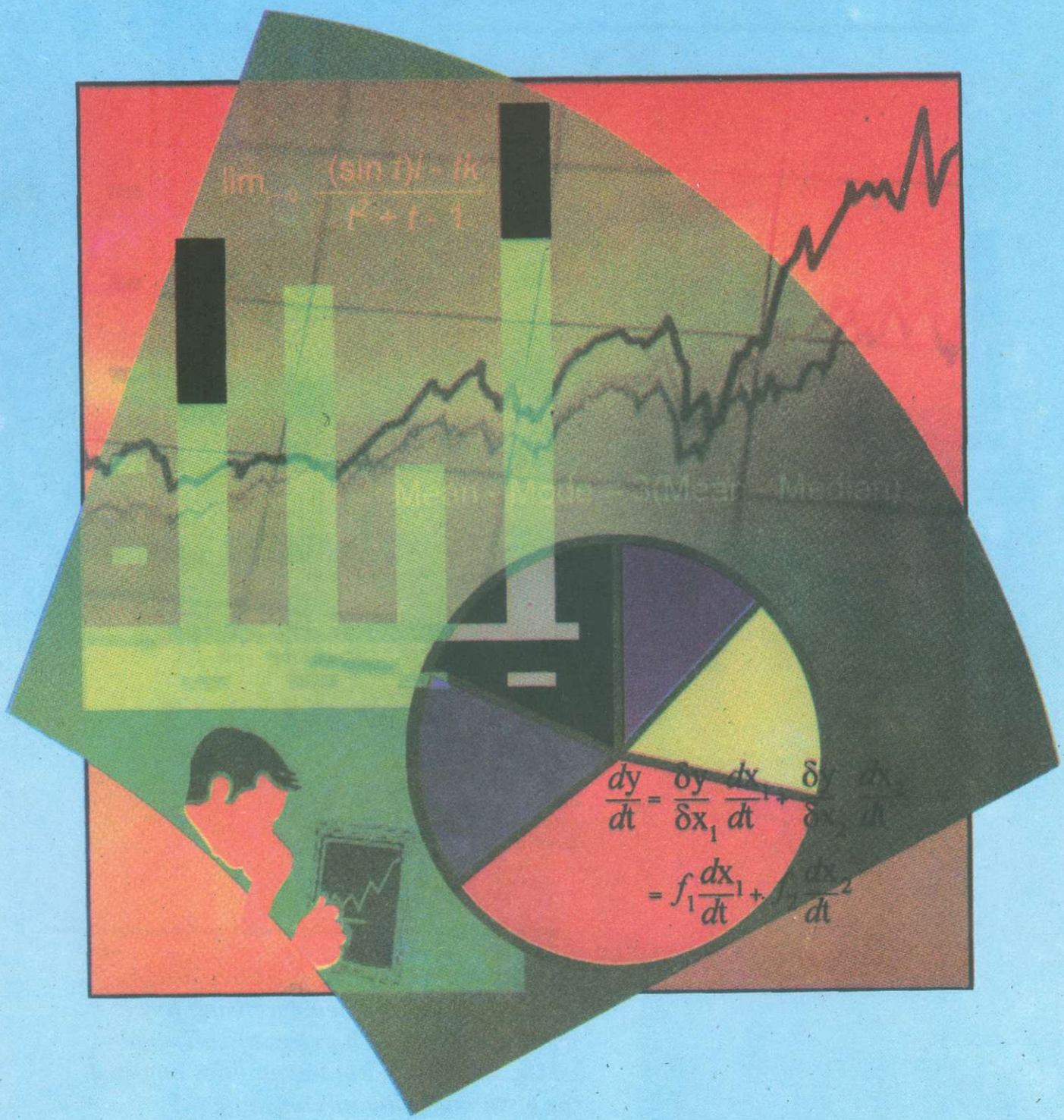
ignou
THE PEOPLE'S
UNIVERSITY

Indira Gandhi National Open University
School of Social Sciences

MEC-103
QUANTITATIVE METHODS



Statistical Methods-II

6

ignou
THE PEOPLE'S
UNIVERSITY

MEC-103
Quantitative Methods

Indira Gandhi
National Open University
School of Social Sciences

Block

6

STATISTICAL METHODS-II

MEC-103-6-10F-7K

Expert Committee

Prof. Bhaswar Moitra
Department of Economics
Jadavpur University
Kolkata

Dr. Naresh Kumar Sharma
University of Hyderabad
School of Economics, Hyderabad

Dr. Anirban Kar
Deptt. of Economics
Delhi School of Economics
University of Delhi

Dr. Indrani Roy Chowdhury
Economics Faculty, Jamia Millia Islamia
New Delhi

Prof. Sudhir Shah
Deptt. of Economics
Delhi School of Economics
University of Delhi

Dr. Manish Gupta
NIPFP,
Special Institutional Area, New Delhi

Prof. Gopinath Pradhan
School of Social Sciences
Indira Gandhi National Open University
New Delhi

Prof. Narayan Prasad
School of Social Sciences
Indira Gandhi National Open University
New Delhi

Prof. Kaustuva Barik
School of Social Sciences
Indira Gandhi National Open University
New Delhi

Prof. B.S. Prakash
School of Social Sciences
Indira Gandhi National Open University
New Delhi

Mr. Saugato Sen
School of Social Sciences
Indira Gandhi National Open University
New Delhi

Course Editor : Prof. Gopinath Pradhan

Programme Coordinators

Prof. Gopinath Pradhan, IGNOU, New Delhi

Prof. Kaustuva Barik, IGNOU, New Delhi

Course Coordinator

Prof. Gopinath Pradhan

Block Preparation Team

Unit	Writer
18, 19 and 20	Biswadeep Basu

PRINT PRODUCTION TEAM

Mr. S. Burman
D.R.(P) MPDD-IGNOU,
Maidan Garhi, Delhi-110068

Mr. Tilak Raj
A.R.(P) MPDD-IGNOU,
Maidan Garhi Delhi-110068

Mr. Yashpal Sharma
S.O.(P) MPDD-IGNOU,
Maidan Garhi Delhi-110068

November 2018 (Reprint)

ISBN : 978-93-86375-17-9

Further information on the Indira Gandhi National Open University Courses may be obtained from the University's Office at Maidan Garhi, New Delhi-110 068 or visit University's http:// www.ignou.ac.in.

Printed and Published on behalf of the Indira Gandhi National Open University, New Delhi by Registrar, MPDD. IGNOU

Printed at: A-One Offset Printers, 5/34, Kirti Nagar Indl. Area, New Delhi-110015

MEC-103-6-10F-7K

BLOCK 6 STATISTICAL METHODS-II

Introduction

This block extends the statistical framework already given for data presentation and discusses methods of data collection, their analysis and inference drawing techniques. **Unit 18** deal with sampling theory covering planning, designing and types of samples; distribution of sample statistic and standared error derivation. Important sampling distributions (discrete and continuous) such as Binomial, Poission, normal, chi-square, t and F are documented in **Unit 19** that help analyst testing the sample coefficients. The last unit of the block **Unit 20** gives the procedure for deriving the tools used for statistical inference. The estimation theory is presented delineating the characteristics of a good estimator and hypothesis formulation and testing. Moreover, the themes such as fixation of level of significance, confidence interval setting, caution against types of errors and critical values for significance test are dealt with to help draw inference from data analysis.

UNIT 18 SAMPLING THEORY

Structure

18.0 OBJECTIVES

After going through this unit, you will be able to answer the following:

- what is sample survey and what are the advantages of it over the total enumeration;
- how to design a sample and what the probable biases that can occur in conducting sample survey;
- different types of sampling and their relative merits and demerits;
- a brief idea of parameter, statistic and standard error; and
- expectations and standard deviation of sample mean and proportion.

18.1 INTRODUCTION

Before giving the notion of sampling, we'll first define population. In a statistical investigation interest generally lies in the assessment of the general magnitude and the study of variation with respect to one or more characteristics

relating to individuals belonging to a group. The group of individuals under study is called population or universe. Thus, in statistics, population is an aggregate of objects, animate or inanimate, under study. The population may be finite or infinite.

It is obvious that for any statistical investigation complete enumeration of the population is rather impracticable. For example, if we want to have an idea of the average per capita (monthly) income of the people in India, we will have to enumerate all the earning individuals in the country, which is rather a very difficult task.

If the population is infinite, complete enumeration is not possible. Also if the units are destroyed in the course of inspection (e.g., inspection of crackers, explosive materials etc.), 100% inspection, though possible, is not at all desirable. But even if the population is finite or the inspection is not destructive, 100% inspection is not taken recourse to because of multiplicity of causes viz., administrative and financial complications, time factor, etc.; in such cases, we take the help of sampling.

A finite subset of statistical individuals in a population is called a sample and the number of individual in a sample is called the sample size.

For the purpose of determining population characteristics, instead of enumerating the entire population, the individuals in the sample are only observed. Then the sample characteristics are utilised to approximately determine or estimate the population. For example, on examining the sample of a particular stuff we arrive at a decision of purchasing or rejecting that stuff. The error involved in such approximation is known as sampling error and is inherent and unavoidable in any sampling scheme. But sampling results in considerable gains, especially in time and cost not only in respect of making observations of characteristics but also in the subsequent handling of the data.

Sampling is quite often used in our day-to-day practical life. For example, in a shop we asses the quality of sugar, wheat or any other commodity by taking a handful it from the bag and then decide to purchase or not. A housewife normally tests the cooked products to find if they are properly cooked and contain the proper quantity of salt.

18.2 ADVANTAGE OF SAMPLE SURVEY

Sample survey has some significant advantages over doing complete enumeration or census study. The following are the some of the advantages of sample survey:

i) Reduction of cost: Since size of the sample is far less than the entire population, so to do the sample survey less number of staff and time is required that reduces the cost associated with it.

ii) Better scope for information: In the sample survey, the surveyor has the scope of interacting more with the sample households, thus can have better information in any particular issue than that in census method. In the census method due to time constraint and financial inadequacy, the surveyor cannot afford much time to any particular household to get better information.

iii) Better quality of data: In the census method, due to time constraint, we do not get good quality of data. But in sample survey, one can have a better quality of data as the survey consists of all the information related to objective of the study.

iv) Gives an idea of the error: For the population, we do not have a standard error, but for the sample we do have a standard error. Given the information of the sample mean and standard error, we can construct the limit within which almost all the sample value will lie.

v) Lastly, the population may be hypothetical or infinite. So to avoid the problems associated with complete enumeration, sample survey is the best alternative to do any statistical analysis.

18.3 SAMPLE DESIGNS

Two things to be required to plan a survey – one is validity, i.e., we must review the valid answers to the questions that we are looking at. And the second is optimising cost and efficiency. It is very obvious that cost increases with the sample size. Whereas efficiency, which is measured by inverse of variance of the estimator (e.g., $v(\bar{x}_n)=\sigma^2/n$) decreases with the sample size. If the sample size increases, then values tend to stick around a central value, thus variance decreases. Now from the cost point of view less number of samples is desirable while from efficiency point of view large number of samples is desirable. So given these two opposite facts, we have to design the sample in order to collectively optimise the constraints.

Sample survey is done in three different stages. The first and the foremost is the planning stage which includes –

- **Defining the objective:** The prime most important thing is to determine the objective of the survey, otherwise, the process cannot be initiated.
- **Defining the population:** It is necessary to define the population of which sample is to be collected so as to make the survey easier otherwise, extra cost will be incurred for an expanded sample set.
- **Determination of the data to be collected:** Before starting the survey the target group has to define, otherwise, the sample collected would not be the representative one.
- **Determining of the method of collecting data:** There can be two methods of collecting data; questionnaire method and interview method. Both of these methods have some demerits. In the questionnaire method the responder may not respond at all or may respond partly. In that case, those observations have to be excluded for better analytical results though there can be a risk of inadequate sample observations.
- **Choice of the sampling units:** Sampling unit has to be chosen on the basis of the objective so that surveying can be done easily.
- **Designing the survey:** It has two parts, (i) conducting a pilot survey, where small scale survey is done before the original survey so as to have a brief idea about the survey; and (ii) deciding on the flexible variables, where

target group should be chosen so as to capture the exact information as far as possible.

- **Drawing the sample:** The easiest way to draw a sample is to identifying each sample unit with a given number; then putting the numbers in a urn and mix them up and draw out the required number of sample size.

18.4 BIASES IN THE SURVEY

There can be two types of biases in the sample survey –

i) **Procedural bias:** Procedural bias can be in the form of response bias, where people do not tend to respond properly; observational bias, where sample chosen are not representative of the population. Very often either of the two occurs. There can be other types of procedural biases also, like, non-response bias, where people do not respond at all; and interviewer bias, where the interviewer collects the information with a biased frame of mind.

ii) **Sampling bias:** There can be three types of sampling biases: (i) wrong choice of type of sampling, where collected information may not have statistical significance; (ii) wrong choice of the statistic, where test statistic chosen is not statistically correct; and (iii) wrong choice of the sampling units, which could make the sampling difficult to conduct.

Check Your Progress 1

1) List the advantages of sample survey.

.....

.....

.....

.....

2) What do you mean by sample designs?

.....

.....

.....

.....

3) What could be the types of bias you face in sample survey?

.....

.....

.....

.....

18.5 TYPES OF SAMPLING

Some of the commonly known and frequently used types of sampling are:

(i) Purposive sampling, (ii) Random sampling, (iii) Stratified sampling, and (iv) Systematic sampling.

Let us explain these terms precisely.

Purposive sampling

Purposive sampling is one in which the sampling units are selected with definite purpose in view. For example, if we want to show that the standard of living has increased in the city, New Delhi, we may take individuals in the sample from rich and posh localities and ignore the localities where low-income group and the middle class families live. This sampling suffers from the drawback of favouritism and nepotism and does not give a representative sample of the population.

Stratified sampling

Here the entire heterogeneous population is divided into a number of homogeneous groups, usually termed as '*strata*', which differ from one another but each of these groups is homogeneous within itself. Then units are sampled in random from each of these stratum, the sample size in each stratum varies accordingly to the relative importance of the stratum in the population. The sample, which is the aggregate of the sampled units of each of the stratum, is termed as stratified sample and the technique of drawing this sample is known as stratified sampling. Such a sampling is by far the best and can safely be considered as representative of the population from which it has been drawn.

Random sampling

In this case, the sample units are selected at random and the drawback of purposive sampling, viz., favouritism or subjective element, is completely overcome. A random sample is one in which each unit of population has an equal chance of being included in it. Suppose we take a sample of size n from a finite population of size N. Then there are ${}^{N}C_{n}$ possible samples. A sampling technique in which each of the ${}^{N}C_{n}$ samples has an equal chance of being selected is known as random sampling and the sample obtained by this technique is termed as random sample.

Proper care has to be taken to ensure that the selected sample is random. Human bias, which varies from individual to individual, is inherent in any sampling scheme administered by human beings. Fairly good random samples can be obtained by the use of *Tippet's random numbers tables* or by throwing of a dice, draw of a lottery, etc. The simplest method, which is normally used, is the lottery system; it is illustrated below by means of an example.

Suppose we want to select r candidates out of n. we assign the numbers 1 to n, one number to each candidate and write this numbers (1 to n) on n slips which are made as homogeneous as possible in shape, size, etc. These slips are then put in a bag and thoroughly shuffled and the r slips are drawn one by one. The r candidates corresponding to the number on the slips drawn will constitute the random sample.

Note: *Tippet's random number tables* consist of 10400 four-digit numbers, giving in all 10400 × 4, i.e., 41600 digits, taken from the British census reports. These tables have proved to be fairly random in character. Any page of the table is selected at random and the number in any row or column or diagonal selected at random may be taken to constitute the sample.

Simple sampling

Simple sampling is random sampling in which each unit of the population has an equal chance, say 'p', of being included in the sample and that this probability is independent of the previous drawings. Thus, a simple sample of size n from a population may be identified with a series of n independent trials with constant probability 'p' of success for each trial.

Note: It should be noted that random sampling does not necessarily imply simple sampling though, obviously, the converse is true. For example, if an urn contains 'a' white balls and 'b' black balls, the probability of drawing a white ball at the first draw is $[a/(a+b)] = p_1$ (say) and if this ball is not replaced the probability of getting a white ball in the second draw is $[(a-1)/(a+b-1)] = p_2 \neq p_1$. This sampling is not simple, but since in the first draw each white ball has the same chance, viz. a/(a+b), of being drawn and in the second draw again each white ball has the same chance, viz. (a-1)/(a+b-1), of being drawn, the sampling is random. Hence in this case, the sampling, though random, is not simple. To ensure the sampling is simple, it must be done with replacement, if population finite. However, in case of infinite population no replacement is necessary.

18.6 PARAMETER AND STATISTIC

In order to avoid verbal confusion with the statistical constants of the population, viz., mean (μ), variance (σ^2), etc., which are usually referred to as 'parameters', statistical measure computed from the sample observations alone, e.g., mean ($\bar{x}$), variance (s^2), etc., have been termed by Professor R. A. Fischer as 'statistic'.

In practice, parameter values are not known and the estimates based on the sample values are generally used. Thus, statistic, which may be regarded as an estimate of parameter, obtained from the sample, is a function of the sample values only. It may be pointed out that a statistic, as it is based on sample values and as there are multiple choices of the samples that can be drawn from a population, varies from sample to sample. These differences in the values of a 'statistic' are called 'sampling fluctuations'. The determination of the characterisation of variation (in the values of the statistic obtained from different samples) that may be attributed to chance or fluctuations of sampling is one of the fundamental problems of the sampling theory.

Note: Now onwards, μ and σ^2 will refer to the population mean and variance respectively while the sample mean and variance will be denoted by $\bar{x}$ and s^2 respectively.

Check Your Progress 2

1) List the important types of sampling.

.....

.....

.....

.....

2) Distinguish between random and stratified sampling.

.....

.....

.....

.....

3) Differentiate between the parameter and statistic.

.....

.....

.....

.....

18.7 SAMPLING DISTRIBUTION OF A STATISTIC

If we draw a sample of size 'n' from a given finite population of size 'N', then the total number of possible samples is:

$$^{N}C_{n} = N!/\{n!\,(N-n)!\} = k, \text{ (say)}.$$

If for each sample, the value of the statistic is calculated, a series of values of the statistic will be obtained. If the number of sample is large, these may be arranged into frequency table. The frequency distribution of the statistic that would be obtained if the number of samples, each of the same size ('n'), were infinite is called the 'sampling distribution' of the statistic. In the case of random sampling, the nature of the sampling distribution of a statistic can be deducted theoretically, provided the nature of the population is given, from considerations of probability theory.

Like any other distribution, a sampling distribution may have its mean, standard deviations and moment of higher orders. Of particular importance is the standard deviation, which is designated as the 'standard error' of the statistic. As an illustration, in the next section we derive for the random sampling the means (expectations) and standard errors of a sample mean and sample proportion.

Some people prefer to use 0.6745 times the standard error, which is called the 'probable error' of the statistic. The relevance of the probable error stems from the fact that for a normally distributed variable x with mean μ and s.d σ,

$P[\mu - 0.6745\sigma \leq x \leq \mu + 0.6745\sigma] = 0.5$ (approximately).

18.8 STANDARD ERROR

The standard deviation of the sampling distribution of a statistic is known as its 'standard error', abbreviated as S.E. The standard errors of some of the well known statistics, for large samples, are given below, where 'n' is the sample

size, σ^2 is the population variance, and P is the population proportion, and Q = 1-P. n_1 and n_2 represent the sizes of two independent random samples respectively drawn from the given population(s).

Sl. No.	Statistic	Standard Error
1.	Sample mean: $\bar{x}$	$\sigma/\sqrt{n}$
2.	Observed sample proportion: 'p'	$\sqrt{\frac{PQ}{n}}$
3.	Sample s.d: 's'	$\sqrt{\sigma^2/2n}$
4.	Sample variance: s^2	$\sigma^2\sqrt{2/n}$
5.	Sample quartile	$1.36263\,\sigma/\sqrt{n}$
6.	Sample median	$1.25331\,\sigma/\sqrt{n}$
7.	Sample correlation coefficient	$(1-p^2)/\sqrt{n}$, p being the population correlation coefficient
8.	Sample moments μ_3	$\sigma^3\sqrt{96/n}$
9.	Sample moments μ_4	$\sigma^4\sqrt{96/n}$
10.	Sample coefficient of variation (v)	$\frac{v}{\sqrt{2n}}\sqrt{1+\frac{2v^3}{10^4}} \approx \frac{v}{\sqrt{2n}}$
11.	Difference of two sample means: $(\bar{x}_1 - \bar{x}_2)$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
12.	Difference of two sample s.d.'s: (s_1-s_2)	$\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}$
13.	Difference of the two sample proportions: (p_1-p_2)	$\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$

18.8.1 Utility of Standard Error

S.E plays a significant role in the large sample theory and forms the basis of the testing of hypothesis.

i) The magnitude of the standard error gives an index of the precision of the estimate of the parameter. The reciprocal of the standard error is taken as the measure of reliability or precision of the statistic.

ii) S.E. enables us to determine the probable limits within which the population parameter may be expected lie.

Check Your Progress 3

1) What are the mean and standard deviation of the sampling distribution of the mean?

.....

.....

.....

.....

2) What is a standard error and why is it important?

.....

.....

.....

.....

3) In a random sample of 400 students of the university teaching departments, it was found that 300 students failed in the examination. In another random sample of 500 students of the affiliated colleges, the number of failures in the same examination was found to be 300. Find out the S. E of the difference between proportion of failures in the university teaching departments and that of in the university teaching departments and affiliated colleges taken together.

.....

.....

.....

.....

18.9 EXPECTATION AND STANDARD ERROR OF SAMPLE MEAN

Suppose a random sample of size 'n' is drawn from a given finite population of size 'N'. Let X_α (α = 1, 2, ..., N) be the value of the variable x for the α^{th} member of the population. Then the population mean of x is $\mu = (1/N)\sum_\alpha X_\alpha$, and the population variance is $\sigma^2 = (1/N)\sum_\alpha (X_\alpha - \mu)^2$.

Again, let us denote by x_i (i=1, 2, ... , n) the value of x for the i^{th} member(i.e., the member selected at the i^{th} drawing) of the sample. The sample mean of x is then $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. For deriving the expectation and standard error of $\bar{x}$, we may consider two distinct cases:

Case I: Random sampling with replacement

For further correspondence, let us recall the following two theorems of the probability theory; (i) If y = bx, then E(y) = bE(x), and (ii) If x and y be two random variables and z a third random variable such that z = x + y, then E (z) = E(x) + E(y).

So from the above two results, it can be written that

$$E(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) \text{ and } \operatorname{var}(\bar{x}) = E\{\bar{x} - E(\bar{x})\}^2$$

$$= (1/n^2)\, E\left[\sum_i \{x_i - E(x_i)\}\right]^2$$

$$= (1/n^2)\sum_i E\{x_i - E(x_i)\}^2 + (1/n^2)\sum_{\substack{i,j \\ i\neq j}} E\{x_i - E(x_i)\}\{x_j - E(x_j)\}$$

$$= (1/n^2)\sum_i \operatorname{var}(x_i) + (1/n^2)\sum_{\substack{i,j \\ i\neq j}} \operatorname{cov}(x_i, x_j)$$

To obtain E (x_i) and var (x_i), we note that x_i can assume the values $X_1, X_2, \ldots, X_n$, each with probability (1/N).

Hence $E(x_i) = \sum_\alpha X_\alpha\, P[x_i = X_\alpha] = \sum_\alpha X_\alpha \times (1/N) = \mu$ and

$$\operatorname{var}(x_i) = E(x_i - \mu)^2 = P[x_i = X_\alpha] = \sum_\alpha (X_\alpha - \mu)^2 \times (1/N) = \sigma^2 \text{ for each } i$$

Again $\operatorname{cov}(x_i, x_j) = E(x_i - \mu)(x_j - \mu)$

$$= \sum_{\alpha,\alpha'} (X_\alpha - \mu)(X_{\alpha'} - \mu)\, P[x_i = X_\alpha, x_j = X_{\alpha'}]$$

Since in sampling with replacements the composition of the population remains the same throughout the sampling process, x_j can take any one of the values $X_1, X_2, \ldots, X_N$, with probability (1/N), irrespective of the value taken by x_i. In other words, for $i \neq j$, x_i and x_j are independent, so that

$$P[x_i = X_\alpha, x_j = X_{\alpha'}] = P[x_i = X_\alpha]\, P[x_j = X_{\alpha'}] = (1/N^2)$$

Hence, $\operatorname{cov}(x_i, x_j) = (1/N^2)\sum_{\alpha,\alpha'} (X_\alpha - \mu)(X_{\alpha'} - \mu)$

$$= (1/N^2)\sum_{\alpha} (X_\alpha - \mu)\sum_{\alpha'} (X_{\alpha'} - \mu) = 0$$

For each i,j ($i \neq j$), since $\sum_\alpha (X_\alpha - \mu) = \sum_{\alpha'} (X_{\alpha'} - \mu)$, being the sum of the deviations of $X_1, X_2, \ldots, X_n$ from their mean is zero.

Hence, we have, finally, $E(\bar{x}) = (1/n) \times n\mu = \mu$

and var($\bar{x}$) = $(1/n^2)\, n\sigma^2 + (1/n^2)\, n(n-1) \times 0 = \sigma^2/n$

The standard error of $\bar{x}$ is, therefore, $\sigma_x = \sigma/\sqrt{n}$

Case II: Random sampling without replacement

As before for each i, E (x_i) = μ and var (x_i) = σ^2,

since here too x_i can take any one of the values $X_1, X_2, \ldots, X_N$, with the same probability (1/N). The covariance term, however, needs special attention.

Here, for $i \neq j$

$P[x_i = X_\alpha, x_j = X_{\alpha'}] = P[x_i = X_\alpha]\, P[x_j = X_{\alpha'} \mid x_i = X_\alpha] = (1/N)(1/(N-1))$ if $\alpha \neq \alpha'$ [since x_j can take any value except X_α, the value which is known to have been already assumed by x_i, with equal probability 1/(N–1)] = 0 if $\alpha = \alpha'$

Hence, cov $(x_i, x_j) = (1/N(N-1)) \sum_{\substack{\alpha,\alpha' \\ \alpha \neq \alpha'}} (X_\alpha - \mu)(X_{\alpha'} - \mu) = (1/N(N-1))$

$$\sum_{\alpha,} (X_\alpha - \mu)\{\sum_{\alpha'} (X_{\alpha'} - \mu) - (X_\alpha - \mu)\} = (1/N(N-1))$$

$$\{\sum_{\alpha,} (X_\alpha - \mu) \sum_{\alpha'} (X_{\alpha'} - \mu) - \sum_{\alpha} (X_\alpha - \mu)^2\}$$

$= -(1/N(N-1))\, N\sigma^2 = -\sigma^2/(N-1)$

Thus, in this case we have E ($\bar{x}$) = $(1/n)\, n\mu = \mu$

and var ($\bar{x}$) = $(1/n^2) \times n\sigma^2 + (1/n^2) \times n(n-1) \times (-\sigma^2/(N-1)) = (\sigma^2/N)\{1 - (n-1)/(N-1)\}$

$= (\sigma^2/N)\{(N-n)/(N-1)\}$

Hence the standard error of $\bar{x}$ is $\sigma_x = (\sigma/\sqrt{n}) \sqrt{1 - \frac{N-n}{N-1}}$

In both the cases, the standard error decreases with increasing n. The standard error of the mean in sampling without replacements is, however, smaller than that in sampling with replacements. But the difference become negligible if N is very large compared to n. Also, in sampling without replacements, the standard error of the sample mean vanishes if n = N, which is to be expected because the sample mean now becomes a constant, i.e., the same as the population mean. However, this is not the case with sampling with replacements.

18.10 EXPECTATION AND STANDARD ERROR OF SAMPLE PROPORTION

Suppose in population of N. there are Np members with a particular character A and Nq members with the character not-A. Then p is the proportion of members in the population having the character A. Let a sample of size n be drawn from the population, and let f be the number of members in the sample

having character A. To find the expectation and standard error of the sample proportion f/n, we adopt the following procedure.

We assign to the α^{th} member of the population the value X_α, which is equal to 1 if, this member possesses the character A and equal to 0 otherwise. Similarly, to the i^{th} member of the sample we assign the value x_i, which is equal to 1 if this member possesses A and is equal to 0 otherwise.

In this way, we get a variable x, which has population mean (1/N) $\sum_\alpha X_\alpha = p$

and the population variance (1/N) $\sum_\alpha X_\alpha^2 - p^2 = p - p^2 = pq$

The sample mean of the variable x, on the other hand, is $\frac{1}{n}\sum_{i=1}^{n} x_i = f/n$

Hence we find, on replacing $\bar{x}$ by f/n, μ by p and σ^2 by pq in the expressions $E(\bar{x})$ and σ_x given in preceding sections,

E (f/n) = p [in case of random sampling with replacement]

= p [in case of random sampling without replacement]

$\sigma_{f/n} = pq/\sqrt{n}$ [in case of random sampling with replacement]

$= \sqrt{\frac{pq}{n}\left(1 - \frac{N-n}{N-1}\right)}$ [in case of random sampling without replacement]

The comments made in connection with the standard error of the mean apply here also.

Check Your Progress 4

1) Discuss the meaning of random sampling with replacement and without replacement.

.....

.....

.....

.....

2) Write the standard error of sample proportion.

.....

.....

.....

.....

18.11 LET US SUM UP

Throughout this unit we have learnt the basic concepts of sampling theory. It tells about different types of sampling along with the method of drawing

sample. Moreover, one can also be able to understand the concept of standard error and mean & standard deviation of sample mean and sample proportion.

18.12 KEY WORDS

Population: In statistics, population is an aggregate of objects, animate or inanimate, under study. The population may be finite or infinite.

Purposive Sampling: Purposive sampling is one in which the sampling units are selected with definite purpose in view.

Random Sampling: A random sampling is one in which each unit of population has an equal chance of being included in it.

Sample: A finite subset of statistical individuals in a population is called a sample and the number of individual in a sample is called the sample size.

Standard Error: The standard deviation of the sampling distribution of a statistic is known as its 'standard error', abbreviated as S.E.

Stratified Sampling: Here the entire heterogeneous population is divided into a number of homogeneous groups, usually termed as *strata*', which differ from one another but each of these groups is homogeneous within itself.

18.13 SOME USEFUL BOOKS

Goon A.M, Gupta M.K & Dasgupta B. (1971), *Fundamental of Statistics*, Volume I, The World Press Pvt. Ltd., Calcutta.

Freund, John E. (2001), *Mathematical Statistics*, Fifth Edition, Prentice-Hall of India Pvt. Ltd., New Delhi.

Das, N.G. (1996), *Statistical Methods*, M.Das & Co. (Calcutta).

18.14 ANSWER OR HINTS TO CHECK YOUR PROGRESS

Check Your Progress 1

1) See Section 18.2

2) See Section 18.3

3) See Section 18.4

Check Your Progress 2

1) See Section 18.5

2) See Section 18.5

3) See Section 18.6

Check Your Progress 3

1) See Section 18.7

2) See Section 18.8

having character A. To find the expectation and standard error of the sample

3) [Hint: n_1 =400 and n_2 =500, p_1 =300/400 =0.75, p_2=300/500=0.6

$p = (n_1p_1+n_2p_2)/(n_1+n_2) = 0.67$, $q = 1-p$; S. E $(p-p_1) = \sqrt{[((pq)/(n_1+n_2))(n_2/n_1)]} = 0.018$]

Check Your Progress 4

1) See Section 18.9

2) See Section 18.10

18.15 EXERCISES

1) A random sample of 500 pineapples was taken from a large consignment and 65 were found to be bad. Show that S.E. of the proportion of bad ones in a sample of this size is 0.015 and deduce that the percentage of bad pineapples in the consignment almost certainly lies between 8.5 and 18.5.

2) How does one get from a sample statistic to an estimate of the population parameter?

3) What is sampling error?

4) What is random sampling error?

5) What is a systematic error?

6) How is the sampling error or standard error determined?

7) What is sample size important?

8) What is sample size?

9) What is bias?

UNIT 19 SAMPLING DISTRIBUTIONS

Structure

19.0 OBJECTIVES

After going through this unit, you will be able to understand:

- the concept of sampling distribution of a statistic;
- various forms of sampling distribution, both discrete (e.g.: Binomial, Possion) and continuous (normal chi-square t and F) various properties of each type of sampling distribution;
- the use of probability density function and also Jacobean transformation in deriving various results of different sampling distribution;
- how to measure the goodness of fit of a test; and
- what should be the way of analysing any sample when it is not randomly distributed?

3) [Hint: n_1 =400 and n_2 =500, p_1 =300/400 =0.75, p_2=300/500=0.6

$p = (n_1p_1+n_2p_2)/(n_1+n_2) = 0.67$, $q = 1-p$; S. E $(p-p_1) = \sqrt{[((pq)/(n_1+n_2))(n_2/n_1)]} = 0.018$]

Check Your Progress 4

1) See Section 18.9

2) See Section 18.10

18.15 EXERCISES

19.1 INTRODUCTION

For a finite sample, it is not a big problem of assigning probabilities to the samples selected from the given population. However, in reality, where the sample size as well as the population is quite large, the number of all possible samples is also large. It becomes difficult to assign probabilities of a specified set of samples. Therefore, we have to think of all possible ways of selecting the samples from the entire population.

19.2 CONCEPT OF SAMPLING DISTRIBUTION OF A STATISTIC

Sampling distribution of a statistic may be defined as the probability law, which the statistic follows, if repeated random samples of a fixed size are drawn from a specified population.

Let us consider a random sample $x_1, x_2,\ldots, x_n$ of size n drawn from a population containing N units. Let us further suppose that we are interested in the sampling distribution of the statistic $\bar{x}$ (i.e., sample mean), where

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots\ldots + x_n)$$

If the population size N is finite, there is a finite number (say k) of possible ways of drawing n units in the sample out of a total of N units in the population. Although the k samples are distinct, the sample means may not be all different, but each of these will occur with equal probability. Thus, we can construct a table showing the set of possible values of the statistic $\bar{x}$ and also the probability that $\bar{x}$ will take each of these values. This probability distribution of the statistic $\bar{x}$ is called 'sampling distribution' of sample mean. The above method is quite general, and the sampling distribution of any other statistic, say, median or standard deviation of the sample, may be obtained.

If, however, the number (N) of units in the population is large, the number (k) of possible distinct samples being even larger, the above method of finding the sampling distribution cannot be applied. In this case, the values of $\bar{x}$ obtained from a large number of samples may be arranged in the form of relative frequency distribution. The limiting form of this relative frequency distribution, when the number of samples considered becomes infinitely large, is called 'sampling distribution of the statistic'. When the population is specified by a theoretical distribution (e.g., binomial or normal), the sampling distribution can be theoretically obtained. The knowledge of sampling distribution is necessary in finding 'confidence limits' for parameters and in 'testing statistical hypothesis'.

In this unit, we will highlight on various properties of different sampling distributions. So we will mainly concentrate on how different sampling distributions work and in doing so we use several statistical formulae. Since our intention is to represent theoretical overview of the topic, number of numerical example is too less than other units. Here, what is important is to have the idea of topic theoretically and the numerical part will be covered subsequently.

19.3 SAMPLING DISTRIBUTION WITH DISCRETE POPULATION DISTRIBUTIONS

We derive some common sampling distributions that arise from an infinite population.

19.3.1 Sampling Distribution of Sample Total: Binomial Parent

Suppose x_1 and x_2 are distributed independently in the binomial form with parameters m_1, P and m_2, P respectively. Consider then the distribution sum can take are 0, 1, 2,...., $m_1 + m_2$.

Also, $$P[x_1 + x_2 = k] = \sum_{k_1=0}^{k} P[x_1 = k_1]P[x_2 = k - k_1]$$

$$= \sum_{k_1=0}^{k} \binom{m_1}{k_1}\binom{m_2}{k-k_1} p^k (1-p)^{m_1+m_2-k}$$

$$= p^k (1-p)^{m_1+m_2-k} \sum_{k_1=0}^{k} \binom{m_1}{k_1}\binom{m_2}{k-k_1}$$

Now, this sum is nothing but the sum of products of the coefficients of t^{k_1} in $(1+t)^{m_1}$ and t^{k-k_1} in $(1+t)^{m_2}$, for varying k_1, and hence equals the coefficient of t^k in $(1+t)^{m_1+m_2}$, which is $\binom{m_1+m_2}{k}$.

Thus, $$P[x_1 + x_2 = k] = \binom{m_1+m_2}{k} p^k (1-p)^{m_1+m_2-k}$$

This shows that x_1+x_2 is itself binomially distributed with parameters $m_1 + m_2$ and p. We also get from the general result that if $x_1, x_2, \ldots, x_n$ are independently distributed binomial variables with parameters m_1, p; m_2, p;....; m_n, p; then the sum x_1+x_2+.....is also a binomial variable with parameters $m_1 + m_2 + ... + m_n$ and p.

This implies that if x_1, x_2, ..., x_n are a random sample from a binomial distribution with parameters of the statistic, $x_1 + x_2 + ... + x_n$ is also binomial with parameters nm and p.

19.3.2 Sampling Distribution of Sample Total: Poisson Parent

Suppose x_1 and x_2 are distributed independently in the Poisson form with parameter x_1 and x_2, respectively. The sum x_1+x_2 can then take the values 0, 1, 2,.....

Also, $$P[x_1 + x_2 = k] = \sum_{k_1=0}^{k} P[x_1 = k_1]P[x_2 = k - k_1]$$

$$= \sum_{k_1=0}^{k} \frac{\exp[-\lambda_1]\lambda_1^{k_1}}{k_1!} \times \frac{\exp[-\lambda_2]\lambda_2^{k-k_1}}{(k-k_1)!}$$

$$= \frac{\exp[-\lambda_1-\lambda_2]}{k!} \sum_{k_1=0}^{k} \binom{k}{k_1} \lambda_1^{k_1} \lambda_2^{k-k_1}$$

$$= \exp[-(\lambda_1+\lambda_2)] \frac{(\lambda_1+\lambda_2)^k}{k!},$$

which shows that x_1+x_2 is itself a Poisson variable with parameter $\lambda_1 + \lambda_2$. It immediately follows that if x_1, x_2, ..., x_n are independently distributed Poisson variables with λ_1, λ_2 ,..., λ_n, then the sum $x_1+x_2+...+x_n$ is also a Poisson variable with parameter $\lambda_1+\lambda_2+ ..... +\lambda_n$.

The above results give, in particular, the sampling distribution of the statistic $x_1+x_2+ .. + x_n$ when x_1, x_2, .., x_n are a random sample from a Poisson distribution with parameter λ. This sampling distribution is also of the Poisson form with parameter $n\lambda$.

Check Your Progress 1

1) The normal distribution is defined by two parameters. What are they?

.....

.....

.....

.....

2) List the discrete and continuous sampling distributions.

.....

.....

.....

.....

3) If the scores are normally distributed with a mean 30 and a standard deviation of 5, what percentage of the scores is

   a) Greater than 30?

   b) Greater than 37?

   c) Between 28 and 34?

.....

.....

.....

.....

4) What is a poisson distribution? Discuss the mean and variance of such a distribution.

.....

.....

.....

.....

19.4 FOUR FUNDAMENTAL DISTRIBUTIONS DERIVED FROM THE NORMAL DISTRIBUTION

19.4.1 Distribution of Standard Normal Variable

A (continuous) random variable, which is normally distributed with mean 0 and variance 1, is called a 'standard normal variable' of a normal deviate. It is generally denoted by τ (or z), so that the p.d.f. of standard normal distribution by be written as

$$f(\tau)=\frac{1}{\sqrt{2\pi}}\exp\left(-\tau^2/2\right),-\alpha<\tau<\alpha$$

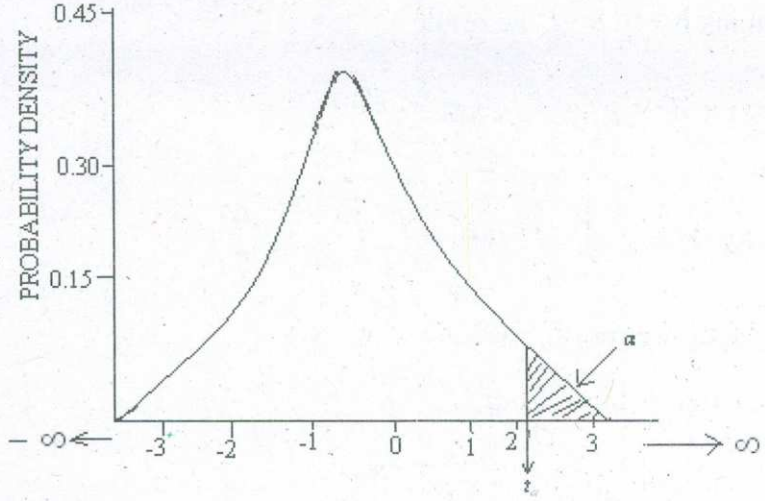


Fig. 19.1: Distribution of a Standard Normal Variable

The distribution is, of course, symmetrical about 0. The probability-density function of standard normal distribution is

$$f(\tau)=\frac{1}{\sqrt{2\pi}}\exp\left(-\tau^2/2\right),-\alpha<\tau<\alpha$$

The properties of this distribution may be deduced from those of a general normal distribution.

We shall denote by τ_α the value of τ such that

$$p[\tau>\tau_a]=\alpha$$

4) The following figures show the distribution of digits in numbers chosen at random from a telephone directory.

Digits:	0	1	2	3	4	5	6	7	8	9	Total
Frequency:	1026	1107	997	966	1075	933	1107	972	964	853	10,000

Test whether the digits may be taken to occur equally frequently in the directory.

.....

.....

.....

.....

19.4.3 't' Distribution

19.4.3.1 Student's 't'

Let x_i, (i = 1, 2, ..., n) be a random sample of size n from a normal population with mean μ and variance σ^2. Then student's 't' is defined by the statistic

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

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is the sample mean and $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is an unbiased estimate of the population variance σ^2,

and it follows t distribution with v = (n -1) df, with probability density function

$$f(t) = \frac{1}{\sqrt{v}\,B\left(\frac{1}{2}, \frac{v}{2}\right)} \cdot \frac{1}{\left[1 + \frac{t^2}{v}\right]^{(v+1)/2}}; -\infty < t < \infty$$

A statistic t following Student's t - distribution with n df will be abbreviated as $t \sim t_n$.

If we take v=1, then $f(t) = \frac{1}{B\left(\frac{1}{2}, \frac{1}{2}\right)} \cdot \frac{1}{(1+t^2)}$

$$= \frac{1}{\pi} \cdot \frac{1}{(1+t^2)}; -\infty < t < \infty$$

$$\left[\because \Gamma\left(\frac{1}{2}\right) \sqrt{\pi}\right]$$

which is the p.d.f. of standard Cauchy distribution.

Hence, when v=1, Student's t distribution reduces to Cauchy distribution.

19.4.3.2 Fisher's 't'

It is the ratio of a standard normal variate to the square roof of an independent chi-square variate divided by its degree of freedom. If τ is a N(0, 1) and χ^2 is an independent chi-square variate with n degrees of freedom, then Fisher's t is given by

$$t = \tau / \sqrt{x^2/n}.$$

and it follows Student's t distribution with n degrees of freedom.

Since τ and χ^2 are independent, their joint density function is given by

$$\frac{1}{\sqrt{2\pi}}.\exp\left(-\tau^2/2\right).\frac{\exp\left(-\chi^2/2\right)\left(\chi^2\right)^{\frac{n}{2}-1}}{2^{n/2}.\Gamma\left(n/2\right)},$$

$$-\infty<\tau<\infty, 0<\chi^2<\infty.$$

Marketing the one-to-one transformation.

$$\left.\begin{array}{l} t = \tau / \sqrt{\chi^2/n} \\ u = x^r \end{array}\right\} (-\infty < t < \infty, 0 < u < \infty),$$

so that $\tau = t\sqrt{u/n}, \chi^2 = u,$

and noting that the Jacobean of the transformation is

$$J = \begin{vmatrix} \frac{\partial\tau}{\partial t} & \frac{\partial\tau}{\partial u} \\ \frac{\partial\chi^2}{\partial t} & \frac{\partial\chi^2}{\partial u} \end{vmatrix} = \begin{vmatrix} \sqrt{u/n} & \frac{t}{2\sqrt{un}} \\ 0 & 1 \end{vmatrix} = \sqrt{u/n}$$

The joint p.d.f. of t and u becomes

$$\frac{1}{\sqrt{2\pi}.2^{\frac{n}{2}}.\Gamma\left(n/2\right).\sqrt{n}} \exp\left\{-\frac{u}{2}\left(1+\frac{t^2}{n}\right)\right\} u^{\frac{n}{2}-\frac{1}{2}}$$

$$-\infty<t<\infty, 0<u<\infty$$

The p.d.f. of t is, therefore,

$$\frac{1}{\sqrt{2\pi}.2^{n/2}.\Gamma\left(n/2\right).\sqrt{n}}\left[\int_0^\infty \exp\left\{-\frac{u}{2}\left(1+\frac{t^2}{n}\right)\right\} u^{(n-1)/2}\,du\right]dt$$

$$= \frac{1}{\sqrt{2\pi}.2^{n/2}.\Gamma(n/2)} \cdot \frac{\Gamma[(n+1)/2]}{\left[\frac{1}{2}\left(1+\frac{t^2}{n}\right)\right]^{(n+1)/2}}$$

$$= \frac{\Gamma[(n+1)/2]}{\sqrt{n}\Gamma(n/2)\Gamma(\frac{1}{2})} \cdot \frac{1}{\left[1+\frac{t^2}{n}\right]^{(n+1)/2}}$$

$-\alpha < t < \alpha$

$$= \frac{1}{\sqrt{n}B\left(\frac{1}{2},\frac{n}{2}\right)\left[1+\frac{t^2}{n}\right]^{(n+1)/2}}, -\alpha < t < \alpha$$

which is same as the probability function of Student's t - distribution with n df.

It should be noted here that, Student's t is a particular case of Fisher's 't'.

Like the standard normal distribution, the t distribution is symmetrical about t = 0. But unlike the normal distribution, it is more peaked than a normal distribution with the same standard deviation.

The symbol $t_{\alpha,n}$ will be used to denote the value of t (with df = n) such that

$$p\left[t > t_{\alpha,n}\right] = \infty$$

$t_{1-\alpha,n} = -t_{\alpha,n}$.

For small n, the t distribution differs considerably from the standard normal distribution, $t_{\alpha,n}$ being always greater than τ_α if $0 < \alpha < \frac{1}{2}$. For large value of n, however, the t distribution tends to the standard normal form and $t_{\alpha,n}$ may then be well approximated by τ_α.

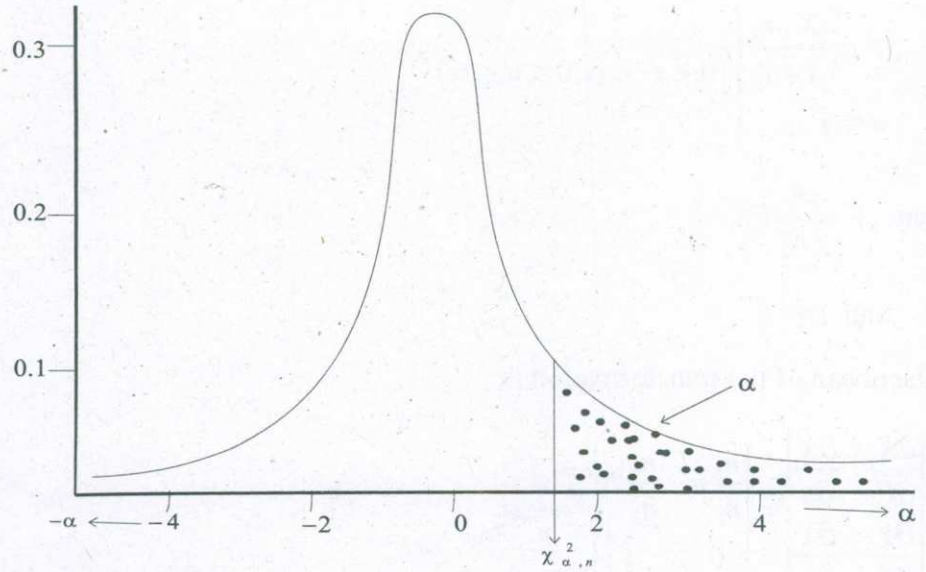


Fig. 19.4: t distribution with 5 Degrees of Freedom (n=5)

19.4.4 'F' Distribution

If X and Y are two independent chi-square variates with n_1 and n_2 df respectively, then F statistic is defined by

$$F = \frac{X/n_1}{Y/n_2}$$

In other words, F is defined as the ratio of two independent chi-square variates divided by their respective degrees of freedom and it follows Suedecor's F distribution with (n_1, n_2) df with probability function given by

$$\frac{\left(n_1/n_2\right)^{n_1/2}}{B\left(n_1/2, n_2/2\right)} \frac{F^{\frac{n_1}{2}-1}}{\left[1+\frac{n_1}{n_2}F\right]^{\left(\frac{n_1+n_2}{2}\right)}}, 0 \le F < \infty$$

Note: i) The sampling distribution of F-statistic does not involve any population parameters and depends only on the degrees of freedom n_1 and n_2.

ii) A statistic F following Suedecor's F distribution with (n_1, n_2) df will be denoted by $F \sim F(n_1, n_2)$.

To derive the above result, note that the joint p.d.f. of X and Y, from $f(\chi^2)$ in χ^2 distribution,

$$g(x,y) = \frac{1}{2^{(n_1+n_2)/2} \cdot \Gamma\left(\frac{n_1}{2}\right)\left(\Gamma\frac{n_2}{2}\right)} \cdot X^{(n_1/2)^{-1}} . Y^{\left(\frac{n_2}{2}\right)-1} . \exp\left[-(X+Y)/2\right]$$

$0 < X < \alpha, 0 < Y, < \alpha$.

Let us now make the one-to-one transformation

$$\left.\begin{aligned} F &= \frac{X/n_1}{Y/n_2} \\ u &= Y \end{aligned}\right\} (0 < F < \alpha, 0 < u < \alpha)$$

So that, $X = \frac{n_1}{n_2} Fu.$

and $Y = u.$

The Jacobean of the transformation is

$$J = \begin{vmatrix} \frac{\partial X}{\partial F} & \frac{\partial X}{\partial u} \\ \frac{\partial Y}{\partial F} & \frac{\partial Y}{\partial u} \end{vmatrix} = \begin{vmatrix} \frac{n_1}{n_2} u & \frac{n_1}{n_2} F \\ 0 & 1 \end{vmatrix} = \frac{n_1}{n_2} u.$$

Hence, the joint p.d.f. of F and u is

$$h(F,u) = \frac{\left(\frac{n_1}{n_2}\right)^{\left(\frac{n_1}{2}\right)}}{2^{(n_1+n_2)/2}\Gamma(n_1/2)\Gamma(n_2/2)} \cdot u^{[(n_1+n_2)/2]-1} . F^{(n_1/2)-1}$$

$$\times \exp\left[\frac{-u}{2}\left(1+\frac{n_1}{n_2}F\right)\right]$$

$0 < F < \alpha,\ 0 < u < \alpha.$

The p.d.f. of F is therefore,

$$f(F) = \int_0^d h(F,u)\,du$$

$$= \frac{\left(n_1/n_2\right)^{(n_1/2)}}{\Gamma\left(n_1/n_2\right)\Gamma\left(n_2/2\right)} \times F^{\left(n_1/2\right)-1} \times \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\left(1+\frac{n_1}{n_2}F\right)^{\frac{n_1+n_2}{2}}}$$

$$= \frac{\left(n_1/n_2\right)^{(n_1/2)}}{B\left(\frac{n_1}{2},\frac{n_2}{2}\right)} \times \frac{F^{\left(\frac{n_1}{2}\right)-1}}{\left(1+\frac{n_1}{n_2}F\right)^{(n_1+n_2)/2}}, 0 < F < \alpha$$

The distribution is highly positively skewed. It is easily seen from the definition of t and F that an F with $n_1 = 1$ is a t^2, t having df = n_2.

As in the previous cases, we shall denote by $F_{\alpha;n_1,n_2}$ the upper α point of the F distribution with df = (n_1, n_2); i.e.,

$$p\left[F > F_{\alpha;n_1 n_2}\right] = \infty$$

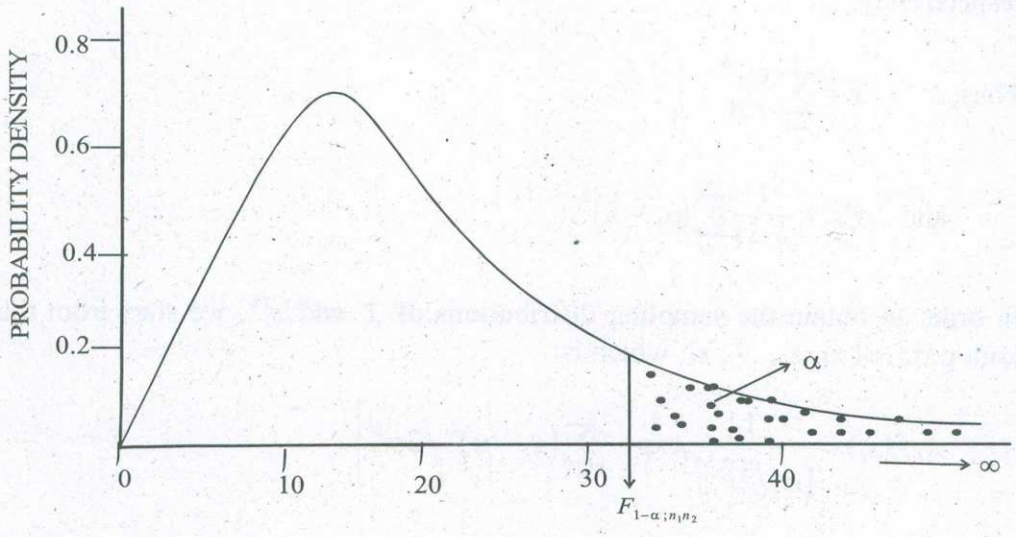


Fig. 19.5: F distribution with (10, 4) Degrees of Freedom (n_1=10, n_2=4)

As regards the lower α point $F_{1-\alpha;n_1 n_2}$, we see that

$$p\left[F < F_{1-\alpha;n_1,n_2}\right] = \alpha$$

or, $$p\left[\frac{1}{F} > \frac{1}{F_{1-\alpha;n_1,n_2}}\right] = \alpha$$

Now, $\frac{1}{F}$, which is of the form $\frac{Y/n_2}{X/n_1}$, is itself distributed as an F with df = (n_2, n_1). It follows that

$$\frac{1}{F_{1-\alpha;n_1,n_2}} = F_{1-\alpha;n_2,n_1}$$

$$F_{1-\alpha;n_1,n_2} = \frac{1}{F_{-\alpha;n_2,n_1}}$$

It is, therefore, unnecessary to tabulate the lower α points of F distributions with various degrees of freedom, once the upper α points are tabulated.

19.5 SAMPLING DISTRIBUTIONS OF MEAN AND VARIANCE IN RANDOM SAMPLING FROM A NORMAL DISTRIBUTION

Let $x_1, x_2, \ldots, x_n$ be a random sample from a normal distribution whose p.d.f. is

$$f(x)\frac{1}{\sigma.\sqrt{2\pi}}\exp.\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], -\infty < x < \infty$$

$$\left.\begin{aligned} F &= \frac{X/n_1}{Y/n_2} \\ u &= Y \end{aligned}\right\} (0 < F < \alpha, 0 < u < \alpha)$$

So that, $X = \frac{n_1}{n_2} Fu.$

and $Y = u.$

We shall denote the sample mean and the sample variance of x by $\bar{x}$ and s'^2, respectively.

Thus, $\bar{x} = \sum_{i=1}^{n} x_i / n$

and $s'^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_1 - \bar{x})^2$

In order to obtain the sampling distributions of $\bar{x}$ and s'^2, we start from the joint p.d.f. of $x_1, x_2, \ldots, x_n$, which is

$$\pi_i f(x_i) = \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n} \exp\left[-\sum_i (x_i - \mu)^2 / 2\sigma^2 \right].$$

We make the following one-to-one transformation from x_i (i=1, 2, …, n) to y_i(i = 1, 2, …., n):

$$y_1 = \frac{(x_1 - \mu)/\sigma + (x_2 - \mu)/\sigma + \ldots + (x_n - \mu)/\sigma}{\sqrt{n}},$$

$$y_i = a_{i1}\left(\frac{(x_1 - \mu)}{\sigma}\right) + a_{i2}\left(\frac{(x_2 - \mu)}{\sigma}\right) + \ldots + a_{in}\left(\frac{(x_n - \mu)}{\sigma}\right), \text{ for i} = 2, 3, \ldots, n.$$

where the (n – 1) vectors $(a_{i1}, a_{i2}, \ldots, a_{in})$ are of unit length, mutually orthogonal and each orthogonal to the vector $\left(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\right)$.

One such set of vectors is:

$\frac{1}{\sqrt{2}}$ (1, -1, 0, 0, …, 0, 0),

$\frac{1}{\sqrt{6}}$ (1, 1, -2, 0, …, 0, 0),

$\frac{1}{\sqrt{n(n-1)}}$ (1, 1, 1, …, 1, -(n-1)).

The Jacobean of the transformation is then J, such that

$$\frac{1}{J} = \begin{vmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & & & \vdots \\ \frac{\partial y_n}{\partial x_1} & \frac{\partial y_n}{\partial x_2} & \cdots & \frac{\partial y_n}{\partial x_n} \end{vmatrix} = \mp \frac{1}{\sigma^n},$$

implying that $J = \mp\sigma^n$ and $|J| = \sigma^n$.

Further, $\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} (x_i - \mu)^2 / \sigma^2$

Hence, the joint p.d.f. of $y_1, y_2, \ldots, y_n$ is

$$\frac{1}{\left(\sqrt{2\pi}\right)^n} \exp\left[-\sum_{i=1}^{n} y_i^2/2\right].$$

This shows that $y_1, y_2, \ldots, y_n$ are independently and identically distributed, each being a standard normal variable.

But $y_i = \frac{1}{\sigma\sqrt{n}} \sum_i (x_i - \mu) = \frac{\sqrt{n}(\bar{x} - \mu)}{\sigma}$,

is linear function of $\bar{x}$. Since y_1 is a standard normal variable, $\bar{x}$ must be a normal variable with mean μ and variance $\frac{\sigma^2}{n}$ (follows from the theorem given in the section 'distribution of standard normal variable').

Thus, the p.d.f. of $\bar{x}$ is

$$g(\bar{x}) = \frac{\sqrt{n}}{\sigma\sqrt{2\pi}} \exp\left[-n(\bar{x} - \mu)^2 / 2\sigma^2\right], -\infty < \bar{x} < \infty$$

Again, $\sum_{i=2}^{n} y_i^2 = \sum_{i=1}^{n} y_i^2 - \bar{y}_i^2$

$$= \frac{1}{\sigma^2}\left[\sum_{i=1}^{n} (x_i - \mu)^2 - n(\bar{x} - \mu)^2\right]$$

$$= \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$= \frac{(n-1)s'^2}{\sigma^2}$$

Now, $\sum_{i=1}^{n} y_i^2$, being the sum of squares of (n-1) independent standard normal variables, is a χ^2 with df =(n-1), and this is distributed independently of y_1. It follows that $(n-1)s'^2/\sigma^2$ is distributed as a χ^2 with df = (n-1) and is independent of $\bar{x}$.

And the p.d.f. of s'^2 is,

$$u(s'^2) = \frac{(n-1)^{(n-1)/2}}{\left(2\sigma^2\right)^{(n-1)/2} \Gamma\left(\frac{n-1}{2}\right)} . \exp\left[-(n-1)s'^2/2\sigma^2\right] \left(s'^2\right)^{n-3/2}, 0 < s'^2 < \infty.$$

Check Your Progress 3

1) Given a test that is normally distributed with mean=30 and a standard deviation=6, what is the probability that a single score drawn at random will be greater than 34.

.....

.....

.....

.....

2) Assume a normal distribution with a mean of 90 and a standard deviation of 7. What limits would include the middle 65% of the cases.

.....

.....

.....

.....

19.6 CENTRAL LIMIT THEOREM

The central limit theorem in the mathematical theory of probability may be expressed as follows:

If x_i (i = 1, 2, ..., n) be independently distributed random variables such that $E(x_i) = \mu_i$ and $V(x_i) = \sigma_i^2$, then it can be proved that under certain very general conditions, the random variables $s_n = x_1 + x_2 + ... + x_n$, is asymptotically normal with mean μ and standard deviation σ where

$$\mu = \sum_{i=1}^{n} \mu_i \text{ and } \sigma^2 = \sum_{i=1}^{n} \mu_i^2$$

This theory helps us in dealing the observations, which are not randomly distributed.

19.7 LET US SUM UP

This unit gives us the idea about different types of sampling distributions concepts and properties of discrete as well as continuous distribution have been given in the unit. Various properties, associated theorem and results are being made clear through the discussion given. The concept of critical value of or significant value of different distributions is given here and the way of analysing probability density function and Jacobean transformation are also made clear. Finally, we have the intuitive idea about central limit theorem, which is useful for the non-randomly distributed sample.

19.8 KEY WORDS

Central Limit Theorem: If xi (i = 1,2,.....,n) be independently distributed random variables such that E(xi)= μ_i and V(xi)=σ_i^2 then, under certain assumption, $s_n\left(=\sum_{i=1}^{n} x\right)$ is asymptotically normal with mean μ and variance σ^2 where

$$\mu=\sum_{i=1}^{n}\mu_i, \sigma^2=\sum_{i=1}^{n}\sigma_i^2 .$$

Chi-square Distribution: The square of a standard normal variate is known as chi-square variate with 1 df.

F Distribution: If $X \sim x^2_{n_1}$ and $Y \sim x^2_{n_2}$ and if X and Y are independent to each other then F statistic is defined by

$$F = \frac{X/n_1}{Y/n_2} \sim F_{(n_1,n_2)}$$

Fisher's 't' Distribution: If t ~ N(0,1) and $x^2 \sim x^2_n$, and if both are independent, then Fisher's 't' is given by $t = t/\sqrt{x^2/n} \sim t_n$

Standard Normal Distribution: A continuous random variable, which is normally distributed with mean zero, variance 1 is called standard normal distribution.

Students's 't' Distribution: If x_i, (i=1,2,…,n) be a random sample of size n from a normal population with mean μ and variance σ^2, then student's 't' is defines by the statistic $t = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim t_{(n-1)}$.

19.9 SOME USEFUL BOOKS

Goon A.M, Gupta M.K & Dasgupta B. (1971), *Fundamental of Statistics*, Volume I, The World Press Pvt. Ltd., Calcutta.

Freund, John E. (2001), *Mathematical Statistics*, Fifth Edition, Prentice-Hall of India Pvt. Ltd., New Delhi.

Das, N.G. (1996), *Statistical Methods*, M.Das & Co. (Calcutta).

19.10 ANSWER OR HINTS TO CHECK YOUR PROGRESS

Check Your Progress 1

1) See Section 19.3

2) See Section 19.3 and 19.4

3) a) 50%

b) 8.08%

c) 44.35

4) See Sub-section 19.3.2

Check Your Progress 2

1) See Section 19.4

2) χ^2=14.5 df=5, p=.017

3) See Section 19.4

4) Hint: Here we set up the null hypothesis that the digit occurs equally frequently in the directory. Under the null hypothesis, the expected frequency for each of the digits 0,1,2,....,9 is 10000/10=1000.

The value of $\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = 58.542$

Here E_i=1000 $\forall$ i = 0,1,.....,9

Oi= observed frequency given in the table.

Degree of freedom = 10-1=9 (since we are give to frequencies subjected to only one linear constraint $\sum_i O_i = \sum_i E_i = 10000$.

The tabulated $\chi^2_{0.05,9}$=16.919 < 58.542

Thus, we conclude that digits are not uniformly distributed.

Check Your Progress 3

1) 0.2524

2) 83.46 and 96.54

19.11 EXERCISES

1) The following table gives the number of aircraft accidents that occur during the various days of the week. Find whether the accidents are uniformly distributed over the week.

Days:	Sun	Mon	Tue	Wed	Thurs	Fri	Sat
No. of Accidents	14	16	8	12	11	9	14

(Given: The value of χ^2 significant at 5,6,7 df are respectively 11.07, 12.59, 14.07 at the 5% level of significance)

2) The theory predicts the proportion of beans in the four groups A, B, C and D should be 9:3:3:1. In an experiment among 1600 beans, the

numbers in the four groups were 882,313,287 and 118. Does the experimental result support the theory?

[Hint: Total no. of bins = 1600

$$E(882) = \frac{9}{16} \times 1600 = 900$$

$$E(313) = \frac{3}{16} \times 1600 = 300 \text{ and so on}$$

$$\chi^2 = \sum \left[\frac{(O-E)^2}{E} \right] = 4.7266 < \chi^2_{0.05,3} (= 7.815)$$

$\therefore$ The null hypothesis is accepted.]

3) What is a normal distribution?

UNIT 20 STATISTICAL INFERENCES

Structure

20.0 OBJECTIVES

After going through this unit, which explains the concepts of estimation theory and hypothesis testing, you will be able to answer the following:

- How characteristics of any population can be inferred on the basis of analysing the sample drawn from that population;
- The likeliness of any characteristic of a population on the basis of analysing the sample drawn from that population;
- What should be the characteristics of an estimator? and
- What are the test criteria under different situations?

20.1 INTRODUCTION

The object of sampling is to study the features of the population on the basis of sample observations. A carefully selected sample is expected to reveal these features, and hence we shall infer about the population from a statistical analysis of the sample. The process is known as 'statistical inference'.

There are two types of problems. First, we may have no information at all about some characteristics of the population, especially the values of the parameters involved in the distribution, and it is required to obtain estimates of these parameters. This is the problem of 'estimation'. Secondly, some information or hypothetical values of the parameters may be available, and it is required test how far the hypothesis is tenable in the light of the information provided by the sample. This is the problem of 'hypothesis testing' or 'test of significance'.

20.2 THEORY OF ESTIMATION

Suppose, we have a random sample $x_1, x_2, \ldots, x_n$ on a variable x, whose distribution in the population involves an unknown parameter θ. It is required to find an estimate of θ on the basis of sample values. The estimation is done in two different ways: (i) Point estimation, and (ii) Interval estimation.

In 'point estimation', the estimated value is given by a single quantity, which is a function of sample observations (i.e., statistic). This function is called 'estimator'.

In 'interval estimation', an interval within which a parameter is expected to lie is given by using two quantities based on sample values. This is known as 'confidence interval', and the quantities, which are used to specify the interval, are known as 'confidence limits'. Since our basic objective is to estimate the parameter associated with the sample conservations, so before going into further details, let us discuss elaborately about the notion of parameter space.

20.2.1 Parameter Space

Let us consider a random variable (r. v) x with p. d. f. $f(x, \theta)$. In most common application, though not always, the functional form of population distribution is

assumed to be known except for the value of some unknown parameter(s) θ, which may take any value on a set Θ. This is expressed by writing the p. d. f. in the form $f(x, \theta)$, $\theta \in \Theta$. The set Θ, which is the set of all possible values of θ, is called the 'parameter space'. Such a situation gives rise not to one probability distributions but a family of probability distribution which we write as $[f(x, \theta), \theta \in \Theta]$. For example, if $X \sim N(\mu, \sigma^2)$, then the parameter space $\Theta = \{(\mu, \sigma^2): -\infty < \mu < \infty; 0 < \sigma < \infty\}$.

In particular, for $\sigma^2 = 1$, the family of probability distribution is given by $\{N(\mu, 1); \mu \in \Theta\}$, where $\Theta = \{\mu: -\infty < \mu < \infty\}$. In the following discussion we shall consider a general family of distributions $\{f(x; \theta_1, \theta_2, \ldots, \theta_k): \theta_i \in \Theta, i = 1, 2, \ldots, k.\}$.

Let us consider a random sample $x_1, x_2, \ldots, x_n$, of size 'n' from a population, with probability function $f(x; \theta_1, \theta_2, \ldots, \theta_k)$, where $\theta 1, \theta_2, \ldots, \theta_k$ are the unknown population parameters. There will then always be an infinite number of functions of sample values, called statistic, which may be proposed as estimates of one or more of the parameters.

Evidently, the best estimates would be one that falls nearest to the true value of the parameter to be estimated. In other words, the statistic whose distribution concentrates as closely as possible near the true value of the parameter may be regarded the best estimate. Hence, the basic problem of estimation in the above case can be formulated as follows:

We wish to determine the functions of the sample observations:

$T_1 = \widehat{\theta_1}(x_1, x_2, \ldots, x_n)$, $T_2 = \widehat{\theta_2}(x_1, x_2, \ldots, x_n)$, $\ldots$, $T_k = \widehat{\theta_k}(x_1, x_2, \ldots, x_n)$, such that their distribution is concentrated as closely as possible near the true value of the parameter. The estimating functions are then referred to as 'estimator'.

20.3 CHARACTERISTICS OF ESTIMATORS

The following are some of the criteria that should be satisfied by a good estimator: (i) Consistency, (ii) Unbiasedness, (iii) Efficiency, and (iv) Sufficiency. We shall now briefly explain these terms one by one.

20.3.1 Consistency

An estimator $T_n = T(x_1, x_2, \ldots, x_n)$, based on a random sample of size 'n' , is said to be consistent estimator of $\gamma(\theta)$, $\theta \in \Theta$, the parameter space, if T_n – converges to $\gamma(\theta)$ in probability i.e., if $T_n \xrightarrow{P} \gamma(\theta)$ as $n \to \infty$. In other words, T_n is a consistent estimator of $\gamma(\theta)$ if for every $\varepsilon > 0$, $\eta > 0$, there exists a positive integer $n \geq m(\varepsilon, \eta)$ such that

$P[|T_n - \gamma(\theta)| < \varepsilon] \to 1$ as $n \to \infty \Rightarrow P[|T_n - \gamma(\theta)| < \varepsilon] > 1 - \eta$; $\forall\, n \geq m$, where m is some very large value of n.

Note: If $X_1, X_2, \ldots, X_n$ is a random sample from a population with finite mean $E(X_i) = \mu < \infty$, then by Khinchine's weak law of large numbers (WLLN), we have $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} x_i \rightarrow E(X_i) = \mu$, as $n \rightarrow \infty$

Hence sample mean ($\bar{X}_n$) is always a consistence estimator of population mean.

20.3.2 Unbiasedness

Obviously, consistency is a property concerning the behaviour of an estimator for indefinitely large values of the sample size 'n', i.e., as $n \rightarrow \infty$. Nothing is regarded of its behavior for finite 'n'.

Moreover, if there exists a consistent estimator, say, T_n of $\gamma(\theta)$, then infinitely many such estimators can be constructed, e.g., $T_n' = (n-a)/(n-b)$, $Tn = [1-(a/n)]/ [1-(b/n)]$, where $T_n' \rightarrow T_n \rightarrow \gamma(\theta)$, as $n \rightarrow \infty$ and hence, for different values of a and b, T_n' is also consistent for $\gamma(\theta)$. Unbiasedness is a property associated with finite 'n'. A statistic $T_n = T(x_1, x_2, \ldots, x_n)$, is said to be an unbiased estimator of $\gamma(\theta)$ if $E(T_n) = \gamma(\theta)$, for all $\theta \in \Theta$

Note: If $E(T_n) > \gamma(\theta)$, T_n is said to be positively biased and if $E(T_n) < \Upsilon(\theta)$, it is said to be negatively biased, the amount of bias $b(\theta)$ being given by $b(\theta) = E(T_n) - \gamma(\theta)$, $\theta \in \Theta$

Sufficient Conditions for Consistency

Let $\{T_n\}$ be a sequence of estimators such that for all $\theta \in \Theta$

i) $E_\theta(T_n) \rightarrow \gamma(\theta)$ as $n \rightarrow \infty$

ii) $Var_\theta(T_n) \rightarrow 0$ as $n \rightarrow \infty$

Then T_n is a consistent estimator of $\gamma(\theta)$.

20.3.3 Efficiency

Even if we confine ourselves to unbiased estimates, there will, in general, more than one consistent estimator of a parameter. There may be found a large number of consistent estimators for μ. Indeed, T be consistent, so are, e.g., $T + a/\psi(n)$ and $T\{1 + a/\psi(n)\}$, where 'a' is any constant independent of 'n' and $\psi(n)$ is any increasing function of 'n'. To choose among these rival estimators, some additional criteria would be needed. Thus, we may consider, together with stochastic convergence, the rate of stochastic convergence, i.e., we may demand not only that T should converge stochastically to $\gamma(\theta)$ but also that it should do so sufficiently rapidly. We shall confine our attention to consistent estimators that are asymptotically normally distributed. In that case the rapidity of convergence will be indicated by the inverse of the variance of the asymptotic distribution. Denoting the asymptotic variance 'avar', we may say that T is the best estimator of $\gamma(\theta)$ if it is consistent and normally distributed and if avar (T) $\leq$ avar (T') whatever the other consistent and asymptotically normal estimator T' may be.

A consistent, asymptotically normal statistic T having this property is called 'efficient'.

20.3.3.1 Most Efficient Estimator

If in a case of consistent estimator for a parameter, there exists one whose sampling variance is less than that of any such estimator, it is called the most efficient estimator. Whenever such an estimator exists, it provides a criterion for measurement of efficiency of the other estimators.

Definition: If T_1 is the most efficient estimator with variance V_1 and T_2 is any other estimator with variance V_2, then the efficiency E of T_2 is defined as: $E = V_1 / V_2$. Obviously, E cannot exceed unity. If $T_1, T_2, \ldots, T_n$ are all estimators of $\gamma(\theta)$ and Var(T) is minimum, then the efficiency of E_i of T_i, $(i = 1, 2, \ldots, n)$ is defined as:

$E_i = \text{Var}(T)/ \text{Var}(T_i)$; obviously $E_i \leq 1$, $(i = 1, 2, \ldots, n)$.

20.3.3.2 Minimum Variance Unbiased Estimator (MVUE)

If a statistic $T = T(x_1, x_2, \ldots, x_n)$, based on sample of size 'n' is such that:

i) T is unbiased for $\gamma(\theta)$, for all $\theta \in \Theta$ and

ii) it has the smallest variance among the class of all unbiased estimators of $\gamma(\theta)$, then T is called the minimum variance unbiased estimator (MVUE) of $\gamma(\theta)$. More precisely, T is MVUE of $\gamma(\theta)$ if $E_\theta(T)=\gamma(\theta)$ for all $\theta \in \Theta$ and $\text{Var}_\theta(T) \leq \text{Var}_\theta(T')$ for all $\theta \in \Theta$ where T' is any other unbiased estimator of $\gamma(\theta)$.

Let us discuss some important theorems concerning MVUE.

- An MVUE is unique in the sense that if T_1 and T_2 are MVUE for $\Upsilon(\theta)$, then $T_1 = T_2$, almost surely.
- Let T_1 and T_2 be unbiased estimators of $\gamma(\theta)$ with efficiencies e_1 and e_2 respectively and $\rho = \rho_\theta$ be the correlation coefficient between them, then $\sqrt{(e_1e_2)} - \sqrt{\{(1-e_1)(1-e_2)\}} \leq \rho \leq \sqrt{(e_1e_2)} + \sqrt{\{(1-e_1)(1-e_2)\}}$.
- If T_1 is an MVUE of $\gamma(\theta)$, $\theta \in \Theta$ and T_2 is any other unbiased estimator $\gamma(\theta)$ with efficiency $e = e_\theta$, then the correlation coefficient between T_1 and T_2 is given by $\rho = \sqrt{e}$, i.e., $\rho_\theta = \sqrt{e_\theta}$, for all $\theta \in \Theta$.

20.3.4 Sufficiency

An estimator is said to be sufficient for a parameter, if it contains all the information in the sample regarding the parameter. More precisely, if $T_n = T(x_1, x_2, \ldots, x_n)$ is an estimator of a parameter θ, based on a sample $x_1, x_2, \ldots, x_n$ of size 'n' from the population with density $f(x, \theta)$ such that the conditional distribution of $x_1, x_2, \ldots, x_n$ given T, is independent of θ, then T is sufficient estimator for θ.

Factorisation Theorem (Neyman)

The necessary and sufficient condition for a distribution to admit sufficient statistic is provided by the 'factorization theorem' due to Neyman.

Statement: $T = t(x)$ is sufficient for θ if and only if the joint density function L (say) of the sample values can be expressed in the form $L = g_\theta [t(x)].h(x)$ where (as indicated) $g_\theta[t(x)]$ depends on θ and x only through the value of $t(x)$ and $h(x)$ is independent of θ.

Note:

i) It should be clearly understood that by 'a function independent of θ' we not only mean that it does not involve θ but also that its domain does not contain θ. For example, the function $f(x) = 1/2a$; $a-\theta < x < a+\theta$ and $-\infty < \theta < \infty$, depends on θ.

ii) It should be noted that the original sample $X = (X_1, X_2, \ldots, X_n)$, is always a sufficient statistic.

iii) The most general form of the distributions admitting sufficient statistic is Koopman's form and is given by $L = L(x, \theta) = g(x).h(\theta).\exp\{a(\theta)\ \psi(x)\}$ where $h(\theta)$ and $a(\theta)$ are functions of the parameter θ only and $g(x)$ and $\psi(x)$ are the functions of the sample observations only. The above equation represents the famous 'exponential family of distributions', of which most of the common distributions like the binomial, the Poisson and the normal with unknown mean and variance are the members.

iv) **Invariance property of sufficient estimator**

If T is a sufficient estimator for the parameter θ, and $\psi(t)$ is a one to one function of T, then $\psi(t)$ is sufficient for $\psi(\theta)$.

v) **Fisher – Neyman Criterion**

A statistic $t_1 = t_1(x_1, x_2, \ldots, x_n)$ is sufficient estimator of parameter θ if and only if the likelihood function (joint p.d.f. of the sample) can be expressed as:

$$L = \prod_{i=1}^{n} f(x_i, \theta)$$

$$= g_1(t_1, \theta).\ k(x_1, x_2, ..., x_n)$$

where $g_1(t_1, \theta)$ is the p.d.f. of statistic t_1 and $k(x_1, x_2, \ldots, x_n)$ is a function of sample observations only, independent of θ.

Note that this method requires working out of the p.d.f. (p.m.f.) of the statistic $t_1(x_1, x_2, \ldots, x_n)$, which is not always easy.

Check Your Progress 1

1) Discuss the meaning of point estimation and interval estimation.

.....

.....

.....

.....

2) List the characteristics of a good estimator.

.....

.....

.....

.....

3) $x_1, x_2, \ldots, x_n$ is a random sample from a normal population N (μ, 1). Show that $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$, is an unbiased estimator of $\mu^2 + 1$.

.....

.....

.....

.....

4) A random sample (X_1, X_2, X_3, X_4, X_5) of size 5 is drawn from a normal population with unknown mean μ. Consider the following estimators t o estimate μ:

i) $t_1 = (X_1 + X_2 + X_3 + X_4 + X_5) / 5$; and (ii) $t_2 = (2X_1 + X_2 + \lambda X_3) / 3$.

Find λ. Are t_1 and t_2 unbiased? State with giving reasons, the estimator which is best among t_1 and t_2.

.....

.....

.....

.....

5) Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with p.d.f $f(x,\theta) = \theta x^{\theta - 1}$; $0<x<1$, $\theta>0$. Show that $t_1 = \prod_{i=1}^{n} X_i$, is sufficient for θ.

.....

.....

.....

.....

20.4 CRAMER-RAO INEQUALITY

If t is an unbiased estimator of $\gamma(\theta)$, a function of parameter θ, then

$$\text{var}(t) \geq \frac{\left[\frac{\partial}{\partial\theta} L(x, \theta)\right]^2}{E\left[\frac{\partial}{\partial\theta} \log L\right]} = \frac{\left[\gamma'(\theta)\right]^2}{I(\theta)} \text{ and } I(\theta) = E\left[\left\{\frac{\partial}{\partial\theta} \log L(x, \theta)\right\}^2\right],$$

where I (θ) is the information on θ, supplied by the sample. In other words, Cramer-Rao inequality provides a lower bound $\frac{\left[\gamma'(\theta)\right]^2}{I(\theta)}$ to the variance of an unbiased estimator of $\gamma(\theta)$.

The Cramer-Rao inequality holds given the following assumptions, which are known as the 'regularity conditions for Cramer-Rao inequality'.

i) The parameter space Θ is a non degenerate open interval on the real line R^1 $(-\infty, \infty)$.

ii) For almost all $x = (x_1, x_2, \ldots, x_n)$, and for all $\theta \in \Theta$, $\frac{\partial}{\partial\theta} L(x, \theta)$ exists, the exceptional set, if any, is independent of θ.

iii) The range of integration is independent of the parameter θ, so that $f(x, \theta)$ is differentiable under integral sign.

iv) The conditions of uniform convergence of integrals are satisfied so that differentiation under the integral sign is valid.

v) $I(\theta) = E\left[\left\{\frac{\partial}{\partial\theta} \log L(x, \theta)\right\}^2\right]$ exists and is positive for all $\theta \in \Theta$.

An unbiased estimator t of $\gamma(\theta)$ for which Cramer-Rao lower bound is attained is called a minimum variance bound (MVB) estimator.

20.4.1 MVUE and Blackwellisation

Cramer-Rao inequality provides us a technique of finding if the unbiased estimator is also an MVUE or not. Here, since the regularity conditions are very strict, its application becomes quite restrictive. Moreover MVB estimator is not the same as MVUE estimator since the Cramer-Rao lower bound may not always be. Moreover, if the regularity conditions are violated, then the least attainable variance may be less than the Cramer-Rao bound. There is a method of obtaining MVUE from an unbiased estimator through the use of sufficient statistic. This technique is called Blackwellisation after D. Blackwell. The result is contained in the following theorem due to C. R. Rao and D. Blackwell.

20.4.2 Rao-Blackwell Theorem

Let X and Y be random variables such that $E(Y) = \mu$ and $var(Y) = \sigma_Y^2 > 0$

Let $E(Y \mid X = x) = \varphi(x)$, then

i) $E[\varphi(x)] = \mu$ and (ii) $var[\varphi(x)] \leq var(Y)$

Thus, Rao-Blackwell theorem enables us to obtain MVUE through sufficient statistic. If a sufficient estimator exists for a parameter, then in our search for MVUE we may restrict ourselves to functions of the sufficient statistic.

The theorem can be stated slightly different way as follows:

Let $U = U(x_1, x_2, \ldots, x_n)$ be an unbiased estimator of parameter $\gamma(\theta)$ and let $T = T(x_1, x_2, \ldots, x_n)$ be sufficient statistic for $\gamma(\theta)$. Consider the function $\varphi(T)$ of the sufficient statistic defined as $\varphi(T) = E(Y \mid T = t)$ which is independent of θ (since T is sufficient for $\gamma(\theta)$. Then $E\varphi(T) = \gamma(\theta)$ and $var\ \varphi(T) \leq var(U)$.

This result implies that starting with an unbiased estimator U, we can improve upon it by defining a function $\varphi(T)$ of the sufficient statistic given as $\varphi(T) = E(Y \mid T = t)$. This technique of obtaining improved estimator is called Blackwellisation.

If in addition, the sufficient statistic T is also complete, then the estimator $\varphi(T)$ discussed above will not only be an improved estimator over U but also the 'best (unique)' estimator.

20.5 TESTS OF SIGNIFICANCE

A very important aspect of the sampling theory is the study of tests of significance, which enables us to decide on the basis of the sample results, if (i) the deviation between the observed sample statistic and the hypothetical parameter value, or (ii) the deviation between two independent sample statistics, is significant or might be attributed to chance or the fluctuations of sampling.

Since, for large 'n', almost all the distributions, e.g., binomial, poisson, negative binomial, hypergeometric, t, F, chi-square, can be approximated very closely by a normal probability curve, we use the 'normal test of significance' for large samples. Some of the well-known tests of significance for studying such differences for small samples are t-test, F-test and Fisher's z-transformation.

20.5.1 Null Hypothesis

The technique of randomization used for the selection of sample units makes the test of significance valid for us. For applying the test of significance we first set up a hypothesis – a definite hypothesis of no difference, is called 'null hypothesis' and usually denoted by H_0. According to Professor R. A. Fischer, null hypothesis is the hypothesis, which is tested for possible rejection under the assumption that is true.

For example, in case of single statistic, H_0 will be that the sample statistic does not differ significantly from the hypothetical parameter value and in the case of two statistics, H_0 will be that the sample statistics do not differ significantly. Having set up the null hypothesis, we compute the probability P that the deviation between the observed sample statistic and the hypothetical parameter value might have occurred due to fluctuations of sampling. If the deviation comes out to be significant (as measured by a test of significance), null hypothesis is refuted or rejected at the particular level of significance adopted and if the deviation is not significant, null hypothesis may be retained at that level.

20.5.2 Alternative Hypothesis

Any hypothesis, which is complementary to the null hypothesis is called an alternative hypothesis, usually denoted by H_1. For example, if we want to test the null hypothesis that the population has a specified mean μ, (say), i.e., H_0: $\mu = \mu_0$, then the alternative hypothesis could be

i) H_1: $\mu \neq \mu_0$ (i.e., $\mu > \mu_0$ or, $\mu < \mu_0$)

ii) H_1: $\mu > \mu_0$

iii) H_1: $\mu < \mu_0$

The alternative hypothesis in (i) is known as 'two-tailed alternative' and the alternatives in (ii) and (iii) are known as 'right-tailed alternative' and 'left-tailed alternative', respectively. The setting of alternative hypothesis is very important since it enables us to decide whether we have to use a single-tailed (right or left) or two-tailed test.

20.5.3 Critical Region and Level of Significance

A region (corresponding to a statistic t) in the sample space S that amounts to rejection of H_0 is termed as 'critical region'. If ω is the critical region and if $t = t(X_1, X_2, \ldots, X_n)$ is the value of the statistic based on a random sample of size 'n', then $P[t \in \omega \mid H_0] = \alpha$ and $P[t \in \omega' \mid H_1] = \beta$, where ω', the complementary set of ω, is called the 'acceptance region'. We have $\omega \cup \omega' = S$ and $\omega \cap \omega' = \varphi$. The probability of α that a random value of the statistic t belongs to the critical region is known as the 'level of significance'. In other words, level of significance is the size of the type I error (or the maximum producer's risk). The levels of significance usually employed in testing of hypothesis are 5% and 1%. The level of significance is always fixed in advance before collecting the sample information.

20.5.4 Confidence Interval and Confidence Limits

Let x_i, ($i = 1, 2, .., n$) be a random sample of 'n' observations from a population involving a single unknown parameter θ (say). Let $f(x, \theta)$ be the probability function of the parent distribution from which the sample is drawn and let us suppose that this distribution is continuous. Let $t = t(x_1, x_2, \ldots, x_n)$, a function of the sample values be an estimate of the population parameter θ, with the sampling distribution given by $g(t, \theta)$.

Having obtained the value of the statistic t from a given sample, the problem is, "Can we make some reasonable probability statements about the unknown parameter θ in the population, from which the sample has been drawn?" this question is very well answered by the technique of 'confidence interval' due to Neyman and is obtained below:

We choose once for all some small value of α (5% or 1%) and then determine two constants, say, c_1 and c_2 such that $P[c_1 < \theta < c_2] = 1-\alpha$. The quantities c_1 and c_2, so determined, are known as the 'confidence limits' and the interval $[c_1, c_2]$ within which the unknown value of the population parameter is expected to lie, is called the 'confidence interval' and $(1-\alpha)$ is called the 'confidence coefficient'. Thus, if we take $\alpha = 0.05$ (or 0.01), we shall get 95% (or 99%) confidence limits.

How to find c_1 and c_2?

Let T_1 and T_2 be two statistics such that $P(T_1 > \theta) = \alpha_1$ and $P(T_2 > \theta) = \alpha_2$ where α_1 and α_2 are constants independent of θ. So, it can be written that $P(T_1 < \theta < T_2) = 1-\alpha$ where $\alpha = \alpha_1 + \alpha_2$. Statistics T_1 and T_2 may be taken as c_1 and c_2 as defined in the last section.

For example, if we take a large sample from a normal population with mean μ and standard deviation σ, then $Z = \dfrac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0, 1)$

and $P(-1.96 < Z < 1.96) = 0.95$ [from normal probability tables]

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

Thus, $\bar{x} \pm 1.96\dfrac{\sigma}{\sqrt{n}}$ are 95% confidence limits for the unknown parameter μ – the population mean and the interval

$[\bar{x} - 1.96\dfrac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\dfrac{\sigma}{\sqrt{n}}]$ is called the 95% confidence interval.

20.5.5 One-Tailed and Two-Tailed Tests

In any test, the critical region is represented by a portion of the area under the probability curve of the sampling distribution of the test statistic.

A test of any statistical hypothesis where the alternative hypothesis is one-tailed (right-tailed or left-tailed) is called a 'one-tailed test'. For example, a test for testing the mean of a population H_0: $\mu = \mu_0$ against the alternative hypothesis: H_1: $\mu > \mu_0$ (right –tailed) or H_1: $\mu < \mu_0$ (left-tailed), is a 'single-tailed test'. In the right-tailed test (H_1: $\mu > \mu_0$), the critical region lies entirely in the right tail of the sampling distribution of $\bar{x}$, while for the left-test (H_1: $\mu < \mu_0$), the critical region is entirely in the left tail of the distribution. Let us consider a test of statistical hypothesis where the alternative hypothesis is two-tailed such as: H_0: $\mu = \mu_0$, against the alternative hypothesis H_1: $\mu \neq \mu_0$ ($\mu > \mu_0$

and $\mu < \mu_0$) is known as 'two-tailed test' and in such a case the critical region is given by the portion of the area lying in both the tails of the probability curve of the test statistic.

In a particular problem, whether one-tailed or two-tailed test is to be applied depends entirely on the nature of the alternative hypothesis. If the alternative hypothesis is two-tailed we apply the two-tailed test and if alternative hypothesis is one-tailed, we apply one-tailed test.

For example, suppose there are two popular brands of bulbs, one manufactured by standard process (with mean life μ_1) and the other manufactured by some new technique (with mean life μ_2). To test if the bulbs differ significantly, our null hypothesis is H_0: $\mu_1 = \mu_2$ and the alternative will be H_1: $\mu_1 \neq \mu_2$, thus giving us a two-tailed test. However, if we want to test if the bulbs produced by new process have higher average life than those produced by standard process, then we have H_0: $\mu_1 = \mu_2$ and H_1: $\mu_1 < \mu_2$, thus giving is a left-tailed test. Similarly, for testing if the product of new process is inferior to that of standard process, we have set: H_0: $\mu_1 = \mu_2$ and H_1: $\mu_1 > \mu_2$, thus giving is a right-tailed test. Accordingly, the decision about applying a two-tailed test or a single-tail test (right or left) will depend on the problem under study.

20.5.6 Critical Values or Significant Values

The value of the test statistic, which separates the critical (or rejection) region and the acceptance region is called the 'critical value' or 'significant value'. It depends upon: (i) the level of significance used, and (ii) the alternative hypothesis, whether it is two-tailed or single-tailed. As has been pointed out earlier, for large samples, the standardized variable asymptotically corresponding to the statistic 't' viz., $Z = \dfrac{t - E(t)}{S.E\,(t)} \sim N\,(0, 1)$, as $n \to \infty$. The value of Z given by the above relation under the null hypothesis is known as the 'test statistic'. The critical value of test statistic at level of significance α for a two-tailed-test is given by z_α where z_α is determined by the equation $P\,[|Z| > z_\alpha] = \alpha$ i.e., z_α is the value so that the total area of the critical region on both tails is α. Since normal probability curve is symmetrical curve, so from $P\,[|Z| > z_\alpha] = \alpha$ we can write,

$P\,[Z > z_\alpha] + P\,[Z < -z_\alpha] = \alpha$

$\Rightarrow P\,[Z > z_\alpha] + P\,[Z < z_\alpha] = \alpha$

$\Rightarrow 2P\,[Z > z_\alpha] = \alpha$

$\Rightarrow P\,[Z > z_\alpha] = \alpha/2$

i.e., the area of each tail is $\alpha/2$.

Thus, z_α is the value such that area to the right of z_α is $\alpha/2$ and to the left of $-z_\alpha$ is $\alpha/2$.

In case of single-tail alternative, the critical value z_α is determined so that total area to the right of it (for right-tailed test) is α and for left-tailed test the total area to the left of $-z_\alpha$ is α, i.e., $P\,[Z > z_\alpha] = \alpha$ (for right-tailed test) and $P\,[Z < -z_\alpha] = \alpha$ (for left tailed test).

Thus, the significant or critical value of Z for a single-tailed test (left or right) at level of significance ' α ' is the same as the critical value of Z for a two-tailed test at level of significance '2α'. The critical values of Z at commonly used levels of significance for both two-tailed and single-tailed test are given in the following table:

Critical values (z_α)	Level of significance (α)		
	1%	5%	10%
Two-tailed test	$\|Z_\alpha\| = 2.58$	$\|Z_\alpha\| = 1.96$	$\|Z_\alpha\| = 1.645$
Right-tailed test	$Z_\alpha = 2.33$	$Z_\alpha = 1.645$	$Z_\alpha = 1.28$
Left-tailed test	$Z_\alpha = -2.33$	$Z_\alpha = -1.645$	$Z_\alpha = -1.28$

Note:

If n is small (usually less than 30), then the sampling distribution of the test statistic Z will not be normal and in that case we cannot use the above significant values, which have been obtained from normal probability curves.

Procedure for Testing of Hypothesis

We now summarise below the various steps in testing of a statistical hypothesis in a systematic manner.

1) **Null Hypothesis:** Set up the null hypothesis H_0.

2) Alternative Hypothesis: Set up the alternative hypothesis H_1. This will enable us to decide whether we have to use a single-tailed (right or left) test or two-tailed test.

3) **Level of Significance:** Choose the appropriate level of significance (α) depending on the reliability of the estimates and permissible risk. This is to be decided before sample is drawn, i.e., α is fixed in advance.

4) **Test Statistic (or Test Criterion):** Compute the test statistic $z = \dfrac{t - E(t)}{S.E(t)}$ under the null hypothesis.

5) **Conclusion:** We compare z, the computed value of Z in step 4 with the significant value (tabulated value) z_α, at the given level of significance ' α '.

If $|Z| < z_\alpha$, i.e., if the calculated value of Z (in modulus value) is less than z_α, we say it is not significant. By this, we mean that the difference t – E(t) is just due to fluctuations of sampling and the sample data do not provide us sufficient evidence against the null hypothesis which may therefore, be accepted.

If $|Z| > z_\alpha$, i.e., if the calculated value of Z is greater than the critical or significant value, then we say that it is significant and the null hypothesis is rejected at level of significance ' α ', i.e., with confidence coefficient (1 - α).

Check Your Progress 2

1) What is Cramer-Rao inequality?

.....

.....

.....

.....

2) What do you mean by test of significance?

.....

.....

.....

.....

3) What is the purpose of hypothesis testing?

.....

.....

.....

.....

20.6 TYPE I AND TYPE II ERROR

The main objective in sampling theory is to draw valid inferences about the population parameters on the basis of the sample results. In practice, we decide to accept or reject the lot after examining the examining the sample from it. As such, we are liable to commit the following two types of errors:

Type I Error: Reject H_0 when it is true.

Type II Error: Accept H_0 when it is wrong, i.e., accept H_0 when H_1 is true.

If we write, P [Reject H_0 when it is true] = P [Reject $H_0 \mid H_0$] = α

and P [Accept H_0 when it is wrong] = P [Accept $H_0 \mid H_1$] = β

then α and β are called the sizes of type I error and type II error respectively.

In practice, type I error amounts to rejecting a lot when it is good and type II error may be regarded as accepting the lot when it is bad.

Thus, P [Reject a lot when it is good] = α

and P [Accept a lot when it is bad] = β

where α and β are referred to as 'Producer's risk' and Consumer's risk' respectively.

The probability of type I error is necessary for constructing a test of significance. It is in fact the 'size of the critical region'. The probability of type II error is used to measure the 'power' of the test in detecting the falsity of the null hypothesis.

It is desirable that the test procedure be so framed, which minimises both the types of error. But this is not possible, because for a given sample size, an attempt to reduce one type of error is generally accompanied by an increase in the other type. The test of significance is designed so as to limit the probability of type I error to a specified value (usually 5% or 1%) and at the same time to minimise the probability of type II error. Note that when the population has a continuous distribution,

Probability of type I error = Level of significance = Size of critical region

20.7 POWER OF THE TEST

The null hypothesis is accepted when the observed value of test statistic lies outside the critical region as determined by the test procedure. Type II error is committed when alternative hypothesis holds even though null hypothesis is not rejected, i.e., the test statistic lies outside the critical region. Hence, the probability of type II error is a function of the value for which alternative hypothesis holds.

If β is the probability of type II error (i.e., probability of accepting H_0 when H_0 is false), then (1- β) is called the 'power function' of the test hypothesis H_0 against the alternative hypothesis H_1. The value of the power function at a parameter point is called the 'power of the test' at that point.

Power = 1 – Probability of Type II error = Probability of rejecting H_0 when H_1 is true.

20.8 OPTIMUM TEST UNDER DIFFERENT SITUATIONS

The discussion till now enables us to obtain the so-called best test under different situations. In any testing problem, the first two steps, viz., the form of the population distribution, the parameter(s) of interest and the framing of H_0 and H_1 should be obvious from the description of the problem. The most crucial step is the choice of the 'best test', i.e., the basic statistic 't' and the critical region W whereby best test we mean one which in addition to controlling α at any desired low level has the minimum type II error β or maximum power (1- β), compared to β of all other tests having this α. This leads to the following definition.

20.8.1 Most Powerful Test

Let us consider the problem of testing a simple hypothesis H_0: $\theta = \theta_0$ against a simple alternative hypothesis H_1: $\theta = \theta_1$

Definition: The critical region W is the most powerful critical region of size α (and the corresponding test as most powerful test of level α) for testing H_0: $\theta = \theta_0$ against H_1: $\theta = \theta_1$ if P $(\mathbf{x} \in W \mid H_0) = \alpha$ and P $(\mathbf{x} \in W \mid H_1) \geq$ P $(\mathbf{x} \in W_1 \mid H_1)$ for every other critical region W_1 satisfying the first condition.

20.8.2 Uniformly Most Powerful Test

Let us now take up the case of testing a simple null hypothesis H_0: $\theta = \theta_0$ against a composite alternative hypothesis H_1: $\theta \neq \theta_0$. In such a case, for a predetermined α, the best test for H_0 is called the uniformly most powerful test of level α.

Definition: The critical region W is called uniformly most powerful critical region of size α (and the corresponding test as uniformly most powerful test of level α) for testing H_0: $\theta = \theta_0$ against H_1: $\theta \neq \theta_0$. If P $(\mathbf{x} \in W \mid H_0) = \alpha$ and P $(\mathbf{x} \in W \mid H_1) \geq$ P $(\mathbf{x} \in W_1 \mid H_1)$ for all $\theta \neq \theta_0$, whatever the other critical region W_1 satisfying the first condition may be.

20.9 TEST PROCEDURE UNDER NORMALITY ASSUMPTION

The general procedure to be followed in testing a statistical hypothesis has been explained in the previous sections.

We shall now take up one by one some of the common tests that are made on the assumption of normality for the underlying random variable or variables.

20.9.1 Problems Regarding the Univariate Normal Distribution

Consider a population where x is normally distributed with mean μ and standard deviation σ. Let $x_1, x_2, \ldots, x_n$ be a random sample obtained from this distribution. We shall denote by $\bar{x}$ the sample mean of x: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and by s'^2 the sample variance of x: $s'^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. The distinction between s^2 and s'^2 is to be noted. In s'^2 the divisor is (n–1), which makes it as unbiased estimator of σ^2.

For $\sum_{i=1}^{n}(x_1 - \bar{x})^2 = \sum_{i}(x_i - \mu)^2 - n(\bar{x} - \mu)^2$,

so that $E(s'^2) = \frac{1}{n-1} E\{\sum_{i=1}^{n}(x_i - \bar{x})^2\} = \frac{1}{n-1} E\{\sum_{i=1}^{n}(x_i - \mu)^2 - n(\bar{x} - \mu)^2\}$

$$= \frac{1}{n-1}\{\sum_{i} \text{var}(x_i) - n\,\text{var}(\bar{x})\} = \frac{1}{n-1}\{n\sigma^2 - n\frac{\sigma^2}{n}\} = \sigma^2.$$

Case I: μ unknown, σ known

Here we may be required to test the null hypothesis H_0: $\mu = \mu_0$. It has already been shown that the test procedure for H_0 in this case is based on the statistic $\sqrt{n}(\bar{x} - \mu_0)/\sigma$, which is distributed as a normal deviate (τ) under this hypothesis.

a) For the alternative H: $\mu > \mu_0$, H_0 is rejected if for the given sample $\tau > \tau_a$ (and is accepted otherwise)

b) For the alternative H: $\mu < \mu_0$, H_0 is rejected if for the given sample $\tau < \tau_{1-a}$ $(= -\tau_a)$.

c) For the alternative H: $\mu \neq \mu_0$, H_0 is rejected if for the given sample $|\tau| > \tau_{a/2}$.

In each case, α denotes the chosen level of significance.

As regards the problem of interval estimation of μ, it has been shown that the limits $(\bar{x} - \tau_{a/2}.\sigma/\sqrt{n})$ and $(\bar{x} + \tau_{a/2}.\sigma/\sqrt{n})$, computed for the given sample, are the confidence limits for μ with confidence coefficient $(1 - \alpha)$.

Case II: μ known, σ unknown

Here one may be interested in testing a hypothesis regarding σ or in estimating σ.

A sufficient statistic for σ is $\sum_i (x_i - \mu)^2$ or, $\sqrt{\sum_i (x_i - \mu)^2}$. It is seen that x_i is a normal variable with mean μ and standard deviation σ.

Hence, $\sum_i \left(\frac{x_i - \mu}{\sigma}\right)^2$, being the sum of squares of n independent normal deviates, is distributed as χ^2 with df = n.

For testing H_0: $\sigma = \sigma_0$, we make use of the fact

$\sum_i \left(\frac{x_i - \mu}{\sigma}\right)^2 = \sum_i (x_i - \mu)^2/\sigma_0^2$ is a χ^2 with df = n under this hypothesis.

a) For the alternative H: $\sigma > \sigma_0$, H_0 is rejected in case for the given sample $\chi^2 > \chi^2_{a,n}$.

b) For the alternative H: $\sigma < \sigma_0$, H_0 is rejected if for the given sample $\chi^2 < \chi^2_{1-a,n}$.

c) For the alternative H: $\sigma \neq \sigma_0$, H_0 is rejected if for the given sample $\chi^2 < \chi^2_{1-a/2,n}$ or, $\chi^2 > \chi^2_{a/2,n}$.

As consistent (but biased) point estimate of σ, we have $\sqrt{\frac{1}{n}\sum_i (X_i - \mu)^2}$. To get a confidence interval of σ, we note that

$$P\left[\chi^2_{1-a/2,n} \le \sum_i \frac{(x_i - \mu)^2}{\sigma^2} \le \chi^2_{a/2,n}\right] = 1 - \alpha$$

$$\text{or, } P\left[\sum_i \frac{(x_i - \mu)^2}{\chi^2_{a/2,n}} \le \sigma^2 \le \sum_i \frac{(x_i - \mu)^2}{\chi^2_{1-a/2,n}}\right] = 1 - \alpha$$

$$\text{or, } P\left[\sum_i \frac{(x_i - \mu)^2}{\chi^2_{a/2,n}} \le \sigma^2 \le \sum_i \frac{(x_i - \mu)^2}{\chi^2_{1-a/2,n}}\right] = 1 - \alpha$$

The confidence limits of σ^2 with confidence coefficient $(1 - \alpha)$ are therefore, $\frac{(x_i - \mu)^2}{\chi^2_{a/2,n}}$ and $\frac{(x_i - \mu)^2}{\chi^2_{1-a/2,n}}$. The confidence limits of σ are just the positive square roots of these quantities, with the same confidence coefficient, $(1 - \alpha)$.

Case III: μ and σ both unknown

In this case, $\bar{x}$ and s' are jointly sufficient for μ and σ. Here to test H_0: $\mu = \mu_0$ or to have confidence limits to μ, one cannot use the statistic $\sqrt{n}(\bar{x} - \mu)/\sigma$ since σ is unknown. σ is in this case is replaced by its sample estimate, s' = $\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$.

The resulting expression will be $\sqrt{n}(\bar{x} - \mu)/s'$.

Now from the discussion made in the last unit, it is clear that $\frac{(n-1)s'^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sigma^2}$ is a χ^2 with df = (n-1) and is distributed independently of $\bar{x}$. Thus, $\sqrt{n}(\bar{x} - \mu)/s' = \frac{\sqrt{n}(\bar{x} - \mu)/\sigma}{\sqrt{\frac{(n-1)s'^2}{\sigma^2}/(n-1)}}$, being of the form $\frac{\tau}{\sqrt{\chi^2/(n-1)}}$, where χ^2 has df = $(n-1)$ and is independent of τ, is distributed as a 't' with df=$(n-1)$

To test H_0: $\mu = \mu_0$ we may, therefore use statistic $t = \sqrt{n}(\bar{x} - \mu_0)/s'$ with df = (n-1). We shall have to compute 't' (computed from the given sample) with $t_{\alpha, n-1}$ or 't' with $-t_{\alpha, n-1}$ or $|t|$ with $t_{\alpha/2, n-1}$, according as the alternative of interest is H: $\mu > \mu_0$, H: $\mu < \mu_0$ or H: $\mu \neq \mu_0$.

In order to obtain confidence limit to μ, we see that

$$P\left[(\bar{x} - t_{a/2,\,n-1})\frac{s'}{\sqrt{n}} \le \mu \le (\bar{x} + t_{a/2,\,n-1})\frac{s'}{\sqrt{n}}\right] = 1-\alpha$$

or, $$P\left[(\bar{x} - t_{a/2,\,n-1})\frac{s'}{\sqrt{n}} \le \mu \le (\bar{x} + t_{a/2,\,n-1})\frac{s'}{\sqrt{n}}\right] = 1-\alpha$$

The $100(1-\alpha)\%$ confidence limits to μ will, therefore, be $(\bar{x} - t_{a/2,\,n-1})\frac{s'}{\sqrt{n}}$ and $(\bar{x} + t_{a/2,\,n-1})\frac{s'}{\sqrt{n}}$, these being computed from the given sample.

The procedure has been called as *Student's t-test.*

In this case, we may have also the problem of testing H_0: $\sigma = \sigma_0$ or the problem of obtaining confidence limits to σ. From what has been said above, it is clear that $\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sigma_0^2} = \frac{(n-1)\,s'^2}{\sigma_0^2}$ is, under the hypothesis H_0, a χ^2 with df = (n–1). This provides us with test for H_0. The value of this χ^2, computed from the given sample, is compared with $\chi^2_{a,n-1}$ or $\chi^2_{1-a,n-1}$, according as the alternative is H: $\sigma > \sigma_0$ or H: $\sigma < \sigma_0$.

For the alternative H: $\sigma \neq \sigma_0$, on the other hand, the computed value is to be compared with both $\chi^2_{1-a/2,n-1}$ and $\chi^2_{a/2,n-1}$, H_0 being rejected if the computed value is smaller than the former or exceeds the latter value.

Since $$P\left[\chi^2_{1-a/2,n-1} \le \frac{(n-1)\,s'^2}{\sigma^2} \le \chi^2_{a/2,n-1}\right] = 1-\alpha$$

i.e., $$P\left[\frac{(n-1)\,s'^2}{\chi^2_{a/2,n-1}} \le \sigma^2 \le \frac{(n-1)\,s'^2}{\chi^2_{1-a/2,n-1}}\right] = 1-\alpha$$

The confidence limits to σ^2 are $\frac{(n-1)\,s'^2}{\chi^2_{a/2,n-1}}$ and $\frac{(n-1)\,s'^2}{\chi^2_{1-a/2,n-1}}$. The confidence limits with the same confidence coefficient $(1-\alpha)$ to σ are, of course, the positive square roots of these quantities.

Check Your Progress 3

1) Which is wider, a 95% or a 99% confidence interval?

.....

.....

.....

.....

2) When you construct a 95% confidence interval, what are you 95% confident about?

.....

.....

.....

.....

3) When computing a confidence interval, when do you use t and when do you use z?

.....

.....

.....

.....

4) What Greek letters are used to represent the Type I and II error rates.

.....

.....

.....

.....

5) What levels are conventionally used for significance testing?

.....

.....

.....

.....

6) When is it valid to use a one-tailed test? What is the advantage of a one-tailed test? Give an example of a null hypothesis that would be tested by a one-tailed test.

.....

.....

.....

.....

7) Distinguish between probability value and significance level?

.....

.....

.....

.....

8) The following are 12 determinations of the melting point of a compound (in degrees centigrade) made by an analyst, the true melting point being 165^0C. would you conclude from these data that her determination is free from bias?

164.4, 161.4, 169.7, 162.2, 163.9, 168.5, 162.1, 163.4, 160.9, 162.9, 160.8, 167.7.

.....

.....

.....

.....

.....

20.9.2 Comparison of Two Univariate Normal Distributions

Let the distribution of x in each of two populations be normal. Suppose the mean and the standard deviation of x for one population are μ_1 and σ_1, for the other they are μ_2 and σ_2, respectively. Suppose further that $x_{11}, x_{12}, \ldots, x_{1n}$ are a random sample from the first distribution, and $x_{21}, x_{22}, \ldots, x_{2n}$ are a random sample for the second set. The first set of observation is also supposed to be independent of the second set. Then $\bar{x}_1 = \sum_j^{n_1} x_{1j} / n_1$ and $s_1' = \sqrt{\sum_j^{n_1} (x_{1j} - \bar{x}_1)^2 / (n_1 - 1)}$ are the mean and standard deviation of x in the first sample, and $\bar{x}_2 = \sum_j^{n_2} x_{2j} / n_2$, $s_2' = \sqrt{\sum_j^{n_2} (x_{2j} - \bar{x}_2)^2 / (n_2 - 1)}$ are the corresponding statistics of the second sample.

Case I: μ_1, μ_2 unknown but σ_1, σ_2 known

In this case one may be concerned with a comparison between the population means. One may have to test the hypothesis that μ_1 and μ_2 differ by a specified quantity, say H_0: $\mu_1 - \mu_2 = \xi_0$, or one may like to obtain confidence limits for the difference $\mu_1 - \mu_2$.

It may be seen that $\bar{x}_1 - \bar{x}_2$, being a linear function of normal variables, is itself normally distributed. It has mean $E(\bar{x}_1 - \bar{x}_2) = E(\bar{x}_1) - E(\bar{x}_2) = \mu_1 - \mu_2$ and variance $\text{var}(\bar{x}_1 - \bar{x}_2) = \text{var}(\bar{x}_1) + \text{var}(\bar{x}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$, and the covariance term being zero since $\bar{x}_1$ and $\bar{x}_2$ are independent. As such $\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})^{1/2}}$ is distributed as a standard normal variable. To test H_0: $\mu_1 - \mu_2 = \xi_0$, we make

use of the statistic $\frac{(\bar{x}_1 - \bar{x}_2) - \xi_0}{(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})^{1/2}}$, which is distributed as a standard normal variable (τ) under H_0. H_0 is to be rejected on the basis of the given samples if $\tau > \tau_a$ or if $\tau < -\tau_a$, according as the alternative hypothesis in which the experimenter is interested in H: $\mu_1 - \mu_2 > \xi_0$ H: $\mu_1 - \mu_2 < \xi_0$.

On the other and, if the alternative is H: $\mu_1 - \mu_2 \neq \xi_0$, H_0 is to be rejected when $|\tau| > \tau_{a/2}$. In the commonest cases, the null hypothesis will be H_0: $\mu_1 = \mu_2$ for which $\xi_0 = 0$. If the problem is one of interval estimation, then it will be found, following the usual mode of argument that the confidence limits to $\mu_1 = \mu_2$ (with confidence coefficient $1 - \alpha$) are $(\bar{x}_1 - \bar{x}_2) - \tau_{a/2}(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})^{1/2}$ and $(\bar{x}_1 - \bar{x}_2) + \tau_{a/2}(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})^{1/2}$.

Case II: μ_1, μ_2 known but σ_1, σ_2 unknown

Here it may be necessary to test the hypothesis that the ratio of the two unknown standard deviations has a specified value, say H_0: $\sigma_1 / \sigma_2 = \xi_0$, or to set confidence limits to this ratio. Since $\sum_{j=1}^{n_1}(x_{1j} - \mu_1)^2 / \sigma_1^2$ and $\sum_{j=1}^{n_2}(x_{2j} - \mu_2)^2 / \sigma_2^2$ are independent χ^2s with n_1 and n_2 degrees of freedom, respectively,

$$P\left[\frac{1}{F_{a/2;\,n2,\,n1}} \leq \frac{\sum_{j=1}^{n_1}(x_{1j} - \mu_1)^2 / n_1\sigma_1^2}{\sum_{j=1}^{n_2}(x_{2j} - \mu_2)^2 / n_2\sigma_2^2} \leq \frac{1}{F_{a/2;\,n1,\,n2}}\right] = 1 - \alpha$$

$\frac{\sum_{j=1}^{n_1}(x_{1j} - \mu_1)^2 / n_1\sigma_1^2}{\sum_{j=1}^{n_2}(x_{2j} - \mu_2)^2 / n_2\sigma_2^2}$ is distributed as an F with (n_1, n_2) df.

Under the hypothesis H_0: $\sigma_1 / \sigma_2 = \xi_0$, therefore, $\frac{\sum_{j=1}^{n_1}(x_{1j} - \mu_1)^2 / \sigma_1^2}{\sum_{j=1}^{n_2}(x_{2j} - \mu_2)^2 / \sigma_2^2} \times \frac{1}{\xi_0^2}$ is an F with df = (n_1, n_2). This provides a test for H_0. When the alternative is H: $\sigma_1 / \sigma_2 > \xi_0$, H_0 is to be rejected if for the given samples $F > F_{\alpha;\,n1,\,n2}$.

If the alternative is H: $\sigma_1 / \sigma_2 < \xi_0$, H0 is to be rejected if for the given samples $F < F_{1-\alpha;\,n1,\,n2}$, i.e., if $1/F > F_{\alpha;\,n2,\,n1}$.

Lastly, when the alternative is $H: \sigma_1 / \sigma_2 \neq \xi_0$, H_0 is to be rejected if the samples in hand give either $F < F_{1-\alpha/2;\, n1,\, n2}$, i.e., $1/F > F_{\alpha/2;\, n2,\, n1}$, or

$F > F_{\alpha/2;\, n2,\, n1}$.

The commonest form of the null hypothesis will be H_0: $\sigma_1 = \sigma_2$ for which $\xi_0 =$ 1, and here $F = \dfrac{\sum_{j=1}^{n_1}(x_{1j} - \mu_1)^2 / n_1}{\sum_{j=1}^{n_2}(x_{2j} - \mu_2)^2 / n_2}$.

For the purpose of setting confidence limits to σ_1 / σ_2, we see that

$$P\left[\frac{1}{F_{a/2;\, n2,\, n1}} \leq \frac{\sum_{j=1}^{n_1}(x_{1j} - \mu_1)^2 / n_1\sigma_1^2}{\sum_{j=1}^{n_2}(x_{2j} - \mu_2)^2 / n_2\sigma_2^2} \leq \frac{1}{F_{a/2;\, n1,\, n2}}\right] = 1 - \alpha$$

$$\text{i.e., } P\left[\frac{1}{F_{a/2;\, n2,\, n1}} \frac{\sum_{j=1}^{n_1}(x_{1j} - \mu_1)^2 / n_1}{\sum_{j=1}^{n_2}(x_{2j} - \mu_2)^2 / n_2} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{1}{F_{a/2;\, n1,\, n2}} \frac{\sum_{j=1}^{n_1}(x_{1j} - \mu_1)^2 / n_1}{\sum_{j=1}^{n_2}(x_{2j} - \mu_2)^2 / n_2}\right] = 1 - \alpha$$

The confidence limits to $\dfrac{\sigma_1^2}{\sigma_2^2}$ (with confidence coefficient $1 - \alpha$) will, therefore,

be $\dfrac{1}{F_{a/2;\, n1,\, n2}} \dfrac{\sum_{j=1}^{n_1}(x_{1j} - \mu_1)^2 / n_1}{\sum_{j=1}^{n_2}(x_{2j} - \mu_2)^2 / n_2}$ and $\dfrac{1}{F_{a/2;\, n2,\, n1}} \dfrac{\sum_{j=1}^{n_1}(x_{1j} - \mu_1)^2 / n_1}{\sum_{j=1}^{n_2}(x_{2j} - \mu_2)^2 / n_2}$.

The corresponding limits to σ_1 / σ_2 will naturally be the positive square roots of these quantities.

Case III: Means and standard deviations all unknown

We shall first consider methods of testing for the difference of the two means and of setting confidence limits to this difference.

For the sake of simplicity, we shall assume that the two unknown standard deviations are equal. Now if σ denotes the common standard deviation, then $\dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{1/2}}$ is a standard normal variable, while

$$\frac{(n_1-1)\,s_1'^2+(n_2-1)\,s_2'^2}{\sigma^2}=\frac{\sum_{j=1}^{n_1}(x_{1j}-\bar{x}_1)^2}{\sigma^2}+\frac{\sum_{j=1}^{n_2}(x_{2j}-\bar{x}_2)^2}{\sigma^2}$$

which is the sum of two independent χ^2 s, one with df = $(n_1 - 1)$ and the other with df = $(n_2 - 1)$, is itself a χ^2 with df = $(n_1 + n_2 - 2)$.

Hence, denoting by s'^2 the pooled variance of the two samples, so that $s'^2=\frac{(n_1-1)\,s_1'^2+(n_2-1)\,s_2'^2}{n_1+n_2-2}$

we have $\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{s'\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}=\frac{\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sigma\left(\frac{1}{n_1}+\frac{1}{n_2}\right)^{1/2}}}{\left\{\frac{(n_1-1)\,s_1'^2+(n_2-1)\,s_2'^2}{\sigma^2(n_1+n_2-2)}\right\}^{1/2}}$

a quantity of the form $\frac{\tau}{\sqrt{\chi^2/(n_1+n_2-2)}}$, where χ^2 is independent of τ and has df = $(n_1 + n_2 - 2)$. As such, the quantity of the left hand side of the above equation is distributed as t with df = $(n_1 + n_2 - 2)$.

A test for H_0: $\mu_1 - \mu_2 = \xi_0$ is then given by the statistic $t=\frac{(\bar{x}_1-\bar{x}_2)-\xi_0}{s'\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$

with df = $(n_1 + n_2 - 2)$. This is called *Fisher's t.*

For acceptance or rejection of the hypothesis H_0, one will have to compare the computed value of t' with the appropriated tabulated value, keeping in view the alternative hypothesis.

Following the usual procedure, it can be found that the confidence limits to $(\mu_1 - \mu_2)$ are $(\bar{x}_1-\bar{x}_2)-(t_{a/2,n_1+n_2-2})\,s'\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ and

$(\bar{x}_1-\bar{x}_2)+(t_{a/2,n_1+n_2-2})\,s'\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$. Obviously, in both cases we are using s'^2 as the estimate of the common variance σ^2.

Next, consider the problem of testing a hypothesis regarding the ratio σ_1/σ_2 or the problem of setting a confidence limits to the ratio. The difference between this problem and the corresponding problem mentioned in case II may be noted. Since μ_1 and μ_2 are unknown in the present case, they are replaced by their estimates $\bar{x}_1$ and $\bar{x}_2$, and we use the fact that

$$\frac{\sum_{j}^{n_1}(x_{1j} - \bar{x}_1)^2/(n_1 - 1)\sigma_1^2}{\sum_{j}^{n_2}(x_{2j} - \bar{x}_2)^2/(n_2 - 1)\sigma_2^2} = \frac{s_1'^2}{s_2'^2} x \frac{\sigma_2^2}{\sigma_1^2}$$ is an F with $(n_1 - 1, n_2 - 1)$ degrees of freedom.

For testing H_0: $\sigma_1 / \sigma_2 = \xi_0$, we will use the F statistic, but now $F = \frac{s_1'^2}{s_2'^2} x \frac{1}{\xi_0^2}$ with $(n_1 - 1, n_2 - 1)$ degrees of freedom. The confidence limits to σ_1^2 / σ_2^2 now will be $\frac{1}{F_{a/2;\ n1-1,\ n2-1}} \frac{s_1'^2}{s_2'^2}$ and $\frac{1}{F_{a/2;\ n2-1,\ n1-1}} \frac{s_1'^2}{s_2'^2}$. The corresponding limits to σ_1 / σ_2 will be the positive square roots of these quantities.

Check Your Progress 4

1) When is a significance test done using z? When is it done using t? What is different about test of differences between proportions?

.....

.....

.....

.....

.....

2) The scores of random sample of 8 students on a physics test are given below. Test to see if the sample mean is significantly different from 65 at the .05 level

60, 62, 67, 69, 70, 72, 75, 80

.....

.....

.....

.....

3) State the effect on the probability of a Type I and of a Type II error of:

a) the difference between population means

b) the variance

c) the sample size

d) the significance level

.....

.....

.....

.....

4) The following data are the lives in hours of two batches of electric bulbs. Test whether there is a significant difference between the batches in respect of average length of life.

Batch I: 1505, 1556, 1801, 1629, 1644, 1607, 1825, 1748.

Batch II: 1799, 1618, 1604, 1655, 1708, 1675, 1728.

.....

.....

.....

.....

20.9.3 Problems Relating to a Bivariate Normal Distribution

Suppose in a given population, the variables x and y are distributed in a bivariate normal form with means μ_x and μ_y, standard deviation σ_x and σ_y and correlation coefficient ρ. Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be random sample of size 'n' drawn from this distribution. We shall assume that all the parameters are unknown.

Test for correlation coefficient:

Here the sample correlation coefficient is $r = \dfrac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\left\{\sum_i (x_i - \bar{x})^2\right\}^{\frac{1}{2}} \left\{\sum_i (y_i - \bar{y})^2\right\}^{\frac{1}{2}}}$

where $\bar{x}$ and $\bar{y}$ are the sample means. When $\rho = 0$, the sampling distribution of r assumes a simple form $f(r) = \dfrac{1}{B\left(\frac{1}{2}, \frac{n-2}{2}\right)} (1 - r^2)^{(n-4)/2}$ and in that case $r\sqrt{n-2}\,/\sqrt{1-r^2}$ can be shown to be distributed as a t with df = (n - 2).

This fact provides us with a test for H_0: $\rho = 0$. As to the general hypothesis H_0: $\rho = \rho_0$, an exact test becomes difficult, because for $\rho \neq 0$ the sample correlation has a complicated sampling distribution. For moderately large n, there is an approximate test, which we have not discussed presently.

Problems regarding the difference between μ_x and μ_y

Information regarding the difference between the means, μ_x and μ_y, may be of some importance when x and y are variables measured in the same units.

To begin with, we note that if we take a new variable, z = x – y, then this z, being a linear function of normal variables, is itself normally distributed with mean $\mu_z = \mu_x - \mu_y$ and variance $\sigma_z^2 = \sigma_x^2 + \sigma_y^2 - 2\rho\sigma_x\sigma_y$.

It will follow, from what we have said in the section for univariate normal distribution, that if we put $z_i = x_i - y_i$, $\bar{z} = \sum_i z_i / n$ and $s_z'^2 = \frac{1}{n-1}\sum_i (z_i - \bar{z})^2$, then $\sqrt{n}(\bar{z} - \mu_z)/s_z'$ will be distributed as a t with df = (n – 1). This will provide us with a test for H_0: $\mu_x - \mu_y = \xi_0$, which is equivalent to H_0: $\mu_z = \xi_0$ and with confidence limits to the difference $\mu_z = \mu_x - \mu_y$. The statistic $\sqrt{n}(\bar{z} - \mu_z)/s_z'$ is often referred to as a *paired t.*

We may, instead, be interested in the ratio $\mu_x / \mu_y = \eta$ (say). In this case, we shall take $z = x - \eta y$, which again is normally distributed with mean $\mu_z = \mu_x - \eta \mu_y = 0$. Hence the statistic $t = \sqrt{n}\,\bar{z}/s_z'$ is distributed as a t (i.e., *paired t*) with df = (n – 1). This can be used for testing the hypothesis H_0: $\mu_x / \mu_y = \eta_0$ or for setting confidence limits to the ratio μ_x / μ_y.

Problems regarding the ratio σ_x / σ_y

When x and y are variable measured in identical units, one may also be interested in the ratio σ_x / σ_y. Let us denote the ratio by ξ.

If we consider the new variables $u = x + \xi y$ and $v = x - \xi y$,

then u and v are jointly normally distributed, like x and y, and

$$\text{cov}(u, v) = \sigma_x^2 - \xi^2 \sigma_y^2 = 0$$

Thus, u and v are uncorrelated normal variables.

In going to test for the hypothesis H_0: $\sigma_x / \sigma_y = \xi_0$, we shall therefore take the new variables $u = x + \xi y$ and $v = x - \xi y$ and shall instead test for the equivalent hypothesis H_0: $\rho_{uv} = 0$. This test will be given by the statistic $t = r_{uv}\sqrt{n-2}\,/\sqrt{1 - r_{uv}^2}$ with df = (n – 2),

r_{uv} being the sample correlation between u and v.

To have confidence limits for ξ, we utilize the fact that, with $u = x + \xi y$ and $v = x - \xi y$

$$P\left[\frac{|r_{uv}|\sqrt{n-2}}{\sqrt{1-r_{uv}^2}} \le t_{\alpha/2,\,n-2}\right] = 1 - \alpha$$

$$\text{or, } P\left[\frac{r_{uv}^2 (n-2)}{1 - r_{uv}^2} \le t_{\alpha/2,n-2}^2\right] = 1 - \alpha$$

By solving the equation $r_{uv}^2 (n-2) = t_{\alpha/2,n-2}^2 (1 - r_{uv}^2)$ or, say, $\psi(\xi) = 0$ for the unknown ration $\xi = \sigma_x / \sigma_y$, two roots will be obtained. In case, the roots, say ξ_1 and ξ_2, are real ($\xi_1 < \xi_2$), this will be the required confidence limits for ξ with confidence coefficient $(1 - \alpha)$.

Again, $\psi(\xi)$ may be either a convex or concave function. In the former case, we shall say $\xi_1 \leq \xi \leq \xi_2$, while in the latter we shall say $0 \leq \xi \leq \xi_1$ or $\xi_2 \leq \xi \leq \infty$.

But the roots may as well be imaginary, in which case we shall say that for the given sample $100(1-\alpha)$ % confidence limits do not exist.

Check Your Progress 5

1) The correlation coefficient between nasal length and stature for a group of 20 Indian adult males was found to be 0.203. Test whether there is any correlation between the characters in the population.

.....

.....

.....

.....

2) a) What proportion of a normal distribution is within one standard deviation of the mean? b) What proportion is more than 1.8 standard deviations from the mean? c) What proportion is between 1 and 1.5 standard deviations above the mean?

.....

.....

.....

.....

3) A test is normally distributed with a mean of 40 and a standard deviation of 7. a) What score would be needed to be in the 85^{th} percentile? b) What score would be needed to be in the 22^{nd} percentile?

.....

.....

.....

.....

4) Assume a normal distribution with a mean of 90 and a standard deviation of 7. What limits would include the middle 65% of the cases.

.....

.....

.....

.....

20.10 LET US SUM UP

In this unit, we have learnt how by using the theory of estimation and test of significance, an estimator can be analysed and how sample observations can be tested for any statistical claim. The unit shows the way of testing various real life problems through using statistical techniques. The basic concepts of hypothesis testing as well as of estimation theory are also made clear.

20.11 KEY WORDS

Alternative Hypothesis: Any hypothesis, which is complementary to the null hypothesis, is called an alternative hypothesis, usually denoted by H_1.

Confidence Interval and Confidence Limits: If we choose once for all some small value of α (5% or 1%) i.e., level of significance and then determine two constants say, c_1 and c_2 such that $P[c_1 < \theta < c_2] = 1 - \alpha$, where θ is the unknown parameter then, the quantities c_1 and c_2, so determined, are known as the 'confidence limits' and the interval $[c_1, c_2]$ within which the unknown value of the population parameter is expected to lie, is called the 'confidence interval' and $(1-\alpha)$ is called the 'confidence coefficient'.

Consistency: T_n is a consistent estimator of $\gamma(\theta)$ if for every $\varepsilon > 0$, $\eta > 0$, there exists a positive integer $n \geq m(\varepsilon, \eta)$ such that

$P[|T_n - \gamma(\theta)| < \varepsilon] \rightarrow 1$ as $n \rightarrow \infty \Rightarrow P[|T_n - \gamma(\theta)| < \varepsilon] > 1 - \eta;\ \forall\, n \geq m$ where m is some very large value of n.

Cramer-Rao Inequality: If t is an unbiased estimator of $\gamma(\theta)$, a function of parameter θ, then

$$\text{var}(t) \geq \frac{\left[\frac{\partial}{\partial\theta} L(x, \theta)\right]^2}{E\left[\frac{\partial}{\partial\theta} \log L\right]} = \frac{[\gamma'(\theta)]^2}{I(\theta)} \quad \text{and} \quad I(\theta) = E\left[\left\{\frac{\partial}{\partial\theta} \log L(x, \theta)\right\}^2\right],$$

where I (θ) is the information on θ, supplied by the sample. In other words, Cramer-Rao inequality provides a lower bound $\frac{[\gamma'(\theta)]^2}{I(\theta)}$ to the variance of an unbiased estimator of $\gamma(\theta)$.

Critical Region: A region (corresponding to a statistic t) in the sample space S, which amounts to rejection of H0, is termed as 'critical region'.

Critical Values or Significant Values: The value of the test statistic, which separates the critical (or rejection) region and the acceptance region is called the 'critical value' or 'significant value'.

Efficiency: T is the best estimator (or efficient estimator) of $\gamma(\theta)$ if it is consistent and normally distributed and if avar (T) $\leq$ avar (T') whatever the other consistent and asymptotically normal estimator T' may be.

Level of Significance: The probability that a random value of the statistic t belongs to the critical region is known as the 'level of significance'.

Minimum Variance Unbiased Estimator: T is Minimum Variance Unbiased Estimator of $\gamma(\theta)$ if $E_\theta(T)=\gamma(\theta)$ for all $\theta \in \Theta$ and $Var_\theta(T) \leq Var_\theta(T')$ for all $\theta \in \Theta$ where T' is any other unbiased estimator of $\gamma(\theta)$.

Most Efficient Estimator: If T_1 is the most efficient estimator with variance V_1 and T_2 is any other estimator with variance V_2, then the efficiency E of T_2 is defined as: $E = V_1 / V_2$. Obviously, E cannot exceed unity. If $T_1, T_2, \ldots, T_n$ are all estimators of $\gamma(\theta)$ and Var(T) is minimum, then the efficiency of E_i of T_i, (i=1,2,..., n) is defined as: $E_i=Var(T)/Var(T_i)$; obviously $E_i \leq 1$, (i=1,2, ..., n).

Most Powerful Test: The critical region W is the most powerful critical region of size α (and the corresponding test as most powerful test of level α) for testing H_0: $\theta = \theta_0$ against H_1: $\theta = \theta_1$ if $P(\mathbf{x} \in W \mid H_0) = \alpha$ and $P(\mathbf{x} \in W \mid H_1) \geq P(\mathbf{x} \in W_1 \mid H_1)$ for every other critical region W_1 satisfying the first condition.

Null Hypothesis: A definite hypothesis of no difference is called 'null hypothesis' and usually denoted by H_0.

One -Tailed and Two - Tailed Tests: A test of any statistical hypothesis where the alternative hypothesis is one-tailed (right-tailed or left-tailed) is called a 'one-tailed test'. For example, a test for testing the mean of a population H_0: $\mu = \mu_0$ against the alternative hypothesis: H_1: $\mu > \mu_0$ (right-tailed) or H_1: $\mu < \mu_0$ (left-tailed), is a 'one-tailed test'.

Let us consider a test of statistical hypothesis where the alternative hypothesis is two-tailed such as: H_0: $\mu = \mu_0$, against the alternative hypothesis H_1: $\mu \neq \mu_0$ ($\mu > \mu_0$ and $\mu < \mu_0$) is known as 'two-tailed test'.

Parameter Space: Let us consider a random variable x with p. d. f. $f(x, \theta)$. The p. d. f. of x can be written in the form $f(x, \theta)$, $\theta \in \Theta$. The set Θ, which is the set of all possible values of θ, is called the 'parameter space'.

Power of the Test: The power of the test can be defined as

Power = 1 – Probability of Type II error

= Probability of rejecting H_0 when H_1 is true.

Sufficiency: If $T_n = T(x_1, x_2, \ldots, x_n)$ is an estimator of a parameter θ, based on a sample $x_1, x_2, \ldots, x_n$ of size 'n' from the population with density $f(x, \theta)$ such that the conditional distribution of $x_1, x_2, \ldots, x_n$ given T, is independent of θ, then T is sufficient estimator for θ.

Type I and Type II Error: Type I Error: Rejecting the null hypothesis H_0 when it is true.

Type II Error: Accepting the null hypothesis H_0 when it is wrong, i.e., accept H_0 when H_1 is true.

Unbiasedness: A statistic $T_n = T(x_1, x_2, \ldots, x_n)$, is said to be an unbiased estimator of $\gamma(\theta)$ if $E(T_n) = \gamma(\theta)$, for all $\theta \in \Theta$

Uniformly Most Powerful Test: The critical region W is called uniformly most powerful critical region of size α (and the corresponding test as uniformly most powerful test of level α) for testing H_0: $\theta = \theta_0$ against H_1: $\theta \neq \theta_0$. if P ($\mathbf{x} \in$ W | H_0) = α and P ($\mathbf{x} \in$ W | H_1) $\geq$ P ($\mathbf{x} \in W_1$ | H_1) for all $\theta \neq \theta_0$, whatever the other critical region W_1 satisfying the first condition may be.

20.12 SOME USEFUL BOOKS

Goon A.M., Gupta M.K. & Dasgupta B. (1971), *Fundamental of Statistics,* Volumen I, The World Press Pvt. Ltd., Calcutta.

Freund, John E. (2001), *Mathematical Statistics*, Fifth Edition, Prentice-Hall of India Pvt. Ltd., New Delhi.

Das, N.G. (1996), *Statistical Methods*, M.Das & Co. (Calcutta).

20.13 ANSWER OR HINTS TO CHECK YOUR PROGRESS

Check Your Progress 1

1) See Section 20.2

2) See Section 20.3

3) Solution: Here we are given E (x_i) = μ, v (x_i) = 1 $\forall$ i = 1, 2, … , n.

Now E (x_i^2) = v (x_i) + E (x_i^2) = 1+ μ^2

$$E(t) = E\left[\frac{1}{n}\sum_{i=1}^{n} x_i^2\right] = \frac{1}{n}\sum_{i=1}^{n} E(x_i^2) = \frac{1}{n}\sum_{i=1}^{n}(1+\mu^2) = 1+\mu^2$$

4) Solution: We are given E (X_i) = μ, var (X_i^2) = σ^2, (say);

Cov (X_i, Y_i) = 0 ($i \neq j = 1, 2, \ldots, n$)

i) $E(t_1) = \frac{1}{5}\sum_{i=1}^{5} E(X_i) = (1/5).5\mu = \mu \Rightarrow t_1$ is an unbiased estimator of μ.

ii) E (t_3) = $\mu \Rightarrow \lambda\mu = 0 \Rightarrow \lambda = 0$

v (t_1) = [v(X_1) + v(X_2) + v(X_3) + v(X_4) + v(X_5)] / 25 = $\sigma^2/5$

v (t_2) = [v(X_1) + v(X_2)] / 4 + v(X_3) = $3\sigma^2/2$

v (t_3) = [4v(X_1) + v(X_2)] / 9 = $5\sigma^2/9$ (since $\lambda = 0$)

Since v (t_1) is the least, t_1 is the best estimator (in the sense of least variance) of μ.

5) Solution: $L(x, \theta) = \prod_{i=1}^{n} f(x_i, \theta) = \theta^n \prod_{i=1}^{n} (x_i^{\theta-1})$

$$= \theta^n \left(\prod_{i=1}^{n} x_i \right)^{\theta} \cdot \frac{1}{\left(\prod_{i=1}^{n} x_i \right)} = g(t_1, \theta).\ h(x_1, x_2, \ldots, x_n), \text{ (say).}$$

Hence by Factorization theorem, $t_1 = \prod_{i=1}^{n} x_i$, is sufficient estimator for θ.

Check Your Progress 2

1) See Section 20.4

2) See Section 20.5

3) See Section 20.5

Check Your Progress 3

1) 95% is wider.

2) You are 95% confident that the interval contains the parameter.

3) You use t when the standard error is estimated and z when it is known. One exception is for a confidence interval for a proportion when z is used even though the standard error is estimated.

4) See Section 20.5

5) See Section 20.6

6) See Section 20.6

7) See Sections 20.6 and 20.7

8) The determination made by the analyst may be said to be unbiased if the mean determination in the population, that could be obtained if she took an infinite number of readings, can be supposed to be 165^0. We have, therefore, to test the null hypothesis H_0: $\mu = 165$ against all the alternatives H: $\mu \neq 165$.

It will be assumed (a) that the population distribution of determinations is of the normal type and (b) that the sample observations are random and mutually independent.

Under these assumptions, a test H_0 is provided by the statistic $t = \sqrt{n}(\bar{x}-165)/s'$, which has df = (n-1).

For the given observations, n = 12, $\bar{x} = 163.992$,

$$s' = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2 = 3.039; \text{ so that } t = -1.149.$$

From t table $t_{0.025,\ 11} = 2.201$ and $t_{0.005,\ 11} = 3.106$. Since for the given sample |t| is smaller than both these tabulated values, H_0 is to be accepted at both the 1% and the 5% level of significance. In other words, we find no reason to suppose that the analyst's determination is not free from bias.

Check Your Progress 4

1) Read standardization of normal variate and answer

2) p=0.101

3) see Section 20.6

4) Here we have to test H_0: $\mu_1 = \mu_2$ against the alternative H: $\mu_1 \neq \mu_2$. The test for H_0 is then provided by the statistic $t = \dfrac{(\bar{x}_1 - \bar{x}_0)}{s\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ which is t with $df = n_1 + n_2 - 2$. And $t_{0.025,\ 13} = 2.160$, $t_{0.005,\ 13} = 3.012$.

Check Your Progress 5

1) The null hypothesis here is H_0: $\rho = 0$, to be tested against all alternatives. As we have seen, under certain assumptions, which may be considered legitimate here, the test is given by $t = r\sqrt{n-2}\ /\sqrt{1-r^2}$ which has df = n- 2.

   Here t = 0.880 and tabulated values are $t_{0.025,\ 18} = 2.101$ and $t_{0.005,\ 18} = 2.878$. The observed value is, therefore, insignificant at both the levels; i.e. the population correlation coefficient may be supposed to be zero.

2) (a) 50%; (b) 8.08%; (c) 44.35

3) (a) 47.25; (b) 34.59

4) 83.46 and 96.4

20.14 EXERCISES

1) If T is an unbiased estimator of θ, show that T^2 is a biased estimator for θ^2.

   [**Hint:** Find var (T) = E (T^2) – θ^2. Since E (T^2) $\neq \theta^2$, T^2 is a biased estimator for θ^2.]

2) X_1, X_2, and X_3 is a random sample of size 3 from a population with mean value μ and variance σ^2, T_1, T_2, T_3 are the estimators used to estimate mean value μ, where

   $T_1 = X_1 + X_2 - X_3$; $T_2 = 2X_1 - 4X_2 + 3X_3$; and $T_3 = (\lambda X_1 + X_2 + X_3) / 3$.

   i) Are T_1 and T_2 unbiased estimators?

   ii) Find the value of λ such that T_3 is unbiased estimator of μ.

iii) With this value of λ is T_3 a consistent estimator?

iv) Which is the best estimator?

[**Hint:** Follow Check Your Progress 2]

3) Let $X_1, X_2, \ldots, X_n$ be a random sample from a Cauchy population:

$$f(x, \theta) = \frac{1}{\pi} \cdot \frac{1}{1 + (x - \theta)^2}; \; -\infty < x < \infty, \; -\infty < \theta < \infty.$$

Examine if there exist sufficient estimate for θ.

[**Hint:** $L(x, \theta) = \prod_{i=1}^{n} f(x_i, \theta) = \frac{1}{\pi^n} \cdot \prod_{i=1}^{n} \left[\frac{1}{1 + (x - \theta)^2} \right] \neq g(t_1, \theta) \cdot h(x_1, x_2, \ldots, x_n)$.

Hence by Factorization Theorem, there is no single statistic, which alone is sufficient for θ.

However, $L(x, \theta) = k_1(X_1, X_2, \ldots, X_n, \theta) \cdot k_2(X_1, X_2, \ldots, X_n) \Rightarrow$ The whole set $(X_1, X_2, \ldots, X_n)$ is jointly sufficient θ.

4) The weights at birth for 15 babies born are given below. Each figure is correct to the nearest tenth of a pound.

6.2, 5.7, 8.1, 6.7, 4.8, 5.0, 7.1, 6.8, 5.8, 6.9, 7.6, 7.9, 7.5, 7.8, 8.5.

Give two limits between which the mean weight at birth for all such babies is likely to lie.

[**Hint:** Let us denote by x the variable: weight at birth per baby. Our problem here is then to find, on the basis of the given sample of 15 babies, confidence limits for the population mean of x. We shall assume (a) that in population x is normally distributed (with a mean μ and standard deviation σ, both of which are unknown) and (b) that the given observations form a random sample from the distribution.

Under these assumptions, the $100(1-\alpha)$ % confidence limits to μ will be $(\bar{x} - t_{\alpha/2, n-1}) \frac{s'}{\sqrt{n}}$ and $(\bar{x} + t_{\alpha/2, n-1}) \frac{s'}{\sqrt{n}}$]

5) It has been said by some educationists that mathematical ability varies widely. To examine this suggestion, 15 students of class IX are given a mathematical aptitude test carrying 100 marks. The score of the students on the test are shown below:

73, 16, 84, 20, 13, 63, 53, 68, 19, 40, 52, 91, 12, 25, 17.

Examine the more specific suggestion that the standard deviation of score per student is higher than 20.

[**Hint:** We shall assume that the random variable x, viz. score per student on the test is distributed normally for students of class IX with some

mean μ and standard deviation σ, both unknown. Further the given set of observations will be regarded as the observed values for a random sample of size n =15 from this distribution. So the problem can be stated as H_0: $\sigma = 20$ against the alternative H: $\sigma > 20$. Under the usual assumption the test is given by $\chi^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sigma_0^2}$.

The value of $\chi^2_{0.05,14} = 23.685$.]

6) Two experimenters, A and B, take repeated measurements on the length of a copper wire. On the basis of the data obtained by them, which are given below, test whether B's measurements are more accurate than A's. (It may be supposed that the readings taken by both are unbiased.)

A's measurements: 12.47, 12.44, 11.90, 12.13, 12.77, 11.86, 11.96, 12.25, 12.78, 12.29.

B's measurements: 12.06, 12.34, 12.23, 12.46, 12.39, 12.46, 11.98, 12.22.

[**Hint:** The problem can be stated as H_0: $\sigma_1 = \sigma_2$, against the alternative H: $\sigma_1 > \sigma_2$. Under the usual assumption the test is given by $F = \frac{s_1'^2}{s_2'^2}$ with $(n_1 - 1, n_2 - 1)$ degrees of freedom. The tabulated values are $F_{0.05;\,9,\,7} = 3.68$; $F_{0.01;\,9,\,7} = 6.72$.]

7) In a certain industrial experiment, a job was performed by 15 workmen according to a particular method (say, Method I) and by 15 other workmen according to a second method (Method II). The time (in minutes) taken by each workman to complete the job is shown below:

Method I: 55, 53, 57, 55, 52, 51, 54, 54, 53, 56, 50, 54, 52, 56, 51.

Method II: 54, 53, 56, 60, 58, 55, 56, 54, 58, 57, 55, 54, 59, 52, 54.

Test if there was difference of time taken between these two method.

[**Hint:** The problem can be stated as H_0: $\mu_1 = \mu_2$, against the alternative H: $\mu_1 < \mu_2$. Under the usual assumption the test is given by $t = \frac{(\bar{x}_1 - \bar{x}_2)}{s'\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

with $(n_1 + n_2 - 2)$ degrees of freedom.

The tabulated values are $-t_{0.05,\,28} = -1.701$; $-t_{0.01,\,28} = -2.467$.]

8) The marks in mathematics (x) and those in English (y) are given below for a group of 14 students:

x: 52, 31, 83, 75, 95, 78, 85, 23, 69, 32, 48, 9, 84, 54.

y: 69, 42, 50, 31, 43, 38, 59, 44, 51, 61, 33, 43, 27, 46.

These will be used to examine the claim of some educationists that mathematical ability and proficiency in English are inversely related (i.e. negatively correlated).

[**Hint:** The problem can be stated as H_0: $\rho=0$, against the alternative H: $\rho < 0$. Under certain assumptions, the test is given by $t = r\sqrt{n-2}/\sqrt{1-r^2}$ which has df = n-2. Here t = -0.495 and tabulated value is $t_{0.05, 12} = -1.782$, H_0 is to be accepted at the 5% level of significance. In other words, we find no evidence in the data to support the claim that x and y are negatively correlated.

9) The weights (in lb.) of 10 boys before they are subjected to a change of diet and after a lapse of six months are recorded below:

Before: 109, 112, 98, 114, 102, 97, 88, 101, 89, and 91.

After: 115, 120, 99, 117, 105, 98, 91, 99, 93, and 89.

Test whether there has been any significant gain in weight as a result of the change in diet.

[**Hint:** The problem can be stated as H_0: $\mu_x = \mu_y$. against the alternative H: $\mu_x > \mu_y$. Under certain assumptions, the test is given by $t = \sqrt{n}\ \bar{z}/s_z'$ with df = n – 1, where z =x - y.

Here, $\bar{z} = 2.5$, $s_z' = 3.171$, t = 2.493 and tabulated value is $t_{0.05, 9} = 1.833$ and $t_{0.01, 9} = 2.821$. The observed value is thus significant at the 5% but insignificant at the 1% level of significance. If we choose the 5% level, then the null hypothesis should be rejected and we should say that the change of diet results in a gain in average weight.

10) What is meant by a level of confidence, or confidence level?

NOTES

MPDD/IGNOU/P.O.7K/November 2018 (Reprint)

ISBN : 978-93-86375-17-9