census

Following are the reasons for preferring sampling over census:

1. ***Reduction of cost*** : Sampling can provide reliable information at less cost than census. Although the cost of collecting an observation may be higher in sample survey, the total cost is expected to be smaller since a sample consists of a part of the total population. If a population is very large, census can be very expensive.

2. ***Greater speed*** : Using sample, the data can be collected more quickly, hence the estimates can be published in time. In census, the collection of data may take so long that the information gathered may be no longer needed or useful.

3. ***Greater scope*** : A complete enumeration or census requires a large administrative set-up and involves many persons in data collection. Some enquiries may even require highly trained personnel or specialized equipment for collecting data which, sometimes, may make census unmanageable, whereas in sample survey we may have better coverage and may not face the problems stated above due to the smaller scale of the investigation.

4. ***Greater accuracy*** : Estimates based on sample are often more

accurate than those based on census, because investigators can be more careful while collecting data for fewer units. In sample, more attention can be given on data quality through the training of personnel and monitoring. A complete enumeration or census requires a large administrative set-up and involves many persons in data collection. With this huge administrative complexity and the pressure to produce timely estimates, many types of errors can easily creep into the census data.

5. *Measure of precision of the estimates* : In complete enumeration there is no way of gauging the magnitude of errors incorporated in the estimates. But a sample, drawn on the basis of a properly designed survey method, permits quantitative assessment of errors involved in the estimates.

6. *Greater applicability* : In some studies the observations are obtained by destroying the units. For example, to obtain the average life-hour of electric bulbs, we have to measure the total hours the bulbs survive until they burnt out. A cookie must be pulverized in order to determine the fat content etc.. In such cases drawing of sample is essential, since census destroys the entire population. While studying an infinite or hypothetical population sampling is inevitable.

However, when it is essential to gather information from all the units of a population irrespective of cost or time, a complete census is required, for example, population census, where relevant information are collected from all the units of the population.

Now we discuss different sampling techniques, viz., simple random sampling, stratified sampling, systematic sampling, cluster samping in the subsequent sections.

In the next section we discuss some important terminology related to sampling theory.

4.3 Basic Terminology

Following definitions will make the notion of sampling more precise:

Population or ***Target population*** : As we have defined earlier, a *population*, or *target population* is the complete collection of units we want to study. Defining the *target population* and the *sampled population* are very important in the context of sample survey. *Sampled population* is the collection of all population units from which the actual sample is drawn. For example, in a political poll the target population could be all adults who are eligible to vote and the sampled population will be all adults enlisted in the voters' list, i.e., the registered voters. In an ideal survey the

sampled population is identical to the target population. A population could be ***finite*** or ***infinite*** depending on whether it contains a finite or infinite number of units. For example, the population of books in the National Library is finite while the population of atmospheric pressures at different points in the atmosphere, the population of possible sizes of paddy crop (i.e., yields) are considered to be infinite, since here the variable can take any numerical value within certain limits. In many cases the number of units in a population is so large that it can be considered as infinite for all practical purposes. Again, the population may be ***existent*** or ***hypothetical***. If the population consists of concrete existent objects, then it is called an ***existent population***. Sometimes the units in the population do not exist in reality but created by some artificial way, like die throwing, coin tossing etc., then the population is called a ***hypothetical population***. The population of values which the bank rate can have in ten years' time, the population of the possible ways in which three balls can be arranged on a square table are further examples of hypothetical populations.

Sample : A *sample* is a subset of the population. A perfect sample is a scaled down version of the population, mirroing every main features of the population. Of course no such ideal sample can exist for a complicated population. But a good sample will reproduce the characteristics of interest in the population as closely as possible. When the population is homogeneous, i.e., when all the units are identical, it is obvious that a sample of just one element is sufficient to provide correct information about the population. For example, in case of blood test, a few drops of blood can give the idea on the nature of blood in the whole body. For a population consisting of a number of homogeneous groups, just one element from each group is enough to constitute a representative sample. Suppose that the population consists of several homogeneous groups and each group is a miniature form of the entire population, then any one group in the population may serve as a sample.

Sampling : *Sampling* is nothing but selecting a sample from the population. In practice, sampling is usually made ***without replacement***. Under this method, selected population units are not eligible to be selected once again and a population unit can appear at most once in the sample. On the contrary, in sampling ***with replacement***, a unit once drawn from the population is returned to the population before the next draw. A without replacement sample is better (in the sense of giving more information) than the sample of same size drawn under sampling with replacement.

Sampling is broadly classified as ***subjective*** and ***objective***. In ***subjective sampling*** the selection of the units in the sample depends on the personal judgement of the investigator. These samples produce biased results that systematically differ from the truth about the population. For example, suppose a truckload of potatoes has just been arrived in and because of convenience, sample of few baskets of potatoes are taken from the top of the truckload to see the quality of the potatoes. Clearly, the potatoes taken from the top may not be representative of the entire truckload since we do not have any potatoes from the bottom or middle of the truckload in the sample which may be damaged in shipment.

Objective sampling is such that there is a definite rule of selecting the population units. Objective sampling can be classified again as ***non-probabilistic***, ***probabilistic*** and ***mixed***. When the population units are drawn according to some specific rule but no probability is assigned to the process of selecting the units, then the sampling is called ***non-probabilistic***. In ***probabilistic sampling*** (also called ***random sampling***) each unit of the population has a preassigned non-zero probability of being selected. Probability sampling can be a sampling with *equal probabilities* or a sampling with *unequal probabilities*. In random sampling, if each unit of the population has equal chance of being selected, then the sampling is called ***simple random sampling***. For *unequal probability* sampling the chances of selecting population units are different. If the sampling is partly probabilistic and partly non-probabilistic, then it is called ***mixed sampling***.

Sample survey : *Sample survey* is a survey carried out on a properly chosen representative sample.

Sampling unit : The ultimate unit sampled from the population and observed is called a *sampling unit*.

Sampling frame : It is the complete list of population units from which the sample is drawn. For household survey , the list of all street addresses, for an agricultural survey, a list of all farmers or a map of areas containing farms can serve as the sampling frame. Some popular sampling frames are census report, voters' list etc.. Failing to include all the units of the target population in the sampling frame is called ***under coverage***.

Parameter and statistic : As discussed earlier the purpose of selecting a sample is to obtain estimates of certain population characteristics or population quantities. These population characteristics are called *parameters*.

Parameter is a function of population values of the variate or character under study. It has a numerical value that would be calculated from all of the units in the population. In sample survey we are mainly interested in the following parameters , viz.,μ, τ, π, λ' , as defined below:

$$\mu = \frac{X_1 + X_2 + \ldots + X_N}{N}, \quad \tau = X_1 + X_2 + \ldots + X_N = N\mu ,$$

$$\pi = \frac{\lambda'}{N} \text{ and } \lambda' = N\pi ,$$

where N = total population size, τ = total value of the variable in the population, μ = population average of the variable, π = population proportion of units having a specific character, λ' = total number of units in the population having a specific character.

A ***statistic*** is a function of sample observations. For example, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, the sample mean is a statistic based on the sample data $x_1, x_2, \ldots, x_n$, sample proportion, $p = \frac{x}{n}$, is a statistic based on x, is the observed number of units having a specific character in the sample. A statistic has got a numerical value calculated from the units in the sample.

Parameter is a fixed value for a population, but the value of a statistic varies from sample to sample. Suppose, our population consists of five individuals. The variable of interest is their monthly income in rupees. The parameter of interest is the average monthly income. The monthly incomes are reported as, say,Rs. 10,000, Rs 15,000, Rs. 12,000, Rs. 8,000, Rs. 15,000. Here we can easily compute the parameter value, since the incomes corresponding to all population units are available. The average income of the population , Rs. 12,000, is fixed for this particular population. Now suppose, we have sampled three individuals from the population and we are interested to have some idea about the population average income on the basis of the sample values. Here we average the three values of income obtained from the selected sample units. The sample average is a statistic, since it is computed from sample units. If the individuals selected have their incomes as Rs. 15,000/-, Rs. 8,000/-, and Rs. 10,000/-, then the sample average will be Rs. 11,000/-. But if three different units are selected such that their incomes are Rs. 10,000, Rs. 12,000, and Rs. 8,000, then the sample average of income will be Rs. 10,000, from which it is evident that the value of a statistic varies from sample to sample. Actually statistic, being based on sample observations, is a random variable.

Sampling fluctuations : Sampling fluctuations are the differences in the values of a statistic observed for different samples.

Sampling distribution of a statistic : Statistic is a random variable. The probability distribution of a statistic is known as its sampling distribution. The *sampling distribution* of a statistic is the probability distribution of all possible values of the statistic that results when random samples of same size (say, n) are repeatedly drawn from the population. The following example demonstrates how to derive the sampling distribution of the statistic 'sample mean $\bar{x}$'.

Unbiased estimator: If a statistic T is used to estimate a parametric θ, then it (T) is called an unbiased estimator of θ if $E(T) = \theta$.

Standard error (s.e.) : Standard deviation of a statistic in its sampling distribution is known as its *standard error*. It is so called because it refers to the precision of the statistic as an estimator.

Example 4.1 *Suppose that a committee consists of 5 members. The number of children of each of them is given below:*

Member	: A	B	C	D	E
No. of children	: 2	1	0	3	2

If a random sample of size $n = 2$ is selected without replacement from the above population, then find the sampling distribution of the sample mean $\bar{x}$.

Solution:

Here possible samples, sample values and possible values of sample mean $\bar{x}$ are given below:

Possible samples	Sample values	Sample mean ($\bar{x}$)
(A, B)	(2, 1)	1.5
(A, C)	(2, 0)	1.0
(A, D)	(2, 3)	2.5
(A, E)	(2, 2)	2.0
(B, C)	(1, 0)	0.5
(B, D)	(1, 3)	2.0
(B, E)	(1, 2)	1.5
(C, D)	(0, 3)	1.5
(C, E)	(0, 2)	1.0
(D, E)	(3, 2)	2.5

Hence the sampling distribution of $\bar{x}$ is as follows:

Values of $\bar{x}$	Probability
0.5	1/10
1.0	2/10
1.5	3/10
2.0	2/10
2.5	2/10
Total	1

Some important sampling distributions derived from the normal population will be discussed later in this chapter.

4.4 Simple Random Sampling (SRS)

The objective of sample survey is to make inference about the population parameter from the information contained in a sample. Two factors affect the quality and quantity of information contained in the sample: (i) representativeness of the sample and (ii) sample size. The representativeness of a sample can be controlled by controlling the way of selecting the sample. The procedure of selecting the sample is known as ***sample survey design***. For a fixed sample size n, we will consider various sampling designs or sampling procedures for selecting representative sample.

If a sample of size n is drawn from a population of size N in such a way that each and every member of the population has equal chance of being selected in the sample, then the sampling is called ***simple random sampling***. The sample, thus obtained is called a ***simple random sample.*** Simple random sampling may be *with* or *without replacement*. In simple random sampling every possible sample of size n has also equal chance of being selected and sometimes this property is considered as the definition of simple random sample.

In a simple random sampling, if the drawings are made one by one and each selected item is returned to the population before the next drawing, then the sampling procedure is known as ***simple random sampling with***

replacement (SRSWR), and if the items are not returned to the population before the next drawing, then the resulting sampling procedure is known as ***simple random sampling without replacement*** (***SRSWOR***). Thus an *SRSWR* sample may contain a particular population unit more than once, but it is not possible that a particular population unit will appear more than once in an *SRSWOR* sample. In that sense, with replacement sampling has an effect of generating an infinite population. For example, in with replacement sampling, there is nothing to stop one from drawing a sample of size 100 from a population of size 10.

In *SRS* procedure if we draw a sample of size n from a population of size N, we have the following results:

- Probability of inclusion of each population unit in the sample is same for all the population units and equal $\frac{1}{N}$.
- Total number of possible samples

$$= \begin{cases} {}^{N}C_{n}, \text{ in case of } SRSWOR \text{ (order ignored)} \\ {}^{N}P_{n}, \text{ in case of } SRSWOR \text{ (order considered)} \\ N^{n}, \text{ in case of } SRSWR \end{cases}$$

Hence,

- Probability of each possible sample being selected

$$= \begin{cases} 1/{}^{N}C_{n}, \text{ in case of } SRSWOR \text{ (order ignored)} \\ 1/{}^{N}P_{n}, \text{ in case of } SRSWOR \text{ (order considered)} \\ 1/N^{n}, \text{ in case of } SRSWR. \end{cases}$$

To draw a simple random sample from the population of interest is not that trivial as it appears to be. A perfect random sample is difficult to achieve in practice. A simple and quite reliable method of selecting a simple random sample is to use ***tables of random numbers*** (vide Table I in Appendix), where digits are generated in such a way that all ten integers (0, 1, ..., 9) occur randomly (i.e., independently) and approximately with equal frequency. Thus, if one number is selected from a random point in the table, it is equally likely to be any of the digits among 0 - 9. We can use one of the available tables of random numbers to select a random sample, viz., (i) Tippett's series, (ii) Fisher and Yates' series,

(iii) Kendall and Smith's series etc. Presently, several computer programs are available which generate random numbers, and we can use computer generated random numbers for drawing a random sample. The numbers thus generated are subject to several statistical tests for randomness.

In a ***random number table***, successive numbers or sequence of numbers are selected by proceeding in a regular predetermined geometric manner, e.g., moving across the rows of the table or down its columns. Any starting point can be used and one can move in any predetermined direction. If more than one sample is to be selected in any problem, then each should have its own unique starting point. If single digit numbers are to be drawn, then we read just one new number, each time, across the rows or down the columns. If two-digit numbers are needed we read two adjacent numbers, each time, across the rows or down the columns, and so on.

Choosing numbers from the table is analogous to drawing numbers out of a box containing those numbers on pieces of papers, when the pieces of papers are thoroughly mixed before the draws. Suppose, we want a simple random sample of four households to be selected from ten. We could number the households from 1 to 10, write one number on one piece of paper and put all the pieces in a box and mix them thoroughly. Then draw the piece of paper one by one, from the box. Thus we can get a *SRS*. The draws could be made without replacement or with replacement.

4.4.1 Method of drawing a simple random sample

Let us consider a population consisting of N units. First it is necessary to assign a unique identification number to each member of the population. Label the population units by $U_1, U_2, \ldots, U_N$. We want to select a random sample of n units from the above finite population.

Case I: $N \neq 10^k$ (*i.e.*, N *is not of the form* 10^k, k *being a positive integer*)

Let N be a d-digit number. Then draw a d-digit random number from the table of random numbers. Let m = maximum d-digit number which is divisible by N. The range of d-digit random numbers between 0 and m is called the *effective range*. Ignore the random numbers drawn if it is more than m or 0. Then, go for another random number. If the random number drawn is less than or equal to m, then denote it by R. When $R \leq N$, the population unit bearing the serial number 1 to N is included in the sample. When $R > N$, select population units on *mod-N basis*, i.e., divide R by N and take the remainder (say, M) as the serial number of the population

unit to be included in the sample (i.e., U_M will be included in the sample). If $M = 0$, then the N th unit, i.e., U_N is selected.

Case II: $N = 10^k$

If N is a d-digit number, then draw a $(d-1)$-digit random number. Then select the population unit on mod-N basis. In other words, we may say, when N is of the form 10^k (i.e.,10, 100, 1000, ...), select the population unit corresponding to the random number drawn. If the random number drawn is zero, then select the N^{th} unit, U_N.

In case, a sample of size n $(0 < n < N)$ is to be drawn using *SRSWR*, we select n random numbers using the method described above, and choose the respective population units for the sample, no matter if the same unit appears more than once or not.

In case of a *SRSWOR* sample of size n, we select n random numbers using the above method and select the respective population units in the sample. But if some unit is repeated, then keep it once and ignore the rest appearances and go for selecting another random number. Continue this until n distinct units are drawn.

***Example* 4.2** *Assume that there are $N = 1000$ patient records in a hospital from which a sample of size $n = 4$ is to be drawn. Determine which records are to be included in the simple random sample of size* 4, *drawn under without replacement scheme.*

***Solution*:**

Let us number the population units (i.e., patient records) serially, from 1 to 1000. The records are denoted as $U_1, U_2, \ldots, U_{1000}$. We draw 3-digit random numbers one by one, and consider the corresponding population unit to be included in the sample. Here we ignore the repeated occurrence of the same number and continue until four distinct records are selected.

Following 3-digit random numbers are drawn from random number table:

721, 911, 105, 721, 023

Hence the selected records are $U_{721}, U_{911}, U_{105}, U_{23}$.

Above four units constitute the random sample under *SRSWOR* procedure.

Example **4.3** *Select a random sample of size 4 from a population of size 400 using with replacement procedure. The random numbers are given as follows:*

7524 7600 7862 0767 2191 1105 7210 2342

Solution:

Here $N = 400$. Let us first number the population units as 001, 002, ..., 400. Next we draw 3-digit random numbers from the random numbers given (starting from the left-most digit 7 in the given set of random numbers and moving rowwise). The highest 3-digit numbers which is divisible by 400 is 800, i.e., here $m = 800$. If the random number drawn is 000 or it exceeds 800, then ignore it, otherwise, select the number. If the selected number is less than 400, then the corresponding population unit is selected in the sample. If the random number drawn exceeds 400, then divide the number by 400 and consider the remainder, say, M, as the serial number of the population unit to be selected in the sample, (i.e., U_M).

3-digit random numbers drawn are 752, 476, 007, 862, 076.

The numbers reduced under mod-400 basis are 352, 76, 7, 76 (862 is ignored here).

So, the population units U_{352}, U_{76}, U_7, U_{76} are selected in the sample. Since we have adopted a with replacement scheme, so we do not have any problem in including 76th population unit twice.

Here the following table shows how the above method ensures same inclusion probability for each unit:

Random number drawn	Population unit selected	Inclusion probability
001, 401	001	$\frac{2}{800} = \frac{1}{400}$
002, 402	002	$\frac{2}{800} = \frac{1}{400}$
.	.	.
.	.	.
.	.	.
400, 800	400	$\frac{2}{800} = \frac{1}{400}$

Note : Strictly speaking, random sampling numbers are applicable only to draw samples from a finite population. But in some situations we can use random sampling numbers to draw samples from an infinite population like drawing samples from a normal population.

4.4.2 Sampling distribution of sample mean

Result 4.1 *Mean and variance of the sampling distribution of sample mean is as follows:*

(*i*) $E(\bar{x}) = \mu$, *for both SRSWR and SRSWOR*

(*ii*) $V(\bar{x}) = \begin{cases} \dfrac{\sigma^2}{n} & , \text{ in case of SRSWR} \\ \dfrac{\sigma^2}{n}\dfrac{N-n}{N-1} & , \text{ in case of SRSWOR} \end{cases}$

and standard error of $(\bar{x})$, $s.e.(\bar{x}) = \begin{cases} \dfrac{\sigma^2}{\sqrt{n}} & , \text{in case of SRSWR} \\ \dfrac{\sigma^2}{\sqrt{n}} \cdot \sqrt{\dfrac{N-n}{N-1}} & , \text{in case of SRSWOR} \end{cases}$

Note: 1. $\sqrt{\dfrac{N-n}{N-1}}$ is known as ***finite population correction (f.p.c.)*** *factor.*

As $N \to \infty$, $f.p.c. = \sqrt{\dfrac{1-\frac{n}{N}}{1-\frac{1}{N}}} \to 1$. Hence, when $f.p.c. \to 1$, $V(\bar{x})$ will have the value $\dfrac{\sigma^2}{n}$ for both the cases – *SRSWR* and *SRSWOR*. Naturally, $f.p.c. \to 1$ as $N \to \infty$, or, if n is very small compared to N, i.e., $n << N$.

2. A statistic is *unbiased* if the centre of its sampling distribution is at the parameter value, i.e.,the statistic T is said to be an unbiased estimator of a parameter θ if $E(T) = \theta$. Sample mean $\bar{x}$ is an unbiased estimator of population mean μ. Theory of estimation will be discussed in details in next chapter.

3. If $(x_1, x_2, \ldots, x_n)$ be a random sample drawn from a population having probability distribution characterized by the probability function

(*p.m.f.* or *p.d.f.*) $f(x)$, then $x_1, x_2, \ldots, x_n$ will be *i.i.d.* and each of $x_1, x_2, \ldots, x_n$ will also be distributed independently with same probability function.

Result 4.2 *Let* $s^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2$ *be the sample variance with divisor* $(n-1)$. *Then*

$$E(s^2) = \begin{cases} \sigma^2 & , \text{in case of SRSWR} \\ \sigma^2 \cdot \frac{N}{N-1} & , \text{in case of SRSWOR} \end{cases}$$

where σ^2 is the population variance, N is populatiopn size, and n is the sample size.

Note: $s^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2$, the sample variance with divisor $(n-1)$, is an unbiased estimator of population variance σ^2, in case of *SRSWR*.

In case of *SRSWOR*, s^2 will be unbiased estimator of $\frac{N\sigma^2}{N-1}$ and if N is large, then s^2 is an unbiased estimator of σ^2.

***Example* 4.4** *Refer to Example* 4.1 *and the sampling distribution of* $\bar{x}$ *obtained based on an SRSWOR sample of size 2. Obtain* $E(\bar{x})$, $V(\bar{x})$ *and* $s.e.(\bar{x})$.

Solution:

Let us obtain the expectation and variance of the sample mean and verify the result given in Result 4.1.

We find the population mean $\mu = \frac{2+1+0+3+2}{5} = 1.6$ and

the population variance $\sigma^2 = \frac{2^2+1^2+0^2+3^2+2^2}{5} - (1.6)^2 = 1.04$

From the table given below we find $E(\bar{x}) = \sum \bar{x}\, p(\bar{x}) = 1.6$

and $V(\bar{x}) = \sum \bar{x}^2 p(\bar{x}) - E^2(\bar{x}) = 2.95 - (1.6)^2 = 0.39.$

Here, $N = 5$, $n = 2$, $\sigma^2 = 1.04$.

Possible values of $\bar{x}$	Probability $p(\bar{x})$	$\bar{x}\,p(\bar{x})$	$\bar{x}^2\,p(\bar{x})$
0.5	0.1	0.05	0.025
1.0	0.2	0.20	0.200
1.5	0.3	0.45	0.675
2.0	0.2	0.40	0.800
2.5	0.2	0.50	1.250
Total	1.0	1.60	2.950

Then, $\frac{\sigma^2}{n}\cdot\frac{N-n}{N}=\frac{1.04}{2}\cdot\frac{3}{4}=0.39$, which is same as $V(\bar{x})$ as calculated.

Hence the results $E(\bar{x})=\mu$ and $V(\bar{x})=\frac{\sigma^2}{n}\cdot\frac{N-n}{N-1}$ are verified.

Here s.e. of $\bar{x}$ is $+\sqrt{0.39}=0.6245$.

***Example* 4.5** *Suppose that a person has three sons. The number of cars they owned for their business are as follows:*

Son	:	1	2	3
No.of cars (x)	:	4	6	3

The person selects 2 sons at random with replacement from 3. Obtain the sampling distribution of mean number of cars and its expectation and variance. Also verify the result stated in Result 4.1. Obtain the s.e. of mean number of cars.

[For the sake of simplicity of calculation we have considered a very small population. Our objective is to illustrate the idea we have discussed so far in this connection.]

***Solution*:**

Here the population mean, $\mu=\frac{4+6+3}{3}=4.33$ and

population variance, $\sigma^2=\frac{4^2+6^2+3^2}{3}-\left(\frac{13}{3}\right)^2=1.6$.

Here, the total number of possible samples is $3^2 = 9$.

All possible samples, values of the sample means and probabilities are shown in the following table:

Possible Sample	Sample values	Sample mean $\bar{x}$	Probability
(1, 1)	(4, 4)	4	$\frac{1}{9}$
(1, 2)	(4, 6)	5	$\frac{1}{9}$
(1, 3)	(4, 3)	3.5	$\frac{1}{9}$
(2, 1)	(6, 4)	5	$\frac{1}{9}$
(2, 2)	(6, 6)	6	$\frac{1}{9}$
(2, 3)	(6, 3)	4.5	$\frac{1}{9}$
(3, 1)	(3, 4)	3.5	$\frac{1}{9}$
(3, 2)	(3, 6)	4.5	$\frac{1}{9}$
(3, 3)	(3, 3)	3	$\frac{1}{9}$
Total	-	-	1

Hence the sampling distribution of $\bar{x}$ is shown below:

Possible values of $\bar{x}$	Probability
3	1/9
3.5	2/9
4	1/9
4.5	2/9
5	2/9
6	1/9
Total	1

Now we compute mean (or expectation) and variance of sample mean $\bar{x}$ from its sampling distribution:

Possible values of $\bar{x}$	Probability $p(\bar{x})$	$\bar{x}\,p(\bar{x})$	$\bar{x}^2 p(\bar{x})$
3	1/9	1/3	1.00
3.5	2/9	7/9	2.72
4	1/9	4/9	1.78
4.5	2/9	1	4.5
5	2/9	10/9	5.56
6	1/9	6/9	4.00
Total	1	$\frac{39}{9} = 4.33$	19.56

$E(\bar{x}) = \sum_{\bar{x}} \bar{x} p(\bar{x}) = 4.33$, which equals μ, the population mean.

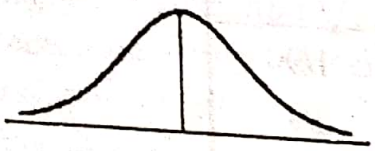
$V(\bar{x}) = \sum_{\bar{x}} \bar{x}^2 p(\bar{x}) - \{E(\bar{x})\}^2 = 19.56 - (4.33)^2 = 0.8.$

Here $N = 3, n = 2, \sigma^2 = 1.6.$

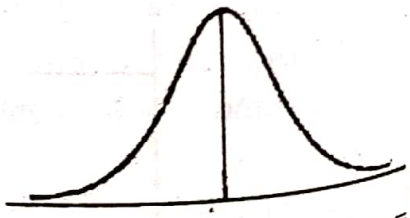
Then $\frac{\sigma^2}{n} = \frac{1.6}{2} = 0.8,$ which is same as $V(\bar{x})$.

Hence, the results $E(\bar{x}) = \mu$ and $V(\bar{x}) = \frac{\sigma^2}{n}$ for *SRSWR*, are verified for this example. Here $s.e.(\bar{x}) = +\sqrt{0.8} = 0.8944$.

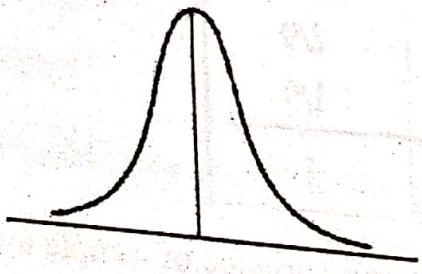
Following figures show the sampling distribution of $\bar{x}$ for different values of n when the samples are drawn from a normal population:



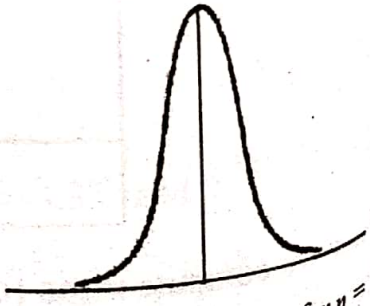
Sampling distribution of $\bar{x}$ for $n = 5$



Sampling distribution of $\bar{x}$ for $n = 16$



Sampling distribution of $\bar{x}$ for $n = 32$



Sampling distribution of $\bar{x}$ for $n = 100$

Note: Clearly, variance of the sampling distribution of $\bar{x}$ depends on n and hence, as the sample size increases, the sampling distribution of $\bar{x}$ becomes more and more clustered around its mean.

4.4.3 Simpling distribution of sample proportion

A population where the population members are classified according to possessing or not possessing a specified characteristic is known as a ***dichotomous population***. Let P be the proportion of individuals in the population possessing a specific characteristic and $Q = 1 - P$ be the proportion of individuals not possessing that characteristic. Many sampling problems are concerned with the proportion of individuals in the population who possess some specific characteristic. The objective is to draw a conclusion about the proportion of individuals or objects in a population that possess the specified character. In such a situation any individual or object that possesses the characteristic of interest is labelled a *success*(S) or a *defective* (D), and one that does not possess the characteristic is termed as a *failure* (F) or a *non-defective* (ND). Let P be the *proportion of* S's (or D's) in the population, e.g., proportion of female population, proportion of smokers in a population etc..

Suppose that a random sample of size n is selected from a population having proportion of individuals possessing a specified characteristic P. Let x be the number of individuals in the sample possessing that characteristic. Then the sample proportion p will be given by $p = \frac{x}{n}$.

Making inference about population proportion P requires knowledge of the sampling distribution of the sample proportion p.

Result 4.3 *Let p be the sample proportion of successes based on a random sample of size n from a population having population proportion of success, P. Then*

(i) $E(p) = P$

(ii) $V(p) = \dfrac{P(1-P)}{n}$, *in case of SRSWR*

$= \dfrac{P(1-P)}{n} \cdot \dfrac{N-n}{N-1}$, *in case of SRSWOR.*

and $s.e.(p) = \sqrt{\dfrac{P(1-P)}{n}}$, *in case of SRSWR*

$= \sqrt{\dfrac{P(1-P)}{n}} \times \sqrt{\dfrac{N-n}{N-1}}$, *in case of SRSWOR.*

Next we discuss an important result, known as *Central Limit Theorem.*

4.10 Central Limit Theorem

When the objective of a study is to make an inference about the population mean, μ, it is natural to consider the sample mean, $\bar{x}$ as the first and most formiddible candidate on which the inference should base.

In order to understand how the inferential procedures based on $\bar{x}$ work, we must study how sampling variablility causes $\bar{x}$ to differ in value from one sample to another. Again, this variablity in the values of $\bar{x}$ will be reflected in its sampling distribution. Sampling distribution of $\bar{x}$ depends on the sample size, n and some other important characteristics of the population, like its shape, mean value μ, standard deviation σ etc.. To study the sampling distribution of $\bar{x}$ we repeat the process of drawing samples of different sizes from the population and observe how the choice of the sample size affects the sampling distribution of $\bar{x}$. We can observe the behaviour of the sampling distribution of $\bar{x}$ for populations of different shapes, symmetric and skewed. We can find the following events:

(i) Regardless of the shape of the population, the sampling distribution of $\bar{x}$ always has a mean identical to the mean of the population and a standard deviation equal to the population standard deviation 'σ' divided by $\sqrt{n}$, i.e.,

$\mu_{\bar{x}}$ = mean of the distribution of $\bar{x} = \mu$,

$\sigma_{\bar{x}}$ = s.d. of the distribution of $\bar{x} = \dfrac{\sigma}{\sqrt{n}}$ = standard error of $\bar{x}$.

Clearly, the spread of the distribution of sample mean is considerably less than the spread of the population from which the sample is drawn. Again, as a consequence the distribution of $\bar{x}$ based on a large n tends to be closer to the population mean μ than does the distribution of $\bar{x}$ based on a small n.

(ii) When the population distribution is normal, the sampling distribution of $\bar{x}$ is also normal for any sample size n.

(iii) When n is sufficiently large ($n \geq 30$), the sampling distribution of $\bar{x}$ is well approximated by a normal distribution, even when the population distribution is not normal. This result is known as ***Central limit theorem.***

> ***Central Limit Theorem (CLT)*** states that when n is sufficiently large, the distribution of $\bar{x}$ is approximately normal with mean μ and s.d. $\sigma/\sqrt{n}$, no matter what the population distribution is provided the population variation is finite.

Now, let us write *CLT* more technically. Let $\bar{X}_n$ be the sample mean of a random sample of size n drawn from a population having probability density function $f(\cdot)$ with mean μ and finite variance σ^2. Let Z_n be defined by

$$Z_n = \frac{\bar{X}_n - E(\bar{X}_n)}{\sqrt{V(\bar{X}_n)}} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}.$$

Then *CLT* states that the distribution of Z_n approaches the standard normal distribution as n approaches infinity.

The result holds irrespective of the form of the population distribution provided that it has a finite variance.

4.9 Sampling Distributions: χ^2, t and F

Let $x_1, x_2, ..., x_n$ be a random sample of size n drawn from a $N(\mu, \sigma^2)$ population. Then the distributions of various statistics used in the context of statistical inference follow some standard distribution, like normal, chi-square t etc.. Sample mean $\bar{x}$ follows $N(\mu, \frac{\sigma^2}{n})$, or, $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ follows ***standard normal distribution***.

Let us now discuss different standard sampling distributions which are useful in determining the distribution of various important statistics.

4.9.1 *Chi-square* (χ^2)*-distribution*

Let $Z_1, Z_2, \cdots, Z_n$ be n independent standard normal variables. Then

$$\chi^2 = \sum_{i=1}^{n} Z_i^2 \sim \chi_n^2,$$

where χ_n^2 denotes a chi-square distribution with n *degrees of freedom.The* ***degrees of freedom*** *(d.f.) of a statistic is the number of independent observations required to define the statistic.* It is actually a parameter of the distribution.

A χ^2-variable with d.f. n has the following p.d.f.:

$$f(\chi^2) = \frac{1}{2^{n/2}\Gamma\left(\frac{n}{2}\right)} e^{-\frac{1}{2}\chi^2} (\chi^2)^{\frac{n}{2}-1}, \quad 0 < \chi^2 < \infty$$

Chi-square distribution is identified by its degrees of freedom and the following figure shows how the distribution looks like for different values of n.

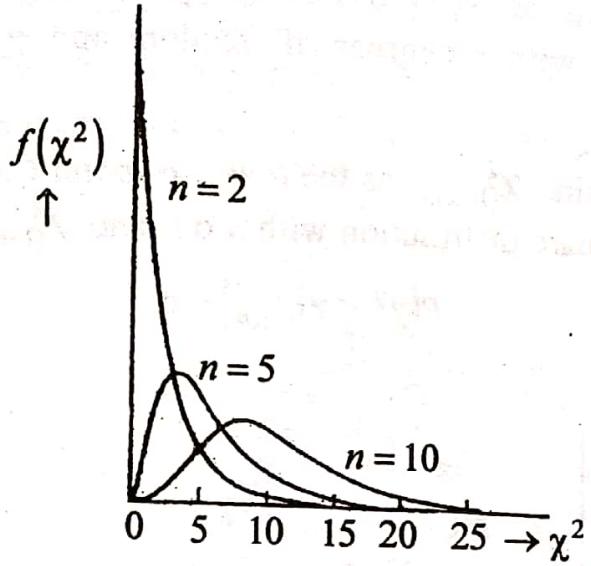


Fig. 4.1 *Shape of the density curve of different chi-square variables*

Important properties of a χ_n^2-distribution

1. $E(\chi^2)=n$ (d.f.), $V(\chi^2)=2n$, when χ^2 is a chi-square variate with n d.f.. Note that $\chi^2>0$ always.

2. χ_n^2-distribution is always positively skewed.

3. The curve has a single peak. The peak shifts to the right and the curve gets flatter when n, the degrees of freedom, increases.

4. For large n, $\sqrt{2\chi^2}$ can be shown to be approximately normal with mean $\sqrt{2n-1}$ and $s.d.$ = 1, i.e., $\sqrt{2\chi^2} \sim N(\sqrt{2n-1}, 1)$ approximately, if n is large.

5. If χ_1^2 and χ_2^2 are two independent chi-square variables with n_1 and n_2 degrees of freedom, respectively, then $\chi_1^2+\chi_2^2 \sim \chi_{n_1+n_2}^2$.

This result can be extended to n independent chi-square variables.

6. Let $X_1, X_2, \cdots, X_n$ be a random sample of size n drawn from a normal population having mean μ and standard deviation σ. Then

(i) $\dfrac{(n-1)s^2}{\sigma^2}$ follows a chi-square distribution with $(n-1)$ degrees of freedom, where s^2 = sample variance $= \dfrac{1}{(n-1)}\sum_{i=1}^{n}(x_i-\bar{x})^2$.

(ii) $\bar{x}$ and s^2 are independently distributed.

We denote $\chi^2_{\alpha,n}$ as the *upper* α*-point* or *upper* $100\alpha\%$ *point* of a chi-square distribution with n degrees of freedom and $\chi^2_{\alpha,n}$ is such that $P(\chi^2 > \chi^2_{\alpha,n}) = \alpha$.

Similarly we define $\chi^2_{(1-\alpha),n}$ as the *lower* α*-point* (i.e., *upper* $(1-\alpha)$ *point*) of the chi-square distribution with n d.f. and $\chi^2_{(1-\alpha),n}$ is such that

$$P(\chi^2 < \chi^2_{1-\alpha,n}) = \alpha$$

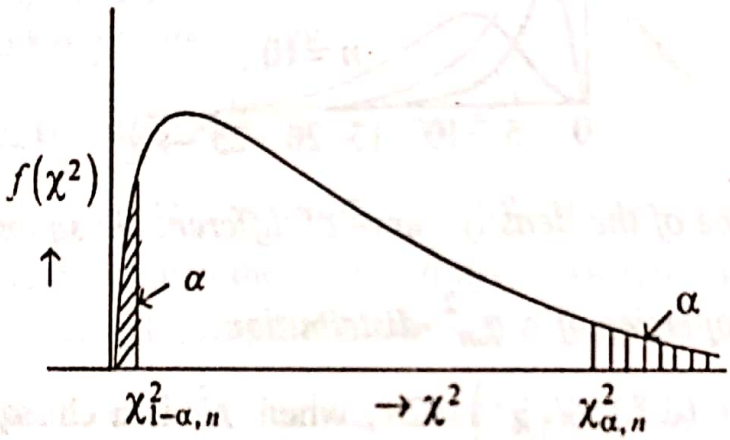


Fig. 4.2 *Percentage points* $\chi^2_{\alpha,n}$ *and* $\chi^2_{1-\alpha,n}$ *of the chi-square distribution with n d.f.*

4.9.2 *t-distribution*

Let Z be a standard normal variable and χ^2 be a chi-square variable with n degrees of freedom. Further assume that Z and χ^2 are independently distributed. Then the new variable defined as

$$t = \frac{Z}{\sqrt{\chi^2/n}} \sim t_n,$$

where t_n denotes a t-distribution with n degrees of freedom. The p.d.f. of t is given by

$$f(t) = \frac{1}{n^{\frac{1}{2}} B\left(\frac{1}{2}, \frac{n}{2}\right)} \frac{1}{\left(1+\frac{t^2}{n}\right)^{\frac{n+1}{2}}}, \quad -\infty < t < \infty$$

The t-distribution was first studied by W.S. Gosset, a British Chemist in 1908 and published his work under the pseudonym 'Student'. That is why the t-distribution is known as *Student's t-distribution.*

Like normal distribution, t-distribution is symmetrical about $t = 0$ and bell shaped, but it has heavier tails than the standard normal curve.

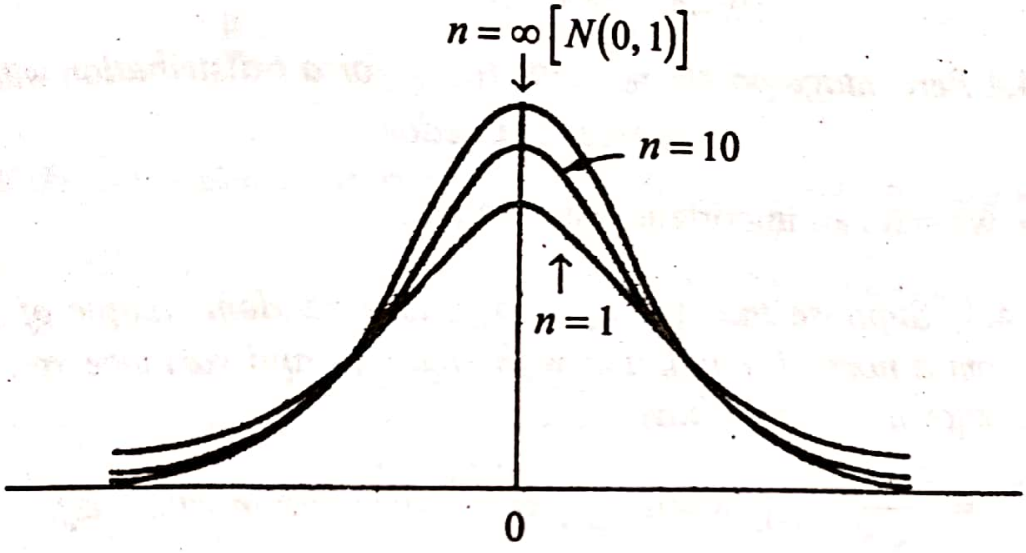


Fig.4.3 *Shape of the density curve of t-distributions with different degrees of freedom*

t-distribution has only one parameter which is its *degrees of freedom* (d.f.).

For a t_n-distribution,

$$E(t) = 0$$
$$V(t) = \frac{n}{n-2}.$$

For a large value of n, the t-distribution approaches normal distribution.

We denote $t_{\alpha,n}$ as the *upper* α*-point* or *upper* $100\alpha\%$ *point of a t-distribution with n degrees of freedom,* where $P(t > t_{\alpha,n}) = \alpha$

Similarly, we define the *lower* α*-point* or *upper* $100(1-\alpha)\%$ *point* $t_{1-\alpha,n}$ as $P(t < t_{1-\alpha,n}) = \alpha$.

Due to symmetry of the distribution $t_{1-\alpha,n} = -t_{\alpha,n}$.

For large value of n $(n \geq 30)$ the value $t_{\alpha,n}$ may be well approximated by z_α, where z_α is the upper α-point of a standard normal variable.

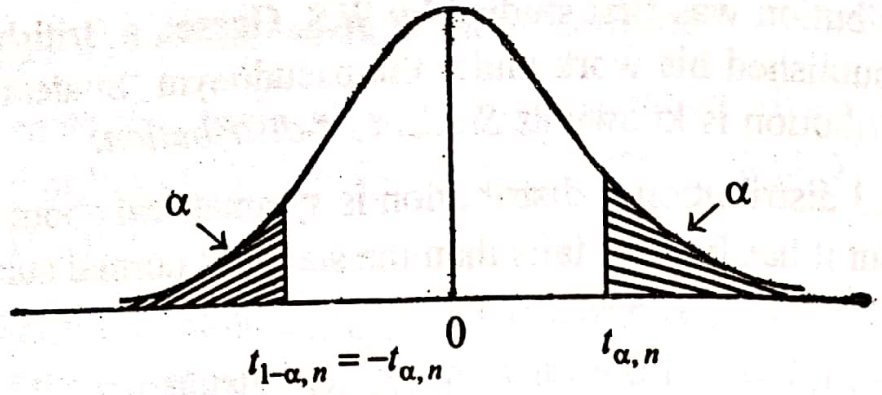


Fig.4.4 *Percentage points* $t_{\alpha,n}$ *and* $t_{1-\alpha,n}$ *for a t-distribution with* n *degrees of freedom*

Now we state an important result below:

Result 4.4 *Suppose that* $(x_1, x_2, \cdots, x_n)$ *is a random sample of size* n *drawn from a normal population with mean* μ *and variance* σ^2, *when* σ^2 *is unknown. Then we have*

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t_{n-1}, \text{ where } t_{n-1} \text{ is a t-distribution with } n \text{ d.f.}$$

As the degrees of freedom n increases, the density curve of t approaches the $N(0,1)$ density more closely. This happens because s^2 estimates σ^2 more accurately as the sample size increases.

4.9.3 *F-distribution*

Let χ_1^2 and χ_2^2 be two independent chi-square variables with m and n degrees of freedom, respectively. Then the random variable defined as the ratio of two mean chi-square variables follows an F-distribution, with d.f. m and n, i.e.,

$$F = \frac{\chi_1^2/m}{\chi_2^2/n}$$

and $F \sim F_{m,n}$, where $F_{m,n}$ denotes a F-distribution with m and n degrees of freedom.

The p.d.f. of F is given by

$$f(F) = \frac{\left(\frac{m}{n}\right)^{\frac{m}{2}}}{B\left(\frac{m}{2}, \frac{n}{2}\right)} \times \frac{F^{\frac{m}{2}-1}}{\left(1 + \frac{m}{n}F\right)^{\frac{m+n}{2}}}, \quad 0 < F < \infty$$

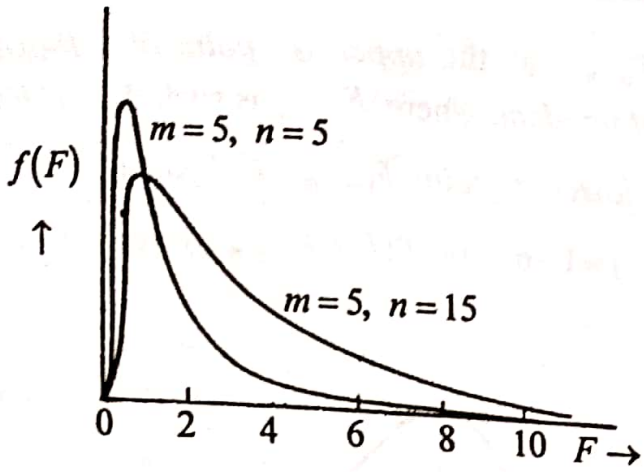


Fig.4.5 *Shape of density curves of F-distributions with different m, n*

F-distribution is named after Sir Ronald Fisher. *F*-distribution has two numbers of degrees of freedom: the numerator degrees of freedom (m) and the denominator degrees of freedom (n). These two numbers representing two degrees of freedom are the parameters of the *F*-distribution. Each combination of degrees of freedom for the numerator and for the denominator gives a different *F*-distribution curve. *F*-distribution is a continuous distribution and skewed to the right. But the skewness decreases as the number of degrees of freedom increases.

For an $F_{m,n}$-variable,

$$\text{Mean} = \frac{n}{n-2} \quad (\text{free of } m)$$

$$\text{variance} = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$$

$$\text{mode} = \frac{n(m-2)}{m(n+2)} = \frac{n}{n+2} \times \frac{m-2}{m},$$

since $F > 0$, mode exists if and only if $m > 2$ and mode is always less than unity.

Note:

1. Again if $m = 1$, then $F_{1,n}$-distribution reduces to the distribution of t^2, where $t \sim t_n$.
2. Also note that if F follows $F_{m,n}$-distribution, then $F' = \frac{1}{F}$ follows $F_{n,m}$-distribution.

We denote $F_{\alpha;m,n}$ as the *upper* α *- point of a F-distribution with* (m,n) *degrees of freedom*, where $F_{\alpha;m,n}$ is such that $P\left(F > F_{\alpha;m,n}\right) = \alpha$.

Similarly, the *lower* α *-point* $F_{1-\alpha;m,n}$ *is defined as*

$$P(F > F_{1-\alpha;m,n}) = 1-\alpha, \text{ or } P(F < F_{1-\alpha;m,n}) = \alpha.$$

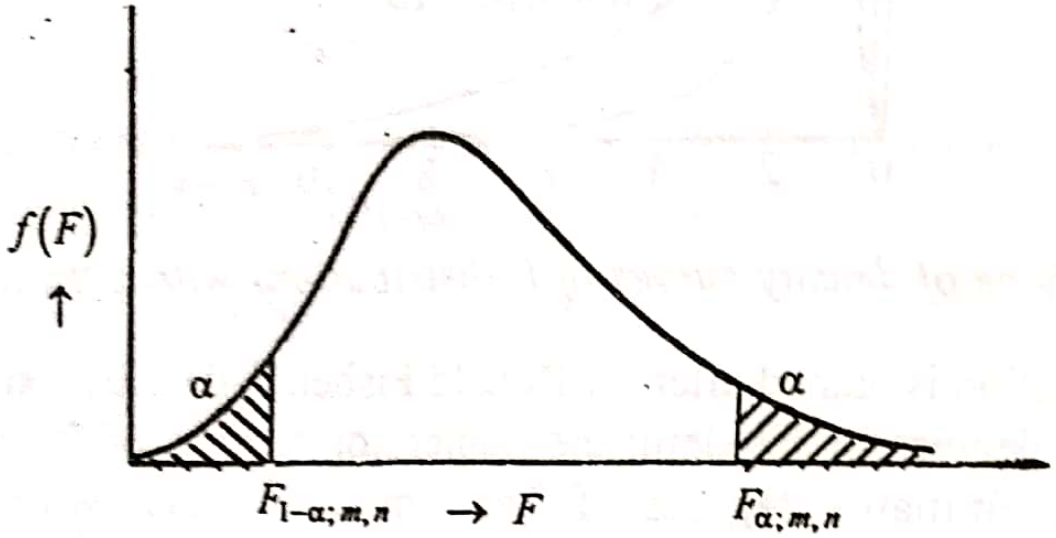


Fig. 4.6 *Percentage points of a* $F_{m,n}$ *-distribution*

Note that $F_{1-\alpha;m,n} = \dfrac{1}{F_{\alpha;\,n,m}}$.

This is the relation between the lower α -point of $F_{m,n}$ -distribution and the upper α -point of $F_{n,m}$ -distribution. This relation is very important and used to obtain the lower α -point of a F-distribution from F-table given in Appendix.