CHAPTER TWELVE

Sampling and Large Sample Tests

12·1. Sampling—Introduction. Before giving the notion of sampling we will first define *population*. In a statistical investigation the interest usually 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. This 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 implications, time factor, etc., and we take the help of *sampling*.

A finite subset of statistical individuals in a population is called a *sample* and the number of individuals 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 only are 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 and every 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 assess the quality of sugar, wheat or any other commodity by taking a handful of it from the bag and then decide to purchase it or not. A housewife normally tests the cooked products to find if they are properly cooked and contain the proper quantity of salt.

12·2. Types of Sampling. Some of the commonly known and frequently used types of sampling are :

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

Below we will precisely explain these terms, without entering into detailed discussion.

12·2·1. Purposive Sampling. Purposive sampling is one in which the sample units are selected with definite purpose in view. For example, if we want to give the picture that the standard of living has increased in the city of New Delhi, we may take individuals in the sample from rich and posh localities like Defence Colony, South Extension, Golf Links, Jor Bagh, Chanakyapuri, Greater Kailash etc. 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.

12·2·2 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 a *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 number tables* or by throwing of a dice, draw of a lottery, etc.

The simplest method, which is normally used, is the *lottery system* which is illustrated below by means of an example.

Suppose we want to select 'r' candidates out of n. We assign the numbers one to n, one number to each candidate and write these 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 then 'r' slips are drawn one by one. The 'r' candidates corresponding to the numbers on the slips drawn, will constitute the random sample.

Remark. *Tippet's Random Numbers.* L.H.C. Tippet's random numbers tables consist of 10400 four-digited 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.

12·2·3. 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.

Remark. It may be pointed out 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 $[(a-1)(a+b-1)] = p_2 \neq p_1$, the 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 that sampling is simple, it must be done with replacement, if population is finite. However, in case of infinite population no replacement is necessary.

12·2·4. 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 homogeous within itself. Then units are sampled at random from each of these stratum, the sample size in each stratum varies according 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 sample is by far the best and can safely be considered as representative of the population from which it has been drawn.

12·3. 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 measures computed from the sample observations alone, *e.g.*, mean ($\bar{x}$), variance (s^2), etc., have been termed by Professor R.A. Fisher as *statistics*.

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. The determination or the characterisaton of the 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.

Remarks 1. 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.

2. Unbiased Estimate. A statistic $t = t(x_1, x_2, \ldots, x_n)$, a function of the sample values $x_1, x_2, \ldots, x_n$ is an unbiased estimate of population parameter θ, if $E(t) = \theta$. In other words, if

$$E(\textit{Statistic}) = \text{Parameter}, \qquad \ldots(12\cdot1)$$

then statistic is said to be an unbiased estimate of the parameter.

12·3·1. 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 :

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

For each of these k samples we can compute some statistic $t = t(x_1, x_2, \ldots, x_n)$, in particular the mean $\bar{x}$, the variance s^2, etc., as given below :

Sample Number	*Statistics*		
	t	$\bar{x}$	s^2
1	t_1	$\bar{x}_1$	s_1^2
2	t_2	$\bar{x}_2$	s_2^2
3	t_3	$\bar{x}_3$	s_3^2
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮
k	t_k	$\bar{x}_k$	s_k^2

The set of the values of the statistic so obtained, one for each sample, constitutes what is called the *sampling distribution* of the statistic. For example, the values $t_1, t_2, t_3, \ldots, t_k$ determine the sampling distribution of the statistic t. In other words, statistic t may be regarded as a random variable which can take the values $t_1, t_2, t_3, \ldots, t_k$ and we can compute the various statistical constants like mean, variance, skewness, kurtosis etc., for its distribution. For example, the mean and variance of the sampling distribution of the statistic t are given by :

$$\bar{t} = \frac{1}{k}(t_1 + t_2 + \ldots + t_k) = \frac{1}{k}\sum_{i=1}^{k} t_i$$

$$\text{Var}(t) = \frac{1}{k}\left[(t_1 - \bar{t})^2 + (t_2 - \bar{t})^2 + \ldots + (t_k - \bar{t})^2\right]$$

$$= \frac{1}{k}\sum_{i=1}^{k}(t_i - \bar{t})^2$$

12·3·2. 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 the population variance, and P 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).

S.No.	*Statistic*	*Standard Error*
1.	Sample mean : $\bar{x}$	$\sigma/\sqrt{n}$
2.	Observed sample proportion 'p'	$\sqrt{PQ/n}$
3.	Sample s.d. : s	$\sqrt{\sigma^2/2n}$
4.	Sample variance : s^2	$\sigma^2\sqrt{2/n}$
5.	Sample quartiles	$1{\cdot}36263\ \sigma/\sqrt{n}$
6.	Sample median	$1{\cdot}25331\ \sigma/\sqrt{n}$

7.	Sample correlation coefficient (r)	$(1-\rho^2)/\sqrt{n}$, ρ being the population correlation coefficient
8.	Sample moment μ_3	$\sigma^3\sqrt{96/n}$
9.	Sample moment μ_4	$\sigma^4\sqrt{96/n}$
10.	Sample coefficient of variation (v)	$\frac{v}{\sqrt{2n}}\sqrt{1+\frac{2v^3}{10^4}} \simeq \frac{v}{\sqrt{2n}}$
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 two sample proportions $(p_1 - p_2)$	$\sqrt{\frac{P_1Q_1}{n_1}+\frac{P_2Q_2}{n_2}}$

Remark on the Utility of Standard Error. S.E. plays a very important role in the large sample theory and forms the basis of the testing of hypothesis. If t is any statistic, then for large samples

$$Z = \frac{t - E(t)}{\sqrt{V(t)}} \sim N(0, 1) \qquad (c.f.\ \S\ 12{\cdot}9)$$

$$\Rightarrow \qquad Z = \frac{t - E(t)}{\text{S.E. } (t)} \sim N(0, 1), \text{ for large samples.}$$

Thus, if the discrepancy between the observed and the expected (hypothetical) value of a statistic is greater than z_α (c.f. § 12·7·2) times its S.E., the null hypothesis is rejected at α level of significance. Similarly, if

$$|\, t - E(t)\,| \leq z_\alpha \times \text{S.E. } (t),$$

the deviation is not regarded significant at 5% level of significance. In other words, the deviation, $t - E(t)$, could have arisen due to fluctuations of sampling and the data do not provide us any evidence against the null hypothesis which may, therefore, be accepted at α level of significance. [For details see § 12·7·3]

(*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.

$$\text{S.E. } (p) = \sqrt{PQ/n} \qquad [c.f.\ (4b)\ \S\ 12{\cdot}9{\cdot}1]$$

and

$$\text{S.E. } (\bar{x}) = \sigma/\sqrt{n} \qquad [c.f.\ \S\ 12{\cdot}2]$$

In other words, the standard errors of p and $\bar{x}$ vary inversely as the square root of the sample size. Thus in order to double the precision, which amounts to reducing the standard error to one half, the sample size has to be increased four times.

(*ii*) S.E. enables us to determine the probable limits within which the population parameter may be expected to lie. For example, the probable limits for population proportion P are given by

$$p \pm 3\sqrt{pq/n} \qquad (c.f.\ \text{Remark § 12·9·1})$$

Remark. S.E. of a statistic may be reduced by increasing the sample size but this results in corresponding increase in cost, labour and time, etc.

12·4. Tests of Significance. A very important aspect of the sampling theory is the study of the *tests of significance*, which enable 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

(*i*) 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 (*c.f.* Chapter 7), t, F (Chapter 14), Chi-square (Chapter 13), can be approximated very closely by a normal probability curve, we use the *Normal Test of Significance* (*c.f.* § 12·9) 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.*

12·5. Null Hypothesis. The technique of randomisation 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 statement about the population parameter. Such a hypothesis, which is usually a hypothesis of no difference, is called *null hypothesis* and is usually denoted by H_0. According to *Prof. R.A. Fisher, null hypothesis is the hypothesis which is tested for possible rejection under the assumption that it is true.*

For example, in case of a 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 (*c.f.* § 12·7). 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 (*c.f.* § 12·7) and if the deviation is not significant, null hypothesis may be retained at that level.

12·5·1. 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_0, (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 a *two tailed alternative* and the alternatives in (*ii*) and (*iii*) are known as *right tailed* and *left-tailed alternatives* 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 [c.f. § 12·7·1].

12·6. Errors in Sampling. 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 a 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.

$$\left.\begin{aligned} P\{\text{Reject } H_0 \text{ when it is true}\} &= P\{\text{Reject } H_0 \mid H_0\} = \alpha \\ \text{and } P\{\text{Accept } H_0 \text{ when it is wrong}\} &= P\{\text{Accept } H_0 \mid H_1\} = \beta \end{aligned}\right\} \quad \ldots(12\cdot2)$$

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.

$$\left.\begin{aligned} \text{Thus} \quad P\{\text{Reject a lot when it is good}\} &= \alpha \\ \text{and} \quad P\{\text{Accept a lot when it is bad}\} &= \beta \end{aligned}\right\} \quad \ldots(12\cdot2a)$$

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

12·7. Critical Region and Level of Significance. A region (corresponding to a statistic t) in the sample space S which amounts to rejection of H_0 is termed as *critical region* or *region of rejection*. 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

$$P(t \in \omega \mid H_0) = \alpha, \; P(t \in \bar{\omega} \mid H_1) = \beta \qquad (12\cdot2b)$$

where $\bar{\omega}$, the complementary set of ω, is called the *acceptance region*.

We have $\omega \cup \bar{\omega} = S$ and $\omega \cap \bar{\omega} = \phi$

The probability '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.

12·7·1. 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 ($H_1 : \mu_1 > \mu_0$), the critical region lies entirely in the right tail of the sampling distribution of $\bar{x}$, while for the left tail test ($H_1 : \mu < \mu_0$), the critical region is entirely in the left tail of the distribution.

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 two-tailed test and if alternative hypothesis is one-tailed, we apply one tailed test.

For example, suppose that there are two population 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). If we want to test if the bulbs differ significantly, then our null hypothesis is $H_0 : \mu_1 = \mu_2$ and 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

$$H_0 : \mu_1 = \mu_2 \quad \text{and} \quad H_1 : \mu_1 < \mu_2,$$

thus giving us a left-tail test. Similarly, for testing if the product of new process is inferior to that of standard process, then we have :

$$H_0 : \mu_1 = \mu_2 \text{ and } H_1 : \mu_1 > \mu_2,$$

thus giving us a right-tail test. Thus, the decision about applying a two-tail test or a single-tail (right or left) test will depend on the problem under study.

12·7·2. Critical Values or Significant Values. The value of 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 standardised variable corresponding to the statistic *t viz.* :

$$Z = \frac{t - E(t)}{S.E.(t)} \sim N(0, 1), \qquad \ldots(*)$$

asymptotically as $n \to \infty$. The value of Z given by (*) under the null hypothesis is known as *test statistic*. The critical value of the 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

$$P(|Z| > z_\alpha) = \alpha \qquad \ldots(12·2c)$$

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 a symmetrical curve, from (12·2*c*), we get

$$P(Z > z_\alpha) + P(Z < -z_\alpha) = \alpha \qquad \text{[By symmetry]}$$

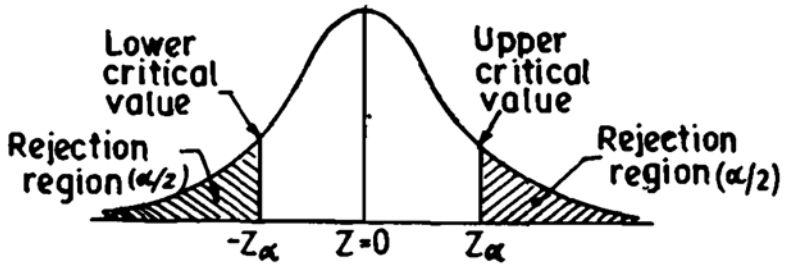
$$\Rightarrow \quad P(Z > z_\alpha) + P(Z > z_\alpha) = \alpha$$

$$\Rightarrow \quad 2P(Z > z_\alpha) = \alpha$$

$$\Rightarrow \quad P(Z > z_\alpha) = \frac{\alpha}{2}$$

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 $\alpha/2$, as shown in the following diagram.

TWO-TAILED TEST
(*Level of Significance 'α'*)



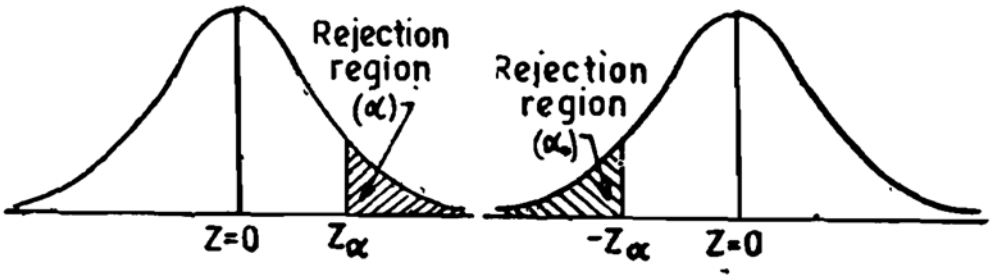
In case of single-tail alternative, the critical value z_α is determined so that total area to the right of it (for right-tailed test) is α and for left-tailed test the total area to the left of $-z_\alpha$ is α (See diagrams below), *i.e.*,

For Right-tail Test : $P(Z > z_\alpha) = \alpha$ ...(12·2*d*)

For Left-tail Test : $P(Z < -z_\alpha) = \alpha$ ...(12·2*e*)

RIGHT-TAILED TEST
(*Level of Significance 'α'*)

LEFT-TAILED TEST
(*Level of Significance 'α'*)



Thus *the significant* or *critical value of Z for a single-tailed test (left or right) at level of significance 'α' is same as the critical value of Z for a two-tailed test at level of significance '2α'.*

We give on page 12·10, the critical values of Z at commonly used levels of significance for both two-tailed and single-tailed tests. These values have been obtained from equations (12·2*c*), (12·2*d*) and (12·2*e*), on using the Normal Probability Tables as explained in § 12·8.

CRITICAL VALUES (z_α) OF Z

Critical Values (z_α)	*Level of significance* (α) 1%	5%	10%
Two-tailed test	$\lvert Z_\alpha \rvert = 2{\cdot}58$	$\lvert Z_\alpha \rvert = 1{\cdot}96$	$\lvert Z_\alpha \rvert = 1{\cdot}645$
Right-tailed test	$Z_\alpha = 2{\cdot}33$	$Z_\alpha = 1{\cdot}645$	$Z_\alpha = 1{\cdot}28$
Left-tailed test	$Z_\alpha = -2{\cdot}33$	$Z_\alpha = -1{\cdot}645$	$Z_\alpha = -1{\cdot}28$

Remark. If n is small, then the sampling distribution of the test statistic Z will not be normal and in that case we can't use the above significant values, which have been obtained from normal probability curves. In this case, *viz.*, n small, (usually less than 30), we use the significant values based on the exact sampling distribution of the statistic Z, [defined in (*), § 12·7·2], which turns out to be t, F, or χ^2 [see Chapters 13, 14]. These significant values have been tabulated for different values of n and α and are given in the Appendix at the end of the book.

12·7·3. 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 (see § 12·5, page 12·6).

2. *Alternative Hypothesis.* Set up the Alternative Hypothesis H_1. Ths 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

$$Z = \frac{t - E(t)}{S.E.(t)}$$

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 $\lvert Z \rvert < z_\alpha$, *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 $\lvert Z \rvert > z_\alpha$, *i.e.*, if the computed value of test statistic 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 $(1 - \alpha)$.

12·8. Test of Significance for Large Samples. In this section we

will discuss the tests of significance when samples are large. We have seen that for large values of n, the number of trials, almost all the distributions, *e.g.*, binomial, Poisson, negative binomial, etc., are very closely approximated by normal distribution. Thus in this case we apply the *normal test*, which is based upon the following fundamental property (*area property*) of the normal probability curve.

$$\text{If } X \sim N(\mu, \sigma^2), \text{ then } Z = \frac{X - \mu}{\sigma} = \frac{X - E(X)}{\sqrt{V(X)}} \sim N(0, 1)$$

Thus from the normal probability tables, we have

$$P(-3 \le Z \le 3) = 0{\cdot}9973, \quad i.e., \quad P(|Z| \le 3) = 0{\cdot}9973$$

$$\Rightarrow \quad P(|Z| > 3) = 1 - P(|Z| \le 3) = 0{\cdot}0027 \qquad \ldots(12{\cdot}3)$$

i.e., in all probability we should expect a standard normal variate to lie between ± 3.

Also from the normal probability tables, we get

$$P(-1{\cdot}96 \le Z \le 1{\cdot}96) = 0{\cdot}95 \quad i.e., \quad P(|Z| \le 1{\cdot}96) = 0{\cdot}95$$

$$\Rightarrow \quad P(|Z| > 1{\cdot}96) = 1 - 0{\cdot}95 = 0{\cdot}05 \qquad \ldots(12{\cdot}3a)$$

and $$P(|Z| \le 2{\cdot}58) = 0{\cdot}99$$

$$\Rightarrow \quad P(|Z| > 2{\cdot}58) = 0{\cdot}01 \qquad \ldots(12{\cdot}3b)$$

Thus the significant values of Z at 5% and 1% level of significance for a two tailed test are 1·96 and 2·58 respectively.

Thus the steps to be used in the normal test are as follows :

(*i*) Compute the test statistic Z under H_0.

(*ii*) If $|Z| > 3$, H_0 is always rejected.

(*iii*) If $|Z| \le 3$, we test its signficance at certain level of significance, usually at 5% and sometimes at 1% level of significance. Thus, for a two-tailed test if $|Z| > 1{\cdot}96$, H_0 is rejected at 5% level of significance.

Similarly if $|Z| > 2{\cdot}58$, H_0 is contradicted at 1% level of significance and if $|Z| \le 2{\cdot}58$, H_0 may be accepted at 1% level of significance.

From the normal probability tables, we have :

$$\begin{aligned} P(Z > 1{\cdot}645) &= 0{\cdot}5 - P(0 \le Z \le 1{\cdot}645) \\ &= 0{\cdot}5 - 0{\cdot}45 \\ &= 0{\cdot}05 \\ P(Z > 2{\cdot}33) &= 0{\cdot}5 - P(0 \le Z \le 2{\cdot}33) \\ &= 0{\cdot}5 - 0{\cdot}49 \\ &= 0{\cdot}01 \end{aligned}$$

Hence for a single-tail test (Right-tail or Left-tail) we compare the computed value of $|Z|$ with 1·645 (at 5% level) and 2·33 (at 1% level) and accept or reject H_0 accordingly.

Important Remark. In the theoretical discussion that follows in the next sections, the samples under consideration are supposed to be large. For practical purposes, sample may be regarded as large if $n > 30$.

12·9. Sampling of Attributes. Here we shall consider sampling from a population which is divided into two mutually exclusive and collectively

exhaustive classes-one class possessing a particular attribute, say A, and the other class not possessing that attribute, and then note down the number of persons in the sample of size n, possessing that attribute. The presence of an attribute in sampled unit may be termed as success and its absence as failure. In this case a sample of n observations is identified with that of a series of n independent Bernoulli trials with constant probability P of success for each trial. Then the probability of x successes in n trials, as given by the binomial probability distribution is

$$p(x) = {}^nC_x P^x Q^{n-x} ; x = 0, 1, 2, \ldots, n.$$

12·9·1. Test for Single Proportion. If X is the number of successes in n independent trials with constant probability P of success for each trial (*c.f.* § 7·2·1)

$$E(X) = nP \quad \text{and} \quad V(X) = nPQ,$$

where $Q = 1 - P$, is the probability of failure.

It has been proved that for large n, the binomial distribution tends to normal distribution. Hence for large n, $X \sim N(nP, nPQ)$ *i.e.*,

$$Z = \frac{X - E(X)}{\sqrt{V(X)}} = \frac{X - nP}{\sqrt{nPQ}} \sim N(0, 1) \qquad \ldots(12\cdot4)$$

and we can apply the normal test.

Remarks 1. In a sample of size n, let X be the number of persons possessing the given attribute. Then

Observed proportion of successes $= X/n = p$, (say).

$$\therefore \qquad E(p) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n}nP = P$$

$$\Rightarrow \qquad E(p) = P \qquad \ldots(12\cdot4a)$$

Thus the sample proportion 'p' gives an unbiased estimate of the population proportion P.

$$\text{Also} \qquad V(p) = V\left(\frac{X}{n}\right) = \frac{1}{n^2}V(X) = \frac{1}{n^2}nPQ = \frac{PQ}{n}$$

$$\therefore \qquad \text{S.E.}(p) = \sqrt{PQ/n} \qquad \ldots(12\cdot4b)$$

Since X and consequently X/n is asymptotically normal for large n, the normal test for the proportion of successes becomes

$$Z = \frac{p - E(p)}{\text{S.E. }(p)} = \frac{p - P}{\sqrt{PQ/n}} \sim N(0, 1) \qquad \ldots(12\cdot4c)$$

2. If we have sampling from a finite population of size N, then

$$S.E.(p) = \sqrt{\left(\frac{N - n}{N - 1}\right) \cdot \frac{PQ}{n}} \qquad \ldots(12\cdot4d)$$

3. Since the probable limits for a normal variate X are $E(X) \pm 3\sqrt{V(X)}$, the probable limits for the observed proportion of successes are :

$$E(p) \pm 3 \text{ S.E. } (p), \textit{ i.e.,} \quad P \pm 3\sqrt{PQ/n}.$$

If P is not known then taking p (the sample proportion) as an estimate of P, the probable limits for the proportion in the population are :

$$p \pm 3\sqrt{pq/n} \qquad \ldots(12\text{·}4e)$$

However, the limits for P at level of significance α are given by :

$$p \pm z_\alpha \sqrt{pq/n}, \qquad \ldots(12\text{·}4f)$$

where z_α is the significant value of Z at level of significance α.

In particular 95% confidence limits for P are given by :

$$p \pm 1\text{·}96\sqrt{pq/n}, \qquad \ldots(12\text{·}4g)$$

and 99% confidence limits for P are given by

$$p \pm 2\text{·}58\sqrt{pq/n} \qquad \ldots(12\text{·}4h)$$

Example 12·1. *A dice is thrown 9,000 times and a throw of 3 or 4 is observed 3,240 times. Show that the dice cannot be regarded as an unbiased one and find the limts between which the probability of a throw of 3 or 4 lies.*

Solution. If the coming of 3 or 4 is called a success, then in usual notations we are given

$$n = 9{,}000;\ X = \text{Number of successes} = 3{,}240$$

Under the null hypothesis (H_0) *that the dice is an unbiased one,* we get

P = Probability of success = Probability of getting a 3 or 4 $= \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$

Alternative hypothesis, $H_1 : p \neq \frac{1}{3}$, (*i.e.,* dice is biased).

We have $Z = \dfrac{X - nP}{\sqrt{nQP}} \sim N(0, 1)$, since n is large.

Now $$Z = \frac{3240 - 9000 \times 1/3}{\sqrt{9000 \times (1/3) \times (2/3)}} = \frac{240}{\sqrt{2000}} = \frac{240}{44\text{·}73} = 5\text{·}36$$

Since $|Z| > 3$, H_0 is rejected and we conclude that the dice is almost certainly biased.

Since dice is not unbiased, $P \neq \frac{1}{3}$. The probable limits for 'P' are given by :

$$\hat{P} \pm 3\sqrt{\hat{P}\hat{Q}/n} = p \pm 3\sqrt{pq/n},$$

where $\hat{P} = p = \dfrac{3240}{9000} = 0\text{·}36$ and $\hat{Q} = q = 1 - p = 0\text{·}64$.

Hence the probable limits for the population proportion of successes may be taken as

$$\hat{P} \pm 3\sqrt{\hat{P}\hat{Q}/n} = 0\text{·}36 \pm 3\sqrt{\frac{0\text{·}36 \times 0\text{·}64}{9000}} = 0\text{·}36 \pm 3 \times \frac{0\text{·}6 \times 0\text{·}8}{30\sqrt{10}}$$

$$= 0\text{·}360 \pm 0\text{·}015 = 0\text{·}345 \text{ and } 0\text{·}375.$$

Hence the probability of getting 3 or 4 almost certainly lies between 0·345 and 0·375.

Example 12·2. *A random sample of 500 pineapples was taken from a large consignment and 65 were found to be bad. Show that the 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 17·5.

Solution. Here we are given $n = 500$

X = Number of bad pineapples in the sample = 65

p = Proportion of bad pineapples in the sample = $\frac{65}{500} = 0{\cdot}13$

$\therefore \quad q = 1 - p = 0{\cdot}87$

Since P, the proportion of bad pineapples in the consignment is not known, we may take (as in the last example)

$$\hat{P} = p = 0{\cdot}13, \quad \hat{Q} = q = 0{\cdot}87$$

S.E. of proportion = $\sqrt{\hat{P}\hat{Q}/n} = \sqrt{0{\cdot}13 \times 0{\cdot}87/500} = 0{\cdot}015$

Thus, the limits for the proportion of bad pineapples in the consignment are :

$$\hat{P} \pm 3\sqrt{\hat{P}\hat{Q}/n} = 0{\cdot}130 \pm 3 \times 0{\cdot}015 = 0{\cdot}130 \pm 0{\cdot}045 = (0{\cdot}085, 0{\cdot}175)$$

Hence the percentage of bad pineapples in the consignment lies almost certainly between 8·5 and 17·5.

Example 12·3. *A random sample of 500 apples was taken from a large consignment and 60 were found to be bad. Obtain the 98% confidence limits for the percentage number of bad apples in the consignment.*

$$\left[\int_0^{2{\cdot}33} \phi(t)\,dt = 0{\cdot}49 \text{ nearly}\right]$$

Solution. We have :

p = Proportion of bad apples in the sample = $\frac{60}{500} = 0{\cdot}12$

Since the significant value of Z at 98% confidence coefficient (level of significance 2%) is given to be 2·33, 98% confidence limits for population proportion are :

$$p \pm 2{\cdot}33\sqrt{pq/n} = 0{\cdot}12 \pm 2{\cdot}33\sqrt{0{\cdot}12 \times 0{\cdot}88/500}$$
$$= 0{\cdot}12 \pm 2{\cdot}33 \times \sqrt{0{\cdot}0002112} = 0{\cdot}12 \pm 2{\cdot}33 \times 0{\cdot}01453$$
$$= 0{\cdot}12000 \pm 0{\cdot}03385 = (0{\cdot}08615, 0{\cdot}15385)$$

Hence 98% confidence limits for percentage of bad apples in the consignment are (8·61, 15·38).

Example 12·4. *In a sample of 1,000 people in Maharashtra, 540 are rice eaters and the rest are wheat eaters. Can we assume that both rice and wheat are equally popular in this State at 1% level of significance ?*

Solution. In the usual notations we are given $n = 1{,}000$

X = Number of rice eaters = 540

$\therefore \quad p$ = Sample proportion of rice eaters = $\frac{X}{n} = \frac{540}{1000} = 0{\cdot}54$

Null Hypothesis, H_0 : Both rice and wheat are equally popular in the State so that

P = Population proportion of rice eaters in Maharashtra = 0·5

$\Rightarrow$ $Q = 1 - P = 0{\cdot}5$

Alternative Hypothesis, $H_1 : P \neq 0{\cdot}5$ (two-tailed alternative).

Test Statistic. Under H_0, the test statistic is

$$Z = \frac{p - P}{\sqrt{PQ/n}} \sim N(0, 1), \text{ (since } n \text{ is large).}$$

Now $$Z = \frac{0{\cdot}54 - 0{\cdot}50}{\sqrt{0{\cdot}5 \times 0{\cdot}5/1000}} = \frac{0{\cdot}04}{0{\cdot}0138} = 2{\cdot}532$$

Conclusion. The significant or critical value of Z at 1% level of significance for two-tailed test is 2·58. Since computed $Z = 2{\cdot}532$ is less than 2·58, it is not significant at 1% level of significance. Hence the null hypothesis is accepted and we may conclude that rice and wheat are equally popular in Maharashtra State.

Example 12·5. *Twenty people were attacked by a disease and only 18 survived. Will you reject the hypothesis that the survival rate, if attacked by this disease, is 85% in favour of the hypothesis that it is more, at 5% level. (Use Large Sample Test.)*

[Patna Univ. B.Sc. (Hons.), 1992; Bombay Univ. B.Sc. 1987]

Solution. In the usual notations, we are given $n = 20$.

X = Number of persons who survived after attack by a disease = 18

p = Proportion of persons survived in the sample = $\frac{18}{20} = 0{\cdot}90$

Null Hypothesis, $H_0 : P = 0{\cdot}85$, *i.e.*, the proportion of persons survived after attack by a disease in the lot is 85%.

Alternative Hypothesis, $H_1 : P > 0{\cdot}85$ (Right-tail alternative).

Test Statistic. Under H_0, the test statistic is :

$$Z = \frac{p - P}{\sqrt{PQ/n}} \sim N(0, 1), \text{ (since sample is large).}$$

Now $$Z = \frac{0{\cdot}90 - 0{\cdot}85}{\sqrt{0{\cdot}85 \times 0{\cdot}15/20}} = \frac{0{\cdot}05}{0{\cdot}079} = 0{\cdot}633$$

Conclusion. Since the alternative hypothesis is one-sided (right-tailed), we shall apply right-tailed test for testing significance of Z. The significant value of Z at 5% level of significance for right-tail test is + 1·645. Since computed value of $Z = 0{\cdot}633$ is less than 1·645, it is not significant and we may accept the null hypothesis at 5% level of significance.

12·9·2. Test of Significance for Difference of Proportions. Suppose we want to compare two distinct populations with respect to the prevalence of a certain attribute, say A, among their members. Let X_1, X_2 be the number of persons possessing the given attribute A in random samples of sizes n_1 and n_2 from the two populations respectively. Then sample proportions are given by

$$p_1 = X_1/n_1 \quad \text{and} \quad p_2 = X_2/n_2$$

If P_1 and P_2 are the population proportions, then

$$E(p_1) = P_1, \; E(p_2) = P_2 \qquad [c.f. \text{ Equation (12·4}a\text{)}]$$

and

$$V(p_1) = \frac{P_1Q_1}{n_1} \quad \text{and} \quad V(p_2) = \frac{P_2Q_2}{n_2}$$

Since for large samples, p_1 and p_2 are asymptotically normally distributed, $(p_1 - p_2)$ is also normally distributed. Then the standard variable corresponding to the difference $(p_1 - p_2)$ is given by

$$Z = \frac{(p_1 - p_2) - E(p_1 - p_2)}{\sqrt{V(p_1 - p_2)}} \sim N(0, 1)$$

Under the *null hypothesis* $H_0 : P_1 = P_2$, *i.e.*, there is no significant difference between the sample proportions, we have

$$E(p_1 - p_2) = E(p_1) - E(p_2) = P_1 - P_2 = 0 \qquad (\text{Under } H_0)$$

Also $V(p_1 - p_2) = V(p_1) + V(p_2)$,

the covariance term $\text{Cov}(p_1, p_2)$ vanishes, since sample proportions are independent.

$$\therefore \qquad V(p_1 - p_2) = \frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2} = PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right),$$

since under $H_0 : P_1 = P_2 = P$, (say), and $Q_1 = Q_2 = Q$.

Hence under $H_0 : P_1 = P_2$, the test statistic for the difference of proportions becomes

$$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1) \qquad \ldots(12\cdot5)$$

In general, we do not have any information as to the proportion of A's in the populations from which the samples have been taken. Under $H_0 : P_1 = P_2 = P$, (say), an unbiased estimate of the population proportion P, based on both the samples is given by

$$\hat{P} = \frac{n_1p_1 + n_2p_2}{n_1 + n_2} = \frac{X_1 + X_2}{n_1 + n_2} \qquad \ldots(12\cdot5a)$$

The estimate is unbiased, since

$$E(\hat{P}) = \frac{1}{n_1 + n_2} E\left[n_1p_1 + n_2p_2\right] = \frac{1}{n_1 + n_2}\left[n_1E(p_1) + n_2E(p_2)\right]$$

$$= \frac{1}{n_1 + n_2}\left[n_1P_1 + n_2P_2\right] = P \qquad [\because P_1 = P_2 = P, \text{ under } H_0]$$

Thus (12·5) along with (12·5a) gives the required test statistic.

Remarks 1. Suppose we want to test the significance of the difference between p_1 and p, where

$$p = \frac{(n_1p_1 + n_2p_2)}{(n_1 + n_2)}$$

gives a pooled estimate of the population proportion on the basis of both the samples. We have

$$V(p_1 - p) = V(p_1) + V(p) - 2\,\text{Cov}\,(p_1, p) \qquad \ldots(*)$$

Since p_1 and p are not independent, Cov $(p_1, p) \neq 0$.

$$\text{Cov}\,(p_1, p) = E[\{p_1 - E(p_1)\}\{p - E(p)\}]$$

$$= E\left[\{p_1 - E(p_1)\}\left\{\frac{1}{n_1 + n_2}\{n_1p_1 + n_2p_2 - E(n_1p_1 + n_2p_2)\}\right\}\right]$$

$$= \frac{1}{n_1 + n_2} E\left[\{p_1 - E(p_1)\}\left\{n_1(p_1 - E(p_1)) + n_2(p_2 - E(p_2))\right\}\right]$$

$$= \frac{1}{n_1 + n_2}\left[n_1 E\left\{p_1 - E(p_1)\right\}^2 + n_2 E\left\{(p_1 - E(p_1))(p_2 - E(p_2))\right\}\right]$$

$$= \frac{1}{n_1 + n_2}\left[n_1 V(p_1) + n_2\,\text{Cov}\,(p_1, p_2)\right]$$

$$= \frac{1}{n_1 + n_2} n_1 V(p_1), \qquad [\because \text{Cov}\,(p_1, p_2) = 0]$$

$$= \frac{n_1}{n_1 + n_2} \cdot \frac{pq}{n_1} = \frac{pq}{n_1 + n_2}$$

$$\text{Also Var}\,(p) = \frac{1}{(n_1 + n_2)^2} E\left[(n_1p_1 + n_2p_2) - E(n_1p_1 + n_2p_2)\right]^2$$

$$= \frac{1}{(n_1 + n_2)^2}\left[n_1^2\,\text{Var}\,(p_1) + n_2^2\,\text{Var}\,(p_2)\right],$$

covariance term vanishes since p_1 and p_2 are independent.

$$\therefore \quad \text{Var}\,(p) = \frac{1}{(n_1 + n_2)^2}\left[n_1^2 \cdot \frac{pq}{n_1} + n_2^2 \cdot \frac{pq}{n_2}\right]$$

$$= \frac{pq}{n_1 + n_2}$$

Substituting in (*) and simplifying, we shall get

$$V(p_1 - p) = \frac{pq}{n_1} + \frac{pq}{n_1 + n_2} - 2\,\frac{pq}{n_1 + n_2} = pq\left[\frac{n_2}{n_1(n_1 + n_2)}\right]$$

Thus, the test statistic in this case becomes

$$Z = \frac{p_1 - p}{\sqrt{\dfrac{n_2}{(n_1 + n_2)} \cdot \dfrac{pq}{n_1}}} \sim N(0, 1) \qquad \ldots(12\cdot5b)$$

2. Suppose the population proportions P_1 and P_2 are given to be distinctly different, *i.e.*, $P_1 \neq P_2$ and we want to test if the difference $(P_1 - P_2)$ in population proportions is likely to be hidden in simple samples of sizes n_1 and n_2 from the two populations respectively.

We have seen that in the usual notations,

$$Z = \frac{(p_1 - p_2) - E(p_1 - p_2)}{\text{S.E.}(p_1 - p_2)} = \frac{(p_1 - p_2) - (P_1 - P_2)}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}} \sim N(0, 1)$$

Here sample proportions are not given. If we set up the *null hypothesis* $H_0 : p_1 = p_2$, *i.e.* the samples will not reveal the difference in the population proportions or in other words the difference in population proportions is likely to be hidden in sampling, the test statistic becomes

$$|Z| = \frac{|P_1 - P_2|}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}} \sim N(0, 1) \qquad \ldots(12\cdot5c)$$

Example 12·6. *Random samples of 400 men and 600 women were asked whether they would like to have a flyover near their residence. 200 men and 325 women were in favour of the proposal. Test the hypothesis that proportions of men and women in favour of the proposal, are same against that they are not, at 5% level.* **[*Agra Univ. M.A., 1992*]**

Solution. *Null Hypothesis* $H_0 : P_1 = P_2 = P$, (say), *i.e.*, there is no significant difference between the opinion of men and women as far as proposal of flyover is concerned.

Alternative Hypothesis, $H_1 : P_1 \neq P_2$ (two-tailed).

We are given :

$n_1 = 400$, X_1 = Number of men favouring the proposal = 200

$n_2 = 600$, X_2 = Number of women favouring the proposal = 325

$\therefore$ p_1 = Proportion of men favouring the proposal in the sample

$$= \frac{X_1}{n_1} = \frac{200}{400} = 0\cdot5$$

p_2 = Proportion of women favouring the proposal in the sample

$$= \frac{X_2}{n_2} = \frac{325}{600} = 0\cdot541$$

Test Statistic. Since samples are large, the test statistic under the Null Hypothesis, H_0 is :

$$Z = \frac{p_1 - p_2}{\sqrt{\hat{P}\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1)$$

where $$\hat{P} = \frac{n_1p_1 + n_2p_2}{n_1 + n_2} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{200 + 325}{400 + 600} = \frac{525}{1000} = 0\cdot525$$

$\Rightarrow$ $\hat{Q} = 1 - \hat{P} = 1 - 0\cdot525 = 0\cdot475$

$\therefore$ $$Z = \frac{0\cdot500 - 0\cdot541}{\sqrt{0\cdot525 \times 0\cdot475 \times \left(\frac{1}{400} + \frac{1}{600}\right)}}$$

$$= \frac{-0{\cdot}041}{\sqrt{0{\cdot}525 \times 0{\cdot}475 \times (10/2{,}400)}}$$

$$= \frac{-0{\cdot}041}{\sqrt{0{\cdot}001039}} = \frac{-0{\cdot}041}{0{\cdot}0323} = -1{\cdot}269$$

Conclusion. Since $|Z| = 1{\cdot}269$ which is less than 1·96, it is not significant at 5% level of significance. Hence H_0 may be accepted at 5% level of significance and we may conclude that men and women do not differ significantly as regards proposal of flyover is concerned.

Example 12·7. *A company has the head office at Calcutta and a branch at Bombay. The personnel director wanted to know if the workers at the two places would like the introduction of a new plan of work and a survey was conducted for this purpose. Out of a sample of 500 workers at Calcutta, 62% favoured the new plan. At Bombay out of a sample of 400 workers, 41% were against the new plan. Is there any significant difference between the two groups in their attitude towards the new plan at 5% level ?*

Solution. In the usual notations, we are given :

$$n_1 = 500,\ p_1 = 0{\cdot}62 \text{ and } n_2 = 400,\ p_2 = 1 - 0{\cdot}41 = 0{\cdot}59$$

Null hypothesis, $H_0 : P_1 = P_2$, *i.e.,* there is no significant difference between the two groups in their attitude towards the new plan.

Alternative hypothesis, $H_1 : P_1 \neq P_2$ (Two-tailed).

Test Statistic. Under H_0, the test statistic for large samples is :

$$Z = \frac{p_1 - p_2}{\text{S.E.}\ (p_1 - p_2)} = \frac{p_1 - p_2}{\sqrt{\hat{P}\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1)$$

where $$\hat{P} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{500 \times 0{\cdot}62 + 400 \times 0{\cdot}59}{500 + 400} = 0{\cdot}607$$

and $$\hat{Q} = 1 - \hat{P} = 0{\cdot}393$$

$$\therefore \quad Z = \frac{0{\cdot}62 - 0{\cdot}59}{\sqrt{0{\cdot}607 \times 0{\cdot}393 \times \left(\frac{1}{500} + \frac{1}{400}\right)}}$$

$$= \frac{0{\cdot}03}{\sqrt{0{\cdot}00107}} = \frac{0{\cdot}03}{0{\cdot}0327} = 0{\cdot}917.$$

Critical region. At 5% level of significance, the critical value of Z for a two-tailed test is 1·96. Thus the critical region consists of all values of $Z \geq 1{\cdot}96$ or $Z \leq -1{\cdot}96$.

Conclusion. Since the calculated value of $|Z| = 0{\cdot}917$ is less than the critical value of Z (1·96), it is not significant at 5% level of significance. Hence the data do not provide us any evidence against the null hypothesis which may be accepted, and we conclude that there is no significant difference between the two groups in their attitude towards the new plan.

Example 12·8. *Before an increase in excise duty on tea, 800 persons out of a sample of 1,000 persons were found to be tea drinkers. After an increase in*

duty, 800 people were tea drinkers in a sample of 1,200 people. Using standard error of proportion, state whether there is a significant decrease in the consumption of tea after the increase in excise duty ?

Solution. In the usual notations, we have $n_1 = 1{,}000$; $n_2 = 1{,}200$

p_1 = Sample proportion of tea drinkers before increase in excise duty

$$= \frac{800}{1000} = 0{\cdot}80$$

p_2 = Sample proportion of tea drinkers after increase in excise duty

$$= \frac{800}{1200} = 0{\cdot}67$$

Null Hypothesis, $H_0 : P_1 = P_2$, *i.e.*, there is no significant difference in the consumption of tea before and after the increase in excise duty.

Alternative Hypothesis, $H_1 : P_1 > P_2$ (Right-tailed alternative).

Test Statistic. Under the null hypothesis, the test statistic is

$$Z = \frac{p_1 - p_2}{\sqrt{\hat{P}\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1) \qquad \text{(Since samples are large)}$$

where

$$\hat{P} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{800 + 800}{1000 + 1200} = \frac{16}{22}, \text{ and } \hat{Q} = 1 - \hat{P} = \frac{6}{22}$$

$$\therefore \quad Z = \frac{0{\cdot}80 - 0{\cdot}67}{\sqrt{\frac{16}{22} \times \frac{6}{22} \times \left(\frac{1}{1000} + \frac{1}{1200}\right)}}$$

$$= \frac{0{\cdot}13}{\sqrt{\frac{16}{22} \times \frac{6}{22} \times \frac{11}{6000}}} = \frac{0{\cdot}13}{0{\cdot}019} = 6{\cdot}842$$

Conclusion. Since Z s much greater than 1·645 as well as 2·33 (since test is one-tailed), it is highly significant at both 5% and 1% levels of significance. Hence,

we reject the null hypothesis H_0 and conclude that there is a significant decease in the consumption of tea after increase in the excise duty.

Example 12·9. *A cigarette manufacturing firm claims that its brand A of the cigarettes outsells its brand B by 8%. If it is found that 42 out of a sample of 200 smokers prefer brand A and 18 out of another random sample of 100 smokers prefer brand B, test whether the 8% difference is a valid claim. (Use 5% level of significance.)*

Solution. We are given.:

$$n_1 = 200, X_1 = 42 \implies p_1 = \frac{X_1}{n_1} = \frac{42}{200} = 0{\cdot}21$$

$$n_2 = 100, X_2 = 18 \implies p_2 = \frac{X_2}{n_2} = \frac{18}{100} = 0{\cdot}18$$

We set up the *Null Hypothesis* that 8% difference in the sale of two brands of cigarettes is a valid claim, *i.e.*, $H_0 : P_1 - P_2 = 0{\cdot}08$.

Alternative Hypothesis : $H_1 : P_1 - P_2 \neq 0{\cdot}08$ (Two-tailed).

Under H_0, the test statistic is (since samples are large)

$$Z = \frac{(p_1 - p_2) - (P_1 - P_2)}{\sqrt{\hat{P}\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1)$$

where $\hat{P} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{42 + 18}{200 + 100} = \frac{60}{300} = 0{\cdot}20 \Rightarrow \hat{Q} = 1 - \hat{P} = 0{\cdot}80$

$$\therefore \quad Z = \frac{(0{\cdot}21 - 0{\cdot}18) - (0{\cdot}08)}{\sqrt{0{\cdot}2 \times 0{\cdot}8\left(\frac{1}{200} + \frac{1}{100}\right)}} = \frac{-0{\cdot}05}{\sqrt{0{\cdot}16 \times 0{\cdot}015}}$$

$$= \frac{-0{\cdot}05}{\sqrt{0{\cdot}0024}} = \frac{-0{\cdot}05}{0{\cdot}04899} = -1{\cdot}02$$

Since $|Z| = 1{\cdot}02 < 1{\cdot}96$, it is not significant at 5% level of significance. Hence null hypothesis may be retained at 5% level of significance and we may conclude that a difference of 8% in the sale of two brands of cigarettes is a valid claim by the firm.

Example 12·10. *On the basis of their total scores, 200 candidates of a civil service examination are divided into two groups, the upper 30 per cent and the remaining 70 per cent. Consider the first question of this examination. Among the first group, 40 had the correct answer, whereas among the second group, 80 had the correct answer. On the basis of these results, can one conclude that the first question is no good at discriminating ability of the type being examined here ?*

Solution. Here, we have

n = Total number of candidates = 200

n_1 = The number of candidates in the upper 30% group

$$= \frac{30}{100} \times 200 = 60$$

n_2 = The number of candidates in the remaining 70% group

$$= \frac{70}{100} \times 200 = 140$$

X_1 = The number of candidates, with correct answer in the first group = 40

X_2 = The number of candidates, with correct answer in the second group = 80

$$\therefore \quad p_1 = \frac{X_1}{n_1} = \frac{40}{60} = 0{\cdot}6666 \text{ and } p_2 = \frac{X_2}{n_2} = \frac{80}{140} = 0{\cdot}5714$$

Null Hypothesis, H_0 : There is no significant difference in the sample proportions, *i.e.*, $P_1 = P_2$, *i.e.*, the first question is no good at dicriminating the ability of the type being examined here.

Alternative Hypothesis, $H_1 : P_1 \neq P_2$.

Test Statistic. Under H_0 the test statistic is :

$$Z = \frac{p_1 - p_2}{\sqrt{\hat{P}\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1) \quad \text{(since samples are large).}$$

where

$$\hat{P} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{40 + 80}{60 + 140} = 0{\cdot}6, \quad \hat{Q} = 1 - \hat{P} = 0{\cdot}4$$

$$\therefore \qquad Z = \frac{0{\cdot}6666 - 0{\cdot}5714}{\sqrt{0{\cdot}6 \times 0{\cdot}4 \left(\frac{1}{60} + \frac{1}{140}\right)}} = \frac{0{\cdot}0953}{0{\cdot}0756} = 1{\cdot}258$$

Conclusion. Since $|Z| < 1{\cdot}96$, the data are consistent with the null hypothesis at 5% level of significance. Hence we conclude that the first question is not good enough to distinguish between the ability of the two groups of candidates.

Example 12·11. *In a year there are 956 births in a town A, of which 52·5% were males, while in towns A and B combined, this proportion in a total of 1,406 births was 0·496. Is there any significant difference in the proportion of male births in the two towns ?*

Solution. We are given

$$n_1 = 956, \; n_1 + n_2 = 1{,}406 \text{ or } n_2 = 1{,}406 - 956 = 450$$

p_1 = Proportion of males in the sample of town $A = 0{\cdot}525$.

Let p_2 be the proportion of males in the sample (of size n_2) of town B. Then

$\hat{P}$ = Proportion of males in both the samples combined.

$$= \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = 0{\cdot}496 \qquad \text{(Given)}$$

$$\therefore \qquad \frac{956 \times 0{\cdot}525 + 450 \times p_2}{1{,}406} = 0{\cdot}496$$

$$\Rightarrow \qquad p_2 = 0{\cdot}434 \qquad \text{(On simplification)}$$

Null Hypothesis, $H_0 : P_1 = P_2$, *i.e.,* there is no significant difference in the proportion of male births in the two towns A and B.

Alternative Hypothesis, $H_1 : P_1 \neq P_2$ (two-tailed).

Test Statistic. Under H_0, the test statistic is :

$$Z = \frac{p_1 - p_2}{\sqrt{\hat{P}\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1) \qquad \text{(Since samples are large)}$$

where $\hat{P} = \dfrac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = 0{\cdot}496, \quad \hat{Q} = 1 - \hat{P} = 0{\cdot}504$

$$\therefore \qquad Z = \frac{0{\cdot}525 - 0{\cdot}434}{\sqrt{0{\cdot}496 \times 0{\cdot}504 \left(\frac{1}{956} + \frac{1}{450}\right)}} = \frac{0{\cdot}091}{0{\cdot}027} = 3{\cdot}368$$

Conclusion. Since $|Z| > 3$, the null hypothesis is rejected, *i.e.,* the data are inconsistent with the hypothesis $P_1 = P_2$ and we conclude that there is significant difference in the proportion of male births in the towns A and B.

Example 12·12. *In two large populations, there are 30 and 25 per cent respectively of blue-eyed people. Is this difference likely to be hidden in samples of 1,200 and 900 respectively from the two populations ?*

[*Delhi Univ. B.Sc., 1992*]

Solution. Here, we are given $n_1 = 1200$, $n_2 = 900$.

P_1 = Proportion of blue-eyed people in the first population

$= 30\% = 0{\cdot}30$.

P_2 = Proportion of blue-eyed people in the second population

$= 25\% = 0{\cdot}25$.

$\therefore \qquad Q_1 = 1 - P_1 = 0{\cdot}70$ and $Q_2 = 1 - P_2 = 0{\cdot}75$

We set up the *null hypothesis* H_0 that $p_1 = p_2$, *i.e.*, the sample proportions are equal, *i.e.*, the difference in population proportions is likely to be hidden in sampling.

Test Statistic. Under $H_0 : p_1 = p_2$, the test statistic is :

$$|Z| = \frac{|P_1 - P_2|}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}} \sim N(0, 1) \qquad \text{(Since samples are large.)}$$

$$\therefore \qquad |Z| = \frac{0{\cdot}30 - 0{\cdot}25}{\sqrt{\frac{0{\cdot}3 \times 0{\cdot}7}{1{,}200} + \frac{0{\cdot}25 \times 0{\cdot}75}{900}}} = \frac{0{\cdot}05}{0{\cdot}0195} = 2{\cdot}56$$

Conclusion. Since $|Z| > 1{\cdot}96$, the null hypothesis ($p_1 = p_2$), is refuted at 5% level of significance and we conclude that the difference in population proportions is unlikely to be hidden in sampling. In other words, these samples will reveal the difference in the population proportions.

Example 12·13. *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 whether the proportion of failures in the university teaching departments is significantly greater than the proportion of failures in the university teaching departments and affiliated colleges taken together.*

Solution. Here we are given : $n_1 = 400$, $n_2 = 500$

$$p_1 = \frac{300}{400} = 0{\cdot}75, \quad p_2 = \frac{300}{500} = 0{\cdot}60$$

$$\therefore \qquad q_1 = 1 - p_1 = 1 - 0{\cdot}75 = 0{\cdot}25 \quad \text{and} \quad q_2 = 1 - p_2 = 0{\cdot}40$$

Here we set up the *null hypothesis* H_0 that p_1 and $\hat{p}$, where $\hat{p}$ is the pooled estimate, *i.e.*, proportion of failures in the university teaching departments and affiliated colleges taken together, do not differ significantly.

$$\text{S.E. of } (\hat{p} - p_1) = \sqrt{\frac{\hat{p}\hat{q}}{n_1 + n_2} \times \frac{n_2}{n_1}} \qquad [c.f.\ (12{\cdot}5b) \text{ page } 12{\cdot}18]$$

$$\text{where} \qquad \hat{p} = \frac{n_1p_1 + n_2p_2}{n_1 + n_2} = \frac{400 \times 0{\cdot}75 + 500 \times 0{\cdot}60}{400 + 500} = 0{\cdot}67$$

$$\hat{q} \doteq 1 - 0{\cdot}67 = 0{\cdot}33$$

$$\therefore \text{ S.E. of } (\hat{p} - p_1) = \sqrt{\frac{0{\cdot}67 \times 0{\cdot}33}{400 + 500} \times \frac{500}{400}} = 0{\cdot}018$$

Test Statistic. Under the null hypothesis H_0, the test statistic is :

$$Z = \frac{\hat{p} - p_1}{\text{S.E. of } (\hat{p} - p_1)} \sim N(0, 1) \qquad \text{(Since samples are large.)}$$

$$Z = \frac{0{\cdot}67 - 0{\cdot}33}{0{\cdot}018} = \frac{0{\cdot}15}{0{\cdot}018} = 8{\cdot}3$$

Conclusion. Since the calculated value of Z is much greater than 3, it is highly signifcant. Hence null hypothesis H_0 is rejected and we conclude that there is significant difference between p_1 and $\hat{p}$.

Example 12·14. *If for one-half of n events, the chance of success is p and the chance of failure is q, while for the other half the chance of success is q and the chance of failure is p, show that the standard deviation of the number of successes is the same as if the chance of successes were p in all the cases, i.e., $\sqrt{npq}$ but that the mean of the number of successes is n/2 and not np.*

Solution. Let X_1 and X_2 denote the number of successes in the first half and the second half of n events respectively. Then according to the given conditions, we have

$$\left.\begin{aligned} E(X_1) &= \frac{n}{2}p \\ V(X_1) &= \frac{n}{2}pq \end{aligned}\right\} \text{and} \quad \left.\begin{aligned} E(X_2) &= \frac{n}{2}q \\ V(X_2) &= \frac{n}{2}pq \end{aligned}\right\}$$

The mean and variance of the number of successes in all the n events are given by

$$E(X_1 + X_2) = E(X_1) + E(X_2) = \frac{n}{2}p + \frac{n}{2}q = \frac{n}{2}$$

and

$$V(X_1 + X_2) = V(X_1) + (X_2) = \frac{n}{2}pq + \frac{n}{2}qp = npq,$$

since the first and second half of events are independent.

Hence the variance is the same as if the probability of success in all the n events is p.

EXERCISE 12(a)

1. (*a*) There are 2 populations and P_1 and P_2 are the proportion of members in the two populations belonging to 'low-income' group. It is disired to test the hypothesis $H_0 : P_1 = P_2$. Explain clearly, the procedure that you would follow to carry out the above test at 5% level of significance.

State the theorem on which the above test is based.

In respect of the above 2 populations, if it is claimed that P_1, the proportion of 'low-income' group in the first population is greater than P_2, how will you modify the procedure to test this claim (at 5% level) ?

(*b*) Take a concrete illustration and in relation to this illustration, explain the following terms :—

(*i*) Null hypothesis and alternative hypothesis.

(*ii*) Type I and Type II errors.

(*iii*) Critical Region.

(*c*) Suggest a possible source of bias in the following :

(*i*) The mean income per family in a certain town is sought to be estimated by sampling from motor owners.

(*ii*) Readers of newspapers are sampled by printing in it an invitation to them to send up their observations on some typical event.

(*iii*) A barrel of apples is sampled by taking a handful from the top.

(*iv*) A set of digits is taken by opening a telephone directory at random and choosing the telephone numbers in the order in which they appear on the page.

2. (*a*) Explain clearly the terms "Standard Error" and "Sampling Distribution." Show that in a series of n independent trials with constant probability p of success, the standard error of the proportion of successes is $\sqrt{pq/n}$, where $q = 1 - p$.

(*b*) n individuals fall into one or the other two categories with probabilities p and q $(=1-p)$, the number in the two categories are x_1 and x_2 $(x_1 + x_2 = n)$. Show that covariance between x_1 and x_2 is $-npq$. Hence obtain the variance of the difference $\left(\frac{x_1}{n} - \frac{x_2}{n}\right)$, between the proportions.

(*c*) Explain clearly the procedure generally followed in testing of a hypothesis. Point out the difference between one-tail and two-tail tests.

(*d*) What do you mean by interval estimation and how would you set up the confidence limits for a parameter from a sample ? Give the formula for 95% confidence limits for mean and proportion. What modifcations do you have to make if the sampling is done from finite population, (*i*) without replacement, (*ii*) with replacement ? [***Calcutta Univ. B.A. (Maths Hons.), 1988***]

3. P_1 and P_2 are the (unknown) proportions of students wearing glasses in two universities A and B. To compare P_1 and P_2, samples of sizes n_1 and n_2 are taken from the two populations and the number of students wearing glasses is found to be x_1 and x_2 respectively. Suggest an unbiased estimate of $(P_1 - P_2)$ and obtain its sampling distribution when n_1 and n_2 are large. Hence explain how to test the hypothesis that $P_1 = P_2$.

4. (*a*) A coin is tossed 10,000 times and it turns up head 5,195 times. Discuss whether the coin may be regarded as unbiased one, explaining briefly the theoretical principles you would use for this purpose. (**Ans.** No.)

(*b*) A biased coin was thrown 400 times and head resulted 240 times. Find the standard error of the observed proportion of heads and deduce that the probability of getting a head in a single throw of the coin lies almost certainly between 0·53 and 0·67. (**Ans.** 0·02445).

(*c*) Experience has shown that 20% of a manufactured product is of the top quality. In one day's production of 400 articles only 50 are of top quality. Show that either the production of the day taken was not a representative sample or the hypothesis of 20% was wrong. (**Ans.** $Z = 3{\cdot}75$)

5. (*a*) In a large consignment of oranges a random sample of 64 oranges revealed that 14 oranges were bad. Is it reasonable to assume that 20% of the oranges were bad?

(*b*) By a mobile court checking in certain buses it was found that out of 1000 people checked on a certain day at Red Fort, 10 persons were found to be ticketless travellers. If daily 1 lakh passengers travel by the buses, find out the estimated limits to the ticketless travellers. (**Ans.** 997 to 1003)

(*c*) In a random sample of 81 items taken from a large consignment some were found to be defective. If the standard error of the proportion of defective items in the sample is 1/18, find 95% confidence limits of the percentage of defective items in the consignment.

[Madras Univ. B.Sc. (Stat. Main), 1991]

6. (*a*) In some dice throwing experiments Weldon threw dice 75,145 times and of these 49,152 yielded a 4, 5 or 6. Is this consistent with the hypothesis that the dice was unbiased?

Hint. H_0 : Dice is unbiased, *i.e.*, $P = \frac{3}{6} = \frac{1}{2} = 0{\cdot}5$; $H_1 : P \neq \frac{1}{2}$

Test Statistic. Under H_0, $Z = \dfrac{p - P}{\sqrt{PQ/n}} = \dfrac{0{\cdot}654 - 0{\cdot}5}{\sqrt{0{\cdot}5 \times 0{\cdot}5/75145}} = \dfrac{0{\cdot}154}{0{\cdot}0018}$

Ans. No.

(*b*) 1,000 apples are taken from a large consignment and 100 are found to be bad. Estimate the percentage of bad apples in the consignment and assign the limits within which the percentage lies.

7. (*a*) A personnel manager claims that 80 per cent of all single women hired for secretarial job get married and quit work within two years after they are hired. Test this hypothesis at 5% level of significance if among 200 such secretaries, 112 got married within two years after they were hired and quit their jobs.

(*b*) A manufacturer claimed that at least 98% of the steel pipes which he supplied to a factory conformed to specifications. An examination of a sample of 500 pieces of pipes revealed that 30 were defective. Test this claim at a significance level of (*i*) 0·05, (*ii*) 0·01.

Hint. X = No. of pipes conforming to specifications in the sample.
$= 500 - 30 = 470$
p = Sample proportion of pipes conforming to specifications
$= \dfrac{470}{500} = 0{\cdot}94$

$H_0 : P = 0{\cdot}98$, *i.e.*, the proportion of pipes conforming to specifications in the lot is 98%.

$H_1 : P < 0{\cdot}98$ (Left-tail alternative)

Test Statistic. $Z = \dfrac{p - P}{\sqrt{PQ/n}} = \dfrac{0{\cdot}94 - 0{\cdot}98}{\sqrt{0{\cdot}98 \times 0{\cdot}02/500}}$

(*c*) A social worker believes that fewer than 25% of the couples in a certain area ever used any form of birth control. A random sample of 120 couples was

contacted. Twenty of them said they had used some method of birth control. Comment on the social worker's belief.

$H_0 : P = 0{\cdot}25, H_1 : P < 0{\cdot}25$ (left-Tailed)

8. In a random sample of 800 adults from the population of a certain large city, 600 are found to have dark hair. In a random sample of 1,000 adults from the habitants of another large city, 700 are dark haired. Show that the difference of the proportion of dark haired people is nearly 2·4 times the standard error of the difference for samples of above sizes.

9. (*a*) In a random sample of 100 men taken from village *A*, 60 were found to be consuming alcohol. In another sample of 200 men taken form village *B*, 100 were found to be consuming alcohol. Do the two villages differ significantly in respect of the proportion of men who consume alcohol ?

[***Delhi Univ. M.A. (Business Eco.), 1987***]

(*b*) In a random sample of 500 men from a particular district of U.P., 300 are found to be smokers. In one of 1,000 men from another district, 550 are smokers. Do the data indicate that the two districts are significantly different with respect to the prevalence of smoking among men ?

Ans. $Z = 1{\cdot}85$, (not significant). (***Delhi Univ. B.Sc., 1991***)

10. A company is considering two different television advertisements for promotion of a new product. Management believed that the advertisement *A* is more effective than advertisement *B*. Two test market areas with virtually identical consumer characteristics are selected; *A* is used in one area and *B* in other area. In a random sample of 60 customers who saw *A*, 18 tried the product. In another random sample of 100 customers who saw *B*, 22 tried the product. Does this indicate that advertisement *A* is more effective than advertisement *B*, if a 5% level of significance is used ? Given critical value at 5% level is 1·96 and at 10% level of significance is 1·645.

[***Delhi Univ. M.C.A., 1990***]

11. (*a*) 1,000 apples kept under one type of storage were found to show rotting to the extent of 4%. 1,500 apples kept under another kind of storage showed 3% rotting. Can it be reasonably concluded that the second type of storage is superior to the first ?

(*b*) In a referendum submitted to the students body at a university, 850 men and 566 women voted. 530 of the men and 304 of the women voted yes. Does this indicate a significant difference of opinion on the matter at 1% level, between men and women students. [**Ans.** $Z = 3{\cdot}2$, (significant).]

(*c*) In a simple sample of 600 high school students from a State, 400 are found to use dot pens. In one of 900 from a neighbouring State, 450 are found to use dot pens. Do the data indicate that the States are significantly different with respect to the habit of using dot pens among the students ? (**Ans.** Yes.)

12. (*a*) A firm, manufacturing dresses for children, sent out advertisement through mail. Two groups of 1,000 each were contacted; the first group having been contacted in white covers while the second in blue covers. 20% from the first while 28% from the second replied.

Do you think that blue envelopes help the sales ?

(*b*) A machine puts out 16 imperfect articles in a sample of 500. After machine is overhauled, it puts out 3 imperfect articles in a batch of 100. Has the machine improved ?

Hint. We are given : $n_1 = 500$, and $n_2 = 100$

$$p_1 = \frac{16}{500} = 0{\cdot}032; \quad p_2 = \frac{3}{100} = 0{\cdot}030$$

Null Hypothesis, $H_0 : P_1 = P_2$, *i.e.*, there is no significant difference in the machine before overhauling and after overhauling. In other words, the machine has not improved after overhauling.

Alternative Hypothesis, $H_1 : P_2 < P_1$ or $P_1 > P_2$.

$$\hat{P} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{16 + 3}{500 + 100} = \frac{19}{600} = 0{\cdot}032$$

$$\text{S.E.}\,(p_1 - p_2) = \sqrt{0{\cdot}032 \times 0{\cdot}968 \left(\frac{1}{500} + \frac{1}{100}\right)} = 0{\cdot}0193$$

$$Z = \frac{0{\cdot}032 - 0{\cdot}030}{0{\cdot}0193} = \frac{0{\cdot}002}{0{\cdot}0193} = 1{\cdot}04$$

Since $Z < 1{\cdot}645$ (Right-tailed test), it is not significant at 5% level of significance.

(*c*) In a large city *A*, 25% of a random sample of 900 school boys had defective eye-sight. In another large city *B*, 15·5% of a random sample of 1,600 school boys had the same defect. Is this difference between the two proportions significant ? (**Ans.** Not significant.)

13. (*a*) A candidate for election made a speech in city *A* but not in *B*. *A* sample of 500 voters from city *A* showed that 59·6% of the voters were in favour of him, whereas a sample of 300 voters from city *B* showed that 50% of the voters favoured him. Discuss whether his speech could produce any effect on voters in city *A*. Use 5% level.

Ans. $|Z| = 2{\cdot}67$. Yes.

(*b*) In a large city, 16 out of a random sample of 500 men were found to be drinkers. After the heavy increase in tax on intoxicants another random sample of 100 men in the same city included 3 drinkers. Was the observed decrease in the proportion of drinkers significant after tax increase ?

Ans. $H_0 : P_1 = P_2$, $H_1 : P_1 > P_2$; $Z = 1{\cdot}04$. Not sigificant.

14. The sex ratio at birth is sometimes given by the ratio of male to female births instead of the proportion of male to total births. If z is the ratio, *i.e.*, $z = p/q$, show that the standard error of z is approximately $\frac{1}{1+z}\sqrt{\left(\frac{z}{n}\right)}$ n being large, so that deviations are small compared with mean.

12·10. Sampling of Variables. In the present section we will discuss in detail the sampling of variables such as height, weight, age, income, etc. In the case of sampling of variables each member of the population provides the value of the variable and the aggregate of these values forms the frequency distribution of the population. From the population, a random sample of size n

can be drawn by any of the sampling methods discussed before which is same as choosing n values of the given variable from the distribution.

12·11. Unbiased Estimate for population Mean (μ) and Variance (σ^2). Let $x_1, x_2, \ldots, x_n$ be a random sample of size n from a large population $X_1, X_2, \ldots X_N$ (of size N) with mean μ and variance σ^2. Then the sample mean ($\bar{x}$) and variance (s^2) are given by

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \text{and} \quad s^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

Now $$E(\bar{x}) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(x_i)$$

Since x_i is a sample observation from the population X_i, ($i = 1, 2, \ldots, N$) it can take any one of the values $X_1, X_2, \ldots, X_N$ each with equal probability $1/N$.

$$\therefore \quad E(x_i) = \frac{1}{N}X_1 + \frac{1}{N}X_2 + \ldots + \frac{1}{N}X_N$$

$$= \frac{1}{N}(X_1 + X_2 + \ldots + X_N) = \mu \qquad \ldots(1)$$

$$\therefore \quad E(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n} (\mu) = \frac{1}{n} n\mu \quad \Rightarrow \quad E(\bar{x}) = \mu \qquad \ldots(12\cdot6)$$

Thus the sample mean ($\bar{x}$) is an unbiased estimate of the population mean (μ).

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

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

We have $V(x_i) = E[x_i - E(x_i)]^2 = E(x_i - \mu)^2$, [From (1)]

$$= \frac{1}{N}\left[(X_1 - \mu)^2 + (X_2 - \mu)^2 + \ldots + (X_N - \mu)^2\right] = \sigma^2 \qquad \ldots(3)$$

Also we know that

$$V(x) = E(x^2) - [E(x)]^2 \Rightarrow E(x^2) = V(x) + \{E(x)\}^2 \qquad \ldots(4)$$

In particular

$$E(x_i^2) = V(x_i) + \{E(x_i)\}^2 = \sigma^2 + \mu^2 \qquad \ldots(5)$$

Also from (4), $E(\bar{x}^2) = V(\bar{x}) + \{E(\bar{x})\}^2$

But $V(\bar{x}) = \frac{\sigma^2}{n}$, where σ^2 is the population variance. [*c.f.* § 12·13]

$$\therefore \quad E(\bar{x}^2) = \frac{\sigma^2}{n} + \mu^2 \qquad \text{[Using (12·6)]} \qquad \ldots(5a)$$

Substituting from (5) and (5*a*) in (2) we get

$$E(s^2) = \frac{1}{n}\sum_{i=1}^{n}(\sigma^2+\mu^2) - \left(\frac{\sigma^2}{n}+\mu^2\right)$$

$$= \frac{1}{n}n(\sigma^2+\mu^2) - \left(\frac{\sigma^2}{n}+\mu^2\right) = \left(1-\frac{1}{n}\right)\sigma^2$$

$$= \frac{n-1}{n}\sigma^2 \qquad \ldots(12\cdot7)$$

Since $E(s^2) \neq \sigma^2$, sample variance is not an unbiased estimate of population variance.

From (12·7), we get

$$\frac{n}{n-1}E(s^2) = \sigma^2 \Rightarrow E\left(\frac{ns^2}{n-1}\right) = \sigma^2$$

$$\Rightarrow E\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right] = \sigma^2 \text{ i.e., } E(S^2) = \sigma^2 \qquad \ldots(12\cdot8)$$

where $$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \qquad \ldots(12\cdot8a)$$

$\therefore$ S^2 is an unbiased estimate of the population variance σ^2.

Aliter for $E(s^2)$.

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

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

But $\sum_i (x_i-\mu) = \sum_i x_i - n\mu = n\bar{x} - n\mu = n(\bar{x}-\mu)$

$$\therefore \quad s^2 = \frac{1}{n}\left\{\sum_{i=1}^{n}(x_i-\mu)^2 - n(\bar{x}-\mu)^2\right\} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 - (\bar{x}-\mu)^2$$

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

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

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

Remarks 1. Here we see that although sample mean is an unbiased estimate of population mean, sample variance is not an unbiased estimate of population variance. However, an unbiased estimate of of σ^2 is given by S^2, given in equation (12·8*a*).

S^2 plays a very important role in sampling theory, particularly in small sampling theory. Whenever σ^2 is not known, its estimate S^2 given by (12·8*a*) is used for practical purposes.

2. We have $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

$$\Rightarrow \qquad ns^2 = (n-1)S^2 \qquad \therefore \quad s^2 = \left(1 - \frac{1}{n}\right)S^2$$

Hence for large samples *i.e.*, for $n \to \infty$, we have $s^2 \to S^2$. In other words, for large samples (*i.e.*, $n \to \infty$), we may take

$$\hat{\sigma}^2 = s^2 \qquad \ldots(12\cdot8b)$$

12·12. Standard Error of Sample Mean. *The variance of the sample mean is σ^2/n, where σ is the population standard deviation and n is the size of the random sample.*

The S.E. of mean of a random sample of size n from a population with variance σ^2 is $\sigma/\sqrt{n}$.

Proof. Let x_i, $(i = 1, 2, \ldots, n)$ be a random sample of size n from a population with variance σ^2, then the sample mean $\bar{x}$ is given by

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

$$\therefore \quad V(\bar{x}) = V\left[\frac{1}{n}(x_1 + x_2 + \ldots + x_n)\right] = \frac{1}{n^2}V(x_1 + x_2 + \ldots + x_n)$$

$$= \frac{1}{n^2}\left[V(x_1) + V(x_2) + \ldots + V(x_n)\right],$$

the covariance terms vanish since the sample observations are independent, [*c.f.* Remark (*ii*) § 6·6]

But $V(x_i) = \sigma^2$, $(i = 1, 2, \ldots, n)$ [From (3) of § 12·11]

$$\therefore \quad V(\bar{x}) = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$$

$$\Rightarrow \qquad \text{S.E.}(\bar{x}) = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}} \qquad \ldots(12\cdot9)$$

12·13. Test of Significance for Single Mean. We have proved that if x_i, $(i = 1, 2, \ldots, n)$ is a random sample of size n from a normal population with mean μ and variance σ^2, then the sample mean is distributed normally with mean μ and variance σ^2/n, *i.e.*, $\bar{x} \sim N(\mu, \sigma^2/n)$. However, this result holds, *i.e.*, $\bar{x} \sim N(\mu, \sigma^2/n)$, even in random sampling from non-normal population provided the sample size n is large [*c.f.* Central Limit Theorem, § 8·10].

Thus for large samples, the *standard normal variate* corresponding to $\bar{x}$ is :

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

Under the *null hypothesis*, H_0 that the sample has been drawn from a population with mean μ and variance σ^2, *i.e.*, there is no significant difference between the sample mean ($\bar{x}$) and population mean (μ), the test statistic (for large samples), is :

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \qquad \ldots(12\cdot9a)$$

Remarks 1. If the population s.d. σ is unknown then we use its estimate provided by the sample variance given by [See (12·8*b*)]:

$$\hat{\sigma}^2 = s^2 \quad \Rightarrow \quad \hat{\sigma} = s \text{ (for large samples).}$$

2. Confidence limits for μ. 95% confidence interval for μ is given by :

$$|Z| \leq 1\cdot96, \textit{ i.e.,} \quad \left| \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \right| \leq 1\cdot96$$

$$\Rightarrow \qquad \bar{x} - 1\cdot96\sigma/\sqrt{n} \leq \mu \leq \bar{x} + 1\cdot96\sigma/\sqrt{n} \qquad \ldots(12\cdot10)$$

and $\bar{x} \pm 1\cdot96\sigma/\sqrt{n}$ are known as 95% confidence limits for μ. Similarly, 99% confidence limits for μ are $\bar{x} \pm 2\cdot58\sigma/\sqrt{n}$ and 98% confidence limits for μ are $\bar{x} \pm 2\cdot33\sigma/\sqrt{n}$.

However, in sampling from a finite population of size N, the corresponding 95% and 99% confidence limits for μ are respectively

$$\bar{x} \pm 1\cdot96 \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \text{ and } \bar{x} \pm 2\cdot58 \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \ldots(12\cdot10a)$$

3. The confidence limits for any parameter (P, μ, etc.) are also known as its *fiducial limits.*

Example 12·15. *A sample of 900 members has a mean 3·4 cms,. and s.d. 2·61 cms. Is the sample from a large population of mean 3·25 cms. and s.d. 2·61 cms. ?*

If the population is normal and its mean is unknown, find the 95% and 98% fiducial limits of true mean.

Solution. *Null hypothesis, (H_0)* : The sample has been drawn from the population with mean $\mu = 3\cdot25$ cms., and S.D. $\sigma = 2\cdot61$ cms.

Alternative Hypothesis, $H_1 : \mu \neq 3\cdot25$ (Two-tailed).

Test Statistic. Under H_0, the test statistic is :

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1), \text{ (since } n \text{ is large)}$$

Here, we are given

$$\bar{x} = 3{\cdot}4 \text{ cms.}, n = 900 \text{ cms.}, \mu = 3{\cdot}25 \text{ cms. and } \sigma = 2{\cdot}61 \text{ cms.}$$

$$Z = \frac{3{\cdot}40 - 3{\cdot}25}{2{\cdot}61/\sqrt{900}} = \frac{0{\cdot}15 \times 30}{2{\cdot}61} = 1{\cdot}73$$

Since $|Z| < 1{\cdot}96$, we conclude that the data don't provide us any evidence against the null hypothesis (H_0) which may, therefore, be accepted at 5% level of significance.

95% fiducial limits for the population mean μ are :

$$\bar{x} \pm 1{\cdot}96\,\sigma/\sqrt{n} \Rightarrow 3{\cdot}40 \pm 1{\cdot}96 \times 2{\cdot}61/\sqrt{900}$$

$\Rightarrow$ $3{\cdot}40 \pm 0{\cdot}1705$, *i.e.*, $3{\cdot}5705$ and $3{\cdot}2295$

98% fiducial limits for μ are given by :

$$\bar{x} \pm 2{\cdot}33 \frac{\sigma}{\sqrt{n}}, \text{ i.e., } 3{\cdot}40 \pm 2{\cdot}33 \times \frac{2{\cdot}61}{30}$$

$\Rightarrow$ $3{\cdot}40 \pm 0{\cdot}2027$ *i.e.*, $3{\cdot}6027$ and $3{\cdot}1973$

Remark. 2·33 is the value z_1 of Z from standard normal probability integrals, such that $P(|Z| > z_1) = 0{\cdot}98 \Rightarrow P(Z > z_1) = 0{\cdot}49$.

Example 12·16. *An insurance agent has claimed that the average age of policyholders who insure through him is less than the average for all agents, which is 30·5 years.*

A random sample of 100 policyholders who had insured through him gave the following age distribution :

Age last birthday	*No. of persons*
16—20	*12*
21—25	*22*
26—30	*20*
31—35	*30*
36—40	*16*

Calculate the arithmetic mean and standard deviation of this distribution and use these values to test his claim at the 5% level of significance. You are given that Z (1·645) = 0·95.

Solution. *Null Hypothesis*, $H_0 : \mu = 30{\cdot}5$ years, *i.e.*, the sample mean ($\bar{x}$) and population mean (μ) do not differ significantly.

Alternative Hypothesis, $H_1 : \mu < 30{\cdot}5$ years (Left-tailed alternative).

CALCULATIONS FOR SAMPLE MEAN AND S.D.

Age last birthday	*No. of persons (f)*	*Mid-point x*	$d = \frac{x - 28}{5}$	*fd*	fd^2
16—20	12	18	−2	−24	48
21—25	22	23	−1	−22	22
26—30	20	28	0	0	0
31—35	30	33	1	30	30
36—40	16	38	2	32	64
Total	$N = 100$			$\Sigma fd = 16$	$\Sigma fd^2 = 164$

$$\bar{x} = 28 + \frac{5 \times 16}{100} = 28{\cdot}8 \text{ years} \quad s = 5 \times \sqrt{\frac{164}{100} - \left(\frac{16}{100}\right)^2} = 6{\cdot}35 \text{ years}$$

Since the sample is large, $\hat{\sigma} \simeq s = 6{\cdot}35$ years.

Test Statistic. Under H_0, the test statistic is

$$Z = \frac{\bar{x} - \mu}{\sqrt{s^2/n}} \sim N(0, 1), \text{ (since sample is large).}$$

Now $$Z = \frac{28{\cdot}8 - 30{\cdot}5}{6{\cdot}35/\sqrt{100}} = \frac{-1{\cdot}7}{0{\cdot}635} = -2{\cdot}681$$

Conclusion. Since computed value of $Z = -2{\cdot}681 < -1{\cdot}645$ or $|Z| = 2{\cdot}681 > 1{\cdot}645$, it is significant at 5% level of significance. Hence we reject the null hypothesis H_0 (Accept H_1) at 5% level of significance and conclude that the insurance agent's claim that the average age of policyholders who insure through him is less than the average for all agents, is valid.

Example 12·17. *As an application of Central Limit Theorem, show that if E is such that $P(|\bar{X} - \mu| < E) > 0{\cdot}95$, then the minimum sample size n is given by $n = \frac{(1{\cdot}96)^2 \sigma^2}{E^2}$, where μ and σ^2 are the mean and variance respectively of the population and $\bar{X}$ is the mean of the random sample.*

Solution. By Central Limit Theorem, we know that $\bar{X} \sim N(\mu, \sigma^2/n)$ asymptotically *i.e.*, for large n.

$$\therefore \quad Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1), \text{ asymptotically } \textit{i.e.}, \text{ for large } n.$$

From normal probability tables, we have

$$P(|Z| \leq 1{\cdot}96) = 0{\cdot}95$$

$$\Rightarrow \quad P\left[\left|\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right| \leq 1{\cdot}96\right] = 0{\cdot}95$$

$$\Rightarrow \quad P\left[|\bar{X} - \mu| \leq 1{\cdot}96 \frac{\sigma}{\sqrt{n}}\right] = 0{\cdot}95 \qquad \ldots(*)$$

We are given that

$$P[|\bar{X} - \mu| < E] > 0{\cdot}95 \qquad \ldots(**)$$

From (*) and (**), we have

$$E > \frac{1{\cdot}96\sigma}{\sqrt{n}} \Rightarrow n > \frac{(1{\cdot}96)^2 \sigma^2}{E^2} = \frac{3{\cdot}84\sigma^2}{E^2}$$

Hence minimum sample size n for estimating μ with 95% confidence coefficient is given by $n = 3{\cdot}84\ \sigma^2/E^2$, where E is the permissible error.

Remark. The minimum sample size for estimating μ with confidence coefficient $(1 - \alpha)$ is given by $\sigma^2 z_\alpha^2/E^2$, where z_α is the significant value of Z at level of significance α and E is the permissible error in the estimate.

Arguing similarly, the minimum sample size for estimating population proportion P with confidence coefficient $(1 - \alpha)$ is given by $n = PQ\, z_\alpha^2/E^2$, where z_α is the significant value of Z at 'α' level of significance and E is the permissible error in the estimate. If P is unknown, we may use $\hat{P} = p$.

Example 12·18. *The mean muscular endurance score of a random sample of 60 subjects was found to be 145 with a s.d. of 40. Construct a 95% confidence interval for the true mean. Assume the sample size to be large enough for normal approximation. What size of sample is required to estimate the mean within 5 of the true mean with a 95% confidence ?*

[*Calicut Univ. B.Sc. (Main Stat.) 1989*]

Solution. We are given : $n = 60$, $\bar{x} = 145$ and $s = 40$.

95% confidence limits for true mean (μ) are :

$$\bar{x} \pm 1{\cdot}96\, s/\sqrt{n} \qquad (\sigma^2 = s^2, \text{ since sample is large})$$

$$= 145 \pm \frac{1{\cdot}96 \times 40}{\sqrt{60}} = 145 \pm \frac{78{\cdot}4}{7{\cdot}75} = 145 \pm 10{\cdot}12 = 134{\cdot}88,\ 155{\cdot}12$$

Hence 95% confidence interval for μ is (134·88, 155·12). In the notations of Example 12·17, we have

$$n = \left(\frac{z_\alpha \cdot \sigma}{E}\right)^2 = \left(\frac{1{\cdot}96 \times 40}{5}\right)^2$$

$$[\because z_{0{\cdot}05} = 1{\cdot}96,\ \hat{\sigma} = s = 40 \text{ and } |\bar{x} - \mu| < 5 = E]$$

$$= (15{\cdot}68)^2 = 245{\cdot}86 \simeq 246.$$

Example 12·19. *The standard deviation of a population is 2·70 inches. Find the probability that in a random sample of size 66 (i) the sample mean will differ from the population mean by 0·75 inch or more and (ii) the sample mean will exceed the population mean by 0·75 inch or more (given that the value of the standard normal probability integral from 0 to 2·25 is 0·4877).*

Solution. Here we are given $n = 66$, $\sigma = 2{\cdot}70$ inches. Since n is large, the sample mean $\bar{x} \sim N(\mu, \sigma^2/n)$.

$$\therefore \qquad Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1) \qquad \ldots(*)$$

We want

(*i*) $$P[|\bar{x} - \mu| \geq 0{\cdot}75] = 1 - P[|\bar{x} - \mu| < 0{\cdot}75]$$

$$= 1 - P\left[\left|\frac{\sigma}{\sqrt{n}} Z\right| < 0{\cdot}75\right] \qquad \text{[From (*)]}$$

$$= 1 - P\left[|Z| < 0{\cdot}75\, \frac{\sqrt{n}}{\sigma}\right]$$

$$= 1 - 2P\left[0 < Z < 0{\cdot}75\, \frac{\sqrt{n}}{\sigma}\right]$$

$$= 1 - 2P\left[0 < Z < 0{\cdot}75 \times \frac{\sqrt{66}}{2{\cdot}70}\right]$$

$$= 1 - 2P\left[0 < Z < \frac{0{\cdot}75 \times 8{\cdot}124}{2{\cdot}70}\right]$$

$$= 1 - 2P[0 < Z < 2{\cdot}25] = 1 - 2 \times 0{\cdot}4877 = 0{\cdot}0246$$

(*ii*) $P[\bar{x} - \mu > 0{\cdot}75] = P(Z > 0{\cdot}75\sqrt{n}/\sigma) = P(Z > 2{\cdot}25)$

$$= 0{\cdot}5 - P(0 < Z < 2{\cdot}25) = 0{\cdot}5 - 0{\cdot}4877 = 0{\cdot}0123$$

Example 12·20. *A normal population has a mean of 0·1 and standard deviation of 2·1. Find the probability that mean of a sample of size 900 will be negative.* ***[Delhi Univ. B.Sc. (Stat. Hons.), 1986]***

Solution. Here we are given that $X \sim N(\mu, \sigma^2)$, where $\mu = 0{\cdot}1$ and $\sigma = 2{\cdot}1$ and $n = 900$.

Since $X \sim N(\mu, \sigma^2)$, the sample mean $\bar{x} \sim N(\mu, \sigma^2/n)$. The standard normal variate corresponding to $\bar{x}$ is given by :

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{x} - 0{\cdot}1}{2{\cdot}1/30} = \frac{\bar{x} - 0{\cdot}1}{0{\cdot}07}$$

$\therefore$ $\bar{x} = 0{\cdot}1 + 0{\cdot}07Z$, where $Z \sim N(0, 1)$

The required probability p, that the sample mean is negative is given by :

$$p = P(\bar{x} < 0) = P(0{\cdot}1 + 0{\cdot}07\,Z < 0)$$

$$= P\left(Z < \frac{-0{\cdot}10}{0{\cdot}07}\right) = P(Z < -1{\cdot}43) = P(Z \geq 1{\cdot}43)$$

$$= 0{\cdot}5 - P(0 < Z < 1{\cdot}43) = 0{\cdot}5 - 0{\cdot}4236 = 0{\cdot}0764$$

(From Normal Probability Tables)

Example 12·21. *The guaranteed average life of a certain type of electric light bulbs is 1000 hours with a standard deviation of 125 hours. It is decided to sample the output so as to ensure that 90 per cent of the bulbs do not fall short of the guaranteed average by more than 2·5 per cent. What must be the minimum size of the sample ?* ***[Madras Univ. B.Sc., Oct. 1991]***

Solution. Here $\mu = 1000$ hours, $\sigma = 125$ hours.

Since we do not want the sample mean to be less than the guaranteed average mean ($\mu = 1000$) by more than 2·5%, we should have

$$\bar{x} > 1000 - 2{\cdot}5\% \text{ of } 1000 \Rightarrow \bar{x} > 1000 - 25 = 975$$

Let n be the given sample size. Then

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1), \text{ since sample is large.}$$

We want $$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} > \frac{975 - 1000}{125/\sqrt{n}} > -\frac{\sqrt{n}}{5} \qquad (\because \bar{x} > 975)$$

According to the given condition, we have

$$P(Z > -\sqrt{n/5}) = 0{\cdot}90 \Rightarrow P(0 < Z < \sqrt{n/5}) = 0{\cdot}40$$

$\therefore$ $\sqrt{n/5} = 1{\cdot}28$ (From Normal Probability Tables)

$\Rightarrow$ $n = 25 \times (1{\cdot}28)^2 = 41$ (approx)

Example 12·22. *A survey is proposed to be conducted to know the annual earnings of the old Statistics graduates of Delhi University. How large should the sample be taken in order to estimate the mean annual earnings within plus and minus Rs. 1,000 at 95% confidence level ? The standard deviation of the annual earnings of the entire population is known to be Rs. 3,000.*

Solution. We are given : σ = Rs. 3,000.

We want : $P[|\bar{x} - \mu| < 1{,}000] = 0{\cdot}95$ ...(*)

We know that, in sampling from normal population or for large samples from any population $\bar{X} \sim N(\mu, \sigma^2/n)$. Hence from normal probability tables, we have :

$$P[|Z| \leq 1{\cdot}96] = 0{\cdot}95$$

$$\Rightarrow \quad P\left[\left|\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq 1{\cdot}96\right|\right] = 0{\cdot}95$$

$$\Rightarrow \quad P[|\bar{x} - \mu| \leq 1{\cdot}96 \times (\sigma/\sqrt{n})] = 0{\cdot}95 \qquad \ldots(**)$$

From (*) and (**), we get

$$\frac{1{\cdot}96 \times \sigma}{\sqrt{n}} = 1000 \Rightarrow \frac{1{\cdot}96 \times 3000}{\sqrt{n}} = 1000$$

$\therefore$ $n = (1{\cdot}96 \times 3)^2 = (5{\cdot}88)^2 = 34{\cdot}56 \simeq 35$

Aliter. Using Remark to Example 12·17,

$$n = \left(\frac{z_\alpha \cdot \sigma}{E}\right)^2 = \left(\frac{1{\cdot}96 \times 3{,}000}{1{,}000}\right)^2 \simeq 35.$$

12·14. Test of Significance for Difference of Means. Let $\bar{x}_1$ be the mean of a random sample of size n_1 from a population with mean μ_1 and variance σ_1^2 and let $\bar{x}_2$ be the mean of an independent random sample of size n_2 from another population with mean μ_2 and variance σ_2^2. Then, since sample sizes are large,

$$\bar{x}_1 \sim N(\mu_1, \sigma_1^2/n_1) \text{ and } \bar{x}_2 \sim N(\mu_2, \sigma_2^2/n_2)$$

Also $\bar{x}_1 - \bar{x}_2$, being the difference of two independent normal variates is also a normal variate. The Z (*S.N.V.*) corresponding to $\bar{x}_1 - \bar{x}_2$ is given by

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - E(\bar{x}_1 - \bar{x}_2)}{S.E.\ (\bar{x}_1 - \bar{x}_2)} \sim N(0, 1)$$

Under the null hypothesis $H_0 : \mu_1 = \mu_2$, *i.e.*, there is no significant difference between the sample means, we get

$$E(\bar{x}_1 - \bar{x}_2) = E(\bar{x}_1) - E(\bar{x}_2) = \mu_1 - \mu_2 = 0;$$

$$V(\bar{x}_1 - \bar{x}_2) = V(\bar{x}_1) + V(\bar{x}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2},$$

the covariance term vanishes, since the sample means $\bar{x}_1$ and $\bar{x}_2$ are independent.

Thus under $H_0 : \mu_1 = \mu_2$, the test statistic becomes (for large samples),

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}} \sim N(0, 1) \qquad \ldots(12\cdot11)$$

Remarks 1. If $\sigma_1^2 = \sigma_2^2 = \sigma^2$, *i.e.*, if the samples have been drawn from the populations with common S.D. σ, then under $H_0 : \mu_1 = \mu_2$,

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{(1/n_1 + 1/n_2)}} \sim N(0, 1) \qquad \ldots[12\cdot11(a)]$$

2. If in (12·11*a*), σ is not known, then its estimate based on the sample variances is used. If the sample sizes are not sufficiently large, then an unbiased estimate of σ^2 is given by

$$\hat{\sigma}^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)},$$

since

$$E(\hat{\sigma}^2) = \frac{1}{n_1 + n_2 - 2}\left[(n_1 - 1)\,E(S_1^2) + (n_2 - 1)\,E(S_2^2)\right]$$

$$= \frac{1}{n_1 + n_2 - 2}\left[(n_1 - 1)\,\sigma^2 + (n_2 - 1)\,\sigma^2\right] = \sigma^2$$

But since sample sizes are large, $S_1^2 \simeq s_1^2$, $S_2^2 \simeq s_2^2$, $n_1 - 1 \simeq n_1$, $n_2 - 1 \simeq n_2$. Therefore in practice, for large samples, the following estimate of σ^2 without any serious error is used :

$$\hat{\sigma}^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2} \qquad \ldots[12\cdot11(b)]$$

However, if sample sizes are small, then a small sample test, t-test for difference of means (*c.f.* Chapter 14) is to be used.

3. If $\sigma_1^2 \neq \sigma_2^2$ and σ_1 and σ_2 are not known, then they are estimated from sample values. This results in some error, which is practically immaterial, if samples are large. These estimates for large samples are given by

$$\left.\begin{aligned}\hat{\sigma}_1^2 &= S_1^2 \simeq s_1^2 \\ \hat{\sigma}_2^2 &= S_2^2 \simeq s_2^2\end{aligned}\right\} \quad \text{(since samples are large).}$$

In this case, (12·11) gives

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}} \sim N(0, 1) \qquad \ldots[12\cdot11(c)]$$

Example 12·23. *The means of two single large samples of 1000 and 2000 members are 67·5 inches and 68·0 inches respectively. Can the samples be regarded as drawn from the same population of standard deviation 2·5 inches ? (Test at 5% level of significance).*

Solution. We are given :

$$n_1 = 1000,\ n_2 = 2000;\ \bar{x}_1 = 67{\cdot}5 \text{ inches},\ \bar{x}_2 = 68{\cdot}0 \text{ inches}.$$

Null hypothesis, $H_0 : \mu_1 = \mu_2$ and $\sigma = 2{\cdot}5$ inches, *i.e.,* the samples have been drawn from the same population of standard deviation 2·5 inches.

Alternative Hypothesis, $H_1 : \mu_1 \neq \mu_2$ (Two tailed.)

Test Statistic. Under H_0, the test statistic is (since samples are large)

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1)$$

$$\text{Now} \quad Z = \frac{67{\cdot}5 - 68{\cdot}0}{2{\cdot}5 \times \sqrt{\frac{1}{1000} + \frac{1}{2000}}} = \frac{-0{\cdot}5}{2{\cdot}5 \times 0{\cdot}0387} = -5{\cdot}1$$

Conclusion. Since $|Z| > 3$, the value is highly significant and we reject the null hypothesis and conclude that samples are certainly not from the same population with standard deviation 2·5.

Example 12·24. *In a survey of buying habits, 400 women shoppers are chosen at random in super market 'A' located in a certain section of the city. Their average weekly food expenditure is Rs. 250 with a standard deviation of Rs. 40. For 400 women shoppers chosen at random in super market 'B' in another section of the city, the average weekly food expenditure is Rs. 220 with a standard deviation of Rs. 55. Test at 1% level of significance whether the average weekly food expenditure of the two populations of shoppers are equal.*

Solution. In the usual notations, we are given that

$$n_1 = 400, \qquad \bar{x}_1 = \text{Rs. } 250, \qquad s_1 = \text{Rs. } 40$$

$$n_2 = 400, \qquad \bar{x}_2 = \text{Rs. } 220 \qquad s_2 = \text{Rs. } 55$$

Null hypothesis, $H_0 : \mu_1 = \mu_2$, *i.e.,* the average weekly food expenditures of the two populations of shoppers are equal.

Alternative Hypothesis, $H_1 : \mu_1 \neq \mu_2$. (Two-tailed)

Test Statistic. Since samples are large, under H_0, the test statistic is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}} \sim N(0, 1)$$

Since σ_1 and σ_2, the population standard deviations are not known, we can take for large samples (*c.f.* § 12·15, Remark 3) :

$$\hat{\sigma}_1^2 = s_1^2 \quad \text{and} \quad \hat{\sigma}_2^2 = s_2^2$$

and then Z is given by

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} = \frac{250 - 220}{\sqrt{\left\{\frac{(40)^2}{400} + \frac{(55)^2}{400}\right\}}} = 8{\cdot}82 \text{ (approx.)}$$

Conclusion. Since $|Z|$ is much greater than 2·58, the null hypothesis $(\mu_1 = \mu_2)$ is rejected at 1% level of significance and we conclude that the average weekly expenditures of two populations of shoppers in markets *A* and *B* differ significantly.

Example 12·25. *The average hourly wage of a sample of 150 workers in a plant 'A' was Rs. 2·56 with a standard deviation of Rs. 1·08. The average wage of a sample of 200 workers in plant 'B' was Rs. 2·87 with a standard deviation of Rs. 1·28. Can an applicant safely assume that the hourly wages paid by plant 'B' are higher than those paid by plant 'A' ?*

Solution. Let X_1 and X_2 denote the hourly wages (in *Rs.*) of workers in plant *A* and plant *B* respectively. Then we are given :

$$n_1 = 150,\ \bar{x}_1 = 2{\cdot}56,\ s_1 = 1{\cdot}08 = \hat{\sigma}_1$$

$$n_2 = 200,\ \bar{x}_2 = 2{\cdot}87,\ s_2 = 1{\cdot}28 = \hat{\sigma}_2$$

Null hypothesis, $H_0 : \mu_1 = \mu_2$, *i.e.,* there is no significant difference between the mean level of wages of workers in plant *A* and plant *B*.

Alternative hypothesis, $H_1 : \mu_2 > \mu_1$ *i.e.,* $\mu_1 < \mu_2$ (Left-tailed test)

Test Statistic. Under H_0, the test statistic (for large samples) is :

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} \sim N(0, 1)$$

$$\therefore \quad Z = \frac{2{\cdot}56 - 2{\cdot}87}{\sqrt{\left\{\frac{(1{\cdot}08)^2}{150} + \frac{(1{\cdot}28)^2}{200}\right\}}} = \frac{-0{\cdot}31}{\sqrt{0{\cdot}016}} = \frac{-0{\cdot}31}{0{\cdot}126} = -2{\cdot}46.$$

Critical region. For a one-tailed test, the critical value of *Z* at 5% level of significance is 1·645. The critical region for left-tailed test thus consists of all values of $Z \leq -1{\cdot}645$.

Conclusion. Since calculated value of *Z* (–2·46) is less than critical value (–1·645), it is significant at 5% level of significance. Hence the null hypothesis is rejected at 5% level of significance and we conclude that the average hourly wages paid by plant '*B*' are certainly higher than those paid by plant '*A*'.

Example 12·26. *In a certain factory there are two independent processes manufacturing the same item. The average weight in a sample of 250 items produced from one process is found to be 120 ozs. with a standard deviation of 12 ozs. while the corresponding figures in a sample of 400 items from the other process are 124 and 14. Obtain the standard error of difference between the two sample means. Is this difference significant ? Also find the 99% confidence limits for the difference in the average weights of items produced by the two processes respectively.*

Solution. We have

$$\left.\begin{aligned} n_1 &= 250,\ \bar{x}_1 = 120\ oz.,\ s_1 = 12\ \text{oz.} = \hat{\sigma}_1 \\ n_2 &= 400,\ \bar{x}_2 = 124\ \text{oz.},\ s_2 = 14\ \text{oz.} = \hat{\sigma}_2 \end{aligned}\right\} \text{(since samples are large).}$$

$$S.E.\ (\bar{x}_1 - \bar{x}_2) = \sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)} = \sqrt{(s_1^2/n_1) + (s_2^2/n_2)}$$

$$= \sqrt{\left(\frac{144}{250} + \frac{196}{400}\right)} = \sqrt{(0{\cdot}576 + 0{\cdot}490)} = 1{\cdot}034$$

Null Hypothesis, $H_0 : \mu_1 = \mu_2$, *i.e.*, the sample means do not differ significantly.

Alternative Hypothesis, $H_1 : \mu_1 \neq \mu_2$ (Two-tailed).

Test Statistic. Under H_0, the test statistic is :

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{S.E.\ (\bar{x}_1 - \bar{x}_2)} = \frac{120 - 124}{1{\cdot}034} \sim N\ (0, 1)$$

$$\therefore \qquad |Z| = \frac{4}{1{\cdot}034} = 3{\cdot}87$$

Conclusion. Since $|Z| > 3$, the null hypothesis is rejected and we conclude that there is significant difference between the sample means.

99% confidence limits for $|\mu_1 - \mu_2|$, *i.e.*, for the difference in the average weights of items produced by two processes, are

$$|\bar{x}_1 - \bar{x}_2| \pm 2{\cdot}58\ S.E.\ (\bar{x}_1 - \bar{x}_2) = 4 \pm 2{\cdot}58 \times 1{\cdot}034$$

$$= 4 \pm 2{\cdot}67 \text{ (approx.)} = 6{\cdot}67 \text{ and } 1{\cdot}33$$

$$\therefore \qquad 1{\cdot}33 < |\mu_1 - \mu_2| < 6{\cdot}67$$

Example 12·27. *The mean height of 50 male students who showed above average participation in college athletics was 68·2 inches with a standard deviation of 2·5 inches; while 50 male students who showed no interest in such participation had a mean height of 67·5 inches with a standard deviation of 2·8 inches.*

(i) Test the hypothesis that male students who participate in college athletics are taller than other male students.

(ii) By how much should the sample size of each of the two groups be increased in order that the observed difference of 0·7 inches in the mean heights be significant at the 5% level of significance.

Solution. Let X_1 and X_2 denote the height (in inches) of athletic participants and non-athletic participants respectively. In the usual notations, we are given :

$$n_1 = 50,\ \bar{x}_1 = 68{\cdot}2,\ s_1 = 2{\cdot}5\ ;\ n_2 = 50,\ \bar{x}_2 = 67{\cdot}5,\ s_2 = 2{\cdot}8$$

Null hypothesis, $H_0 : \mu_1 = \mu_2$.

Alternative hypothesis, $H_1 : \mu_1 > \mu_2$ (Right-tailed).

Test Statistic. Under H_0, the test statistic for large samples is :

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} \sim N\ (0, 1)$$

$$\therefore \qquad Z = \frac{68{\cdot}2 - 67{\cdot}5}{\sqrt{\left\{\frac{(2{\cdot}5)^2}{50} + \frac{(2{\cdot}8)^2}{50}\right\}}} = \frac{0{\cdot}7}{\sqrt{0{\cdot}282}} = \frac{0{\cdot}7}{0{\cdot}53} = 1{\cdot}32$$

For a right-tailed test, the critical (significant) value of Z at 5% level of significance is 1·645.

(*i*) Since the calculated value of $Z(1{\cdot}32)$ is less than the critical value (1·645), it is not significant at 5% level of significance. Hence the null hypothesis is accepted and we conclude that the college athletes are not taller than other male students.

(*ii*) The difference between the mean heights of two groups, each of size n will be significant at 5% level of significance if $Z \geq 1{\cdot}645$

$$\Rightarrow \quad \frac{68{\cdot}2 - 67{\cdot}5}{\sqrt{\left\{\frac{(2{\cdot}5)^2}{n} + \frac{(2{\cdot}8)^2}{n}\right\}}} \geq 1{\cdot}645$$

$$\Rightarrow \quad \frac{0{\cdot}7}{\sqrt{14{\cdot}09/n}} \geq 1{\cdot}645 \Rightarrow \frac{0{\cdot}7}{3{\cdot}754/\sqrt{n}} \geq 1{\cdot}645$$

$$\Rightarrow \quad n \geq \left(\frac{1{\cdot}645 \times 3{\cdot}754}{0{\cdot}7}\right)^2 = (8{\cdot}8219)^2 = 77{\cdot}83 \simeq 78$$

Hence the sample size of each of the two groups should be increased by at least 78 – 50 = 28, in order that the difference between the mean heights of the two groups is significant.

12·15. Test of Significance for the Difference of Standard Deviations. If s_1 and s_2 are the standard deviations of two independent samples, then under null hypothesis, $H_0 : \sigma_1 = \sigma_2$, *i.e.*, *i.e.*, the sample standard deviations don't differ significantly, the statistic

$$Z = \frac{s_1 - s_2}{\text{S.E.}\,(s_1 - s_2)} \sim N(0, 1) \text{ for large samples.}$$

But in case of large samples, the S.E of the difference of the sample standard deviations is given by

$$\text{S.E.}\,(s_1 - s_2) = \sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}$$

$$\therefore \quad Z = \frac{s_1 - s_2}{\sqrt{\left(\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}\right)}} \sim N(0, 1) \qquad \ldots(12{\cdot}12)$$

σ_1^2 and σ_2^2 are usually unknown and for large samples, we use their estimates given by the corresponding sample variances. Hence the test statistic reduces to

$$Z = \frac{s_1 - s_2}{\sqrt{\left(\frac{s_1^2}{2n_1} + \frac{s_2^2}{2n_2}\right)}} \sim N(0, 1) \qquad \ldots(12{\cdot}13)$$

Example 12·18. *Random samples drawn from two countries gave the following data relating to the heights of adult males* :

	Country A	*Country B*
Mean height (in inches)	*67·42*	*67·25*
Standard deviation (in inches)	*2·58*	*2·50*
Number in samples	*1000*	*1200*

(i) Is the difference between the means significant ?

(ii) Is the difference between the standard deviations significant ?

Solution. We are given :

$$n_1 = 1000, \quad \bar{x}_1 = 67{\cdot}42 \text{ inches}, \quad s_1 = 2{\cdot}58 \text{ inches},$$

$$n_2 = 1200, \quad \bar{x}_2 = 67{\cdot}25 \text{ inches}, \quad s_2 = 2{\cdot}50 \text{ inches}.$$

As in the last examples (since sample sizes are large), we can take

$$\hat{\sigma}_1 = s_1 = 2{\cdot}58, \quad \hat{\sigma}_2 = s_2 = 2{\cdot}50$$

(*i*) $H_0 : \mu_1 = \mu_2$, *i.e.*, the sample means do not differ significantly.

$$H_1 : \mu_1 \neq \mu_2 \text{ (Two tailed).}$$

Under the *Null hypothesis* H_0, the test statistic is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}} \sim N(0, 1), \text{ since samples are large.}$$

Now $$Z = \frac{67{\cdot}42 - 67{\cdot}25}{\sqrt{\left\{\frac{(2{\cdot}58)^2}{1000} + \frac{(2{\cdot}50)^2}{1200}\right\}}} = \frac{0{\cdot}17}{\sqrt{\left(\frac{6{\cdot}66}{1000} + \frac{6{\cdot}25}{1200}\right)}} = 1{\cdot}56$$

Conclusion. Since $|Z| < 1{\cdot}96$, null hypothesis may be accepted at 5% level of significance and we may conclude that there is no significant difference between the sample means.

(*ii*) Under H_0 : that there is no significant difference between sample standard deviations,

$$Z = \frac{s_1 - s_2}{\text{S.E. } (s_1 - s_2)} \sim N(0, 1), \text{ since samples are large.}$$

$$\text{Now } S.E. (s_1 - s_2) = \sqrt{\left(\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}\right)} = \sqrt{\left(\frac{s_1^2}{2n_1} + \frac{s_2^2}{2n_2}\right)},$$

if σ_1 and σ_2 are not known and $\hat{\sigma}_1 = s_1$, $\hat{\sigma}_2 = s_2$.

$$\therefore \quad S.E. (s_1 - s_2) = \sqrt{\left\{\frac{(2{\cdot}58)^2}{2 \times 1000} + \frac{(2{\cdot}50)^2}{2 \times 1200}\right\}} = 0{\cdot}07746$$

Hence $$Z = \frac{2{\cdot}58 - 2{\cdot}50}{0{\cdot}07746} = \frac{0{\cdot}08}{0{\cdot}07746} = 1{\cdot}03$$

Conclusion. Since $|Z| < 1{\cdot}96$, the data don't provide us any evidence against the null hypothesis which may be accepted at 5% level of significance. Hence the sample standard deviations do not differ significantly.

Example 12·29. *Two populations have their means equal, but S.D. of one is twice the other. Show that in the samples of size 2000 from each drawn under simple sampling conditions, the difference of means will, in all probability, not exceed $0{\cdot}15\sigma$, where σ is the smaller S.D. What is the probability that the difference will exceed half this amount ?*

Solution. Let the standard deviations of the two populations be σ and 2σ respectively and let μ be the mean of each of the two populations. Also we are given $n_1 = n_2 = 2000$. If $\bar{x}_1$ and $\bar{x}_2$ be the two sample means then, since samples are large,

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - E(\bar{x}_1 - \bar{x}_2)}{S.E.\ (\bar{x}_1 - \bar{x}_2)} \sim N(0, 1)$$

Now $E(\bar{x}_1 - \bar{x}_2) = E(\bar{x}_1) - E(\bar{x}_2) = \mu - \mu = 0$ and

$$S.E.\ (\bar{x}_1 - \bar{x}_2) = \sqrt{\left\{\frac{\sigma^2}{n_1} + \frac{(2\sigma)^2}{n_2}\right\}} = \sigma.\sqrt{\left(\frac{1}{2000} + \frac{4}{2000}\right)} = 0{\cdot}05\sigma$$

$$\therefore \qquad Z = \frac{\bar{x}_1 - \bar{x}_2}{S.E.\ (\bar{x}_1 - \bar{x}_2)} \sim N(0, 1)$$

Under simple sampling conditions, we should in all probability have

$$|Z| < 3 \Rightarrow |\bar{x}_1 - \bar{x}_2| < 3\ S.E.\ (\bar{x}_1 - \bar{x}_2)$$

$$\Rightarrow \qquad |\bar{x}_1 - \bar{x}_2| < 0{\cdot}15\sigma,$$

which is the required result.

We want $p = P\left[|\bar{x}_1 - \bar{x}_2| > \frac{1}{2} \times 0{\cdot}15\sigma\right]$

$$\therefore \quad p = P\left[0{\cdot}05\sigma\,|Z| > 0{\cdot}075\sigma\right] \quad \left[\because\ Z = \frac{\bar{x}_1 - \bar{x}_2}{0{\cdot}05\sigma} \sim N(0, 1)\right]$$

$$= P[|Z| > 1{\cdot}5] = 1 - P[|Z| \leq 1{\cdot}5]$$

$$= 1 - 2P(0 \leq Z \leq 1{\cdot}5) = 1 - 2 \times 0{\cdot}4332 = 0{\cdot}1336$$

EXERCISE 12·2

1. Define sampling distribution and standard error. Obtain standard error of mean when population is large.

2. Find the standard error of a linear function of a number of variables. Deduce the standard error of the mean of n uncorrelated variables following the same distribution.

3. Derive the expressions for the standard error of

(*i*) the mean of a random sample of size n, and

(*ii*) the difference of the means of two independent random samples of sizes n_1 and n_2.

4. (*a*) What is meant by a statistical hypothesis ? What are the two types of errors of decision that arise in testing a hypothesis ? Briefly explain how a statistical hypothesis is tested.

The manufacturer of television tubes knows from past experience that the

average life of a tube is 2,000 hours with a standard deviation of 200 hours. A sample of 100 tubes has an average life of 1950 hours. Test at the 0·05 level of significance if this sample came from a normal population of mean 2,000 hours.

State your null and alternative hypothesis and indicate clearly whether a one-tail or a two-tail test is used and why ? Is the result of the test significant ?

[Calcutta Univ. B.Sc. (Maths. Hons.), 1990]

(*b*) A sample of 100 items, drawn from a universe with mean value 64 and S.D. 3 has a mean value 63·5. Is the difference in the means significant ? What will be your inference, if the sample had 200 items ?

[Madras Univ. B.E., Nov. 1990]

(*c*) A sample of 400 individuals is found to have a mean height of 67·47 inches. Can it be reasonably regarded as a sample from a large population with mean height of 67·39 inches and standard deviation 1·30 inches ?

Ans. Yes, $Z = 1{\cdot}23$.

(*d*) The mean breaking strength of cables supplied by a manufacturer is 1800 with a standard deviation 100. By a new technique in the manufacturing process it is claimed that the breaking strength of the cables has increased. In order to test this claim a sample of 50 cables is tested. It is found that the mean breaking strength is 1850. Can we support the claim at 0·01 level of significance ?

Ans. $H_0 : \mu = 1800$, $H_1 : \mu > 1800$, $Z = 3{\cdot}535$.

(*e*) An ambulance service claims that it takes on the average 8·9 minutes to reach its destination in emergency calls. To check on this claim, the agency which licenses ambulance services has them timed on 50 emergency calls, getting a mean of 9·3 minutes with a standard deviation of 1·6 minutes. What can they conclude at the level of significance $\alpha = 0{\cdot}05$?

Ans. $Z = 1{\cdot}768$.

(*f*) A paper mill in U.P. has agreed to buy waste paper for recycling from a waste collection firm, under the agreement that the waste collection firm will supply the waste paper in packages of 300 kg each, for which the paper mill will pay by the package. To speed up their work the waste collection firm is making packages by some *approximation* procedure. The paper mill does not object to this procedure as long as it gets 300 kg. per package on the average. The waste collection firm has an interest not to exceed 300 kg. per package, because it is not being paid for more, and not to go under 300 kg. because the paper mill might terminate the agreement if it does. To estimate the mean weight of waste paper per package, the waste collection firm weighed 75 randomly selected packages and found that the mean weight was 290 kg and standard deviation was 15 kg. Can we infer that the mean weight per package in the entire supply was 300 kg ? ***[Delhi Univ. M.A. (Eco.), 1987]***

Ans. $H_0 : \mu = 300$ kg; $H_1 : \mu \neq 300$ kg. (Two-tailed).

$$Z = \frac{290 - 300}{15/\sqrt{75}} = 5{\cdot}77 \text{ ; Significant.}$$

(*g*) The wages of a factory's workers are assumed to be normally distributed with mean μ and variance 25. A random sample of 25 workers gives the total wages equal to 1250 units.

Test the hypothesis : $\mu = 52$, against the alternative : $\mu = 49$, at 1% level of significance.

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-2\cdot32} \exp\left(-\frac{1}{2}u^2\right) du = 0\cdot01.$$

[***Calcutta Univ. B.Sc.(Maths. Hons.), 1988***]

Ans. $H_0 : \mu = 52$, $H_1 : \mu = 49 < 52$, (Left-tailed test).

$Z = -2$, Not significant.

5. (*a*) A sample of 450 items is taken from a population whose standard deviation is 20. The mean of the sample is 30. Test whether the sample has come from a population with mean 29. Also calculate the 95% confidence limits for the population mean.

(*b*) A sample of 400 observations has mean 95 and standard deviation 12. Could it be a random sample from a population with mean 98 ? What can be the maximum value of the population mean ?

6. (*a*) If the mean age at death of 64 men engaged in an occupation is 52·4 years with standard deviation of 10·2 years, what are the 98% confidence limits for the mean age of all men in that population ?

[***Calicut Univ. B.Sc. (Subs.), 1989***]

(*b*) The weights of 1500 ball bearings are normally distributed with mean 22·40 and standard deviation 0·048. If 300 random samples of size 36 each are drawn from this population, determine the expected mean and standard deviation of the sampling distribution of means, if sampling is done with replacement.

How many of the random samples in the above problem would have their means between 22·39 and 22·41 ? [***Madras Univ. B.E., April 1989***]

Hint. $E(\bar{X}) = \mu = 22\cdot40$; $S.E.(\bar{X}) = \sigma/\sqrt{n} = 0\cdot048/\sqrt{36} = 0\cdot008$

Required number of samples (out of 300) is : $300 \times P(22\cdot39 < \bar{X} < 22\cdot41)$

$$= 300 \times P\left(\frac{22\cdot39 - 22\cdot40}{0\cdot008} < Z < \frac{22\cdot41 - 22\cdot40}{0\cdot008}\right); Z \sim N(0, 1)$$

$$= 300 \times P(-1\cdot25 < Z < 1\cdot25) = 600 \times P(0 < Z < 1\cdot25) \simeq 237$$

7. (*a*) A random sample of 500 is drawn from a large number of freshly minted coins. The mean weight of the coins in the sample is 28·57 gm. and the standard deviation is 1·25 gm. What are the limits which have a 49 to 1 chance of including the mean weight of all the coins ? How large a sample would have to be drawn to make these limits differ by only 0·1 gm, assuming that the standard deviation of the whole distribution is 1·25 gm.

(*b*) A research worker wishes to estimate the mean of a population by using sufficiently large sample. The probability is 0·95 that the sample mean will not differ from the true mean of a normal population by more than 25% of the standard deviation. How large a sample should be taken ? (**Ans.** $n = 62$.)

8. (*a*) A normal distribution has mean 0·5 and standard deviation 2·5. Find :

(*i*) The probability that the mean of a random sample of size 16 from the population is positive.

(*ii*) The probability that the mean of a sample of size 90 from the population will be negative.

(*b*) The mean of a certain normal distribution is equal to the standard error of the mean of a random sample of 100 from that distribution. Find the probability, (in terms of an integral), that the mean of a sample of 25 from the distribution will be negative. (**Ans.** 0·3085.)

(*c*) The average value $\bar{x}$ of a random sample of observations from a certain population is normally distributed with mean 20 and standard deviation $5/\sqrt{n}$. How large a sample should be drawn in order to have a probability of at least 0·90 that $\bar{x}$ will lie between 18 and 22. (***Karnataka Univ. B.E. 1991***)

9. (*a*) From a population of 169 units it is desired to choose a simple random sample of size n. If the population standard deviation is 2, determine the smallest 'n' for which the probability that the sample mean differs from the population mean by more than 0·75 is controlled at 0·05.

(*b*) An economist would like to estimate the mean income (μ) in a large city. He has decided to use the sample mean as an estimate of μ and would like to ensure that the error in estimation is not more than Rs. 100 with probability 0·90. How large a sample should he take if the standard deviation is known to be Rs. 1,000 ? [***Delhi Univ. M.A. (Eco.), 1986***]

Ans. $n = \left[\frac{z_\alpha \cdot \sigma}{E}\right]^2 = \left[\frac{1{\cdot}645 \times 1000}{100}\right]^2 = 270{\cdot}6 \simeq 271$

(*c*) The management of a manufacturing firm wishes to determine the average time required to complete a certain manual operation. There should be 0·95 confidence that the error in the estimate will not exceed 2 minutes.

What sample size is required if the standard deviation of the time needed to complete the manual operation is estimated by a time and motion study expert as (*i*) 10 minutes, (*ii*) 16 minutes ? Explain intuitively (without referring to the formula) why the sample size is large in (*ii*) than in (*i*).

(*Given* $Z_{0{\cdot}975} = 1{\cdot}96$ and $Z_{0{\cdot}95} = 1{\cdot}645$)

[***Delhi Univ. M.C.A., 1987***]

Ans. (*i*) $n_1 = \left(\frac{z_\alpha \cdot \sigma}{E}\right)^2 = \left(\frac{1{\cdot}96 \times 10}{2}\right)^2 = 96$, (*ii*) $n_2 = \left(\frac{1{\cdot}96 \times 16}{2}\right)^2 = 246$.

10. (*a*) Two populations have the same mean, but the standard deviation of one is twice that of the other. Show that in samples of 500 each drawn under simple random conditions, the difference of the means will, in all probability, not exceed $0{\cdot}3\sigma$, where σ is the smaller standard deviation, and assuming the distribution of the difference of the means to be normal, find the probability that it exceeds half that amount. (**Ans.** 0·1336.)

(*b*) A simple sample of heights of 6,400 Englishmen has a mean of 67·85 inches and S.D. 2·56 inches, while a simple sample of heights of 1,600 Australians has a mean of 68·55 inches and a S.D. of 2·52 inches. Do the data indicate that Australians are, on the average, taller than Englishmen ?

Ans. $H_0 : \mu_1 = \mu_2, H_1 : \mu_1 < \mu_2, Z = 9{\cdot}2$, (significant).

(*c*) In a random sample of 500, the mean is found to be 20. In another independent sample of 400, the mean is 15. Could the samples have been drawn from the same population with standard deviation 4 ?

11. (*a*) The following table presents data on the values of a harvested crop stored in the open and inside a godown :

	Sample size	*Mean*	$\Sigma (x - \bar{x})^2$
Outside	40	117	8,685
Inside	100	132	27,315

Assuming that the two samples are random and they have been drawn from normal populations with equal variances, examine if the mean value of the harvested crop is affected by weather conditions.

Ans. $Z = 0{\cdot}342$; Not significant.

(*b*) Samples of students were drawn from two universities and from their weights in kgm., means and standard deviations are calculated. Make a large sample test to test the significance of the difference between the means.

	Mean	*S.D.*	*Size of sample*
University *A*	55	10	400
University *B*	57	15	100

Ans. $Z = 1{\cdot}2648$; Not significant.

(*c*) A storekeeper wanted to buy a large quantity of light bulbs from two orands labelled 'one' and 'two'. He bought 100 bulbs from each brand and found by testing that brand 'one' had mean lifetime of 1120 hours and the standard deviation of 75 hours; and brand 'two' had mean lifetime of 1062 hours and standard deviation of 82 hours. Examine whether the difference of means is significant.

12. The mean yield of two sets of plots and their variability are as given below. Examine

(*i*) whether the difference in the mean yields of the two sets of plots is significant, and

(*ii*) whether the difference in the variability in yields is significant.

	Set of 40 plots	*Set of 60 plots*
Mean yield per plot	1258 lb.	1243 lb.
S.D. per plot	34 lb.	28 lb.

Ans. (*i*) $Z = 2{\cdot}3$, (*ii*) $Z = 1{\cdot}3$.

13. (*a*) In a survey of incomes of two classes of workers, two random samples gave the following details. Examine whether the differences between the (*i*) means and (*ii*) the standard deviations, are significant.

Sample	*Size*	*Mean annual income (in rupees)*	*Standard deviation (in rupees)*
I	100	582	24
II	100	546	28

Examine also whether the first sample could have come from a population with annual mean income of 500 rupees.

(*b*) The electric light tubes of manufacturer *A* have a lifetime of 1400 hours, with a standard deviation of 200 hours, while of manufacture *B* have a mean lifetime of 1200 hours with a standard deviation of 100 hours. If random samples of 125 tubes of each batch are tested, what is the probability that the brand *A* tubes will have a mean time which is at least (*i*) 160 hours more than the brand *B* tubes, and (*ii*) 250 hours more than the brand *B* tubes ?

Hint. Under the assumption of normal population, the sampling distribution of $(\bar{x}_1 - \bar{x}_2)$ would have mean; $\mu_1 - \mu_2 = 1400 - 1200 = 200$ hours and standard deviation :

$$\text{S.E. } (\bar{x}_1 - \bar{x}_2) = \sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)} = \sqrt{\left\{\frac{(100)^2}{125} + \frac{(200)^2}{125}\right\}} = 20 \text{ hours.}$$

(*i*) The required probability is given by :

$$P\{(\bar{x}_1 - \bar{x}_2) \geq 160\} = P\left[\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\text{S.E. } (\bar{x}_1 - \bar{x}_2)} \geq \frac{160 - 200}{20}\right]$$

$$= P(Z \geq -2) = 0{\cdot}5 + P(-2 < Z < 0)$$

(*ii*) The required probability is given by :

$$P\{(\bar{x}_1 - \bar{x}_2) \geq 250\} = P(Z \geq 2{\cdot}5) = 0{\cdot}5 - P(0 < Z < 2{\cdot}5)$$

$$= 0{\cdot}5 - 0{\cdot}4938 = 0{\cdot}0062$$

14. A random sample of 1,200 men from one State gives the mean pay as Rs. 400 p.m.with a standard deviation of Rs. 60, and a random sample of 1,000 men from another State gives the mean pay as Rs. 500 p.m., with a standard deviation of Rs. 80.

Discuss, (stating clearly the result or theorem used), whether the mean levels of pay of men from the two States differ significantly.

15. (*a*) A normal population has a mean 0·1 and a standard deviation 2·1. Find the probability that the mean of a sample of size 900 will be negative, it being given that the probability that the absolute value of a standard normal variate exceeds 1·43 is 0·153.

(*b*) A random sample of 100 articles selected from a batch of 2,000 articles shows that the average diameter of the articles is 0·354 with a standard deviation 0·048. Find 95% confidence interval for the average of this batch of 2,000 articles.

Hint. We are given $n = 100, N = 2{,}000, \bar{x} = 0{\cdot}354, s = 0{\cdot}048$.

The Standard Error of sample mean $\bar{x}$ in random sampling from the batch of $N = 2{,}000$ is given by : [c.f. (16·23)].

$$S.E.(\bar{x}) = \sqrt{\frac{N-n}{N-1}} \times \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{N-n}{N-1}} \times \frac{s}{\sqrt{n}} \quad (\because \hat{\sigma} = s, \text{ for large } n)$$

$$= \sqrt{\frac{2000 - 100}{2000 - 1}} \times \frac{0{\cdot}048}{\sqrt{100}} = 0{\cdot}00468$$

Hence 95% confidence limits for μ are given by :

$\bar{x} \pm 1{\cdot}96\ S.E.(\bar{x}) = 0{\cdot}354 \pm 1{\cdot}96 \times 0{\cdot}00468 = (0{\cdot}3448,\ 0{\cdot}3632)$

16. (*a*) Explain the terms :

(*i*) Statistic and Parameter

(*ii*) Sampling distribution of a statistic, and

(*iii*) Standard error of a statistic.

(*b*) Explain why a random sample of size 30 is to be preferred to a random sample of size 25 to estimate the population mean.

17. (*a*) Obtain the expressions for the standard error of sampling distributions of : (*i*) sample mean ($\bar{x}$), and (*ii*) sample variance (s^2), in random sampling from a large population. Assume that n, the sample size, is large.

(*b*) Let $X_1, X_2, \ldots, X_n$ be a random sample from a population which has a finite fourth moment $\mu_r = E(X_i - \mu)^r$, $r = 4$; $E(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2$; and

let : $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ and $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$.

Show that : (*i*) $S^2 = \frac{1}{2n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} (X_i - X_j)^2$,

(*ii*) $\text{Var}(S^2) = \frac{1}{n}\left[\mu_4 - \left(\frac{n-3}{n-1}\right)\sigma^4\right]$,

(*iii*) $\text{Cov}(\bar{X}, S^2) = \mu_3 / n$