PART B

Unit - VIII

Sampling Theory

8.1 Introduction

In our day to day life it becomes quite often necessary to draw some valid and reasonable conclusions concerning a large mass of individuals or things. It becomes practically impossible to examine every individual or the entire group known as *population*. Therefore we may prefer to examine a small part of this population known as a *sample* with the motive of drawing some conclusion about the entire population based on the information/result revealed by the sample. The entire process known as ***statistical inference*** aims in ascertaining maximum information about the population with minimum effort and time. Poll prediction is a good example for statistical inference.

8.2 Random Sampling

A large collection of individuals or attributes or numerical data can be understood as a *population* or *universe*.

A finite subset of the universe is called a ***sample***. The number of individuals in a sample is called a sample size. If the sample size (n) is less than or equal to 30 the sample is said to be small, otherwise it is a large sample.

The process of selecting a sample from the population is called as ***sampling***.

The selection of an individual or item from the population in such a way that each has the same chance of being selected is called as ***random sampling***. Suppose we take a sample of size n from a finite population of size N, then we will have N_{C_n} possible samples. Random sampling is a technique in which each of the N_{C_n} samples has an equal chance of being selected.

Sampling where a member of the population may be selected more than once is called as *sampling with replacement*, on the other hand if a member cannot be chosen more than once is called as *sampling without replacement*.

Simple sampling is a special case of random sampling in which trials are independent and the probability of success is a constant.

The word *statistic* is often used for the random variable or for its values.

8.3 Sampling Distributions

Let us suppose that we have different samples of size n drawn from a population. For each and every sample of size n we can compute quantities like mean, standard deviation etc. Obviously these will not be the same. Suppose we group these characteristics according to their frequencies, the frequency distributions so generated are called ***sampling distributions***. These can be distinguished as sampling distribution of mean, standard deviation etc. *The sampling distribution of large samples is assumed to be a normal distribution.*

The standard deviation of a sampling distribution is also called the *Standard Error* (S.E). The reciprocal of the standard error is called *precision.*

8.31 Sampling distribution of the means

Sample mean is a statistic and we discuss the sampling distribution of this statistic.

We consider all possible random samples of size n and determine the mean of each one of these samples. We discuss the sampling distribution of the sample means for the two possible types of random sampling (with/without replacment) associated with a finite population.

Case - (1) ***Random sampling with replacement***

Here the items are drawn one by one and are put back to the population before the next draw. If N is the size of the finite population and n is the size of the sample then we have N^n samples.

The mean $\mu_{\bar{x}}$ of the frequency distribution of the sample means will be equal to the population mean (μ).

The variance $\sigma_{\bar{x}}^2$ of the frequency distribution of the sample means will be equal to σ^2/n where σ^2 is the variance of the population.

Thus we have, $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}}^2 = \sigma^2/n$

$\sigma_{\bar{x}}$ is also called the *standard error* of the means.

Case - (2) Random sampling without replacement

Here the items are drawn one by one and are not put back to the population before the next draw. In this case there will be N_{C_n} samples and we have the following results in accordence with the notations used in the *case - (1)*

$$\mu_{\bar{x}} = \mu \ ; \ \sigma_{\bar{x}}^2 = \left[\frac{N-n}{N-1}\right]\frac{\sigma^2}{n} = C\,\frac{\sigma^2}{n}$$

where $C = \dfrac{N-n}{N-1}$ is called the *finite population correction factor.*

Remark : *If N is very large, then C is closer to 1 as we have* $\lim_{N \to \infty} \dfrac{N-n}{N-1} = 1$

In practice if N is large the correction factor can be omitted.

Illustrative Example

Let $\{1, 2, 3\}$ constitute a population. We form the sampling distribution of the sample means in the case of (1) random samples of size 2 with replacement (2) random samples of size 2 without replacement.

>> Here N (*size of the finite population*) $= 3$

Population mean $(\mu) = \dfrac{1+2+3}{3} = 2$

Population variance $(\sigma^2) = \dfrac{1}{3}\left\{(1-2)^2+(2-2)^2+(3-2)^2\right\} = \dfrac{2}{3}$

Case - (i) Random samples of size 2 $(n = 2)$ with replacement

The various possible samples are :

(1, 1) (1, 2) (1, 3) ; (2, 1), (2, 2), (2, 3) ; (3, 1), (3, 2), (3, 3)

These are $N^n = 3^2 = 9$ in number.

The mean of these are respectively

1, 1.5, 2, 1.5, 2, 2.5, 2, 2.5, 3

We prepare the frequency distribution of these means where x is the variate and f is the frequency.

x	1	1.5	2	2.5	3
f	1	2	3	2	1

We shall compute the mean and variance of this frequency distribution.

$$\mu_{\bar{x}} = \frac{\sum fx}{\sum f} = \frac{1+3+6+5+3}{9} = 2$$

$$\sigma_{\bar{x}}^2 = \frac{\sum f(x-\mu_{\bar{x}})^2}{\sum f}$$

$$\sigma_{\bar{x}} = \frac{1}{9}\left\{1(1-2)^2+2(1.5-2)^2+3(2-2)^2+2(2.5-2)^2+1(3-2)^2\right\}$$

$$\sigma_{\bar{x}} = \frac{1}{9}(1+0.5+0+0.5+1) = \frac{1}{3}$$

Thus we have $\mu_{\bar{x}} = 2 = \mu$; $\sigma_{\bar{x}}^2 = \frac{1}{3}$ and $\sigma^2/n = \frac{(2/3)}{2} = \frac{1}{3}$

Case - (ii) Random sample of size 2 without replacement.

We have $3_{C_2} = 3$ samples. The three samples are $(1, 2), (2, 3), (3, 1)$

The associated means are 1.5 , 2.5 , 2 . Further we have,

$$\text{Mean} = \mu_{\bar{x}} = \frac{1.5+2.5+2}{3} = 2$$

$$\text{Variance} = \sigma_{\bar{x}}^2 = \frac{1}{3}\left\{(1.5-2)^2+(2.5-2)^2+(2-2)^2\right\} = \frac{0.5}{3} = \frac{1}{6}$$

Here we note that $\mu_{\bar{x}} = \mu = 2$

Also $\left[\frac{N-n}{N-1}\right]\frac{\sigma^2}{n} = \left[\frac{3-2}{3-1}\right]\frac{2/3}{2} = \frac{1}{6} = \sigma_{\bar{x}}^2$

Thus $\sigma_{\bar{x}}^2 = \left[\frac{N-n}{N-1}\right]\frac{\sigma^2}{n} = \frac{1}{6}$

Note : *Suppose N is large say 500 and n is small say 5 then the correction factor (C) becomes*

$$C = \frac{N-n}{N-1} = \frac{500-5}{500-1} = \frac{495}{499} = 0.992 \approx 1$$

8.4 Testing of Hypothesis

In order to arrive at a decision regarding the population through a sample of the population we have to make certain assumption referred to as **hypothesis** *which may or may not be true.* Much depends on the framing of hypothesis.

The hypothesis formulated for the purpose of its rejection under the assumption that it is true is called the **Null Hypothesis** denoted by H_0.

Any hypothesis which is complimentary to the null hypothesis is called **Alternative Hypothesis** denoted by H_1.

Examples

(1) To test whether a process B is better than a process A we can formulate the hypothesis as *there is no difference between the process A and B.*

(2) To test whether there is a relationship between two variates we can formulate the hypothesis as *there is no relationship between them.*

8.41 Errors

In a test process there can be four possible situations of which two of the situations leads to the two types of errors and the same is presented as follows.

	Accepting the hypothesis	Rejecting the hypothesis
Hypothesis true	Correct decision	Wrong decision (Type I error)
Hypothesis false	Wrong decision (Type II error)	Correct decision

In order to minimize both these types of errors we need to increase the sample size. It is further important to note that acceptance or non acceptance of a hypothesis is purely based on the information revealed by the sample and what is indicated by a particular sample may not always be true in respect of the population.

A region which amounts to the rejection of null hypothesis is called critical region or region of rejection.

8.42 Significance level

The probability level, below which leads to the rejection of the hypothesis is known as the *significance level.* This probability is conventionally fixed at 0.05 or 0.01 i.e., 5% or 1%. These are called *significance levels.*

We feel confident in rejecting a hypothesis at 1% level of significance than at 5% level of significance. 5% level of significance can also be understood as, the probability of committing errors of either types, (Type I or Type II) is 0.05.

8.5 Tests of significance and Confidence intervals

The process which helps us to decide about the acceptance or rejection of the hypothesis is called the *test of significance.*

Let us suppose that we have a normal population with mean μ and S.D σ. If $\bar{x}$ is the sample mean of a random sample of size n the quantity z defined by

$$z = \frac{\bar{x} - \mu}{(\sigma / \sqrt{n})} \qquad \ldots (1)$$

is called the *Standard Normal Variate (S.N.V).*

From the table of normal areas we find that 95% of the area lies between $z = -1.96$ and $z = +1.96$. In other words we can say with 95% confidence that z liesbetween -1.96 and $+1.96$. Further 5% level of significance is denoted by $z_{0.05}$. Thus we can write the verbal statement in the mathematical form,

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

i.e., $$\frac{-\sigma}{\sqrt{n}}(1.96) \le \bar{x}-\mu \le \frac{\sigma}{\sqrt{n}}(1.96)$$

$$\Rightarrow \quad \mu \le \bar{x} + \frac{\sigma}{\sqrt{n}}(1.96) \text{ and } \bar{x} - \frac{\sigma}{\sqrt{n}}(1.96) \le \mu$$

Thus we can write by combining the two results in the form,

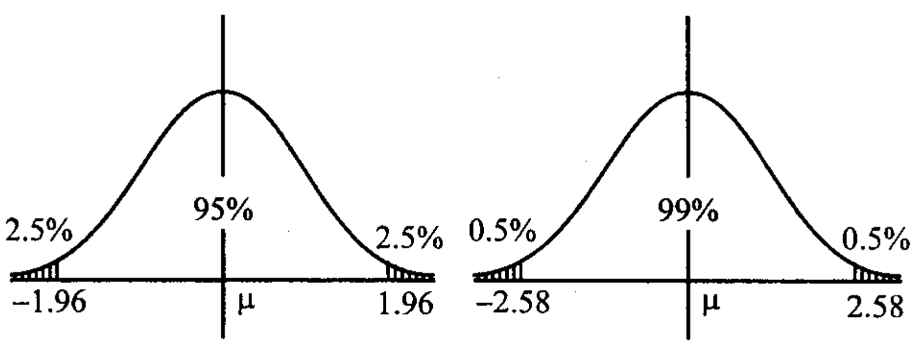
$$\bar{x} - 1.96\left(\frac{\sigma}{\sqrt{n}}\right) \le \mu \le \bar{x} + 1.96\left(\frac{\sigma}{\sqrt{n}}\right) \qquad \ldots (2)$$

Similarly, from the table of normal areas 99% of the area lies between -2.58 and $+2.58$. This is equivalent to the form,

$$\bar{x} - 2.58\left(\frac{\sigma}{\sqrt{n}}\right) \le \mu \le \bar{x} + 2.58\left(\frac{\sigma}{\sqrt{n}}\right) \qquad \ldots (3)$$

Thus we can say that (2) is the 95% confidence interval and (3) is the 99% confidence interval.

The constants 1.96, 2.58 etc. in the confidence limits are called confidence coefficients denoted by z_c. From confidence levels we can find confidence coefficients and vice-versa.



As reflected in the figure, we can say with 95% confidence that if the hypothesis is true, the value of z for an actual sample lies between -1.96 to 1.96 since the area under the normal curve between these values is 0.95. However if the value of z for random sample lies outside this range we can conclude that the probability of the happening of such an event is only 0.05 if the given hypothesis is true.

The total shaded area 0.05 being the level of significance of the test, represents the probability of making type-I error. *(rejecting the hypothesis when it should have been accepted).* The set of values of z outside the range -1.96, 1.96 constitutes the critical (*significant*) region or the region of rejecting the hypothesis whereas the values of z within the same range constitutes the insignificant region or the region of acceptance of the hypothesis.

8.51 One tailed and two tailed tests

In our test of acceptance or non acceptance of a hypothesis we concentrated on the value of z on both sides of the mean. This can be categorically stated that the focus of attention lies in the two "tails" of the distribution and hence such a test is called a *two tailed test.*

Sometimes we will be interested in the extreme values to only one side of the mean in which case the region of significance will be a region to one side of the distribution. Obviously the area of such a region will be equal to the level of significance itself. Such a test is called a *one tailed test.*

The critical values of $z : -1.96, 1.96$ as already stated with reference to 5% and 1% level of significance can be understood as the values in respect of a two tailed test. However the critical values of z in respect of a one tailed test (*as found in the table of areas under a normal curve*) are $[-1.645, 1.645]$; $[-2.33, 2.33]$ at 5% and 1% level of significance respectively.

The following table will be useful for working problems.

Test	Critical values of z	
	5% level	1% level
One-tailed test	–1.645 or 1.645	–2.33 or 2.33
Two-tailed test	–1.96 and 1.96	–2.58 and 2.58

8.6 Tests of significance for large samples

8.61 Test of significance of proportions

In the discussion of probability distributions we have remarked that the normal distribution is the limiting form of the binomial distribution when n is large and neither p nor q is small. Let us suppose that we take N samples, each having n members. Let p be the probability of success of each member and q of failure so that $p + q = 1$. The frequencies of samples with successes 0, 1, 2, . . . n are the terms of the binomial expansion of $N(q + p)^n$. Thus the binomial distribution is regarded as the sampling distribution of the number of successes in the sample.

We know that the mean of this distribution is np and S.D is $\sqrt{npq}$.

Let us consider the proportion of successes.

(1) mean proportion of successes = $\frac{np}{n} = p$

(2) S.D or S.E proportion of successes $= \frac{\sqrt{npq}}{n} = \sqrt{pq/n}$

Let x be the observed number of successes in a sample size of n and $\mu = np$ be the expected number of successes. Let the associated standard normal variate Z be defined by

$$Z = \frac{x-\mu}{\bar{\sigma}} = \frac{x-np}{\sqrt{npq}}$$

If $|Z| > 2.58$ we conclude that the difference is highly significant and reject the hypothesis. Since p is the probability of success and $\sqrt{pq/n}$ is the S.E proportion of successes, $p \pm 2.58\sqrt{pq/n}$ are the probable limits.

8.62 Test of significance for difference of means

Let μ_1 and μ_2 be the mean of two populations.

Let $(\bar{x}_1, \sigma_1); (\bar{x}_2, \sigma_2)$ be the mean and S.D of two large samples of size n_1 and n_2 respectively. We wish to test the null hypothesis H_0 that there is no difference between the population means. That is $H_0 : \mu_1 = \mu_2$.

The statistic for this test is given by

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

Also confidence limits for the difference of means of the population are

$$(\bar{x}_1 - \bar{x}_2) \pm z_c\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$$

We adopt the same procedure for testing the null hypothesis by using one tailed test or two tailed test.

Corollary: If the samples are drawn from the same population then $\sigma_1 = \sigma_2 = \sigma$

$$Z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sigma\sqrt{(1/n_1 + 1/n_2)}}$$

8.63 Test of significance for difference of properties (attributes) for two samples

Let p_1 and p_2 be the sample proportions in respect of an attribute corresponding to two large samples of size n_1 and n_2 drawn from two populations.

We wish to test the null hypothesis H_0 that there is no difference between the population with regard to the attribute.

The statistic for this test is given by

$$Z = \frac{p_1 - p_2}{\sqrt{pq\ (1/n_1 + 1/n_2)}} \quad \text{where} \quad p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} \quad \text{and} \quad q = 1 - p$$

We adopt the same procedure for testing the null hypothesis by using one tailed test or two tailed test.

8.7 Test of significance for small samples

In the case of large samples, sampling distribution follows a normal distribution which is not true in the case of small samples. We introduce the concept of *degrees of freedom* for discussing *Student's t distribution*.

8.71 Degrees of freedom

The number of degrees of freedom $(d \cdot f)$ usually denoted by ν is the number of values in a set which may be assigned arbitrarily. It can be interpreted as the number of independent values generated by a sample of small size for estimating a population parameter.

Examples : Let us suppose that we need to find 3 numbers whose sum is 25. That is to find a, b, c such that $a + b + c = 25$. We can arbitrarily assign values to any *two* of the variables a, b, c and hence these are the degrees of freedom. That is to say that $d \cdot f(\nu) = 2$. If there are n observations $d \cdot f$ is equal to $(n - 1)$.

Suppose that we are finding the mean of a sample of size n comprising values $x_1, x_2, \cdots x_n$. We use all the n values to compute the sample mean $\bar{x}$. Then $\bar{x}$ is said to have n degrees of freedom.

Suppose we are finding the sample variance, we use the n values
$(x_1 - \bar{x})^2, (x_2 - \bar{x})^2, \cdots (x_n - \bar{x})^2$.

But these values do not have n degrees of freedom as they all depend on a fixed value $\bar{x}$ which has already been computed. Hence the sample variance is said to have $(n - 1)\ d \cdot f$. If we compute another statistic based on the sample mean and variance, that statistic is said to have $(n - 2)\ d \cdot f$ and so on. In general the number of degrees of freedom $\nu = n - k$ where n is the number of observations in the sample and k is the number of constraints / number of values which are pre determined.

8.8 Student's *t* Distribution

Sir william Gosset under the pen name '***Student***' derived a theoretical distribution to test the significance of a sample mean where the small sample is drawn from a normal population.

Let $x_i\,(i = 1, 2, \cdots n)$ be a random sample of size n drawn from a normal population with mean μ and variance σ^2. The statistic t is defined as follows.

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

Here $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean.

$$s^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2$$ is the sample variance.

$\nu = (n-1)$ denote the number of degrees of freedom of t.

The statistic t follows the Student's t distribution with $(n-1)$ $d \cdot f$ having the probability density function

$$y = f(t) = \frac{y_0}{\left[1 + t^2/\nu\right]^{(\nu+1)/2}}$$

where y_0 is a constant such that the area under the curve is unity.

Note : 1. *Statistic t is also defined as follows.*

$$t = \frac{\bar{x} - \mu}{\sigma}\sqrt{n-1}$$

2. *The constant y_0 present in p.d.f is given by*

$$y_0 = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\;\Gamma(\nu/2)}$$ *so that the p.d.f of the Student's t distribution with ν degrees of freedom is given by*

$$y = f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\;\Gamma(\nu/2)}\left[1 + \frac{t^2}{\nu}\right]^{-(\nu+1)/2} \quad ; -\infty < t < \infty$$

3. *If ν is large $(\nu \geq 30)$ the graph of $f(t)$ closely approximates standard normal curve. In other words we can say that t is normally distributed for large samples.*

8.81 Student's t test for a sample mean

We need to test the hypothesis, whether the sample mean ($\bar{x}$) differs significantly from the population mean (μ) / hypothetical value (μ).

We compute $t = \dfrac{\bar{x} - \mu}{s} \sqrt{n}$ and consider $|t|$.

We also take a note of the value of t for the given $d \cdot f$ from the table of standard values.

If $|t| > t_{.05}$ the difference between $\bar{x}$ and μ is said to be significant at 5% level of significance.

If $|t| > t_{.01}$ the difference is said to be significant at 1% level of significance.

If $|t|$ is less than the table value at a certain level of significance, the data is said to be conformal / consistent with the hypothesis that μ is the mean of the population.

8.82 Confidence limits for the population mean μ

If $t_{.05}$ is the tabulated value of t for $(n-1)\, d \cdot f$ at 5% level of significance, it implies that.

$$P[\,|t| > t_{.05}\,] = 0.05$$

$$\Rightarrow \quad P[\,|t| \le t_{.05}\,] = 1 - 0.05 = 0.95$$

Now consider $|t| \le t_{.05}$

i.e., $$\left| \frac{\bar{x} - \mu}{s} \sqrt{n} \right| \le t_{.05}$$

or $$\left| \frac{\bar{x} - \mu}{(s/\sqrt{n})} \right| \le t_{.05}$$

i.e., $$-t_{.05} \le \frac{\bar{x} - \mu}{(s/\sqrt{n})} \le t_{.05}$$

i.e., $$\frac{-s}{\sqrt{n}}\, t_{.05} \le \bar{x} - \mu \le \frac{s}{\sqrt{n}}\, t_{.05}$$

$$\Rightarrow \quad \mu \le \bar{x} + \frac{s}{\sqrt{n}}\, t_{.05} \quad \text{and} \quad \bar{x} - \frac{s}{\sqrt{n}}\, t_{.05} \le \mu$$

Combining these two results we can write in the form

$$\bar{x} - \frac{s}{\sqrt{n}}\, t_{.05} \le \mu \le \bar{x} + \frac{s}{\sqrt{n}}\, t_{.05}$$

Thus we have 95% confidence limits for μ given by $\bar{x} \pm \frac{s}{\sqrt{n}} t_{.05}$

Similarly 99% confidence limits for μ are given by $\bar{x} \pm \frac{s}{\sqrt{n}} t_{.01}$

Note : *Confidence limits are also called Fiducial limits.*

8.83 Test of significance of difference between sample means

Consider two independent samples $x_i (i = 1, 2, \cdots n_1)$ and $y_j (j = 1, 2, \cdots n_2)$ drawn from a normal population.

Let $(\bar{x}, \sigma_x)$ and $(\bar{y}, \sigma_y)$ respectively be the mean and variance of the two samples. Let μ be the population mean and σ be the population variance. We need to test the hypothesis whether the difference between the sample means is significant.

We compute $t = \dfrac{\bar{x} - \bar{y}}{S . \sqrt{1/n_1 + 1/n_2}}$

where $S^2 = \dfrac{1}{n_1 + n_2 - 2} \left\{ \sum_{i=1}^{n_1} (x_i - \bar{x})^2 + \sum_{j=1}^{n_2} (y_j - \bar{y})^2 \right\}$

and degrees of freedom $v = n_1 + n_2 - 2$

if $|t| > t_{.05}$ the difference between the sample means is said to be significant at 5% level of significance. Similarly for 1% level of significance also.

Table for values of $|t|$ is given at the end of the book.

8.9 Chi - Square distribution

Chi - Square distribution provides a measure of correspondence between the theoretical frequencies and observed frequencies.

If $0_i (i = 1, 2, \ldots n)$ and $E_i (i = 1, 2, \ldots n)$ respectively denotes a set of observed and estimated frequencies, the quantity **chi - square** denoted by χ^2 is defined as follows.

$$\chi^2 = \sum_{i=1}^{n} \frac{(0_i - E_i)^2}{E_i} \quad ; \quad \text{degrees of freedom} = n - 1$$

Note : *If the expected frequencies are less than 10, we group them suitably for computing the value of chi. square.*

8.91 Chi - Square test as a test of goodness of fit

It is possible to test the hypothesis about the association of two attributes. We have already discussed the fitting of Binomial distribution, Normal distribution, Poisson distribution to a given data. It is easily possible to find the theoretical frequencies from the distribution of fit.

Chi - Square test helps us to test the goodness of fit of these distributions.

If the calculated value of χ^2 is less than the table value of χ^2 at a specified level of significance the hypothesis is accepted, otherwise the hypothesis is rejected.

WORKED PROBLEMS

Sampling distribution of the means

1. *A population consists of five numbers* 2, 3, 6, 8, 11. *Consider all possible samples of size 2 which can be drawn with replacement from this population. Find **(a)** the mean and S.D of the population. **(b)** the mean and standard deviation of the sampling distribution of means. **(c)** Considering samples without replacement find the mean and S.D of the sampling distribution of means.*

>> (a) Population mean $\mu = \dfrac{2+3+6+8+11}{5} = 6$

Population variance $\sigma^2 = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$

i.e., $\sigma^2 = \dfrac{1}{5}\left\{(2-6)^2+(3-6)^2+(6-6)^2+(8-6)^2+(11-6)^2\right\}$

$\sigma^2 = \dfrac{54}{5} = 10.8$

Thus $\mu = 6$ **and** $\sigma = \sqrt{10.8}$

(b) Let us consider all samples of size 2 with replacement. $N = 5$, $n = 2$. There will be $N^n = 5^2 = 25$ samples which are as follows.

(2, 2) (2, 3) (2, 6) (2, 8) (2, 11)

(3, 2) (3, 3) (3, 6) (3, 8) (3, 11)

(6, 2) (6, 3) (6, 6) (6, 8) (6, 11)

(8, 2) (8, 3) (8, 6) (8, 8) (8, 11)

(11, 2) (11, 3) (11,6) (11, 8) (11, 11)

∴

The mean of these samples in the respective order is as follows.

(2, 2.5, 4, 5, 5.5) ; (2.5, 3, 4.5, 5.5, 7)

(4, 4.5, 6, 7, 8.5) ; (5, 5.5, 7, 8, 9.5) (6.5, 7, 8.5, 9.5, 11).

Note : *The mean and S.D can be found as we have done in the case of population or by forming a frequency distribution as follows.*

x	2	2.5	3	4	4.5	5	5.5	6	6.5	7	8	8.5	9.5	11
f	1	2	1	2	2	2	2	1	2	4	1	2	2	1

$$\mu_{\bar{x}} = \frac{\Sigma f x}{\Sigma f} = \frac{150}{25} = 6$$

$$\sigma^2_{\bar{x}} = \frac{\Sigma f x^2}{\Sigma f} - [\mu_{\bar{x}}]^2 = \frac{1035}{25} - (6)^2 = 5.4$$

Thus $\mu_{\bar{x}} = 6$ and $\sigma_{\bar{x}} = \sqrt{5.4}$

Remark : *We observe that* $\mu_{\bar{x}} = \mu$ *and* $\sigma^2_{\bar{x}} = \sigma^2/n$ *in this case of random sampling with replacement.*

(c) Let us consider random samples without replacement.

$N_{C_n} = 5_{C_2} = 10$ samples are as follows.

(2, 3) (2, 6) (2, 8) (2, 11) (3, 6) (3, 8)
(3, 11) (6, 8) (6, 11) (8, 11)

The mean of these samples are respectively

2.5, 4, 5, 6.5, 4.5, 5.5, 7, 7, 8.5, 9.5

$$\therefore \quad \mu_{\bar{x}} = \frac{1}{10}(2.5 + 4 + 5 + \cdots + 9.5) = \frac{60}{10} = 6$$

$$\sigma^2_{\bar{x}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_{\bar{x}})^2 = \frac{1}{10}(40.5) = 4.05$$

Thus $\mu_{\bar{x}} = 6$ **and** $\sigma_{\bar{x}} = \sqrt{4.05}$

Remark : *We observe that* $\mu_{\bar{x}} = \mu$. *Also the result*

$$\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) \text{ can be verified as we have}$$

$$\text{RHS} = \frac{10.8}{2}\left(\frac{5-2}{5-1}\right) = \frac{10.8}{2} \times \frac{3}{4} = 4.05 = \sigma^2_{\bar{x}} = \text{LHS}$$

2. *A population consists of 4 numbers 3, 7, 11, 15.*

(a) *Find the mean and S.D of the sampling distribution of means by considering samplings of size 2 with replacement.*

(b) *If N,n denotes respectively the population size and sample size, σ and $\sigma_{\bar{x}}$ respectively denotes population S.D and S.D of the sampling distribution of means without replacement verify that*

(i) $\sigma_{\bar{x}}^2 = \dfrac{\sigma^2}{n}\left[\dfrac{N-n}{N-1}\right]$

(ii) $\mu_{\bar{x}} = \mu$ *where* $\mu_{\bar{x}}$ *is the mean of this distribution and* μ *is the population mean.*

>> Population mean $\mu = \dfrac{1}{4}(3+7+11+15) = 9$

Population variance $\sigma^2 = \dfrac{1}{4}\{(3-9)^2+(7-9)^2+(11-9)^2+(15-9)^2\} = 20$

Thus $\mu = 9$ **and** $\sigma = \sqrt{20}$

(a) Let us consider samples of size 2 with replacement. They are as follows.

(3, 3) (3, 7) (3, 11) (3, 15)

(7, 3) (7, 7) (7, 11) (7, 15)

(11, 3) (11, 7) (11, 11) (11, 15)

(15, 3) (15, 7) (15, 11) (15, 15)

Sampling means are as follows.

(3, 5, 7, 9);(5, 7, 9, 11);(7, 9, 11, 13):(9, 11, 13, 15)

The frequency distribution of the sampling means is as follows.

x	3	5	7	9	11	13	15
f	1	2	3	4	3	2	1

$$\mu_{\bar{x}} = \frac{\Sigma fx}{\Sigma f} = \frac{144}{16} = 9$$

$$\sigma_{\bar{x}}^2 = \frac{\Sigma fx^2}{\Sigma f} - [\mu_{\bar{x}}]^2 = \frac{1456}{16} - (9)^2 = 10$$

Thus $\mu_{\bar{x}} = 9$ **and** $\sigma_{\bar{x}} = \sqrt{10}$

Remark : $\mu_{\bar{x}} = \mu$ *and* $\sigma_{\bar{x}}^2 = \sigma^2/n$ *where* $\sigma^2 = 20,\ n = 2$

(b) Let us consider samples without replacement. They are as follows.

(3, 7) (3, 11) (3, 15) (7, 11) (7, 15) (11, 15)

The sampling means are 5, 7, 9, 9, 11, 13

$$\therefore \quad \mu_{\bar{x}} = \frac{1}{6}(5+7+9+9+11+13) = 9 \text{ ; Thus } \mu_{\bar{x}} = \mu$$

$$\sigma_{\bar{x}}^2 = \frac{1}{6}\left\{(5-9)^2+(7-9)^2+\cdots+(13-9)^2\right\} = \frac{40}{6} = \frac{20}{3}$$

Consider $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}\left[\frac{N-n}{N-1}\right]$

$$\text{RHS} = \frac{20}{2}\left[\frac{4-2}{4-1}\right] = 10 \times \frac{2}{3} = \frac{20}{3} = \sigma_{\bar{x}}^2 = \text{LHS}$$

---

3. *The weights of 1500 ball bearings are normally distributed with a mean of 635 gms. and S.D of 1.36 gms. If 300 random samples of size 36 are drawn from this population, determine the expected mean and S.D of the sampling distribution of means if sampling is done*

(a) with replacement (b) without replacement.

>> Here $N = 1500$, $\mu = 635$, $\sigma = 1.36$, $n = 36$

(a) Expected mean $\mu_{\bar{x}} = \mu = \mathbf{635}$

Expected S.D $\sigma_{\bar{x}} = \sqrt{\sigma^2/n} = \frac{\sigma}{\sqrt{n}} = \frac{1.36}{6} = \mathbf{0.227}$

(b) Expected mean $\mu_{\bar{x}} = \mu = 635$

Expected variance $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}\left[\frac{N-n}{N-1}\right] = \frac{(1.36)^2}{36}\left[\frac{1500-36}{1500-1}\right] = 0.05$

Thus $\sigma_{\bar{x}} = \sqrt{0.05} = \mathbf{0.224}$

---

4. *Consider the data as in the previous problem. In the case of random sampling with replacement find how many random samples would have their mean **(a)** between 634.76 gms and 635.24 gms **(b)** greater than 635.5 gms **(c)** less than 634.2 gms **(d)** less than 634.5 gms or more than 635.24 gms.*

>> We assume that the population is a normal population and hence the sampling distribution of means is also taken to be distributed normally.

The standard normal variate $z = \frac{x-\mu}{\sigma}$ in the case of sampling distribution of means is in the equivalent form

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} \text{ where we have } \mu_{\bar{x}} = 635, \ \sigma_{\bar{x}} = 0.227$$

Hence we have $z = \dfrac{\bar{x} - 635}{0.227}$ ...(1)

(a) Probability of a sample having mean between 634.76 and 635.24 is represented by $P(634.76 < \bar{x} < 635.24)$

Now from (1), if $\bar{x} = 634.76$, $z = \dfrac{-0.24}{0.227} = -1.06$

if $\bar{x} = 635.24$, $z = \dfrac{0.24}{0.227} = 1.06$

Hence we have to find $P(-1.06 < z < 1.06)$

ie., $= 2P(0 < z < 1.06)$

$= 2\phi(1.06) = 2(0.3554)$ by using tables.

$= 0.7108$

Thus we have corresponding to 300 samples, the expected number of samples having their mean between 634.76 gms and 635.24 gms is given by

$300 \times 0.7108 = 213.24 \approx$ **213 samples**

(b) To find $P(\bar{x} > 635.5) \times 300$

If $\bar{x} = 635.5$ then $z = \dfrac{635.5 - 635}{0.227}$ from (1). That is $z = 2.203$

$\therefore$ $P(z > 2.203) = P(z > 0) - P(0 < z < 2.203)$

$= 0.5 - \phi(2.2)$

$= 0.5 - 0.4861 = 0.0139$

Thus $P(\bar{x} > 635.5) \times 300 = 4.17 \approx$ **4 samples**

(c) To find $P(\bar{x} < 634.2) \times 300$

If $\bar{x} = 634.2$ then $z = \dfrac{634.2 - 635}{0.227}$ from (1). That is, $z = -3.52$

$\therefore \quad P(z < -3.52) = P(z > 3.52)$

$= P(z > 0) - P(0 < z < 3.52)$

$= 0.5 - \phi(3.52)$

$= 0.5 - 0.4998 = 0.0002$

Thus $P(\bar{x} < 634.2) \times 300 = 0.06 \approx$ **0 samples**

(d) To find $\left[P(\bar{x} < 634.5) + P(\bar{x} > 635.24)\right] \times 300$

If $\quad \bar{x} = 634.5$ then $z = -2.2$

$\bar{x} = 635.24$ then $z = 1.06$; by using (1).

$\therefore \quad P(z < -2.2) + P(z > 1.06)$

$= P(z > 2.2) + P(z > 1.06)$

$= \{P(z > 0) - P(0 < z < 2.2)\} + \{P(z > 0) - P(0 < z < 1.06)\}$

$= \{0.5 - \phi(2.2)\} + \{0.5 - \phi(1.06)\}$

$= 1 - \{\phi(2.2) + \phi(1.06)\}$

$= 1 - (0.4861 + 0.3554) = 0.1585$

Multiplying this value by the sample size 300 we get 47.55 $\approx$ **48 samples.**

---

5. *Certain tubes manufactured by a company have mean life time of 800 hours and S.D of 60 hours. Find the probability that a random sample of 16 tubes taken from the group will have a mean life time*
   (a) between 790 hours and 810 hours. (b) less than 785 hours.
   (c) more than 820 hours. (d) between 770 hours and 830 hours.

>> By data $\mu = 800$, $\sigma = 60$, $n = 16$

$\therefore \quad \sigma_{\bar{x}} = \sigma/\sqrt{n} = 60/4 = 15$

We have $z = \dfrac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \dfrac{\bar{x} - 800}{15}$ ...(1)

(a) To find $P(790 < \bar{x} < 810)$

If $\quad \bar{x} = 790$, $z = -0.67$

$\bar{x} = 810$, $z = 0.67$; by using (1).

$\therefore \quad P(-0.67 < z < 0.67) = 2P(0 < z < 0.67)$

$= 2\phi(0.67) = 2(0.2486) = 0.4972$

Thus $\quad$ **$P(790 < \bar{x} < 810) = 0.4972$**

(b) To find $P(\bar{x} < 785)$

If $\bar{x} = 785$ then $z = -1$ from (1).

$\therefore$ $P(z < -1) = P(z > 1)$

$= P(z > 0) - P(0 < z < 1)$

$= 0.5 - \phi(1) = 0.5 - 0.3413 = 0.1587$

Thus $\boldsymbol{P(\bar{x} < 785) = 0.1587}$

(c) To find $P(\bar{x} > 820)$

If $\bar{x} = 820$ then $z = 1.33$ from (1)

$\therefore$ $P(z > 1.33) = P(z > 0) - P(0 < z < 1.33)$

ie., $= 0.5 - \phi(1.33) = 0.5 - 0.4082 = 0.0918$

Thus $\boldsymbol{P(\bar{x} > 820) = 0.0918}$

(d) To find $P(770 < \bar{x} < 830)$

If $\bar{x} = 770$ then $z = -2$ and $\bar{x} = 830$ then $z = 2$; from (1).

$\therefore$ $P(-2 < z < 2) = 2P(0 < z < 2) = 2\phi(2) = 2(0.4772) = 0.9544$

Thus $\boldsymbol{P(770 < \bar{x} < 830) = 0.9544}$

Test of significance of proportions

6. *A coin is tossed 1000 times and head turns up 540 times. Decide on the hypothesis that the coin is unbiassed.*

>> Let us suppose that the coin is unbiassed.

p = probability of getting a head in one toss = 1/2

Since $p + q = 1,\ q = 1/2.$

Expected number of heads in 1000 tosses = $np = 1000 \times 1/2 = 500$

Actual number of heads = 540

The difference = $x - np = 540 - 500 = 40.$

Consider $Z = \dfrac{x - np}{\sqrt{npq}} = \dfrac{40}{\sqrt{1000 \times 1/2 \times 1/2}} = 2.53 < 2.58$

Thus we can say that the coin is unbiassed.

---

7. *Result extracts revealed that in a certain school over a period of five years 725 students had passed and 615 students had failed. Test the hypothesis that success and failure are in equal proportions.*

>> Total number of students = 725 + 615 = 1340

Observed proportion of success = $\frac{725}{1340}$ = 0.54

Suppose that success and failure are in equal proportion. Then $p = 1/2$

$\therefore$ difference in proportion = 0.54 − 0.5 = 0.04

S.D proportion of success = $\frac{\sqrt{npq}}{n} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.5 \times 0.5}{1340}} \approx 0.014$

In terms of proportion $Z = \frac{x - \mu}{\sigma} = \frac{0.04}{0.014} \approx 2.86 > 2.58$

Thus the hypothesis that success and failure are in equal proportion is rejected.

Aliter : Observed number of successes = 725

Expected number of successes = 1340/2 = 670

$$Z = \frac{x - \mu}{\sqrt{npq}} = \frac{725 - 670}{\sqrt{1340 \times 1/2 \times 1/2}} = 3.005 > 2.58$$

Hence the hypothesis is rejected.

---

8. *A sample of 900 days was taken in a coastal town and it was found that on 100 days the weather was very hot. Obtain the probable limits of the percentage of very hot weather.*

>> Probability of very hot weather = $p = \frac{100}{900} = \frac{1}{9} \quad \therefore \quad q = \frac{8}{9}$

Probable limits = $p \pm 2.58\sqrt{pq/n}$

$$= 0.111 \pm (2.58)\sqrt{\frac{1}{9} \times \frac{8}{9} \times \frac{1}{900}} = 0.111 \pm 0.027$$

$$= 0.084 \text{ and } 0.138$$

Probable limits of very hot weather is 8.4% to 13.8%

---

9. *In a sample of 500 men it was found that 60% of them had over weight. What can we infer about the proportion of people having over weight in the population ?*

>> Probability of persons having over weight = $p = \frac{60}{100} = 0.6$ & $q = 1 - p = 0.4$

Probable limits = $p \pm 2.58\sqrt{pq/n}$

Probable limits $= 0.6 \pm 2.58 \quad \dfrac{(0.6)(0.4)}{500}$

Probable limits $= 0.6 \pm 0.0565 = 0.5435$ and 0.6565

Thus the probable limits of people having over weight is 54.35% to 65.65%

---

10. *A survey was conducted in a slum locality of 2000 families by selecting a sample of size 800. It was revealed that 180 families were illiterates. Find the probable limits of the illiterate families in the population of 2000.*

>> Probability of illiterate families $= p = \dfrac{180}{800} = 0.225 \therefore q = 0.775$

Probable limits of illiterate families $= p \pm (2.58)\sqrt{pq/n}$

i.e., $= 0.225 \pm (2.58)\sqrt{\dfrac{(0.225)(0.775)}{800}} = 0.225 \pm 0.038$

$= 0.187$ and 0.263

$\therefore$ the probable limits of illiterate families in the population of 2000 is

$(0.187)\,2000$ and $(0.263)\,2000$

Thus **374 to 526 are probably illiterate families.**

---

11. *To know the mean weights of all 10 year old boys in Delhi a sample of 225 was taken. The mean weight of the sample was found to be 67 pounds with S.D of 12 pounds. What can we infer about the mean weight of the population ?*

>> Sample mean $(\bar{x}) = 67$, Sample size $n = 225$, S.D $(\sigma) = 12$

95% confidence limits for the mean of the population corresponding to a given sample is $\bar{x} \pm 1.96(\sigma/\sqrt{n})$ and 99% confidence limits for the mean is $\bar{x} \pm 2.58(\sigma/\sqrt{n})$.

We have $\sigma/\sqrt{n} = 12/15 = 0.8$

95% confidence limits : $67 \pm 1.96(0.8) = 65.432$ and 68.568

99% confidence limits : $67 \pm 2.58(0.8) = 64.936$ and 69.064

We can say with 95% confidence that the mean weight of the population lies between 65.4 pounds and 68.6 pounds. Also with 99% confidence we can say that the mean weight lies between 64.9 pounds to 69.1 pounds.

---

12. *In a hospital 230 females and 270 males were born in a year.*
Do these figures confirm the hypothesis that sexes are born in equal proportions ?

>> Total number of births, $n = 230 + 270 = 500$

Observed proportion of females $= \dfrac{230}{500} = \dfrac{23}{50} = 0.46.$

Assuming that sexes are born in equal proportion, probability of female birth is equal to $1/2$.

Thatis, $p = 1/2$ and hence $q = 1/2$

$\therefore$ difference in proportions $= 0.5 - 0.46 = 0.04$

$$\text{S.D proportion of females} = \frac{\sqrt{npq}}{n} = \sqrt{pq/n}$$

$$= \sqrt{1/2 \times 1/2 \times 1/500} = 0.0224$$

$$Z = \frac{\text{Difference}}{S.D} = \frac{0.04}{0.0224} = 1.786 < 2.58$$

Thus we conclude that the figures are conformal with the hypothesis that the sexes are born in equal proportions.

---

13. *In 324 throws of a six faced 'die', an odd number turned up 181 times. Is it reasonable to think that the 'die' is an unbiassed one ?*

>> Probability of the turn up of an odd number is $p = 3/6 = 1/2$

Hence $q = 1 - p = 1/2$

Expected number of successes $= 1/2 \times 324 = 162$

Observed number of successes $= 181$

$\therefore$ difference $= 181 - 162 = 19$

$$\text{Consider } Z = \frac{x - np}{\sqrt{npq}} = \frac{19}{\sqrt{324 \times 1/2 \times 1/2}} = \frac{19}{9} = 2.11 < 2.58$$

Thus we conclude that the die is unbiassed .

---

14. *A'die' is thrown 9000 times and a throw of 3 or 4 was observed 3240 times. Show that the die cannot be regarded as an unbiassed one.*

>> Probability of getting 3 or 4 is a single throw is $p = 2/6 = 1/3$ and $q = 1 - p = 2/3$

$\therefore$ expected number of successes $= 1/3 \times 9000 = 3000$

Observed number of successes $= 3240$

The difference $= 3240 - 3000 = 240$

$$\text{Consider } Z = \frac{x - np}{\sqrt{npq}} = \frac{240}{\sqrt{9000 \times 1/3 \times 2/3}} = \frac{240}{\sqrt{2000}} = 5.37$$

Since $Z = 5.37 > 2.58$ **we conclude that the die is biassed.**

---

15. *A sample of 100 days is taken from meteorological records of a certain district and 10 of them are found to be foggy. What are the probable limits of the percentage of foggy days in the district.*

>> p = proportion of foggy days in a sample of 100 days is given by $10/100 = 0.1$ Hence $q = 1 - p = 0.9$

∴ probable limits of foggy days

$$= p \pm 2.58 \sqrt{pq/n}$$

$$= 0.1 \pm 2.58 \sqrt{(0.1 \times 0.9)/100}$$

$$= 0.1 \pm 0.0774 = 0.0226 \text{ and } 0.1774$$

$$= 2.26\ \% \text{ and } 17.74\ \%$$

Thus the percentage of foggy days lies between 2.26 and 17.74

---

16. *A random sample of 500 apples was taken from a large consignment and 65 were found to be bad. Estimate the proportion of bad apples in the consignment as well as the standard error of the estimate. Also find the percentage of bad apples in the consignment.*

>> p = proportion of bad apples in the sample is given by $65/500 = 0.13$.

Hence $q = 1 - p = 0.87$

S.E proportion of bad apples $= \sqrt{pq/n} = \sqrt{(0.13 \times 0.87)/500} = 0.015$

Probable limits of bad apples in the consignment

$$= p \pm 2.58 \sqrt{pq/n}$$

$$= 0.13 \pm 2.58\ (0.015) = 0.13 \pm 0.0387$$

$$= 0.0913 \text{ and } 0.1687$$

$$= 9.13\ \% \text{ and } 16.87\ \%$$

Thus the required percentage of bad apples in the consignment lies between 9.13 and 16.87

---

17. *In a locality of 18000 families a sample of 840 families was selected at random. Of these 840 families, 206 families were found to have monthly income of Rs.2500 or less. It was desired to estimate how many of the 18,000 families have monthly income of Rs.2500 or less. Within what limits would you place your estimate.*

>> Proportion of families having monthly income of Rs.2500 or less is given by

$$p = 206/840 = 0.245. \text{ Hence } q = 1 - p = 0.755$$

S.E proportion $= \sqrt{pq/n} = \sqrt{(0.245 \times 0.755)/840} = 0.015$

Probable limits of families having monthly income of Rs.2500 or less are $p \pm 2.58 \sqrt{pq/n}$. That is,

$= 0.245 \pm (2.58)(0.015)$

$= 0.245 \pm 0.0387$

$= 0.2063$ and 0.2837 or $20.63\ \%$ and $28.37\ \%$

Hence the probable limits in respect of 18,000 families is given by

$0.2063 \times 18{,}000$ and $0.2837 \times 18{,}000$

That is 3713.4 and 5106.6 or 3713 and 5107

Thus we say that 3713 to 5107 families are likely to have monthly income of Rs.2500 or less.

Remark : *We have said that the value 2.58 is the confidencce coefficient corresponding to the 99 % confidence level. However we can even take this value to be equal to 3 corresponding to 99.73 % confidence level so that the probable limits can be taken as* $p \pm 3 \cdot \sqrt{pq/n}$

Accordingly we have the following answers in the previous 3 problems.

In problem - 15, we have, $0.1 \pm 3(0.03) = 0.01$ *to* 0.19 *or 1% to 19%*

In problem - 16, we have, $0.13 \pm 3(0.015) = 0.085$ *to* 0.175 *or 8.5 to 17.5%*

In problem - 17, we have, $0.245 \pm 3(0.015) = 0.2$ *to* 0.29

Multiplying by 18, 000 we get 3600 to 5220.

---

18. *The mean and S.D of the maximum loads supported by 60 cables are 11.09 tonnes and 0.73 tonnes respectively. Find (a) 95% (b) 99% confidence limits for mean of the maximum loads of all cables produced by the company.*

>> By data $\bar{x} = 11.09$, $\sigma = 0.73$

(a) 95 % confidence limits for the mean of maximum loads are given by

$\bar{x} \pm 1.96(\sigma/\sqrt{n})$

$= 11.09 \pm 1.96(0.73/\sqrt{60})$

$= 11.09 \pm 0.18$

Thus 10.91 tonnes to 11.27 tonnes are the 95 % confidence limits for the mean of maximum loads.

(b) 99 % confidence limits for the mean of maximum loads are given by

$\bar{x} \pm 2.58(\sigma/\sqrt{n})$

$= 11.09 \pm 2.58(0.73/\sqrt{60})$

$= 11.09 \pm 0.24 = 10.85$ and 11.33

Thus 10.85 tonnes to 11.33 tonnes are the 99 % confidence limits for the mean of maximum loads.

---

19. *The mean and S.D of the diameters of a sample of 250 rivet heads manufactured by a company are 7.2642 mm and 0.0058 mm respectively. Find (a) 99 % (b) 98 % (c) 95 % (d) 90 % (e) 50 % confidence limits for the mean diameter of all the rivet heads manufactured by the company.*

>> By data $\bar{x} = 7.2642$, $\sigma = 0.0058$, $n = 250$

Confidence limits for the mean is given by $\bar{x} \pm (z_c)(\sigma/\sqrt{n})$ where z_c is the confidence coefficient corresponding to the confidence level. We have the following from the normal probability tables.

Confidence level	99%	98%	95%	90%	50%
z_c	2.58	2.33	1.96	1.645	0.6745

Now $\dfrac{\sigma}{\sqrt{n}} = \dfrac{0.0058}{\sqrt{250}} = 0.00037$

Confidence limits for various confidence level respectively are as follows.

(a) $7.2642 \pm 2.58\,(0.00037) =$ **7.2642 ± 0.00095**

(b) $7.2642 \pm 2.33\,(0.00037) =$ **7.2642 ± 0.00086**

(c) $7.2642 \pm 1.96\,(0.00037) =$ **7.2642 ± 0.00073**

(d) $7.2642 \pm 1.645\,(0.00037) =$ **7.2642 ± 0.00061**

(e) $7.2642 \pm 0.6745\,(0.00037) =$ **7.2642 ± 0.00025**

---

20. *An unbiassed coin is thrown n times. It is desired that the relative frequency of the appearence of heads should lie between 0.49 and 0.51. Find the smallest value of n that will ensure this result with (a) 95% confidence (b) 90% confidence*

>> p = probability of getting a head $= 1/2$; $q = 1 - p = 1/2$

S.E proportion of heads $= \sqrt{pq/n} = \sqrt{1/4n} = (1/2)\sqrt{n}$

(a) Probable limits for 95% confidence level is given by $p \pm 1.96\sqrt{pq/n}$ which should be 0.51 and 0.49

ie $0.5 \pm 1.96\,(1/2\sqrt{n}) = 0.51$ or 0.49

ie., $0.5 + \dfrac{1.96}{2\sqrt{n}} = 0.51$ and $0.5 - \dfrac{1.96}{2\sqrt{n}} = 0.49$

$\Rightarrow \dfrac{1.96}{2\sqrt{n}} = 0.01$ or $\sqrt{n} = \dfrac{1.96}{0.02} = 98$

Thus **$n = 9604$**

(b) Taking the confidence coefficient equal to 1.645 for 90 % confidence level we have as before

$$\frac{1.645}{2\sqrt{n}} = 0.01 \quad \text{or} \quad \sqrt{n} = \frac{1.645}{0.02} = 82.25 \qquad \text{or} \qquad n = 6765.0625 \approx 6765 \ .$$

Thus n = **6765**

Test of significance of a sample mean

21. *A manufacturer claimed that atleast 95% of the equipment which he supplied to a factory conformed to specifications . An examination of a sample of 200 pieces of equipment revealed that 18 of them were faulty. Test his claim at a significance level of 1% and 5%.*

>> Let p be the probability of success which being the probability of the equipment supplied to the factory conformal to the specifications.

$\therefore$ by data $p = 0.95$ and hence $q = 0.05$

$H_0 : p = 0.95$ and the claim is correct.

$H_1 : p < 0.95$ and the claim is false.

We choose the one tailed test to determine whether the supply is conformal to the specification.

$$\mu = np = 200 \times 0.95 = 190$$

$$\sigma = \sqrt{npq} = \sqrt{200 \times 0.95 \times 0.05} = 3.082$$

Expected number of equipments according to the specification = 190

Actual number = 182 since 18 out of 200 were faulty

$\therefore$ difference = $190 - 182 = 8$

Now $Z = \dfrac{x - np}{\sqrt{npq}} = \dfrac{8}{3.082} = 2.6$

The value of Z is greater than the critical value 1.645 at 5% level and 2.33 at 1% level of significance. The claim of the manufacturer (*null hypothesis that the claim is correct*) is **rejected** at 5% as well as at 1% level of significance in accordence with the one tailed test.

22. *It has been found from experience that the mean breaking strength of a particular brand of thread is 275.6 gms with standard deviation of 39.7 gms. Recently a sample of 36 pieces of thread showed a mean breaking strength of 253.2 gms. Can one conclude at a significance level of **(a)** 0.05 **(b)** 0.01 that the thread has become inferior?*

>> We have to decide between the two hypothesis

$H_0 : \mu = 275.6$ gms, mean breaking strength

$H_1 : \mu < 275.6$ gms, inferior in breaking strength.

We choose the one tailed test.

Mean breaking strength of a sample of 36 pieces = 253.2

$\therefore$ difference= $275.6 - 253.2 = 22.4$; $n = 36$

$$Z = \frac{\text{difference}}{(\sigma/\sqrt{n})} = \frac{22.4}{39.7/6} = 3.38$$

The value of Z is greater than the critical value of Z = 1.645 at 5% level and 2.33 at 1% level of significance.

Under the hypothesis H_1 that the **thread has become inferior is accepted at both 0.05 and 0.01 levels in accordence with one tailed test.**

---

23. *In an examination given to students at a large number of different schools the mean grade was 74.5 and S.D grade was 8. At one particular school where 200 students took the examination the mean grade was 75.9. Discuss the significance of this result from the view point of (a) one tailed test (b) two tailed test at both 5% and 1% level of significance.*

>> $H_0 : \mu = 74.5$ and there is no change in the mean grade.

$H_1 : \mu \neq 74.5$ *i.e.,* $\mu > 74.5$ and $\mu < 74.5$.

$\mu = 74.5$ and mean of a sample of size 200 (n) is 75.9

$\therefore$ difference = $75.9 - 74.5 = 1.4$

$$Z = \frac{\text{difference}}{(\sigma/\sqrt{n})} = \frac{1.4}{8/\sqrt{200}} = 2.475$$

We have the table for the critical values of Z in the case of one and two tailed tests.

Test	$Z_{0.05}$	$Z_{0.01}$
One tailed	± 1.645	± 2.33
Two tailed	± 1.96	± 2.58

The calculated value of Z is more than $Z_{0.05}$, $Z_{0.01}$ in one tailed test as well as $Z_{0.05}$ in two tailed test.

Thus we conclude that the difference in the mean grade is significant in these tests but the same is not significant in the two tailed test at 1% level of significance.

Test of significance of difference between means

24. *In an elementary school examination the mean grade of 32 boys was 72 with a standard deviation of 8, while the mean grade of 36 girls was 75 with a standard deviation of 6. Test the hypothesis that the performance of girls is better than boys.*

>> We have $\bar{x}_B = 72,\ \sigma_B = 8,\ n_B = 32$ [Boys]

$\bar{x}_G = 75,\ \sigma_G = 6,\ n_G = 36$ [Girls]

Consider $$Z = \frac{(\bar{x}_G - \bar{x}_B)}{\sqrt{\sigma_G^2/n_G + \sigma_B^2/n_B}}$$

$$= \frac{(75-72)}{\sqrt{36/36 + 64/32}} = \frac{3}{\sqrt{3}} = \sqrt{3} = 1.73$$

$$\therefore \quad Z = 1.73 \begin{cases} > Z_{.05} = 1.645 \text{ (one tailed test)} \\ < Z_{.01} = 2.33 \text{ (one tailed test)} \end{cases}$$

The difference in the performance of girls and boys in the examination is **significant at 5% level but not at 1% level.**

25. *A sample of 100 bulbs produced by a company A showed a mean life of 1190 hours and a standard deviation of 90 hours. Also a sample of 75 bulbs produced by a company B showed a mean life of 1230 hours and a standard deviation of 120 hours. Is there a difference between the mean life time of the bulbs produced by the two companies at*
(a) 5% level of significance (b) 1% level of significance.

>> By data $\bar{x}_A = 1190,\ \sigma_A = 90,\ n_A = 100$ [Company *A*]

$\bar{x}_B = 1230,\ \sigma_B = 120,\ n_B = 75$ [Company *B*]

Consider $$Z = \frac{(\bar{x}_B - \bar{x}_A)}{\sqrt{\sigma_B^2/n_B + \sigma_A^2/n_A}}$$

$$= \frac{(1230-1190)}{\sqrt{(120)^2/75 + (90)^2/100}} = 2.42$$