SOLVED PROBLEMS INTRODUCTION TO STATISTICAL THEORY

PART II

(Prof. Sher Muhammad Chaudhary
Dr. Shahid Kamal)

CONTENTS

CHAPTER 14

SURVEY SAMPLING AND SAMPLING DISTRIBUTIONS

Q.14.1. (a) Population. In Statistics, *a population* (or universe) is defined as an aggregate or totality of units of interest or objects, whether animate or inanimate, concrete or abstract, *e.g.* population of College students, population of heights, population of botanical plants, population of opinions, etc. A statistical population thus consists of all measurements or counts of a variable.

A population may be finite or infinite. A *finite population* contains a finite number of units or objects such as the population of students in a college, the population of all licensed motor drivers, the population of houses in a country, etc. The size of a finite population is usually denoted by the letter N. A population containing an infinite number of units, is called an *infinite population* which is usually regarded as a conceptual device because nobody can ever actually enumerate all the units. The population of all points on a line, the population of all the results obtained by throwing of two dice, the population of all heights between 5 feet and 6 feet, etc. are the examples of infinite population.

Sample. A sample is a part of a population, selected for study. The number of units included in the sample is called the *size* of the sample and is denoted by the letter n. Sometimes, a sample may include the entire population. Samples are selected from population to provide estimates of population parameters. One cannot obtain valid conclusions if the sample is not representative of the population.

Sampling Frame. A list of elements or form of identification of the elements in the population that is used to select a sample, is called *sampling frame.*

Parameter. A numerical quantity used to describe the characteristics of a population is called a parameter.

A Statistic is a numerical descriptive measure computed from a sample.

(b) **Sampling.** The procedure of selecting a sample from the population is called *sampling*, and different procedures give different types of sampling.

Objects of Sampling. The fundamental objects of sampling are:

(i) to get the maximum information about a population without examining each and every unit in it; and

(ii) to find the reliability of the estimates obtained from the sample.

Q.14.16. The population consists of $X_1=2$, $X_2=4$, $X_3=6$, $X_4=8$, $X_5=10$. The mean, μ and the variance, σ^2 of the population are computed as below:

$$\mu = \frac{\sum X_i}{N} = \frac{2+4+6+8+10}{5} = \frac{30}{5} = 6, \text{ and}$$

$$\sigma^2 = \frac{\sum (X_i-\mu)^2}{N} = \frac{(2-6)^2+(4-6)^2+(6-6)^2+(8-6)^2+(10-6)^2}{5}$$

$$= \frac{16+4+0+4+16}{5} = \frac{40}{5} = 8.$$

These five members of the population can be used to simulate an infinite population, if samples are drawn with replacement. Theoretically an infinite number of samples could be drawn, many of them would be identical. All the possible random samples of size 2 that could result when sampling is with replacement, from the population are $(5)^2 = 25$.

But the number of all possible *distinct* samples (*i.e.* when sampling is without replacement) is ${}^5C_2$, *i.e.* $m=10$. These ten distinct samples of size 2 together with their means and variances are given as follows:

No.	Members of Samples		Sample Mean ($\bar{x}_i$)	Sample Variance (S^2)
1	X_1, X_2	2, 4	$\frac{2+4}{2} = 3$	$\frac{(2-3)^2+(4-3)^2}{2} = 1$
2	X_1, X_3	2, 6	4	4
3	X_1, X_4	2, 8	5	9
4	X_1, X_5	2, 10	6	$\frac{(2-6)^2+(10-6)^2}{2} = 16$
5	X_2, X_3	4, 6	5	1
6	X_2, X_4	4, 8	6	4
7	X_2, X_5	4, 10	7	9
8	X_3, X_4	6, 8	7	1
9	X_3, X_5	6, 10	8	4
10	X_4, X_5	8, 10	9	1
Total	--	--	60	50

Now the mean of sample means, $\mu_{\bar{x}}$, is

$$\mu_{\bar{x}} = \frac{\sum \bar{x}_i}{m} = \frac{60}{10} = 6 = \mu$$

Thus the mean of sample means is equal to the mean of the population. This property is of fundamental importance.

The mean of sample variance is

$$E(S^2) = \frac{50}{10} = 5$$

which is not identical with the population variance. The sampling variance is a biased estimator.

Q.14.17. We are required to select a random sample of 10 households from a list of 250 households, by using a table of random sampling numbers.

We first list the households and assign a 3-digit serial number to each household from 001 to 250 as 250 is a 3-digit number. Then using the first 3 columns from Random Numbers

Table (Table XXXIII page 136, Statistical Tables by Fisher and Yates, 6th edition), we find the following numbers:

221, 193, 167, 784, 032, 932, 787, 236, 153, 587, 573, 485, 619, 369, 188, 885, 092, 129, 859, 386, 534, 407, 021, 951, etc.

Hence we select the following 10 households for our sample, skipping a number larger than 250.

221, 193, 167, 32, 236, 153, 188, 92, 129 and 21.

In case of sampling with replacement, a household may be selected more than once but when sampling is done without replacement no household is selected again. Thus in such a case, a random sampling number if appears again, is to be skipped.

Q.14.18. First of all, we assign 2-digit number to each of the 100-students from 00 to 99. Thus the 5-students whose heights are 60–62 inches are assigned the numbers 00–04, the 18 students with heights 63–65 inches are given the numbers 05–22, and so on.

Height (inches)	Mid point (X_i)	f_i	Sampling Numbers
60–62	61	5	00–04
63–65	64	18	05–22
66–68	67	42	23–64
69–71	70	27	65–91
72–74	73	8	92–99
Total	--	100	

Then we draw the sampling numbers from the Random Number Tables (Table XXXIII on page 136, Statistical Tables by Fisher and Yates). As the 100 students have been assigned 2-digit numbers, we therefore use the first 2 columns (and then the next 2 columns) and find the following numbers.

22, 19, 16, 78, 03, 93, 78, 23, 15, 58, etc.

Each of these random sampling numbers, gives the height of a particular student. Thus the first random sampling number 22 corresponds to a student whose height is 63–65 inches, which we take as 64 inches, the midpoint of the concerned height group.

Similarly the random sampling numbers 19, 16, 78, 03 correspond to heights of 64, 64, 70, 61 inches. A student is to be selected more than once as the sampling is with replacement. Proceeding in this way, we draw 30 samples of size 3 each. These samples with their mean heights are presented in the following table:

No.	Sample Number Drawn	Corresponding Heights	Mean Heights
1	22, 19, 16	64, 64, 64	64
2	78, 03, 93	70, 61, 73	68
3	78, 23, 15	70, 67, 64	67
4	58, 57, 48	67, 67, 67	67
5	61, 36, 18	67, 67, 64	66
6	88, 09, 12	70, 64, 64	66
7	85, 38, 53	70, 67, 67	68
8	40, 02, 95	67, 61, 73	67
9	35, 26, 77	67, 67, 70	68
10	46, 37, 61	67, 67, 67	67
11	93, 21, 95	73, 64, 73	70
12	97, 69, 04	73, 70, 61	68
13	61, 85, 21	67, 70, 64	67
14	15, 02, 87	64, 61, 70	65
15	98, 10, 47	73, 64, 67	68
16	22, 67, 27	64, 70, 67	67
17	33, 13, 17	67, 64, 64	65
18	36, 77, 43	67, 70, 67	68
19	28, 22, 76	67, 64, 70	67
20	68, 39, 71	70, 67, 70	69
21	35, 50, 96	67, 67, 73	69
22	93, 87, 56	73, 70, 73	70
23	72, 96, 94	70, 73, 73	72
24	64, 44, 76	67, 67, 70	68
25	17, 17, 76	64, 64, 70	66
26	29, 80, 40	67, 70, 67	68
27	56, 65, 43	67, 70, 67	68
28	96, 20, 86	73, 64, 70	69
29	92, 31, 06	73, 67, 64	68
30	93, 74, 69	73, 70, 70	71

To find the mean of means, we construct the following frequency table:

Sample mean ($\bar{x}_i$)	Tally	f_i	$f_i \bar{x}_i$
64	I	1	64
65	II	2	130
66	III	3	198
67	𝍸 II	7	469
68	𝍸 𝍸	10	680
69	III	3	207
70	II	2	140
71	I	1	71
72	I	1	72
Σ	--	30	2031

Hence $\mu_{\bar{x}} = \dfrac{\sum f_i \bar{x}_i}{\sum f_i} = \dfrac{2031}{30} = 67.7$ inches.

Q.14.19. (i) Given X has a binomial distribution with $p=0.4$ and $n=5$, i.e.

$$P(X=x) = \binom{5}{x} (0.4)^x (0.6)^{5-x} \text{ for } x = 0, 1, 2, 3, 4, 5$$

We first calculate the probabilities associated with each value of x. The probabilities and the assigned numbers are shown below:

x	$P(X=x)$	Cumulative $P(X\leq x)$	Assigned Numbers
0	0.0778	0.0778	0000 – 0777
1	0.2592	0.3370	0778 – 3369
2	0.3456	0.6826	3370 – 6825
3	0.2304	0.9130	6826 – 9129
4	0.0768	0.9898	9130 – 9897
5	0.0102	1.0000	9898 – 9999

We now consult a table of random numbers to select a sample of 10 by finding 10-four digit numbers. Let us select four columns, say columns 11, 12, 13 and 14 of Table 14.1, page 8 of Text. Then going down the four columns, we select the first 10 numbers. These numbers and the x-values (in brackets) corresponding to them are listed below:

6132 (2), 9900 (5), 0672 (0), 6551 (2), 0191 (0),

7150 (3), 4822 (2), 0110 (0) 5154 (2), 6148 (2),

The sample results are shown in the following table:

x	0	1	2	3	4	5
$f(x)$	3	0	5	1	0	1

(ii) Given X has a Poisson distribution with

$$P(X=x) = \frac{e^{-4}(4)^x}{x!}, \text{ for } x = 0, 1, 2, ...$$

The probabilities and the assigned numbers are shown below:

x	$P(X=x)$	Cumulative $P(X\leq x)$	Assigned Numbers
0	0.0183	0.0183	0000 – 0182
1	0.0733	0.0916	0183 – 0915
2	0.1465	0.2381	0916 – 2380
3	0.1954	0.4335	2381 – 4334
4	0.1954	0.6289	4335 – 6288
5	0.1563	0.7852	6289 – 7851
6	0.1042	0.8894	7852 – 8893
7	0.0595	0.9489	8894 – 9488
8	0.0293	0.9787	9489 – 9786
9	0.0132	0.9919	9787 – 9918
10	0.0053	0.9972	9919 – 9971
11+	0.0028	1.0000	9972 – 9999

Now we select a sample of 10 by finding 10-four digit numbers from a table of random numbers. Let us select the last four columns of Table 14.1, page 8 of the Text. Going down these columns, the following 10 numbers are found (the corresponding x values are shown in brackets):

1995 (2), 8060 (6), 0049 (0), 8865 (6), 3707 (3),

8698 (6), 2329 (2), 8380 (6), 1897 (2), 4257 (3).

Thus the sample consists of the following x-values:

0, 2, 2, 2, 3, 3, 6, 6, 6, 6.

Q.14.21. (b) According to proportional allocation, *i.e.*, $n_h = \frac{N_h}{N}$, we would select the following sub-samples:

$$n_1 = n\,(N_1/N) = 8 \times \frac{4}{16} = 2.$$

$$n_2 = n\,(N_2/N) = 8 \times \frac{6}{16} = 3, \text{ and}$$

$$n_3 = n\,(N_3/N) = 8 \times \frac{6}{16} = 3.$$

Using a table of random numbers, we get from

Stratum I: $X_{11} = 3$, $X_{13} = 4$.

Stratum II: $X_{21} = 10$, $X_{24} = 16$, $X_{26} = 20$.

Stratum III: $X_{32} = 18$, $X_{34} = 22$, $X_{35} = 26$.

Hence the sample mean $= \frac{1}{n}\sum_{h=1}^{k}\sum_{i=1}^{n_h} X_{hi}$ (k = no. of strata)

$$= \frac{1}{8}[3+4+10+16+20+18+22+26]$$

$$= \frac{119}{8} = 14.9.$$

Q.14.22. (a) Calculation of sample size for each stratum by Proportional Allocation:

Classification	No. of students (N_i)	Sample size (n_i)
B.Sc.	150	$n_1 = \frac{N_1}{N} \times n = 8$
B.A.	163	$n_2 = \frac{163}{728} \times 40 = 9$
F.Sc.	195	$n_3 = \frac{195}{728} \times 40 = 11$
F.A.	220	$n_4 = \frac{22}{728} \times 40 = 12$
Total	N = 728	$n = 40$

(b) A sample of 2% of all the employees means that $n = \frac{2}{100} \times 300{,}000 = 6{,}000$.

Now, using the proportional allocation, *i.e.* $n_i = n \cdot \frac{N_i}{N}$, the following subsample sizes are computed:

$$n_1 = n \cdot \frac{N_1}{N} = 6{,}000 \times \frac{45{,}000}{300{,}000} = 900,$$

$$n_2 = n \cdot \frac{N_2}{N} = 6{,}000 \times \frac{90{,}000}{300{,}000} = 1{,}800,$$

$$n_3 = n \cdot \frac{N_3}{N} = 6{,}000 \times \frac{75{,}000}{300{,}000} = 1{,}500,$$

$$n_4 = n \cdot \frac{N_4}{N} = 6{,}000 \times \frac{60{,}000}{300{,}000} = 1{,}200,$$

$$n_5 = n \cdot \frac{N_5}{N} = 6{,}000 \times \frac{30{,}000}{300{,}000} = 600.$$

Q.14.23. (a) Sampling Distribution. A *sampling distribution* is a probability (realtive frequency) distribution formed by the values of a *statistic* such as the sample proportion, the sample correlaton coefficient, etc., computed from all the different possible samples of size n, which can be drawn either with replacement or without replacement from the same population. The mean, the variance etc. can be computed from a sampling distribution. The sampling distributions are of great importance in statistical inference.

Properties of the Sampling Distribution of means. The important properties of the sampling distribution of means are:

(i) The mean of the sampling distribution of mean is always equal to the mean of the population, *i.e.*,

$$\mu_{\bar{x}} = \mu$$

(ii) The variance of the sampling distribution of mean is given by

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$$

when the sampling is without replacement from a finite population of size N, or by

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n},$$

when the sampling is with replacement or random sampling from an infinite population, and where σ^2 is the population variance.

(iii) The sampling distribution of mean will be normal or approximately normal for a reasonably large sample size.

(b) Finite Correction Factor. When a sample of size n is drawn without replacement from a finite population of size N, the variance of the sampling distribution of say, mean is given by

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1},$$

where σ^2 is the population variance.

The factor $\frac{N-n}{N-1}$ is usually called the *finite correction factor* (fcf) or *finite population correction* (fpc) for the variance. The finite correction factor approaches unity as the population becomes larger and larger.

The formula for $\sigma_{\bar{x}}^2$ is, in fact, exact and needs no correction. However, when the proportion of the population sampled – $\frac{n}{N}$ (called the sampling fraction)-is small, say less than 5 per cent, many statisticians are of the view that the finite correction factor may be neglected.

Q.14.26. (b) Let A, B, C, D and E stand for children. There are $(5)^2=25$ samples which can be drawn with replacement. The samples and the mean ages are given below:

No.	Members of sample	Ages	Mean Age	No.	Members of sample	Ages	Mean Age
1	A, A	4, 4	4	14	C, D	6, 7	6.5
2	A, B	4, 5	4.5	15	C, E	6, 8	7
3	A, C	4, 6	5	16	D, A	7, 4	5.5
4	A, D	4, 7	5.5	17	D, B	7, 5	6
5	A, E	4, 8	6	18	D, C	7, 6	6.5
6	B, A	5, 4	4.5	19	D, D	7, 7	7
7	B, B	5, 5	5	20	D, E	7, 8	7.5
8	B, C	5, 6	5.5	21	E, A	8, 4	6
9	B, D	5, 7	6	22	E, B	8, 5	6.5
10	B, E	5, 8	6.5	23	E, C	8, 6	7
11	C, A	6, 4	5	24	E, D	8, 7	7.5
12	C, B	6, 5	5.5	25	E, E	8, 8	8
13	C, C	6, 6	6				

(i) The theoretical sampling distribution of $\bar{X}$, the mean age of two children in any sample, is given as follows:

$\bar{x}_i$	Tally	f_i	$f_i\bar{x}_i$	$f_i\bar{x}_i^2$
4	I	1	4	16
4.5	II	2	9	40.5
5	III	3	15	75
5.5	IIII	4	22	121
6	卌	5	30	180
6.5	IIII	4	26	169
7	III	3	21	147
7.5	II	2	15	112.5
8	I	1	8	64
Total	---	25	150	925.0

(ii) The mean, $\mu_{\bar{x}}$ and the standard error of $\bar{X}$, $\sigma_{\bar{x}}$ are

$$\mu_{\bar{x}} = \frac{\sum f_i \bar{x}_i}{\sum f_i} = \frac{150}{25} = 6; \text{ and}$$

$$\sigma^2_{\bar{x}} = \frac{\sum f_i \bar{x}_i^2}{\sum f_i} - \left(\frac{\sum f_i \bar{x}_i}{\sum f_i}\right)^2 = \frac{925}{25} - \left(\frac{150}{25}\right)^2$$

$$= 37 - 36 = 1,$$

so that $\sigma_{\bar{x}} = 1$.

Q.14.27. (i) The possible number of samples of size $n = 2$ that could be drawn without replacement is $\binom{6}{2} = 15$.

(ii) Let A, B, C, D, E and F stand for the population values 2, 4, 8, 8, 10 and 10. Then the samples of size 2 and their means are:

No.	Members of sample	Ages	Mean Age	No.	Members of sample	Ages	Mean Age
1	A, B	2, 4	3	9	B, F	4, 10	7
2	A, C	2, 8	5	10	C, D	8, 8	8
3	A, D	2, 8	5	11	C, E	8, 10	9
4	A, E	2, 10	6	12	C, F	8, 10	9
5	A, F	2, 10	6	13	D, E	8, 10	9
6	B, C	4, 8	6	14	D, F	8, 10	9
7	B, D	4, 8	6	15	E, F	10, 10	10
8	B, E	4, 10	7				

The sampling distribution of the means is given below:

$\bar{x}$	3	5	6	7	8	9	10
$f(\bar{x})$	1/15	2/15	4/15	2/15	1/15	4/15	1/15

Q.14.28. Total possible samples of size 4 with replacement from the population 2, 4, 6 are $3^4 = 81$ samples.

The 81 possible samples and their means are listed below:

Samples	Mean	Samples	Mean	Samples	Mean
2,2,2,2	2	4,2,2,2	2.5	6,2,2,2	3.0
2,2,2,4	2.5	4,2,2,4	3.0	6,2,2,4	3.5
2,2,2,6	3	4,2,2,6	3.5	6,2,2,6	4.0
2,2,4,2	2.5	4,2,4,2	3.0	6,2,4,2	3.5
2,2,4,4	3.0	4,2,4,4	3.5	6,2,4,4	4.0
2,2,4,6	3.5	4,2,4,6	4.0	6,2,4,6	4.5
2,2,6,2	3.0	4,2,6,2	3.5	6,2,6,2	4.0
2,2,6,4	3.5	4,2,6,4	4.0	6,2,6,4	4.5
2,2,6,6	4.0	4,2,6,6	4.5	6,2,6,6	5.0
2,4,2,2	2.5	4,4,2,2	3.0	6,4,2,2	3.5
2,4,2,4	3.0	4,4,2,4	3.5	6,4,2,4	4.0
2,4,2,6	3.5	4,4,2,6	4.0	6,4,2,6	4.5
2,4,4,2	3.0	4,4,4,2	3.5	6,4,4,2	4.0
2,4,4,4	3.5	4,4,4,4	4.0	6,4,6,2	4.5
2,4,4,6	4.0	4,4,4,6	4.5	6,4,6,4	5.0
2,4,6,2	3.5	4,4,6,2	4.0	6,4,4,6	4.5
2,4,6,4	4.0	4,4,6,4	4.5	6,4,6,4	5.0
2,4,6,6	4.5	4,4,6,6	5.0	6,4,6,6	5.5
2,6,2,2	3.0	4,6,2,2	3.5	6,6,2,2	4.0
2,6,2,4	3.5	4,6,2,4	4.0	6,6,2,4	4.5
2,6,2,6	4.0	4,6,2,6	4.5	6,6,2,6	5.0
2,6,4,2	3.5	4,6,4,2	4.0	6,6,4,2	4.5
2,6,4,4	4.0	4,6,4,4	4.5	6,6,4,4	5.0
2,6,4,6	4.5	4,6,4,6	5.0	6,6,4,6	5.5
2,6,6,2	4.0	4,6,6,2	4.5	6,6,6,2	5.0
2,6,6,4	4.5	4,6,6,4	5.0	6,6,6,4	5.5
2,6,6,6	5.0	4,6,6,6	5.5	6,6,6,6	6.0

The sampling distribution of the Mean is obtained below:

$\bar{x}_i$	Tally	f	$f\bar{x}$	$f\bar{x}^2$
2.0	I	1	2.0	4.0
2.5	IIII	4	10.0	25.0
3.0	卌 卌	10	30.0	90.0
3.5	卌 卌 卌 I	16	56.0	196.0
4.0	卌 卌 卌 IIII	19	76.0	304.0
4.5	卌 卌 卌 I	16	72.0	324.0
5.0	卌 卌	10	50.0	250.0
5.5	IIII	4	22.0	121.0
6.0	I	1	6.0	36.0
Total	---	81	324.0	1350

$$\therefore \quad \mu_{\bar{x}} = \frac{\sum f\bar{x}}{\sum f} = \frac{324.0}{81} = 4.$$

$$\sigma^2_{\bar{x}} = \frac{\sum f\bar{x}^2}{\sum f} - \left(\frac{\sum f\bar{x}}{\sum f}\right)^2 = \frac{1350}{81} - \left(\frac{324}{81}\right)^2 = 0.67.$$

For verification, we first calculate mean and variance of the given population.

$$\mu = \frac{2+4+6}{3} = 4$$

$$\sigma^2 = \frac{(2-4)^2 + (4-4)^2 + (6-4)^2}{3} = \frac{4+0+4}{3} = 2.67$$

Now (i) $\mu = 4 = \mu_{\bar{x}}$ (ii) $\frac{\sigma^2}{n} = \frac{2.67}{4} = 0.67 = \sigma^2_{\bar{x}}$

Hence the result.

Q.14.29. (i) All possible random samples of size n=2 that could be drawn without replacement, are:

(2, 2), (2, 4), (2, 4), (2, 6), (2, 8), (2, 10), (2, 4), (2, 4), (2, 6), (2, 8), (2, 10), (4, 4) (4, 6), (4, 8), (4, 10), (4, 6), (4, 8), (4; 10), (6, 8), (6, 10), (8, 10).

The sample means are 2, 3, 3, 4, 5, 6, 3, 3, 4, 5, 6, 4, 5, 6, 7, 5, 6, 7, 7, 8, 9.

(ii) $$\mu = \frac{2+2+4+4+6+8+10}{7} = \frac{36}{7} = 5.14$$

$$\sigma = \sqrt{\frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2} = \sqrt{\frac{240}{7} - \left(\frac{36}{7}\right)^2}$$

$$= \sqrt{34.2857 - 26.4490} = \sqrt{7.8367} = 2.7994$$

$$\mu_{\bar{x}} = \frac{\sum \bar{X}_i}{\text{No. of samples}} = \frac{108}{21} = 5.14 = \mu$$

$$\sigma_{\bar{x}} = \sqrt{\frac{\sum \bar{x}_i^2}{21} - \left(\frac{\sum \bar{x}_i}{21}\right)^2} = \sqrt{\frac{624}{21} - \left(\frac{108}{21}\right)^2}$$

$$= \sqrt{29.7143 - 26.4490} = \sqrt{3.2653} = 1.8070$$

Now $$\frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} = \frac{2.7994}{\sqrt{2}} \cdot \sqrt{\frac{7-2}{7-1}}$$

$$= \frac{(2.7994)\,(2.2361)}{3.4641} = 1.8070 = \sigma_{\bar{x}}.$$

Hence the result.

(iii) The Chebyshev inequality says, "at least $1 - \frac{1}{k^2}$ of the observations lie within mean $\pm$ $k(s.d.)$". The problem says", at least $\frac{8}{9}$ of the means", so $\frac{8}{9}$ is $1 - \frac{1}{k^2}$, which gives $k = 3$.

Hence we would expect *at least* $\frac{8}{9}$ of the sample means to fall between $\mu_{\bar{x}} - 3\sigma_{\bar{x}}$ and $\mu_{\bar{x}} + 3\sigma_{\bar{x}}$, that is between $5.14 - 3 \times (1.807)$ and $5.14 + 3\,(1.807)$ or between 0 and 10.

Q.14.30. The possible samples of size n=3 are $\binom{6}{3} = 20$.

To compute the sampling distribution of means, we first calculate the sample means as follows:

Samples	$\bar{x}$	Samples	$\bar{x}$
0, 3, 6	3	3, 6, 12	7
0, 3, 12	5	3, 6, 15	8
0, 3, 15	6	3, 6, 18	9
0, 3, 18	7	3, 12, 15	10
0, 6, 12	6	3, 12, 18	11
0, 6, 15	7	3, 15, 18	12
0, 6, 18	8	6, 12, 15	11
0, 12, 15	9	6, 12, 18	12
0, 12, 18	10	6, 15, 18	13
0, 15, 18	11	12, 15, 18	15

$\bar{x}$	f	$f(\bar{x})$	$\bar{x}\,f(\bar{x})$	$\bar{x}^2 f(\bar{x})$
3	1	1/20	3/20	9/20
5	1	1/20	5/20	25/20
6	11	2/20	12/20	72/20
7	111	3/20	21/20	147/20
8	11	2/20	16/20	128/20
9	11	2/20	18/20	162/20
10	11	2/20	20/20	200/20
11	111	3/20	33/20	363/20
12	11	2/20	24/20	288/20
13	1	1/20	13/20	169/20
15	1	1/20	15/20	225/20
$\sum$	20	1	180/20=9	1788/20

$E(\bar{X}) = \mu_{\bar{x}} = 9,$

$$\sigma_{\bar{x}} = \sqrt{\frac{1788}{20} - (9)^2} = \sqrt{89.4 - 81} = 2.8983$$

Q.14.31. Total number of possible samples with replacement = $N^n = 4^3 = 64$ samples. The samples together with their means are given below:

Samples	Mean	Samples	Mean	Samples	Mean	Samples	Mean
3,3,3	3	6,3,3	4	9,3,3	5	12,3,3	6
3,3,6	4	6,3,6	5	9,3,6	6	12,3,6	7
3,3,9	5	6,3,9	6	9,3,9	7	12,3,9	8
3,3,12	6	6,3,12	7	9,3,12	8	12,3,12	9
3,6,3	4	6,6,3	5	9,6,3	6	12,6,3	7
3,6,6	5	6,6,6	6	9,6,6	7	12,6,6	8
3,6,9	6	6,6,9	7	9,6,9	8	12,6,9	9
3,6,12	7	6,6,12	8	9,6,12	9	12,6,12	10
3,9,3	5	6,9,3	6	9,9,3	7	12,9,3	8
3,9,6	6	6,9,6	7	9,9,6	8	12,9,6	9
3,9,9	7	6,9,9	8	9,9,9	9	12,9,9	10
3,9,12	8	6,9,12	9	9,9,12	10	12,9,12	11
3,12,3	6	6,12,3	7	9,12,3	8	12,12,3	9
3,12,6	7	6,12,6	8	9,12,6	9	12,12,6	10
3,12,9	8	6,12,9	9	9,12,9	10	12,12,9	11
3,12,12	9	6,12,12	10	9,12,12	11	12,12,12	12

The sampling distribution of mean is obtained below:

$\bar{x}$	Tally	f	$f\bar{x}$	$f\bar{x}^2$
3	I	1	3	9
4	III	3	12	48
5	𝍸 I	6	30	150
6	𝍸 𝍸	10	60	360
7	𝍸 𝍸 II	12	84	588
8	𝍸 𝍸 II	12	96	768
9	𝍸 𝍸	10	90	810
10	𝍸 I	6	60	600
11	III	3	33	363
12	I	1	12	144
Total	---	64	480	3840

$$\mu_{\bar{x}} = \frac{\sum f\bar{x}}{\sum f} = \frac{480}{64} = 7.5$$

$$\sigma_{\bar{x}}^2 = \frac{\sum f\bar{x}^2}{\sum f} - \left(\frac{\sum f\bar{x}}{\sum f}\right)^2 = \frac{3840}{64} - \left(\frac{480}{64}\right)^2 = 3.75$$

For verification, we first calculate mean and the variance of the given population. Thus

$$\mu = \frac{3+6+9+12}{4} = \frac{30}{4} = 7.5, \text{ and}$$

$$\sigma^2 = \frac{(3-7.5)^2+(6-7.5)^2+(9-7.5)^2+(12-7.5)^2}{4}$$

$$= \frac{20.25+2.25+2.25+20.25}{4} = 11.25.$$

Hence the statements and verification of the relation desired are:

(i) $\mu_{\bar{x}} = \mu$ *i.e.* $7.5 = 7.5$

(ii) $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} = \frac{11.25}{3} = 3.75$

Q.14.32. (a) Computation of population mean and variance.

$$\mu = \frac{\sum(X_i)}{N} = \frac{4+5+7+9+10}{5} = \frac{35}{5} = 7, \text{ and}$$

$$\sigma^2 = \frac{\sum(X_i-\mu)^2}{N}$$

$$= \frac{(4-7)^2+(5-7)^2+(7-7)^2+(9-7)^2+(10-7)^2}{5}$$

$$= \frac{9+4+0+4+9}{5} = \frac{26}{5}(=5.2)$$

(b) (i) When sampling is done without replacement, then

$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}\cdot\frac{N-n}{N-1}$, where the letter have their usual meangings

$$= \frac{26}{5}\times\frac{1}{3}\times\frac{5-3}{5-1} = \frac{13}{15}(=0.87)$$

(ii) When sampling is done without replacement, then

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} = \frac{26}{5}\times\frac{1}{3} = \frac{26}{15}(=1.73)$$

(c) There are $^5C_3 = 10$ samples of size 3, which can be drawn without replacement. The samples and their means are given in the table below:

No.	Members of Sample	Sample mean ($\bar{x}_i$)	$\bar{x}_i-\mu_{\bar{x}}$	$(\bar{x}_i-\mu_{\bar{x}})^2$
1	4, 5, 7	$5\frac{1}{3}$	$-1\frac{2}{3}$	25/9
2	4, 5, 9	6	−1	1
3	4, 5, 10	$6\frac{1}{3}$	$-\frac{2}{3}$	4/9
4	4, 7, 9	$6\frac{2}{3}$	$-\frac{1}{3}$	1/9
5	4, 7, 10	7	0	0
6	4, 9, 10	$7\frac{2}{3}$	$\frac{2}{3}$	4/9
7	5, 7, 9	7	0	0
8	5, 7, 10	$7\frac{1}{3}$	$\frac{1}{3}$	1/9
9	5, 9, 10	8	1	1
10	7, 9, 10	$8\frac{2}{3}$	$1\frac{2}{3}$	25/9
Σ	--	70	0	78/9

Now, $\mu_{\bar{x}} = \frac{\sum\bar{x}_i}{m} = \frac{70}{10} = 7$, and

$$\sigma_{\bar{x}}^2 = \frac{\sum(\bar{x}_i-\mu_{\bar{x}})^2}{\text{No. of samples}} = \frac{1}{10}\left(\frac{78}{9}\right) = \frac{13}{15}(=0.87)$$

(d) There are $(5)^3 = 125$ samples of size 3, which can be drawn with replacement. Listing all the samples and forming the sampling distribution of means, we find that

$\mu_{\bar{x}} = 7$ and $\sigma_{\bar{x}}^2 = \frac{26}{15}$ as before.

Q.14.33. There are $(4)^3 = 64$ samples of size 3 which can be drawn with replacement from the population 2, 4, 6, 8. The possible samples together with their means and medians are given below:

No.	Sample	Mean	Median	No.	Sample	Mean	Median
1	2,2,2	2	2	33	6,2,2	10/3	2
2	2,2,4	8/3	2	34	6,2,4	4	4
3	2,2,6	10/3	2	35	6,2,6	14/3	6
4	2,2,8	4	2	36	6,2,8	16/3	6
5	2,4,2	8/3	2	37	6,4,2	4	4
6	2,4,4	10/3	4	38	6,4,4	14/3	4
7	2,4,6	4	4	39	6,4,6	16/3	6
8	2,4,8	14/3	2	40	6,4,8	6	6
9	2,6,2	10/3	4	41	6,6,2	14/3	6
10	2,6,4	4	4	42	6,6,4	16/3	6
11	2,6,6	14/3	6	43	6,6,6	6	6
12	2,6,8	16/3	2	44	6,6,8	20/3	6
13	2,8,2	4	2	45	6,8,2	16/3	6
14	2,8,4	14/3	4	46	6,8,4	6	6
15	2,8,6	16/3	6	47	6,8,6	20/3	6
16	2,8,8	6	8	48	6,8,8	22/3	8
17	4,2,2	8/3	2	49	8,2,2	4	2
18	4,2,4	10/3	4	50	8,2,4	14/3	4
19	4,2,6	4	4	51	8,2,6	16/3	6
20	4,2,8	14/3	4	52	8,2,8	6	8
21	4,4,2	10/3	4	53	8,4,2	14/3	4
22	4,4,4	4	4	54	8,4,4	16/3	4
23	4,4,6	14/3	4	55	8,4,6	6	6
24	4,4,8	16/3	4	56	8,4,8	20/3	8
25	4,6,2	4	4	57	8,6,2	16/3	6
26	4,6,4	14/3	4	58	8,6,4	6	6
27	4,6,6	16/3	6	59	8,6,6	20/3	6
28	4,6,8	6	6	60	8,6,8	22/3	8
29	4,8,2	14/3	4	61	8,8,2	6	8
30	4,8,4	16/3	4	62	8,8,4	20/3	8
31	4,8,6	6	6	63	8,8,6	22/3	8
32	4,8,8	20/3	8	64	8,8,8	8	8

The following table gives the sampling distribution of sample means:

$\bar{x}_i$	Tally	f_i	$f_i\bar{x}_i$	$f_i\bar{x}_i^2$
2	I	1	2	4
8/3	III	3	8	$21\frac{1}{3}$
10/3	𝍸 I	6	20	$66\frac{2}{3}$
4	𝍸 𝍸	10	40	160
14/3	𝍸 𝍸 II	12	56	$261\frac{1}{3}$
16/3	𝍸 𝍸 II	12	64	$341\frac{1}{3}$
6	𝍸 𝍸	10	60	360
20/3	𝍸 I	6	40	$266\frac{2}{3}$
22/3	III	3	22	$161\frac{1}{3}$
8	I	1	8	64
Total	---	64	320	$1706\frac{2}{3}$

$$\therefore \quad \mu_{\bar{x}} = \frac{\sum f_i \bar{x}_i}{\sum f_i} = \frac{320}{64} = 5, \text{ and}$$

$$\sigma_{\bar{x}}^2 = \frac{\sum f_i \bar{x}_i^2}{\sum f_i} - \left(\frac{\sum f_i \bar{x}_i}{\sum f_i}\right)^2$$

$$= \frac{1}{64}\left(1706\,\frac{2}{3}\right) - \left(\frac{320}{64}\right)^2 = \frac{5120}{192} - 25 = \frac{5}{3}.$$

The following table gives the sampling distribution of sample medians:

Median (y_i)	Tally	f_i	$f_i y_i$	$f_i y_i^2$
2	𝍸 𝍸	10	20	40
4	𝍸 𝍸 𝍸 𝍸 II	22	88	352
6	𝍸 𝍸 𝍸 𝍸 II	22	132	792
8	𝍸 𝍸	10	80	640
Total	---	64	320	1824

$$\therefore \quad \mu_{(med)} = \frac{\sum f_i v_i}{\sum f_i} = \frac{320}{64} = 5, \text{ and}$$

$$\sigma^2_{(med)} = \frac{\sum f_i v_i^2}{\sum f_i} - \left(\frac{\sum f_i v_i}{\sum f_i}\right)^2 = \frac{1824}{64} - \left(\frac{320}{64}\right)^2$$

$$= \frac{57}{2} - 25 = \frac{7}{2}$$

Population mean is

$$\mu = \frac{2+4+6+8}{4} = \frac{20}{4} = 5, \text{ and}$$

Population varinace is

$$\sigma^2 = \frac{(2-5)^2 + (4-5)^2 + (6-5)^2 + (8-5)^2}{4} = \frac{20}{4} = 5.$$

Comparing the variances, we find that the sampling distribution of means has a small variance.

The mean of the sampling distribution of means equals the mean of the population. The mean of the sampling distribution of medians is also equal to the population mean.

Q.14.34. Population distribution:

x	1	2	3	4	Total
$f(x)$	1/7	3/7	2/7	1/7	1
$xf(x)$	1/7	6/7	6/7	4/7	17/7
$x^2f(x)$	1/7	12/7	18/7	16/7	47/7

Now $\mu = \frac{17}{7} = 2.4286$, and

$$\sigma^2 = \frac{47}{7} - \left(\frac{17}{7}\right)^2 = 6.7143 - 5.8996 = 0.8147$$

The population values as given in the distribution can be written as

1, 2, 2, 2, 3, 3 and 4.

Let *A*, *B*, *C*, *D*, *E*, *F* and *G* stand for population values 1, 2, 2, 2, 3, 3, 4. Then the possible samples of size n=3 and their means are given as follows:

Samples		Means	Samples		Means
ABC	1,2,2	5/3	*BCG*	2,2,4	8/3
ABD	1,2,2	5/3	*BDE*	2,2,3	7/3
ABE	1,2,3	6/3	*BDF*	2,2,3	7/3
ABF	1,2,3	6/3	*BDG*	2,2,4	8/3
ABG	1,2,4	7/3	*BEF*	2,3,3	8/3
ACD	1,2,2	5/3	*BEG*	2,3,4	9/3
ACE	1,2,3	6/3	*BFG*	2,3,4	9/3
ACF	1,2,3	6/3	*CDE*	2,2,3	7/3
ACG	1,2,4	7/3	*CDF*	2,2,3	7/3
ADE	1,2,3	6/3	*CDG*	2,2,4	8/3
ADF	1,2,3	6/3	*CEF*	2,3,3	8/3
ADG	1,2,4	7/3	*CEG*	2,3,4	9/3
AEF	1,3,3	7/3	*CFG*	2,3,4	9/3
AEG	1,3,4	8/3	*DEF*	2,3,3	8/3
AFG	1,3,4	8/3	*DEG*	2,3,4	9/3
BCD	2,2,2	6/3	*DFG*	2,3,4	9/3
BCE	2,2,3	7/3	*EFG*	3,3,4	10/3
BCF	2,2,3	7/3			

The sampling distribution of means and calculation of $\sigma^2_{\bar{x}}$.

$\bar{x}_i$	5/3	6/3	7/3	8/3	9/3	10/3	Total
$f(\bar{x})$	3/35	7/35	10/35	8/35	6/35	1/35	1
$\bar{x}f(\bar{x})$	10/105	42	70	64	54	10	255/105
$\bar{x}^2f(\bar{x})$	75/315	252	490	512	486	100	1915/315

Now $\sigma^2_{\bar{x}} = \sum \bar{x}^2 f(\bar{x}) - [\sum \bar{x} f(\bar{x})]^2 = \frac{1915}{315} - \left(\frac{255}{105}\right)^2$

$$= 6.0794 - (2.4286)^2 = 6.0794 - 5.8981 = 0.1813$$

And $\frac{\sigma^2}{n} \cdot \frac{N-n}{N-1} = \frac{0.8147}{3} \cdot \frac{7-3}{7-1} = \frac{0.8147}{3} \times \frac{4}{6} = 0.1810$

(Slight difference due to rounding off)

Q.14.35. Here $n = 100$. $\mu = 20$, $\sigma = 5$.

(i) Sampling Distribution of $\bar{X}$ will be approximately normal as sample size is large enough.

(ii) $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \dfrac{\bar{X} - 20}{5/\sqrt{100}} = \dfrac{\bar{X} - 20}{0.5}$ is approximately $N(0, 1)$.

At $\bar{X} = 20.75$, we find that $z = \dfrac{20.75 - 20}{0.5} = \dfrac{0.75}{0.5} = 1.5$

$\therefore \quad P(\bar{X} > 20.75) = P(Z > 1.5) = 0.5 - 0.4332 = 0.0668$

Q.14.36. The Sampling distribution is normal with $\mu_{\bar{x}} = 60$ and $\sigma_{\bar{x}} = \dfrac{10}{\sqrt{64}}$.

Then $\quad Z = \dfrac{\bar{X} - 60}{10/8}$

Now at $\bar{X} = 59$, we find that $z = \dfrac{59 - 60}{10/8} = \dfrac{-1}{1.25} = -0.80$, and

at $\bar{X} = 61$, $z = \dfrac{61 - 60}{1.25} = \dfrac{1}{1.25} = +0.80$

$\therefore \quad P(59 < \bar{X} < 61) = P(-0.80 < Z < 0.80) = 2(0.2881) = 0.5762$

Q.14.37. Given a normally distributed population with the following data:

N = 1,000, μ = 68.5 inches, σ = 2.7 inches

(a) Determination of expected mean, $\mu_{\bar{x}}$ and the standard deviation of the sampling distribution of the mean, $\sigma_{\bar{x}}$.

$\mu_{\bar{x}} = \mu = 68.5$ inches

$\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{2.7}{5}$ inches. when sampling is with replacement, or

$\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}} \sqrt{\dfrac{N-n}{N-1}}$. when sampling is without replacement.

$= \dfrac{2.7}{5} \sqrt{\dfrac{1000 - 25}{1000 - 1}} = 0.54 \times 0.9879 = 0.5335$ inches.

(b) The sample mean $\bar{X}$ in standard units is given by

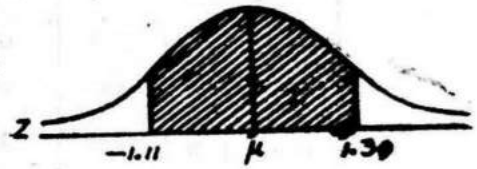
$$Z = \frac{\bar{X} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{X} - 68.5}{0.54}$$

Then at $\bar{x}$=67.9, $z = \dfrac{67.9 - 68.5}{0.54}$

$= \dfrac{-0.6}{.54} = -1.11$, and

at $\bar{x}$=69.2, $z = \dfrac{67.9 - 68.5}{0.54}$

$= \dfrac{0.7}{0.54} = 1.30$



Thus the proportion of samples with means between 67.9 inches and 69.2 inches = Area under normal curve between $z = -1.11$ and $z = +1.30$

= (Area between $z = 0$ to $z = -1.11$) + (Area between z=0 to z=1.30)

$= 0.3665 + 0.4032 = 0.7697$

Hence the expected number of samples = 200×0.7697=154.

Q.14.38. Given μ=1.14 m, σ=0.25m and n=100.

Then $\quad Z = \dfrac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \dfrac{\bar{X} - 1.14}{0.25 / 10}$ is $N(0, 1)$.

(i) $P(\bar{X} \geq 1.16) = P\left(Z \geq \dfrac{1.16 - 1.14}{0.025}\right)$

$= P(Z \geq 0.8) = 0.5 - 0.2881 = 0.2119$

$\therefore$ The number of samples with mean sample greater than 1.16m

$= (50)\,(0.2119) = 11.$

(ii) $P(1.13 \leq \bar{X} \leq 1.18) = P\left(\dfrac{1.13 - 1.14}{0.025} \leq Z \leq \dfrac{1.18 - 1.14}{0.025}\right)$

$= P(-0.4 \leq Z \leq 1.6)$

$= 0.1554 + 0.4452 = 0.6006$

Hence the number of samples with the means being between 1.13m and 1.18m = (50) (0.6006) = 30.

Q.14.39 (a) Computation of standard error.

I.Q.	f	u	fu	fu^2
80 – 89	30	–3	–90	270
90 – 99	52	–2	–104	208
100 – 109	75	–1	–75	75
110 – 119	109	0	–269	0
120 – 129	65	1	65	65
130 – 139	42	2	84	168
140 – 149	27	3	81	243
Σ	400	--	230 / –39	1029

$$\bar{x} = 114.5 + \frac{(-39)}{400} \times 10 = 114.5 - 0.975 = 113.525$$

$$s = h\sqrt{\frac{\sum fu^2}{n} - \left(\frac{\sum fu}{n}\right)^2} = 10\sqrt{\frac{1029}{400} - (-39/400)^2}$$

$$= 10\sqrt{2.5729 - 0.0095} = 10\,(1.6011) = 16.011$$

$$S.E. = \frac{s}{\sqrt{n}} = \frac{16.011}{20} = 0.80$$

(b) Computation of Mean and Variance.

x	$f(x)$	$xf(x)$	$x^2f(x)$
4	0.2	0.8	3.2
5	0 4	2.0	10.0
6	0.3	1.8	10.8
7	0.1	0.7	4.9
Σ	1.0	5.3	28.9

Now $\mu = \sum xf(x) = 5.3$, and

$$\sigma^2 = E(X^2) - [E(X)]^2 = \sum x^2 f(x) - [\sum xf(x)]^2$$

$$= 28.9 - (5.3)^2 = 28.9 - 28.09 = 0.81$$

For a random sample of $n = 36$, we have

(i) $\mu_{\bar{x}} = \mu = 5.3$, and

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} = \frac{0.81}{36} = 0.0225$$

(ii) The sample size is large enough to assume that the sampling distribution of $\bar{X}$ is approximately normal, i.e.

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{X} - 5.3}{0.15} \text{ is approximately } N(0, 1)$$

At $\bar{X} = 5.5$, $z = \dfrac{5.5 - 5.3}{0.15} = \dfrac{0.2}{0.15} = 1.33.$

$$\therefore \quad P(\bar{X}<5.5) = P(Z<1.33) = 0.5 + 0.4082 = 0.9082$$

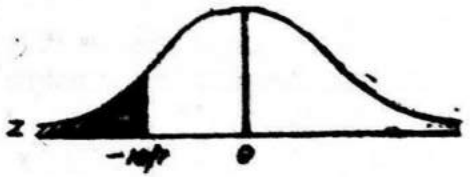
Q.14.40. (b) Given a normal distribution with μ=0.1, σ=2.1 and sample size, n = 900. Let $\bar{x}$ be the mean of the sample. Then the standard normal variate is given by

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{x} - 0.1}{2.1/30}$$

$$\therefore \quad \bar{x} = 0.1 + \frac{2.1}{30}(Z) = \frac{1}{10} + \frac{7}{100}(Z)$$

Now $\bar{x}$ will be negative if

$$\frac{1}{10} + \frac{7}{100}(Z) < 0$$



i.e. $Z < -\dfrac{10}{7}$

Hence $P\left(Z < -\dfrac{10}{7}\right)$ = shaded area in the figure

$$= 0.5 - P(0 < Z < 10/7)$$

$$= 0.5 - 0.4236 = 0.0764.$$

Q.14.41 (a) Let X denote the Binomial distribution with n=40 and p=0.5. Then

$$\mu = np = (40)\,(0.5) = 20, \text{ and}$$

$$\sigma = \sqrt{npq} = \sqrt{(40)\,(0.5)\,(0.5)} = \sqrt{10} = 3.162$$

By the *central limit theorem*, $\bar{X}$ is approximately $N(20, \frac{10}{100})$

Thus $Z = \dfrac{\bar{X} - 20}{\sqrt{10/100}} = \dfrac{\bar{X} - 20}{0.3162}$

(i) We require $P(\bar{X} > 20.5)$

$\therefore \quad P(\bar{X} > 20.5) = P\left(Z > \dfrac{20.5 - 20}{0.3162}\right) = P(Z > 1.58)$

$= 0.5 - 0.4429 = 0.0571$

(ii) We require $P(\bar{X} < 19.3)$

$\therefore \quad P(\bar{X} < 19.3) = P\left(Z < \dfrac{19.3 - 20}{0.3162}\right)$

$= P(Z < -2.21) = 0.0136$

(iii) We require $P(19.3 < \bar{X} < 20.5)$

Thus $P(19.3 < \bar{X} < 20.5) = P(-2.21 < Z < 1.58)$

$= 0.4864 + 0.4429 = 0.9293$

(b) Given $\mu = 2$ and $\sigma = 3$.

As the sample size $n = 36$ is large enough to use the *central limit theorem*, therefore the sampling distribution of $\bar{X}$ is approximately normal. Thus $Z = \dfrac{\bar{X} - 2}{3/\sqrt{36}} = \dfrac{\bar{X} - 2}{0.5}$, and $\bar{X} = 2 + 0.5Z$.

We are required to find the probability that $\bar{X}$ will be negative.

Now $\bar{X}$ will be negative if $(2 + 0.5Z) < 0$, i.e. if $Z < -4$.

$\therefore \quad P(Z < -4) = 0$ approximately.

Q.14.42. (b) $\mu_1 = 10$, $\mu_2 = 8$, $\sigma_1 = 2$, $\sigma_2 = 1$

(i) $Z = \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} = \dfrac{(\bar{X}_1 - \bar{X}_2) - (10 - 8)}{\sqrt{4/100 + 1/100}} = \dfrac{(\bar{X}_1 - \bar{X}_2) - 2}{2.236/10}$

Now at $\bar{X}_1 - \bar{X}_2 = 1.5$, we find that $z = \dfrac{1.5 - 2}{0.2234} = -2.236$.

$\therefore \quad P(\bar{X}_1 - \bar{X}_2 < 1.5) = P(Z < -2.236) = 0.0127$.

(ii) At $\bar{X}_1 - \bar{X}_2 = 1.75$, we find that $z = \dfrac{1.75 - 2.0}{0.2236} = -1.118$ and

At $\bar{X}_1 - \bar{X}_2 = 2.5$, $z = \dfrac{2.5 - 2.0}{0.2236} = 2.236$.

$\therefore \quad P(1.75 < \bar{X}_1 - \bar{X}_2 < 2.5) = P(-1.118 < Z < 2.236)$

$= P(-1.12 < Z < 0) + P(0 < Z < 2.24)$

$= 0.3686 + 0.4875 = 0.8561$.

Q.14.43. There are $(3)^2 = 9$ possible samples of size 2, which can be drawn with replacement from each population. These two sets of samples and their means are given below:

From Population I			From Population II		
No.	Sample	$\bar{x}_1$	No.	Sample	$\bar{x}_2$
1	3, 3	3	1	A, A: 1, 1	1
2	3, 4	3.5	2	A, B: 1, 1	1
3	3, 5	4	3	A, C: 1, 3	2
4	4, 3	3.5	4	B, A: 1, 1	1
5	4, 4	4	5	B, B: 1, 1	1
6	4, 5	4.5	6	B, C: 1, 3	2
7	5, 3	4	7	C, A: 3, 1	2
8	5, 4	4.5	8	C, B: 3, 1	2
9	5, 5	5	9	C, C: 3, 3	3

(a) The 81 possible differences $\bar{X}_1 - \bar{X}_2$ are presented in the following table:

	Differences of Independent means								
$\bar{x}_2$ \ $\bar{x}_1$	3	3.5	4	3.5	4	4.5	4	4.5	5
1	2	2.5	3	2.5	3	3.5	3	3.5	4
1	2	2.5	3	2.5	3	3.5	3	3.5	4
2	1	1.5	2	1.5	2	2.5	2	2.5	3
1	2	2.5	3	2.5	3	3.5	3	3.5	4
1	2	2.5	3	2.5	3	3.5	3	3.5	4
2	1	1.5	2	1.5	2	2.5	2	2.5	3
2	1	1.5	2	1.5	2	2.5	2	2.5	3
2	1	1.5	2	1.5	2	2.5	2	2.5	3
3	0	0.5	1	0.5	1	1.5	1	1.5	2

(b) The Sampling Distribution of $\bar{X}_1 - \bar{X}_2$ and computation of its mean and variance.

$\bar{X}_1 - \bar{X}_2 (=d_i)$	Tally	f_i	$f_i d_i$	$f_i d_i^2$
0	I	1	0	0
0.5	II	2	1	0.5
1	₩ II	7	7	7.0
1.5	₩ ₩	10	15	22.5
2	₩ ₩ ₩ II	17	34	68.0
2.5	₩ ₩ ₩ I	16	40	100.0
3	₩ ₩ ₩ I	16	48	144.0
3.5	₩ III	8	28	98.0
4	IIII	4	16	64.0
Σ	---	81	189	504.0

$$\therefore \text{Mean, } \mu_{\bar{x}_1 - \bar{x}_2} = \frac{\Sigma f_i d_i}{\Sigma f_i} = \frac{189}{81} = 2.33, \text{ and}$$

$$\sigma^2_{\bar{x}_1 - \bar{x}_2} = \frac{\Sigma f_i d_i^2}{\Sigma f_i} - \left(\frac{\Sigma f_i d_i}{\Sigma f_i}\right)^2$$

$$= \frac{504}{81} - \left(\frac{189}{81}\right)^2 = \frac{5103}{6561} = \frac{7}{9}$$

(c) **Verification.** The mean and variance of the first population are:

$$\mu_1 = \frac{3+4+5}{3} = \frac{12}{3} = 4, \text{ and}$$

$$\sigma_1^2 = \frac{(3-4)^2 + (4-4)^2 + (5-4)^2}{3} = \frac{2}{3}.$$

The mean and variance of the second population are

$$\mu_2 = \frac{1+1+3}{3} = \frac{5}{3} = 1.67, \text{ and}$$

$$\sigma_2^2 = \frac{11}{3} - \frac{25}{9} = \frac{8}{9}.$$

Now, $\mu_{\bar{x}_1 - \bar{x}_2} = 2.33 = 4.00 - 1.67 = \mu_1 - \mu_2$, and

$$\sigma^2_{\bar{x}_1 - \bar{x}_2} = \frac{7}{9} = \frac{1}{3} + \frac{4}{9}$$

$$= \frac{1}{2}\left(\frac{2}{3}\right) + \frac{1}{2}\left(\frac{8}{9}\right) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Hence the result.

Q.14.44. Given that

$\mu_1 = 6.5$, $\mu_2 = 6.0$, $\sigma_1 = 0.9$, $\sigma_2 = 0.8$

$n_1 = 36$, $n_2 = 49$ and $\bar{x}_1 - \bar{x}_2 \geq 1$

The sampling distribution of the difference between means $(\bar{X}_1 - \bar{X}_2)$ is normally distributed with a mean of $\mu_1 - \mu_2$ and standard deviation equal to $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

This means that

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \text{ is } N(0, 1)$$

Now $$z = \frac{1 - (6.5 - 6.0)}{\sqrt{\frac{(0.9)^2}{36} + \frac{(0.8)^2}{49}}} = \frac{1 - 0.5}{0.19} = \frac{0.5}{0.19} = 2.63$$

$\therefore \quad P(\bar{X}_1 - \bar{X}_2 \geq 1) = P(Z \geq 2.63) = 0.5 - P(0 \leq Z \leq 2.63)$

$= 0.5 - 0.4957 = 0.0043.$

Q.14.45. Given $n_1 = 25, \quad \mu_1 = 80, \quad \sigma_1 = 5;$

$n_2 = 36, \quad \mu_2 = 75, \quad \sigma_2 = 3;$

and we are required to find $P[3.4 \leq \bar{x}_1 - \bar{x}_2 < 5.9]$

As the populations are normal, therefore the sampling distribution of the difference in means will be normal with mean

$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2 = 80 - 75 = 5$, and standard deviation

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{(5)^2}{25} + \frac{(3)^2}{36}}$$

$$= \sqrt{\frac{25}{25} + \frac{9}{36}} = \sqrt{1 + 0.25} = \sqrt{1.25} = 1.12,$$

Thus the variable

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_{\bar{x}_1 - \bar{x}_2}}{\sigma_{\bar{x}_1 - \bar{x}_2}} = \frac{(\bar{x}_1 - \bar{x}_2) - 5}{1.12} \text{ is } N(0, 1)$$

Now 3.4 in standard units $= \frac{3.4 - 5}{1.12} = -1.43$, and

5.9 in standard units $= \frac{5.9 - 5}{1.12} = 0.80$

Hence $P[3.4 \leq \bar{x}_1 - \bar{x}_2 < 5.9] = P(-1.43 \leq Z \leq 0.80)$

$= P(-1.43 \leq Z \leq 0) + P(0 \leq Z \leq 0.80)$

$= 0.4236 + 0.2881 = 0.7117$

Q.14.47. The sample proportions of even number in 20 possible samples.

Samples	$\hat{P}$	Samples	$\hat{P}$
0, 3, 4	2/3	3, 4, 6	2/3
0, 3, 6	2/3	3, 4, 9	1/3
0, 3, 9	1/3	3, 4, 15	1/3
0, 3, 15	1/3	3, 6, 9	1/3
0, 4, 6	3/3	3, 6, 15	1/3
0, 4, 9	2/3	3, 9, 15	0/3
0, 4, 15	2/3	4, 6, 9	2/3
0, 6, 9	2/3	4, 6, 15	2/3
0, 6, 15	2/3	4, 9, 15	1/3
0, 9, 15	1/3	6, 9, 15	1/3

The sampling distribution of proportions is:

$\hat{P}$	0/3	1/3	2/3	3/3	Total
$f(\hat{p})$	1/20	9/20	9/20	1/20	1
$\hat{p}\, f(\hat{p})$	0/60	9/60	18/60	3/60	30/60=0.5
$\hat{p}^2 f(\hat{p})$	0/180	9/180	36/180	9/180	54/180

$$\mu_{\hat{p}} = \frac{3}{6} = 0.5$$

$$\text{Var}(\hat{p}) = \sum \hat{p}^2 f(\hat{p}) - [\sum \hat{p} f(\hat{p})]^2$$

$$= \frac{54}{180} - \left(\frac{30}{60}\right)^2 = 0.3 - 0.25 = 0.05$$

and $$\frac{pq}{n} \cdot \frac{N-n}{N-1} = \frac{(0.5)(0.5)}{3} \cdot \frac{6-3}{6-1}$$

$$= \frac{0.25}{3} \cdot \frac{3}{5} = 0.05 = \text{Var}(\hat{p})$$

Q.14.48. Let A, B, C, D, E, F and G stand for population values 1, 1, 2, 3, 4, 4 and 5 respectively. Then the possible samples of size n=3 that could be drawn without replacemnet and the proportion of odd numbers in the samples are given below. (Possible samples are 35)

Samples		Sample Proportion ($\hat{p}$)	Samples		Sample Proportion ($\hat{p}$)
ABC	1,1,2	2/3	BCG	1,2,5	2/3
ABD	1,1,3	3/3	BDE	1,3,4	2/3
ABE	1,1,4	2/3	BDF	1,3,4	2/3
ABF	1,1,4	2/3	BDG	1,3,5	3/3
ABG	1,1,5	3/3	BEF	1,4,4	1/3
ACD	1,2,3	2/3	BEG	1,4,5	2/3
ACE	1,2,4	1/3	BFG	1,4,5	2/3
ACF	1,2,4	1/3	CDE	2,3,4	1/3
ACG	1,2,5	2/3	CDF	2,3,4	1/3
ADE	1,3,4	2/3	CDG	2,3,5	2/3
ADF	1,3,4	2/3	CEF	2,4,4	0/3
ADG	1,3,5	3/3	CEG	2,4,5	1/3
AEF	1,4,4	1/3	CFG	2,4,5	1/3
AEG	1,4,5	2/3	DEF	3,4,4	1/3
AFG	1,4,5	2/3	DEG	3,4,5	2/3
BCD	1,2,3	2/3	DFG	3,4,5	2/3
BCE	1,2,4	1/3	EFG	4,4,5	1/3
BCF	1,2,4	1/3			

The Sampling distribution of proportion and other calculations are given below:

$\hat{p}$	Tally	f	$f(\hat{p})$	$\hat{p}\,f(\hat{p})$	$\hat{p}^2 f(\hat{p})$
0/3	I	1	1/35	0	0
1/3	𝍸 𝍸 II	12	12/35	4/35	4/105
2/3	𝍸 𝍸 𝍸 III	18	18/35	12/35	8/35
3/3	IIII	4	4/35	4/35	4/35
Σ	---	35	1	20/35	40/105

Now $\mu_{\hat{p}} = \Sigma \hat{p} f(\hat{p}) = \frac{20}{35} = \frac{4}{7}$ and

$$\sigma^2_{\hat{p}} = \Sigma \hat{p}^2 f(\hat{p}) - [\Sigma \hat{p} f(\hat{p})]^2 = \frac{40}{105} - \left(\frac{4}{7}\right)^2 = \frac{40}{735} = \frac{8}{147}$$

Again $p = \frac{4}{7}$ and $\frac{pq}{n} \cdot \frac{N-n}{N-1} = \frac{1}{3}\left(\frac{4}{7}\right)\left(\frac{3}{7}\right)\frac{7-3}{7-1} = \frac{8}{147}$.

Hence the result.

Q.14.49. (a) Using the Sampling Distribution of Sample Proportion.

Let $\hat{P}$ be the r.v. the proportion of diseased plants in the sample. Then, (as the sample size (n=250) is large enough) $\hat{p}$ is approximately normally distributed with mean

$$\mu_{\hat{p}} = p = 0.02, \text{ and standard error}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.02)\,(0.98)}{250}} = 0.008854$$

Thus $Z = \dfrac{\hat{p} - 0.02}{0.008854}$ is approximately $N(0, 1)$

(i) We require $P(\hat{p} < 0.01)$

$$P(\hat{p}<0.01) \Rightarrow P\left(\hat{p}<0.01 - \frac{1}{2(250)}\right) \text{ (continuity correction)}$$

$$= P\left(\frac{\hat{p} - 0.02}{0.008854} < \frac{(0.01 - 1/500) - 0.02}{0.008854}\right)$$

$$= P(Z < -1.355) = 0.5 - 0.4123 = 0.0877$$

Thus the probability that less than 1% of the trees are diseased is 0.0877.

(ii) We require $P(\hat{p} > 0.04)$

Now $P(\hat{p}>0.04) \Rightarrow P\left(\hat{p}>0.04 + \frac{1}{2(250)}\right)$ (continuity correction)

$$= P\left(Z > \frac{(0.04 + 1/500) - 0.02}{0.008854}\right)$$

$$= P(Z > 2.485) = 0.5 - 0.4935 = 0.0065$$

Hence the probability that more than 4% of the trees are diseased is 0.0065.

Alternatively. *Using the Normal approximation to the Binomial Distribution.*

Let X be the r.v., the number of diseased plants in the sample. Then

X is approximately $N(np, npq)$, where n=250, p=0.02 and q=0.98.

Now $np = 0.02 \times 250 = 5$ and $npq = (0.02)\,(0.98)\,(25) = 4.90$

That is X is $N(5, 490)$

(i) As 1% of 250 = 2.5, so we require $P(X<2.5)$.

$$P(X<2.5) \Rightarrow P(X < 2.5 - 0.5) \quad \text{(continuity correction)}$$

$$= P\left(\frac{X-5}{\sqrt{4.9}} < \frac{(2.5-0.5)-5}{2.2136}\right)$$

$$= P(Z < -1.355) = 0.5 - 0.4123 = 0.0877$$

(ii) As 4% of 250 = 10, so we require $P(X>10)$.

$$P(X>10) \Rightarrow P(X > 10 + 0.5) \quad \text{(continuity correction)}$$

$$= P\left(\frac{X-5}{2.2136} > \frac{(10.5-5)}{2.2136}\right)$$

$$= P(Z > 2.485) = 0.5 - 0.4935 = 0.0065$$

(b) Here n = 150, p = 0.6, so that

$$S.E. = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.24}{150}} = 0.04$$

Using Normal approximation, we get

$$P(\hat{p} < 0.52) = P\left(Z < \frac{0.52 - 0.6}{0.04}\right) = P(Z < -2) = 0.0228$$

Q.14.50. $\mu_{\hat{p}} = 0.7$, $\sigma_{\hat{p}} = \sqrt{\frac{(0.7)(0.3)}{400} \cdot \frac{4500 - 4000}{4500 - 1}}$

$= (0.0229)(0.9546) = 0.02186$

$$P[-0.05 \le \hat{p} - p \le 0.05] = P\left[\frac{-0.05}{0.0219} < \frac{\hat{p} - p}{\sigma_{\hat{p}}} < \frac{0.05}{0.0219}\right]$$

$= P[-2.28 < Z < 2.28)] = 2(0.4887)$

$= 0.9774$

Q.14.51. (b) $n_1 = 40$, $n_2 = 45$, $p = 0.70$

$$\therefore \quad S.E. = \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{(0.7)(0.3)\left(\frac{1}{40} + \frac{1}{45}\right)}$$

$= \sqrt{(0.21)(0.0472)} = 0.0996$, and

the standard variable is

$$Z = \frac{\hat{P}_1 - \hat{P}_2}{S.E.} = \frac{\hat{P}_1 - \hat{P}_2}{0.0996}.$$

At $\hat{P}_1 - \hat{P}_2 = -0.1$, we find $z = \frac{-0.1}{0.0996} = -1.004$

Hence $P(-1 < \hat{P}_1 - \hat{P}_2 < 1) = P(-1.004 < Z < 1.004)$

$= P(-1.004 < Z < 0) + P(0 < Z < 1.004)$

$= 2(0.3413) = 0.6826$

Q.14.52. Computation of the variance of the population.

$$\mu = \frac{\sum X_i}{N} = \frac{1 + 2 + 3 + 4 + 5}{5} = \frac{15}{5} = 3,$$

$$\sigma^2 = \frac{\sum (X_i - \mu)^2}{N}$$

$$= \frac{(1-3)^2 + (2-3)^2 + \ldots + (5-3)^2}{5} = \frac{10}{5} = 2.$$

The possible samples of size 2 with replacement and their variances are given as follows:

No.	Members of Sample	Mean $\bar{x}_i$	Variance S_i^2
1	1, 1	1	0
2	1, 2	1.5	0.25
3	1, 3	2	1
4	1, 4	2.5	2.25
5	1, 5	3.	4
6	2, 1	1.5	0.25
7	2, 2	2	0
8	2, 3	2.5	0.25
9	2, 4	3	1
10	2, 5	3.5	2.25
11	3, 1	2	1
12	3, 2	2.5	025
13	3, 3	3	0
14	3, 4	3.5	0.25
15	3, 5	4	1
16	4, 1	2.5	2.25
17	4, 2	3	1
18	4, 3	3.5	0.25
19	4,4	4	0
20	4, 5	4.5	0.25
21	5, 1	3	4
22	5, 2	3.5	2.25
23	5, 3	4	1
24	5, 4	4.5	0.25
25	5, 5	5	0

The following table gives the Sampling Distribution of sample variances.

S_i^2	Tally	f_i	$f_i S_i^2$
0	𝍸	5	0
0.25	𝍸 III	8	2
1	𝍸 I	6	6
2.25	IIII	4	9
4	II	2	8
Total	--	25	25

Now the mean of the sampli..g distribution of variance is

$$\mu_{S^2} = \frac{\sum f_i S_i^2}{\sum f_i} = \frac{25}{25} = 1.$$

But this is not equal to the population variance. We may write

$$\mu_{S^2} = 1 = \frac{(2-1)}{2} \cdot 2 \quad (\because \quad n = 2)$$

$$= \frac{n-1}{n} \cdot \sigma^2$$

That is why sampling variance needs a modified definition to make it an unbiased estimetor of the population variance. Hence the desirability of using definition of the type

$$s^2 = \frac{\sum (x - \bar{x})^2}{n-1}$$

rather than the definition of the type

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n}$$

❖❖❖❖❖❖❖❖

CHAPTER 15

STATISTICAL INFERENCE: ESTIMATION

Q.15.4. (a) Computation of Mean, Variance and Third Mean Moment.

X	$X-\bar{X}$	$(X-\bar{X})^2$	$(X-\bar{X})^3$
8	2.1	4.41	9.261
4	−1.9	3.61	−6.859
10	4.1	16.81	68.921
5	−0.9	0.81	−0.729
5	−0.9	0.81	−0.729
4	−1.9	3.61	−6.859
9	3.1	9.61	29.791
4	−1.9	3.61	−6.859
3	−2.9	8.41	−24.389
7	1.1	1.21	1.331
59	0	52.90	62.880

$\therefore$ Mean, $\bar{X} = \frac{\sum X}{n} = \frac{59}{10} = 5.9,$

Var(X), $m_2 = \frac{\sum (X-\bar{X})^2}{n} = \frac{52.90}{10} = 5.29$, and

$$m_3 = \frac{\sum (X-\bar{X})^3}{n} = \frac{62.880}{10} = 6.288$$

Hence the point estimates of mean, variance and third moment are 5.9, 5.29 and 6.29 respectively.

(b) Estimation of μ and Standard Error.

(i) Now $\bar{x} = \frac{\sum x_i}{n} = \frac{852}{70} = 12.17$, which is the point estimate of μ.

$$S = \sqrt{\frac{1}{n}\sum(x_i - \bar{x})^2} = \sqrt{\frac{215}{70}} = \sqrt{3.0714} = 1.75,$$ which is the point estimate of σ.

$\therefore$ $S_{\bar{x}} = \frac{S}{\sqrt{n}} = \frac{1.75}{\sqrt{70}} = \frac{1.75}{8.37} = 0.21$, is the required estimated standard error.

(ii) The point estimate of μ is

$\bar{x} = \frac{\sum x_i}{n} = \frac{1985}{160} = 12.41$, and the point estimate of σ is

$$S = \sqrt{\frac{1}{n}\sum(x_i - \bar{x})^2} = \sqrt{\frac{475}{160}} = \sqrt{2.9688} = 1.72$$

Thus the estimated standard error is

$$S_{\bar{x}} = \frac{S}{\sqrt{n}} = \frac{1.72}{\sqrt{160}} = \frac{1.72}{12.65} = 0.14.$$

Q.15.5. (b) Population consists of 10, 12, 14, 16, 18 and 20.

Now $\mu = \frac{\sum X_i}{N} = \frac{10+12+14+16+18+20}{6} = \frac{90}{6} = 15$; and

$$\sigma^2 = \frac{\sum(X_i-\mu)^2}{N}$$

$$= \frac{(10-15)^2+(12-15)^2+\ldots+(18-15)^2+(20-15)^2}{6}$$

$$= \frac{70}{6} = 11.67$$

The possible samples of size 2 which can be drawn with replacement together with the values of $\bar{x}$ and s^2 are given as follows:

No.	Sample	$\bar{x}_i$	s^2	No.	Sample	$\bar{x}_i$	s^2
1	10, 10	10	0	19	16, 10	13	18
2	10, 12	11	2	20	16, 12	14	8
3	10, 14	12	8	21	16, 14	15	2
4	10, 16	13	18	22	16, 16	16	0
5	10, 18	14	32	23	16, 18	17	2
6	10, 20	15	50	24	16, 20	18	8
7	12, 10	11	2	25	18, 10	14	32
8	12, 12	12	0	26	18, 12	15	18
9	12, 14	13	2	27	18, 14	16	8
10	12, 16	14	8	28	18, 16	17	2
11	12, 18	15	18	29	18, 18	18	0
12	12, 20	16	32	30	18, 20	19	2
13	14, 10	12	8	31	20, 10	15	50
14	14, 12	13	2	32	20, 12	16	32
15	14, 14	14	0	33	20, 14	17	18
16	14, 16	15	2	34	20, 16	18	8
17	14, 18	16	8	35	20, 18	19	2
18	14, 20	17	18	36	20, 20	20	0

Now $\bar{X}$ is an unbiased estimator of μ if $E(\bar{X}) = \mu$. To find $E(\bar{X})$, we obtain the smapling distribution of $\bar{X}$ as below:

$\bar{X}$	Tally	f	Probability: $f(\bar{x})$	$\bar{x}\,f(\bar{x})$
10	I	1	1/36	10/36
11	II	2	2/36	22/36
12	III	3	3/36	36/36
13	IIII	4	4/36	52/36
14	~~IIII~~	5	5/36	70/36
15	~~IIII~~ I	6	6/36	90/36
16	~~IIII~~	5	5/36	80/36
17	IIII	4	4/36	68/36
18	III	3	3/36	54/36
19	II	2	2/36	38/36
20	I	1	1/36	20/36
Total	--	36	1	540/36

Therefore $E(\bar{X}) = \sum \bar{x} f(\bar{x}) = \frac{540}{36} = 15 = \mu$.

Again s^2 is an unbiased estimator of the population variance σ^2 if $E(s^2) = \sigma^2$. To find $E(s^2)$, we have the sampling distribution of s^2 as follows:

s^2	Tally	f	fs^2
0	⧻ I	6	0
2	⧻ ⧻	10	20
8	⧻ III	8	64
18	⧻ I	6	108
32	IIII	4	128
50	II	2	100
Σ	--	36	420

$\therefore E(s^2) = \frac{1}{\sum f}(\sum fs^2) = \frac{420}{36} = 11.67 = \sigma^2$. Hence the result.

Q.15.6. (b) To find the relative efficiency, we compare the expected values and the variances of these two statistics:

Now $$E(T_1) = E\left[\frac{X_1 + 2X_2 + X_3}{4}\right]$$

$$= \frac{1}{4}[E(X_1) + 2E(X_2) + E(X_3)]$$

$$= \frac{1}{4}[\mu + 2\mu + \mu] = \mu$$

$\therefore$ T_1, *i.e.* the weighted sample mean, is an unbiased estimator of μ.

Again $$E(T_2) = E(\bar{X}) = E\left[\frac{X_1 + X_2 + X_3}{4}\right] = \mu.$$

We see that T_2 is also an unbiased estimator of μ.

Next we find their variances as:

$$\text{Var}(T_1) = \text{Var}\left[\frac{X_1 + 2X_2 + X_3}{4}\right]$$

$$= \text{Var}\left(\frac{X_1}{4}\right) + \text{Var}\left(\frac{2X_2}{4}\right) + \text{Var}\left(\frac{X_3}{4}\right)$$

$$= \frac{1}{16}\text{Var}(X_1) + \frac{4}{16}\text{Var}(X_2) + \frac{1}{16}\text{Var}(X_3)$$

$$= \frac{1}{16}\sigma^2 + \frac{4}{16}\sigma^2 + \frac{1}{16}\sigma^2 = \frac{6}{16}\sigma^2; \text{ and}$$

$$\text{Var}(T_2) = \text{Var}\left[\frac{X_1 + X_2 + X_3}{3}\right]$$

$$= \frac{1}{9}[\text{Var}(X_1) + \text{Var}(X_2) + \text{Var}(X_3)] = \frac{1}{3}\sigma^2.$$

$\therefore$ Efficiency of $T_1 = \frac{\text{Var}\,(T_2)}{\text{Var}\,(T_1)} = \frac{\sigma^2}{3} \times \frac{16}{6\sigma^2} = \frac{8}{9} = 89\%$.

Q.15.7. (b) The possible samples with the values of $\bar{x}$ and S^2 are given below:

No.	Sample	$\bar{x}_i$	$S^2 = \frac{1}{n}\sum (x_i - \bar{x})^2$
1	3, 3	3	0
2	3, 5	4	1
3	3, 2	2.5	0.25
4	5, 3	4	1
5	5, 5	5	0
6	5, 2	3.5	2.25
7	2, 3	2.5	0.25
8	2, 5	3.5	2.25
9	2, 2	2	0
Σ	--	30.0	7.00

Now, $\mu = \frac{3 + 5 + 2}{3} = 3\frac{1}{3}$, and

$$\sigma^2 = \frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2 = \frac{38}{3} - \left(\frac{10}{3}\right)^2 = \frac{14}{9}$$

Thus $E(S^2) = \frac{7}{9} \neq \frac{14}{9} \neq \sigma^2$.

Q.15.8. (b) Let $T = \dfrac{X_1 + 2X_2 + 3X_3 + \ldots + nX_n}{n(n+1)/2}$. Then T is an unbiased estimator for μ, if $E(T) = \mu$.

Now $E(T) = E\left[\dfrac{X_1 + 2X_2 + 3X_3 + \ldots + nX_n}{n(n+1)/2}\right]$

$$= \frac{1}{n(n+1)/2}[E(X_1) + 2E(X_2) + 3E(X_3) + \ldots + nE(X_n)]$$

$$= \frac{1}{n(n+1)/2}[\mu + 2\mu + 3\mu + \ldots + n\mu]$$

$$= \frac{\mu}{n(n+1)/2}[1 + 2 + 3 + \ldots + n] = \mu$$

Thus T is an unbiased estimator for μ.

Again, T will be a consistent estimator for μ if $\text{Var}(T)$ approaches zero as n tends to ∞. Thus

$$\text{Var}(T) = \text{Var}\left[\frac{X_1 + 2X_2 + 3X_3 + \ldots + nX_n}{n(n+1)/n}\right]$$

$$= \frac{1}{[n(n+1)/2]^2}[\text{Var}(X_1) + 4\text{Var}(X_2) + 9\text{Var}(X_3) + \ldots + n^2\,\text{Var}(X_n)]$$

$$= \frac{1}{n^2(n+1)^2/4}[1^2 + 2^2 + 3^2 + \ldots + n^2]\,\sigma^2$$

$$= \frac{4\sigma^2}{n^2(n+1)^2}\left[\frac{n(n+1)(2n+1)}{6}\right]$$

$$= \frac{2(2n+1)}{3n(n+1)}\sigma^2$$

Clearly, $\text{Var}(T)$ approaches zero as $n \to \infty$. Hence T is a consistent estimator for μ.

Q.15.9. (a) (i) Now $E(\bar{X}_1) = E\left[\frac{1}{3}X_1 + \frac{1}{3}X_2 + \frac{1}{3}X_3\right]$

$$= \frac{1}{3}[E(X_1) + E(X_2) + E(X_3)]$$

$$= \frac{1}{3}[\mu + \mu + \mu] = \mu$$

Thus $\bar{X}_1$ is an unbiased estimator of μ.

$$E(\bar{X}_2) = E\left[\frac{5}{8}X_1 + \frac{1}{8}X_2 + \frac{1}{4}X_3\right]$$

$$= \frac{5}{8}E(X_1) + \frac{1}{8}E(X_2) + \frac{1}{4}E(X_3)$$

$$= \frac{5}{8}\mu + \frac{1}{8}\mu + \frac{1}{4}\mu = \mu$$

So $\bar{X}_2$ is also an unbiased estimator of μ.

Again $E(\bar{X}_3) = E[0.2X_1 + 0.3X_2 + 0.4X_3]$

$$= 0.2\,E(X_1) + 0.3\,E(X_2) + 0.4\,E(X_3)$$

$$= 0.2\mu + 0.3\mu + 0.4\mu = 0.9\mu$$

This shows that $\bar{X}_3$ as defined in the problem is not an unbiased estimator of μ.

(ii) To find the efficiency of the unbiased estimators, we find their variances. Thus

$$\text{Var}(\bar{X}_1) = \text{Var}\left[\frac{1}{3}X_1 + \frac{1}{3}X_2 + \frac{1}{3}X_3\right]$$

$$= \frac{1}{9}[\text{Var}(X_1) + \text{Var}(X_2) + \text{Var}(X_3)]$$

$$= \frac{1}{9}[\sigma^2 + \sigma^2 + \sigma^3] = \frac{\sigma^3}{3}, \text{ and}$$

$$\text{Var}(\bar{X}_2) = \text{Var}\left[\frac{5}{8}X_1 + \frac{1}{8}X_2 + \frac{1}{4}X_3\right]$$

$$= \frac{25}{64}\text{Var}(X_1) + \frac{1}{64}\text{Var}(X_2) + \frac{1}{16}\text{Var}(X_3)]$$

$$= \frac{25}{64}\sigma^2 + \frac{\sigma^2}{64} + \frac{\sigma^2}{16} = \frac{15\sigma^2}{32}$$

Hence, efficiency of $\bar{X}_1 = \dfrac{\text{Var}(\bar{X}_2)}{\text{Var}(\bar{X}_1)} = \dfrac{15\sigma^2}{32} \div \dfrac{\sigma^2}{3} = \dfrac{45}{32} = 140.6\%$

and, efficiency of $\bar{X}_2 = \dfrac{\text{Var}(\bar{X}_1)}{\text{Var}(\bar{X}_2)} = \dfrac{\sigma^2}{3} \div \dfrac{15\sigma^2}{32} = \dfrac{32}{45} = 71.1\%$

$\bar{X}_1$ is a better estimator as $\text{Var}(\bar{X}_1) < \text{Var}(\bar{X}_2)$

(b) An estimator T is unbiased if $E(T) = \mu$.

Now $E(T_1) = E\left[\frac{X_1 + X_2 + X_3}{3}\right] = \frac{1}{3}(3\mu) = \mu;$

$$E(T_2) = E\left[\frac{2X_2 - X_4 + 4X_5 + X_8}{6}\right]$$

$$= \frac{1}{6}[2E(X_2) - E(X_4) + 4E(X_5) + E(X_8)]$$

$$= \frac{1}{2}[2\mu - \mu + 4\mu + \mu] = \frac{1}{6}(6\mu) = \mu;$$

$$E(T_3) = E\left[\frac{X_8 - X_1}{8}\right] = \frac{1}{8}[E(X_8) - E(X_1)]$$

$$= \frac{1}{8}[\mu - \mu]] = 0; \text{ and}$$

$$E(T_4) = E(T_5) = \mu.$$

We see that T_1, T_2 and T_4 are unbiased estimators for μ, but T_3 is not an unbiased estimator as $E(T_3) \neq \mu$.

The most efficient estimator is one that has the minimum variance. Then

$$\text{Var}(T_1) = \text{Var}\left[\frac{X_1+X_2+X_3}{3}\right] = \frac{1}{9}[\text{Var}(X_1)+\text{Var}(X_2)+\text{Var}(X_3)]$$

$$= \frac{1}{9}[\sigma^2 + \sigma^2 + \sigma^2] = \frac{\sigma^2}{3};$$

$$\text{Var}(T_2) = \text{Var}\left[\frac{2X_2 - X_4 + 4X_5 + X_8}{6}\right]$$

$$= \frac{1}{36}[4\text{Var}(X_2) + \text{Var}(X_4) + 16\,\text{Var}(X_5) + \text{Var}(X_8)]$$

$$= \frac{1}{36}[4\sigma^2 + \sigma^2 + 16\sigma^2 + \sigma^2] = \frac{11}{18}\sigma^2; \text{ and}$$

$\text{Var}(T_4) = \text{Var}(X_5) = \sigma^2$.

We see that $\text{Var}(T_1)$ is the smallest, so T_1 is the best estimator.

Q.15.10. (b) Given $f(x; p) = p^x q$ for $x = 0, 1, 2, \ldots$

The joint distribution for the sample from the population is

$$f(x_1, x_2, \ldots, x_n; p) = \prod_{i=1}^{n} p^{xi}\, q = q^n\, p^{\Sigma xi}$$

$$= g\,(\Sigma x_i; p)\, h(x),$$

where $g(\Sigma x_i, p) = (1-p)^n\, p^{\Sigma xi}$ and $h(x) = 1$.

This shows that $f(x; p)$ factors into two functions which satisfy the factorization criterion of sufficiency. Hence ΣX_i is a sufficient statistic.

(c) The normal distribution with mean μ and variance σ^2, is given by

$$f(x; \mu) = \frac{1}{\sqrt{2\pi}\ \sigma} e^{-(X-\mu)^2/2\sigma^2}$$

The joint distribution for the sample from the population is

$$f(x_1, x_2, \ldots, x_n; \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\ \sigma} e^{-(x-\mu)^2/2\sigma^2}$$

$$= \frac{1}{(\sqrt{2\pi})^{n/2} \cdot \sigma^n} e^{-\Sigma(X-\mu)^2/2\sigma^2}$$

The expression $\Sigma(X-\mu)^2$ may be written as

$$\Sigma(X-\mu)^2 = \Sigma(X-\bar{X})^2 + n(\bar{X}-\mu)^2$$

$$\therefore \quad f(x_1, x_2, \ldots, x_n; \mu) = \frac{1}{(\sqrt{2\pi})^{n/2} \cdot \sigma^n} e^{-(\bar{X}-\mu)^2/2\sigma^2} \cdot e^{-\Sigma(X-\bar{X})^2/2\sigma^2}$$

$$= g(\bar{x}, \mu)\, h(x_1, x_2, \ldots, x_n)$$

That is the first factor depends on $\bar{X}$ and μ, and the second part depends on X and $\bar{X}$, but not on μ. Hence $\bar{X}$ is a sufficient estimator for the parameter μ.

Q.15.12. Given $T_1 = \dfrac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1 + n_2}$ and $T_2 = \dfrac{\bar{X}_1 + \bar{X}_2}{2}$

(a) $E(T_1) = \dfrac{1}{n_1+n_2}[n_1 E(\bar{X}_1) + n_2 E(\bar{X}_2)] = \mu$; and

$$\text{Var}(T_1) = \text{Var}\left[\frac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1 + n_2}\right]$$

$$= \frac{1}{(n_1+n_2)^2}[n_1^2 \text{Var}(\bar{X}_1) + n_2^2 \text{Var}(\bar{X}_2)]$$

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

$$= \frac{\sigma^2}{(n_1+n_2)^2}[n_1 + n_2] = \frac{\sigma^2}{n_1+n_2}.$$

$\text{Var}(T_1)$ approaches zero as $n_1, n_2 \to \infty$. Thus T_1 is a consistent pooled estimator for μ.

(b) (i) $E(T_2) = E\left[\dfrac{\bar{X}_1 + \bar{X}_2}{2}\right] = \dfrac{1}{2}E(\bar{X}_1) + \dfrac{1}{2}E(\bar{X}_2) = \dfrac{2\mu}{2} = \mu$

$\therefore$ T_2 is also an unbiased estimator for μ.

(ii) $\text{Var}(T_2) = \text{Var}\left[\dfrac{\bar{X}_1 + \bar{X}_2}{2}\right] = \dfrac{1}{4}[\text{Var}(\bar{X}_1) + \text{Var}(\bar{X}_2)]$

$$= \frac{1}{4}\left[\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}\right] = \frac{\sigma^2}{4}\left[\frac{1}{n_1} + \frac{1}{n_2}\right]$$

$\text{Var}(T_2)$ tends to zero as $n_1, n_2 \to \infty$. This shows that T_2 is also a consistent estimator for μ.

(c) Now $\dfrac{\text{Var}\,(T_1)}{\text{Var}\,(T_2)} = \dfrac{\sigma^2}{n_1+n_2} \div \dfrac{\sigma^2}{4}\left[\dfrac{1}{n_1} + \dfrac{1}{n_2}\right]$

$$= \frac{\sigma^2}{n_1+n_2} \times \frac{4}{\sigma^2}\left(\frac{n_1 n_2}{n_1+n_2}\right) = \frac{4n_1n_2}{(n_1+n_2)^2}$$

In general, $\text{Var}(T_1) < \text{Var}(T_2)$, if

$$(n_1 + n_2)^2 > 4n_1n_2,$$

i.e. $(n_1 + n_2)^2 - 4n_1n_2 > 0,$

i.e. $n_1^2 + n_2^2 + 2n_1 n_2 - 4n_1n_2 > 0$

i.e. $(n_1 - n_2)^2 > 0$, which is true.

Q.15.15 (b) (i) Since a single observation is taken, so the <u>likelihood function</u> is

$$L(p) = f(x; p) = \binom{n}{x} p^x q^{n-x}$$

The natural logarithm (ln) of the likelihood function is

$$ln\, L(p) = ln \binom{n}{x} + x\, ln\, p + (n-x)\, ln\, (1-p)$$

Differenting w.r.t. p, we obtain

$$\frac{\partial}{\partial p}[ln\, L(p)] = \frac{x}{p} - \frac{n-x}{1-p}$$

Equating to zero, we get

$$\frac{x}{p} - \frac{n-x}{1-p} = 0 \text{ or } \frac{x(1-p) - p(n-x)}{p(1-p)} = 0$$

or $x - np = 0$, which gives $\hat{p} = \dfrac{X}{n}$

(ii) The likelihood function for a sample of m observations $X_1, X_2, .., X_m$ from the binomial distribution with parameter n and p is

$$L(p) = \prod_{i=1}^{m} \binom{n}{x_i} p^{x_i} (1-p)^{n-x_i}$$

The natural logarithm (ln) of the likelihood function is

$$ln\, L(p) = \sum_{i=1}^{m} [ln \binom{n}{x_i} + x_i\, ln\, p + (n-x_i)\, ln\, (1-p)]$$

$$= \sum_{i=1}^{m} ln \binom{n}{x_i} + \sum_{i=1}^{m} x_i\, ln\, p + (mn - \sum_{i=1}^{m} x_i)\, ln\, (1-p)]$$

Differentiating w.r.t. p and equating to zero, we get

$$\frac{d}{dp}[ln\ L(p)] = \frac{\sum x_i}{p} - \frac{mn - \sum x_i}{1-p} = 0$$

or $(1-p)\sum x_i - p(mn - \sum x_i) = 0$

or $\sum x_i - p\sum x_i - pmn + p\sum x_i = 0$

This gives $\hat{p} = \frac{\sum x_i}{mn} = \frac{\bar{x}}{n}$ as the M.L.E. of p.

Q.15.16. (b) The maximum likelihood estimates for μ and σ^2 are $\bar{x}$ and S^2, where

$$\bar{X} = \frac{\sum X}{n} \text{ and } S^2 = \frac{1}{n}\sum (X - \bar{X})^2$$

Now $\bar{X} = \frac{\sum X_i}{n} = \frac{501.4}{15} = 33.43$, and

$$S^2 = \frac{1}{n}\sum (X-\bar{X})^2 = \frac{\sum X^2}{n} - \left(\frac{\sum X}{n}\right)^2$$

$$= \frac{16836.6}{15} - \left(\frac{501.4}{15}\right)^2 = 1122.44 - 1117.34 = 5.10$$

Q.15.17. (a) The likelihood function for a sample of n observations from the negative exponential distribution, is

$$L(p) = f(x_1, \lambda)\, f(x_2, \lambda) \ldots f(x_n, \lambda)$$

$$= \frac{1}{\lambda} e^{-x_1/\lambda} \cdot \frac{1}{\lambda} e^{-x_2/\lambda} \ldots \frac{1}{\lambda} e^{-x_n/\lambda}$$

$$= \prod_{i=1}^{n} \frac{1}{\lambda} e^{-x_i/\lambda} = \frac{1}{(\lambda)^n} e^{-n\bar{x}/\lambda}$$

$\therefore$ $ln\ L(p) = -n\ ln\ \lambda - n\bar{x}\ /\ \lambda$

Differentiating w.r.t. λ and equating to zero, we get

$$-\frac{n}{\lambda} + \frac{n\bar{x}}{\lambda^2} = 0 \text{ or } \bar{x} = \lambda$$

Hence the sample mean $\bar{x}$ is the MLE of λ.

(b) The likelihood function for a sample of n observations from the normal distribution with mean zero and variance σ^2, is

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-x_i^2/2\sigma^2}$$

$$= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \cdot e^{-\Sigma x_i^2/2\sigma^2}$$

The natural logarithm of the likelihood function is

$$ln\ L(\theta) = -\frac{n}{2} ln\ (2\pi) - \frac{n}{2} ln\ \sigma^2 - \frac{\sum x_i^2}{2\sigma^2}$$

Differentiating w.r.t. σ^2 and equating to zero, we get

$$\frac{\partial}{\partial\sigma^2}[ln\ L(\theta)] = -\frac{n}{2\sigma^2} + \frac{\sum x_i^2}{2\sigma^4} = 0$$

or $\frac{1}{2} \cdot \frac{1}{\sigma^4}[\sum x_i^2 - n\sigma^2] = 0$

Solving for σ^2, we find $\hat{\sigma}^2 = \frac{\sum x^2}{n}$

Hence the MLE of σ^2 is $\hat{\sigma}^2$.

Q.15.18. (i) We need one equation as there is only one unknown parameter.

The first sample moment about zero is

$$m_1' = \frac{1}{n}\sum x = \bar{x}$$

The corresponding moment of the Poisson distribution is $\mu_1' = \lambda$.

Matching, we get $\hat{\lambda} = \bar{x}$

(ii) Estimation of λ by the method of ML. The *p.f.* of the Poisson distribution is

$$f(x, \lambda) = \frac{e^{-\lambda} \lambda^x}{x!}$$

The *likelihood function* for a sample of size n with values $x_1, x_2, \ldots, x_n$ is

$$L(\theta) = f_1(x_1, \lambda) f_2(x_2, \lambda) \ldots f_n(x_n, \lambda)$$

$$= \frac{e^{-n\lambda} (\lambda)^{\Sigma x}}{x_1! x_2! \ldots x_n!}$$

The natural logarithm (ln) of the likelihood function is

$$ln\, L(\theta) = -n\lambda + \Sigma x\, (ln\, \lambda) - ln\, (x_1!) - ln\, (x_2!) \ldots$$

Differentiating *w.r.t.* λ, we get

$$\frac{\partial [ln\, L(\theta)]}{\partial \lambda} = -n + \frac{\Sigma x}{\lambda}$$

Equating to zero and solving for λ, we obtain

$$\hat{\lambda} = \frac{\Sigma x}{n} = \bar{x}, \text{ the sample mean.}$$

Q.15.20. (b) Here $n = 20$, $\bar{X} = 81.2$ and $\sigma^2 = 80$.

The 95% confidence interval for μ would be

$$\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}.$$

Substituting the values, we get

$$81.2 \pm 1.96 \sqrt{\frac{80}{20}}$$

i.e. $81.2 \pm (1.96)\ (20)$ or 81.2 ± 3.92

Hence the population mean μ lies in the interval 77.28 t 85.12.

Q.15.21. (b) Here n = 25, $\sigma = 10$ and $\bar{x} = 67.53$.

The 95% confidence interval for μ would be

$$\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

Substituting the values, we get

$$67.53 \pm 1.96 \frac{10}{\sqrt{25}} \text{ } i.e. \text{ } 67.53 \pm 3.92$$

Hence the 95% C.I. for μ is (63.61, 71.45).

Q.15.22. Given n = 25, $\bar{x} = 100$ lbs, and $\sigma = 15$ lbs.

The 90% confidence interval for the population mean, μ, would be

$$\bar{X} \pm 1.645 \frac{\sigma}{\sqrt{n}}$$

Substituting the values, we get

$$100 \pm 1.645 \left(\frac{15}{\sqrt{25}}\right) \text{ } i.e. \text{ } 100 \pm 4.94$$

Hence the population mean, μ, lies in the interval 95.06 lbs to 104.94 lbs.

Q.15.23. (b) Here $\bar{x} = \frac{2.3 + (-0.2) + (-0.4) + (-0.9)}{4} = \frac{0.8}{4} = 0.2$

$n = 4$ and $\sigma = 3$.

The 90% confidence interval for μ, would be

$$\bar{x} \pm 1.645 \frac{\sigma}{\sqrt{n}}$$

Substituting the values, we obtain

$$0.2 \pm 1.645 \left(\frac{3}{\sqrt{4}}\right) i.e.\ 0.2 \pm 2.47$$

Hence the confidence interval for μ is -2.27 to 2.67.

Q.15.24(a) The $100(1-\alpha)\%$ confidence limits are given by

$$\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \text{ and } \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

Substituting the given values, we get

$$\bar{X} - z_{\alpha/2}\frac{21}{\sqrt{50}} = 866.11,$$

$$\bar{X} + z_{\alpha/2}\frac{21}{\sqrt{50}} = 875.89.$$

Subtracting, we get

$$2z_{\alpha/2}\frac{21}{\sqrt{50}} = 875.89 - 866.11$$

$$\therefore \quad z_{\alpha/2} = \frac{9.78}{2} \times \frac{\sqrt{50}}{21} = 1.646.$$

From area tables, we find that the area to the right of the value $z_{\alpha/2} = 1.646$ is $0.5 - 0.45 = 0.05$, implying that $\alpha/2 = 0.05$ or $\alpha = 0.10$.

$$\therefore \quad 1 - \alpha = 1 - 0.10 = 0.90$$

Hence a 90% confidence co-efficient was used.

(b) The 95% confidence interval for μ is given by $\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$. Then

$$\bar{x} + 1.96\left(\frac{\sigma}{6}\right) = 1101.7, \text{ and } \bar{x} - 1.96\left(\frac{\sigma}{6}\right) = 1023.3$$

Adding, we get

$$2(\bar{x}) = 2125.0 \text{ or } \bar{x} = 1062.5$$

Subtracting, we obtain

$$2(1.96)\left(\frac{\sigma}{6}\right) = 78.4$$

$$\therefore \quad \sigma = \frac{78.4 \times 6}{(2)\,(1.96)} = 120$$

Now, the 99% confidence interval for the mean length of life of light bulbs, is

$$\bar{X} \pm 2.58\left(\frac{\sigma}{\sqrt{n}}\right).$$

Substituting the values, we get

$$1062.5 \pm 2.58\left(\frac{120}{6}\right)$$

i.e. 1062.5 ± 51.6, *i.e.* (1010.9, 1114.1)

Hence the population mean lies in the interval 1010.9*h* to 1114.1*h*.

Q.15.25. (b) Here n = 50, $\bar{x} = 190$ and $S^2 = 800$.

The 95% C.I. for μ would be

$$\bar{X} \pm 1.96\frac{S}{\sqrt{n}} \qquad (\because \ n \text{ is large})$$

Substituting the values, we get

$$190 \pm 1.96\sqrt{\frac{800}{50}} \text{ } i.e. \text{ } 190 \pm (1.96)\,(4)$$

i.e. 190 ± 7.84.

Hence the desired confidence interval is (182.16, 197.84).

(c) Here n = 100, $\bar{x} = \frac{3978.7}{100} = 39.787$

and $S = \sqrt{15830.983 - (39.787)^2} = 119.36$

The 98% C.I. for μ would be

$$\bar{X} \pm 2.326\frac{S}{\sqrt{n}} \qquad (\because \ n \text{ is large})$$

Substituting the values, we get

$$39.787 \pm (2.326)\left(\frac{119.36}{10}\right) \text{ } i.e. \text{ } 39.787 \pm 27.763.$$

Hence the desired confidence interval is (12.02, 67.55).

Q.15.26. (a) Given n = 36, $\bar{x}$ = Rs. 35.00 and σ = Rs. 13.77.

The 95% confidence interval for the population mean, μ, would be

$$\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}.$$

Substituting the values, we obtain

$$\text{Rs. } 35.00 \pm 1.96 \left(\frac{\text{Rs. } 13.77}{\sqrt{36}}\right), \text{ i.e. Rs. } 35.00 \pm \text{Rs. } 4.50.$$

Hence the required limits are Rs. 30.50 to Rs. 39.50.

(b) As the sample is selected without replacement and its size $n = 100$ is not less than 5% of population size $N = 500$, therefore the 90% C.I. for μ would be

$$\bar{X} \pm 1.645 \frac{S}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

Substituting the values, we get

$$3.5 \pm 1.645 \frac{0.1}{\sqrt{100}} \sqrt{\frac{500-100}{500-1}}$$

i.e. $3.5 \pm (1.645)\ (0.01)\ (0.895)$, *i.e.*, 3.5 ± 0.015

Hence the 90% C.I. for μ is (3.485, 3.515).

Q.15.27. Here n = 64, $\bar{x}$ = 172 and S^2 = 299.

The 99% confidence interval for μ would be

$$\bar{x} \pm 2.58 \frac{S}{\sqrt{n}} \quad (n \text{ is large})$$

Substituting the values, we get

$$172 \pm 2.58 \sqrt{\frac{299}{64}}$$

i.e. $172 \pm (2.58)\ (2.16)$, *i.e.*, 172 ± 5.57

Hence the mean weight of rocks on the lunar surface lies between 166.43 and 177.57 ounces.

Q.15.28. We first compute the sample means and sample variances.

For sample 1: $\bar{x}_1 = \frac{\Sigma X_i}{n_1} = \frac{5452.8}{64} = 85.20,$

$$S_1^2 = \frac{\Sigma (X - \bar{X})^2}{n_1} = \frac{973.44}{64} = 15.21$$

For sample 2: $\bar{x}_2 = \frac{\Sigma fx}{\Sigma f} = \frac{8501}{100} = 85.01,$

$$S_2^2 = \left(\frac{\Sigma fx^2}{\Sigma f}\right) - \left(\frac{\Sigma fx}{\Sigma f}\right)^2 = \frac{722869}{100} - \left(\frac{8501}{100}\right)^2$$

$$= 7228.69 - 7226.7001 = 1.9899$$

Next, we calculate using the information given by the two samples, unbiased estimates of μ and σ^2. Thus for the combined samples,

$$\bar{x}_c = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2} = \frac{5452.8 + 8501}{64 + 100} = \frac{13953.8}{164} = 85.08, \text{ and}$$

$$s_p^2 = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2} = \frac{973.44 + 198.99}{64 + 100 - 2} = \frac{1172.43}{162} = 7.2372.$$

The 97% C.I. for μ based on the combined samples, is

$$\bar{x}_c \pm 2.17 \left(\frac{s_p}{\sqrt{n}}\right)$$

Substituting the values, we get

$$85.08 \pm 2.17 \sqrt{\frac{7.2372}{164}} \text{ i.e. } 85.08 \pm 0.456$$

Hence the desired C.I. for μ is (84.62, 85.54).

Q.15.29. (b) Given $\bar{x} = 4.8$, $\bar{y} = 5.6$, $S_1^2 = 8.64$, $S_2^2 = 7.88$ and $n_1 = n_2 = 100$:

The 95% confidence interval for $(\mu_1 - \mu_2)$ would be

$$(\bar{x}-\bar{y}) \pm 1.96\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$$

Substituting the values, we get

$$(4.8-5.6) \pm 1.96\sqrt{\frac{8.64}{100}+\frac{7.88}{100}}$$

i.e., $(-0.8) \pm (1.96)\ (0.4064)$ *i.e.*, $(-0.8) \pm (0.80)$

Hence the required confidence interval for $(\mu_1-\mu_2)$ would be -1.6 to 0.

Q.15.31. (b) Let μ_1 and μ_2 denote the mean output of the two departments respectively. Then the point estimate for the true difference between the mean outputs of the two departments (*i.e.*, $\mu_1-\mu_2$) is

$$\mu_1 - \mu_2 = \bar{x}_1 - \bar{x}_2 = 100 - 90 = 10.$$

The 95% confidence limits for the true difference, *i.e.*, $\mu_1-\mu_2$ would be

$$(\bar{x}_1-\bar{x}_2) \pm 1.96\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

Substituting the values, we get

$$(100-90) \pm 1.96\sqrt{\frac{256}{64}+\frac{196}{49}}$$

i.e., $10 \pm (1.96)\ (2.828)$ or 10 ± 5.54

Hence the 95% confidence limits for $\mu_1-\mu_2$ are 4.46 and 15.54.

Q.15.32. Given: Males: $n_1 = 1720$, $\bar{x}_1 = 33.93$ yrs, $S_1 = 14.20$ years;

Females: $n_2 = 1230$, $\bar{x}_2 = 27.44$ years, $S_2 = 10.79$ years.

(a) The 95% confidence interval for mean age (μ_1) of all male operatives is

$$\bar{x}_1 \pm 1.96\frac{S_1}{\sqrt{n_1}} = 33.93 \pm (1.96) = \frac{14.20}{\sqrt{1720}}$$

$$= 33.93 \pm (1.96)(0.3424) = 33.93 \pm 0.67$$

Thus the 95% C.I. for μ_1 is (33.26, 34.60).

(b) The 95% confidence interval for mean age of all female operatives (μ_2) is

$$\bar{x}_2 \pm 1.96\frac{S_2}{\sqrt{n_2}} = 27.44 \pm (1.96)\frac{10.79}{\sqrt{1230}}$$

$$= 27.44 \pm (1.96)\ (0.3077) = 27.44 \pm 0.60$$

$\therefore$ The 95% *C.I.* for μ_2 is (26.84, 28.04).

(c) The 95% confidence interval for the difference between their mean ages, *i.e.*, $\mu_1-\mu_2$ is

$$(\bar{x}_1-\bar{x}_2) \pm 1.96\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$$

$$= (33.93-27.44) \pm 1.96\sqrt{\frac{(14.20)^2}{1720}+\frac{(10.79)^2}{1230}}$$

$$= 6.49 \pm (1.96)\ (0.46) = 6.49 \pm 0.90$$

Hence the 95% confidence interval for the difference between mean ages is (5.59, 7.39).

Q.15.33. (a) Let μ_1 and μ_2 denote the mean weekly incomes of the workers in factory *A* and *B* respectively. Then the *ML* estimate of the difference in mean incomes, *i.e.*, $\mu_1-\mu_2$, is

$$\mu_1-\mu_2 = \bar{x}_1 - \bar{x}_2 = 12.80 - 11.25 = 1.55$$

(b) The 90% confidence interval estimate for $\mu_1-\mu_2$ would be

$$(\bar{x}_1-\bar{x}_2) \pm (1.645)\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$$

Substituting the values, we get

$$(12.80 - 11.25) \pm 1.645\sqrt{\frac{64}{160} + \frac{47}{220}}$$

i.e., $1.55 \pm (1.645)\ (0.7833)$

i.e., 1.55 ± 1.29 or $0.26, 2.84$

Hence the required confidence interval for $\mu_1 - \mu_2$ is Rs. 0.26 to Rs. 2.84.

Q.15.34. Given $n_1 = 100$, $\bar{x}_1 = 509$, $S_1^2 = 950$; $n_2 = 100$, $\bar{x}_2 = 447$ **and** $S_2^2 = 876$.

As the sample sizes are large, the 95% confidence interval for $\mu_1 - \mu_2$ would be

$$(\bar{x}_1 - \bar{x}_2) \pm 1.96\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

Substituting the values, we get

$$(509 - 447) \pm 1.96\sqrt{\frac{950}{100} + \frac{876}{100}}$$

i.e., $62 \pm (1.96)\ (4.27)$, *i.e.*, 62 ± 8.37 or $53.63, 70.37$

Hence the required confidence interval is (53.63, 70.37).

Q.15.35. (b) Given $n = 400$, $\hat{p} = \dfrac{32}{400} = 0.08$ **so that** $\hat{q} = 0.92$.

The 95% confidence interval for p, the proportion unemployed in the region would be

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\,\hat{q}}{n}} \qquad (n \text{ is large})$$

Substituting the values, we get

$$0.08 \pm 1.96\sqrt{\frac{(0.08)\ (0.92)}{400}}$$

i.e., $0.08 \pm (1.96)\ (0.0141)$

i.e., 0.08 ± 0.0276 or $0.0524, 0.1076$

(c) The 95% confidence interval for p, the fraction of students who have cars on campus, is

$$\hat{p} \pm z_{0.025}\sqrt{\frac{\hat{p}\,\hat{q}}{n}}, \text{ where } \hat{p} = \frac{16}{75} = 0.2133$$

Substituting the values, we get

$$0.2133 \pm (1.96)\sqrt{\frac{(0.2133)\ (0.7867)}{75}}$$

i.e., $0.2133 \pm (1.96)\ (0.0473)$

i.e., 0.2133 ± 0.0927, *i.e.*, 0.1206 to 0.3060.

Q.15.36. (a) Given X = 60, n = 100. Therefore $\hat{p} = 0.6$

The 95% confidence interval for p (large sample) would be

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}\,\hat{q}}{n}}$$

Substituting the values, we get

$0.6 \pm 1.96\sqrt{(0.6)\ (0.4)\ /\ 100}$ *i.e.* $0.6 \pm (1.96)\ (0.049)$

i.e., 0.6 ± 0.096

Hence the required *C.I.* for p is (0.504, 0.696).

(b) Here $\hat{p} = \dfrac{628}{1000} = 0.628$ **so that** $\hat{q} = 0.372$.

The 98% confidence interval for the fraction of homes that are heated by natural gas is

$$0.628 \pm 2.326\sqrt{\frac{(0.628)\ (0.372)}{100}} \qquad (z_{0.02} = 2.326)$$

i.e., $0.628 \pm (2.326)\ (0.0153)$, *i.e.*, 0.628 ± 0.03559

Hence the 98% *C.I.* for p is (0.5924, 0.6636).

(c) The 90% confidence interval for p, is given by

$$\hat{p} \pm 1.645\sqrt{\frac{\hat{p}\,\hat{q}}{n}}, \text{ where } \hat{p} = 0.76$$

Substituting the values, we get

$$0.76 \pm (1.645)\sqrt{\frac{(0.76)\,(0.24)}{144}}$$

or $0.76 \pm (1.645)\,(0.0356)$

or 0.76 ± 0.059, or 0.701 to 0.819.

The 90% *C.I.* implies that we are 90 per cent confident that the true population proportion p will be in the interval (0.70, 0.82).

Q.15.37. (b) The 95% confidence interval for $p_1 - p_2$ would be

$$(\hat{p}_1 - \hat{p}_2) \pm 1.96\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

Substituting the values, we get

$$(0.54 - 0.49) \pm 1.96\sqrt{\frac{(0.54)\,(0.46)}{100} + \frac{(0.49)\,(0.51)}{100}}$$

i.e., $0.05 \pm (1.96)\sqrt{0.002484 + 0.002499}$

i.e., $0.05 \pm (1.96)\,(0.0706)$

i.e., 0.05 ± 0.138 or $-0.088,\ 0.188$

Hence the required confidence interval is $(-0.088,\ 0.188)$.

Q.15.38. Let p_1 and p_2 denote the true proportions of residents in the city and its suburbs, respectively, favouring the proposal. Then

$$\hat{p}_1 = \frac{2400}{5000} = 0.48,\quad \hat{p}_2 = \frac{1200}{2000} = 0.60.$$

The 90% confidence interval for $p_1 - p_2$ would be

$$(\hat{p}_1 - \hat{p}_2) \pm 1.645\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

Substituting the values, we get

$$(0.48 - 0.60) \pm 1.645\sqrt{\frac{(0.48)\,(0.52)}{5000} + \frac{(0.60)\,(0.40)}{2000}}$$

i.e., $-0.12 \pm (1.645)\sqrt{0.00005 + 0.00012}$

i.e., $-0.12 \pm (1.645)\,(0.013)$

i.e., -0.12 ± 0.0214 or $-0.1414,\ -0.0986$

Hence the required confidence interval is $(-0.1414,\ -0.0986)$, where both end points of the interval are negative. It can be concluded that the proportion of suburban residents favouring the proposal is greater than the proportion of city residents favouring the proposal.

Q.15.39(a) The maximum mean diameter of the population, μ, is the upper limit of the one-sided interval. Thus

$$\mu < \bar{x} + z_\alpha \frac{s}{\sqrt{n}}$$

i.e., $\mu < 0.51 + (2.33)\,\dfrac{0.16}{\sqrt{65}}$ $\quad (\because\ z_{0.01} = 2.33)$

i.e., $\mu < 0.51 + 0.046$ *i.e.*, $\mu < 0.556$.

(b) Here $n = 20$, $\bar{x} = 1014$ and $\sigma = 25$.

The 95% lower interval (in one-sided *C.I.*) is

$$\bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}}.$$

Substituting the values, we get

$$1014 - (1.645)\left(\frac{25}{\sqrt{20}}\right)$$

or $1014 - (1.645)\,(9.2)$ or 1005.

Q.15.40. (a) The normal deviate, Z, corresponding to the given probability is 1.2816, the error, $e = 0.4$, and $\sigma^2 = 10$.

Substituting these values in the formula

$$n = \left(\frac{Z \cdot \sigma}{e}\right)^2,$$

where n stands for sample size, we get

$$n = \frac{10\times(1.2816)^2}{(0.4)^2} = \frac{16.425}{0.16} = 102.66.$$

Hence the required sample size should be 103.

(b) Here $\sigma=100$, the error, $e=20$ and the normal deviate, Z, corresponding to the given probability is 1.96.

Let n denote the sample size. Then

$$n = \left(\frac{Z\,.\,\sigma}{e}\right)^2,$$

$$= \left(\frac{1.96\times 100}{20}\right)^2 = \frac{3.8416\times(100)^2}{400} = 96.04$$

Hence the required sample size should be 97.

(c) Here $e=0.06, p = 0.3$ and $z_{0.025} = 1.96 \quad (\alpha/2=0.025)$

Substituting these values in the formula

$$n = \left(\frac{z_{\alpha/2}}{e}\right)^2 \hat{p}\,\hat{q}, \text{ we get}$$

$$n = \left(\frac{1.96}{0.06}\right)^2 (0.3)\,(0.7) = \frac{0.806736}{0.0036} = 224.09$$

Hence the required sample size is 225.

CHAPTER 16

STATISTICAL INFERENCE: HYPOTHESIS TESTING

Q.16.5(b) The null and alternative hypotheses are given as

$$H_0 : p = 0.6 \text{ and } H_1 : p < 0.6$$

Let X denote the number of families buying milk from company A.

Then the test-statistic is the binomial distribution with $p=0.6$ and $n=10$.

The rejection region, as given, consists of all values from $X=0$ to $X=3$.

Thus the probability of making Type I-Error, *i.e.*, α, consists of $P(X\le 3)$.

Hence $\alpha = P(X \le 3$ when $p=0.6$ and $n=10)$

$$= \sum_{x=0}^{3} b(x;\, 10,\, 0.6)$$

= 0.0548 (From Binomial probability tables)

To compute β, the probability of Type II-Error, we need a specific alternative hypothesis.

We are given $H_0 : p=0.6$ and H_1 : (i) $p=0.3$, (ii) $p=0.4$, (iii) $p=0.5$.

(i) Now a Type II-Error occurs if any value of the distribution under $H_1 : p = 0.3$ falls in the region $X=4$ to $X=10$, the acceptance region of the distribution under $H_0 : p=0.6$.

Hence $\beta = P(4 \le X \le 10$ when $H_1 : p = 0.3)$

$$= \sum_{x=4}^{10} b(x;\, 10,\, 0.3) = 1 - \sum_{x=0}^{3} b(x;\, 10,\, 0.3)$$

$= 1 - 0.6496 = 0.3504$ (From Binomial probability tables)

(ii) When $H_1 : p = 0.4$, we have

$$\beta = P(4 \le X \le 10 \text{ when } p = 0.4)$$

$$= \sum_{x=4}^{10} b(x; 10, 0.4) = 1 - \sum_{x=0}^{3} b(x; 10, 0.4)$$

$= 1 - 0.3823 = 0.6177$ (From Binomial probability tables)

(iii) Similarly, when $H_1 : p = 0.5$, we have

$$\beta = 1 - \sum_{x=0}^{3} b(x; 10, 0.5)$$

$$= 1 - 0.1719 = 0.8281$$

Q.16.6 (a) A type-I error is made by rejecting the null hypothesis, H_0 when it is true. The probability of a type-I error is denoted by the Greek letter α.

A type-II error is made by accepting the null hypothesis, H_0 when it is false and some alternative H_1 is true. The probability of a type-II error is denoted by the Greek letter β.

Given $H_0 : \mu \le 14$, $H_1 : \mu > 14$, $n = 2$, $\sigma^2 = 2.8$ and $\alpha = 0.05, 0.01$

We first find the value of $\bar{x}$, the critical point, which would lead to rejection of H_0 by the test-statistic $Z = \dfrac{\bar{X} - \mu}{\sigma / \sqrt{n}}$.

(i) The critical value of Z for one tailed test at $\alpha = 0.05$ from the normal curve area is 1.645. Thus

$$1.645 = \frac{\bar{X} - 14}{\sqrt{2.8 / 2}} \text{ gives } \bar{x} = 15.95$$

To compute β, the probability of type II error, we use the H_1-distribution $N\left(16.5, \dfrac{2.8}{2}\right)$. Thus

$$\beta = P(\text{Type II error} / \mu = 16.5)$$

$$= P(\bar{x} < 15.95) = P\left(Z < \frac{15.95 - 16.5}{\sqrt{1.4}}\right)$$

$$= P(Z < -0.46) = 0.3228$$

(ii) The critical value of Z for one-tailed test at $\alpha = 0.01$ from the normal curve area is 2.326. Thus

$$2.326 = \frac{\bar{x} - 14}{\sqrt{1.4}} \text{ gives } \bar{x} = 16.75$$

$\therefore$ $\beta = P(\text{type II error} / \mu = 16.5)$

$$= P(\bar{x} < 16.75) = P\left(Z < \frac{16.75 - 16.5}{\sqrt{1.4}}\right)$$

$$= P(Z < 0.21) = 0.5832$$

(b) Given $H_0 : \mu \ge 200$, $H_1 : \mu < 200$, $n = 100$, $\alpha = 0.023$, $\sigma = 25$.

(i) To find the value of $\bar{x}$ (the critical point) that would lead to acceptance of the hypothesis $H_0 : \mu \ge 200$, we use the test-statistic (assuming normal population) given by

$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}.$$

The critical value of Z for one-tailed test at $\alpha = 0.023$ from the table of normal curve is -2, ($\bar{x}$ lies to the left of $\mu = 200$). Thus

$$-2 = \frac{\bar{x} - 200}{25 / \sqrt{100}}$$

Simplifying, we get $\bar{x} = 195$.

(ii) To compute β, the probability of Type II error, we use the distribution under the alternative hypothesis $H_1 : \mu = 191$.

Now at $\bar{x} = 195$, we find $z = \dfrac{195 - 191}{25 / \sqrt{100}} = 1.6$

$\therefore \beta =$ Area between $z=\infty$ and $z=1.6$, *i.e.* area in the acceptance region of the distribution under H_0, when H_1 is $\mu=191$.

$= 0.5 - P(0 < z < 1.6) = 0.5 - 0.4452 = 0.0548.$

(iii) Power $= 1 - \beta = 1 - 0.0548 = 0.9452.$

This is the probability of rejecting the null hypothesis $H_0 : \mu \geq 200$ when it is actually false.

Q.16.8. Given $H_0 : \mu = 67$, $H_1 = \mu > 67$, $\alpha = 0.05$ and $\sigma=3$. Since $H_1 : \mu > 67$, so we find, using one-sided test, the critical value for the decision rule as

$$c = \mu_0 + 1.645 \frac{\sigma}{\sqrt{n}} = 67 + 1.645 \left(\frac{3}{\sqrt{25}}\right)$$

$$= 67 + 0.987 = 67.99$$

Since 4 values in H_1 are specified, so we associate the variables $\bar{X}_1$, $\bar{X}_2$, $\bar{X}_3$ and $\bar{X}_4$ with each of the four H_1-distributions. Using the H_1-distribution with $\mu=68.5$, we calculate the value of β, the probability of Type-II error, (say $\beta_{\mu=68.5}$) as

$$\beta_{\mu=68.5} = P(\text{Type II error} / \mu = 68.5)$$

$$= P(\bar{X}_1 < 67.99) = P\left(Z < \frac{67.99 - 68.5}{3/5}\right)$$

$$= P(Z < -0.85) = 0.1977$$

Again, using the H_1-distribution with $\mu=68.0$, the value of β is

$$\beta_{\mu=68.0} = P(\text{Type II error} / \mu = 68.0)$$

$$= P(\bar{X}_2 < 67.99) = P\left(Z < \frac{67.99 - 68.0}{3/5}\right)$$

$$= P(Z < -0.02) = 0.4920$$

Similarly,

$$\beta_{\mu=67.5} = P(\text{Type II error} / \mu = 67.5)$$

$$= P(\bar{X}_3 < 67.99) = P\left(Z < \frac{67.99 - 67.50}{3/5}\right)$$

$$= P(Z < 0.82) = 0.7939, \text{ and}$$

$$\beta_{\mu=66} = P(\text{Type II error} / \mu = 66)$$

$$= P(\bar{X}_4 < 67.99) = P\left(Z < \frac{67.99 - 66}{3/5}\right)$$

$$= P(Z < 3.32) = 0.9995.$$

The power of the test for $\mu=\mu_1$, *i.e.* $P_w(\mu_1)$ is given by $1-\beta_{\mu=\mu_1}$. Thus

$$P_w(68.5) = 1 - \beta_{\mu=68.5} = 1 - 0.1977 = 0.8023,$$

$$P_w(68.0) = 1 - \beta_{\mu=68.0} = 1 - 0.4920 = 0.5080,$$

$$P_w(67.5) = 1 - \beta_{\mu=67.5} = 1 - 0.7939 = 0.2061,$$

$$P_w(66.0) = 1 - \beta_{\mu=66.0} = 1 - 0.9995 = 0.0005.$$

Q.16.9. The sample size n is given by

$$n = \frac{(z_\alpha + z_\beta)^2 \sigma^2}{(\mu_1 - \mu_0)^2},$$

where z_α, the value of Z under the μ_0-distribution at $\alpha=0.05$ is 1.645, and

z_β, the value of Z under the μ_1-distribution at $\beta=0.01$ is 2.326.

Substituting the values, we get

$$n = \frac{(1.645 + 2.326)^2 (12)^2}{(32 - 28)^2} = 142.$$

Q.16.10. Given: $H_0 : \mu=100$, $\sigma=10$, $\alpha=0.05$ **and** $n=100$. **(a) Let us find the value of** $\bar{x}$ **(the critical point) which would lead to rejection of the hypothesis** H_0 **by the test statistic**

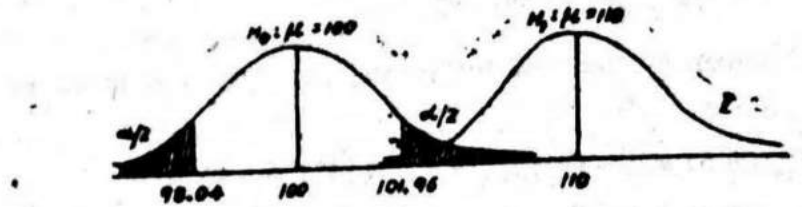
$$Z = \frac{\bar{x} - \mu}{10 / \sqrt{n}}.$$

The critical values of Z for two tailed test at $\alpha = 0.05$ from the normal curve areas, are -1.96 and $+1.96$. Thus

$$\pm 1.96 = \frac{\bar{x} - 100}{10 / \sqrt{100}}$$

Simplifying, we get $\bar{x} = 98.04$ and 101.96.

That is, the hypothesis $H_0 : \mu = 100$ would be rejected if $\bar{x} < 98.04$ or $\bar{x} > 101.96$. (See figure)



To compute β, the probability of type II error, we use the distribution under the alternative hypothesis $H_1 : \mu = 110$.

$\therefore$ 101.96 in standard units $= \dfrac{101.96 - 110}{10 / \sqrt{100}} = -8.04$, and

98.04 in standard units $= \dfrac{98.04 - 110}{10 / \sqrt{100}} = -11.96$.

Thus β = Area under normal curve between $Z = -8.04$ and $Z = -11.96$

< 0.2

Hence β would result in less than 0.2 when in fact $H_1 : \mu = 110$.

(b) In this case, the sample size, n, is given by the relation

$$n = \frac{(Z_0 + Z_1)^2 \cdot \sigma^2}{(\mu_1 - \mu_0)^2},$$

where $Z_0 = 1.96$, the value of Z under the μ_0-distribution at $\alpha = .05$, and

$Z_1 = 2.33$, the value of Z under the μ_1-distribution at $\alpha = .01$.

Thus $n = \dfrac{(1.96 + 2.33)^2 \cdot (10)^2}{(110 - 100)^2} = 18.52$

Hence the sample size must be at least 19.

Q.16.11. Let μ_0 and μ_1 be the mean stated in the null hypothesis and the alternative hypothesis respectively, and the values of the normal deviate Z be Z_0 and Z_1.

Then the sample size, n, is given by the relation

$$n = \frac{\sigma^2 (Z_0 + Z_1)^2}{(\mu_1 - \mu_0)^2}$$

The value of Z under the μ_0-distribution at $\alpha = 0.05$ is $Z_0 = 1.645$, and the value of Z under the μ_1-distribution at $\alpha = 0.10$ is $Z_1 = 1.28$.

Substituting the values, we get

$$n = \frac{(15)^2 (1.28 + 1.645)^2}{(205 - 200)^2} = \frac{225 \times 8.56}{25} = 77.04$$

Thus the appropriate sample size is 78.

Q.16.12. (b) (i) We state the hypotheses as

$$H_0 : \mu = 75 \text{ and } H_1 : \mu \neq 75.$$

(ii) Let us choose the level of significance at $\alpha = 0.10$.

(iii) The test-statistic is

$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

which under H_0 is a standard normal variable.

(iv) Computations: Here $n = 16$, $\bar{x} = 82$ and $\sigma = \sqrt{36} = 6$.

$$\therefore \quad z = \frac{82 - 75}{6 / \sqrt{16}} = 4.67.$$

(v) The critical region is $|Z| \geq 1.645$.

(vi) **Conclusion.** The computed value of $z = 4.67$ falls in the critical region, we therefore reject H_0 and conclude that this group of 16 students is not typical. (In other words, it is superior).

Q.16.13. (b) (i) We state the hypotheses as

$$H_0 : \mu = 45 \text{ and } H_1 : \mu < 45.$$

(ii) The level of significance is set at $\alpha = 0.05$ and one-sided test is used.

(iii) The test-statistic is

$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

which under H_0, is a standardized normal variable.

(iv) Computations.

Here $n = 36$, $\sigma^2 = 25$ and $\bar{x} = 42.6$.

Thus $$z = \frac{42.6 - 45}{5 / \sqrt{36}} = \frac{(-2.4)\,6}{5} = -2.88.$$

(v) The critical region is $Z < -1.645$.

(vi) The computed value of Z falls in the critical region, we therefore reject the hypothesis $H_0 : \mu = 45$ and accept the alternative hypothesis $H_1 : \mu < 45$.

Q.16.14. (a) (i) We state our hypotheses as

$H_0 : \mu \leq 67.39$ inches and $H_1 : \mu > 67.39$ inches.

(ii) We use a significance level of 0.05 and a one sided test.

(iii) The test statistic is

$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

which under H_0, has a normal distribution with zero mean and unit variance.

(iv) Computation:

Here $n = 400$, $\bar{x} = 67.47$ inches and $\sigma = 1.3$ inches

$$\therefore \quad z = \frac{67.47 - 67.39}{1.30 / \sqrt{400}} = \frac{(0.08)\,20}{1.30} = 1.23$$

(v) The critical region is $Z \geq 1.645$.

(vi) The calculated value of $Z = 1.23$ is less than 1.645, so we accept H_0 and may conclude that the sample might be regarded drawn from a population with $\mu = 67.39$ inches.

(b) (i) We state our hypotheses as

$H_0 : \mu \geq 123$ and $H_1 : \mu < 123$. (One-tailed test)

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic is

$$Z = \frac{\bar{X} - \mu_0}{S / \sqrt{n}}$$

which under H_0 is approximately standard normal.

(iv) The critical region is $Z < -z_{0.05} = -1.645$

(v) Computations: Here $\mu_0 = 123$, $n = 49$, $\bar{x} = 120.67$ and $S = 8.44$.

$$\therefore \quad z = \frac{120.67 - 123}{8.44 / \sqrt{49}} = \frac{(-2.33)\,(7)}{8.44} = -1.93.$$

(vi) **Conclusion.** Since the calculated value of $z = -1.93$ falls in the critical region, so we reject H_0.

Q.16.15. (a) (i) We state our hypotheses as

$$H_0 : \mu = 72 \text{ and } H_1 : \mu \neq 72 \quad \text{(Two-tailed)}$$

(ii) The significance level is set at $\alpha = 0.01$

(iii) The test-statistic to use is

$$Z = \frac{\bar{X} - \mu_0}{S / \sqrt{n}}$$

which, if H_0 is true, is approximately standard normal as the sample size is large enough.

(iv) The critical region is $|Z| \geq z_{0.05} = 2.58$

(v) Computations: Here $\mu_0 = 72$, $\bar{x} = 71$, $n = 40$ and $S^2 = 200$

$$\therefore \quad z = \frac{71 - 72}{\sqrt{200 / 40}} = \frac{(-1)}{2.236} = -0.45$$

(vi) **Conclusion.** As the calculated value of $z = -0.45$ does no fall in the critical region, we therefore accept H_0.

(b) (i) We state our hypotheses are

$$H_0 : \mu = 15 \text{ and } H_1 : \mu \neq 15$$

(ii) The significance level is set at $\alpha = 0.05$

(iii) The test-statistic to use is

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

which, if H_0 is true, is a standard normal variable.

(iv) The critical region is $|Z| \geq z_{\alpha/2} = 1.96$.

(v) Computations. Substituting the values, we get

$$z = \frac{12 - 15}{\sqrt{144 / 64}} = \frac{(-3)\ (8)}{12} = -2$$

(vi) **Conclusion.** Since the calculated value of $z = -2$ falls in the critical region, we therefore reject our null hypothesis $H_0 : \mu = 15$ in favour of $H_1 : \mu \neq 15$.

Q.16.16. We make H_1, what is claimed, and the negation of the claim as H_0. Thus our hypotheses are stated as

(i) $H_0 : \mu \leq 20{,}000$ and $H_1 : \mu > 20{,}000$. (one-tailed)

(ii) The significance level is set at $\alpha = 0.01$

(iii) The test-statistic to use is

$$Z = \frac{\bar{X} - \mu_0}{S / \sqrt{n}}$$

which, if H_0 is true, has an approximate standard normal distribution.

(iv) The critical region is $Z \geq z_{0.01} = 2.33$

(v) Computations. Here $\mu_0 = 20{,}000$, $\bar{x} = 23{,}500$, $S = 3900$ and $n = 100$.

$$\therefore \quad z = \frac{23{,}500 - 20{,}000}{3{,}900 / \sqrt{100}} = 8.97$$

(vi) **Conclusion.** Since our calculated value of $z = 8.97$ falls in the critical region, so we reject H_0. We may conclude that $\mu > 20{,}000$ kilometers, i.e. the claim is justified.

Q.16.17. (a) (i) We state our hypotheses as

$$H_0 : \mu = 2.9 \text{ inches and } H_1 : \mu \neq 2.9 \text{ inches.}$$

(ii) We use a level of significance of 0.05, and a two sided test.

(iii) The test-statistic is

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}},$$

which under H_0, is a standardized normal variable.

(iv) Computations.

Here $n = 900, \bar{x} = 2.4$ inches and $\sigma = 3.2$ inches.

$$\therefore \quad z = \frac{2.4 - 2.9}{3.2 / \sqrt{900}} = \frac{(-0.5) \times 30}{3.2} = -4.69$$

(v) The critical region is $Z < -1.96$ and $Z > +1.96$.

(vi) The computed value of Z from sample data falls in the critical region, so the null hypothesis $H_0 : \mu = 2.9$ inches can be rejected. The sample cannot be regarded as being a simple sample from a large population with $\mu = 2.9$ at the 0.05 level of significance.

(b) (i) Our hypotheses are stated as

$$H_0 : \mu = 6.2'' \text{ and } H_1 : \mu \neq 6.2''$$

(ii) We use a level of significance, $\alpha = 0.05$, and a two-sided test.

(iii) The test statistic is

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}},$$

which under H_0, is a standardized normal variable.

(iv) Computations.

Here $n = 400, \bar{x} = 6.0''$ and $\sigma = 2.25''$

$$\therefore \quad z = \frac{6.0 - 6.2}{2.25 / \sqrt{400}} = \frac{(-0.2) \times 20}{2.25} = -1.78$$

(v) The critical region is $Z < -1.96$ and $Z > 1.96$.

(vi) The calculated value of Z from sample data falls in the acceptance region, so we accept H_0 and conclude that the sample can be regarded as being drawn from a large population with $\mu = 6.2''$.

Q.16.18. (a) (i) We state the hypotheses as

$$H_0 : \mu = 6 \text{ oz and } H_1 : \mu \neq 6 \text{ oz.}$$

(ii) We use a level of significance of $\alpha = 0.05$ and use a two-sided test.

(iii) The test-statistic is

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}},$$

which under H_0, is a standardized normal variable.

(iv) Computations.

Here $n = 100, \bar{x} = 6.1$ oz and $\sigma = 0.2$ oz

$$\therefore \quad z = \frac{6.1 - 6.0}{0.2 / \sqrt{100}} = \frac{(0.1)\ (10)}{0.2} = 5.$$

(v) The critical region is $Z < -1.96$ and $Z > 1.96$.

(vi) The computed value of Z from sample data is greater than 1.96, so we reject H_0. There is sufficient evidence to conclude that the process is out of control.

(b) (i) The hypotheses are

$$H_0 : \mu \geq 50 \text{ and } H_1 : \mu < 50 \quad \text{(one-tailed)}$$

(ii) The significance level is set at $\alpha = 0.01$.

(iii) The test-statistic to use is

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}},$$

which under H_0 is standard normal.

(iv) Computations.

Here $n = 36, \bar{x} = 48.7$ years and $\sigma = 3.1$ years

Substituting these values, we get

$$z = \frac{48.7 - 50}{3 / \sqrt{36}} = \frac{-7.8}{3.1} = -2.52$$

(v) The critical region is $Z < -2.326$.

(vi) **Conclusion.** The computed value of Z falls in the critical region. Reject H_0 (i.e. the claim).

Q.16.19. (a) (i) Our hypotheses are

$H_0 : \mu_1 - \mu_2 = 0$ and $H_1 : \mu_1 - \mu_2 \neq 0$.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic to use is

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}},$$

which, if H_0 is true, has a standard normal distribution.

(iv) The critical region is $|Z| \geq 1.96$.

(v) Computations. Here $\bar{x}_1 = 15$, $\sigma_1^2 = 24$, $n_1 = 6$;

$\bar{x}_2 = 13$, $\sigma_2^2 = 80$, $n_2 = 8$.

Substituting the values, we get

$$z = \frac{15 - 13}{\sqrt{\frac{24}{6} + \frac{80}{8}}} = \frac{2}{3.74} = 0.53$$

(vi) **Conclusion.** Since the calculated value of $z = 0.53$ does not fall in the critical region, so we accept H_0.

(b) (i) We state our hypotheses as

$H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 \neq \mu_2$

(ii) The significance level is set at $\alpha = 0.06$.

(iii) The test-statistic under H_0 is

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}},$$

which has a standard normal distribution.

(iv) The critical region is $|Z| \geq z_{0.03} = 1.88$.

(v) Computations. Here $n_1 = 25$, $\sigma_1 = 5.2$, $\bar{x}_1 = 81$,

$n_2 = 36$, $\sigma_2 = 3.4$, $\bar{x}_2 = 76$.

Substituting these values, we get

$$z = \frac{81 - 76}{\sqrt{\frac{(5.2)^2}{25} + \frac{(3.4)^2}{36}}}$$

$$= \frac{5}{\sqrt{1.0816 + 0.3211}} = \frac{5}{1.184} = 4.22$$

(vi) **Conclusion.** Since the calculated value of $z = 4.22$ falls in the critical region, we therefore reject H_0.

Q.16.20. (i) We state our hypotheses as

$H_0 : \mu_A = \mu_B$ and $H_1 : \mu_A \neq \mu_B$.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic under H_0 is

$$Z = \frac{\bar{X}_A - \bar{X}_B}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

which has a standard normal distribution.

(iv) The critical region is $|Z| \geq 1.96$.

(v) Computations. $\bar{X}_A = 21.7$, $\bar{X}_B = 24.3$ and $\sigma = 2.5$

$$\therefore \quad z = \frac{21.7 - 24.3}{2.5\sqrt{\frac{1}{10} + \frac{1}{10}}} = \frac{-2.6}{1.118} = -2.33.$$

(vi) **Conclusion.** Since the computed value of $z = -2.33$ falls in the critical region, so we reject H_0.

Q.16.21. Let μ_1 and μ_2 denote the mean grades of the first class and the second class students respectively. Then the hypotheses are

(i) $H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 \neq \mu_2$

(ii) The levels of significance are $\alpha = 0.05$ and $\alpha = 0.01$.

(iii) The test-statistic under H_0 is

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}},$$

which is approximately normal.

(iv) Computations. Here $n_1 = 40,\ \bar{X}_1 = 74,\ S_1 = 8,$

$n_2 = 50,\ \bar{X}_2 = 78,\ S_2 = 7.$

$$\therefore \quad z = \frac{74 - 78}{\sqrt{\frac{64}{40} + \frac{49}{50}}} = \frac{-4}{\sqrt{1.6 + 0.98}} = \frac{-4}{1.61} = -2.48$$

(v) The critical regions are $|Z| \geq z_{0.05} = 1.96$ and $|Z| \geq z_{0.01} = 2.58.$

(vi) **Conclusion.** Reject H_0 at 5% but accept at 1% level of significance. This means that the difference is significant at 5% but is insignificant at 1% level.

Q.16.22. (i) We state our hypotheses are

$$H_0 : \mu_1 = \mu_2 \text{ and } H_1 : \mu_1 \neq \mu_2.$$

(ii) The significance level is set at $\alpha = 0.01.$

(iii) Since the sample sizes are large, we therefore substitute the sample standard deviations for the population standard deviations. Thus the test-statistic to be used under H_0 is

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}},$$

which is approximately standard normal.

(iv) The critical region is $|Z| \geq 2.58.$

(v) Computations. Substituting the values, we get

$$z = \frac{150 - 153}{\sqrt{\frac{(10)^2}{50} + \frac{(5)^2}{100}}} = \frac{-3}{1.5} = -2.00$$

(vi) **Conclusion.** Since the computed value of $z = -2.00$ does not fall in the critical region, so we accepts H_0.

Q.16.23. (a) (i) The hypotheses are stated as

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

(ii) We use a level of significance of $\alpha = 0.05$ and a one-sided test.

(iii) Since the sample sizes are sufficiently large and the population variances are not known, we take the given standard deviations as estimate of σ_1 and σ_2. Then the test statistic is

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}},$$

which under H_0 is a S.N.V.

(iv) Computations.

Here $n_1 = 1{,}600 \qquad n_2 = 6{,}400$

$\bar{x}_1 = 68.55$ inches $\qquad \bar{x}_2 = 67.85$ inches

$S_1 = 2.52$ inches and $S_2 = 2.56$ inches.

Thus $$z = \frac{68.55 - 67.85}{\sqrt{\frac{(2.56)^2}{6400} + \frac{(2.52)^2}{1600}}}$$

$$= \frac{+0.70}{\sqrt{0.001024 + 0.003969}} = \frac{+0.70}{0.0706} = +9.92$$

(v) The critical region is $Z > 1.645.$

(vi) The computed value of Z lies in the critical region. We reject H_0 in favour of H_1, concluding that Australians are on the average taller than Englishmen.

(b) (i) We state the hypotheses as

$$H_0 : \mu_A = \mu_B \text{ and } H_1 : \mu_A \neq \mu_B$$

(ii) We use a significance level of $\alpha = 0.05$ & a two-tailed test.

(iii) The test-statistic under H_0 is

$$Z = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{\frac{S_A^2}{n_1} + \frac{S_B^2}{n_2}}},$$

which has a standard normal distribution.

(iv) The critical region is $|Z| \geq 1.96$.

(v) Computations. Substituting the values, we get

$$z = \frac{1282 - 1208}{\sqrt{\frac{(80)^2}{50} + \frac{(94)^2}{50}}}$$

$$= \frac{74}{\sqrt{128 + 176.72}} = \frac{74}{17.46} = 4.24.$$

(vi) **Conclusion.** Since the computed value of $z = 4.24$ falls in the critical region, we therefore reject H_0 in favour of H_1. Thus it seems quite certain that the two brands differ in quality as far as mean burning time is concerned and that brand A is to be preferred.

Q.16.24. Let μ_A and μ_B denote the average life time of bulbs manufactured by company A and by company B respectively. Then our hypotheses are

(i) $H_0 : \mu_A - \mu_B \leq 200$, and $H_1 : \mu_A - \mu_B > 200$

(ii) The significance level is set at $\alpha = 0.01$.

(iii) The test-statistic under H_0 is

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - 200}{\sqrt{\frac{S_A^2}{n_1} + \frac{S_B^2}{n_2}}},$$

which is approximately normal.

(iv) Computations. Here $n_1 = 80$, $\bar{X}_1 = 1258$, $S_A = 94$

$n_2 = 60$, $\bar{X}_2 = 1029$, $S_B = 68$

$$\therefore \quad z = \frac{(1258 - 1029) - 200}{\sqrt{\frac{(94)^2}{80} + \frac{(68)^2}{60}}} = \frac{229 - 200}{\sqrt{110.45 + 70.0667}}$$

$$= \frac{29}{13.44} = 2.16$$

(v) The critical region is $Z \geq z_{0.01} = 2.58$.

(vi) **Conclusion.** Since the calculated value of Z does not fall in the critical region, we therefore accept H_0. This means that the difference is insignificant.

Q.16.25. (b) (i) We state our hypotheses as

$H_0 : p = 0.60$ and $H_1 : p > 0.60$

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic to use is

$$Z = \frac{(X \pm \frac{1}{2}) - np_0}{\sqrt{np_0q_0}}, \quad \text{(with continuity correction)}$$

which, if H_0 is true, is approximately a standard normal.

(iv) The critical region is $Z > 1.645$.

(v) Computations. Substituting the values, we get

$$z = \frac{(70 - \frac{1}{2}) - 60}{\sqrt{100 \times 0.6 \times 0.4}} = \frac{9.5}{4.9} = 1.94$$

(vi) **Conclusion.** Since the computed value of $z = 1.94$ falls in the critical region, we therefore reject H_0. It could be concluded that his shooting has improved.

Q.16.26. (a) (i) The hypotheses would be stated as

$H_0 : p = \frac{1}{2}$, *i.e.*, the coin is unbiased, and

$H_1 : p \neq \frac{1}{2}$, *i.e.*, the coin is biased.

(ii) We use a level of significance of $\alpha = 0.05$, and a two-sided test.

(iii) The test-statistic is

$$Z = \frac{(X \pm \frac{1}{2}) - np_0}{\sqrt{np_0q_0}}, \quad \text{(with continuity correction)}$$

which under H_0, is approximately a S.N.V.

(iv) Computation.

$$z = \frac{(490 - \frac{1}{2}) - 900 \times \frac{1}{2}}{\sqrt{900 \times \frac{1}{2} \times \frac{1}{2}}}, \quad (\text{using } x - \tfrac{1}{2} \text{ as } x > np_0)$$

$$= \frac{39.5}{15} = 2.63$$

(v) The rejection region is $Z < -1.96$ and $Z > 1.96$.

(vi) The computed value of Z falls in the rejection region. We reject H_0 in favour of H_1. We may conclude that this result does not support the hypothesis that the coin is unbiased.

(b) (i) The hypothesis would be stated as

$H_0 : p = \frac{1}{2}$, *i.e.*, there is an equal sex division in the population, and

$H_1 : p \neq \frac{1}{2}$, *i.e.*, there is not an equal sex division in the population.

(ii) We use a level of significance of $\alpha = 0.05$, and a two sided test.

(iii) The test-statistic would be

$$Z = \frac{(X \pm \frac{1}{2}) - np_0}{\sqrt{np_0q_0}}, \quad \text{(with continuity correction)}$$

which under H_0, is approximately a S.N.V.

(iv) Computation.

$$z = \frac{(52 - \frac{1}{2}) - 98 \times \frac{1}{2}}{\sqrt{98 \times \frac{1}{2} \times \frac{1}{2}}}; \quad (\text{using } x - \tfrac{1}{2} \text{ as } x > np_0)$$

$$= \frac{2.5}{4.95} = 0.51$$

(v) The critical region is $Z < -1.96$ and $Z > 1.96$.

(vi) The computed value of Z does not fall in the critical region. We accept H_0 and may conclude that the data are consistent with an equal sex division in the population.

Q.16.27. (i) We state our hypotheses as

$$H_0 : p \leq 0.50 \text{ and } H_1 : p > 0.50 \quad \text{(one-tailed)}$$

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic under H_0 is

$$Z = \frac{(X \pm \frac{1}{2}) - np_0}{\sqrt{np_0q_0}}, \quad \text{(with continuity correction)}$$

which is approximately a standard normal.

(iv) The critical region is $Z \geq 1.645$.

(v) Computations. Substituting the values, we get

$$z = \frac{(5180 - \frac{1}{2}) - 10{,}000\ (0.50)}{\sqrt{10{,}000\ (0.50)\ (0.50)}} = \frac{179.5}{50} = 3.59$$

(vi) **Conclusion.** Since the computed value of $z = 3.59$ falls in the critical region, so we reject H_0 and conclude that the proportion of the voters who favour the candidate is more than 50%.

Q.16.28. (a) (i) We state the hypotheses as

$$H_0 : p = 0.05 \text{ and } H_1 : p \neq 0.05$$

(ii) We use a level of significance of $\alpha = 0.05$ and a two sided test.

(iii) The test-statistic would be

$$Z = \frac{(X \pm \frac{1}{2}) - np_0}{\sqrt{np_0q_0}}, \quad \text{(with continuity correction)}$$

which under H_0, is approximately a standardized normal variable.

(iv) Computations.

$$z = \frac{(12 - \frac{1}{2}) - (200)\,(0.05)}{\sqrt{200\,(0.05)\,(0.95)}}, \quad (\text{using } x - \tfrac{1}{2} \text{ a } x > np_0)$$

$$= \frac{1.5}{3.08} = 0.487.$$

(v) The rejection region is $Z < -1.96$ and $Z > 1.96$.

(vi) The computed value does not fall in the rejection region. We therefore accept H_0 and may conclude that the given information is consistent with an average of 5 percent set as a standard.

(b) (i) We state our hypotheses as

$$H_0 : p = 0.55 \text{ and } H_1 : p < 0.55. \quad \text{(one-tailed)}$$

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic under H_0 is

$$Z = \frac{X - np_0}{\sqrt{np_0q_0}}$$

which is approximately a standard normal.

(iv) The critical region is $Z \leq -1.645$

(v) Computations. Substituting the values, we get

$$z = \frac{35 - 78\,(0.55)}{\sqrt{78\,(0.55)\,(0.45)}} = \frac{-7.9}{4.39} = -1.80$$

(vi) **Conclusion.** Since the computed value of $z = -1.80$ falls in the critical region, we therefore reject H_0.

Q.16.29. (a) (i) The hypotheses would be stated as

$H_0 : p = 0.9$, and the claim is legitimate, and

$H_1 : p < 0.9$, and the claim is false.

(ii) We use a level of significance of $\alpha = 0.01$, and a one-sided test.

(iii) The test-statistic would be

$$Z = \frac{X - np_0}{\sqrt{np_0q_0}}, \quad \text{(without continuity correction)}$$

which is assumed to be a S.N.V.

(iv) Computations.

$$Z = \frac{160 - (200)\,(0.9)}{\sqrt{(200)\,(0.9)\,(0.1)}} \quad \text{(without continuity correction)}$$

$$= \frac{160 - 180}{\sqrt{18}} = \frac{-20}{4.23} = -4.73$$

(v) The critical region is $Z < -2.33$.

(vi) The computed value of Z falls in the critical region. We reject H_0. We may conclude that the claim is not legitimate and that the sample results are highly significant.

(b) Let p denote the proportion of men *who can tell* the difference beteen the two brands of cheese and let X be the number of men in the sample who can tell.

Then we need to test the hypothesis

(i) $H_0 : p \leq 0.1$ against $H_1 : p > 0.1$

(ii) We use a level of significance of $\alpha = 0.05$ and $\alpha = 0.01$ and one tailed test.

(iii) The test-statistic would be

$$Z = \frac{(X \pm \frac{1}{2}) - np_0}{\sqrt{np_0q_0}}, \quad \text{(with continuity correction)}$$

which is approximately standard normal.

(iv) Computations. Here $n = 500$ and $X = 72$

$$z = \frac{(72 - \frac{1}{2}) - (500)\,(0.1)}{\sqrt{500\,(0.1)\,(0.9)}} = \frac{71.5 - 50}{\sqrt{45}} = 3.21$$

(v) The critical region is $Z > z_{0.05} = 1.645$, and

$$z > z_{0.01} = 2.326$$

(vi) **Conclsuion.** The computed value of z exceeds the table value both at 5% and at 1% level of significance, we therefore reject the claim.

Q.16.30. (i) We state our hypotheses as

$$H_0 : p = 0.95 \text{ and } H_1 : p < 0.95$$

(ii) The significance level is set at $\alpha = 0.05$

(iii) The test-statistic under H_0 is

$$Z = \frac{X - np_0}{\sqrt{np_0q_0}}$$

which is approximately a standard normal.

(iv) The critical region is $Z \leq -1.645$

(v) Computations: Substituting the values, we get

$$z = \frac{355 - (400)\,(0.95)}{\sqrt{(400)(0.95)(0.05)}}, \text{ (} \because \text{ 355 conform to specification)}$$

$$= \frac{-25}{4.36} = -5.73$$

(vi) **Conclusion.** Since the computed value of $z = -5.73$ falls in the critical region, so we reject H_0. The company's claim that at least 95% of the parts conformed to specifications is not justified.

Q.16.31. (a) (i) We state our hypotheses as

$$H_0 : p_1 = p_2 \text{ and } H_1 : p_1 \neq p_2.$$

(ii) We choose a significance level of $\alpha = 0.05$.

(iii) The test-statistic under H_0 is

$$Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{p}_c\,\hat{q}_c\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where $\hat{p}_c = \dfrac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$ and Z for large sample sizes is approximately standard normal.

(iv) The critical region is $|Z| \geq 1.96$.

(v) Computations. Here $\hat{p}_1 = \frac{12}{150} = 0.08$, $\hat{p}_2 = \frac{4}{100} = 0.04$, and

$$\hat{p}_c = \frac{12 + 4}{150 + 100} = \frac{16}{250} = 0.064, \text{ so that } \hat{q}_c = 0.936.$$

Substituting these values, we get

$$z = \frac{0.08 - 0.04}{\sqrt{(0.064)\,(0.936)\left(\frac{1}{150} + \frac{1}{100}\right)}}$$

$$= \frac{0.04}{\sqrt{(0.0599)\,(0.0167)}} = \frac{0.04}{0.0316} = 1.27.$$

(vi) **Conclusion.** Since the computed value of $z = 1.27$ does not fall in the critical region, we therefore accept H_0 and conclude that the difference between the proportions of the two firms is not significant.

(b) (i) We state our hypotheses as

$$H_0 : p_1 = p_2 \text{ and } H_1 : p_1 \neq p_2.$$

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic under H_0 is

$$Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{p}_c\,\hat{q}_c\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where $\hat{p}_c = \dfrac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$ and which has approximately a standard normal.

(iv) The critical region is $|Z| \geq 1.96$.

(v) Computations. Here $\hat{p}_1 = \frac{225}{500} = 0.45, \hat{p}_2 = \frac{275}{500} = 0.55$

and $\hat{p}_c = \frac{225 + 275}{500 + 500} = \frac{500}{1000} = 0.5.$

$$\therefore \quad z = \frac{0.45 - 0.55}{\sqrt{(0.5)\,(0.5)\left(\frac{1}{500} + \frac{1}{500}\right)}}$$

$$= \frac{-0.10}{\sqrt{(0.25)\,(0.004)}} = \frac{-0.10}{0.032} = -3.12$$

(vi) **Conclusion.** Since the computed value of $z = -3.12$ falls in the critical region, so we reject H_0.

Q.16.32. (a) (i) We state the hypotheses as

H_0 : The machine has not been improved, *i.e.* $p_1 = p_2$

H_1 : The machine has been improved, *i.e.* $p_1 > p_2$.

(ii) We use a level of significance of $\alpha = 0.05$, and a one sided test.

(iii) The test-statistic would be

$$Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{p}_c\,\hat{q}_c\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where $\hat{p}_c$ is estimated from sample proportions $\hat{p}_1$ and $\hat{p}_2$ by the relation

$$\hat{p}_c = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

Under H_0 the Z-statistic is approximately a S.N.V.

(iv) Computations. Here

$\hat{p}_1$ = proportion of imperfect articles in the first sample

$$= \frac{16}{500} = 0.032$$

$\hat{p}_2$ = proportion of imperfect articles in the second sample

$$= \frac{3}{100} = 0.03$$

Under $H_0 : \hat{p}_c = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$

$$= \frac{(500)\,(0.032) + 100\,(0.03)}{500 + 100} = \frac{19}{600}$$

and $\hat{q}_c = 1 - \hat{p}_c = 1 - \frac{19}{600} = \frac{581}{600}$

$$\therefore \quad z = \frac{0.032 - 0.03}{\sqrt{\frac{19}{600} \times \frac{581}{600}\left(\frac{1}{500} + \frac{1}{100}\right)}}$$

$$= \frac{0.002}{\sqrt{(0.030664)\,(0.012)}} = \frac{0.002}{0.019} = 0.105$$

(v) The critical region is $Z > 1.645$.

(vi) The calculated of Z does not fall in the critical region. We therefore cannot reject H_0. We may conclude that it is likely that the machine has not been improved.

(b) Let $\hat{p}_1$ denote the proportion of orders received from white return envelope and $\hat{p}_2$ denote the proportion of orders received from blue envelope. Then

$$\hat{p}_1 = 0.10 \text{ and } \hat{p}_2 = 0.13$$

(i) We wish to test the hypotheses

$H_0 : p_2 - p_1 = 0$ against $H_1 : p_2 - p_1 > 0$

(ii) We use a level of significance of $\alpha = 0.05$ and a one-sided test.

(iii) The test-statistic is

$$Z = \frac{\hat{P}_2 - \hat{P}_1}{\sqrt{\hat{p}_c\,\hat{q}_c\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where $\hat{p}$ = an estimte of the common population proportion on the assumption that the two types of envelope are alike with respect to orders received, *i.e.*

$$\hat{p}_c = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} \quad \text{and} \quad \hat{q}_c = 1 - \hat{p}_c$$

The statistic Z is approximately a S.N.V.

(iv) Computations.

Now $\hat{p}_c = \dfrac{1000\,(0.10) + 1000\,(0.13)}{1000 + 1000} = \dfrac{230}{2000} = 0.115$, and

$$\hat{q}_c = 1 - 0.115 = 0.885.$$

Substituting the values, we get

$$z = \frac{0.13 - 0.10}{\sqrt{(0.115)\,(0.885)\left(\frac{1}{1000} + \frac{1}{1000}\right)}}$$

$$= \frac{0.03}{\sqrt{(0.101775)\,(0.002)}} = \frac{0.03}{\sqrt{0.000204}} = \frac{0.03}{0.0143} = 2.10$$

(v) The critical region is $Z > z_{0.05} = 1.645$.

(vi) Since the computed value falls in the critical region, so we reject H_0 in favour of H_1. There is evidence to conclude that the blue envelope will help sales.

Q.16.33. Let $\hat{p}_1$ denote the proportion of correct answers in the first group and $\hat{p}_2$, the proportion of correct answers in the second group.

(i) The hypotheses would be stated as

$H_0: p_1 - p_2 = 0$, *i.e.* the proportion of correct answers in both the groups is the same and equals

$$\hat{p}_c = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}, \text{ and}$$

$H_1: p_1 - p_2 \neq 0$.

(ii) We use a level of significance of $\alpha = 0.05$, and a two sided test.

(iii) The test-statistic would be

$$Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{p}_c\,\hat{q}_c\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

which under H_0 is approximately a S.N.V.

(iv) Computations.

Here $\hat{p}_1 = \dfrac{40}{60} = 0.67$ and $\hat{p}_2 = \dfrac{80}{140} = 0.57$

$\therefore$ $\hat{p}_c = \dfrac{40 + 80}{60 + 140} = \dfrac{120}{200} = 0.6$ and $\hat{q}_c = 1 - 0.6 = 0.4$.

Substituting these values, we get

$$z = \frac{0.67 - 0.57}{\sqrt{(0.6)\,(0.4)\left(\frac{1}{60} + \frac{1}{140}\right)}}$$

$$= \frac{0.10}{\sqrt{(0.24)\,(0.023810)}} = \frac{0.10}{\sqrt{0.005714}} = \frac{0.10}{0.0756} = 1.32$$

(v) The critical region is $|Z| > Z_{0.05} = 1.96$.

(vi) Since the computed value of Z does not fall in the critical region, so we accept H_0. Thus the first question fails to show the discriminating ability and therefore should be deleted from the examination.

Q.16.34. (i) We state our hypotheses as

$H_0: \sigma_1 = \sigma_2$ and $H_1: \sigma_1 \neq \sigma_2$

(ii) We use a significance level of $\alpha = 0.05$ and a two-sided test.

(iii) The test-statistic under H_0 would be

$$Z = \frac{S_1 - S_2}{\sqrt{\frac{S_1^2}{2n_1} + \frac{S_2^2}{2n_2}}}$$

which is approximately a standard normal for large sample sizes.

(iv) The critical region is $|Z| \geq 1.96$.

(v) Computations. Here $n_1 = 1{,}000$, $S_1 = 5.9$ years, $n_2 = 900$ and $S_2 = 6.1$ years

$$\therefore \quad z = \frac{5.9 - 6.1}{\sqrt{\frac{(5.9)^2}{2(1000)} + \frac{(6.1)^2}{2(900)}}}$$

$$= \frac{-0.2}{\sqrt{0.0174 + 0.0207}} = \frac{-0.2}{0.195} = -1.03.$$

(vi) **Conclusion.** Since the computed value of $z = -1.03$ does not fall in the critical region, we therefore accept H_0 and conclude that the samples can be reasonably regarded as drawn from *equally variable* normal populations.

Q.16.35. (i) Let p denote the probability of heads in a single toss of the coin. Then our null hypothesis that the coin is fair, will be formulated as

$H_0 : p = 0.5$, and the alternative hypothesis would be

$H_1 : p \neq 0.5$

(ii) The significance level is approximately 0.05.

(iii) The test-statistic to be used is x, the number of heads.

(iv) The critical region. First we compute the probabilities associated with X, the number of heads, by using the binomial distribution

$$P(X = x) = \binom{10}{x} p^x q^{10-x}$$

Under $H_0 : p = 0.5$, $P(X = x) = \binom{10}{x}\left(\frac{1}{2}\right)^{10}$

Now $$P(X = 0) = \binom{10}{0}\left(\frac{1}{2}\right)^{10} = 0.0010$$

$$P(X = 1) = \binom{10}{1}\left(\frac{1}{2}\right)^{10} = 0.0097$$

$$P(X = 2) = \binom{10}{2}\left(\frac{1}{2}\right)^{10} = 0.0440$$

.....

$$P(X = 8) = \binom{10}{8}\left(\frac{1}{2}\right)^{10} = 0.0440$$

$$P(X = 9) = \binom{10}{9}\left(\frac{1}{2}\right)^{10} = 0.0097$$

$$P(X = 10) = \binom{10}{10}\left(\frac{1}{2}\right)^{10} = 0.0010$$

Since $\alpha = 0.5$, so $\alpha/2 = 0.025$ (area in each tail)

We observe that $P(X \leq 2) = 0.0547 > 0.025$, and

$P(X \geq 8) = 0.0547 > 0.025$.

Therefore the true significance level is

$$\alpha = P(X \leq 1) + P(X \geq 9) = 0.0107 + 0.0107 = 0.0214$$

Thus the critical region is $X \leq 1$ and $X \geq 9$.

(v) Computation: $x = 8$.

(vi) **Conclusion.** Since $x = 8$ does not fall in the critical region, so we accept H_0 and conclude that the coin is fair.

❖❖❖❖❖❖❖❖❖

CHAPTER 17

THE CHI-SQUARE DISTRIBUTION AND STATISTICAL INFERENCE

Q.17.3 Let Z_i be normally distributed with mean zero and unit variance, where

$$Z_i = \frac{X_i - \mu}{\sigma}. \quad \text{Then} \quad \chi^2 = \sum_{i=1}^{n} Z_i^2$$

The *m.g.f.* of Z^2 is

$$M_0(t) = E[e^{tZ^2}] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tz^2} e^{-z^2/2}\, dz$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(1-2t)z^2/2}\, dz$$

$$= (1-2t)^{-1/2} \qquad \text{for } t < 1/2$$

We can obtain the distribution of a sum of squares of n independent standard normal variables $Z_1, Z_2, \ldots, Z_n$.

Since $\chi^2 = Z_1^2 + Z_2^2 + \ldots + Z_n^2$, therefore the m.g.f. of χ^2 is the nth power of the m.g.f. of a single term.

Hence $M_{\chi^2}(t) = [(1-2t)^{-1/2}]^n = (1-2t)^{-n/2}$

which is the m.g.f. of a chi-square distribution wih n degrees of freedom.

Let $X_1, X_2, \ldots, X_n$ denote the sample from a normal distribution with mean μ and variance σ^2. Then

$$Z_i = \frac{X_i - \mu}{\sigma}, \text{ and}$$

$$\bar{Z} = \frac{1}{n}\sum \frac{X_i - \mu}{\sigma} = \frac{\bar{X} - \mu}{\sigma} \text{ which is } N\left(0, \frac{1}{n}\right).$$

$$\text{Now} \quad \sum (Z_i - \bar{Z})^2 = \sum \left[\frac{X_i - \mu}{\sigma} - \frac{\bar{X} - \mu}{\sigma}\right]^2 = \sum \left(\frac{X_i - \bar{X}}{\sigma}\right)^2$$

$$\text{Moreover} \quad \sum (Z_i - \bar{Z})^2 = \sum Z_i^2 + n\bar{Z}^2 - 2\bar{Z}\sum Z_i$$

$$= \sum Z_i^2 - n\bar{Z}^2$$

$$\text{Thus} \quad \sum Z_i^2 = \sum (Z_i - \bar{Z})^2 + n\bar{Z}^2.$$

Let us accept the independence of $\sum(Z_i - \bar{Z})^2$ and $n\bar{Z}^2$ (which has been theoretically proved). Then applying the m.g.f. technique, we get

$$M_{\sum Z_i^2}(t) = M_{\sum (Z_i - \bar{Z})^2}(t)\, M_{n\bar{Z}^2}(t)$$

$$\text{or} \quad M_{\sum (Z_i - \bar{Z})^2}(t) = \frac{M_{\sum Z_i^2}(t)}{M_{n\bar{Z}^2}(t)}$$

$$= \frac{(1-2t)^{-n/2}}{(1-2t)^{-1/2}} = (1-2t)^{-(n-1)/2}$$

The factor $n\bar{Z}^2$ has a chi-square distribution with one d.f. as $\sqrt{n}\,\bar{Z}$ is a standard normal distribution. Thus the m.g.f. of $\sum(Z_i - \bar{Z})^2$ i.e. $\sum \left(\frac{X_i - \bar{X}}{\sigma}\right)^2$ is that of a chi-square with $n-1$ degrees of freedom. We kow that the term *degrees of freedom* refers to the number of independent squares in the sum. But the quantity $\sum(X_i - \bar{X})^2$ has only $n-1$ independent squares because $\sum_{i=1}^{n} (X_i - \bar{X}) = 0$

Hence the number of the degrees of freedom is $n-1$, which is in conformity with *the concept of* the term degrees of freedom.

Q.17.4. (a) We are required to show that

$$\sqrt{2\chi^2} - \sqrt{2n-1} \text{ is } N(0, 1)$$

when χ^2 has the chi-square distribution with n degrees of freedom and n is large.

Let $Y = \sqrt{2\chi^2} - \sqrt{2n-1}$

Then its distribution function is

$$P(Y<x) = P(\sqrt{2\chi^2} - \sqrt{2n-1} < x)$$

The expression $\sqrt{2\chi^2} - \sqrt{2n-1} < x$ may be written as

$$\sqrt{2\chi^2} < x + \sqrt{2n-1}$$

or $$2\chi^2 < x^2 + (2n-1) + 2x\sqrt{2n-1}$$

or $$\chi^2 < \frac{x^2-1}{2} + n + x\sqrt{2n-1}$$

or $$\chi^2 - n < \frac{x^2-1}{2} + x\sqrt{2n-1}$$

or $$\frac{\chi^2-n}{2n} < x + \frac{x^2-1}{2\sqrt{2n}} + x\left[\sqrt{1-\frac{1}{2n}} - 1\right]$$

Thus $$P(Y<x) = P\left[\frac{\chi^2-n}{2n} < x + \frac{x^2-1}{2\sqrt{2n}} + x\left(\sqrt{1-\frac{1}{2n}} - 1\right)\right]$$

$$\cong P\left(\frac{\chi^2-n}{2n} < x\right)$$

This approximation holds for large n. Hence Y has asymptotically the standard normal distribution, and for any α,

$$\sqrt{2\chi_\alpha^2} - \sqrt{2n-1} \cong Z_\alpha$$

or $$\chi_\alpha^2 \cong \frac{1}{2}\left[Z_\alpha + \sqrt{2n-1}\right]^2$$

(b) (i) **From Fisher's approximation $Z = \sqrt{2\chi^2} - \sqrt{2n-1}$ for any α, we get**

$$\chi_\alpha^2 = \frac{1}{2}\left[Z_\alpha + \sqrt{2n-1}\right]^2$$

Substituting $\alpha = 0.05$, $n = 40$ and $Z_{0.05} = 1.645$, we get

$$\chi_{0.05}^2 = \frac{1}{2}\left[1.645 + \sqrt{2(40)-1}\right]^2$$

$$= \frac{1}{2}[1.645 + 8.888]^2 = \frac{1}{2}[110.944] = 55.47;$$

for $n = 60$, we have

$$\chi_{0.05}^2 = \frac{1}{2}\left[1.645 + \sqrt{2(60)-1}\right]^2$$

$$= \frac{1}{2}[1.645 + 10.909]^2 = \frac{1}{2}[157.6029] = 78.80, \text{ and}$$

for $n = 150$, we have

$$\chi_{0.05}^2 = \frac{1}{2}\left[1.645 + \sqrt{2(105)-1}\right]^2$$

$$= \frac{1}{2}[1.645 + 14.457]^2 = \frac{1}{2}[259.2744] = 129.64.$$

(ii) From Wilson-Hilferty approximation, for any α, we obtain

$$\chi_\alpha^2 = n\left[\left(1 - \frac{2}{9n}\right) + Z_\alpha\sqrt{\frac{2}{9n}}\right]^3$$

Substituting $n = 40$, $\alpha = 0.05$, and $Z_{0.05} = 1.645$, we get

$$\chi_{0.05}^2 = 40\left[\left(1 - \frac{2}{9\times 40}\right) + 1.645\sqrt{\frac{2}{9\times 40}}\right]^3$$

$$= 40\,[1 - 0.0056 + 1.645\,(0.0745)]^3$$

$$= 40\,[1 - 0.0056 + 0.1226]^3$$

$$= 40\,(1.1170)^3 = 40(1.3937) = 55.75;$$

for $n = 60$, we get

$$\chi^2_{0.05} = 60\left[\left(1 - \frac{2}{9(60)}\right) + 1.645\sqrt{\frac{2}{9(60)}}\right]^3$$

$$= 60\,[1 - 0.0037 + 1.645\,(0.0609)]^3$$

$$= 60\,[1 - 0.0037 + 0.1002]^3$$

$$= 60\,[1.0965]^3 = 60(1.3183) = 79.10, \text{ and}$$

for $n = 105$, we obtain

$$\chi^2_{0.05} = 105\left[1 - \frac{2}{9(105)} + 1.645\sqrt{\frac{2}{9(105)}}\right]^3$$

$$= 105\,[1 - 0.0021 + 1.645\,(0.046)]^3$$

$$= 105\,[1 - 0.0021 + 0.0757]^3$$

$$= 105\,[1.0736]^3 = 105(1.2374) = 129.93$$

The values obtained by Fisher's approximation agree very well with the tabulated values of $\chi^2_{0.05(40)} = 55.76$ and $\chi^2_{0.05(60)} = 79.08$, while the values obtained by Wilson-Hilfirty approximation are almost identical with the tabulated ones.

Q.17.5. (b) The 98% confidence limits for σ^2 are given by

$$\frac{n\,S^2}{\chi^2_{0.01,(19)}} < \sigma^2 < \frac{n\,S^2}{\chi^2_{0.99,(19)}}$$

Substituting the values, we get

$$\frac{20 \times (5)^2}{36.19} < \sigma^2 < \frac{20 \times (5)^2}{7.63}$$

i.e., $\quad 13.82 < \sigma^2 < 65.53$

Hence the required confidence limits on the population variance are 13.82 to 65.53.

Q.17.6. (a) The 95% confidence interval for σ^2 is given by

$$\frac{\Sigma(X_i - \bar{X})^2}{\chi^2_{0.025,(9)}} < \sigma^2 < \frac{\Sigma(X_i - \bar{X})^2}{\chi^2_{0.975,(9)}}$$

where $\Sigma(X_i - \bar{X})^2 = \Sigma X_i^2 - \dfrac{(\Sigma X_i)^2}{n} = 21273.12 - \dfrac{(461.2)^2}{10}$

$$= 21273.12 - 21270.544 = 2.576,$$

$\chi^2_{0.025,(9)} = 19.023$ and $\chi^2_{0.975,(9)} = 2.700$

Substituting these values, we get

$$\frac{2.576}{19.023} < \sigma^2 < \frac{2.576}{2.700}$$

or $\quad 0.135 < \sigma^2 < 0.954.$

(b) The 95% confidence interval for σ^2 is given by

$$\frac{\Sigma(X_i - \bar{X})^2}{\chi^2_{0.025,(9)}} < \sigma^2 < \frac{\Sigma(X_i - \bar{X})^2}{\chi^2_{0.975,(9)}}$$

where $\Sigma(X_i - \bar{X})^2 = \Sigma X_i^2 - \dfrac{(\Sigma X)^2}{n} = 1012.58 - \dfrac{(100.6)^2}{10}$

$$= 1012.58 - 1012.036 = 0.544,$$

$\chi^2_{0.025,(9)} = 19.023$ and $\chi^2_{0.975,(9)} = 2.700$

Substituting these values, we get

$$\frac{0.544}{19.023} < \sigma^2 < \frac{0.544}{2.700} \text{ or } 0.0286 < \sigma^2 < 0.2015$$

Q.17.7. (a) The 96% confidence limits on σ^2 are given by

$$\frac{\Sigma(n_i\,S_i^2)}{\chi^2_{\alpha/2,(\Sigma n_i - k)}} < \sigma^2 < \frac{\Sigma(n_i\,S_i^2)}{\chi^2_{1-\alpha/2,(\Sigma n_i - k)}}$$

Substituting the values, we get

$$\frac{(5\times25) + (5\times36) + (10\times16)}{\chi^2_{0.02,(20-3)}} < \sigma^2 < \frac{(5\times25) + (5\times36) + (10\times16)}{\chi^2_{0.98,(20-3)}}$$

i.e. $\quad \dfrac{465}{31.00} < \sigma^2 < \dfrac{465}{7.26}$

i.e. $15.00 < \sigma^2 < 64.05$

Hence the required 96% confidence limits on σ^2 are (15,64).

(b) The 95% confidence interval for σ^2 is given by

$$\frac{\sum n_i S_i^2}{\chi^2_{0.025,(\Sigma n_i - k)}} < \sigma^2 < \frac{\sum n_i S_i^2}{\chi^2_{0.975,(\Sigma n_i - k)}}$$

where $\sum n_i S_i^2 = (3\times25) + (5\times16) + (7\times9) = 218$,

$\sum n_i - k = 12$, $\chi^2_{0.025,(12)} = 23.34$, and $\chi^2_{0.975,(12)} = 4.40$

Substituting these values, we get

$$\frac{218}{23.34} < \sigma^2 < \frac{218}{4.40} \text{ or } 9.34 < \sigma^2 < 49.55$$

Q.17.8. The pooled unbiased estimate of σ^2 is given by

$$s_p^2 = \frac{n_1 S_1^2 + n_2 S_2^2 + \ldots + n_k S_k^2}{(n_1 + n_2 + \ldots + n_k) - k} = \frac{\sum n_i S_i^2}{\sum n_i - k},$$

where k is the number of samples.

Thus $s_p^2 = \dfrac{9(23.5 + 30.6 + 29.3 + \ldots + 26.5)}{90 - 10} = \dfrac{(9)\,(274.0)}{80}$

$$= \frac{2466.0}{80} = 30.83$$

The 90% confidence limits for σ^2 are given by

$$\frac{(\sum n_i - k)\, s_p^2}{\chi^2_{\alpha/2,(\Sigma n_i - k)}} < \sigma^2 < \frac{(\sum n_i - k)\, s_p^2}{\chi^2_{1-\alpha/2,(\Sigma n_i - k)}}$$

Substituting the values, we obtain

$$\frac{80 \times 30.83}{\chi^2_{0.05,(80)}} < \sigma^2 < \frac{80 \times 30.83}{\chi^2_{0.95,(80)}}$$

i.e. $\dfrac{2466}{101.88} < \sigma^2 < \dfrac{2466}{60.39}$ *i.e.* $24.20 < \sigma^2 < 40.83$.

Hence the required confidence limits for σ^2 are (24.20, 40.83).

Q.17.9. (b) (i) Our null hypothesis is $H_0 : \sigma^2 = 4$. Let the alternative hypothesis be $H_1 : \sigma^2 > 4$ (one-tailed test)

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic is $\chi^2 = \dfrac{nS^2}{\sigma_0^2}$,

which has a χ^2-distribution with $(n-1)$ degrees of freedom.

(iv) The critical region is $\chi^2 \geq \chi^2_{0.05,(8)} = 15.51$.

(v) Computations. $\sum X = 5 + 7 + 2 + \ldots + 5 = 54$.

$$\sum X^2 = (5)^2 + (7)^2 + \ldots + (5)^2 = 364$$

$$\therefore \quad S^2 = \frac{\sum X^2}{n} - \left(\frac{\sum X}{n}\right)^2 = \frac{364}{9} - \left(\frac{54}{9}\right)^2$$

$$= 40.4444 - 36 = 4.4444, \text{ and}$$

$$\chi^2 = \frac{(9)\,(4.4444)}{4} = 10.$$

(vi) **Conclusion.** Since the computed value of $\chi^2 = 10$ does not fall in the critical region, we therefore accept H_0 and conclude that the process has the variance equal to 4 (inches)^2.

Q.17.10. (a) (i) **We state the hypotheses as**

$$H_0 : \sigma^2 = 20 \text{ and } H_1 : \sigma^2 \neq 20$$

(ii) We use a level of significance of $\alpha = 0.05$, and a two-tailed test.

(iii) The test-statistic would be

$$\chi^2 = \frac{n\,S^2}{\sigma^2},$$

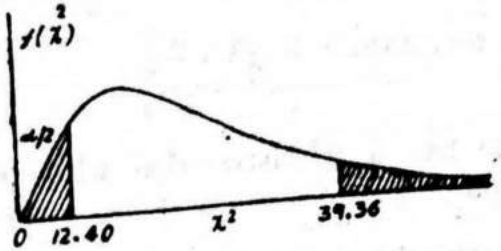
which has a chi-square distribution with $(n-1)$ *d.f.* assuming that the population is normal.

(iv) Computations.

$$\chi^2 = \frac{n\,S^2}{\sigma^2} = \frac{25 \times 12.6}{20} = 15.75$$

(v) The critical region is $\chi^2 \geq \chi^2_{0.025,(24)} = 39.36$, and

$$\chi^2 \leq \chi^2_{0.975,(24)} = 12.40$$



(vi) **Conclusion.** Since the computed value does not fall in the critical region, so we accept our null hypothesis $H_0 : \sigma^2 = 20$.

(b) (i) We wish to decide between the hypotheses

$$H_0 : \sigma^2 = 9 \quad \text{and} \quad H_1 : \sigma^2 \neq 9$$

(ii) We use a level of significance of $\alpha = 0.05$, and a two-sided test.

(iii) The test-statistic would be

$$\chi^2 = \frac{n S^2}{\sigma^2},$$

which has a chi-square distribution with $(n-1)$ *d.f.*, assuming that the population is normal.

(iv) Computations.

$$\chi^2 = \frac{25 \times 12}{9} = 33.33 \qquad (\because\ n = 25,\ S^2 = 12)$$

(v) The critical region is $\chi^2 \geq \chi^2_{0.025,(24)} = 39.36$ and

$$\chi^2 \leq \chi^2_{0.975,(24)} = 12.40$$

(vi) **Conclusion.** Since the computed value of χ^2 falls in the acceptance region, so we accept H_0. There is no evidence to doubt that the score of all the students in the school would have a variance of 9.

Q.17.11. (a) (i) We have to decide between the hypotheses $H_0 : \sigma^2 = 0.12$ and $H_1 : \sigma^2 > 0.12$ (one tailed)

(ii) The levels of significance are $\alpha = 0.05$ and $\alpha = 0.01$.

(iii) The test-statistic would be

$$\chi^2 = \frac{n S^2}{\sigma^2} \text{ or } \chi^2 = \frac{\sum (X - \bar{X})^2}{\sigma^2},$$

which has a chi-square distribution with $(n-1)$ *d.f.* under the assumptions that the population is normal and the sample is random.

(iv) Computations.

$$\sum X = 146.70,\ \sum X^2 = 1437.6190$$

$$\therefore \quad \sum (X - \bar{X})^2 = \sum X^2 - (\sum X)^2/n = 1437.6190 - (146.70)^2/15$$

$$= 1437.6190 - (1434.7260 = 2.8930$$

Thus $\chi^2 = \dfrac{2.8930}{0.12} = 24.108$

(v) The critical regions are $\chi^2 \geq \chi^2_{0.05,(14)} = 23.68$, and

$$\chi^2 \geq \chi^2_{0.01,(14)} = 29.14$$

(vi) **Conclusion.** We reject H_0 at 5% level but accept at 1% level of significance.

(b) (i) We are required to test the hypotheses

$$H_0 : \sigma^2 = (2.792)^2 \text{ against } H_1 : \sigma^2 > (2.792)^2$$

(ii) We use a level of significance of $\alpha = 0.05$, and a one-sided test.

(ii) The test-statistic would be

$$\chi^2 = \frac{n S^2}{\sigma^2},$$

which has a chi-square distribution with $(n-1)$ degrees of freedom, under the assumptions that the population is normal and the sample is random.

(iii) Computations.

$$\Sigma X = 67.50 + 70.75 + 72.00 + 63.25 + 65.25 + 68.75 + 69.25 + 68.50 + 66.50 + 64.75 = 676.50$$

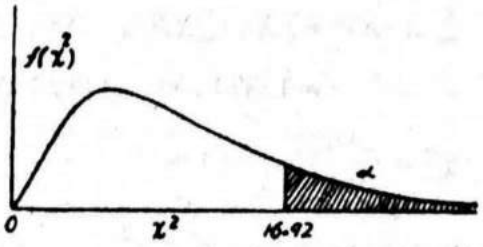
$$\Sigma X^2 = (67.50)^2 + (70.75)^2 + \ldots + (64.75)^2 = 45,833.1250$$

$$\therefore S^2 = \frac{\Sigma X^2}{n} - \left(\frac{\Sigma X}{n}\right)^2 = \frac{45,833.1250}{10} - \left(\frac{676.50}{10}\right)^2$$

$$= 4,583.3125 - 4,576.5225 = 6.79, \text{ and hence}$$

$$\chi^2 = \frac{n S^2}{\sigma^2} = \frac{10 \times 6.79}{(2.792)^2} = \frac{67.9}{7.8} = 8.71$$

(v) The critical region is $\chi^2 \geq \chi^2_{0.05,(9)} = 16.92$



(vi) **Conclusion.** Since the computed value does not fall in the critical region, so we accept H_0.

Q.17.12 (a) (i) We state our null and alternative hypotheses as

$$H_0 : \sigma^2 = 0.02 \text{ and } H_1 : \sigma^2 < 0.02.$$

(ii) The level of significance is set at $\alpha = 0.01$.

(iii) The test-statistic is $\chi^2 = \frac{n S^2}{\sigma^2}$,

which has a χ^2-distribution with $(n-1)$ degrees of freedom.

(iv) The critical region is $\chi^2 \leq \chi^2_{1-\alpha,(9)} = 2.09$

(v) Computations. $\Sigma X = 14.2 + 13.7 + \ldots + 14.3 = 141.6$

$$\Sigma X^2 = (14.2)^2 + (13.7)^2 + \ldots + (14.3)^2 = 2005.98$$

$$\therefore \quad \Sigma(X-\bar{X})^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{n}$$

$$= 2005.98 - \frac{(141.6)^2}{10} = 0.924$$

$$\text{Now} \quad \chi^2 = \frac{n S^2}{\sigma_0^2} = \frac{\Sigma (X - \bar{X})^2}{\sigma_0^2} = \frac{0.924}{0.02} = 46.2$$

(vi) **Conclusion.** Since our computed value of $\chi^2 = 46.2$ does not fall in the critical region, we therefore accept H_0.

(b) (i) We have to decide between the hypotheses

$$H_0 : \sigma = 0.9 \text{ and } H_1 : \sigma > 0.9$$

(ii) The significance level is set at $\alpha = 0.05$

(iii) The test-statistic is

$$\chi^2 = \frac{n S^2}{\sigma_0^2},$$

which under H_0, has a chi-square distribution with $(n-1)$ *d.f.*, assuming that the life of batteries is normally distributed.

(iv) The critical region is $\chi^2 \geq \chi^2_{0.05,(9)} = 16.92$

(v) Computations. $n = 10$ and $S = 1.2$,

$$\therefore \quad \chi^2 = \frac{(10)\ (1.2)^2}{(0.9)^2} = 17.78$$

(If 1.2 is taken as s, then $\chi^2 = \frac{(9)\ (1.2)^2}{(0.9)^2} = 16$)

(vi) **Conclusion.** Since the calculated value of χ^2 falls in the critical region, we therefore reject H_0. There is evidence to conclude that $\sigma > 0.9$ years.

Q.17.13. (b) (i) The hypotheses would be stated as

$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_5^2$; *i.e.* All the population variances are the smae, and

H_1 : Not all the variances are the same.

(ii) We use a level of significance of $\alpha = 0.05$, and one-sided test.

(iii) The test statistic would be

$$u = \frac{2.3026 \left(v \log_{10} s_p^2 - \sum_{i=1}^{k} v_i \log_{10} s_i^2\right)}{1 + \frac{1}{3(k-1)}\left[\sum_{i=1}^{k} \frac{1}{v_i} - \frac{1}{v}\right]}$$

where $s_p^2 = \frac{\sum_{i=1}^{k} v_i s_i^2}{\sum v_i}$,

v = total number of degrees of freedom and

k = no. of samples.

The statistic u is approximated by the chi-square distribution having $(k-1)$ *d.f.*

(iv) Computations for Bartlett's test for homogeneity of variances.

s_i^2	v_i	$1/v_i$	$v_i s_i^2$	$\log s_i^2$	$v_i \log s_i^2$
3.8	5	0.2000	19.0	0.5798	2.8990
4.4	8	0.1250	35.2	0.6435	5.1480
8.1	6	0.1667	48.6	0.9085	5.4510
6.1	7	0.1428	42.7	0.7853	5.4971
9.4	4	0.2500	37.6	0.9731	3.8924
Σ	30	0.8845	183.1	---	22.8875

The pooled unbiased estimate of σ^2, *i.e.*,

$$s_p^2 = \frac{\sum v_i s_i^2}{\sum v_i} = \frac{183.1}{30} = 6.103, \text{ and}$$

$$v \log s_p^2 = 30 \times 0.7855 = 23.5650.$$

Substituting these values, we get

$$u = \frac{2.3026\,(23.5650 - 22.8875)}{1 + \frac{1}{3\,(5-1)}\,[0.8845 - 0.0333]}$$

$$= \frac{2.3026 \times 0.6775}{1 + \frac{1}{12}\,(0.8512)} = \frac{1.5600}{1.0709} = 1.46.$$

(v) The critical region is $u \geq \chi^2_{0.05,(4)} = 9.49$.

(vi) Since the computed value of u does not fall in the cirtical region, so we cannot reject H_0. All the populaton variances may be regarded as homogeneous.

Q.17.14. (b) (i) The hypotheses would be stated as

$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2$ and H_1 : All variances are not equal.

(ii) We use a level of significance of $\alpha = 0.05$, and a one tailed test.

(iii) The test-statistic is

$$u = \frac{2.3026 \left(v \log_{10} s_p^2 - \sum v_i \log_{10} s_i^2\right)}{1 + \frac{1}{3(k-1)}\left[\sum_{i=1}^{k} \frac{1}{v_i} - \frac{1}{v}\right]}, \quad k = \text{no. of samples}$$

which is distributed approximately as a χ^2-variate with $(k-1)$ *d.f.*

(iv) Computations for the unbiased estimates of variances.

X_{1i}	X_{1i}^2	X_{2i}	X_{2i}^2	X_{3i}	X_{3i}^2
34	1156	40	1600	46	2116
40	1600	59	3481	93	8649
47	2209	60	3600	100	10000
60	3600	67	4489		
84	7056	86	7396		
		92	8464		
		95	9025		
		98	9604		
		108	11664		
Σ 265	15,621	705	59,323	239	20,765

$$s_1^2 = \frac{1}{n_1 - 1}\left[\sum X_{1i}^2 - \frac{(\sum X_{1i})^2}{n_1}\right] = \frac{1}{4}\left[15621 - \frac{(265)^2}{5}\right]$$

$$= \frac{1}{4}[15621 - 14045] = \frac{1576}{4} = 394,$$

$$s_2^2 = \frac{1}{n_2 - 1}\left[\sum X_{2i}^2 - \frac{(\sum X_{2i})^2}{n_2}\right] = \frac{1}{8}\left[59323 - \frac{(705)^2}{9}\right]$$

$$= \frac{1}{8}[59323 - 55225] = \frac{4098}{8} = 512.25, \text{ and}$$

$$s_3^2 = \frac{1}{n_3 - 1}\left[\sum X_{3i}^2 - \frac{(\sum X_{3i})^2}{n_3}\right] = \frac{1}{2}\left[20765 - \frac{(239)^2}{3}\right]$$

$$= \frac{1}{2}[20765 - 19040.33] = \frac{1724.67}{2} = 862.34.$$

Computations for Bartlett's test.

Sample	s_i^2	v_i (= d.f.)	$v_i s_i^2$	$1/v_i$	$\log_{10} s_i^2$	$v_i \log_{10} s_i^2$
1	394.00	4	1576.00	0.2500	2.5955	10.3820
2	512.25	8	4098.00	0.1250	2.7095	21.6760
3	862.34	2	1724.68	0.5000	2.9357	5.8714
Σ	--	14	7398.68	0.8750	--	37.9294

The pooled unbiased estimate of variance,

$$s_p^2 = \frac{\sum v_i s_i}{\sum v_i} = \frac{7398.68}{14} = 528.48$$

and $v \log_{10} s_p^2 = 14 \log 528.48 = 14 \times 2.7230 = 38.122.$

Substituting these values, we obtain

$$u = \frac{2.3026\,(38.122 - 37.9294)}{1 + \frac{1}{3\times 2}[0.8750 - 0.0714]}$$

$$= \frac{2.3026 \times 0.1926}{1 + 0.1339} = \frac{0.44348}{1.1339} = 0.39.$$

(v) The critical region is $u \geq \chi^2_{0.05,(2)} = 5.99$.

(vi) We accept H_0 as the computed value of u falls in the acceptance region. The apparent difference in variances is not significant.

Q.17.15. (a) (i) The hypotheses are stated as

$$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_5^2 = \sigma_6^2, \text{ and}$$

H_1 : Not all the variances are equal.

(ii) We use a level of significance of $\alpha = 0.05$, and a one-sided test.

(iii) The test-statistic is

$$u = 2.3026 \frac{q}{c}, \qquad \text{(Bartlett's test)}$$

where $q = (n-k) \log s_p^2 - \sum (n_i - 1) \log s_i^2$, and

$$c = 1 + \frac{1}{3\,(k-1)}\left(\sum \frac{1}{n_i - 1} - \frac{1}{n-k}\right).$$

The statistic u under H_0 has approximately a chi-square distribution with $(k-1)$ *d.f.*

(iv) Computations. We are given sample variances S_i^2, but for Bartlett's test, we need unbiased estimates of variances, which we obtain by the relation $s_i^2 = \dfrac{n_i S_i^2}{n_i - 1}$. The unbiased estimates of variances, *i.e.* s_i^2's are

13.00, 17.25, 14.625, 24.125, 20.50 and 19.75

Now $\Sigma(n_i-1)\, s_i^2 = \Sigma n_i\, S_i^2 = 437$, and the pooled unbiased estimate is $s_p^2 = \dfrac{\Sigma (n_i - 1)\, s_i^2}{\Sigma (n_i - 1)} = \dfrac{\Sigma n_i S_i^2}{\Sigma n_i - k} = \dfrac{437}{24} = 18.2083$

$\Sigma(n_i-1) \log s_i^2 = 4\ [1.1139 + 1.2368 + 1.1651 + 1.3825 + 1.3118 + 1.2956] = 4(7.5057) = 30.0228,$

$q = (n-k) \log 18.2083 - \Sigma (n_i - 1) \log s_i^2$

$= (24)\ (1.26027) - 30.0228 = 30.24648 - 30.0228$

$= 0.22368,$

$c = 1 + \dfrac{1}{3(6-1)}[1.50 - 0.0417] = 1 + 0.0667(1.4583) = 1.0973$

$\therefore\quad u = 2.3026 \left(\dfrac{0.22368}{1.0973}\right) = 0.47$

(v) The critical region is $u \geq \chi^2_{0.05,(5)} = 11.07$

(vi) **Conclusion.** Since the computed value of u does not fall in the critical region, we therefore accept H_0 and conclude that all the variances are equal.

(b) (i) The hypotheses are stated as

$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2$ and

H_1 : Not all the variances are equal.

(ii) The level of significance is set at $\alpha = 0.05$.

(iii) The test-statistic is

$$u = 2.3026 \frac{q}{c}, \qquad \text{(Bartlett's test)}$$

where $q = (n-k) \log s_p^2 - \Sigma (n_i - 1) \log s_i^2$, and

$$c = 1 + \frac{1}{3\,(k-1)} \left(\Sigma \frac{1}{n_i - 1} - \frac{1}{n-k} \right).$$

The statistic u under H_0 has approximately a χ^2-distribution with $(k-1)$ *d.f.*

(iv) Computations:

Sample 1: $\Sigma X_{1i} = 23,\ \Sigma X_{1i}^2 = 137,\ n_1 = 4,$

Sample 2: $\Sigma X_{2i} = 21,\ \Sigma X_{2i}^2 = 85,\ n_2 = 6,$

Sample 3: $\Sigma X_{3i} = 39,\ \Sigma X_{3i}^2 = 279,\ n_3 = 6.$

$s_1^2 = \dfrac{1}{3}\left[137 - \dfrac{(23)^2}{4}\right] = \dfrac{1}{3}[137 - 132.25] = \dfrac{4.75}{3} = 1.58,$

$s_2^2 = \dfrac{1}{5}\left[85 - \dfrac{(21)^2}{6}\right] = \dfrac{1}{5}[85 - 73.5] = \dfrac{11.5}{5} = 2.3,$

$s_3^2 = \dfrac{1}{5}\left[279 - \dfrac{(39)^2}{6}\right] = \dfrac{1}{5}[379 - 253.5] = \dfrac{25.5}{5} = 5.1,$

Sample	n_i-1	s_i^2	$(n_i-1) \times s_i^2$	$\log s_i^2$	$(n_i-1) \times \log s_i^2$	$1/(n_i-1)$
1	3	1.58	4.74	0.19866	0.59598	0.3333
2	5	2.3	11.5	0.36173	1.80865	0.2000
3	5	5.1	25.5	0.70757	3.53785	0.2000
Σ	13	--	41.74	--	5.94248	0.7333

Now $s_p^2 = \dfrac{41.74}{13} = 3.1923,$

$q = (n-k) \log 3.1923 - \Sigma(n_i - 1) \log s_i^2$

$= (13)\ (0.5041) - 5.94248 = 0.61082,$

$$c = 1 + \frac{1}{3(3-1)}[0.7333 - 0.07692]$$

$$= 1 + 0.1667\,(0.65638) = 1.094$$

$$\therefore \quad u = 2.3026\left(\frac{0.61082}{1.1094}\right) = 1.27$$

(v) The critical region is $u \geq \chi^2_{0.05,(2)} = 5.99$

(vi) **Conclusion.** The calculated value of u does not fall in the critical region. We accept H_0 and may conclude that the variances are equal.

Q.17.16. (i) The hypotheses are stated as

$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \ldots = \sigma_{10}^2$, and

H_1 : Not all the variances are equal.

(ii) We use a level of significance of $\alpha = 0.05$.

(iii) The test-statistic is

$$u = 2.3026\left(\frac{q}{c}\right), \qquad \text{(Bartlett's test)}$$

where $q = (n-k)\log s_p^2 - \sum (n_i - 1)\log s_i^2$, and

$$c = 1 + \frac{1}{3(k-1)}\left[\sum \frac{1}{n_i - 1} - \frac{1}{n-k}\right].$$

The statistic u under H_0 has approximately a chi-square distribution with $(k-1)$ *d.f.*

(iv) **Computations.** Here sample variances S_i^2 are given, but we need unbiased estimates of the variances to apply Bartlett's test. The unbiased estimates of variances are

27, 34.875, 32.625, 31.5, 31.5, 29.25, 33.75,
34.875, 24.75, 29.25 $\qquad (\because s_i^2 = (9/8)\, S_i^2)$

Now $\quad \sum (n_i - 1)\, s_i^2 = \sum n_i S_i^2 = 2475,$

and the pooled unbiased estimate of variance is

$$s_p^2 = \frac{\sum (n_i - 1)\, s_i^2}{\sum (n_i - 1)} = \frac{2475}{80} = 30.9375$$

$$\sum (n_i - 1)\log s_i^2 = 8[1.43136 + 1.54251 + 1.51356 + 1.49831 + 1.49831 + 1.46613 + 1.52827 + 1.54251 + 1.39358 + 1.46613]$$
$$= 8\,(14.88067) = 119.04536$$

$$q = 80 \log 30.9375 - 119.04536$$

$$= (80)\,(1.49049) - 119.04536$$

$$= 119.2392 - 119.04536 = 0.19384, \text{ and}$$

$$c = 1 + \frac{1}{3(9)}[1.25 - 0.0125] = 1.04584$$

$$\therefore \quad u = 2.3026\left(\frac{0.19384}{1.04584}\right) = 0.43$$

(v) The critical region is $\mu \geq \chi^2_{0.05,(9)} = 16.92$

(vi) **Conclusion.** The calculated value of u does not fall in the critical region, so we accept H_0 of equal variances.

Q.17.18. (a) Uses of the chi-square distribution. Chi-square is used to

(i) compute the confidence interval for σ^2, the variance of a Normal population;

(ii) test the variance of a normal population and the homogeneity of k $(k > 2)$ variances,

(iii) test the homogeneity of several correlation coefficients;

(iv) test the goodness of fit of observed distribution (data) to theoretical distribution (data);

(v) test the independence of the two variables of classification, *i.e.* of data in contingency tables;

(vi) test the hypothesis about equality of two or several proportions.

We state our hypotheses as

(i) $H_0 : p_1 = p^2$, $p_2 = 2pq$, $p_3 = q^2$ for a trinomial distribution involving three distinct classes with $p = 0.04$ and $n = 105$; and

$H_1 : p_i \neq p_{i0}$ for at least one value of $i = 1,2,3$

(ii) We use a significance level of $\alpha = 0.05$ and a one-tailed test.

(iii) The test statistic is

$$\chi^2 = \sum_{i=1}^{3} \frac{(n_i - np_i)^2}{np_i},$$

which is approximated by a chi-square distribution having $(k-1)$ *d.f.*

(iv) Computations. The expected frequencies under H_0 are:

$$np_1 = 105 \times (0.4)^2 = 16.8$$

$$np_2 = 105 \times 2(0.4)\,(0.6) = 50.4, \text{ and}$$

$$np_3 = 105 \times (0.6)^2 = 37.8.$$

Hence $$\chi^2 = \frac{(9-16.8)^2}{16.8} + \frac{(51-50.4)^2}{50.4} + \frac{(45-37.8)^2}{37.8}$$

$$= \frac{60.84}{16.8} + \frac{0.36}{50.4} + \frac{51.84}{37.8}$$

$$= 3.62 + 0.01 + 1.37 = 5.00.$$

and degrees of freedom = 3 − 1 = 2.

(v) The critical region is $\chi^2 \geq \chi^2_{0.05,(2)} = 5.99$.

(vi) Since the computed value of χ^2 does not fall in the critical region, so we do not reject H_0. The data are almost consistent with the genetic model with $p = 0.4$.

(b) (i) The hypotheses would be stated as

$H_0 : p_1 = q^2, p_2 = p^2 + 2pq, p_3 = r^2 + 2qr, p_4 = 2pr$ for a multinomial distribution involving 4 cells, with $p = 0.4$, $q = 0.4$, $r = 0.2$, $p+q+r = 1$ and $n = 770$.

$H_1 : p_i \neq p_{io}$ for at least one value of $i = 1, 2, 3, 4$.

(ii) We use a significance level of $\alpha = 0.05$, and a one sided test.

(iii) The test-statistic would be

$$\chi^2 = \sum_{i=1}^{4} \frac{(n_i - np_i)^2}{np_i},$$

which is approximated by a chi-square distribution with $(k-1)$ *d.f.*

(iv) Computations. Under H_0, the expected frequencies are:

$$np_1 = 770 \times (0.4)^2 = 123.2,$$

$$np_2 = 770\,[(0.4)^2 + 2(0.4)\,(0.4)] = 369.6,$$

$$np_3 = 770\,[(0.4)^2 + 2(0.4)\,(0.2)] = 154.0, \text{ and}$$

$$np_4 = 770 \times 2(0.4)\,(0.2) = 123.2.$$

Then χ^2 is computed as below:

	n_i	np_i	$n_i - np_i$	$(n_i - np_i)^2$	$(n_i - np_i)^2/np_i$
	180	123.2	56.8	3226.24	26.19
	360	369.6	−9.6	92.16	0.25
	132	154.0	−22.0	484.00	3.14
	98	123.2	−25.2	635.04	5.15
Σ	770	770	--	--	$\chi^2 = 34.73$

(v) The critical region is $\chi^2 \geq \chi^2_{0.05,(3)} = 7.82$.

(vi) Since the computed value of χ^2 falls in the critical region, so we reject H_0. Our null hypothesis of proportions of individuals possessing the four blood types with $p = 0.4$, $q = 0.4$, and $r = 0.2$ is not supported.

Q.17.19. (a) (i) We state our hypotheses as

$H_0 : p_1 = \frac{1}{4}$, $p_2 = \frac{2}{4}$, $p_3 = \frac{1}{4}$ for a trinomial distribution involving 3 cells and for which $n = 162$, and

$H_1 : p_i \neq p_{io}$ for at least one of $i = 1, 2, 3$.

(ii) We use a significance level of $\alpha = 0.05$ and a one-sided test.

(iii) The test-statistic is

$$\chi^2 = \sum_{i=1}^{3} \frac{(n_i - np_i)^2}{np_i},$$

which is approximated by a chi-square distribution with $(k-1)$ *d.f.*

(iv) Computations. Under H_0 the expected frequencies are:

Children of type $M = np_1 = 162 \times \frac{1}{4} = 40.5$,

Children of type $MN = np_2 = 162 \times \frac{2}{4} = 81.0$,

Children of type $N = np_3 = 162 \times \frac{1}{4} = 40.5$;

while the corresponding observed frequencies (n_i) are

$$n_1 = 162 \times \frac{284}{1000} = 46,$$

$$n_2 = 162 \times \frac{42}{100} = 68,$$

and $n_3 = 48$.

Hence $\chi^2 = \frac{(46-40.5)^2}{40.5} + \frac{(68-81)^2}{81} + \frac{(48-40.5)^2}{40.5}$

$$= \frac{30.25}{40.5} + \frac{169}{81} + \frac{56.25}{40.5}$$

$$= 0.75 + 2.09 + 1.39 = 4.23.$$

(v) The critical region is $\chi^2 \geq \chi^2_{0.05,(9)} = 5.99$.

(vi) Since the computed value of χ^2 falls in the acceptance region so we accept H_0. The data provide sufficient evidence to accept the truth of the genetic theory.

(b) Let p_1, p_2, p_3 and p_4 represent the mixing proportions of peanuts, hazelnuts, cashews and pecans respectively. Then the hypotheses are

(i) $H_0 : p_1 = \frac{5}{10}$, $p_2 = \frac{2}{10}$, $p_3 = \frac{2}{10}$ and $p_4 = \frac{1}{10}$; and

H_1 : The proportions differ from those specified in H_0.

(ii) We use a level of significance of $\alpha = 0.05$.

(iii) The test statistic is

$$\chi^2 = \sum_{i=1}^{4} \frac{(o_i - e_i)^2}{e_i} \text{ with } \upsilon = (k-1) \text{ d.f.}$$

(iv) Computations. The expected numbers, e_i under H_0 are calculated as below:

expected number of peanuts $(e_1) = 500 \times \frac{5}{10} = 250$

expected number of hazelnuts $(e_2) = 500 \times \frac{2}{10} = 100$

expected number of cashews $(e_3) = 500 \times \frac{2}{10} = 100$

expected number of pecans $(e_4) = 500 \times \frac{1}{10} = 50$

$$\therefore \quad \chi^2 = \frac{(269-250)^2}{250} + \frac{(112-100)^2}{100} + \frac{(74-100)^2}{100} + \frac{(45-50)^2}{50}$$

$$= 1.444 + 1.440 + 6.760 + 0.500 = 10.144,$$

and $\upsilon = 4 - 1 = 3$.

(v) The critical region is $\chi^2 \geq \chi^2_{0.05,(3)} = 7.81$

(vi) **Conclusion.** Since calculated value of χ^2 is greater than table value, so we reject H_0 and may conclude that the machine is not mixing nuts in the ratio 5 : 2 : 2 : 1.

Q.17.20. (i) We state our null and alternative hypotheses as

H_0: $p_1 = 0.45$, $p_2 = 0.5$, $p_3 = 0.04$, $p_4 = 0.01$ for a multinomial distribution involving four cells and with $n = 1{,}000$; and

$H_1 : p_i \neq p_{i0}$ for at leat one value of $i = 1, 2, 3, 4$.

(ii) Let the significance level be set at $\alpha = 0.05$.

(iii) The test-statistic under H_0 is

$$\chi^4 = \sum_{i=1}^{4} \frac{(n_i - np_{i0})^2}{np_{i0}},$$

which has an approximate chi-square distribution with 3 *d.f.*

(iv) Computations. Under H_0, the expected frequencies are:

$np_{10} = 1000 \times 0.45 = 450$; $np_{20} = 1000 \times 0.50 = 500$;

$np_{30} = 1000 \times 0.04 = 40$ and $np_{40} = 1000 \times 0.01 = 10$.

The value of χ^2 is then computed as below:

$$\chi^2 = \frac{(452-450)^2}{450} + \frac{(494-500)^2}{500} + \frac{(38-40)^2}{40} + \frac{(16-10)^2}{10}$$

$$= 0.0089 + 0.0720 + 0.1000 + 3.6000 = 3.78$$

(v) The critical region is $\chi^2 \geq \chi^2_{0.05,(3)} = 7.82$.

(vi) **Conclusion.** Since the calculated value of $\chi^2 = 3.78$ does not fall in the critical region, we therefore accept H_0 and conclude that the data are consistent with theory.

Q.17.21. (a) (i) We state our hypotheses as

$H_0 : p = \frac{1}{2}$ and the coin is fair, and

$H_1 : p \neq \frac{1}{2}$ and the coin is not fair.

(ii) The significance level is set at $\alpha = 0.05$

(iii) The test-statistic is $\chi^2 = \sum \frac{(0_i - e_i)^2}{e_i}$, which has approximately a chi-square distribution with one degree of freedom.

(iv) Computations. Under H_0, the expected number of heads = expected number of tails = 100.

$$\therefore \quad \chi^2 = \frac{(115-100)^2}{100} + \frac{(85-100)^2}{100} = 4.50$$

(v) The critical region is $\chi^2 \geq \chi^2_{0.05,(1)} = 3.84$.

(vi) **Conclusion.** Since the computed value of χ^2 falls in the critical region, so we reject H_0. The data indicate that the coin is not fair.

(b) (i) We state the hypotheses as

$H_0 : p_i = \frac{1}{6}$, $(i = 1,2,...,6)$ and the dice are fair, and

H_1 : not all p's are $\frac{1}{6}$.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic is $\chi^2 = \sum_i \frac{(0_i - e_i)^2}{e_i}$, which is approximated by a chi-square distribution with 1 *d.f.*

(iv) Computations. Under H_0,

P(getting a "Seven" with 2 dice) $= \frac{6}{36}$, and

P(getting an "Eleven" with 2 dice) $= \frac{2}{36}$

$\therefore$ Expected number of "sevens" $= 360 \times \frac{6}{36} = 60$, and

Expected number of "elevens" $= 360 \times \frac{2}{36} = 20$.

Thus $\chi^2 = \frac{(74-60)^2}{60} + \frac{(24-20)^2}{20} = \frac{196}{60} + \frac{16}{20} = 4.07$

(v) The critical region is $\chi^2 \geq \chi^2_{0.05,(1)} = 3.84$.

(vi) Since the computed value of χ^2 falls in the critical region, so we reject H_0.

Q.17.22. (a) Using Chi-square Approximation.

(i) We state the hypotheses in multinomial notation as

H_0: $p_1 = \frac{1}{2}$, $p_2 = \frac{1}{2}$ for a multinomial distribution involving 2 classes and with $n = 1600$; and

H_1: Both p's are not equal to 1/2.

(ii) We use a level of significance of $\alpha = 0.05$.

(iii) The test-statistic under H_0 is

$$\chi^2 = \frac{(n_1 - np_{10})^2}{np_{10}} + \frac{(n_2 - np_{20})^2}{np_{20}}$$

where n_1 and n_2 are the number of boys (52) and the number of girls (46). The test-statistic has an approximate χ^2-distribution with 1 *d.f.*

(iv) Computations. Under H_0, the expected numbers are

$$np_{10} = 98 \times \tfrac{1}{2} = 49, \quad np_{20} = 98 \times \tfrac{1}{2} = 49$$

$$\therefore \quad \chi^2 = \frac{(52-49)^2}{49} + \frac{(46-49)^2}{49} = \frac{18}{49} = 0.37$$

(v) The critical region is $\chi^2 \geq \chi^2_{0.05,(1)} = 3.84$

(vi) **Conclusion.** Since the calculated value of χ^2 does not fall in the critical region, we therefore do not reject our null hypothesis of equal sex division in the population.

Using Normal Approximation (to Binomial).

(i) We set up our hypotheses in binomial notation as

$$H_0 : p = \frac{1}{2} \text{ and } H_1 : p \neq \frac{1}{2}.$$

(ii) We use a level of significance of $\alpha = 0.05$.

(iii) The test-statistic under H_0 is

$$Z = \frac{x - np_0}{\sqrt{n\,p_0\,q_0}} \text{ (without continuity correction).}$$

where x is the number of boys and Z is approximately standard normal.

(iv) Computations.

$$z = \frac{52 - 49}{\sqrt{98\,(\frac{1}{2})\,(\frac{1}{2})}} = \frac{3}{4.95} = 0.61$$

(v) The critical region is $|Z| \geq 1.96$.

(vi) **Conclusion.** Since the calculated value $z = 0.61$ does not fall in the critical region, so we accept our null hypothesis of equal sex division in the population.

(b) We can apply χ^2 method to test whether or not the treatment is effective. We calculate the value of χ^2 with 40% mortality as below:

	Die	Not-die	Total
Observed	1	9	10
Expected	4	6	10

$$\chi^2 = \sum \frac{(|0-e|-1/2)^2}{e} = \frac{(|1-4|-1/2)^2}{4} + \frac{(|9-6|-1/2)^2}{6} = 2.60,$$

which does not exceed 3.84 at $\alpha = 0.05$. We cannot conclude that the treatment is effective.

Q.17.23. (a) (i) We state our hypotheses as

H_0: The births in four quarterly periods occur in the ratio : 2 : 1 : 1 :1, *i.e.*

$$p_1 = \frac{2}{5}, \; p_2 = p_3 = p_4 = \frac{1}{5}, \text{ with } n = 300.$$

$H_1 : p_i \neq p_{i0}$ for at least one value of $i = 1,2,3,4$.

(ii) The significance level is set at $\alpha = 0.10$.

(iii) The test-statistic under H_0 is

$$\chi^2 = \sum_{i=1}^{4} \frac{(n_i - np_{i0})^2}{np_{i0}},$$

which has an approximate chi-square distribution with 3 *d.f.*

(iv) Computations. Under H_0, the expected frequencies are

$$np_{10} = \frac{2}{5} \times 300 = 120; \; np_{20} = \frac{1}{5} \times 300 = 60 = np_{30} = np_{40}.$$

Now $$\chi^2 = \frac{(110-120)^2}{120} + \frac{(57-60)^2}{60} + \frac{(53-60)^2}{60} + \frac{(80-60)^2}{60}$$

$$= 0.8333 + 0.1500 + 0.8167 + 6.6667 = 8.47.$$

(v) The critical region is $\chi^2 \geq \chi^2_{0.10,(3)} = 6.25$

(vi) **Conclusion.** Since the computed value of $\chi^2 = 8.47$ falls in the critical region, we therefore reject H_0 and conclude that the data contradict the stated null hypothesis.

(b) (i) We state our hypotheses as

H_0 : The distribution of grades is uniform, and

H_1 : The distribution of grades is not uniform.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic under H_0 is

$$\chi^2 = \sum \frac{(n_i - np_{i0})^2}{np_{i0}},$$

which has an approximate chi-square distribution with 4 *d.f.*

(iv) Computations. Under H_0, the expected frequency of each grade is $\frac{1}{5}$ of $100 = 20$.

$$\therefore \quad \chi^2 = \frac{(14-20)^2}{20} + \frac{(18-20)^2}{20} + \frac{(32-20)^2}{20} + \frac{(20-20)^2}{20} + \frac{(16-20)^2}{20}$$

$$= 1.8 + 0.2 + 7.2 + 0 + 0.8 = 10.0$$

(v) The critical region is $\chi^2 \geq \chi^2_{0.05,(4)} = 9.49$.

(vi) **Conclusion.** Since the computed value of $\chi^2 = 10.0$ falls in the critical region, we therefore reject H_0, and conclude that the distribution of grades is not uniform.

Q.17.24. (a) (i) We state our hypotheses as

H_0 : The digits are uniformly distributed, and

H_1 : The digits are not equally distributed.

(ii) We use the significance level of $\alpha = 0.05$, and a one-sided test.

(iii) The test-statistic is $\chi^2 = \sum \frac{(n_i - np_{i0})^2}{np_{i0}}$, which is approximated by a chi-square distribution with $(k-1)$ *d.f.*

(iv) Computations. Under H_0, the expected frequency of each digit is $\frac{250}{10} = 25$.

$$\therefore \quad \chi^2 = \frac{1}{25}[(17-25)^2 + (31-25)^2 + (29-25)^2 + \ldots + (36-25)^2]$$

$$= \frac{1}{25}[64 + 36 + 16 + 49 + 121 + 25 + 100 + 25 + 25 + 121]$$

$$= \frac{582}{25} = 23.28, \text{ and } d.f. = 10-1 = 9.$$

(v) The critical region is $\chi^2 \geq \chi^2_{0.05,(9)} = 16.92$.

(vi) Since the computed value of χ^2 falls in the critical region, so we reject H_0. The observed distribution differs significantly from the expected distribution.

(b) (i) We state our hypotheses as

H_0: Equal numbers die in all age-groups, *i.e.* $p_i = \frac{1}{7}$ for $i = 1,2, \ldots, 7$ with $n = 147$.

H_1 : Some age-groups are more likely to die.

(ii) Let the significance level be set at $\alpha = 0.05$.

(iii) The test-statistic under H_0 is

$$\chi^2 = \sum \frac{(0_i - e_i)^2}{e_i},$$

which has an approximate chi-square distribution with 6 *d.f.*

(iv) Computations. Under H_0, the expected number of deaths in each age-group is $147 \div 7 = 21$.

$$\therefore \chi^2 = \frac{(40-21)^2}{21} + \frac{(35-21)^2}{21} + \ldots + \frac{(4-21)^2}{21}$$

$$= 17.1905 + 9.3333 + 5.7619 + 5.7619 + 3.0476 + 3.0476 + 13.7619$$

$$= 57.90.$$

(v) The critical region is $\chi^2 \geq \chi^2_{0.05,(6)} = 12.59$.

(vi) **Conclusion.** Since the computed value of $\chi^2 = 57.90$ falls in the critical region, we therefore reject H_0 and conclude that some age-group is more likely to die of a narcotics overdose than some other age group.

Q.17.25. (a) (i) We state our hypotheses as

$H_0: p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = \frac{1}{6}$, *i.e.* the die is balanced, and

H_1: The die is not balanced.

(ii) The level of significance is set at $\alpha = 0.01$.

(iii) The test-statistic under H_0 is

$$\chi^2 = \sum \frac{(0_i - e_i)^2}{e_i}, \text{ with } (k-1) \text{ } d.f.$$

(iv) Computations. Under H_0, each expected frequency is $180 \div 6 = 30$.

$$\therefore \quad \chi^2 = \frac{(20-30)^2}{30} + \frac{(36-30)^2}{30} + \ldots + \frac{(22-30)^2}{30}$$

$$= \frac{1}{30}[100 + 36 + 256 + 25 + 81 + 64] = \frac{562}{30} = 18.73.$$

(v) The critical region is $\chi^2 \geq \chi^2_{0.01,(5)} = 15.086$.

(vi) **Conclusion.** Since the computed value of χ^2 falls in the critical region, we therefore reject H_0 and may conclude that the die is not balanced.

(b) (i) We state our null hypothesis as

H_0: The total number of births is evenly distributed over the period and there is no seasonality.

(ii) We use a significance level of $\alpha = 0.05$.

(iii) The test-statistic is $\chi^2 = \sum \frac{(0_i - e_i)^2}{e_i}$, which is approximated by a chi-square distribution with $(k-1)$ *d.f.*

(iv) Computations. Under H_0, the expected number of births (on the basis of number of days in various months) is 50,959 for a month of 31 days, 49,315 for a month of 30 days and 46,027 for February. The value of χ^2 is then computed as:

Month	0_i	e_i	$0_i - e_i$	$(0_i - e_i)^2/e_i$
January	50,759	50,959	−200	0.78
February	46,472	46,027	445	4.30
March	51,419	50,959	460	4.15
April	49,670	49,315	355	2.56
May	51,371	50,959	412	3.33
June	47,388	49,315	−1,927	75.30
July	49,995	50,959	−964	18.24
August	51,043	50,959	84	0.14
September	52,162	49,315	2,847	164.36
October	50,824	50,959	−135	0.36
November	47,768	49,315	−1,547	48.53
December	51,129	59,959	170	0.57
Total	600,000	600,000	--	$\chi^2 = 322.62$

(v) The critical region is $\chi^2 \geq \chi^2_{0.05,(11)} = 19.68$

(vi) **Conclusion.** Since the computed value of χ^2 is far greater than the critical value, we therefore confidently reject our null hypothesis. The data strongly suggest the presence of seasonality.

Q.17.26. (b) (i) The hypotheses would be stated as

H_0: The population distribution is a binomial with $n=4$ and $p = 1/2$, and

H_1: The population distribution is not a binomial with $p = 1/2$.

(ii) We use a significance level of $\alpha = 0.05$, and a one-sided test.

(iii) The test-statistic under H_0, is

$$\chi^2 = \sum \frac{(0_i - e_i)^2}{e_i},$$

which is approximated by a chi-square distribution with $(k-1)$ *d.f.* where k denote the number of classes. Both the parameters n and p are known.

(iv) Computations. Under H_0, the expected frequencies for the 5 cells corresponding to 0, 1, 2, 3, 4 are obtained by the expansion of the binomial $800\ (\frac{1}{2} + \frac{1}{2})^4$. Expanding, we get

$$800\left(\frac{1}{2}\right)^4 \left[\binom{4}{0} + \binom{4}{1} + \binom{4}{2} + \binom{4}{3} + \binom{4}{4}\right]$$

$$= 50\,[1+4+6+4+1] = 50+200+300+200+50.$$

The value of χ^2 is computed below:

No. of male births	Families (0_i)	(e_i)	$0_i - e_i$	$(0_i - e_i)^2/e_i$
0	32	50	−18	6.48
1	178	200	−22	2.42
2	290	300	−10	0.33
3	236	200	36	6.48
4	64	50	14	3.92
Σ	800	800	--	$\chi^2 = 19.63$

As p is not estimated, therefore the number of *d.f.* is $5-1=4$.

(v) The critical region is $\chi^2 \geq \chi^2_{0.05,(4)} = 9.49$.

(vi) The computed value of χ^2 falls in the critical region. We therefore reject the hypothesis of a binomial distribution with $p = \frac{1}{2}$. Hence the data are not binomially distributed.

Q.17.27. (b) (i) We state our hypotheses as

H_0: The data conform to a binomial distribution with $n=3$ and $p = 1/3$, and

H_1: The data do not conform to a binomial distribution with $n=3$ and $p = 1/3$.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic under H_0 is

$$\chi^2 = \sum \frac{(0_i - e_i)^2}{e_i},$$

which has an approximate chi-square distribution with $(k-1)$ *d.f.* as p is not to be estimated.

(iv) Computations. Under H_0, the expected frequencies are the successive terms in the expansion of the binomial $648\ (\frac{2}{3} + \frac{1}{3})^3$. Expanding, we get 192, 288, 144, 24.

$$\therefore\ \chi^2 = \frac{(179-192)^2}{192} + \frac{(298-288)^2}{288} + \frac{(141-144)^2}{144} + \frac{(30-24)^2}{24}$$

$$= 0.8802 + 0.3472 + 0.0625 + 1.5000 = 2.79$$

(v) The critical region is $\chi^2 \geq \chi^2_{0.05,(3)} = 7.82$.

(vi) **Conclusion.** Since the calculated value of $\chi^2 = 2.79$ does not fall in the critical region, we therefore accept H_0 and conclude that the data conform to binomial distribution with $p = \frac{1}{3}$

Q.17.28. (i) Let p denote the probability of getting a 6. Then we state our hypotheses as

$H_0 : p = \frac{1}{6}$ and the dice are unbiased, and

$H_1 : p \neq \frac{1}{6}$ and the dice are biased.

(ii) We use a significance level of $\alpha = 0.05$, and a one-tailed test.

(iii) The test-statistic is

$$\chi^2 = \sum \frac{(0_i - e_i)^2}{e_i},$$

which is approximated by a chi-square distribution with $(k-1-m)$ *d.f.*, where k denotes the effective number of classes and m, the number of estimated parameters.

(iv) Computations. Under H_0, the expected frequencies are the successive terms in the expansion of the binomial $4096\left(\frac{5}{6} + \frac{1}{6}\right)^{12}$, which are given in column 3 in the following table:

Number of successes	Observed frequency	Expected frequency	$0_i - e_i$	$(0_i - e_i)^2/e_i$
0	447	459	−12	0.31
1	1145	1103	42	1.60
2	1181	1213	−32	0.84
3	796	809	−13	0.21
4	380	364	16	0.70
5	115	116	−1	0.01
6	24	27	−3	0.33
7 & over	8	5	3	1.80
Total	4096	4096	--	$\chi^2 = 5.80$

Here the number of degrees of freedom is $8-1=7$ as we lose only one degree of freedom associated with the restriction of total and p is not estimated.

(v) The critical region is $\chi^2 \geq \chi^2_{0.05,(7)} = 14.07$.

(vi) Since the computed value of χ^2 does not fall in the critical region, we do not reject the null hypothesis that the dice are unbiased. That is, so far as the χ^2-test is concerned there is no reason to suspect that the dice were biased.

Q.17.29. (a) Let p denote the probability of a head when a well balanced coin is tossed. Then our hypotheses would be stated as

(i) $H_0 : p = \frac{1}{2}$ and the coins are well-balanced, and

$H_1 : p \neq \frac{1}{2}$ and the coins are not balanced.

(ii) The level of significance is set at $\alpha = 0.01$.

(iii) The test-statistic is

$$\chi^2 = \sum \frac{(0_i - e_i)^2}{e_i},$$

which, if H_0 is ture, has an approximate chi-square distribution with $(k-1)$ *d.f.*

(iv) Computations. Under H_0, the expected frequencies are the successive terms in the expansion of $640\left(\frac{1}{2} + \frac{1}{2}\right)^6$, which are 10, 60, 150, 200, 150, 60 and 10.

$$\therefore \chi^2 = \frac{(13-10)^2}{10} + \frac{(70-60)^2}{60} + \frac{(137-150)^2}{150} + \frac{(210-200)^2}{200} + \frac{(145-150)^2}{150} + \frac{(56-60)^2}{60} + \frac{(9-10)^2}{10}$$

$$= 0.9 + 1.67 + 1.13 + 0.5 + 0.17 + 0.27 + 0.1 = 4.74$$

(v) The critical region is $\chi^2 \geq \chi^2_{0.01,(6)} = 16.81$

(vi) **Conclusion.** Since the computed value of χ^2 does not fall in the critical region, we therefore accept H_0 and conclude that the coins are well-balanced.

(b) To decide which probability distribution will be appropriate, we estimate the mean and variance. The mean is $\bar{x} = 0.628$ and $s^2 = 0.730$. As the mean and the variance are approximately equal, so the Poisson distribution becomes appropriate.

(i) We state our hypotheses as

H_0 : The data follow a Poisson distribution, and

H_1 : The data do not follow a Poisson distribution.

(ii) Let the significance level be set at $\alpha = 0.05$

(iii) The test-statistic under H_0 is

$$\chi^2 = \sum_{i=1}^{k} \frac{(0_i - e_i)^2}{e_i},$$

which has an approximate χ^2-distribution with (k−1−no. of estimated parameters) *d.f.*

(iv) Computations. To compute the expected frequencies under H_0, we consider the mean from the data to be estimate of μ, the Poisson parameter. Thus the fitted Poisson distribution is

$$p(x; 0.628) = \frac{e^{-0.628}\,(0.628)^x}{x!}, \text{ for } x = 1, 2, 3, \ldots$$

where $e^{-0.628} = 0.5337$. We calculate the expected frequencies (column 3) and the value of χ^2 as follows:

x	o_i	Expected $f(e_i)$	$o_i - e_i$	$(o_i - e_i)^2/e_i$
0	139	$250 \times e^{-0.028} = 133.4$	5.6	0.24
1	76	$250 \times e^{-0.628}(0.628) = 83.8$	−7.8	0.73
2	28	26.3	1.7	0.11
3	4	5.5		
4	2	0.9	0.5	0.04
5	1	0.1		
Σ	250	250		$\chi^2 = 1.12$

And *d.f.* = 4 − 1 − 1 = 2

(v) The critical region is $\chi^2 \geq \chi^2_{0.05,(2)} = 5.99$

(vi) **Conclusion.** Since the computed value of $\chi^2 = 1.12$ does not fall in the critical region, we therefore accept H_0 and conclude that the data follow a Poisson distribution.

Q.17.30. (i) We state our hypotheses as

H_0 : The population has a Poisson distribution, and

H_1 : The population does not have a Poisson distribution.

(ii) Let the significance level be set at $\alpha = 0.05$.

(iii) The test-statistic to use is

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

which under H_0 has an approximate χ^2-distribution with *d.f.* = k − 1 − no. of estimated parameters.

(iv) Computations. To compute the expected frequencies under H_0, we find the mean from the data and consider it to be the estimate of μ.

Thus $$\bar{x} = \frac{\Sigma fx}{\Sigma f} = \frac{100}{100} = 1$$

Then the Poisson distribution becomes

$$p(x; 1) = \frac{e^{-1}\,(1)^x}{x!}, \text{ for } x = 0, 1, 2, \ldots,$$

and where $e^{-1} = 0.3679$.

The calculations regarding the expected frequencies and χ^2 are shown as follows:

X	Obs. f	Expected f (e_i)	$o_i - e_i$	$(o_i-e_i)^2/e_i$
0	40	$100 \times e^{-1} = 36.79$	3.21	0.28
1	32	$100 \times e^{-1} = 36.79$	−4.79	0.62
2	18	$100 \times e^{-1} \times \frac{(1)^2}{2} = 18.39$	−0.39	0.01
3	8	$100 \times e^{-1} \times \frac{(1)^3}{6} = 6.13$		
4	2	$100 \times e^{-1} \times \frac{(1)^4}{24} = 1.53$	1.97	0.48
5+	0	$100\,(1-0.9963) = 0.37$		
Σ	100	100	--	$1.39 = \chi^2$

$d.f. = 4 - 1 - 1 = 2$

(v) The critical region is $\chi^2 \geq \chi^2_{0.05,(2)} = 5.99$

(vi) **Conclusion.** Since the computed value of $\chi^2 = 1.39$ does not fall in the critical region, we therefore accept H_0 and conclude that the data conform to a Poisson distribution.

Q.17.31. (i) We state our null hypothesis as

H_0 : The data fit a Poisson Distribution.

(ii) We use a significance level of $\alpha = 0.05$, and a one-tailed test.

(iii) The test-statistic is

$$\chi^2 = \Sigma \frac{(o_i - e_i)^2}{e_i}$$

which is distributed approximately as a χ^2-variate with $(k-2)$ $d.f.$

(iv) Computations. To compute the expected frequencies under H_0, we find the mean from the data and consider it to be the estimate of μ, the Poisson parameter. Thus

$$\bar{x} = \frac{\Sigma fx}{\Sigma f} = \frac{1}{1000}\,[0\times305 + 1\times365 + 2\times210 + \ldots + 7\times1]$$

$$= \frac{1201}{1000} = 1.20$$

The the fitted Poisson distribution is

$$p(x;\, 1.20) = \frac{e^{-1.20}\,(1.20)^x}{x!}, \text{ for } x = 0, 1, 2, \ldots$$

and $e^{-1.20} = 0.3012$.

The necessary computations are shown below:

x	Obs. f	Expected f (e_i)	$o_i - e_i$	$(o_i-e_i)^2/e_i$
0	305	$1000 \times e^{-1.2} = 301.2$	3.8	0.05
1	365	$1000 \times e^{-1.2} \times 1.2 = 361.4$	3.6	0.04
2	210	$1000 \times e^{-1.2} \times \frac{(1.2)^2}{2} = 216.8$	−6.8	0.21
3	80	$1000 \times e^{-1.2} \times \frac{(1.2)^3}{6} = 86.7$	−6.7	0.52
4	28	$1000 \times e^{-1.2} \times \frac{(1.2)^4}{24} = 26.0$	2.0	0.15
5	9	$1000 \times e^{-1.2} \times \frac{(1.2)^5}{120} = 6.2$		
6	2	$1000 \times e^{-1.2} \times \frac{(1.2)^6}{6!} = 1.2$	4.1	1.40
7+	1	$1000 \times (1-0.9995) = 0.5$		
Σ	1000	1,000	--	2.37

The last three frequencies are grouped together as no expected frequency should be < 5. Then $d.f. = 6-2 = 4$.

(v) The critical region is $\chi^2 \geq \chi^2_{0.05,(4)} = 9.49$.

(vi) Since the computed value of $\chi^2 = 2.37$ does not fall in the critical region, so we do not reject H_0. Thus there is evidence that these data arise with Poisson distribution.

Q.17.32. Fitting of Normal Distribution.

The midpoints of the classes are 5, 6, 7, ..., 19. Taking $D = x - 12$, we find that $\Sigma fD = 6$ and $\Sigma fD^2 = 6898$. Thus

$$\bar{x} = a + \frac{\Sigma fD}{n} = \text{Rs. } 12 + \frac{6}{1000} = \text{Rs. } 12.006 \text{ and}$$

$$s = \sqrt{\frac{\Sigma fD^2}{n} - \left(\frac{\Sigma fD}{n}\right)^2} = \sqrt{\frac{6898}{1000} - \left(\frac{6}{1000}\right)^2}$$

$$= \sqrt{6.897964} = \text{Rs. } 2.626.$$

To compute the expected frequencies, we consider $\bar{x}$ and s to be the estimates of μ and σ. The expected frequencies are then calculated as follows:

Upper Class boundary	$z_i = \frac{ucb - \bar{x}}{s}$	$P(Z < z)$	$\hat{p}_i$	Expected frequency e_i $(n\hat{p}_i)$
5.5	−2.48	0.0066	0.0066	6.6
6.5	−2.10	0.0179	0.0113	11.3
7.5	−1.72	0.0427	0.0248	24.8
8.5	−1.34	0.0901	0.0474	47.4
9.5	−0.95	0.1711	0.0810	81.0
10.5	−0.57	0.2843	0.1132	113.2
11.5	−0.19	0.4247	0.1404	140.4
12.5	+0.19	0.5753	0.1506	150.6
13.5	0.57	0.7157	0.1404	140.4
14.5	0.95	0.8289	0.1132	113.2
15.5	1.33	0.9082	0.0793	79.3
16.5	1.71	0.9564	0.0482	48.2
17.5	2.09	0.9817	0.0253	25.3
18.5	2.47	0.9932	0.0115	11.5
∞	+∞	1.0000	0.0068	6.8
Σ	--	--	--	1,000

To test the goodness of fit, we proceed as follows:

(i) We state our null hypothesis as

H_0 : The data fit a normal distribution.

(ii) We use a significance level of $\alpha = 0.05$, and a one sided test.

(iii) The test-statistic would be

$$\chi^2 = \Sigma \frac{(o_i - e_i)^2}{e_i},$$

which has an approximate chi-square distribution with $(k-3)$ degrees of freedom.

(iv) Computation of χ^2.

	Obs. f	Exp. f	$o_i - e_i$	$(o_i - e_i)^2$	$(o_i - e_i)^2/e_i$
	6	6.6	−0.6	0.36	0.05
	17	11.3	5.7	32.49	2.88
	35	24.8	10.2	104.04	4.20
	48	47.4	0.6	0.36	0.01
	65	81.0	−16.0	256.00	3.16
	90	113.2	−23.2	538.24	4.75
	131	140.4	−9.4	88.36	0.63
	173	150.6	22.4	501.76	3.33
	155	140.4	14.6	213.16	1.52
	117	113.2	3.8	14.44	0.13
	75	79.3	−4.3	18.49	0.23
	52	48.2	3.8	14.44	0.30
	21	25.3	−4.3	18.49	0.73
	9	11.5	−2.5	6.25	0.54
	6	6.8	−0.8	0.64	0.09
Σ	1,000	1,000	--	--	22.55

The degrees of freedom = 15−3 = 10.

(v) The critical region is $\chi^2 \geq \chi^2_{0.05,(12)} = 21.03$

(vi) Since the computed value of the test-statistic falls in the critical region, so we reject H_0. There is evidence to indicate that these data do not fit a Normal Distribution.

Q.17.33. (i) We state our hypotheses as

H_0 : The data follow a Normal Distribution, and

H_1 : The data do not follow a normal distribution.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic to use is

$$\chi^2 = \Sigma \frac{(o_i - e_i)^2}{e_i},$$

which under H_0, has an approximate chi-square distribution.

(iv) Computations. First we fit the normal distribution, for which we find the values of mean and standard deviation. Thus

$$\bar{x} = \frac{\Sigma fx}{\Sigma f} = \frac{13594}{200} = 67.97, \text{ and}$$

$$s = \sqrt{\frac{\Sigma fx^2}{\Sigma f} - \left(\frac{\Sigma fx}{\Sigma f}\right)^2} = \sqrt{\frac{927728}{200} - \left(\frac{13594}{200}\right)^2}$$

$$= \sqrt{4638.64 - 4619.9209} = 4.33.$$

The expected frequencies and the value of χ^2 are computed as follows:

Upper class boundery	$z_i = \frac{ucb - \bar{x}}{s}$	$P(Z < z)$	$\hat{p}_i$	Expected frequency $e_i (= n\hat{p}_i)$
60.5	−1.73	0.0418	0.0418	8.4
63.5	−1.03	0.1515	0.1097	21.9
66.5	−0.34	0.3669	0.2154	43.1
69.5	0.35	0.6368	0.2699	54.0
72.5	1.05	0.8531	0.2163	43.3
75.5	1.74	0.9591	0.1060	21.2
+∞	+∞	1.0000	0.0409	8.2
Σ	--	--	--	200.0

Now $\chi^2 = \frac{(9-8.4)^2}{8.4} + \frac{(20-21.9)^2}{21.9} + \ldots + \frac{(11-8.2)^2}{8.2} = 2.10,$

and $d.f. = 7 - 1 - 2 = 4.$

(v) The critical region is $\chi^2 \geq \chi^2_{0.05,(4)} = 9.49.$

(vi) **Conclusion.** Since the computed value of $\chi^2 = 2.10$ does not fall in the critical region, we therefore accept H_0 and conclude that a normal distribution gives a satisfactory fit to the given data at $\alpha = 0.05$.

Q.17.34. (b) The necessary computations are given below:

Now $(A) + (\alpha) = N = 500$ and $(A) - (\alpha) = 80$

Adding, we get

$$2(A) = 580 \quad \text{or} \quad (A) = 290$$

Again $(B) + (\beta) = N = 500$ and $(B) - (\beta) = 200.$

Adding, we get

$$2(B) = 700 \quad \text{or} \quad (B) = 350$$

$$(C) = N - (\gamma) = 500 - 125 = 375.$$

$(AB) = (A) - (A\beta) = 290 - 18 = 272$

$(BC) = (B) - (B\gamma) = 350 - 98 = 252.$

$(\alpha\beta\gamma) = \alpha\beta\gamma \,.\, N = (1-A)\,(1-B)\,(1-C)\,.\,N$

$= N-(A)-(B)-(C)+(AB) + (AC) + (BC) - (ABC)$

$\therefore\ 25 = 500 - 290 - 350 - 375 + 272 + (AC) + 252 - 240$

or $(AC) = 231 + 25 = 256$

Now, we find the ultimate class-frequencies as follows:

$(\alpha BC) = \alpha BC \,.\, N = (1-A)\,BC \,.\, N = (BC) - (ABC)$

$= 252 - 240 = 12$

$(A\beta C) = A\beta C \,.\, N = A\,(1-B)\,C \,.\, N = (AC) - (ABC)$

$= 256 - 240 = 16$

$(AB\gamma) = AB\gamma \,.\, N = AB\,(1-C) \,.\, N = (AB) - (ABC)$

$= 272 - 240 = 32$

$(A\beta\gamma) = A\beta\gamma \,.\, N = A(1-B)\,(1-C) \,.\, N$

$= (A)-(AB)-(AC)+(ABC) = 290-272-256+240 = 2.$

$(\alpha\beta C) = \alpha\beta C \,.\, N = (1-A)\,(1-B)\,C \,.\, N$

$= (C) - (AC)-(BC)+(ABC) = 375-256-252+240 = 107$

$(\alpha B\gamma) = \alpha B\gamma \,.\, N = (1-A)\,B(1-C) \,.\, N$

$= (B)-(AB)-(BC)+(ABC) = 350-272-252+240 = 66$

$(\alpha\beta\gamma) = 25$ and $(ABC) = 240.$

Q.17.35. Let A, B and C denote the success in the first Terminal, in the second Terminal and in the Annual examination respectively. Then the given data reduce to

$(A) = 120,\quad (B) = 112,\quad (C) = 144,\quad (ABC) = 38,$

$(\alpha\beta\gamma) = 69,\quad (AB\gamma) = 44,\quad (\alpha\beta C) = 63$ and $N = 300.$

We are required to find the number of students who passed *at least* two (*i.e.* two or more) examinations. In other words, we are required to find the value of $(AB\gamma) + (A\beta C) + (\alpha BC) + (ABC).$

Now $(C) = (ABC) + (\alpha BC) + (A\beta C) + (\alpha\beta C)$

$\therefore\quad (ABC) + (\alpha BC) + (A\beta C) = (C) - (\alpha\beta C)$

Adding $(AB\gamma)$ to both sides, we get

$(ABC) + (\alpha BC) + (A\beta C) + (AB\gamma) = (C)-(\alpha\beta C)+(AB\gamma)$

$= 144-63+44 = 125$

Hence 125 students passed at least two examinations.

Q.17.36. (b) Let A, B and C denote males, married and college graduates respectively. Then the given data reduce to

$N = 1{,}000;\ (A) = 312,\ (B) = 470,\ (C) = 525,$

$(AC) = 42,\ (BC) = 147,\ (AB) = 86$ and $(ABC) = 25$

The criterion for consistence of data is that no ultimate class-frequency should be negative. The information would be regarded incorrect if at least one of the ultimate class-frequencies happens to be negative. Thus

$(\alpha\beta\gamma) = \alpha\beta\gamma \,.\, N = (1-A)\,(1-B)\,(1-C)\,.\,N$

$= N-(A)-(B)-(C)+(AB)+(AC)+(BC)-(ABC)$

$= 1{,}000-312-470-525+86+42+147-25$

$= -57$, which is negative.

Hence the numbers reported in the various groups are not consistent. In other words, the given information is incorrect.

Q.17.37. Two attributes A and B are said to be *independent* if the observed frequency equals the expected one, that is, if

$$(AB) = \frac{(A)\,(B)}{N}$$

or if $(AB)\ (\alpha\beta) = (A\beta)\ (\alpha B)$, when the ultimate class-frequencies on the two attributes A and B are given.

Two attributes A and B are said to be *positively associated* or simply associated, if $(AB) > \frac{(A)\ (B)}{N}$, *i.e.*, observed frequency > expected frequency. Two attributes A and B are said to be *negatively associated*, or briefly *dissociated* if $(AB) < \frac{(A)\ (B)}{N}$, *i.e.* observed frequency < expected frequency.

Now $(A) = (AB) + (A\beta) = 256 + 48 = 304,$

$(B) = (AB) + (\alpha B) = 256 + 768 = 1024,$ and

$(N) = (AB) + (A\beta) + (\alpha B) + (\alpha\beta) = 1216.$

Thus the expected frequency of $AB = \frac{(A)\ (B)}{N}$

$= \frac{(304)\ (1024)}{1216} = 256$

$= (AB) =$ observed frequency.

Hence the two attributes A and B are independent.

Q.17.38. (b) (i) We find the expected frequency of AB as

Expected frequency of $AB = \frac{(A)\ (B)}{N}$

$= \frac{2350 \times 3100}{5000} = 1457$

Since observation, *i.e.* $(AB) > 1457$, *i.e.* the expectation, therefore A and B are positively associated.

(ii) Here $(A) = 490,$

$(B) = (AB) + (\alpha B) = 294 + 380 = 674,$ and

$(N) = (A) + (\alpha) = 490 + 570 = 1060.$

$\therefore$ Expected $(AB) = \frac{(A)\ (B)}{N}$

$= \frac{490 \times 674}{1060} = 311.16,$ and

Observed $(AB) = 294$ which is < 311.6.

Hence A and B are negatively associated.

(iii) In case of ultimate class-frequencies, the criterion of independence of the two attributes A and B is

$$(AB)\ (\alpha\beta) = (A\beta)\ (\alpha B).$$

Now $(AB)\ (\alpha\beta) = (256)\ (144)$

$= 36864 = (48)\ (768) = (A\beta)\ (\alpha B).$

Hence A and B are independent.

Q.17.41. (b) (i) We state our hypotheses as

H_0: The results of the final test are independent of those of the preliminary test, and

H_1: The results of the final test are not independent of those of the preliminary test.

(ii) We use a significance level of $\alpha = 0.05$, and a one-tailed test.

(iii) The test-statistic is $\chi^2 = \sum_i \sum_j (o_{ij} - e_{ij})^2 / e_{ij}$, which is approximated by a chi-square distribution with $(r-1)\ (c-1)$ *d.f.*

(iv) Computations. Under H_0, we determine the expected frequencies, which are

$$e_{11} = \frac{800 \times 740}{1000} = 592, \quad e_{21} = \frac{800 \times 260}{1000} = 208,$$

$$e_{12} = \frac{200 \times 740}{1000} = 148, \quad e_{22} = \frac{200 \times 260}{1000} = 52.$$

Then the χ^2-value is obtained as under:

	o_{ij}	e_{ij}	$(o_{ij}-e_{ij})$	$(o_{ij}-e_{ij})^2$	$(o_{ij}-e_{ij})^2/e_{ij}$
	605	592	13	169	0.285
	195	208	−13	169	0.812
	135	148	−13	169	1.142
	65	52	13	169	3.250
Σ	1000	1000	--	--	$\chi^2 = 5.489$

(v) The critical region is $\chi^2 \geq \chi^2_{0.05,(1)} = 3.841$

(vi) Since the computed value exceeds the critical value, we therefore reject H_0. The results of the final test are almost certainly associated with those of the preliminary test.

Q.17.42 (a) (i) We state our hypotheses as

H_0 : The two variables of classification are independent, and

H_1 : The two classifications are associated.

(ii) We use a significance level of $\alpha = 0.05$ and a one tailed test.

(iii) The test-statistic is $\chi^2 = \sum\sum (o_{ij}-e_{ij})^2/e_{ij}$ which is approximately distributed as a χ^2-variate with 1 *d.f.*

(iv) Computations. Under H_0, we determine the first cell expected frequency as $e_{11} = \frac{1318\times553}{1518} = 480$. The other three expected frequencies are obtained by subtracting 480 from the marginal totals, etc. Then the value of χ^2 is:

	o_{ij}	e_{ij}	$o_{ij}-e_{ij}$	$(o_{ij}-e_{ij})^2$	$(o_{ij}-e_{ij})^2/e_{ij}$
	528	480	48	2304	4.80
	790	838	−48	2304	2.75
	25	73	−48	2304	31.56
	175	127	48	2304	18.14
Σ	1518	1518	--	--	$\chi^2=57.25$

(v) The critical region is $\chi^2 \geq \chi^2_{0.05,(1)} = 3.841$

(vi) The computed value far exceeds the critical value. We therefore confidently reject H_0 and conclude that the two classifications are associated.

(b) (i) The hypotheses would be stated as

H_0 : The two variables of classification are independent, and

H_1 : The two variables of classification are not independent, *i.e.* they are associated.

(ii) We use a significance level of $\alpha = 0.05$; and a one tailed test.

(iii) The test-statistic is $\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(o_{ij}-e_{ij})^2}{e_{ij}}$, which is approximated by a chi-square distribution with $(r-1)(c-1)$ *d.f.*

(iv) Computations. Under H_0, the expected frequencies are

$$e_{11} = \frac{600\times555}{1000} = 333, \quad e_{12} = \frac{400\times555}{1000} = 222,$$

$$e_{21} = \frac{600\times445}{1000} = 267, \quad e_{22} = \frac{400\times445}{1000} = 178.$$

and the value of χ^2 is obtained as

	o_{ij}	e_{ij}	$o_{ij}-e_{ij}$	$(o_{ij}-e_{ij})^2$	$(o_{ij}-e_{ij})^2/e_{ij}$
	350	333	17	289	0.87
	205	222	−17	289	1.30
	250	267	−17	289	1.08
	195	178	17	289	1.62
Σ	1000	1000	--	--	$\chi^2=4.87$

(v) The critical region is $\chi^2 \geq \chi^2_{0.05,(1)} = 3.84$.

(vi) Since the computed value of χ^2 falls in the critical region, so we reject H_0. These data provide evidence to indicate association between the two variables of classification.

Q.17.43. (i) The hypotheses would be stated as

H_0 : The two classifications are independent, and

H_1 : The two classifications are not independent.

(ii) We use a significance level of $\alpha = 0.05$, and a one-tailed test.

(iii) The test-statistic is $\chi^2 = \sum_i \sum_j (o_{ij}-e_{ij})^2/e_{ij}$, which has an approximate chi-square distribution with $(r-1)(c-1)$ degrees of freedom.

(iv) Computations: Under H_0, the expected frequencies are

$$e_{11} = \frac{350\times600}{1000} = 210, \quad e_{12} = \frac{500\times600}{1000} = 300,$$ the rest being obtained by subtracting these numbers from the marginal totals as they will not change. Thus χ^2-value is obtained as

o_{ij}	e_{ij}	$o_{ij}-e_{ij}$	$(o_{ij}-e_{ij})^2$	$(o_{ij}-e_{ij})^2/e_{ij}$
215	210	5	25	0.12
135	140	−5	25	0.18
325	300	25	625	2.08
175	200	−25	625	3.12
60	90	−30	900	10.00
90	60	30	900	15.00
Σ 1000	1000	--	--	$\chi^2=30.50$

Number of $d.f.$ = (3−1) (2−1) = 2.

(v) The critical region is $\chi^2 \geq \chi^2_{0.05,(2)} = 5.99$

(vi) Since the computed value of χ^2 falls in the critical region, so the hypothesis that the two variables of classification are independent, is rejected. Hence there is sufficient evidence to indicate association between the two classifications.

Q.17.44. (a) (i) We state our hypotheses as

H_0: The two variables of classification are independent, and

H_1: The two variables of classification are not independent.

(ii) We use a significance level of $\alpha = 0.05$, and a one-sided test.

(iii) The test-statistic is $\chi^2 = \sum_i \sum_j (o_{ij}-e_{ij})^2/e_{ij}$, which is approximated by a chi-square distribution with $(r-1)(c-1)$ $d.f.$

(iv) Computations. Under H_0, we determine the expected frequencies as

$$e_{11} = \frac{562\times930}{1600} = 326.7, \quad e_{12} = \frac{498\times930}{1600} = 289.5, \text{ etc.}$$

Then the χ^2-value is obtained as

o_{ij}	e_{ij}	$o_{ij}-e_{ij}$	$(o_{ij}-e_{ij})^2$	$(o_{ij}-e_{ij})^2/e_{ij}$
337	326.7	10.3	106.09	0.32
225	235.3	−10.3	106.09	0.45
291	289.5	1.5	2.25	0.01
207	208.5	−1.5	2.25	0.01
302	313.8	11.8	139.24	0.44
238	226.2	−11.8	139.24	0.62
Σ 1600	1600	--	--	$\chi^2=1.85$

and the number of $d.f.$ = (3−1) (2−1) = 2.

(v) The critical region is $\chi^2 \geq \chi^2_{0.05,(2)} = 5.99$.

(vi) Since the computed value of the test-statistic does not fall in the critical region, so we accept the null hypothesis of independence of two classifications.

(b) (i) We state our hypotheses as

H_0: The home ownership is independent of the family income level, and

H_1: The home ownership is associated with the family income level.

(ii) The level of significance is set at $\alpha = 0.01$.

(iii) The test-statistic to use is

$$\chi^2 = \sum_i \sum_j \frac{(o_{ij}-e_{ij})^2}{e_{ij}},$$

which is approximated by a chi-square distribution with $(r-1)(c-1)$ *d.f.*

(iv) Computatioins. The observed numbers are first obtained by multiplying the respective percentages by 400, and they are:

	A_1	A_2	A_3	Total
Home owner	20 (40)	140 (120)	40 (40)	200
Renter	60 (40)	100 (120)	40 (40)	200
Total	80	240	80	400

Next, we calculate under H_0, the expected frequencies which are shown in brackets in the table given above.

$$\text{Now} \quad \chi^2 = \frac{(20-40)^2}{40} + \frac{(60-40)^2}{40} + \frac{(140-120)^2}{120} + \frac{(100-120)^2}{120} + \frac{(40-40)^2}{40} + \frac{(40-40)^2}{40}$$

$$= 10 + 10 + 3.33 + 3.33 + 0 + 0 = 26.66, \text{ and}$$

$$d.f. = (2-1)(3-1) = 2$$

(v) The critical region is $\chi^2 \geq \chi^2_{0.01,(2)} = 9.21$

(vi) **Conclusion.** Since the computed value of χ^2 falls in the critical region, we therefore reject H_0 and conclude that home ownership is associated with the family income level.

Q.17.45. (i) We state our hypotheses as

H_0: The drug is no better than sugar pills for curing colds, and

H_1: The drug is better than sugar pills for curing colds.

(ii) We use a significance level of $\alpha = 0.05$, and a one-tailed test.

(iii) The test-statistic is $\chi^2 = \sum_i \sum_j (o_{ij}-e_{ij})^2/e_{ij}$, which is approximated by a chi-square distribution with $(r-1)(c-1)$ degrees of freedom.

(iv) Computations. Under H_0, we determine the expected frequencies as

$$e_{11} = \frac{96\times82}{164} = 48, \; e_{12} = \frac{22\times82}{164} = 11, \; e_{13} = \frac{46\times82}{164} = 23,$$

$e_{21}=48$, $e_{22}=11$, and $e_{23}=23$. The χ^2-value is obtained as

	o_{ij}	e_{ij}	$o_{ij}-e_{ij}$	$(o_{ij}-e_{ij})^2$	$(o_{ij}-e_{ij})^2/e_{ij}$
	52	48	4	16	0.33
	44	48	−4	16	0.33
	10	11	−1	1	0.09
	12	11	1	1	0.09
	20	23	−3	9	0.39
	26	23	3	9	0.39
Σ	164	164	--	--	$\chi^2=1.62$

(v) The critical region is $\chi^2 \geq \chi^2_{0.05,(2)} = 5.99$

(vi) Since the computed value of χ^2 does not fall in the critical region, so the null hypothesis that the drug is no better than sugar pills for curing colds, is not rejected.

Q.17.46. (i) The hypotheses are stated as

H_0: There is no association between total income and televisoin ownership, and

H_1: There is association between total income and television ownership.

(ii) We use a level of significance of $\alpha = 0.05$.

(iii) The test-statistic is

$$\chi^2 = \sum_i \sum_j \frac{(o_{ij}-e_{ij})^2}{e_{ij}},$$

which, if H_0 is true, has an approximate χ^2-distribution with $(3-1)(3-1) = 4$ *d.f.*

(iv) Computations. Applying the method given by J.Skory, we calculate χ^2 as below:

(a) $$T_1 = \frac{(55)^2}{200} + \frac{(118)^2}{700} + \frac{(26)^2}{100} = 41.776,$$

$$T_2 = \frac{(51)^2}{200} + \frac{(207)^2}{700} + \frac{(42)^2}{100} = 91.858,$$

$$T_3 = \frac{(93)^2}{200} + \frac{(375)^2}{700} + \frac{(32)^2}{100} = 254.378;$$

(b) $$R = \frac{41.776}{200} + \frac{91.858}{300} + \frac{254.378}{500} = 1.0238$$

(c) $\therefore \quad \chi^2 = 1{,}000\,(1.0238-1) = 23.8.$

(v) The critical region is $\chi^2 \geq \chi^2_{0.05,(4)} = 9.49.$

(vi) **Conclusion.** Since the calculated value of χ^2 falls in the critical region, we therefore reject H_0 and conclude that the data provide evidence of a statistical association between total income and television ownership.

Q.17.47. Computation of chi-square value. Assuming that the two variables of classification are independent, we find the expected frequencies as below:

$$e_{11} = \frac{350\times70}{1000} = 24.5,\ e_{12} = \frac{370\times70}{1000} = 25.9,\ e_{13} = \frac{280\times70}{1000} = 19.6,$$

$$e_{21} = \frac{350\times700}{1000} = 245,\ e_{22} = \frac{370\times700}{1000} = 259,\ e_{23} = \frac{280\times700}{1000} = 196,$$

$$e_{31} = \frac{350\times230}{1000} = 80.5,\ e_{32} = \frac{370\times230}{1000} = 85.1,\ e_{33} = \frac{280\times230}{1000} = 64.4$$

Table of observed and expected frequencies (in parentheses).

Attributes	A_1	A_2	A_3	Total
B_1	44 (24.5)	22 (25.9)	4 (19.6)	70
B_2	265 (245)	257 (259)	178 (196)	700
B_3	41 (80.5)	91 (85.1)	98 (64.4)	230
Total	350	370	280	1,000

The value of χ^2 is computed by the formula $\chi^2 = \sum\sum (o_{ij}-e_{ij})^2/e_{ij}$

o_{ij}	e_{ij}	$o_{ij}-e_{ij}$	$(o_{ij}-e_{ij})^2$	$(o_{ij}-e_{ij})^2/e_{ij}$
44	24.5	19.5	380.25	15.52
265	245.0	20.0	400.00	1.63
41	80.5	−39.5	1560.25	19.38
22	25.9	−3.9	15.21	0.59
257	259.0	−2.0	4.00	0.02
91	85.1	5.9	34.81	0.41
4	19.6	−15.6	243.36	12.42
178	196.0	−18.0	324.00	1.65
98	64.4	33.6	1128.96	17.53
Σ 1,000	1,000	--	--	$\chi^2 = 69.15$

Since the computed value of $\chi^2 = 69.15$ exceeds the critical value of $\chi^2_{0.01,(4)} = 13.28$, we therefore reject the null hypothesis of independence.

Q.17.48. We state our hypotheses as

H_0 : Intelligence is independent of family income level, and

H_1 : Intelligence is not independent of family income level.

(ii) The significance level is set at $\alpha = 0.01$.

(iii) The test-statistic to use is

$$\chi^2 = \sum \frac{(o_i-e_i)^2}{e_i},$$

which is approximately a chi-square distribution with $(r-1)\times(c-1)$ *d.f.*

(iv) Computations. Applying shorter method (given by Skory), we get

(a) $$T_1 = \frac{(81)^2}{636} + \frac{(141)^2}{751} + \frac{(127)^2}{338} = 84.5076$$

$$T_2 = \frac{(322)^2}{636} + \frac{(457)^2}{751} + \frac{(163)^2}{338} = 519.7262$$

$$T_3 = \frac{(233)^2}{636} + \frac{(153)^2}{751} + \frac{(48)^2}{338} = 123.3471$$

(b) $$R = \frac{84.5076}{349} + \frac{519.7262}{942} + \frac{123.3471}{434} = 1.078$$

(c) $\therefore \quad \chi^2 = 1725\,(1.078 - 1) = 134.55$, and

$d.f. = (3-1)\,(3-1) = 4.$

(v) The critical region is $\chi^2 > \chi^2_{0.01,(4)} = 13.28.$

(vi) **Conclusion.** Since the computed value of $\chi^2 = 134.55$ falls in the critical region, we therefore reject H_0 and conclude that intelligence and family income level are associated.

Q.17.49. (i) **We state our hypotheses as**

H_0: The claim status is independent of the policyholder's age, and

H_1: The claim status and age are not independent.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic under H_0 is

$$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i},$$

which is approximately a chi-square distribution with $(r-1)\times(c-1)$ degrees of freedom.

(iv) Computations. Under H_0, we find the expected frequencies (e_i's).

Age →	Under 25,	25–40,	40–55,	over 55	Total
Reported claim	93	72	53	63	281
No claim	115	155	265	184	719
Total	208	227	318	247	1000

$$e_{11} = \frac{(208)\,(281)}{1000} = 58.44, e_{12} = \frac{(227)\,(281)}{1000} = 63.78$$

$$e_{13} = \frac{(318)\,(281)}{1000} = 89.35, e_{14} = \frac{(227)\,(281)}{1000} = 69.40$$

$$e_{21} = \frac{(208)\,(719)}{1000} = 149.55, e_{22} = \frac{(227)\,(719)}{1000} = 163.21,$$

$$e_{23} = \frac{(318)\,(719)}{1000} = 228.64, e_{24} = \frac{(247)\,(719)}{1000} = 177.59.$$

Now $$\chi^2 = \frac{(93-58.44)^2}{58.44} + \frac{(72-63.78)^2}{63.78} + \ldots + \frac{(184-177.59)^2}{177.59}$$

$= 51.28$ and $d.f. = (2-1)\,(4-1) = 3$

(v) The critical region is $\chi^2 \geq \chi^2_{0.05,(3)} = 7.82$

(vi) **Conclusion.** Since the computed value of $\chi^2 = 51.28$ falls in the critical region, we therefore reject H_0 and conclude that age and claim status are not independent.

Q.17.50. (i) The hypotheses are stated as

H_0: The size of a family is independent of the level of education attained by the father, and

H_1: The size of a family is associated with the level of education attained.

(ii) The level of significance is set at $\alpha = 0.05$.

(iii) The test-statistic to use is

$$\chi^2 = \sum_i \sum_j \frac{(o_{ij} - e_{ij})^2}{e_{ij}},$$

which, if H_0 is true, has an approximate χ^2-distribution with $(3-1)\,(3-1) = 4$ $d.f.$

(iv) Computations. Under H_0, we calculate the expected frequencies by the formula

$$e_{ij} = \frac{(i\text{th Row Total})\,(j\text{th Column Total})}{\text{Total number of observations}}$$

Table of observed and expected frequencies (in parentheses) is:

Education	Number of children			Total
	0 – 1	2 – 3	Over 3	
Elementary	14 (18.675)	37 (39.84)	32 (24.485)	83
Secondary	19 (17.55)	42 (37.44)	17 (23.01)	78
College	12 (8.775)	17 (18.72)	10 (11.505)	39
Total	45	96	59	200

$$\therefore \quad \chi^2 = \frac{(14-18.675)^2}{18.675} + \frac{(19-17.55)^2}{17.55} + \ldots + \frac{(10-11.505)^2}{11.505}$$

$$= 1.1703 + 0.1198 + 1.1853 + 0.2024 + 0.5554 + 0.1580 + 2.3065 + 1.5698 + 0.1969 = 7.4644$$

(v) The critical region is $\chi^2 \geq \chi^2_{0.05,(4)} = 9.49$

(vi) **Conclusion.** Since the computed value of χ^2 does not fall in the critical region, we therefore cannot reject H_0, and may conclude that the size of a family is independent of the level of education attained by the father.

Q.17.52. Computation of the co-efficient of contingency.

	Fair	Grey	Brown	Total
Blue	69	49	28	146
Black	91	56	27	174
Dark Blue	57	34	33	124
Total	217	139	88	444

First, we calculate χ^2 by the method of Skory.

(i) $$T_1 = \frac{(69)^2}{146} + \frac{(91)^2}{174} + \frac{(57)^2}{124} = 106.4032$$

$$T_2 = \frac{(49)^2}{146} + \frac{(56)^2}{174} + \frac{(34)^2}{124} = 43.7908$$

$$T_3 = \frac{(28)^2}{146} + \frac{(27)^2}{174} + \frac{(33)^2}{124} = 18.3418$$

(ii) $$R = \frac{106.4032}{217} + \frac{43.7908}{139} + \frac{18.3418}{88} = 1.0138$$

(iii) $\chi^2 = 444\,(1.0138 - 1) = 6.13.$

Then $C = \sqrt{\dfrac{\chi^2}{n + \chi^2}} = \sqrt{\dfrac{6.13}{444 + 6.13}} = 0.12$

Q.17.53. (b) (i) We state our hypotheses as

H_0 : The two classifications are independent or the treatment does not give any protection against infection, and

H_1 : The two classifications are not independent, or the treatment gives any protection against infection.

(ii) We use a significance level of $\alpha = 0.05$, and a one-tailed test.

(iii) The test-statistic is $\chi^2 = \sum(o_{ij} - e_{ij})^2/e_{ij}$, which for a 2×2 table becomes

$$\chi^2 = \frac{n\,(ad - bc)^2}{(a+b)(c+d)(b+d)(a+c)}, \text{ without Yates' correction, and}$$

$$\chi^2 = \frac{n\left(|ad-bc| - \dfrac{n}{2}\right)^2}{(a+b)(c+d)(b+d)(a+c)}, \text{ with Yates' correction.}$$

where $n = a + b + c + d$.

This statistic has an approximate chi-square distribution with 1 *d.f.*

(iv) Computations. The given data are presented in the following table:

	Infected	Not Infected	Total
Immunized	15	50	65
Not Immunized	95	160	255
Total	110	210	320

$$\therefore \quad \chi^2 \text{ (without Yates' correction)} = \frac{320\,(15\times160 - 50\times95)^2}{110\times210\times65\times255}$$

$$= \frac{706880}{153153} = 4.62, \text{ and}$$

$$\chi^2 \text{ (with Yates' correction)} = \frac{320\,[|15\times160-50\times95| - 320/2]^2}{110\times210\times65\times255}$$

$$= \frac{320\,(2350-160)^2}{110\times210\times65\times255} = \frac{341056}{85085} = 4.01,$$

(v) The critical region is $\chi^2 \geq \chi^2_{0.05,(1)} = 3.84$,

(vi) **Conclusion.** Since the computed value in both the cases falls in the critical region, so we can safely reject H_0. Hence there is evidence to indicate that treatment gives protection against infection.

Q.17.54. (a) Given a modified contingency table as below:

	A_1	A_2	Total
B_1	$a+\frac{1}{2}$	$a-\frac{1}{2}$	$a+b$
B_2	$c-\frac{1}{2}$	$d+\frac{1}{2}$	$c+d$
Total	$a+c$	$b+d$	n

Substituting these values in the short-cut method of computing χ^2 for a 2×2 table, viz., $\chi^2 = \frac{n\,(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)}$, we get

$$\chi^2 = \frac{n\,[(a+\frac{1}{2})(d+\frac{1}{2})-(b-\frac{1}{2})\,(c-\frac{1}{2})]^2}{(a+c)(b+d)(a+b)(c+d)}, \quad (\because\ n=a+b+c+d)$$

$$= \frac{n\,[ad-bc+\frac{1}{2}\,(a+b+c+d)]^2}{(a+c)(b+d)(a+b)(c+d)} = \frac{n\,[ad-bc+\frac{n}{2}]^2}{(a+c)(b+d)(a+b)(c+d)}$$

Since $ad < bc$, the expression $[ad-bc+\frac{n}{2}]^2$ is therefore equivalent to $[\,|ad-bc| - \frac{n}{2}]^2$.

Hence $\chi^2 = \frac{n\,(|ad-bc| - \frac{n}{2})^2}{(a+c)(b+d)(a+b)(c+d)}$

(b) (i) We state our hypotheses as

H_0 : A person's sex and time watching television are independent, and

H_1 : A person's sex and time watching television are not independent.

(ii) The significance level is set at $\alpha = 0.01$

(iii) The test-statistic to use is

$$\chi^2 = \frac{n\,(|ad-bc| - \frac{n}{2})^2}{(a+c)(b+d)(a+b)(c+d)}$$

as the cell frequencies are small.

(iv) Computations. Substituting the values, we get

$$\chi^2 = \frac{30\,(|5\times7-9\times9| - 30/2)^2}{14\times16\times14\times16}$$

$$= \frac{30\,(46-15)^2}{14\times16\times14\times16} = \frac{28830}{50176} = 0.57$$

(v) The critical region is $\chi^2 \geq \chi^2_{0.01,(1)} = 6.64$

(vi) **Conclusion.** Since the computed value of $\chi^2 = 0.57$ does not fall in the critical region, we therefore accept H_0.

Q.17.55. (b) (i) H_0 : The two variables of classification are independent.

(ii) The significance level is $\alpha = 0.05$.

(iii) We calculate the exact probabilities as below:

Toxity Present	Tumer Regression		Total
	Yes	No	
Yes	5	2	7
No	1	7	8
Total	6	9	15

The smallest frequency is $d=1$. Therefore probability for $d=1$, denoted p_1, is

$$p_1 = \frac{7!\,8!\,6!\,9!}{5!\,1!\,2!\,7!\,15!} = 0.0335$$

As the range of variation of d is from 0 to 1, therefore the other possible 2×2 table is

6	1	7
0	8	8
6	9	15

The probability of this table for $d = 0$ is

$$p_0 = \frac{7!\,8!\,6!\,9!}{6!\,0!\,1!\,9!\,15!} = 0.0014$$

$\therefore$ Total probability $= p_0 + p_1 = 0.0014 + 0.0335 = 0.0349$, and

$2P = 2(0.0349) = 0.0698$, which is greater then 0.05.

Hence we reject H_0.

(c) We calculate the exact probability by using the Fisher-Irwin exact test to test the hypothesis of independence.

Deaths	Ward		Total
	A	*B*	
Yes	2	6	8
No	18	14	32
Total	20	20	40

The smallest frequency is $d=2$. Therefore probability for $d=2$, denoted p_2, is

$$p_2 = \frac{8!\,32!\,20!\,20!}{2!\,6!\,18!\,14!\,40!} = \frac{1520}{15873} = 0.09576,$$

As the range of variation of d is from 0 to 2, therefore the other two possible 2×2 tables are

1	7	8
19	13	32
20	20	40

and

0	8	8
20	12	32
20	[illegible]	40

Thus the probabilities of these tables for $d=1$ and $d=0$ are

$$p_1 = \frac{8!\,32!\,20!\,20!}{1!\,7!\,19!\,13!\,40!} = \frac{320}{15873} = 0.02016, \text{ and}$$

$$p_0 = \frac{8!\,32!\,20!\,20!}{0!\,8!\,20!\,12!\,40!} = \frac{2}{1221} = 0.00164$$

Total probability $P = 0.09576 + 0.02016 + 0.00164 = 0.11756$

$\therefore$ $2P = 2(0.11756) = 0.235$, which is not negligible.

Hence we reject the hypothesis of independence.

Q.17.56. (a) (i) The hypotheses are stated as

H_0 : The brittleness of polyethylene bars does not vary with the two heat treatments, and

H_1 : The brittleness of polyethylene bars *varies* with the two heat treatments.

(ii) The level of significance is set at $\alpha = 0.05$

(iii) The test-statistic is Fisher's exact test for a 2×2 contingency table.

(iv) Computations. Given

	Brittle	Tough	Total
Treatment 1	2	8	10
Treatment 2	6	3	9
Total	8	11	19

The smallest frequency is $d=2$. Therefore probability for $d=2$, denoted p_2, is

$$p_2 = \frac{10!\,9!\,8!\,11!}{2!\,8!\,6!\,3!\,19!} = \frac{35}{4199} = 0.008335,$$

As the range of variation of d is from 0 to 2, therefore the other two possible 2×2 tables are

1	9	10
7	2	9
8	11	19

and

0	10	10
8	1	9
8	11	19

Thus the probabilities of these tables for $d=1$ and $d=0$ are

$$p_1 = \frac{10!\,9!\,8!\,11!}{1!\,9!\,7!\,2!\,19!} = \frac{20}{4199} = 0.004763, \text{ and}$$

$$p_0 = \frac{10!\,9!\,8!\,11!}{0!\,10!\,8!\,1!\,19!} = \frac{1}{8398} = 0.000119.$$

The total porbability $P = 0.008335 + 0.004763 + 0.000119$

$= 0.013217$

$\therefore$ $2P = 2(0.013217) = 0.0264$ which is less than 0.05.

(iv) **Conclusion.** Hence we cannot reject H_0 and we may conclude that the brittleness of polyethylene bars does not vary with the two heat treatments.

(b) (i) H_0: Inoculation is independent of immunity from attack.

(ii) The significance level is $\alpha = 0.05$.

(iii) We calculate the exact probabilites as below:

	Inoculated	Not-inoculated	Total
Attacked	9	2	11
Not attacked	7	6	13
Total	16	8	24

The smallest frequency is $d=2$. Therefore probability for $d=2$, denoted p_2, is

$$p_2 = \frac{11!\,13!\,16!\,8!}{9!\,2!\,7!\,6!\,24!} = 0.1284$$

As the range of variation of d is from 0 to 2, therefore the other two possible 2×2 tables are

10	1	11
6	7	13
16	8	24

and

11	0	11
5	8	13
16	8	24

Thus the probabilities of these tables for $d=1$ and $d=0$ are

$$p_1 = \frac{11!\,13!\,16!\,8!}{10!\,1!\,6!\,7!\,24!} = 0.0256, \text{ and}$$

$$p_0 = \frac{11!\,13!\,16!\,8!}{11!\,0!\,5!\,8!\,24!} = 0.0017$$

The total porbability $P = 0.1284 + 0.0256 + 0.0017$

$= 0.1557$

$\therefore$ $2P = 2(0.1557) = 0.3114$, which is not negligible.

Hence we reject the hypothesis of independence.

Q.17.57. (a) (i) The hypotheses are stated as

H_0: The proportion of defectives is the same for all three shifts, and

H_1: The proportion of defectives is *not* the same for all three shifts.

(ii) The level of signficance is set at $\alpha = 0.025$

(iii) The test-statistic is

$$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i},$$

the summation being over all the $2k$ cells of the table, and which, if H_0 is true, has an approximate χ^2-distribution with (3–1), *i.e.* 2 *d.f.*

(iv) Computations. We calculate the expected frequencies under H_0, for each cell by the formula

$$e_{ij} = \frac{\text{(Row Total) (Column Total)}}{n},$$

Table of observed and expected frequencies (in parentheses) is:

Shift →	Day	Evening	Night	Total
Defectives	45 (56.97)	55 (56.67)	70 (56.37)	170
Non-defectives	905 (893.03)	890 (888.33)	870 (883.67)	2665
Total	950	945	910	2835

$$\therefore \quad \chi^2 = \frac{(45-56.97)^2}{56.97} + \frac{(55-56.97)^2}{56.67} + \frac{(70-56.37)^2}{56.37} + \frac{(905-893.03)^2}{893.03} + \frac{(890-888.33)^2}{888.33} + \frac{(870-883.63)^2}{883.63}$$

$$= 2.5150 + 0.0492 + 3.2957 + 0.1604 + 0.0031 + 0.2102$$

$$= 6.2336$$

(v) The critical region is $\chi^2 \geq \chi^2_{0.025,(2)} = 7.38$

(vi) **Conclusion.** Since the calculated value of χ^2 does not fall in the critical region, we therefore accept H_0 and conclude that the proportion of defectives is the same for all three shigts.

(b) (i) We state our null hypothesis as

$H_0 : p$ (proportion of success) $= \frac{4}{5}$ or pass : failure = 4 : 1

(ii) We use a significance level of $\alpha = 0.05$.

(iii) The test-statistic is $\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i}$, which is approximated by a chi-square distribution with $(k-1)$ *d.f.*

(iv) Computations. Under H_0, the expected number of pass in each class is 40 and that of failure is 10. The expected numbers are given within the parentheses in the following table:

Class	Pass		Failure		$\frac{(o_i-e_i)^2}{e_i}$	$\frac{(o_i-e_i)^2}{e_i}$
	o_i	(e_i)	o_i	(e_i)		
1	42	(40)	8	(10)	4/40 = 0.100	4/10 = 0.4
2	45	(40)	5	(10)	25/40 = 0.625	25/10 = 2.5
3	43	(40)	7	(10)	9/40 = 0.225	9/10 = 0.9
4	45	(40)	5	(10)	25/40 = 0.625	25/10 = 2.5
5	45	(40)	5	(10)	25/40 = 0.625	25/10 = 2.5
Σ	220	(200)	30	(50)	2.200	8.8

$\therefore \quad \chi^2 = 2.2 + 8.8 = 11.0$, and

d.f. = 5−1 = 4, as we consider the 5 classes as 5 samples.

(v) The critical region is $\chi^2 \geq \chi^2_{0.05,(4)} = 9.49$.

(vi) **Conclusion.** Since the computed value falls in the critical region, so we reject H_0.

Note: In this problem, the χ^2 value may also be computed for each class (sample) and test of significance applied, *e.g.* for class I, we have

$$\chi^2 = \frac{(42-40)^2}{40} + \frac{(8-10)^2}{10} = 0.5$$

without continuity correction,

or $$\chi^2 = \frac{(|42-40| - \frac{1}{2})^2}{40} + \frac{(|8-10| - 0.5)^2}{10}$$

$= 0.281$, with continuity correction.

This individual value is not significant as the critical region is $\chi^2 \geq \chi^2_{0.05,(1)} = 3.841$. Thus the null hypothesis is supported in respect of class I, etc.

Q.17.58. Computations of χ^2 value by the Brandt-Snedecor formula.

No. of Rows of Kernels	Present (a_i)	Absent	Total (c_i)	$\frac{a_i^2}{c_i}$
8	18	15	33	9.82
10	37	26	63	21.73
12	27	43	70	10.41
14	0	4	4	0
Total	82 (A)	88 (B)	170 (N)	41.96

$$\therefore \quad \chi^2 = \frac{N^2}{AB}\left[\sum_i \frac{a_i^2}{c_i} - \frac{A^2}{N}\right]$$

$$= \frac{(170)^2}{82 \times 88}\left[41.96 - \frac{(82)^2}{70}\right]$$

$$= \frac{28900}{7216}(41.96 - 39.55) = 4.005 \times 2.41 = 9.65.$$

Q.17.59. (b) (i) We state our hypotheses as

H_0 : The two groups are homogeneous, and

H_1 : The two groups are not homogeneous.

(ii) Let the significance le /el be set at $\alpha = 0.05$.

(iii) The test-statistic to use is the Brandt-Snedecor formula, *i.e.*,

$$\chi^2 = \frac{N^2}{AB}\left[\sum \frac{a_i^2}{c_i} - \frac{A^2}{N}\right]$$

which has an approximate χ^2-distribution with $(n-1)$ *d.f.*

(iv) Computations. We calculate the value of χ^2 as follows:

Score	0-9,	10-19,	20-29,	30-39,	40-49,	50-59,	60-69,	70-79,	80-89,	90-99,	Total
a_i	71	68	66	47	51	39	43	39	33	18	475 = A
b_i	22	8	14	12	3	13	3	14	12	10	111 = B
c_i	93	76	80	59	54	52	46	53	45	28	586 = N

Then $\chi^2 = \frac{(586)^2}{(457)\,(111)}\left[\frac{(71)^2}{93} + \frac{(68)^2}{76} + \ldots + \frac{(18)^2}{28} - \frac{(475)^2}{586}\right]$

$= (6.51)\,(4.293) = 27.95$

and $d.f. = 10 - 1 = 9$.

(v) The critical region is $\chi^2 \geq \chi^2_{0.05,(9)} = 16.92$

(vi) **Conclusion.** Since the computed value of $\chi^2 = 27.95$ falls in the critical region, we therefore reject H_0 and conclude that there is significant difference in college aptitude test between the groups.

Q.17.60. (i) We state our hypotheses as

H_0 : There is no significant difference in achievements of the two groups, and

H_1 : There is a significant difference in achievements of the two groups.

(ii) We use a level of significance of $\alpha = 0.05$.

(iii) The test-statistic to use is the Brandt-Snedecor formula, *i.e.*

$$\chi^2 = \frac{N^2}{AB}\left[\sum \frac{a_i^2}{c_i} - \frac{A^2}{N}\right]$$

which has an approximate χ^2-distribution with $(n-1)$ *d.f.*

(iv) Computations. We calculate the value of χ^2 as follows:

Score	40-49,	50-59,	60-69,	70-79,	80-89,	90-99,	Total
Teacher instructed (a_i)	21	40	55	38	10	2	166 = A
Machine instructed (b_i)	18	35	42	46	19	4	164 = B
Total: c_i	39	75	97	84	29	6	330 = N

$\therefore \quad \chi^2 = \frac{(330)^2}{(166)\,(164)}\left[\frac{(21)^2}{39} + \frac{(40)^2}{75} + \ldots + \frac{(2)^2}{6} - \frac{(166)^2}{330}\right]$

$= (4.0001)\,(11.3077 + 21.3333 + 31.1856 + 17.1905 + 3.4483 + 0.6667 - 83.5030)$

$= (4.0001)\,(1.6291) = 6.52$

and $d.f. = 6 - 1 = 5$.

(v) The critical region is $\chi^2 \geq \chi^2_{0.05,(5)} = 11.07$.

(vi) **Conclusion.** Since the computed value of χ^2 does not fall in the critical region, we therefore accept H_0 and conclude that there is no significant difference in achievements of the two groups.

CHAPTER 18

THE STUDENT'S t-DISTRIBUTION AND STATISTICAL INFERENCE

Q.18.1 (b) Definition of "Student's t". Let Z be a normally distributed variable with zero mean and unit variance, and U be an independent variable having a χ^2 distribution with v degrees of freedom. Then the ratio

$$t = \frac{Z}{\sqrt{U/v}},$$

is usually called the "Student's t". This t is a random variable. It is also referred to a test-statistic.

The probability density function of t is

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

which is called Student's distribution with v degrees of freedom.

Assumptions:

(i) The sample consisting of n observations, $X_1, X_2, \ldots, X_n$ is randomly selected. Random sampling is essential for the validity of the t-test.

(ii) The population of observations is normally distributed. It has, however, been shown that slight departures from normality do not seriously affect the test.

(iii) In case of two samples chosen randomly from two normal populations, the populations must have the same variances.

Use and Importance. The Student's t (or the t-distribution) being independent of population variance, is

extremely important as it makes possible the drawing of inferences wholly from sample data. That is why we use the Student's t-statistic when the sample size is small ($n < 30$) and the population variance is unknown. The t-statistic is therefore used

(i) to test the hypothesis that the population mean μ is equal to a specific constant μ_0 or zero.

(ii) to test the hypothesis that the means of two normal populations differ significantly from each other,

(iii) to test the significance of the regression coefficient, correlation coefficient, etc.

(iv) to obtain various confidence intervals such as for population mean, μ, for the difference between two population means, $\mu_1 - \mu_2$, for regression coefficient, B, etc.

To explain its application to a sample mean, let us consider a sample $x_1, x_2, \ldots, x_n$ from a $N(\mu, \sigma^2)$ where σ^2 is unknown. We know that the distribution of $\bar{x}$ is $N\left(\mu, \frac{\sigma^2}{n}\right)$. As variance is unknown its unbiased estimate s^2 is calculated from the sample values by the relation $s^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2$. The quantity s^2 is a random variable such that $\frac{(n-1)\, s^2}{\sigma^2}$ has a χ^2-distribution with $n-1$ degrees of freedom. It has been observed that $\bar{x}$ and s^2, though computed from the same sample values, are independently distributed. Thus

$$t = \frac{Z}{\sqrt{U/v}} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \frac{1}{\sqrt{s^2/\sigma^2}} = \frac{\bar{x} - \mu}{s / \sqrt{n}}.$$

which has "Student's" t-distribution with $n-1$ degrees of freedom and hence has a great practical use.

Q.18.3. (c) The Student's t-distribution with n d.f. is given by the relation

$$f_n(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{t^2}{n}\right)^{-(n+1)/2} \qquad -\infty < t < \infty$$

The factor $\left(1+\frac{t^2}{n}\right)^{-(n+1)/2}$ may be expresed as

$$\left(1+\frac{t^2}{n}\right)^{-(n+1)/2} = \left\{\left(1+\frac{t^2}{n}\right)^{(n/t^2)}\right\}^{-(t^2/2)}\left(1+\frac{t^2}{n}\right)^{-1/2}.$$

Now, as $n \to \infty$, the expression on R.H.S. $\to e^{-t^2/2}$

To simplify the constant term $\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}$, we make use of

Stirling's formula which for large p, is $\Gamma(p) \cong \sqrt{2\pi}\ p^{p+1/2} \cdot e^{-p}$.

This relation implies that $\frac{\Gamma(p+h)}{\Gamma(p)} \cong p^h$.

Thus $\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}$ tends to $\frac{1}{\sqrt{n\pi}}\left(\frac{n}{2}\right)^{1/2} = \frac{1}{\sqrt{2\pi}}$

Hence, when $n \to \infty$, $f_n(t) \to \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$, which is normal $(o,1)$.

Q.18.4. Now $t = \frac{\bar{X}-\mu}{s/\sqrt{n}}$, where $\mu = 5$.

(i) If $n = 25$, $\bar{x} = 3$ and $s = 2$, then

$$t = \frac{3-5}{2/\sqrt{25}} = \frac{-2}{2/5} = -5.$$

(ii) From $t = \frac{\bar{X}-\mu}{s/\sqrt{n}}$, we get $s = \frac{(\bar{X}-\mu)\sqrt{n}}{t}$

Substitution gives

$$s = \frac{(2-5)\sqrt{9}}{-2} = \frac{(-3)(3)}{-2} = 4.5$$

(iii) From $t = \frac{\bar{X}-\mu}{s/\sqrt{n}}$, we get $\bar{X} = \mu + t \cdot \frac{s}{\sqrt{n}}$

Putting the given values, we obtain

$$\bar{x} = 5 + 2 \cdot \frac{10}{\sqrt{25}} = 5 + 4 = 9$$

(iv) The formula $t = \frac{\bar{X}-\mu}{s/\sqrt{n}}$ gives $\sqrt{n} = \frac{ts}{(\bar{X}-\mu)}$

Substituting the values, we get

$$\sqrt{n} = \frac{3 \times 15}{14-5} = 5, \text{ giving } n = 25.$$

Q.18.5 (b) The 95% confidence limits for the mean breaking strength, μ, would be

$$\bar{x} \pm t_{0.025,(n-1)} \cdot \frac{s}{\sqrt{n}},$$

where $\bar{x} = \frac{1}{n}\Sigma x_i$ and $s^2 = \frac{1}{n-1}\Sigma(x_i-\bar{x})^2 = \frac{1}{n-1}\left[\Sigma D^2 - \frac{(\Sigma D)^2}{n}\right]$

Computation of mean and standard deviation.

	x	$D(=x-570)$	D^2
	578	8	64
	572	2	4
	570	0	0
	568	−2	4
	572	2	4
	570	0	0
	570	0	0
	572	2	4
	596	26	676
	584	14	196
Σ	5752	52	952

$$\bar{x} = a + \frac{\Sigma D}{n} = 570 + \frac{52}{10} = 575.2$$

$$s^2 = \frac{1}{n-1}\left[\Sigma D^2 - \frac{(\Sigma D)^2}{n}\right] = \frac{1}{9}\left[952 - \frac{(52)^2}{10}\right]$$

$$= \frac{1}{9}[952 - 270.4] = \frac{681.6}{9} = 75.7333, \text{ so that}$$

$s = \sqrt{75.7333} = 8.702$, and $t_{0.025}$ for 9 *d.f.* = 2.262

Substituting the values, we get

$$575.2 \pm 2.262\left(\frac{8.702}{\sqrt{10}}\right)$$

i.e. $575.2 \pm 2.262\ (2.752)$ *i.e.* 575.2 ± 26.2

Hence the required 95% confidence limits for μ are 569 to 581.

Q.18.6 (a) The 90% confidence interval for the population mean, μ, when σ is known, would be

$$\bar{x} \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}.$$

Now $\bar{x} = \frac{\Sigma x_i}{n} = \frac{9 + 14 + \ldots + 12}{8} = \frac{88}{8} = 11.$

Substituting the values, we get

$$11 \pm 1.645\left(\frac{2}{\sqrt{8}}\right), \text{ i.e. } 11 \pm \frac{(1.645)\ (2)}{2.828}, \text{ i.e. } 11 \pm 1.16$$

Thus μ lies between 9.84 and 12.16.

The 90% confidence interval for the population mean, μ, when σ is unknown, would be

$$\bar{x} \pm t_{.05,(8-1)} \cdot \frac{s}{\sqrt{n}}..$$

Now $s^2 = \frac{1}{n-1}\Sigma(x_i - \bar{x})^2 = \frac{(9-11)^2 + (14-11)^2 + \ldots + (12-11)^2}{7}$

$$= \frac{36}{7} = 5.1429, \text{ so that } s = \sqrt{5.1429} = 2.27$$

and $t_{.05(7)} = 1.895.$

Substituting these values, we get

$$11 \pm 1.895\left(\frac{2.27}{2.828}\right), \text{ i.e. } 11 \pm 1.52$$

Hence the 90% confidence interval for μ when σ is unknown, is 9.48 to 12.52.

(b) The 90% confidence interval for μ when $\sigma = 3$ would be

$$\bar{x} \pm 1.645\frac{\sigma}{\sqrt{n}}, \text{ where } n = 4, \text{ and}$$

$$\bar{x} = \frac{2.3 + (-0.2) + (-0.4) + (-0.9)}{4} = \frac{0.8}{4} = 0.2$$

Substituting the values, we get

$0.2 \pm 1.645(3) / \sqrt{4}$ or 0.2 ± 2.47 or -2.27 to 2.67

When σ is unknown, it is estimated from sample values, and the 90% confidence interval would become

$$\bar{x} \pm t_{\alpha/2,(n-1)} \cdot \frac{s}{\sqrt{n}}, \text{ where}$$

$$s = \sqrt{\frac{1}{n-1}\left[\Sigma X^2 - \frac{(\Sigma X)^2}{n}\right]} = \sqrt{\frac{1}{3}\left[6.30 - \frac{0.64}{4}\right]}$$

$$= \sqrt{(6.14) / 3} = \sqrt{2.0467} = 1.43, \text{ and}$$

$t_{0.05,(3)} = 2.353$

Substituting the values, we get the 90% confidence interval for μ as

$$0.2 \pm (2.353)\frac{1.43}{\sqrt{4}}, \text{ i.e. } 0.2 \pm 1.68 \text{ or } -1.48 \text{ to } 1.88.$$

Q.18.7 (a) The 95% confidence limits for the population mean, μ, would be

$$\bar{x} \pm t_{.025(n-1)} \cdot \frac{s}{\sqrt{n}},$$

where $\bar{x} = \frac{1}{n}\sum_i x_i$ and $s = \sqrt{\frac{1}{n-1}\sum(x_i-\bar{x})^2}$

Computation of $\bar{x}$ and s.

	x_i	$D_i(=x_i-100)$	D_i^2
	70	−30	900
	120	20	400
	110	10	100
	101	1	1
	88	−12	144
	83	−17	289
	95	−5	25
	107	7	49
	100	0	0
	98	−2	4
Σ	972	−28	19 2

$$\bar{x} = \frac{\Sigma x}{n} = \frac{972}{10} = 97.2$$

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

$$= \frac{1}{9}\left[1912 - \frac{(-28)^2}{10}\right] = \frac{1}{9}[1912-78.4]$$

$$= \frac{1833.6}{9} = 203.7333$$

$\therefore$ $s = \sqrt{203.7333} = 14.27$, and $t_{.025}$ for 9 *d.f.* = 2.262

Substituting the values, we get

$$97.2 \pm 2.262\left(\frac{14.27}{\sqrt{10}}\right)$$

i.e. $97.2 \pm 2.262\ (4.513)$, *i.e.* 97.2 ± 10.2

Hence the required 95% confidence limits for the populationl mean, μ, are 87.0 to 107.4.

(b) The 90% confidence interval for the mean, μ, is given by

$$\bar{x} \pm t_{\alpha/2,(v)}\frac{s}{\sqrt{n}}.$$

Here $\bar{x}=14.5$, $s=5$, $n=16$, $v = 16-1 = 15$ and $t_{0.05(15)} = 1.75$.
Substituting the values, we get

$$14.5 \pm 1.753\left(\frac{5}{\sqrt{16}}\right)$$

i.e. $14.5 \pm (1.753)\ (1.25)$ *i.e.* 14.5 ± 2.2 or 12.3 to 16.7

Hence the 90% confidence interval for μ calculated from the given information is (12.3, 16.7).

Q.18.8. The 90% confidence interval for the mean mass of the population, μ, is given by

$$\bar{x} \pm t_{\alpha/2,(v)}\frac{s}{\sqrt{n}}.$$

Now $\bar{x} = \frac{\Sigma x_i}{n} = \frac{412.8}{12} = 34.4,$

$$s = \sqrt{\frac{1}{n-1}[\Sigma x^2 - (\Sigma x)^2/n]} = \sqrt{\frac{1}{11}[14231.98-(412.8)^2/12]}$$

$$= \sqrt{\frac{31.66}{11}} = 1.70,$$

$v = 12 - 1 = 11$ and $t_{0.05(11)} = 1.796$

Substituting these values, we get

$$34.4 \pm 1.796\left(\frac{1.70}{\sqrt{12}}\right), \text{ i.e. } 34.4 \pm (1.796)\ (0.49)$$

i.e. 34.4 ± 0.88 or 33.52 to 35.28

Hence the 90% C.I. for the mean mass of the population is (33.52, 35.28).

Q.18.9 (a) The 95% confidence interval for $\mu_1-\mu_2$ is given by

$$(\bar{x}_1-\bar{x}_2) \pm t_{\alpha/2,(v)} \cdot s_p \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

where $s_p^2 = \dfrac{(n_1-1)\,s_1^2 + (n_2-1)\,s_2^2}{n_1+n_2-2}$.

Here $\bar{x}_1 = 64, \bar{x}_2 = 59, s_1 = 6, s_2 = 5, n_1 = 9$ and $n_2 = 16$.

Now $s_p^2 = \dfrac{(n_1-1)\,s_1^2 + (n_2-1)\,s_2^2}{n_1+n_2-2} = \dfrac{(9-1)(5)^2+(16-1)(6)^2}{9+16-2}$

$= \dfrac{663}{23} = 28.8261$, so that $s_p = \sqrt{28.8261} = 5.369$ and

$t_{0.025(23)} = 2.069$

Substituting these values, we get

$$(64-59) \pm (2.069)\,(5.369)\sqrt{\frac{1}{9}+\frac{1}{16}}$$

i.e., $5 \pm (2.069)\,(5.369)\,(0.4167)$

i.e., 5 ± 4.6 or 0.4 to 9.6.

Hence the 95% C.I. for $\mu_1-\mu_2$ is (0.4, 9.6).

(b) The 95% confidence interval for $\mu_1-\mu_2$ is given by

$$(\bar{x}_1-\bar{x}_2) \pm t_{\alpha/2(v)} \cdot s_p \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

Now $\bar{x}_1 = \dfrac{184}{5} = 36.8, \bar{x}_2 = \dfrac{218}{7} = 31.1,$

$$s_p^2 = \frac{1}{n_1+n_2-2}\left[\left(\Sigma x_1^2 - \frac{(\Sigma x_1)^2}{n_1}\right)+\left(\Sigma x_2^2 - \frac{(\Sigma x_2)^2}{n_2}\right)\right]$$

$$= \frac{1}{5+7-2}\left[\left(7024 - \frac{(184)^2}{5}\right)+\left(7158 - \frac{(218)^2}{7}\right)\right]$$

$= \dfrac{1}{10}[252.8 + 368.9] = 62.17$, so that $s_p = 7.88$,

$t_{0.025(10)} = 2.228$

Substituting these values, we get

$$(36.8 - 31.1) \pm (2.228)\,(7.8848)\sqrt{\frac{1}{5}+\frac{1}{7}}$$

i.e., $5.7 \pm (2.228)\,(7.8848)\,(0.5855)$

i.e., $5.7 \pm (2.228)\,(4.62)$

i.e., 5.7 ± 10.3 or -4.6 to 16.0.

Hence the 95% confidence interval for $\mu_1-\mu_2$ is (−4.6, 16.0).

Q.18.10. The 90% confidence interval for the difference between the average grades, *i.e.*, $\mu_2-\mu_1$, is given by

$$(\bar{x}_2-\bar{x}_1) \pm t_{\alpha/2}(v) \cdot s_p \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}.$$

Here $\bar{x}_1 = \dfrac{720}{10} = 72, \bar{x}_2 = \dfrac{1012}{12} = 84.3,$

$$s_p^2 = \frac{1}{n_1+n_2-2}\left[\Sigma(x_1-\bar{x}_1)^2 + \Sigma(x_2-\bar{x}_2)^2\right]$$

$$= \frac{1}{10+12-2}[324.67 + 120.00] = 22.2335$$

$\therefore$ $s_p = \sqrt{22.2335} = 4.72$ and $t_{0.05(20)} = 1.725$.

Substituting these values, we get

$$(84.3-72) \pm (1.725)\,(4.72)\sqrt{\frac{1}{10}+\frac{1}{12}}$$

i.e., $12.3 \pm (1.725)\,(4.72)\,(0.43)$

i.e., 12.3 ± 3.5 or 8.8 to 15.8.

Hence the 90% C.I. for $\mu_2-\mu_1$ is (8.8, 15.8).

Q.18.11. Let X_1 and X_2 denote the aberrant and offtype measurements respectively. Then

$$\Sigma X_{1i} = 80.92,\ \Sigma X_{1i}^2 = 561.6402,\ \Sigma X_{2j} = 84.25$$

and $\Sigma X_{2j}^2 = 478.9779$.

The 90% confidence interval for $\mu_1 - \mu_2$ is given by

$$(\bar{x}_1 - \bar{x}_2) \pm t_{0.05,(v)} \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \text{ where}$$

$$\bar{x}_1 = \frac{\Sigma X_{1i}}{n_1} = \frac{80.92}{12} = 6.74; \; \bar{x}_2 = \frac{\Sigma X_{2j}}{n_2} = \frac{84.25}{15} = 5.62,$$

$$\Sigma(X_{1i} - \bar{X}_1)^2 = \Sigma X_{1i}^2 - (\Sigma X_{1i})^2 / n_1$$

$$= 561.6402 - (80.92)^2/12 = 15.9697,$$

$$\Sigma(X_{2j} - \bar{X}_2)^2 = \Sigma X_{2j}^2 - (\Sigma X_{2j})^2 / n_2$$

$$= 478.9779 - (84.25)^2/15 = 5.7737;$$

$$s_p^2 = \frac{\Sigma(X_{1i} - \bar{X}_1)^2 + \Sigma(X_{2j} - \bar{X}_2)^2}{n_1 + n_2 - 2} = \frac{15.9697 + 5.7739}{12 + 15 - 2}$$

$= 0.8697$, so that $s_p = \sqrt{0.8697} = 0.93$; and $t_{0.05,(25)} = 1.708$.

Substitution gives

$$(6.74 - 5.62) \pm (1.708)\,(0.93) \sqrt{\frac{1}{12} + \frac{1}{15}}$$

or $\quad 1.12 \pm (1.708)\,(0.93)\,(0.3873)$

or $\quad 1.12 \pm 0.62$ or 0.50 to 1.74.

Thus the 90% confidence interval for $\mu_1 - \mu_2$ is (0.50, 1.74)

Q.18.13. For all the samples as the sample sizes are less than 30 and the population variances are unknown, we therefore use t-test for testing the given null hypotheses.

Sample (a). Given $H_0 : \mu = 10$ and $H_1 : \mu > 10$ (one-tailed test)

$$\text{Here} \quad t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} = \frac{12 - 10}{6 / \sqrt{9}} = 1.$$

The critical region is $t \geq t_{0.05(8)} = 1.86$.

Since the computed value of $t = 1$ does not fall in the critical region, we therefore accept H_0.

Sample (b). Given $H_0 : \mu = 10$ and $H_1 : \mu \neq 10$ (two-tailed test)

$$\text{Here} \quad t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} = \frac{13 - 10}{8 / \sqrt{16}} = 1.5$$

The critical region is $|t| \geq t_{0.025(15)} = 2.131$

Since the computed value of $t = 1.5$ does not fall in the critical region, we therefore accept H_0.

Sample (c). Given $H_0 : \mu \leq 10$ and $H_1 : \mu > 10$ (one-tailed test)

$$\text{Here} \quad t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} = \frac{11 - 10}{9 / \sqrt{16}} = 0.44$$

The critical region is $t \geq t_{0.01(15)} = 2.602$

Since the computed value of $t = 0.44$ does not fall in the critical region, we therefore accept H_0.

Sample (d). Given $H_0 : \mu \geq 10$ and $H_1 : \mu < 10$ (one-tailed test)

$$\text{Here} \quad t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} = \frac{8 - 10}{8 / \sqrt{25}} = -1.25$$

The critical region is $t \leq -t_{0.01(24)} = -2.492$

Since the computed value of $t = -1.25$ does not lie in the critical region, we therefore accept H_0.

Sample (e). Given $H_0 : \mu = 10$ and $\mu \neq 10$ (two-tailed test)

$$\text{Here} \quad t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} = \frac{9 - 10}{7 / \sqrt{25}} = -0.714$$

The critical region is $|t| \geq t_{0.01(24)} = 2.92$

Since the computed value of $t = -0.714$ does not fall in the critical region, we therefore accept H_0.

Q.18.14. (i) We state the hypotheses as $H_0 : \mu = 8$ oz and $H_1 : \mu \neq 8$ oz.

(ii) We use a significance level of $\alpha = 0.05$, and a two-sided test.

(iii) The test-statistic would be

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

which under H_0, has Student's t-distribution with $(n-1)$ *d.f.*

(iv) Computation:

$$\bar{x} = \frac{1}{n}(\Sigma x_i) = \frac{91.7}{12} = 7.64$$

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

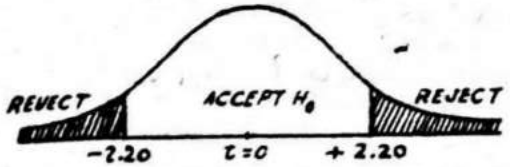
$$= \frac{1}{11}\left[701.61 - \frac{(91.7)^2}{12}\right] = \frac{1}{11}[701.61 - 700.7408]$$

$$= \frac{0.8692}{11} = 0.079018, \text{ so that } s = \sqrt{0.079018} = 0.28$$

Weight (x)	x^2
8.2	67.24
8.0	64.00
7.6	57.76
7.6	57.76
7.7	59.29
7.5	56.25
7.3	53.29
7.4	54.76
7.5	56.25
8.0	64.00
7.4	54.76
7.5	56.25
Σ 91.7	701.61

$$\therefore \quad t = \frac{7.64 - 8.00}{0.28 / \sqrt{12}} = \frac{(-0.36)(3.464)}{0.28} = -4.45, \text{ and } d.f. = 11.$$

(v) The critical region is $t \leq -t_{0.025(11)} = -2.20$ and $t \geq t_{.025(11)} = 2.20$



(vi) The computed value of t falls in the critical region, so we reject H_0. We may conclude that the given values are not consistent with the population mean of 8 oz.

Q.18.15. (a) (i) The hypotheses would be stated as

$$H_0 : \mu = 47.5 \text{ and } H_1 : \mu \neq 47.5$$

(ii) We use a significance level of $\alpha = 0.05$, and a two-sided test.

(iii) The test-statistic would be

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}},$$

which under H_0, has Student's t-distribution with $(n-1)$ *d.f.*

(iv) Computations:

x_i	45, 47, 50, 52, 48, 47, 49, 53, 51	442
D_i	0, 2, 5, 7, 3, 2, 4, 8, 6	37
D_i^2	0, 4, 25, 49, 9, 4, 16, 64, 36	207

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{442}{9} = 49.11$$

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

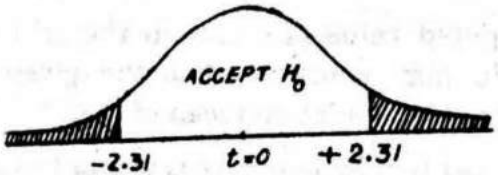
$$= \frac{1}{8}\left[207 - \frac{(37)^2}{9}\right] = \frac{1}{8}[207 - 152.11] = \frac{54.89}{8} = 6.8612$$

$$\therefore \quad s = \sqrt{6.8612} = 2.62$$

Now $t = \dfrac{49.11 - 47.5}{2.62 / \sqrt{9}} = \dfrac{3(1.61)}{2.62} = \dfrac{4.83}{2.62} = 1.84$ and $d.f. = 8$.

(v) The critical region is $t \le -t_{0.25(8)} = -2.31$ and

$$t \ge t_{.025(8)} = 2.31.$$



(vi) The computed value of t does not fall in the critical region, so we cannot reject H_0. We may conclude that the mean of the nine items does not differ from the assumed population mean of 47.5.

(b) (i) The hypotheses would be stated as

$$H_0 : \mu = 16 \text{ oz} \quad \text{and} \quad H_1 : \mu \neq 16 \text{ oz}.$$

(ii) We use a significance level of $\alpha = 0.05$, and a two-sided test.

(iii) The test-statistic under H_0, would be

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}},$$

which has Student's t-distribution with $(n-1)$ *d.f.*

(iv) Computations:

Here $n = 10$, $\bar{x} = 15.90$ oz,

$$s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}} = \sqrt{\frac{0.276}{9}} = \frac{0.525}{3} = 0.175$$

$$\therefore \quad t = \frac{15.90 - 16.00}{0.175 / \sqrt{9}} = \frac{3(-0.10)}{0.175} = -1.71$$

and $d.f. = 10 - 1 = 9$.

(v) The critical region is $t \le -t_{0.025(9)} = -2.262$ and

$$t \ge t_{.025(9)} = 2.262.$$

(vi) Since the computed value of t does not fall in the critical region, so decision would be, "We cannot reject H_0". The sample mean does not differ significantly from the intended weight of 16 oz.

Q.18.16. (i) We state the hypotheses as

$$H_0 : \mu = 43.5 \text{ inches and } H_1 : \mu \neq 43.5 \text{ inches}.$$

(ii) We use a significance level of $\alpha = 0.05$, and a two-sided test.

(iii) The test-statistic under H_0, would be

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}},$$

which under H_0, has a Student's t-distribution with $(n-1)$ *d.f.*

(iv) Computations.

Here $n = 16$, $\bar{x} = 41.5$, and

$$s = \sqrt{\frac{1}{n-1}\Sigma(x_i - \bar{x})^2} = \sqrt{\frac{135}{15}} = \sqrt{9} = 3$$

$$\therefore \quad t = \frac{41.5 - 43.5}{3 / \sqrt{16}} = \frac{(-2.0)\,(4)}{3} = -2.67$$

and $d.f. = 16 - 1 = 15$

(v) The critical region is $t \le -t_{0.025(15)} = -2.13$ and

$$t \ge t_{.025(15)} = 2.13.$$

(vi) Since the computed value falls in the critical region, so we reject H_0. Hence there is evidence to indicate that the

assumption of a mean of 43.5 inches for the population is not reasonable.

The 95% confidence interval for the population mean, μ, would be

$$\bar{x} \pm t_{.025(n-1)} \cdot \frac{s}{\sqrt{n}}$$

..Substituting the values, we obtain

$$41.5 \pm 2.13\left(\frac{3}{\sqrt{16}}\right) \textit{ i.e., } 41.5 \pm 2.13\left(\frac{3}{4}\right) \textit{ i.e. } 41.5 \pm 1.6$$

Hence the 95% confidence limits for the population mean, μ, are 39.9 inches to 43.1 inches.

Q.18.17. (i) The hypotheses would be stated as

$H_0 : \mu = 0.050$ inches and

$H_1 : \mu \neq 0.050$ inches.

(ii) The levels of significance are $\alpha = 0.05$ and $\alpha = 0.01$. We use a two-sided test.

(iii) The test-statistic would be

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}},$$

which under H_0, has a Student's t-distribution with $(n-1)$ *d.f.*

(iv) Computations:

$$t = \frac{0.053 - 0.050}{0.003 / \sqrt{10}} = \frac{(3.162)\ (0.003)}{0.003} = 3.16,$$

and $d.f. = 10 - 1 = 9$.

(v) The critical regions are at 0.05 level

$$t \leq -t_{.025(9)} = -2.262 \text{ and } t \geq t_{.025(9)} = 2.262$$

and at 0.01 level, $t \leq -t_{.005(9)} = -3.25$ and $t \geq t_{.005,(9)} = 3.25$.

(vi) The calculated value of t falls in the critical region at 0.05 level but does not fall in the critical region at 0.01 level of significance. We reject H_0 at 0.05 level but not at 0.01 level. It is advisable to check the machine.

Q.18.18. (i) We formulate our hypotheses as

$H_0 : \mu = 1500$ hours, and $H_1 : \mu \neq 1500$ hours

(ii) Let us choose the significance level of $\alpha = 0.05$.

(iii) The test-statistic to use is

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}},$$

which under H_0 has a t-distribution with $v = 5$ *d.f.*

(iv) The critical region is $|t| \geq t_{0.025(5)} = 2.571$

(v) Computations: Here $\bar{x} = \dfrac{\sum x}{n} = \dfrac{8909}{6} = 1484.8$ hrs.

$$s^2 = \frac{1}{n-1}\sum (x - \bar{x})^2 = \frac{1}{5}[13280481 - 13228380.17]$$

$$= \frac{52100.83}{5} = 10420.17, \text{ so that } s = 102.08 \text{ hrs.}$$

$$\therefore \quad t = \frac{1484.8 - 1500}{102.08 / \sqrt{6}} = \frac{(-15.20)\ (2.45)}{102.08} = -0.36$$

(vi) **Conclusion.** Since the computed value of $t = -0.36$ does not fall in the critical region, we therefore accept H_0. The manufacturer's claim is vindicated.

Q.18.19. (b) (i) We state our hypotheses as

$H_0 : \mu_2 - \mu_1 = 10$ and $H_1 : \mu_2 - \mu_1 > 10$. (one-tailed)

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic under H_0 is

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 10}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

which has a Student's t-distribution with $v = 12+18-2 = 28$ *d.f.*

(iv) The critical region is $t \geq t_{0.05(28)} = 1.701$

(v) Computations. Given $s_1^2 = 1200$, $s_2^2 = 900$, $n_1 = 12$ and $n_2 = 18$.

Now $s_p^2 = \dfrac{(n_1-1)\, s_1^2 + (n_2-1)\, s_2^2}{n_1 + n_2 - 2} = \dfrac{(11)\,(1200)+(17)(900)}{12+18-2}$

$= \dfrac{28500}{28} = 1017.8571$ so that $s_p = \sqrt{1017.8571} = 31.9039$

Substituting the values, we get

$$t = \frac{(25-10) - 10}{31.9039 \sqrt{\frac{1}{18} + \frac{1}{12}}} = \frac{5}{11.89} = 0.42$$

(vi) **Conclusion.** Since the computed value of $t = 0.42$ does not fall in the critical region, we therefore accept H_0.

Q.18.20. (i) Let μ_1 and μ_2 be the mean weights of two populations of males and females respectively. Then we are required to decide between the hypotheses

$H_0 : \mu_1 - \mu_2 = 350$ grams and $H_1 : \mu_1 - \mu_2 > 350$ grams.

(ii) The significance level is set at $\alpha = 0.05$. (one-tailed test)

(iii) The test-statistic under H_0 is

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 350}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

which has a t-distribution with (n_1+n_2-2) *d.f.*

(iv) Computations.

Males (weight in grams)			Females (weight in grams)		
X_{1i}	D_{1i} $(X_{1i}-1450)$	D_{1i}^2	X_{2j}	D_{2j} $(X_{2j}-1000)$	D_{2j}^2
1293	−157	24649	1061	61	3721
1380	−70	4900	1065	65	4225
1614	164	26896	1092	92	8464
1497	47	2209	1017	17	289
1310	−110	12100	1021	21	441
1630	193	37249	1138	138	19044
1466	16	256	1143	143	20449
1627	177	31329	1094	94	8836
1383	−67	4489	1270	270	72900
1711	261	68121	1028	28	784
14,954	454	212,198	10,929	929	139,153

Now $\bar{X}_1 = 1495.4$ grams, $\bar{X}_2 = 1092.9$ grams,

$\Sigma(X_{1i}-\bar{X}_1)^2 = \Sigma D_{1i}^2 - \dfrac{(\Sigma D_{1i})^2}{n_1} = 212{,}198 - 20{,}611.6 = 191{,}586.4,$

$\Sigma(X_{2j}-\bar{X}_2)^2 = \Sigma D_{2j}^2 - \dfrac{(\Sigma D_{2j})^2}{n_2} = 139{,}153 - 85{,}304.1 = 52{,}848.9,$

$\therefore \quad s_p^2 = \dfrac{191{,}586.4 + 52{,}848.9}{10 + 10 - 2} = \dfrac{244{,}435.3}{18} = 13{,}579.7389,$

so that $s_p = \sqrt{13{,}579.7389} = 116.53$

Thus $t = \dfrac{(1495.4 - 1092.9) - 350}{(116.53)\sqrt{\frac{1}{10} + \frac{1}{10}}} = \dfrac{52.5}{(116.53)\,(0.447)}$

$= \dfrac{52.5}{52.09} = 1.008$

(v) The critical region is $t \geq t_{0.05,(18)} = 1.734$

(vi) **Conclusion.** Since the computed value of $t = 1.008$ does not fall in the critical region, we therefore accept H_0 of a difference of 350 grams between population means in favour of males against the alternative of a greater difference.

Q.18.21 (a) (i) The hypotheses are stated as

$H_0 : \mu_1 = \mu_2$, *i.e.* soldiers are not taller than sailors, and

$H_1 : \mu_1 > \mu_2$, *i.e.* soldiers are taller than sailors.

(ii) We use a significance level of $\alpha = 0.05$ and a one-sided test.

(iii) The test-statistic under the null hypothesis $H_0 : \mu_1 = \mu_2$, would be

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

where $$s_p^2 = \frac{1}{n_1 + n_2 - 2} [\Sigma(x_{1i} - \bar{x}_1)^2 + \Sigma(x_{2j} - \bar{x}_2)^2]$$

This statistic has Student's t-distribution with $(n_1 + n_2 - 2)$ *d.f.*

(iv) Computations.

Soldiers			Sailors		
Height (x_{1i})	$(x_{1i} - \bar{x}_1)$	$(x_{1i} - \bar{x}_1)^2$	Height (x_{2j})	$(x_{2j} - \bar{x}_2)$	$(x_{2j} - \bar{x}_2)^2$
61	-6.8	46.24	63	-5	25
62	-5.8	33.64	65	-3	9
65	-2.8	7.84	68	0	0
66	-1.8	3.24	69	1	1
69	1.2	1.44	71	3	9
69	1.2	1.44	72	4	16
70	2.2	4.84	--	--	--
71	3.2	10.24	--	--	--
72	4.2	17.64	--	--	--
73	5.2	27.04	--	--	--
Σ 678	0	153.60	408	0	60

Now $\bar{x}_1 = \frac{678}{10} = 67.8$ inches, $\bar{x}_2 = \frac{408}{6} = 68$ inches,

$$s_p^2 = \frac{\Sigma(x_{1i} - \bar{x}_1)^2 + \Sigma(x_{2j} - \bar{x}_2)^2}{n_1 + n_2 - 2} = \frac{153.60 + 60}{10 + 6 - 2} = \frac{213.60}{14}$$

$= 15.2571$, so that $s_p = \sqrt{15.2571} = 3.906$

$$\therefore \quad t = \frac{67.8 - 68.0}{3.906 \sqrt{\frac{1}{10} + \frac{1}{6}}} = \frac{-0.2}{(3.906)(0.516)} = \frac{-0.2}{2.015} = -0.099$$

and $d.f. = 10 + 6 - 2 = 14$.

(v) The critical region is

$t \geq t_{.05(14)} = 1.76$

(vi) Since the computed value of t does not fall in the critical region, so we cannot reject H_0. The difference in the mean heights is not sufficient to demonstrate that soldiers are on the average taller than sailors.

(b) (i) The hypotheses would be stated as

$H_0 : \mu_1 - \mu_2 = 0$, and electrification does not exert any effect on the tillering and $H_1 : \mu_1 - \mu_2 \neq 0$, and electrification does exert some effect on the tillering.

(ii) We use a significance level of $\alpha = 0.05$, and a two-sided test.

(iii) The test-statistic under H_0, would be

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

where $$s_p^2 = \frac{1}{n_1 + n_2 - 2} [\Sigma(x_{1i} - \bar{x}_1)^2 + \Sigma(x_{2j} - \bar{x}_2)^2]$$

This statistic has Student's distribution with $(n_1 + n_2 - 2)$ *d.f.*

(iv) Computations.

Caged		Electrified	
x_{1i}	x_{1i}^2	x_{2j}	x_{2j}^2
17	289	16	256
27	729	16	256
18	324	20	400
25	625	16	256
27	729	21	441
29	841	17	289
27	729	15	225
23	529	20	400
17	289	--	--
Σ 210	5084	141	2523

Now $\bar{x}_1 = \frac{210}{9} = 23.333, \quad \bar{x}_2 = \frac{141}{8} = 17.625,$

$$\Sigma(x_{1i}-\bar{x}_1)^2 = \Sigma x_{1i}^2 - \frac{(\Sigma x_{1i})^2}{n_1} = 5084 - \frac{(210)^2}{9}$$

$$= 5084 - 4900 = 184,$$

$$\Sigma(x_{2j}-\bar{x}_2)^2 = \Sigma x_{2j}^2 - \frac{(\Sigma x_{2j})^2}{n_2} = 2523 - \frac{(141)^2}{8}$$

$$= 2523 - 2485.125 = 37.875.$$

$$\therefore \quad s_p^2 = \frac{184+37.875}{9+8-2} = \frac{221.875}{15} = 14.7917, \text{ so that}$$

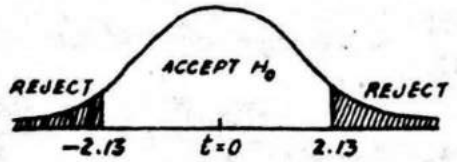
$$s_p = \sqrt{14.7917} = 3.846$$

Thus $$t = \frac{23.333 - 17.625}{3.846\sqrt{\frac{1}{9}+\frac{1}{8}}}$$

$$= \frac{5.708}{(3.846)\,(0.4859)} = \frac{5.708}{1.869} = 3.05$$

and $d.f. = 9 + 8 - 2 = 15.$

(v) The critical region is $|t| \geq t_{.025,(15)} = 2.13$



(vi) Since the computed value of t falls in the critical region, so we reject H_0 in favour of H_1. We may conclude that electrification does exert some effect on the tillering.

Q.18.22. (i) We have to decide between the hypotheses

$$H_0 : \mu_A = \mu_B \quad \text{and} \quad H_1 : \mu_A \neq \mu_B$$

(ii) We use a significance level of $\alpha = 0.05$, and a two-sided test.

(iii) The test-statistic under the null hypothesis $H_0 : \mu_A = \mu_B$, would be

$$t = \frac{\bar{x}_A - \bar{x}_B}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

where $$s_p^2 = \frac{1}{n_1+n_2-2}\left[\Sigma(x_{Ai}-\bar{x}_A)^2 + (\Sigma x_{Bj}-\bar{x}_B)^2\right].$$

This statistic has a t-distribution with (n_1+n_2-2) $d.f.$

(iv) Computations.

Diet A		Diet B	
x_{Ai}	x_{Ai}^2	x_{Bj}	x_{Bj}^2
25	625	44	1936
30	900	34	1156
28	784	22	484
34	1156	8	64
24	576	47	2209
25	625	31	961
13	169	40	1600
32	1024	30	900
24	576	32	1024
30	900	35	1225
31	961	18	324
35	1225	21	441
--	--	35	1225
--	--	29	841
--	--	22	484
Σ 331	9521	448	14874

Now $\bar{x}_A = \frac{331}{12} = 27.58$, $\bar{x}_B = \frac{448}{15} = 29.87$,

$$\Sigma(x_{Ai}-\bar{x}_A)^2 = \Sigma x_{Ai}^2 - \frac{(\Sigma x_{Ai})^2}{n_1} = 9521 - \frac{(331)^2}{12}$$

$$= 9521 - 9130.08 = 390.92$$

$$\Sigma(x_{Bj}-\bar{x}_B)^2 = \Sigma x_{Bj}^2 - \frac{(\Sigma x_{Bj})^2}{n_2} = 14874 - \frac{(448)^2}{15}$$

$$= 14874 - 13380.27 = 1493.73$$

$$\therefore \quad s_p^2 = \frac{390.92+1493.73}{12+15-2} = \frac{1884.65}{25} = 75.386$$

so that $s_p = \sqrt{75.386} = 8.68$ (an estimate of the common σ)

Thus $$t = \frac{27.58 - 29.87}{8.68\sqrt{\frac{1}{12}+\frac{1}{15}}}$$

$$= \frac{-2.29}{(8.68)\,(0.387)} = \frac{-2.29}{3.359} = -0.68$$

(v) The critical region is $t \le -t_{.025(25)} = -2.06$

and $t \ge t_{.025(25)} = 2.06$.

(vi) The computed value of t does not fall in the critical region. We cannot reject H_0. The difference between the sample means (gains in weights from two diets) is not sufficient to demonstrate that any diet is better than the other.

Q.18.23. (b) (i) We state our hypotheses as

$H_0 : \mu = 115$ and $H_1 : \mu \neq 115$

(ii) Let the significance level be set at $\alpha = 0.05$

(iii) The test-statistic under H_0 is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}},$$

which has a Student's t-distribution with $(n-1)$ *d.f.*

(iv) Computations. $\bar{x} = \frac{\Sigma x}{n} = \frac{1422}{12} = 118.5$,

$$s^2 = \frac{1}{n-1}[\Sigma x^2 - (\Sigma x)^2/n]$$

$$= \frac{1}{11}[169350 - (1422)^2/12] = \frac{843}{11} = 76.64, \text{ so that}$$

$s = \sqrt{76.64} = 8.75$,

$$\therefore \quad t = \frac{118.5 - 115}{8.75/\sqrt{12}} = \frac{(3.5)\,(3.464)}{8.75} = 1.386$$

(v) The critical region is $|t| \ge t_{0.025(11)} = 2.201$

(vi) **Conclusion.** Since the computed value of $t = 1.386$ does not fall in the critical region, we therefore accept H_0.

(c) (i) We state our hypotheses as

$$H_0 : \mu_1 - \mu_2 = 0 \text{ and } H_1 : \mu_1 - \mu_2 \neq 0,$$

where μ_1 and μ_2 represent the mean of the first population and that of the second population of children respectively.

(ii) Let the significance level be set at $\alpha = 0.05$.

(iii) The test-statistic under H_0 is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

where $\bar{x}_1$ is the mean $1Q$ of the first group and $\bar{x}_2$ is the average $1Q$ of the second group; and the statistic has a Student's t-distribution with *d.f.* $= n_1 + n_2 - 2$.

(v) Computations. Now, $\bar{x}_2 = \frac{1103}{10} = 110.3$,

$$s_2^2 = \frac{1}{9}[121879 - (1103)^2/10] = \frac{236.1}{9} = 26.23$$

$$\therefore \quad s_p^2 = \frac{(n_1-1)\,s_1^2 + (n_2-1)\,s_2^2}{n_1 + n_2 - 2} = \frac{(11)(76.64)+(9)(26.23)}{12 + 10 - 2}$$

$$= \frac{843.04 + 236.07}{20} = \frac{1079.11}{20} = 53.9555, \text{ so that}$$

$$s_p = \sqrt{53.9555} = 7.35$$

Thus $$t = \frac{118.5 - 110.3}{7.35\sqrt{\frac{1}{12} + \frac{1}{10}}} = \frac{8.2}{3.147} = 2.61$$

(v) The critical region is $|t| \geq t_{0.025(20)} = 2.086$

(vi) **Conclusion.** Since the computed value of $t = 2.61$ falls in the critical region, we therefore reject H_0 and conclude that the second group significantly differs from the first group.

Q.18.24. (a) (i) We state our hypotheses as

$$H_0 : \mu_E - \mu_C = 0 \text{ and } H_1 : \mu_E - \mu_C \neq 0.$$

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic under H_0 is

$$t = \frac{\bar{X}_E - \bar{X}_C}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

which has a t-distribution with $d.f. = n_1 + n_2 - 2$.

(iv) Computations. Here $\bar{X}_E = \frac{140}{10} = 14, \bar{X}_C = \frac{117}{9} = 13,$

$$s_p^2 = \frac{1}{n_1+n_2-2}\left[\{\Sigma X_E^2 - (\Sigma X_E)^2/n_1\} + \{\Sigma X_C^2 - (\Sigma X_C)^2/n_2\}\right]$$

$$= \frac{1}{10+9-2}\left[\{1980 - \frac{(140)^2}{10}\} + \{1551 - \frac{(177)^2}{9}\}\right]$$

$$= \frac{20 + 30}{17} = 2.9412, \text{ so that } s_p = 1.715$$

Substituting these values, we get

$$t = \frac{14 - 13}{1.715\sqrt{\frac{1}{10} + \frac{1}{9}}} = \frac{1}{0.79} = 1.27$$

(v) The critical region is $|t| \geq t_{0.025(17)} = 2.11$.

(vi) **Conclusion.** Since the computed value of $t = 1.27$ does not fall in the critical region, we therefore reject H_0 and conclude that the means of the two groups do not differ significantly at the 5% significance level.

(b) (i) The hypotheses would be stated as

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

(ii) We use a significance level of $\alpha = 0.05$, and a two-sided test.

(iii) The test-statistic for the null hypothesis $H_0 : \mu_1 = \mu_2$, would be

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

where $$s_p^2 = \frac{1}{n_1+n_2-2}\left[\sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1)^2 + \sum_{j=1}^{n_2}(x_{2j}-\bar{x}_2)^2\right].$$

This statistic has a t-distribution with (n_1+n_2-2) $d.f.$

(iv) Computations.

Cotton yarn		Coir	
x_{1i}	x_{1i}^2	x_{2j}	x_{2j}^2
7.5	56.25	8.3	68.89
5.4	29.16	6.1	37.21
10.6	112.36	9.6	92.16
9.0	81.00	10.4	108.16
6.1	37.21	6.4	40.96
10.2	104.04	10.0	100.00
7.9	62.41	7.9	62.41
9.7	94.09	8.9	79.21
7.1	50.41	7.5	56.25
8.5	72.25	9.7	94.09
Σ 82.0	699.18	84.8	739.34

Now $\bar{x}_1 = \frac{82.0}{10} = 8.20, \quad \bar{x}_2 = \frac{84.8}{10} = 8.48,$

$$\sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1)^2 = \sum_i x_{1i}^2 - \frac{(\sum x_{1i})^2}{n_1} = 699.18 - \frac{(82.0)^2}{10}$$

$$= 699.18 - 672.40 = 26.78, \text{ and}$$

$$\sum_{j=1}^{n_2}(x_{2j}-\bar{x}_2)^2 = \sum_j x_{2j}^2 - \frac{(\sum x_{2j})^2}{n_2} = 739.34 - \frac{(84.8)^2}{10}$$

$$= 739.34 - 719.104 = 20.236$$

$$\therefore \quad s_p^2 = \frac{26.78 + 20.236}{10 + 10 - 2} = \frac{47.016}{18} = 2.612, \text{ so that}$$

$$s_p = \sqrt{2.612} = 1.616$$

Thus $$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{8.20 - 8.48}{1.616\sqrt{\frac{1}{10}+\frac{1}{10}}}$$

$$= \frac{-0.28}{1.616\,(0.447)} = \frac{-0.28}{0.72} = -0.39$$

(v) The critical region is $t \le -t_{.025(18)} = -2.10$ and

$$t \ge t_{.025(18)} = 2.10$$

(vi) Since the computed value does not fall in the critical region, so we cannot reject H_0. There is no significant difference in the strength of the two types of ropes.

Q.18.25. (i) The hypotheses would be stated as

$H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 \neq \mu_2$.

(ii) We use a significance level of $\alpha = 0.05$, and a two-sided test.

(iii) The test-statistic under $H_0 : \mu_1 = \mu_2$, would be

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

where $s_p^2 = \frac{1}{n_1 + n_2 - 2}[\sum(x_{1i} - \bar{x}_1)^2 + \sum(x_{2j} - \bar{x}_2)^2]$.

This statistic has a t-distribution with (n_1+n_2-2) *d.f.*

(iv) Computations:

1st Group		2nd Group	
x_{1i}	x_{1i}^2	x_{2j}	x_{2j}^2
2.6	6.76	3.5	12.25
1.5	2.25	2.5	6.25
4.0	16.00	1.5	2.25
1.0	1.00	2.5	6.25
3.5	12.25	3.0	9.00
3.4	11.56	2.0	4.00
2.5	6.25	3.0	9.00
3.0	9.00	2.0	4.00
4.0	16.00	1.5	2.25
3.5	12.25	2.5	6.25
Σ 29.0	93.32	24.0	61.50

Now $\bar{x}_1 = \frac{\sum x_{1i}}{n_1} = \frac{29.0}{10} = 2.9,$

$$\bar{x}_2 = \frac{\sum x_{2j}}{n_2} = \frac{24.0}{10} = 2.4,$$

$$\sum(x_{1i}-\bar{x}_1)^2 = \sum x_{1i}^2 - \frac{(\sum x_{1i})^2}{n_1} = 93.32 - 84.10 = 9.22,$$

$$\sum(x_{2j}-\bar{x}_2)^2 = \sum x_{2j}^2 - \frac{(\sum x_{2j})^2}{n_2} = 61.50 - 57.60 = 3.90$$

$$\therefore \quad s_p^2 = \frac{\sum(x_{1i}-\bar{x}_1)^2 + \sum(x_{2j}-\bar{x}_2)^2}{n_1 + n_2 - 2} = \frac{9.22 + 3.90}{10 + 10 - 2}$$

$$= \frac{13.12}{18} = 0.7289, \text{ so that } s_p = \sqrt{0.7289} = 0.854$$

$$\therefore \quad t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{2.9 - 2.4}{0.854\sqrt{\frac{1}{10} + \frac{1}{10}}}$$

$$= \frac{0.5}{(0.854)\,(0.447)} = \frac{0.5}{0.38} = 1.32,$$

and $d.f. = 10 + 10 - 2 = 18.$

(v) The critical region is $|t| \geq t_{.025(18)} = 2.10$

(vi) The computed value does not fall in the critical region. We cannot reject H_0. The difference between the two sample means is not significant.

Q.18.26. Given $n_1 = 9$, $\bar{x}_1 = 75$, $s_1 = 13.61$,
$n_2 = 16$, $\bar{x}_2 = 60$, $s_2 = 11.05$,

Populations are normal, but $\sigma_1^2 \neq \sigma_2^2$.

(i) We are required to decide between the hypotheses

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

(ii) The level of significance is $\alpha = 0.05$, and we use a one-tailed test.

(iii) Since the populations have unequal variances, the test statistic would be

$$t' = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

which has approximately a t-distribution with v degrees of freedom where

$$v = \frac{[(s_1^2/n_1) + (s_2^2/n_2)]^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

(iv) Computations.

$$t' = \frac{75 - 60}{\sqrt{\frac{(13.61)^2}{9} + \frac{(11.05)^2}{16}}}$$

$$= \frac{15}{\sqrt{20.5813 + 7.6314}} = \frac{15}{5.31} = 2.82, \text{ and}$$

$$v = \frac{[(13.61)^2/9 + (11.05)^2/16]^2}{\frac{[(13.61)^2/9]^2}{8} + \frac{[(11.05)^2/16]^2}{15}}$$

$$= \frac{(20.5813 + 7.6314)^2}{52.95 + 3.88} = \frac{795.96}{56.83} = 14$$

(v) The critical region is $t' \geq t_{.05(14)} = 1.76$.

(vi) Since the calculated value of t' falls in the critical region, so we reject the hypothesis of equal means.

Q.18.27. (b) (i) The hypotheses would be stated as

$H_0 : \mu_2 - \mu_1 = 0$, and giving up smoking has no effect on person's weight, and

$H_1 : \mu_2 - \mu_1 \neq 0$, and giving up smoking has effect on person's weight

(ii) We use a significance level of $\alpha = 0.05$, and a two-sided test.

(iii) The test-statistic under H_0 would be

$$t = \frac{\bar{d} - 0}{s_d / \sqrt{n}}, \quad \text{(paired observations)}$$

where $\bar{d}$ is the mean of the differences and s_d is the unbiased standard deviation of differences. Assuming populations to be normal, this statistic has the Student's distribution with $(n-1)$ $d.f.$

(iv) Computation:

Person	Weight Before (x_1)	Weight After (x_2)	$d_i(=x_2-x_1)$	d_i^2
1	148	154	6	36
2	176	176	0	0
3	153	151	−2	4
4	116	121	5	25
Σ	593	602	9	65

$$\therefore \quad \bar{d} = \frac{\Sigma d_i}{n} = \frac{9}{4} = 2.25$$

$$s_d^2 = \frac{1}{n-1}\Sigma(d_i-\bar{d})^2 = \frac{1}{n-1}\left\{\Sigma d_i^2 - \frac{(\Sigma d_i)^2}{n}\right\}$$

$$= \frac{1}{3}\{65 - 20.25\} = \frac{44.75}{3} = 14.9167, \text{ so that}$$

$$s_d = \sqrt{14.9167} = 3.862$$

Thus $$t = \frac{2.25 - 0}{3.862/\sqrt{4}} = \frac{4.50}{3.862} = 1.17$$

(v) The critical region is $t \le -t_{.025(3)} = -3.18$ and

$$t \ge t_{.025(3)} = 3.18..$$

(vi) Since the computed value of t does not fall in the critical region, so we cannot reject H_0. This experiment does not provide sufficient evidence to conclude that giving up smoking has no effect on a person's weight.

Q.18.28. (a) (i) We state our hypotheses as

$H_0 : \mu_2 - \mu_1 \le 0$ and $H_1 : \mu_2 - \mu_1 > 0$

(ii) The significance level is set at $\alpha = 0.05$

(iii) The test-statistic under H_0 is

$$t = \frac{\bar{d} - 0}{s_d/\sqrt{n}}, \quad \text{(paired observations)}$$

which has a t-distribution with $(n-1)$ $d.f.$

(iv) Computations.

Grades Before (X_{1i})	After (X_{2i})	$d_i = X_{2i} - X_{1i}$	d_i^2
44	53	9	81
40	38	−2	4
61	69	8	64
52	57	5	25
32	46	14	196
44	39	−5	25
70	73	3	9
41	48	7	49
67	73	6	36
72	74	2	4
53	60	7	49
72	78	6	36
648	708	60	578

Now $$\bar{d} = \frac{\Sigma d}{n} = \frac{60}{12} = 5, \text{ and}$$

$$s_d^2 = \frac{1}{n-1}[\Sigma d_i^2 - (\Sigma d)^2/n] = \frac{1}{11}[578 - (60)^2/12]$$

$$= \frac{278}{11} = 25.27, \text{ so that } s_d = 5.03.$$

Ths $$t = \frac{5}{5.03/\sqrt{12}} = 3.44, \text{ and } d.f. = 12 - 1 = 11.$$

(v) The critical region is $t \ge t_{0.05(11)} = 1.796$

(vi) **Conclusion.** Since the computed value of $t = 3.44$ falls in the the critical region, we therefore reject H_0 and conclude that the course has improved the performance.

(b) Observations not considered paired

(i) We state our hypotheses as

$H_0 : \mu_2 - \mu_1 = 0$ and $H_1 : \mu_2 - \mu_1 > 0$

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic under H_0 is

$$t = \frac{\bar{X}_2 - \bar{X}_1}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

which has a t-distribution with n_1+n_2-2 *d.f.*

(iv) Computations.

$$\bar{X}_1 = \frac{\Sigma X_{1i}}{n_1} = \frac{648}{12} = 54, \quad \bar{X}_2 = \frac{\Sigma X_{2i}}{n_2} = \frac{708}{12} = 59,$$

$$\Sigma(X_{1i}-\bar{X}_1)^2 = \Sigma X_{1i}^2 - \frac{(\Sigma X_{1i})^2}{n_1} = 37168 - \frac{(648)^2}{12} = 2176,$$

$$\Sigma(X_{2i}-\bar{X}_2)^2 = \Sigma X_{2i}^2 - \frac{(\Sigma X_{2i})^2}{n_2} = 44022 - \frac{(708)^2}{12} = 2250,$$

$$\therefore \quad s_p^2 = \frac{2176 + 2250}{12 + 12 - 2} = 201.1818, \text{ so that } s_p = \sqrt{201.1818} = 14.18$$

$$\text{Thus} \quad t = \frac{59 - 54}{14.18 \sqrt{\frac{1}{12} + \frac{1}{12}}} = \frac{5}{(14.18)\,(0.4082)} = 0.86.$$

(v) The critical region is $t \geq t_{0.05(22)} = 1.717$

(vi) **Conclusion.** Since the computed value of $t = 0.86$ does not fall in the critical region, we therefore accept H_0.

It is to be noted that the same conclusion has not been reached.

Q.18.29. (a) Given two independent samples from two normal populations with means μ_A, μ_B and having common variance σ^2. That is the observations are not paired.

(i) We wish to test the hypothesis

$H_0 : \mu_B - \mu_A = 0$, *i.e.* Food B is not better than food A, against

$H_1 : \mu_B - \mu_A > 0$, *i.e.* Food B is better than food A.

(ii) We use a significance level of $\alpha = 0.05$, and a one-sided test.

(iii) For the null hypothesis, the test-statistic would be

$$t = \frac{\bar{x}_B - \bar{x}_A}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

where $$s_p^2 = \frac{1}{n_1 + n_2 - 2} [\Sigma(x_{A_i} - \bar{x}_A)^2 + \Sigma(x_{B_j} - \bar{x}_B)^2]$$

This statistic has t-distribution with (n_1+n_2-2) *d.f.*

(iv) Computations.

Sheep No.	Increase in weights: Food A (x_{A_i})	Increase in weights: Food B (x_{B_j})	D_{1i} ($x_{A_i}-50$)	D_{1i}^2	D_{2j} ($x_{B_j}-50$)	D_{2j}^2
1	49	52	−1	1	2	4
2	53	55	3	9	5	25
3	51	52	1	1	2	4
4	52	53	2	4	3	9
5	47	50	−3	9	0	0
6	50	54	0	0	4	16
7	52	54	2	4	4	16
8	53	53	3	9	3	9
Σ	407	423	7	37	23	83

Now $\bar{x}_A = \frac{407}{8} = 50.875$, $\bar{x}_B = \frac{423}{8} = 52.875$

$$\Sigma(x_{A_i} - \bar{x}_A)^2 = \Sigma D_{1i}^2 - \frac{(\Sigma D_{1i})^2}{n_1} = 37 - 6.125 = 30.875.$$

$$\Sigma(x_{B_j} - \bar{x}_B)^2 = \Sigma D_{2j}^2 - \frac{(\Sigma D_{2j})^2}{n_2} = 83 - 66.125 = 16.875.$$

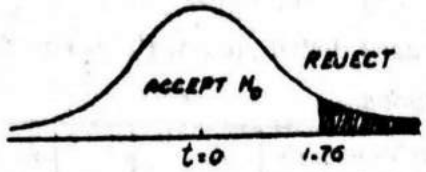
$$\therefore \quad s_p^2 = \frac{30.875 + 16.875}{8 + 8 - 2} = \frac{47.75}{14} = 3.4107, \text{ so that}$$

$$s_p = \sqrt{3.4107} = 1.85$$

Thus $$t = \frac{52.875 - 50.875}{1.85\sqrt{\frac{1}{8} + \frac{1}{8}}} = \frac{2.00}{(1.85)(0.5)} = \frac{2.00}{0.925} = 2.16$$

and $d.f. = 8 + 8 - 2 = 14$.

(v) The critical region is $t \geq t_{.05(14)} = 1.76$



(vi) Since the computed value of t falls in the critical region, so we reject H_0 in favour of H_1. Hence the experiment provides an evidence to conclude that food B is better than food A.

(b) The same set of eight sheep were used in both the foods. This satement implies that the data are to be treated as paired obervations.

(i) The hypotheses would be stated as

$H_0 : \mu_B - \mu_A = 0$, *i.e.* Food B is not better than food A, and

$H_1 : \mu_B - \mu_A > 0$, *i.e.* Food B is better than food A.

(ii) We use a significance level of $\alpha = 0.05$, and a one-sided test.

(iii) The test-statistic under H_0, would be

$$t = \frac{\bar{d} - 0}{s_d / \sqrt{n}},$$

where $\bar{d}$ = the mean of the differences, and

s_d = unbiased standard deviation of the differences.

This statistic has a t-distribution with $(n-1)$ $d.f.$

(iv) Computations.

Sheep No.	Increase in weights		$d_i\ (x_{B_j} - x_{A_i})$	$d_i - \bar{d}$	$(d_i - \bar{d})^2$
	Food A (x_{A_i})	Food B (x_{B_j})			
1	49	52	3	1	1
2	53	55	2	0	0
3	51	52	1	−1	1
4	52	53	1	−1	1
5	47	50	3	1	1
6	50	54	4	2	4
7	52	54	2	0	0
8	53	53	0	−2	4
Σ	407	423	16	0	12

Now $\bar{d} = \frac{\Sigma d_i}{n} = \frac{16}{8} = 2$, and

$$s_d = \sqrt{\frac{1}{n-1}\Sigma(d_i - \bar{d})^2} = \sqrt{\frac{12}{7}} = \sqrt{1.7143} = 1.31$$

$$\therefore \quad t = \frac{2 - 0}{1.31 / \sqrt{8}} = \frac{(2)(2.828)}{1.31} = \frac{5.656}{1.31} = 4.32, \text{ and}$$

$d.f. = 8 - 1 = 7$.

(v) The critical region is $t \geq t_{.05(7)} = 1.895$

(vi) The computed value of t falls in the critical region. We reject H_0. There is sufficient evidence to conclude that food B is better than food A.

Q.18.30. (i) We state our hypotheses as

$H_0 : \mu_2 - \mu_1 = 0$ and $H_1 : \mu_2 - \mu_1 \neq 0$.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic under H_0 (paired vlaues) is

$$t = \frac{\bar{d} - 0}{s_d / \sqrt{n}},$$

which has a t-distribution with $(n-1)$ *d.f.*

(vi) Computations.

Station	1	2	3	4	5	6	7	8	9	Total
Variety 1	38	23	35	41	44	29	37	31	38	316
Variety 2	45	25	31	38	50	33	35	40	43	340
d_i (=2–1)	7	2	–4	–3	6	4	–2	9	5	24
d_i^2	49	4	16	9	36	16	4	81	25	240

Now $\bar{d} = \dfrac{\sum d_i}{n} = \dfrac{24}{9} = 2.67$, and

$$s_d^2 = \frac{1}{n-1}\left[\sum d_i^2 - \frac{(\sum d_i)^2}{n}\right] = \frac{1}{8}[240\text{–}64] = 22, \text{ so that } s_d = 4.69$$

Thus $t = \dfrac{2.67}{4.69 / \sqrt{9}} = 1.71$

(v) The critical region is $|t| \geq t_{0.025(8)} = 2.31$.

(vi) **Conclusion.** Since the computed value of $t = 1.71$ does not fall in the critial region, we therefore accept H_0.

Q.18.31. Let μ_X and μ_Y denote the gasoline consumption in km per liter by the use of radial tires and regular belted tires respectively. Then the hypotheses are stated as

(i) $H_0 : \mu_X - \mu_Y \leq 0$ and $H_1 : \mu_X - \mu_Y > 0$

(ii) The level of significance is set at $\alpha = 0.025$.

(iii) The test-statistic under H_0 (paired values) is

$$t = \frac{\bar{d} - 0}{s_d / \sqrt{n}},$$

which has a t-distribution with $(n-1)$ *d.f.*

(iv) Computations.

Radial tires (X)	Belted tires (Y)	$d = X - Y$	d^2
4.2	4.1	0.1	0.01
4.7	4.9	–0.2	0.04
6.6	6.2	0.4	0.16
7.0	6.9	0.1	0.01
6.7	6.8	–0.1	0.01
4.5	4.4	0.1	0.01
5.7	5.7	0	0
6.0	5.8	0.2	0.04
7.4	6.9	0.5	0.25
4.9	4.7	0.2	0.04
6.1	6.0	0.1	0.01
5.2	4.9	0.3	0.09
	Σ	1.7	0.67

Now $\bar{d} = \dfrac{\sum d}{n} = \dfrac{1.7}{12} = 0.1417$, and

$$s_d^2 = \frac{1}{n-1}\left[\sum d^2 - \frac{(\sum d)^2}{n}\right] = \frac{1}{11}\left[0.67 - \frac{(1.7)^2}{12}\right]$$

$$= \frac{1}{11}[0.67 - 0.2408] = \frac{0.4292}{11} = 0.0390$$

$$s_d = \sqrt{0.0390} = 0.1975$$

Thus $t = \dfrac{0.1417}{0.1975/\sqrt{12}} = \dfrac{(0.1417)\,(3.46)}{0.1975} = 2.482$

(v) The critical region is $t > t_{0.025,(11)} = 2.201$.

(vi) **Conclusion.** Since the computed value of t falls in the critical region, we therefore reject H_0. We may conclude that cars equipped with radial tires give better fuel economy than those equipped with belted tires.

Q.18.32. Here two large samples have been taken from the population, assumed normal, with σ unknown, the t-test may therefore be correctly applied.

(i) The hypotheses are stated as

$H_0 : \mu_1 = \mu_2$, *i.e.* there is no significant difference in achievement of the two groups, and

$H_1 : \mu_1 \neq \mu_2$

(ii) We choose the level of significance of $\alpha = 0.05$

(iii) The test-statistic under H_0 is

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$$

which has a t-distribution with (n_1+n_2-2) *d.f.* .

(iv) Computations.

Teacher Instructed:

Here $n_1 = 166$, $\Sigma f_1 x = 10527$, $\Sigma f_1 x^2 = 689381.50$

$$\therefore \quad \bar{X}_1 = \frac{\Sigma f_1 x}{n_1} = \frac{10527}{166} = 63.42, \text{ and}$$

$$n_1 S_1^2 = \Sigma f_1 x^2 - (\Sigma f_1 x)^2/166$$

$$= 689381.50 - \frac{(10527)^2}{166} = 21804.82$$

Machine Instructed:

Here $n_2 = 164$, $\Sigma f_2 x = 10828$, $\Sigma f_2 x^2 = 741031$

$$\therefore \quad \bar{X}_2 = \frac{\Sigma f_2 x}{n_2} = \frac{10828}{164} = 66.02, \text{ and}$$

$$n_2 S_2^2 = \Sigma f_2 x^2 - (\Sigma f_2 x)^2/164$$

$$= 741031 - \frac{(10828)^2}{164} = 26118.90$$

$$\text{Now} \quad s_p^2 = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2} = \frac{21804.82 + 26118.90}{166 + 164 - 2}$$

$$= \frac{47923.72}{328} = 146.1089, \text{ so that}$$

$$s_p = \sqrt{146.1089} = 12.09$$

$$\text{Thus} \quad |t| = \frac{66.02 - 63.42}{12.09\sqrt{\dfrac{1}{166} + \dfrac{1}{164}}} = \frac{2.60}{(12.09)\,(0.1101)} = 1.95$$

(v) The critical region is $|t| \geq t_{0.025(328)} = 1.96$

(vi) **Conclusion.** Since the calculated value of t does not fall in the critical region, we therefore accept H_0 and may conclude that there is no significant difference in achievement of the two groups.

❖❖❖❖❖❖❖❖

CHAPTER 19

THE F-DISTRIBUTION AND STATISTICAL INFERENCE

Q.19.3. (a) The F-ratio may be written as

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} = \frac{U/\nu_1}{V/\nu_2}$$

where $U = \nu_1 s_1^2/\sigma_1^2$ and $V = \nu_2 s_2^2/\sigma_2^2$ have chi-square distributions with ν_1 and ν_2 degrees of freedom respectively and are independent. The mean and variance are found as follows:

$$E(F) = E\left[\frac{U/\nu_1}{V/\nu_2}\right] = \frac{\nu_2}{\nu_1} E(U)\, E\left(\frac{1}{V}\right)$$

Since U is a chi-square variate with ν_1 *d.f.*, therefore

$$E(U) = \nu_1, \text{ and}$$

$$E\left(\frac{1}{V}\right) = \frac{1}{\Gamma(\nu_2/2)}\left(\frac{1}{2}\right)^{\nu_2/2}\int_0^\infty \frac{1}{v} v^{(\nu_2-2)/2} e^{-v/2}\, dv$$

$$= \frac{1}{\Gamma(\nu_2/2)}\left(\frac{1}{2}\right)^{\nu_2/2}\int_0^\infty v^{(\nu_2-4)/2} e^{-v/2}\, dv$$

$$= \frac{1}{\Gamma(\nu_2/2)}\left(\frac{1}{2}\right)^{\nu_2/2}\Gamma[(\nu_2-2)/2]\left(\frac{1}{2}\right)^{-(\nu_2-2)/2}$$

$$= \frac{\Gamma[(\nu_2-2)/2]}{(\nu_2-2)/2\,\Gamma[(\nu_2-2)/2]}\left(\frac{1}{2}\right) = \frac{1}{\nu_2-2}$$

$$\therefore \quad E(F) = \frac{\nu_2}{\nu_1}\cdot\frac{\nu_1}{\nu_2-2} = \frac{\nu_2}{\nu_2-2}, \text{ for } \nu_2 > 2.$$

Thus we find that the mean depends only on the degrees of freedom of the denominator (or does not involve ν_1, the degrees of freedom of the numerater).

For variance, we first find $E(F^2)$, where

$$E(F^2) = E\left(\frac{U^2/\nu_1^2}{V^2/\nu_2^2}\right) = \frac{\nu_2^2}{\nu_1^2} E[U^2]\, E\left(\frac{1}{V^2}\right)$$

But $E[U^2] = \nu_1(\nu_1+2)$ as U is a χ^2-variate with ν_1 *d.f.*, and

$$E\left[\frac{1}{V^2}\right] = \frac{1}{\Gamma(\nu_2/2)}\left(\frac{1}{2}\right)^{\nu_2/2}\int_0^\infty \frac{1}{v^2} v^{(\nu_2-2)/2} e^{-v/2}\, dv$$

$$= \frac{1}{\Gamma(\nu_2/2)}\left(\frac{1}{2}\right)^{\nu_2/2}\int_0^\infty v^{(\nu_2-6)/2} e^{-v/2}\, dv$$

$$= \frac{1}{\Gamma(\nu_2/2)}\left(\frac{1}{2}\right)^{\nu_2/2}\Gamma[(\nu_2-4)/2]\left(\frac{1}{2}\right)^{-(\nu_2-4)/2}$$

$$= \frac{\Gamma[(\nu_2-4)/2]}{[(\nu_2-2)/2]\,[(\nu_2-4)/2]\,\Gamma[(\nu_2-4)/2]}\left(\frac{1}{2}\right)^{(\nu_2-\nu_2+4)/2}$$

$$= \frac{4}{(\nu_2-2)\,(\nu_2-4)}\left(\frac{1}{2}\right)^2 = \frac{1}{(\nu_2-2)\,(\nu_2-4)}$$

$$\therefore \quad E(F^2) = \frac{\nu_2^2}{\nu_1^2}\cdot\frac{\nu_1\,(\nu_1+2)}{(\nu_2-2)\,(\nu_2-4)} = \frac{\nu_2^2\,(\nu_1+2)}{\nu_1(\nu_2-2)(\nu_2-4)}$$

But $\text{var}(F) = E(F^2) - [E(F)]^2$

$$= \frac{\nu_2^2\,(\nu_1+2)}{\nu_1\,(\nu_2-2)\,(\nu_2-4)} - \left(\frac{\nu_2}{\nu_2-2}\right)^2$$

$$= \frac{2\nu_2^2\,(\nu_1+\nu_2-2)}{\nu_1\,(\nu_2-2)^2\,(\nu_2-4)}, \text{ for } \nu_2 > 4.$$

Thus the variance depends only on the degrees of freedom. Hence we see that the distribution of F does not depend on the

population variance σ^2, but depends on v_1 and v_2, the degrees of freedom.

(b) $$F(v_1,v_2) = \frac{{v_1}^{v_1/2} \cdot {v_2}^{v_2/2} F^{(v_1/2)-1}}{\beta\left(\frac{v_1}{2}, \frac{v_2}{2}\right)(v_2+v_1 F)^{(v_1+v_2)/2}}$$

Let $F' = \frac{1}{F}$ so that $|J| = \left|\frac{1}{F^2}\right|$

$$\therefore \quad F'(v_1,v_2) = \frac{{v_1}^{v_1/2} \cdot {v_2}^{v_2/2} (1/F)^{(v_1/2)-1} F^{(v_1+v_2)/2}}{\beta\left(\frac{v_1}{2}, \frac{v_2}{2}\right)(v_2F+v_1)^{(v_1+v_2)/2}} \frac{1}{F^2} \qquad [F' = \frac{1}{F}]$$

$$= \frac{{v_1}^{(v_1/2)} \cdot {v_2}^{v_2/2} F^{(v_2/2)-1}}{\beta\left(\frac{v_2}{2}, \frac{v_1}{2}\right)(v_1+v_2F)^{(v_1+v_2)/2}}$$

Hence $\frac{1}{F(v_1, v_2)} = F(v_2, v_1)$

Q.19.4.(b) The 90% confidence interval for $\frac{\sigma_1^2}{\sigma_2^2}$ would be

$$\frac{s_1^2}{s_2^2} \cdot \frac{1}{F_{.05(v_1,v_2)}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2} \cdot F_{.05(v_1,v_2)} \qquad (\alpha = 0.10)$$

Substituting the values, we get

$$\frac{50}{16} \cdot \frac{1}{2.40} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{50}{16} \cdot (2.40) \qquad (v_1 = v_2 = 15)$$

i.e. $$1.30 < \frac{\sigma_1^2}{\sigma_2^2} < 7.50$$

Hence the required 90% confidence interval for $\frac{\sigma_1^2}{\sigma_2^2}$ is 1.30 to 7.50.

(c) The 98% confidence interval for the ratio $\frac{\sigma_1^2}{\sigma_2^2}$ would be

$$\frac{s_1^2}{s_2^2} \cdot \frac{1}{F_{.01(v_1,v_2)}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2} \cdot F_{.01(v_2,v_1)} \qquad (\alpha = 0.02)$$

Substituting the values, we get

$$\frac{15.6}{6.3} \times \frac{1}{3.62} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{15.6}{6.3}(2.66) \qquad \begin{cases} F_{.01,(40,12)} = 3.62 \\ F_{.01,(12,40)} = 2.66 \end{cases}$$

i.e. $$0.68 < \frac{\sigma_1^2}{\sigma_2^2} < 6.59$$

Hence the required 98% confidence interval for $\frac{\sigma_1^2}{\sigma_2^2}$ is 0.68 to 6.59.

Q.19.5 (b) Given $n_1 = 9, n_2 = 16, s = 6$ and $s_2 = 5$.

The 98% confidence interval for $\frac{\sigma_1^2}{\sigma_2^2}$ is given by

$$\frac{s_1^2}{s_2^2} \cdot \frac{1}{F_{0.01(v_1,v_2)}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2} \cdot F_{0.01(v_2,v_1)} \qquad (\alpha = 0.02)$$

Substituting the values, we get

$$\frac{36}{25} \cdot \frac{1}{4.00} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{36}{25}(5.52)$$

i.e., $$0.36 < \frac{\sigma_1^2}{\sigma_2^2} < 7.95.$$

The confidence interval for $\frac{\sigma_1}{\sigma_2}$ is obtained by taking the square root of the endpoints (0.36, 7.95). Thus we get

$$0.600 < \frac{\sigma_1}{\sigma_2} < 2.819$$

(c) The 90% confidence interval for $\frac{\sigma_2^2}{\sigma_1^2}$ is given by

$$\frac{s_2^2}{s_1^2} \cdot \frac{1}{F_{0.05(7,9)}} < \frac{\sigma_2^2}{\sigma_1^2} < \frac{s_2^2}{s_1^2} \cdot F_{0.05(9,7)}$$

Now $s_1^2 = 7.2$, $s_2^2 = 3.6$, $F_{0.05(7,9)} = 3.29$ and $F_{0.05(9,7)} = 3.68$

Substituting these value, we get

$$\frac{3.6}{7.2} \cdot \frac{1}{3.29} < \frac{\sigma_2^2}{\sigma_1^2} < \frac{3.6}{7.2}(3.68)$$

i.e., $$0.152 < \frac{\sigma_2^2}{\sigma_1^2} < 1.840$$

Q.19.6. (b) (i) Given $H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$ and $H_1 : \frac{\sigma_1^2}{\sigma_2^2} > 1$. (one tailed)

$$F = \frac{s_1^2}{s_2^2} = \frac{50}{16} = 3.125$$

The critical region is $F \geq F_{0.05(15,15)} = 2.40$

Since the computed value of $F = 3.125$ falls in the critical region, we therefore reject H_0.

(ii) Given $H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$ and $H_1 : \frac{\sigma_1^2}{\sigma_2^2} < 1$.

The alternative hypothesis may be written as $H_1 = \frac{\sigma_2^2}{\sigma_1^2} > 1$

Then $F = \frac{s_2^2}{s_1^2} = \frac{15.6}{6.3} = 2.48$

The critical region is $F \geq F_{0.01(40,12)} = 3.62$

Since the computed value of $F = 2.48$ does not fall in the critical region, we therefore accept H_0.

(iii) Given $H_0 : \sigma_1^2 = \sigma_2^2$ and $H_1 : \sigma_1^2 \neq \sigma_2^2$ (two-tailed)

Now $F = \frac{s_1^2}{s_2^2} = \frac{8.0}{17.0} = 0.47$

The critical region is $F \geq F_{0.01(59,119)} = 1.62$, and

$$F \leq \frac{1}{F_{0.01(119,59)}} = \frac{1}{1.74} = 0.57.$$

Since the computed value of $F = 0.47$ falls in the critical region, we therefore reject H_0.

Alternatively: Since s_2^2 is larger than s_1^2 and $H_1 : \sigma_1^2 \neq \sigma_2^2$, we may interchange the role of the two samples in order to apply one tailed test. Then

$$F = \frac{s_2^2}{s_1^2} = \frac{17.0}{8.0} = 2.125, \text{ and}$$

the critical region becomes $F \geq F_{0.01(119,59)}$.

Q.19.7. (a) (i) The null and alternative hypotheses are given as

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ and } H_1 : \sigma_1^2 \neq \sigma_2^2$$

(ii) The significance level is set at $\alpha = 0.10$, and we use a two-tailed test. Then $\alpha/2 = 0.05$ for each of the rejection areas.

(iii) The test-statistic under H_0 is

$$F = \frac{s_1^2}{s_2^2}, \qquad (s_1^2 > s_2^2)$$

which has an F-distribution with v_1 and v_2 degrees of freedom.

(iv) Computations. $F = \frac{16}{3} = 5.33$

(v) The critical regions are $F \geq F_{0.05(9,6)} = 4.10$, and

$$F \leq \frac{1}{F_{0.05(6,9)} = 3.37} = 0.297.$$

(vi) **Conclusion.** Since the computed value of $F = 5.33$ falls in the critical region, we therefore reject H_0.

Again, to test $H_0 : \sigma_1^2 = \sigma_2^2$ against $H_1 : \sigma_1^2 > \sigma_2^2$, we are given $\alpha = 0.05$. As $s_1^2 > s_2^2$, therefore

$$F = \frac{16}{3} = 5.33$$

The critical region is $F \geq F_{0.05(9,6)} = 4.10$.

Since the computed value of $F = 5.33$ falls in the critical region, so we reject H_0.

To test $H_0 : \sigma_1^2 = \sigma_2^2$ against $H_1 : \sigma_1^2 < \sigma_2^2$, we would compute

$$F = \frac{s_2^2}{s_1^2} = \frac{3}{16} = 0.18.$$

The critical region would be $F \geq F_{0.05(6,9)} = 3.37$

Since the computed value of $F = 0.18$ does not fall in the critical region, we therefore accept H_0.

(b) (i) We state our hypotheses as

$H_0 : \sigma_1^2 = \sigma_2^2$, *i.e.* the two methods of teaching are equally variable; and

$H_1 : \sigma_1^2 > \sigma_2^2$, *i.e.* the two methods of teaching are not equally variable.

(ii) We use a significance level of $\alpha = 0.05$

(iii) The test-statistic under H_0 is

$$F = \frac{s_1^2}{s_2^2}, \text{ where } s_1^2 > s_2^2, \text{ and}$$

which has an F-distribution with v_1 and v_2 *d.f.*

(iv) Computations.

$$F = \frac{47}{30} = 1.57; \text{ and } v_1 = 7, \; v_2 = 9$$

(v) The critical region is $F > F_{0.05(7,9)} = 3.29$

(vi) **Conclusion.** Since the calculated value of F does not fall in the critical region, we therefore accept H_0 and conclude that the two methods of teaching are equally variable.

Q.19.8. (i) The hypotheses are stated as

$H_0 : \sigma_1^2 = \sigma_2^2$ and $H_0 : \sigma_1^2 \neq \sigma_2^2$ (two-tailed test)

(ii) The level of significance is set at $\alpha = 0.02$.

(iii) The test-statistic under H_0, is

$$F = \frac{s_1^2}{s_2^2},$$

which has an F-distribution with v_1 and v_2 *d.f.*

(iv) Computations. Given $n_1 = 25$, $n_2 = 16$,

$\bar{x}_1 = 82$, $\bar{x}_2 = 78$, $S_1 = 8$ and $S_2 = 7$

Now $s_1^2 = \frac{n_1}{n_1-1} \cdot S_1^2 = \frac{25}{25-1}(8)^2 = 66.67$, and

$$s_2^2 = \frac{n_2}{n_2-1} \cdot S_2^2 = \frac{16}{16-1}(7)^2 = 52.27$$

$$\therefore \quad F = \frac{s_1^2}{s_2^2} = \frac{66.67}{52.27} = 1.28$$

(v) The critical region is $F > F_{0.01(24,15)}$ and $F < F_{0.99(24,15)}$

Now $F_{0.01(24,15)} = 3.29$ and

$$F_{0.99(24,15)} = \frac{1}{F_{0.01(15,24)}} = \frac{1}{2.89} = 0.35$$

Thus the acceptance region is $0.35 < F < 3.29$.

(vi) **Conclusion.** Since the computed value of F falls in the acceptance region, we therefore accept H_0 at 2% level of significance; and thus we may conclude that the variances are equal.

Q.19.9. (i) The hypotheses are stated as

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ and } H_1 : \sigma_1^2 > \sigma_2^2$$

(ii) The level of significance is $\alpha = 0.05$

(iii) The test-statistic under H_0 is

$$F = \frac{s_1^2}{s_2^2}, \qquad (s_1^2 > s_2^2)$$

which has an F-distribution with v_1 and v_2 *d.f.*

(iv) Computations. Let X_{1i} and X_{2j} denote the observations from sample 2 and sample 1 respectively. Then $\Sigma X_{1i} = 21.8$, $\Sigma X_{1i}^2 = 130.40$, $\Sigma X_{2j} = 14.5$ and $\Sigma X_{2j}^2 = 45.35$.

Now $s_1^2 = \frac{1}{4}[130.40 - (21.8)^2/5] = \frac{1}{4}(130.40 - 95.048)$

$= \frac{1}{4}[35.352] = 8.84$; and

$s_2^2 = \frac{1}{5}[45.35 - (14.5)^2/6] = \frac{1}{5}(45.35 - 35.0417)$

$= \frac{1}{5}(10.308) = 2.06$

$\therefore \quad F = \frac{8.84}{2.06} = 4.29$

(v) The critical region is $F > F_{0.05(4,5)} = 5.19$

(vi) **Conclusion.** Since the computed value of F does not fall in the critical region, we therefore do not reject H_0. We may conclude that the data do not provide sufficient evidence to indicate a difference between the population variances.

Q.19.10. (a) To test whether there is a significant difference between the two samples as regards variability, we proceed as below:

(i) We state our hypotheses as

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ and } H_1 : \sigma_1^2 > \sigma_2^2$$

(ii) We choose a significance level of $\alpha = 0.05$

(iii) The test-statistic under H_0 is

$$F = \frac{s_1^2}{s_2^2},$$

which has an F-distributioin with v_1 and v_2 degrees of freedom.

(iv) Computations.

X_{1i}	X_{1i}^2	X_{2j}	X_{2j}^2
15.7	246.49	12.3	151.29
10.3	106.09	13.7	187.69
12.6	158.76	10.4	108.16
14.5	210.25	11.4	129.96
12.6	158.76	14.9	222.01
13.8	190.44	12.6	158.76
11.9	141.61		
91.4	1212.40	75.3	957.17

$$s_1^2 = \frac{1}{n_1-1}\left[\Sigma X_{1i}^2 - (\Sigma X_{1i})^2/n_1\right] = \frac{1}{6}[1212.40 - (91.4)^2/7]$$

$$= \frac{1}{6}[1212.40 - 1193.42] = 3.16;$$

$$s_2^2 = \frac{1}{n_2 - 1}[\Sigma X_{2j}^2 - (\Sigma X_{2j})^2/n_2] = \frac{1}{5}[957.87 - (75.3)^2/6]$$

$$= \frac{1}{5}[957.87 - 945.02] = 2.57.$$

$$\therefore \quad F = \frac{s_1^2}{s_2^2} = \frac{3.16}{2.57} = 1.23$$

(v) The critical region is $F \geq F_{0.05(6,5)} = 4.95$.

(vi) **Conclusion.** Since the computed value of $F = 1.23$ does not fall in the critical region, we therefore accept H_0 and conclude that there is no significant difference between the two samples as regards variability.

(b) Since the variation between the two samples is not significant, we use t to test whether there is a significant difference between the two samples as regards the mean percentage extension.

Thus $$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (\text{under } H_0 : \mu_1 = \mu_2)$$

where $$\bar{X}_1 = \frac{91.4}{7} = 13.06, \quad \bar{X}_2 = \frac{75.3}{6} = 12.55.$$

$$s_p^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2} = \frac{(6)\,(3.16) + (5)\,(3.57)}{7 + 6 - 2}$$

$$= \frac{18.96 + 12.85}{11} = 2.8918, \text{ so that } s_p = 1.70$$

Substituting these values, we get

$$t = \frac{13.06 - 12.55}{1.70\sqrt{\frac{1}{7} + \frac{1}{6}}} = \frac{0.51}{(1.70)\,(0.556)} = 0.54$$

The critical region is $|t| \geq t_{0.025(11)} = 2.201$.

Since the computed value of $t = 0.54$ does not fall in the critical region, so we accept $H_0 : \mu_1 = \mu_2$, and conclude that the two samples do not differ significantly as regards the mean percentage extension.

Q.19.11. (i) We state our hypotheses as

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ and } H_1 : \sigma_1^2 < \sigma_2^2.$$

(ii) Let the significance level be set at $\alpha = 0.05$.

(iii) The test-statistic under H_0 is

$$F = \frac{s_2^2}{s_1^2}, \qquad (s_2^2 > s_1^2)$$

which has an F-distribution with v_2 and v_1 *d.f.*

(iv) Computations. Now $\Sigma X_{1i} = 21.4$, $\Sigma X_{1i}^2 = 76.50$.

$$\therefore \quad s_1^2 = \frac{1}{n_1 - 1}[\Sigma X_{1i}^2 - (\Sigma X_{1i})^2/n_1] = \frac{1}{5}[76.50 - (21.4)^2/6]$$

$$= \frac{1}{5}[76.50 - 76.33] = \frac{0.17}{5} = 0.034$$

Again, $\Sigma X_{2j} = 32.0$, $\Sigma X_{2j}^2 = 129.40$,

$$\therefore \quad s_2^2 = \frac{1}{n_2 - 1}[\Sigma X_{2j}^2 - (\Sigma X_{2j})^2/n_2] = \frac{1}{7}[129.40 - (32.0)^2/8]$$

$$= \frac{1}{7}[129.40 - 128.00] = 0.20$$

Thus $$F = \frac{s_2^2}{s_1^2} = \frac{0.20}{0.034} = 5.88$$

(v) The critical region is $F \geq F_{0.05(7,5)} = 4.88$.

(vi) **Conclusion.** Since the computed value of $F = 5.88$ falls in the critical region, we therefore reject H_0 and conclude that the variances are not equal.

As the variances are not equal, so we donot use t-test to test the equality of means. The assumption of equal variances is necessary to test $H_0 : \mu_1 = \mu_2$ by the use of t-statistic.

Q.19.12. (i) We are required to decide between the hypotheses

$$H_0 : \sigma_1^2 = 1.25\,\sigma_2^2 \quad \text{and} \quad H_1 : \sigma_1^2 > 1.25\,\sigma_2^2$$

(ii) We use a level of significance of $\alpha = 0.05$, and one-tailed test.

(iii) The test-statistic under the null hypothesis would be

$$F = \frac{s_1^2}{1.25\,s_2^2}$$

which has an F-distribution with $v_1 = n_1 - 1$, $v_2 = n_2 - 1$ degrees of freedom.

(iv) Computation. Given $n_1 = n_2 = 21$, $\bar{y}_1 = 50, \bar{y}_2 = 53$

$$\Sigma(y_{1i} - \bar{y}_1)^2 = 720, \quad \Sigma(y_{2j} - \bar{y}_2)^2 = 340$$

Then $$s_1^2 = \frac{1}{n_1 - 1}\Sigma(y_{1i} - \bar{y}_1)^2 = \frac{720}{20} = 36,$$

$$s_2^2 = \frac{1}{n_2 - 1}\Sigma(y_{2j} - \bar{y}_2)^2 = \frac{340}{20} = 17.$$

$$\therefore \quad F = \frac{36}{(1.25)\,(17)} = \frac{36}{21.25} = 1.69$$

(v) The critical region is $F > F_{.05(20,20)} = 2.12$

(vi) Since the computed value of F does not fall in the critical region, so we do not reject H_0. Hence the data do not provide sufficient evidence to contradict our null hypothesis.

❖❖❖❖❖❖❖❖

CHAPTER 20

THE ANALYSIS OF VARIANCE

Q.20.6. (i) $\sum_{i=1}^{r}\sum_{j=1}^{c} (\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})\,(X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X}_{\cdot\cdot})$

$$= \sum_{i=1}^{r} (\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot}) \left[\sum_{j=1}^{c} (X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X}_{\cdot\cdot})\right]$$

$$= \sum_{i=1}^{r} (\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})(\bar{X}_{i\cdot} - c\bar{X}_{i\cdot} - \sum_{j=1}^{c} \bar{X}_{\cdot j} + c\bar{X}_{\cdot\cdot})$$

$$= \sum_{i=1}^{r} (\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})\,(c\bar{X}_{i\cdot} - c\bar{X}_{i\cdot} - c\bar{X}_{\cdot\cdot} + c\bar{X}_{\cdot\cdot})$$

$$= \sum_{i=1}^{r} (\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})\,(0) = 0.$$

(ii) $\sum_{i=1}^{r}\sum_{j=1}^{c} (\bar{X}_{\cdot j} - \bar{X}_{\cdot\cdot})\,(X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X}_{\cdot\cdot})$

$$= \sum_{j=1}^{c} (\bar{X}_{\cdot j} - \bar{X}_{\cdot\cdot}) \left[\sum_{i=1}^{r} (X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X}_{\cdot\cdot})\right]$$

$$= \sum_{j=1}^{c} (\bar{X}_{\cdot j} - \bar{X}_{\cdot\cdot})\,(\bar{X}_{\cdot j} - \sum_{i=1}^{r} \bar{X}_{i\cdot} - r\bar{X}_{\cdot j} + r\bar{X}_{\cdot\cdot})$$

$$= \sum_{j=1}^{c} (\bar{X}_{\cdot j} - \bar{X}_{\cdot\cdot})\,(r\bar{X}_{\cdot j} - r\bar{X}_{\cdot\cdot} - r\bar{X}_{\cdot j} + r\bar{X}_{\cdot\cdot})$$

$$= \sum_{j=1}^{c} (\bar{X}_{\cdot j} - \bar{X}_{\cdot\cdot})\,(0) = 0.$$

(iii) $\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{h=1}^{n} (X_{ijh} - \bar{X}_{ij\cdot})\,(\bar{X}_{i\cdot\cdot} - \bar{X}_{\cdots})$

$$= \sum_{i=1}^{k}\sum_{j=1}^{k} (\bar{X}_{i..} - \bar{X}_{...}) \left[\sum_{h=1}^{n} (X_{ijh} - \bar{X}_{ij.})\right]$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{k} (\bar{X}_{i..} - \bar{X}_{...}) [n\bar{X}_{ij.} - n\bar{X}_{ij.}]$$

The terms within brackets become zero. Hence the result.

Q.20.7 (a) Computation for analysis of variance.

	Sample Number				Total
	1	2	3	4	
	11(121) 4 (16) 6 (36)	13(169) 9 (81) 14(196)	21(441) 18(324) 15(225)	10(100) 4 (16) 19(361)	
$T_{.j}$	21	36	54	33	144
$T_{.j}^2$	441	1296	2916	1089	5742
$\sum_i X_{ij}^2$	173	446	990	477	2086

$$\text{Total } SS = \sum_i\sum_j X_{ij}^2 - \frac{T_{..}^2}{n} = 2086 - \frac{(144)^2}{12}$$

$$= 2086 - 1728 = 358,$$

$$\text{Between Samples } SS = \sum_j \frac{T_j^2}{m} - \frac{T_{..}^2}{n} = \frac{5742}{3} - 1728$$

$$= 1914 - 1728 = 186, \text{ and}$$

Within Samples SS = Total SS − Between Samples SS

$= 358 - 186 = 172.$

Hence the analysis of variance table is

Source of variation	d.f.	Sum of Squares	Mean Square	F-ratio
Between Samples	3	186	62.0	2.88
Within Samples	8	172	21.5	--
Total	11	358	---	--

(b) (i) The hypotheses are stated as

$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$; and

H_1 : Not all the variances are equal.

(ii) We use a level of significance of $\alpha = 0.05$.

(iii) The test statistic (Bartlett's test) is

$$u = 2.3026 \frac{q}{c}, \text{ where}$$

$$q = (n-k) \log s_p^2 - \Sigma(n_i - 1) \log s_i^2, \text{ and}$$

$$c = 1 + \frac{1}{3(k-1)} \left(\Sigma \frac{1}{n_i - 1} - \frac{1}{n-k}\right)$$

(iv) Computations.

$$s_1^2 = \left(\frac{1}{n_1 - 1}\right) \left[\Sigma X_{1i}^2 - \frac{(\Sigma X_{1i})^2}{n_1}\right] = \frac{1}{2}\left[173 - \frac{(21)^2}{3}\right]$$

$$= \frac{1}{2}[173 - 147] = 13;$$

$$s_2^2 = \left(\frac{1}{n_2 - 1}\right) \left[\Sigma X_{2i}^2 - \frac{(\Sigma X_{2i})^2}{n_2}\right] = \frac{1}{2}\left[446 - \frac{(36)^2}{3}\right]$$

$$= \frac{1}{2}[446 - 432] = 7;$$

$$s_3^2 = \left(\frac{1}{n_3 - 1}\right) \left[\Sigma X_{3i}^2 - \frac{(\Sigma X_{3i})^2}{n_3}\right] = \frac{1}{2}\left[990 - \frac{(54)^2}{3}\right]$$

$$= \frac{1}{2}[990 - 972] = 9;$$

$$s_4^2 = \left(\frac{1}{n_4 - 1}\right) \left[\Sigma X_{4i}^2 - \frac{(\Sigma X_{4i})^2}{n_4}\right] = \frac{1}{2}\left[477 - \frac{(33)^2}{3}\right]$$

$$= \frac{1}{2}[477 - 363] = 57.$$

$$s_p^2 = \frac{\sum(n_i-1)\,s_i^2}{\sum(n_i-1)} = \frac{2(13+7+9+57)}{8} = \frac{172}{8} = 21.5,$$

$$\sum(n_i-1)\log s_i^2 = 2\,(\log 13 + \log 7 + \log 9 + \log 57)$$

$$= 2\,(1.1139 + 0.8451 + 0.9542 + 1.7559)$$

$$= 9.3382.$$

Now $q = (n-k)\log s_p^2 - \sum(n_i-1)\log s_i^2$

$= (8)\log 21.5 - 9.3382$

$= (8)\,(1.3324) - 9.3382 = 1.3210$, and

$$c = 1 + \frac{1}{3(4-1)}\left[\frac{1}{2}+\frac{1}{2}+\frac{1}{2}+\frac{1}{2}-\frac{1}{8}\right]$$

$$= 1 + \frac{1}{9}\left(\frac{15}{8}\right) = 1.2083$$

$$\therefore \quad u = 2.3026\left(\frac{1.3210}{1.2083}\right) = 2.517$$

(v) The critical region is $u \geq \chi^2_{0.05(3)} = 7.81$

(vi) **Conclusion.** Since the computed value of u dos not fall in the critical region, we therefore cannot reject H_0, and may conclude that the assumption of equal variances is satisfied.

Q.20.8. (a) (i) We set up our hypotheses as

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ and

H_1 : At least two of the means are not equal.

(ii) We use a significance level of $\alpha = 0.05$.

(iii) The test-statistic is $F = \dfrac{s_b^2}{s_w^2}$, which under H_0, has an F-distribution with $n_1 = 3$ and $n_2 = 16$ degrees of freedom.

(iv) Computations. Taking the origin at $X = 50$, the computations are carried out as follows:

	Machine Number				Total
	1	2	3	4	
	14(196)	-9(81)	15(225)	-5(25)	
	-11(121)	-2 (4)	7 (49)	1 (1)	
	15(225)	-9(81)	26(676)	5(25)	
	-4 (16)	-1 (1)	22(484)	-2 (4)	
	13(169)	7(49)	14(196)	-3 (9)	
$T_{\cdot j}$	27	−14	84	−4	93
T_j^2	729	196	7056	16	7997
$\sum_i X_{ij}^2$	727	216	1630	64	2637

$$\text{Total } SS = \sum_i\sum_j X_{ij}^2 - \frac{T_{..}^2}{n} = 2637 - \frac{(93)^2}{20}$$

$$= 2637 - 432.45 = 2204.55,$$

$$\text{Between Machines } SS = \sum_j \frac{T_j^2}{m} - \frac{T_{..}^2}{n} = \frac{7997}{5} - \frac{(93)^2}{20}$$

$$= 1599.4 - 432.45 = 1166.95, \text{ and}$$

Within Machines SS = Total SS − Between Machines SS

$= 2204.55 - 1166.95 = 1037.60.$

The ANOVA-Table for these results is

Source of variation	*d.f.*	Sum of Squares	Mean Square	*F*-ratio
Between Machines	3	1166.95	388.98	6.00
Within Machines	16	1037.60	64.85	--
Total	19	2204.55	---	--

(v) The critical region is $F > F_{0.05;(3,16)} = 3.24$

(vi) Since the computed value of F is larger than the corresponding critical value of F, so we reject H_0. The data provide sufficient evidence to conclude that the machines are significantly different with respect to items produced.

(b) **Assumptions.** In the above analysis, the following assumptions are involved:

(i) The four machine samples have been drawn randomly.

(ii) The populations being sampled are normally distributed with equal variances.

(iii) The effects are additive. That is, the model assumes that each observation X_{ij} is the sum of the true mean effect μ, the true effect of the treatment τ_j and the random error ε_{ij} associated with the observation.

Q.20.9 (a) (i) We are required to test the hypothesis

$H_0 : \mu_1 = \mu_2 = \mu_3$ against

H_1 : Not all three means are equal.

(ii) We use a significance level of $\alpha = 0.05$.

(iii) The test-statistic is $F = \dfrac{s_b^2}{s_w^2}$, which under H_0, has an F-distribution with $n_1 = k-1 = 2$, $n_2 = n-k = 6$ *d.f.*

(iv) Computations. Taking origin at $x = 160$, the computations proceed as follows:

	Salesmen			Total
	A	*B*	*C*	
	-8 (64) 15(225) 20(400)	21 (441) 11 (121) 43(1849)	0 (0) -30 (900) -36(1296)	
$T_{\cdot j}$	27	75	−66	36
T_j^2	729	5625	4356	10710
$\sum X_{ij}^2$	689	2411	2196	5296

$$\text{Total } SS = \sum_i \sum_j X_{ij}^2 - \frac{T_{..}^2}{n} = 5296 - \frac{(36)^2}{9}$$

$$= 5296 - 144 = 5152,$$

$$\text{Between Salesmen } SS = \sum_j \frac{T_j^2}{m} - \frac{T_{..}^2}{n}$$

$$= \frac{10710}{3} - \frac{(36)^2}{9} = 3570 - 144 = 3426, \text{ and}$$

Within Salesmen SS = Total SS − Between SS

$= 5152 - 3426 = 1726$

The ANOVA-Table for these results is

Source of variation	*d.f.*	Sum of Squares	Mean Square	*F*-ratio
Between Salesmen	2	3426	1713	5.95
Within Salesmen	6	1726	287.67	--
Total	8	5152	---	--

(v) The critical region is $F > F_{0.05;(2,6)} = 5.14$

(vi) Since the computed value of F falls in the critical region, so we reject H_0. The differences between salesmen are significant.

(b) (i) We wish to test the hypothesis

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ against

H_1 : Not all four means are equal.

(ii) We use a significance level of $\alpha = 0.05$.

(iii) The test-statistic is $F = \dfrac{s_b^2}{s_w^2}$, which under H_0, has an F-distribution with $n_1 = k-1$, $n_2 = n-k$ *d.f.*

(iv) Computations. Taking working origin at $X = 10$, the computations are as below:

	Treatments				Total
	1	2	3	4	
	1 (1) −6(36) −6(36) −5(25)	−4(16) −6(36) −7(49) −4(16)	−2 (4) −4(16) −6(36) 1 (1)	4 (16) 17(289) −2 (4) 8 (64)	
$T_{\cdot j}$	−16	−21	−11	27	−21
T_j^2	256	441	121	729	1547
$\sum X_{ij}^2$	98	117	57	373	645

$$\text{C.F.} = \frac{T_{..}^2}{n} = \frac{(-21)^2}{16} = \frac{441}{16} = 27.56$$

$$\text{Total } SS = \sum_i \sum_j X_{ij}^2 - \text{C.F.} = 645 - 27.56 = 617.44$$

$$\text{Between Treatments } SS = \sum_j \frac{T_j^2}{m} - \text{C.F.} = \frac{1547}{4} - 27.56$$

$$= 386.75 - 27.56 = 359.19$$

Within Treatments SS = Total SS − Between SS
= 617.44 − 359.19 = 258.25

The ANOVA-Table is set up as below:

Source of variation	*d.f.*	Sum of Squares	Mean Square	*F*-ratio
Between Treatments	3	359.19	119.73	5.56
Within Treatments	12	258.25	21.52	--
Total	15	617.44	---	--

(v) The critical region is $F > F_{0.05;(3,12)} = 3.49$.

(vi) The computed F is greater than the critical F at the 0.05 level of significance. We reject the hypothesis and conclude that there is sufficient evidence to indicate a difference in treatment means.

Q.20.10 Let μ_A, μ_B and μ_C denote the mean warm-up time in seconds for tube types A, B and C respectively. Then the hypotheses are stated as

(i) $H_0 : \mu_A = \mu_B = \mu_C$, and

H_1 : Not all three tube types have the same mean warm up time.

(ii) We use a significance level of $\alpha = 0.05$.

(iii) The test-statistic is $F = \dfrac{s_b^2}{s_{wp}^2}$, which under H_0, has an F-distribution with $n_1 = k-1$ and $n_2 = n-k$ *d.f.*

(iv) Computations.

	Tube-Types			Total
	A	*B*	*C*	
	19 (361)	20 (400)	16(256)	
	23 (529)	20 (400)	15(225)	
	26 (676)	32(1024)	18(324)	
	18 (324)	27 (729)	26(676)	
	20 (400)	40(1600)	19(361)	
	20 (400)	24 (576)	17(289)	
	18 (324)	22 (484)	19(361)	
	35(1225)	18 (324)	18(324)	
$T_{\cdot j}$	179	203	148	530
T_j^2	32041	41209	21904	95154
$\sum_i X_{ij}^2$	4239	5537	2816	12592

$$\text{Total } SS = \sum_i \sum_j X_{ij}^2 - \frac{T_{..}^2}{n} = 12592 - \frac{(530)^2}{24}$$

$$= 12592 - 11704.17 = 887.83,$$

$$\text{Between Tube Types } SS = \sum_j \frac{T_j^2}{m} - \frac{T_{..}^2}{n}$$

$$= \frac{95154}{8} - \frac{(530)^2}{24} = 11894.25 - 11704.17 = 190.08$$

Error SS = Total SS − Between Tube Types SS
= 887.83 − 190.08 = 697.75

The ANOVA-Table is set up below:

Source of variation	*d.f.*	Sum of Squares	Mean Square	*F*-ratio
Between Tube Types	2	190.08	95.04	2.86
Error	21	697.75	33.23	--
Total	23	887.83	---	--

(v) The critical region is $F > F_{0.05;(2,21)} = 3.47$

(vi) **Conclusion.** Since the computed value of F does not fall in the critical region, we therefore accept H_0 and may conclude that all the three types require the same average warm-up time.

Q.20.11 (i) We wish to test the hypothesis

(i) $H_0 : \mu_A = \mu_B = \mu_C = \mu_D = \mu_E$, against

H_1 : Not all μ_i's are equal.

(ii) We use a significance level of $\alpha = 0.05$.

(iii) The test-statistic is $F = \dfrac{s_b^2}{s_w^2}$, which under H_0, has an F-distribution with $n_1 = k-1$ and $n_2 = n-k$ degrees of freedom.

(iv) Computations. Taking origin at $X = 40$, the computations are carried out as below:

	Samples					Total
	A	B	C	D	E	
	−14 (196)	0 (0)	10(100)	−3 (9)	10(100)	
	0 (0)	−23(529)	25(625)	18(324)	−6 (36)	
	−3 (9)	−8 (64)	−2 (4)	−6 (36)	−4 (16)	
	−7 (49)	6 (36)	1 (1)	7 (49)	−13(169)	
	−21(441)	−16(256)	−4 (16)	5 (25)	1 (1)	
	12(144)	−12(144)	5 (25)	−12(144)	−6 (36)	
	−10(100)	−4 (16)	8 (64)	0 (0)	7 (49)	
	3 (9)	−15(225)	13(169)	7 (49)	−5 (25)	
$T_{\cdot j}$	−40	−72	56	16	−16	−56
T_j^2	1600	5184	3136	256	256	10432
$\sum_i X_{ij}^2$	948	1270	1004	636	432	4290

$$\text{Total } SS = \sum_i \sum_j X_{ij}^2 - \frac{T_{..}^2}{n} = 4290 - \frac{(-56)^2}{40}$$

$$= 4290 - 78.4 = 4211.6,$$

$$\text{Between Samples } SS = \sum_j \frac{T_j^2}{m} - \frac{T_{..}^2}{n} = \frac{10432}{8} - \frac{(-56)^2}{40}$$

$$= 1304 - 78.4 = 1225.6$$

With Samples SS = Total SS − Between Samples SS

$$= 4211.6 - 1225.6 = 2986.0$$

These results are displayed in the following ANOVA-Table:

Source of variation	$d.f.$	Sum of Squares	Mean Square	F-ratio
Between Samples	4	1225.6	306.4	3.59
Within Samples	35	2986.0	85.31	--
Total	39	4211.6	---	--

(v) The critical region is $F > F_{0.05;(4,35)} = 2.65$.

(vi) Since the computed value of F is greater than the corresponding critical value of F, so we reject H_0. There is evidence to indicate that the samples do not come from populations having the same means.

Q.20.12. (i) We are required to test the hypothesis

$H_0 : \mu_1 = \mu_2 = \mu_3$ against

H_1 : Not all three means are equal.

(ii) We use a significance level of $\alpha = 0.05$.

(iii) The test-statistic is $F = \dfrac{s_b^2}{s_w^2}$, which under H_0, has an F-distribution with $n_1 = k-1$, $n_2 = n-k$ $d.f.$

(iv) Computations. The necessary computations are given below:

	Groups			Total
	A	B	C	
	4 (16)	6 (36)	9 (81)	
	9 (81)	8 (64)	13(169)	
	10(100)	10(100)	15(225)	
	11(121)	11(121)	20(400)	
	17(289)	12(144)	23(529)	
	19(361)	12(144)		
		15(225)		
$T_{\cdot j}$	70	74	80	224
T_j^2	4900	5476	6400	--
$\sum_i X_{ij}^2$	968	834	1404	3206

$$\text{C.F.} = \frac{T_{..}^2}{n} = \frac{(224)^2}{18} = \frac{25088}{9} = 2787.56$$

$$\text{Total } SS = \sum_i \sum_j X_{ij}^2 - \text{C.F.} = 3206 - 2787.56 = 418.44$$

$$\text{Between Groups } SS = \sum_j \frac{T_j^2}{m_j} - \text{C.F.}$$

$$= \left[\frac{4900}{6} + \frac{5476}{7} + \frac{6400}{5}\right] - 2787.56$$

$$= (816.67 + 782.29 + 1280) - 2787.56$$

$$= 2878.96 - 2787.56 = 91.40$$

$$\text{Within Groups } SS = \text{Total } SS - \text{Between Groups } SS$$
$$= 418.44 - 91.40 = 327.04$$

The ANOVA-Table is set up below:

Source of variation	*d.f.*	Sum of Squares	Mean Square	*F*-ratio
Between Groups	2	91.40	45.70	2.10
Within Groups	15	327.24	21.80	--
Total	17	418.44	---	--

(v) The critical region is $F > F_{0.05;(2,15)} = 3.68$.

(vi) The computed value of F does not fall in the critical region. We accept the hypothesis H_0. There is sufficient evidence to conclude that means of Groups are equal.

Q.20.13. (i) We state our hypotheses as

$H_0 : \mu_1 = \mu_2 = \mu_3$ and

H_1 : Not all three means are equal.

(ii) We use a significance level of $\alpha = 0.05$.

(iii) The test-statistic is $F = \dfrac{s_b^2}{s_w^2}$, which under H_0, has an F-distribution with $n_1 = k-1$, $n_2 = n-k$ *d.f.*

(iv) Computations. Taking the origin at $X = 45$, the computations are carried out as below:

	Methods			Total
	1	2	3	
	2.2 (4.84) 4.8(23.04) 3.5(12.25) 3.7(13.69) -- ----	5.1(26.01) 4.3(18.49) 6.5(42.25) 5.9(34.81) -- ----	4.1(16.81) 8.2(67.24) 6.2(38.44) 7.8(60.84) 7.3(53.29)	
$T_{\cdot j}$	14.2	21.8	33.6	69.6
T_j^2	201.64	475.24	1128.96	--
$\sum_i X_{ij}^2$	53.82	121.56	236.62	412.00

$$\text{Total } SS = \sum_i \sum_j X_{ij}^2 - \frac{T_{..}^2}{n} = 412.00 - \frac{(69.6)^2}{13} = 412.00 - 372.63 = 39.37$$

$$\text{Between Methods } SS = \sum_j \frac{T_j^2}{m_j} - \frac{T_{..}^2}{n}$$

$$= \left(\frac{201.64}{4} + \frac{475.24}{4} + \frac{1128.96}{5}\right) - \frac{(69.6)^2}{13}$$

$$= 50.41 + 118.81 + 225.79 - 372.63 = 22.38,$$

and Within SS is obtained by subtractioin.

The ANOVA-Table for these results is set up below:

Source of variation	*d.f.*	Sum of Squares	Mean Square	*F*-ratio
Between Methods	2	22.38	11.19	6.58
Within Methods	10	16.99	1.70	--
Total	12	39.37	---	--

(v) The critical region is $F > F_{0.05;(2,10)} = 4.10$.

(vi) Since the computed value of F is greater than the corresponding critical F, we therefore reject H_0. The data provide

sufficient evidence to conclude that the Methods differ significantly.

Q.20.14. (i) We are required to test the hypothesis

$H_0 : \mu_1 = \mu_2 = \mu_3$ against H_1:Not all three means are equal.

(ii) The level of significance is set at $\alpha = 0.05$.

(iii) The test-statistic is $F = \dfrac{s_b^2}{s_w^2}$, which under H_0, has an F-distribution with $n_1 = 2$ and $n_2 = 29$ degrees of freedom.

(iv) Computations. Taking the origin at $X = 70$, the computations proceed as below:

	Teacher			
	A	B	C	Total
	5 (25)	−11 (121)	−4 (16)	
	21 (441)	13 (169)	7 (49)	
	14 (196)	29 (841)	−19(361)	
	−25 (625)	7 (49)	20(400)	
	12 (144)	−5 (25)	3 (9)	
	5 (25)	11 (121)	20(400)	
	−2 (4)	−36(1296)	1 (1)	
	−23 (529)	11 (121)	−2 (4)	
	25 (625)	7 (49)	−1 (1)	
	−32(1024)	18 (324)		
		24 (576)		
		−19 (361)		
		12 (144)		
$T_{.j}$	0	61	25	86
$\sum_i X_{ij}^2$	3638	4197	1241	9076

$$\text{Total } SS = \sum_i \sum_j X_{ij}^2 - \frac{T_{..}^2}{n} = 9076 - \frac{(86)^2}{32} = 9076 - 231.12 = 8844.88$$

$$\text{Between Teachers } SS = \sum_j \frac{T_j^2}{m_j} - \frac{T_{..}^2}{n}$$

$$= \left[\frac{(0)^2}{10} + \frac{(61)^2}{13} + \frac{(25)^2}{9}\right] - \frac{(86)^2}{32}$$

$$= (0 + 286.23 + 69.44) - 231.12 = 124.55, \text{ and}$$

Within Teachers SS = Total SS − Between Teacher SS

$$= 8844.88 - 124.55 = 8720.33$$

These results are displayed in the following ANOVA-Table.

Source of variation	*d.f.*	Sum of Squares	Mean Square	*F*-ratio
Between Teachers	2	124.55	62.28	0.21
Within Teachers	29	8720.33	300.70	--
Total	31	8844.88	---	--

(v) The critical region is $F > F_{0.05;(2,29)} = 3.33$

(vi) Since the computed value of F is greater than the corresponding critical F, we accept H_0. Hence there is not sufficient evidence to indicate a difference in the average grades given by the three teachers.

Q.20.15. (a) To test for equality of variances, we use the Bartlett's chi-square test, given by

$$u = \frac{2.3026\,(v \log s_p^2 - \sum v_i \log s_i^2)}{1 + \dfrac{1}{3(k-1)}\left[\sum \dfrac{1}{v_i} - \dfrac{1}{v}\right]}$$

which is approximated by a chi-square distribution with (k−1) d.f. (k = number of samples).

	Samples								Total
	A		B		C		D		
	X_{i1}	X_{i1}^2	X_{i2}	X_{i2}^2	X_{i3}	X_{i3}^2	X_{i4}	X_{i4}^2	
	49	2401	49	2401	44	1936	58	3364	
	42	1764	44	1936	57	3249	54	2916	
	47	2209	50	2500	34	1156	64	4096	
	76	5776	58	3364	48	2304	60	3600	
	69	4761	70	4900	50	2500	53	2809	
	58	3364	--	--	--	--	64	4096	
	--	--	--	--	--	--	52	2704	
	--	--	--	--	--	--	42	1764	
$T_{\cdot j}$	341	--	271	--	233	--	447	--	1292
$\sum_i X_{ij}^2$	--	20275	--	15101	--	11145	--	25349	71870
$(\Sigma X_{ij})^2/m_j$	19380.17		14688.2		10857.8		24976.12		69902.29
$X_{ij}^2 - (\Sigma X_{ij})^2/m_j$	894.38		412.8		287.2		372.88		1967.71
s_i^2	178.97		103.2		71.8		53.27		$s_p^2 = 98.39$

$\therefore$ Numerator $= 2.3026[20 \log 98.39 - (5 \log 178.97 + 4 \log 103.2 + 4 \log 71.8 + 7 \log 53.27)]$

$= 2.3026\,[20(1.9930) - \{5(2.2528) + 4(2.0137) + 4(1.8561) + 7(1.7265)\}]$

$= 2.3026[39.86 - (11.264 + 8.0548 + 7.4244 + 12.0855)]$

$= 2.3026(1.0313) = 2.3747$, and

$$\text{Denominator} = 1 + \frac{1}{3(4-1)}\left[\frac{1}{5} + \frac{1}{4} + \frac{1}{4} + \frac{1}{7} - \frac{1}{20}\right]$$

$$= 1 + \frac{1}{9}\left[\frac{111}{140}\right] = 1 + 0.0881 = 1.0881.$$

Hence $u = \dfrac{2.3747}{1.0881} = 2.18$

Since $\chi^2_{0.05,(3)} = 7.82$ is greater than $u = 2.18$, we therefore conclude that the variances are homogeneous.

(b) To test for equality of means by the analysis of variance, we proceed as below:

(i) We state the hypotheses as

$H_0 : \mu_A = \mu_B = \mu_C = \mu_D$, and H_1 : Not all μ_i's are equal.

(ii) We use a significance level of $\alpha = 0.05$.

(iii) The test-statistic is $F = \dfrac{s_b^2}{s_w^2}$, which under H_0, has an F-distribution with $n_1 = 3$, $n_2 = 20$ degrees of freedom. This statistic depends upon the following assumptions.

(a) the four samples have been randomly selected.

(b) the populations sampled are normally distributed with equal variances σ^2 and means μ_A, μ_B, μ_C, μ_D.

(c) All the effects are additive.

(iv) Computations.

$$\text{Total } SS = \sum_i \sum_j X_{ij}^2 - \frac{T_{..}^2}{n} = 71870 - \frac{(1292)^2}{24}$$

$$= 71870 - 69552.67 = 2317.33,$$

$$\text{Between Samples } SS = \sum_j \frac{T_j^2}{m_j} - \frac{T_{..}^2}{n} = 69902.29 - 69552.67$$

$$= 349.62, \text{ and}$$

Within Samples SS = Total SS – Between Samples SS

$$= 2317.33 - 349.62 = 1967.71$$

The analysis of variance table is then set up as below:

Source of variation	$d.f.$	Sum of Squares	Mean Square	F-ratio
Between Samples	3	349.62	116.54	1.18
Within Samples	20	1967.71	98.39	--
Total	23	2317.33	---	--

(v) The critical region is $F > F_{0.05;(3,20)} = 3.10$

(vi) Since the computed value of F does not fall in the critical region, so we accept H_0. Hence the data provide sufficient evidence to indicate that the four means are almost certainly equal.

Q.20.16 (b) (i) We state our hypotheses as

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$, and

H_1 : Not all four means are equal.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic to use is

$$F = \frac{s_b^2}{s_w^2},$$

which, if H_0 is true, has an F-distribution with $v_1 = k-1$ and $v_2 = n-k$ degrees of freedom.

(iv) Computations. To compute the necessary sums of squares, we have to first compute the grand mean $\bar{X}..$ as

$$\bar{X}.. = \frac{n_1\bar{X}_1 + n_2\bar{X}_2 + n_3\bar{X}_3 + n_4\bar{X}_4}{n_1 + n_2 + n_3 + n_4}$$

$$= \frac{4(58)+6(57)+7(43)+3(42)}{4+6+7+3} = \frac{1001}{20} = 50.05$$

Now Between $SS = \sum_j n_j (\bar{X}_{.j} - \bar{X}..)^2$

$= [4(58-50.05)^2 + 6(57-50.05)^2$
$+7(43-50.05)^2+3(42-50.05)^2]$

$= 252.81+289.815+347.9175+194.4075$

$= 1084.95$, and

Within $SS = (n_1-1)\, s_1^2+(n_2-1)\, s_2^2 + (n_3-1)\, s_3^2 + (n_4-1)\, s_4^2$

$= 3(10) + 5(30.4) + 6(5.67) + 2(9) = 234.02$

Then the ANOVA Table becomes

Source of variation	d.f.	Sum of Squares	Mean Square	F-ratio
Between Samples	3	1084.95	361.65	24.72
Within Samples	16	234.02	14.63	--
Total	19	1318.97	---	--

(v) The critical region is $F \geq F_{0.05(3,16)} = 3.24$.

(vi) **Conclusion.** Since the computed value of $F = 3.24$ falls in the critical region, we therefore reject H_0.

Q.20.17. (a) Let μ_N, μ_O and μ_A denote the mean rubber content of normal, offtype and aberrant type plants respectively. Then the hypotheses are stated as

(i) H_0 : $\mu_N = \mu_O = \mu_A$, *i.e.* there is no difference in mean rubber content for the three types of plants; and

H_1 : Not all three means are equal.

(ii) The level of significance is set at $\alpha = 0.01$

(iii) The test-statistic is $F = \dfrac{\text{Between Mean Square}}{\text{Within Mean Square}}$,

which under H_0, is distributed as F-distribution with $n_1 = k-1$ and $n_2 = n-k$ *d.f.*

(iv) Computations. Let X_N, X_O and X_A denote the normal, offtype and aberrant measurements respectively. Then

$\Sigma X_N = 183.88$, $\Sigma X_N^2 = 1258.3610$, $\bar{X}_N = 6.8104$,

$\Sigma X_O = 84.25$, $\Sigma X_O^2 = 478.9779$, $\bar{X}_O = 5.6167$,

$\Sigma X_A = 80.92$, $\Sigma X_A^2 = 561.6402$, $\bar{X}_A = 6.7433$,

$T.. = \Sigma X_N + \Sigma X_O + \Sigma X_A = 349.05$

$\therefore$ Total $SS = \sum X_{ij}^2 - \dfrac{T..^2}{n} = (\Sigma X_N^2 + \Sigma X_O^2 + \Sigma X_A^2) - \dfrac{T..^2}{n}$

$$= 2298.9791 - \frac{(349.05)^2}{54}$$

$$= 2298.9791 - 2256.2204 = 42.7587,$$

$$\text{Between } SS = \Sigma \frac{T_j^2}{n_j} - \frac{T_{..}^2}{n}$$

$$= \frac{(183.88)^2}{27} + \frac{(84.25)^2}{15} + \frac{(80.92)^2}{12} - \frac{(349.05)^2}{54}$$

$$= 1252.2909 + 473.2042 + 545.6705 - 2256.2204$$

$$= 2271.1656 - 2256.2204 = 14.9452, \text{ and}$$

Error SS is obtained by difference.

Then the ANOVA Table is as below:

Source of variation	*d.f.*	Sum of Squares	Mean Square	*F*-ratio
Between Types of plants	2	14.9452	7.4726	13.70
Error	51	27.8135	0.5454	--
Total	53	42.7587	---	--

(v) The critical region is $F \geq F_{0.01(2,51)} = 5.07$.

(vi) **Conclusion.** Since the computed value of F falls in the critical region, we therefore reject H_0 and may conclude that the rubber content of the three types of plants is different.

(b) (i) To test $H_0 : \mu_N = \mu_A$, we calculate the value of Student's t as below:

$$\text{Under } H_0,\ t = \frac{\bar{X}_N - \bar{X}_A}{\sqrt{s_e^2 \left(\frac{1}{27} + \frac{1}{12}\right)}} = \frac{6.8104 - 6.7433}{\sqrt{0.5454 \left(\frac{39}{324}\right)}}$$

$$= \frac{0.0671}{0.2562} = 0.26$$

This value of t is clearly less than table value with 51 *d.f.* so we accept H_0.

(ii) To test $H_0 : \mu_0 = \frac{1}{2}(\mu_N + \mu_A)$, we calculate t value under H_0, as

$$t = \frac{\bar{X}_0 - \frac{1}{2}(\bar{X}_N + \bar{X}_A)}{\sqrt{s_e^2 \left[\frac{1}{15} + \frac{1}{4}\left(\frac{1}{27} + \frac{1}{12}\right)\right]}}$$

$$= \frac{5.6167 - (6.8104 + 6.7433)/2}{\sqrt{(0.5454)\ (1881)/19440}} = \frac{-1.1602}{\sqrt{0.0527725}}$$

$$= \frac{-1.1602}{0.2297} = -5.05 \text{ with } 51 \text{ } d.f.$$

The calculated value is highly significant and we reject H_0.

Q.20.18. (i) The hypotheses would be stated as

$H_0 : \mu_1 = \mu_2 = \mu_3$, *i.e.* there is no difference between the means of students in arithmetic computation in different type of schools.

H_1 : Not all three means are equal.

(ii) We use a significance level of $\alpha = 0.05$.

(iii) The test-statistic is $F = \dfrac{\text{Between Mean Square}}{\text{Within Mean Square}}$, which under H_0, is distributed as F-distribution with $n_1 = k-1$, $n_2 = n-k$ *d.f.*

(iv) Computations.

Mid values x_i	Residential			Non-Residential			Mission		
	f_1	$f_1 x_i$	$f_1 x_i^2$	f_2	$f_2 x_i$	$f_2 x_i^2$	f_3	$f_3 x_i$	$f_3 x_i^2$
4.5	1	4.5	20.25	4	18.0	81.00	--	--	--
14.5	3	43.5	630.75	7	101.5	1471.75	4	58.0	841.00
24.5	10	245.0	6002.50	25	612.5	15006.25	8	196.0	4802.00
34.5	63	2173.5	74985.75	37	1276.5	44039.25	17	586.5	20234.25
44.5	38	1691.0	75249.50	13	578.5	25743.25	29	1290.5	57427.25
54.5	5	272.5	14851.25	4	218.0	11881.00	7	381.5	20791.75
Σ	120	4430.0	171740.00	90	2805.0	98222.50	65	2512.5	104096.25

The sums and sums of squares of scores obtained from each type of school, are tabulated below to perform further computations:

Type of School	No. of Students (Σf)	Sums of Scores (Σfx)	SS of Scores (Σfx^2)
Residential	120	4430.0	171740.00
Non-Resid.	90	2805.0	98222.50
Mission	65	2512.5	104096.25
Total	275	9747.5	374058.75

$$\text{C.F.} = \frac{(9747.5)^2}{275} = \frac{95013756.25}{275} = 345504.57$$

$$\text{Total } SS = \sum_i \sum_j x_{ij}^2 - \text{C.F.} = 374058.75 - 345504.57 = 28554.18$$

$$\text{Between School } SS = \sum_j \frac{T_j^2}{m_j} - \text{C.F.}$$

$$= \frac{(4430)^2}{120} + \frac{(2805)^2}{90} + \frac{(2512.5)^2}{65} - \text{C.F.}$$

$$= 163540.83 + 87422.50 + 97117.79 - 345504.57$$

$$= 2576.55$$

Within School SS is obtained by subtraction.

Then the ANOVA Table is setup below:

Source of variation	$d.f.$	Sum of Squares	Mean Square	F-ratio
Between Schools	2	2,576.55	1288.275	13.49
Within Schools	272	25,977.63	95.51	--
Total	274	28,554.18	---	--

(v) The critical region is $F \geq F_{.05(2,272)} = 3.00$

(vi) As the computed value of F is greater than the critical value, so we reject H_0 and conclude that the difference between the means of students in the different types of schools is significant.

Q.20.19. (i) We wish to test the hypotheses

H_0' : $\mu_{\cdot 1} = \mu_{\cdot 2} = \mu_{\cdot 3} = \mu_{\cdot 4}$

H_0'' : $\mu_{A\cdot} = \mu_{B\cdot} = \mu_{C\cdot}$. against

H_1' : Not all four group means are equal.

H_1'' : Not all three ration means are equal.

(ii) We use a significance level of $\alpha = 0.05$.

(iii) The test-statistics are $F_1 = \frac{s_1^2}{s_3^2}$, $F_2 = \frac{s_2^2}{s_3^2}$, which under the null hypotheses have F-distributions with $n_1 = 3$, $n_2 = 6$ and $n_1 = 2$, $n_2 = 6$ $d.f.$

(iv) The necessary computations are given below:

Ration	Groups I	Groups II	Groups III	Groups IV	$T_{i\cdot}$	$T_{i\cdot}^2$
A	7.0(49.00)	16.0(256.00)	10.5(110.25)	13.5(182.25)	47.0	2209.00
B	14.0(196.00)	15.5(240.25)	15.0(225.00)	21.0(441.00)	65.5	4290.25
C	8.5 (72.25)	16.5(272.25)	9.5(90.25)	13.5(182.25)	48.0	2304.00
$T_{\cdot j}$	29.5	48.0	35.0	48.0	160.5	8803.25
$T_{\cdot j}^2$	870.25	2304.00	1225.00	2304.00	6703.25	
$\sum_i X_{ij}^2$	317.25	768.50	425.50	805.50	2316.75	

$$\text{C.F.} = \frac{T_{\cdot\cdot}^2}{cr} = \frac{(160.5)^2}{12} = \frac{25760.25}{12} = 2146.67$$

$$\text{Total } SS = \sum_i \sum_j X_{ij}^2 - \text{C.F.} = 2316.75 - 2146.67 = 170.08$$

$$\text{Between Groups } SS = \sum_j \frac{T_j^2}{r} - \text{C.F.} = \frac{6703.25}{3} - \text{C.F.}$$

$$= 2234.42 - 2146.67 = 87.75$$

Between Rations $SS = \sum_i \frac{T_{i.}^2}{c} - \text{C.F.} = \frac{8803.25}{4} - \text{C.F.}$

$= 2200.81 - 2146.67 = 54.14$

Error SS = Total SS – (Group SS + Ration SS)

$= 170.08 - (87.75 + 54.14) = 28.19$

The ANOVA-Table is set up below:

Source of variation	*d.f.*	Sum of Squares	Mean Square	*F*-ratio
Between Groups	3	87.75	$s_1^2 = 29.25$	6.22
Between Rations	2	54.14	$s_2^2 = 27.07$	5.76
Error	6	28.19	$s_3^2 = 4.70$	--
Total	1	170.08	---	--

(v) The critical regions are $F_1 \geq F_{0.05;(3,6)} = 4.76$

$F_2 \geq F_{0.05;(2,6)} = 5.14$

(vi) There is significant difference between Rations at 5% level as the computed value of F is greater than the corresponding critical value of F.

Q.20.20. (i) We wish to test the hypotheses

H_0' : $\mu_{.1} = \mu_{.2} = \mu_{.3} = \mu_{.4}$,

H_0'' : $\mu_{1.} = \mu_{2.} = \mu_{3.} = \mu_{4.}$ against

H_1' : Not all $\mu_{.j}$ are equal.

H_1'' : Not all $\mu_{i.}$ are equal.

(ii) We use a significance level of $\alpha = 0.05$.

(iii) The test-statistics are $F_1 = \frac{s_1^2}{s_3^2}$, $F_2 = \frac{s_2^2}{s_3^2}$, which under null hypotheses, have F-distributions with $n_1 = 4$, $n_2 = 12$ and $n_1 = 3$, $n_2 = 12$ *d.f.* respectively.

(iv) Computations. Taking the origin at $X = 30$, the computations are carried out as below:

Factor A	Factor B 1	2	3	4	$T_{i.}$	$T_{i.}^2$
1	-15(225)	1 (1)	-10(100)	0 (0)	-24	576
2	-8 (64)	-19(361)	15(225)	-4 (16)	-16	256
3	3 (9)	7 (49)	0 (0)	14(196)	24	576
4	-12(144)	1 (1)	19(361)	4 (16)	12	144
5	7 (49)	0 (0)	6 (36)	-9 (81)	4	16
$T_{.j}$	-25	-10	30	5	0	1568
$T_{.j}^2$	625	100	900	25	1650	---
$\sum_i X_{ij}^2$	491	412	722	309	1934	--

$$\text{C.F.} = \frac{T_{..}^2}{cr} = \frac{(0)^2}{20} = 0$$

Total $SS = \sum_i \sum_j X_{ij}^2 - \text{C.F.} = 1934 - 0 = 1934$

Factor A $SS = \sum_i \frac{T_{i.}^2}{c} - \text{C.F.} = \frac{1568}{4} - 0 = 392$

Factor B $SS = \sum_j \frac{T_{.j}^2}{r} - \text{C.F.} = \frac{1650}{5} - 0 = 330$

Error SS = Total SS – (Factor A SS + Factor B SS)

$= 1934 - (392 + 330) = 1212$

The ANOVA-Table is set up below:

Source of variation	*d.f.*	Sum of Squares	Mean Square	*F*-ratio
Factor A	4	392	$s_1^2 = 98$	0.97
Factor B	3	330	$s_2^2 = 110$	1.09
Error	12	1212	$s_3^2 = 101$	--
Total	19	1934	---	--

(v) The critical regions are (a) $F \geq F_{0.05;(4,12)} = 3.26$

(b) $F \geq F_{0.05;(3,12)} = 3.49$

(vi) The computed values of F are less than the corresponding critical values of F. Hence both the hypotheses are accepted.

Q.20.21. (i) We are required to test the hypotheses

H_0': $\mu_{.1} = \mu_{.2} = \mu_{.3} = \mu_{.4}$ against H_1' : Not all $\mu_{.j}$ are equal.

H_0'': $\mu_{1.} = \mu_{2.} = \mu_{3.} = \mu_{4.}$ against H_1'' : Not all $\mu_{i.}$ are equal.

(ii) We use a significance level of $\alpha = 0.05$.

(iii) The test-statistics are $F_1 = \dfrac{s_1^2}{s_3^2}$, $F_2 = \dfrac{s_2^2}{s_3^2}$, which under null hypotheses, have F-distributions with $n_1 = 3$, $n_2 = 6$ and $n_1 = 2$, $n_2 = 6$ *d.f.*

(iv) Computations. Taking the origin at $X = 40$, the computations are carried out as below:

Ration	Breeds of Cattle				$T_{i.}$	$T_{i.}^2$
	B_1	B_2	B_3	B_4		
R_1	6.5 (42.25)	22.0(484.00)	1.0 (1.00)	5.0 (25.00)	34.5	1190.25
R_2	7.5 (56.25)	1.5 (2.25)	- 18.0(324.00)	-8.5(72.25)	-17.5	306.25
R_3	10.0(100.00)	0 (0)	- 14.5(210.25)	-11.5(132.25)	-16.0	256.00
$T_{.j}$	24.0	23.5	-31.5	-15.0	1.0	1752.50
$T_{.j}^2$	576.00	552.25	992.25	225.00	2345.50	---
$\sum_i X_{ij}^2$	198.50	486.25	535.25	229.50	1449.50	--

$$\text{C.F.} = \frac{T_{..}^2}{cr} = \frac{(1)^2}{12} = 0.08$$

$$\text{Total } SS = \sum_i \sum_j X_{ij}^2 - \text{C.F.} = 1449.50 - 0.08 = 1449.42$$

$$\text{Between Breeds } SS = \sum_j \frac{T_{.j}^2}{r} - \text{C.F.} = \frac{2345.50}{3} - \text{C.F.}$$

$$= 781.83 - 0.08 = 781.75$$

$$\text{Between Rations } SS = \sum_i \frac{T_{i.}^2}{c} - \text{C.F.} = \frac{1752.50}{4} - \text{C.F.}$$

$$= 438.12 - 0.08 = 438.04$$

Error SS = Total SS – (Breeds SS + Rations SS)

= 1449.42 – (781.75 + 438.04)

= 1449.42 – 1219.79 = 229.63

The ANOVA-Table is set up below:

Source of variation	*d.f.*	SS	MS	*F*-ratio
Between Breeds	3	781.75	260.58	$F_1 = 6.81$
Between Rations	2	438.04	219.02	$F_2 = 5.72$
Error	6	229.63	38.27	--
Total	11	1449.42	---	--

(v) The critical regions are $F_1 \geq F_{0.05;(3,6)} = 4.76$

$F_2 \geq F_{0.05;(2,6)} = 5.14$

(vi) There is evidence of significant difference both in breeds and between rations as the computed values of F-ratios in both the cases fall in the critical regions.

Q.20.22. (i) We wish to test the hypotheses

H_0': $\mu_{.1} = \mu_{.2} = \mu_{.3} = \mu_{.4} = \mu_{.5}$ against H_1': Not all $\mu_{.j}$ are equal.

H_0'': $\mu_{1\cdot} = \mu_{2\cdot} = \mu_{3\cdot} = \mu_{4\cdot}$ against H_1'' : Not all $\mu_{i\cdot}$ are equal.

(ii) We use a significance level of $\alpha = 0.05$.

(iii) The test-statistics are $F_1 = \dfrac{s_1^2}{s_3^2}$, $F_2 = \dfrac{s_2^2}{s_3^2}$, which under null hypotheses, have F-distributions with $n_1 = 4$, $n_2 = 12$ and $n_1 = 3$, $n_2 = 12$ *d.f.* respectively.

(iv) Computations. Multiplying each observation by 10 and then taking the origin at $X = 20$, the computations are carried out as below:

Variety	Fertilizer 1	2	3	4	5	$T_{i\cdot}$	$T_{i\cdot}^2$
1	−1 (1)	2(4)	6(36)	−2 (4)	1 (1)	6	36
2	5(25)	−1(1)	3 (9)	6(36)	2 (4)	15	225
3	−3 (9)	−1(1)	2 (4)	0 (0)	1 (1)	−1	1
4	1 (1)	−2(4)	5(25)	3 (9)	4(16)	11	121
$T_{\cdot j}$	2	−2	16	7	8	31	383
T_j^2	4	4	256	49	64	377	---
$\sum_i X_{ij}^2$	36	10	74	49	22	191	--

$$\text{C.F.} = \frac{T_{\cdot\cdot}^2}{cr} = \frac{(31)^2}{20} = \frac{961}{20} = 48.05$$

$$\text{Total } SS = \sum_i \sum_j X_{ij}^2 - \text{C.F.} = 191.00 - 48.05 = 142.95$$

$$\text{Between Fertilizers } SS = \sum_j \frac{T_j^2}{r} - \text{C.F} = \frac{377}{4} - 48.05$$

$$= 94.25 - 48.05 = 46.20$$

$$\text{Between Varieties } SS = \sum_i \frac{T_{i\cdot}^2}{c} - \text{C.F.} = \frac{383}{5} - 48.05$$

$$= 76.60 - 48.05 = 28.55$$

$$\text{Error } SS = \text{Total } SS - (\text{Fertilizers } SS + \text{Varieties } SS)$$

$$= 142.95 - (46.20 + 28.55)$$

$$= 142.95 - 74.75 = 68.20$$

The ANOVA-Table is set up below:

Source of variation	*d.f.*	SS	MS	*F*-ratios
Between Fertilizers	4	46.20	$s_1^2 = 11.55$	$F_1 = 2.03$
Between Varieties	3	28.55	$s_2^2 = 9.52$	$F_2 = 1.68$
Error	12	68.20	$s_3^2 = 5.68$	--
Total	19	142.95	---	--

(v) The critical regions are $F_1 \geq F_{0.05;(4,12)} = 3.26$

$$F_2 \geq F_{0.05;(3,12)} = 3.49$$

(vi) The computed values of F are less than the corresponding critical values of F, we accept both the hypotheses and may conclude that data could have arisen from a population in which there was no difference between the yields of varieties and the fertilizers did not differ in their effect.

Q.20.23. (i) The hypotheses are stated as

H_0 : $\mu_{1\cdot} = \mu_{2\cdot} = \mu_{3\cdot} = \mu_{4\cdot}$, i.e. there is no difference between the treatment means, and

H_1 : Not all means $\mu_{i\cdot}$ are equal.

(ii) The level of significance is set at $\alpha = 0.05$.

(iii) The test-statistic is $F = \dfrac{MS\ (\text{Treatments})}{MS\ (\text{Error})}$, which under H_0, has the F-distribution with $n_1 = 3$ and $n_2 = 12$ *d.f.*

(iv) Computations.

Treatment	Subject 1	2	3	4	5	$T_{i.}$	$T_{i.}^2$
1	12.8 (163.84)	10.6 (112.36)	11.7 (136.89)	10.7 (114.49)	11.0 (121.00)	56.8	3226.24
2	11.7 (136.89)	14.2 (201.64)	11.8 (139.24)	9.9 (98.01)	13.8 (190.44)	61.4	3769.96
3	11.5 (132.25)	14.7 (216.09)	13.6 (184.96)	10.7 (114.49)	15.9 (252.81)	66.4	4408.96
4	12.6 (158.76)	16.5 (272.25)	15.4 (237.16)	9.6 (92.16)	17.1 (292.41)	71.2	5069.44
$T_{.j}$	48.6	56.0	52.5	40.9	57.8	255.8	16474.60
$T_{.j}^2$	2361.96	3136.00	2756.25	1672.81	3340.84	13267.86	---
$\sum_i X_{ij}^2$	591.74	802.34	698.25	419.15	856.66	3368.14	--

$$\text{C.F.} = \frac{T_{..}^2}{n} = \frac{(255.8)^2}{20} = 3271.68$$

$$\text{Total } SS = \sum_i \sum_j X_{ij}^2 - \text{C.F.} = 3368.14 - 3271.68 = 94.46$$

$$\text{Subjects } SS = \sum_j \frac{T_j^2}{r} - \text{C.F.} = \frac{13267.84}{4} - 3271.68 = 45.28,$$

$$\text{Treatments } SS = \sum_i \frac{T_{i.}^2}{c} - \text{C.F.} = \frac{16474.6}{5} - 3271.68 = 23.24$$

Error SS is obtained by difference.

The analysis of variance is as follows:

Source of variation	d.f.	SS	MS	F-ratios
Between Subjects	4	45.28	11.32	--
Between Treatments	3	23.24	7.75	3.33
Error	12	27.94	2.33	--
Total	19	96.46	---	--

(v) The critical region is $F \geq F_{0.05;(3,12)} = 3.49$

(vi) **Conclusion.** Since the calculated value of F does not fall in the critical region, we therefore cannot reject H_0. We may conclude that there is no difference among the treatment means.

Q.20.24. (i) We wish to test the hypotheses

H_0' : $\mu_{.1} = \mu_{.2} = \mu_{.3} = \mu_{.4}$, against H_1':Not all $\mu_{.j}$ are equal.

H_0'' : $\mu_{1.} = \mu_{2.} = \mu_{3.}$ against H_1'':Not all $\mu_{i.}$ are equal.

(ii) We use a significance level of $\alpha = 0.05$

(iii) The test-statistics are $F_1 = \frac{s_1^2}{s_3^2}$, $F_2 = \frac{s_2^2}{s_3^2}$, which under the null hypotheses, have the F-distributioins with $n_1=3$, $n_2=6$ and $n_1=2$, $n_2=6$ *d.f.* respectively.

(iv) Computations. Taking origin at $X=60$, the computations proceed as below:

Districts	Salesmen A	B	C	D	$T_{i.}$	$T_{i.}^2$
K	-30 (900)	10 (100)	-30 (900)	-30 (900)	-80	6400
O	20 (400)	-10 (100)	-20 (400)	10 (100)	0	0
S	40 (1600)	0 (0)	20 (400)	20 (400)	80	6400
$T_{.j}$	30	0	-30	0	0	12800
$T_{.j}^2$	900	0	900	0	1800	---
$\sum_i X_{ij}^2$	2900	200	1700	1400	6200	--

$$\text{C.F.} = \frac{T_{..}^2}{cr} = \frac{(0)^2}{12} = 0$$

$$\text{Total } SS = \sum_i \sum_j X_{ij}^2 - \text{C.F.} = 6200$$

$$\text{Between Salesmen } SS = \sum_j \frac{T_{.j}^2}{r} - \text{C.F.} = \frac{1800}{3} = 600$$

$$\text{Between Districts } SS = \sum_i \frac{T_{i.}^2}{c} - \text{C.F.} = \frac{12800}{4} = 3200$$

Error SS = Total SS – (Between Salesmen SS + Between Districts SS)

$$= 6200 - (600 + 3200) = 2400$$

These results are shown in the following ANOVA-Table.

Source of variation	*d.f.*	SS	MS	*F*-ratios
Between Salesmen	3	600	200	$F_1 = 0.5$
Between Districts	2	3200	1600	$F_2 = 4.0$
Error	6	2400	$s_3^2 = 400$	--
Total	11	6200	---	--

(v) The critical regions are $F_1 \geq F_{0.05;(3,6)} = 4.76$,

$$F_2 \geq F_{0.05;(2,6)} = 5.14.$$

(vi) The computed values of F are less than the corresponding critical values of F at 5% level of significance. We therefore accept the hypotheses. There is sufficient evidence to conclude that the salesmen were equally capable and that all districts were equally profitable to work.

Q.20.25. (i) The hypotheses would be stated as

H_0' : The courses are of equal difficulty, *i.e.* $\mu_{.1} = \mu_{.2} = \mu_{.3}$

H_0'': The students have equal ability, *i.e.*

$$\mu_{1.} = \mu_{2.} = \mu_{3.} = \mu_{4.} = \mu_{5.}$$

(ii) We use a significance level of $\alpha = 0.05$.

(iii) The test-statistics are $F_1 = \dfrac{s_1^2}{s_3^2}$, $F_2 = \dfrac{s_2^2}{s_3^2}$, which under the null hypotheses, follow F-distributioins with $n_1 = 2$, $n_2 = 8$ and $n_1 = 4$, $n_2 = 8$ *d.f.* respectively.

(iv) Computations. Taking origin at $X = 70$, the computations are carried out as below:

Students	Subjects			$T_{i.}$	$T_{i.}^2$
	English	Statistics	Economics		
1	-23 (529)	3 (9)	-9 (81)	-29	841
2	10 (100)	18 (324)	16 (256)	44	1936
3	1 (1)	7 (49)	-11 (121)	-3	9
4	-8 (64)	-10 (100)	-4 (16)	-22	484
5	-14(196)	25 (625)	17 (289)	28	784
$T_{.j}$	-34	43	9	18	4054
$T_{.j}^2$	1156	1849	81	3086	---
$\sum_i X_{ij}^2$	890	1107	763	2760	--

$$\text{C.F.} = \frac{T_{..}^2}{cr} = \frac{(18)^2}{15} = \frac{324}{15} = 21.6$$

$$\text{Total } SS = \sum_i \sum_j X_{ij}^2 - \text{C.F.} = 2760 - 21.6 = 2738.4$$

$$\text{Between Subjects } SS = \sum_j \frac{T_j^2}{r} - \text{C.F} = \frac{3086}{5} - 21.6$$

$$= 617.2 - 21.6 = 595.6$$

$$\text{Between Students } SS = \sum_i \frac{T_{i.}^2}{c} - \text{C.F} = \frac{4054}{3} - 21.6$$

$$= 1351.33 - 21.6 = 1329.73$$

Error SS = Total SS – (Between Subjects SS + Between Districts SS)

$$= 2738.4 - (595.6 + 1329.73) = 813.07$$

The ANOVA-Table is set up below:

Source of variation	*d.f.*	SS	MS	*F*-ratios
Between Subjects	2	595.60	297.80	$F_1 = 2.93$
Between Students	4	1329.73	332.43	$F_2 = 3.27$
Error	8	813.07	101.63	--
Total	14	2738.40	--	--

(v) The critical regions are $F_1 \geq F_{0.05;(2,8)} = 4.46$.

$F_2 \geq F_{0.05;(4,8)} = 3.84$.

(vi) The computed values of F in both the cases are less than the corresponding critical values of F. We therefore accept both the null hypotheses.

Q.20.26. (i) The null hypotheses are stated as:

H_0 : (a) The column means are equal,

(b) The row means are equal,

(c) The columns and rows do not interact.

(ii) The level of significance is set at $\alpha = 0.05$.

(iii) The test-statistic is the usual variance-ratio F which under H_0, has F-distribution.

(iv) **Computations.** To compute the sums of squares, a table containing the totals of the observations in the ijth cell, *i.e.* $T_{ij.}$ is first constructed and then the necessary computations are carried out as below:

Columns

Rows	1	2	3	$T_{i..}$	$T_{i..}^2$	$\sum_j T_{ij.}^2$
1	16 (256)	7 (49)	15 (225)	38	1444	530
2	25 (625)	20 (400)	25 (625)	70	4900	1650
$T_{.j.}$	41	27	40	108	6344	2180
$T_{.j.}^2$	1681	729	1600	4010	11664	--
$\sum_i T_{ij.}^2$	881	449	850	2180	--	--

$$\text{C.F.} = \frac{T_{...}^2}{rcn} = \frac{(108)^2}{18} = 648$$

$$\text{Total } SS = \sum_i \sum_j \sum_k X_{ijk}^2 - \text{C.F.} = 4^2 + 7^2 + \ldots + 8^2 + 7^2 - \text{C.F.}$$

$$= 774 - 648 = 96$$

$$\text{Between Columns } SS = \sum_j \frac{T_{j.}^2}{rn} - \text{C.F.} = \frac{4010}{6} - 648$$

$$= 668.33 - 648 = 20.33$$

$$\text{Between Rows } SS = \sum_i \frac{T_{i..}^2}{cn} - \text{C.F.} = \frac{6344}{9} - 648$$

$$= 70.89 - 648 = 56.89$$

$$\text{Interaction } SS = \sum_i \sum_j \frac{T_{ij.}^2}{n} - \sum_i \frac{T_{i..}^2}{cn} - \sum_j \frac{T_{j.}^2}{rn} + \text{C.F.}$$

$$= \frac{2180}{3} - \frac{6344}{9} - \frac{4010}{6} + 648 = 1.45$$

$$\text{Error } SS = SST - SSC - SSR - SS(RC)$$

$$= 96 - 20.33 - 56.89 - 1.45 = 17.33$$

The analysis of variance table is set up as below:

Source of variation	*d.f.*	SS	MS	*F*-ratios
Between Columns	2	20.33	10.165	7.04
Between Rows	1	56.89	56.89	39.40
Interaction	2	1.45	0.725	0.50
Error	12	17.33	1.444	--
Total	17	96.00	--	--

(v) The critical regions are a) $F \geq F_{0.05(2,12)} = 3.88$

b) $F \geq F_{0.05(1,12)} = 4.75$, and

c) $F \geq F_{0.05(2,12)} = 3.88$

(vi) **Conclusions.** Comparing the calculated values of F with the table values, we find that

a) The column means are different.

b) The row means are different.

c) The columns and rows do not interact.

Q.20.27. (i) We state our hypotheses as

H_0' : there is no difference between diets,

H_0'' : there is no difference between drugs, and

H_0''' : there is no interaction of drugs and diets, against

H_1' : the diets are not equal.

H_1'' : the drugs are not equal.

H_1''' : drugs and diets interact.

(ii) We use a significance level of $\alpha = 0.05$

(iii) The statsitic to be used is the usual variane-ratio, which under the hypotheses, has F-distribution.

(iv) To compute the sum of squares, we first construct a table containing the totals of the observations in the ijth cell, *i.e.* T_{ij}. The other necessary computations are also carried out below:

Table of Cell-Totals, Squares, etc.

Drug	Diet 1	Diet 2	Diet 3	Diet 4	$T_{i\cdot\cdot}$	$T_{i\cdot\cdot}^2$	$\sum_j T_{ij.}^2$
A	12.8 (163.84)	19.1 (364.81)	10.8 (116.64)	9.4 (88.36)	52.1	2714.41	733.65
B	9.7 (94.09)	16.3 (265.69)	9.0 (81.00)	5.9 (34.81)	40.9	1672.81	475.59
C	13.8 (190.44)	19.1 (364.81)	12.0 (144.00)	9.2 (84.64)	54.1	2926.81	783.89
$T_{\cdot j\cdot}$	36.3	54.5	31.8	24.5	147.1	7314.03	1993.13
$T_{\cdot j\cdot}^2$	1317.69	2970.25	1011.24	600.25	5899.43	-----	Check
$\sum T_{ij.}^2$	448.37	995.31	341.64	207.81	1993.13		

$$\text{C.F.} = \frac{T_{\cdots}^2}{rcn} = \frac{(147.1)^2}{24} = \frac{21638.41}{24} = 901.60$$

$$\text{Total } SS = \sum_i \sum_j \sum_k X_{ijk}^2 - \text{C.F.}$$

$$= (6.0)^2 + (6.8)^2 + \ldots + (4.5)^2 + (4.7)^2 - \text{C.F.}$$

$$= 998.45 - 901.60 = 96.85$$

$$\text{Between Diet } SS = \sum_j \frac{T_{\cdot j\cdot}^2}{rn} - \text{C.F.} = \frac{5899.43}{6} - \text{C.F.}$$

$$= 983.24 - 901.60 = 81.64$$

$$\text{Between Drug } SS = \sum_i \frac{T_{i\cdot\cdot}^2}{cn} - \text{C.F.} = \frac{7314.03}{8} - \text{C.F.}$$

$$= 914.25 - 901.60 = 12.65$$

$$\text{Interaction } SS = \sum_i \sum_j \frac{T_{ij\cdot}^2}{n} - \sum_i \frac{T_{i\cdot\cdot}^2}{cn} - \sum_j \frac{T_{\cdot j\cdot}^2}{rn} + \text{C.F.}$$

$$= \frac{1993.13}{2} - 914.25 - 983.24 + 901.60 = 0.68$$

$$\text{Error } SS = \text{Total } SS - \text{Diet } SS - \text{Drug } SS - \text{Interaction } SS$$

$$= 96.85 - (81.64 + 12.65 + 0.68) = 1.88$$

The ANOVA-Table is set up below:

Source of variation	*d.f.*	SS	MS	*F*-ratios
Diets	3	81.64	27.213	$F_1 = 173.33$
Drugs	2	12.65	6.325	$F_2 = 40.29$
Interaction	6	0.68	0.113	$F_3 = 0.72$
Error	12	1.88	0.157	--
Total	23	96.85	--	--

(v) The critical regions are a) $F_1 > F_{0.05;(3,12)} = 3.49$

b) $F_2 > F_{0.05;(2,12)} = 3.88$

c) $F_3 > F_{0.05;(6,12)} = 3.00$

(vi) The difference between drugs are significant but the differences due to the interaction of drugs and diets are not significant.

Q.20.29. In Exercise 20.11, we found that

$s^2 = MSE = 85.31$, $r = 8$, $k = 5$, and v for $MSE = 35$.

We now find $\sqrt{MSE/r} = \sqrt{\frac{85.31}{8}} = \sqrt{10.66375} = 3.266$

Arranging the treatment means ($\bar{X}_i$) in increasing order of magnitude, we get

$\bar{X}_B$	$\bar{X}_A$	$\bar{X}_E$	$\bar{X}_D$	$\bar{X}_C$
31	35	38	42	47

The values of $q_{0.05(p,35)}$ for $p=2,3,4,5$ taken from Duncan's table of significant ranges and the least significant ranges, R_p, obtained by multiplying $q_{0.05}(p,35)$ by $\sqrt{MSE/r}$, are shown below:

p	$q_{0.05}(p,35)$	R_p
2	2.875	9.39 (R_2)
3	3.025	9.88 (R_3)
4	3.110	10.16 (R_4)
5	3.185	10.40 (R_5)

Comparing the differences between all pairs of means with the least significant ranges, R_p, beginning with the largest $\bar{X}_C$ against the smallest $\bar{X}_B$, we have the following results:

$\bar{X}_C$ versus $\bar{X}_B$: $47 - 31 = 16 > 10.40$ (R_5)

$\bar{X}_C$ versus $\bar{X}_A$: $47 - 35 = 12 > 10.16$ (R_4)

$\bar{X}_C$ versus $\bar{X}_E$: $47 - 38 = 9 < 9.88$ (R_3)

$\bar{X}_C$ versus $\bar{X}_D$: $47 - 42 = 5 < 9.39$ (R_2)

$\bar{X}_D$ versus $\bar{X}_B$: $42 - 31 = 11 > 10.16$ (R_4)

$\bar{X}_D$ versus $\bar{X}_A$: $42 - 35 = 7 < 9.88$ (R_3)

$\bar{X}_D$ versus $\bar{X}_E$: $42 - 38 = 4 < 9.33$ (R_2)

$\bar{X}_E$ versus $\bar{X}_B$: $38 - 31 = 7 < 9.88$ (R_3)

$\bar{X}_E$ versus $\bar{X}_A$: $38 - 35 = 3 < 9.39$ (R_2)

$\bar{X}_A$ versus $\bar{X}_B$: $35 - 31 = 4 < 9.39$ (R_2)

The pairs of means whose differences are greater than the corresponding least significant ranges, R_p, are significantly different. Drawing a line under means which are significantly different, we have

$\bar{X}_B$	$\bar{X}_A$	$\bar{X}_E$	$\bar{X}_D$	$\bar{X}_C$
31	35	38	42	47

Q.20.30 (a) (i) We state our hypotheses as

H_0 : $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$, i.e. the differences among the average yields are not significant, and

H_1 : At least two average yields differ significantly.

(ii) The significance level is set at $\alpha = 0.05$

(iii) The test-statistic to use is the usual variance ratio, which has F-distribution.

(iv) Computations.

	V_1	V_2	V_3	V_4	V_5	Total
	46.2 51.9 48.7	49.2 58.6 57.4	60.3 58.7 60.4	48.9 51.4 44.6	52.5 54.0 49.3	
$T_{\cdot j}$	146.8	165.2	179.4	144.9	155.8	792.1
$T^2_{\cdot j}$	21550.24	27291.04	32184.36	20996.01	24273.64	
Means	48.93	55.07	59.80	48.30	51.93	

$$\text{C.F.} = \frac{T^2_{..}}{n} = \frac{(792.1)^2}{15} = 41828.16,$$

$$\text{Total } SS = \sum_i \sum_j X^2_{ij} - \text{C.F.}$$

$$= (46.2)^2 + (51.9)^2 + \ldots + (44.6)^2 - \text{C.F.}$$

$$= 42204.11 - 41828.16 = 375.95,$$

$$\text{Variety } SS = \sum_j \frac{T_j^2}{r} - \text{C.F.} = \frac{126295.29}{3} - \text{C.F.}$$

$$= 42098.43 - 41828.16 = 270.27,$$

Error SS = Total SS − Variety SS

$$= 375.95 - 270.27 = 105.68.$$

The analysis of variance table is:

Source of variation	*d.f.*	SS	MS	*F*-ratios
Varieties	4	270.27	67.57	6.39
Error	10	105.68	10.57	--
Total	14	375.95	--	--

(v) The critical region is $F \geq F_{0.05;(4,10)} = 3.48$

(vi) **Conclusion.** Since the computed value of $F = 6.39$ falls in the critical region, we therefore reject H_0 and conclude that the differences among the average yields are statistically significant.

(b) Use of Duncan's multiple range test.

Arranging the variety means ($\bar{V}_i$) in increasing order of magnitude, we get

$\bar{V}_3$	$\bar{V}_1$	$\bar{V}_5$	$\bar{V}_2$	$\bar{V}_3$
48.30	48.93	51.93	55.07	59.80

The values of $q_{0.05}(p,10)$ for $p = 2,3,4,5$ taken from Duncan's table of significant ranges and the least significant ranges, R_p, obtained by multiplying $q_{0.05}(p,10)$ by $\sqrt{MSE/r} = \sqrt{\frac{10.57}{3}} = 1.877$, are shown below:

p	$q_{0.05}(p,10)$	R_p
2	3.15	5.91 (R_2)
3	3.30	6.19 (R_3)
4	3.37	6.33 (R_4)
5	3.43	6.44 (R_5)

Comparing the differences between all pairs of means with the least significant ranges, R_p, beginning with the largest $\bar{V}_3$ against the smallest $\bar{V}_4$, we have the following results:

$\bar{V}_3$ versus $\bar{V}_4$: $59.80 - 48.30 = 11.50 > 6.44$ (R_5)

$\bar{V}_3$ versus $\bar{V}_1$: $59.80 - 48.93 = 10.87 > 6.33$ (R_4)

$\bar{V}_3$ versus $\bar{V}_5$: $59.80 - 51.93 = 7.87 > 6.19$ (R_3)

$\bar{V}_3$ versus $\bar{V}_2$: $59.80 - 55.07 = 4.73 < 5.91$ (R_2)

$\bar{V}_2$ versus $\bar{V}_4$: $55.07 - 48.30 = 6.77 > 6.33$ (R_4)

$\bar{V}_2$ versus $\bar{V}_1$: $55.07 - 48.93 = 6.14 < 6.19$ (R_3)

$\bar{V}_2$ versus $\bar{V}_5$: $55.07 - 51.93 = 3.14 < 5.91$ (R_2)

$\bar{V}_5$ versus $\bar{V}_4$: $51.93 - 48.30 = 3.63 < 6.19$ (R_3)

$\bar{V}_5$ versus $\bar{V}_1$: $51.93 - 48.93 = 3.00 < 5.91$ (R_2)

$\bar{V}_1$ versus $\bar{V}_4$: $48.93 - 48.30 = 0.63 < 5.91$ (R_2)

The pairs of means whose differences are greater than the corresponding least significant ranges, R_p, are significantly different. Drawing a line under means which are not significantly different, we have

$\bar{V}_4$	$\bar{V}_1$	$\bar{V}_5$	$\bar{V}_2$	$\bar{V}_3$
48.30	48.93	51.93	55.07	59.80

Q.20.31 (b) (i) We state our hypotheses as

H_0 : There is no difference between Coater Types, and

H_1 : There is a significant difference between Coater Types.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic to use is the usual variance-ratio, which has *F-distribution.*

(iv) Computations.

Day	Graphite Coater Type				$T_{i\cdot}$
	M	*A*	*K*	*L*	
1	4.0	4.8	5.0	4.6	18.4
2	4.8	5.0	5.2	4.6	19.6
3	4.0	4.8	5.6	5.0	19.4
$T_{\cdot j}$	12.8	14.6	15.8	14.2	57.4

$$\text{Total } SS = \sum_i \sum_j X_{ij}^2 - \frac{T_{..}^2}{n}$$

$$= (4.0)^2 + (4.8)^2 + \ldots + (5.0)^2 - \frac{(57.4)^2}{12}$$

$$= 276.84 - 274.56 = 2.28,$$

$$\text{Between Coater Types } SS = \sum_j \frac{T_j^2}{r} - \frac{T_{..}^2}{n}$$

$$= \frac{(12.8)^2 + \ldots + (14.2)^2}{3} - \frac{(57.4)^2}{12}$$

$$= 276.09 - 274.56 = 1.53,$$

$$\text{Between Days } SS = \sum_i \frac{T_{i.}^2}{c} - \frac{T_{..}^2}{n}$$

$$= \frac{(18.4)^2 + (19.6)^2 + (19.4)^2}{4} - \frac{(57.4)^2}{12}$$

$$= 274.77 - 274.56 = 0.21.$$

Error $SS = 2.28 - (1.53 + 0.21) = 0.54$

These results are shown in the following *ANOVA*-Table.

Source of variation	*d.f.*	SS	MS	*F*-ratios
Between Coater Types	3	1.53	0.51	5.67
Between Days	2	0.21	0.10	--
Error	6	0.54	0.09	--
Total	11	2.28	--	--

(v) The critical region is $F \geq F_{0.05;(3,6)} = 4.76$.

(vi) **Conclusion.** Since the computed value of $F = 5.67$ falls in the critical region, we therefore reject H_0 and conclude that there is significant difference between Coater types

The least significant difference (*LSD*) is given by

$$LSD = t_{0.025(6)} \sqrt{\frac{2s^2}{r}}$$

$$= (2.447) \sqrt{\frac{2(0.09)}{3}} = (2.447)\,(0.2449) = 0.60$$

Arranging the means of Coater Types in ascending order of magnitude, and drawing a line under any pair of adjacent means (or set of means) that are not significantly differnet, we get

	M	*L*	*A*	*K*
$\bar{X}_{.j}$	4.27	4.73	4.87	5.27

Q.20.33(b) To minimize the function

$$S = \sum_{i=1}^{r} \sum_{j=1}^{c} (X_{ij} - \mu - \alpha_i - \beta_j)^2,$$

we find the partial derivatives of S w.r.t. μ, α_i, β_j and set them equal to zero. Thus

$$\frac{\partial S}{\partial \mu} = 2 \sum_i \sum_j (X_{ij} - \mu - \alpha_i - \beta_j)\,(-1) = 0,$$

$$\frac{\partial S}{\partial \alpha_i} = 2 \sum_j (X_{ij} - \mu - \alpha_i - \beta_j)\,(-1) = 0, \text{ and}$$

$$\frac{\partial S}{\partial \beta_j} = 2 \sum_i (X_{ij} - \mu - \alpha_i - \beta_j)\,(-1) = 0.$$

Simplifying, we get the least squares normal equations as

$$\sum_i \sum_j (X_{ij} - \mu - \alpha_i - \beta_j) = 0$$

$$\sum_j (X_{ij} - \mu - \alpha_i - \beta_j) = 0, \text{ and}$$

$$\sum_i (X_{ij} - \mu - \alpha_i - \beta_j) = 0.$$

We solve these equations with the supporting conditions that

$$\sum_{i=1}^{r} \alpha_i = 0 \text{ and } \sum_{j=1}^{c} \beta_j = 0$$

From the first equation, we get

$$\sum_i \sum_j X_{ij} - rc\mu - c\sum_i \alpha_i - r\sum_j \beta_j = 0$$

or $$\sum_i \sum_j X_{ij} - rc\,\mu = 0 \qquad (\sum_i \alpha_i = 0 = \sum_j \beta_j)$$

This gives $\hat{\mu} = \frac{1}{rc}\sum_i \sum_j X_{ij} = \bar{X}..$

From the second equation, we obtain

$$\sum_j X_{ij} - c\mu - c\alpha_i - \sum_j \beta_j = 0$$

or $$\hat{\alpha}_i = \frac{1}{c}\sum_j X_{ij} - \hat{\mu} \qquad (\sum_j \beta_j = 0)$$

$$= \bar{X}_{i.} - \bar{X}..$$

From the third equation, we get

$$\sum_i X_{ij} - r\mu - \sum_i \alpha_i - r\,\beta_j = 0$$

or $$\hat{\beta}_j = \frac{1}{r}\sum_i X_{ij} - \hat{\mu} \qquad (\sum_i \alpha_j = 0)$$

$$= \bar{X}_{.j} - \bar{X}..$$

where μ, α_i and β_j have been replaced by $\hat{\mu}$, $\hat{\alpha}_i$ and $\hat{\beta}_j$, as they are estimates of the parameters.

Q.34. Let X_{ij} ($i = 1,2,...,n$; $j = 1, 2, ... T$) denote the ith observation of the jth treatment.

	Treatments						
	1	2	...	j	...	T	
	X_{11}	X_{12}	...	X_{1j}	...	X_{1T}	
	X_{21}	X_{22}	...	X_{2j}	...	X_{2T}	
	⋮	⋮		⋮		⋮	
	X_{i1}	X_{i2}	...	X_{ij}	...	X_{iT}	
	⋮	⋮		⋮		⋮	
	X_{n1}	X_{n2}	...	X_{nj}	...	X_{nT}	Total
Total	$T_{.1}$	$T_{.2}$	...	$T_{.j}$	...	$T_{.T}$	$T..$
Mean	$\bar{X}_{.1}$	$\bar{X}_{.2}$	...	$\bar{X}_{.j}$		$\bar{X}_{.T}$	$\bar{X}..$

Further, let $T_{.j}$ denote the total of observations under treatment j, $\bar{X}_{.j}$, the mean of the jth treatment, $T..$, the total of all nT observations and $\bar{X}..$, the grand mean.

Partitioning of Sum of Square:

Now the model is either

$$X_{ij} = \mu + \tau_j + \varepsilon_{ij}, \text{ where } \varepsilon_{ij}\text{'s are NID } (0, \sigma^2).$$

or $$X_{ij} = \mu + (\mu_{.j} - \mu) + (X_{ij} - \mu_{.j})$$

i.e. $$X_{ij} - \mu = (\mu_{.j} - \mu) + (X_{ij} - \mu_{.j})$$

Substituting sample statistics, the identity becomes

$$(X_{ij} - \bar{X}..) = (\bar{X}_{.j} - \bar{X}..) + (X_{ij} - \bar{X}_{.j})$$

Squaring and summing over both i and j, we get

$$\sum_{i=1}^{n}\sum_{j=1}^{T}(X_{ij} - \bar{X}..)^2 = \sum_{i=1}^{n}\sum_{j=1}^{T}(\bar{X}_{.j} - \bar{X}..)^2 + \sum_{i=1}^{n}\sum_{j=1}^{T}(X_{ij} - \bar{X}..)^2 + 2\sum_i\sum_j (\bar{X}_{.j} - \bar{X}..)(X_{ij} - \bar{X}_{.j})$$

Now the cross product term is

$$\sum_i\sum_j (\bar{X}_{.j} - \bar{X}..)(X_{ij} - \bar{X}_{.j}) = \sum_{i=1}^{n}(\bar{X}_{.j} - \bar{X}..)\left[\sum_{j=1}^{T}(X_{ij} - \bar{X}_{.j})\right]$$

The term in brackets equals zero as the sum of the deviations from the mean within a given treatment is equal to zero. Hence

$$\sum_{i=1}^{n}\sum_{j=1}^{T}(X_{ij} - \bar{X}..)^2 = \sum_{j=1}^{T} n(\bar{X}_{.j} - \bar{X}..)^2 + \sum_{i=1}^{n}\sum_{j=1}^{T}(X_{ij} - \bar{X}_{.j})^2$$

i.e. Total SS = Treatment SS + Residual (or Error) SS.

Expected Mean Square.

(i) Treatment $SS = n\sum_{j=1}^{T}(\bar{X}_{.j} - \bar{X}..)^2$, where

$$\bar{X}_{.j} = \frac{1}{n}\sum_{i=1}^{n} X_{ij} = \frac{1}{n}\sum_{i=1}^{n}(\mu + \tau_j + \varepsilon_{ij})$$

$$= \frac{1}{n}(n\mu + n\tau_j + \sum_{i=1}^{n}\varepsilon_{ij})$$

$$= \mu + \tau_j + \bar{\varepsilon}_{.j}, \text{ and } \qquad (\bar{\varepsilon}_{.j} = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_{ij})$$

$$\bar{X}_{..} = \frac{1}{nT}\sum_{i=1}^{n}\sum_{j=1}^{T}X_{ij} = \frac{1}{nT}\sum_{i=1}^{n}\sum_{j=1}^{T}(\mu + \tau_j + \varepsilon_{ij})$$

$$= \frac{1}{nT}\left[nT\mu + n\sum_{j=1}^{T}\tau_j + \sum_{i=1}^{n}\sum_{j=1}^{T}\varepsilon_{ij}\right]$$

$$= \mu + \frac{1}{T}\sum_{j=1}^{T}\tau_j + \bar{\varepsilon}_{..}, \text{ where } \bar{\varepsilon}_{..} = \frac{1}{nT}\sum\sum_{i\ j}\varepsilon_{ij}$$

$$\bar{X}_{.j} - \bar{X}_{..} = (\mu + \tau_j + \bar{\varepsilon}_{.j}) - (\mu + \frac{1}{T}\sum_{j=1}^{T}\tau_j + \bar{\varepsilon}_{..})$$

$$= \tau_j - \frac{1}{T}\sum_{j=1}^{T}\tau_j + \bar{\varepsilon}_{.j} - \bar{\varepsilon}_{..}$$

Squaring, we get

$$(\bar{X}_{.j} - \bar{X}_{..})^2 = (\tau_j - \frac{1}{T}\sum_{j=1}^{T}\tau_j)^2 + (\bar{\varepsilon}_{.j} - \bar{\varepsilon}_{..})^2 + \text{cross product term}$$

Multiplying by n and summing over j, we get

$$\sum_{j=1}^{T} n(\bar{X}_{.j} - \bar{X}_{..})^2 = \sum_{j=1}^{T} n(\tau_j - \frac{1}{T}\sum_{j=1}^{T}\tau_j)^2 + \sum_{j=1}^{T} n(\bar{\varepsilon}_{.j} - \bar{\varepsilon}_{..})^2 + \sum_{j=1}^{T} n \text{ (cross product term)}$$

Taking the expected value, we obtain

$$E(\text{Treatment } SS) = E\left[\sum_{j=1}^{T} n(\bar{X}_{.j} - \bar{X}_{..})^2\right]$$

$$= nE\left[\sum_{j=1}^{T}(\tau_j - \frac{1}{T}\sum_{j=1}^{T}\tau_j)^2\right] + nE\left[\sum_{j=1}^{T}(\bar{\varepsilon}_{.j} - \bar{\varepsilon}_{..})^2\right]$$

($\because$ the expected value of the cross product term equal zero)

(a) **T-treatments are fixed.** Now τ_j are assumed to be fixed constants, *i.e.*

$$\sum_{j=1}^{T}\tau_j = \sum_{j=1}^{T}(\mu_{.j} - \mu) = 0$$

$\therefore$ $E(\text{Treatment } SS) = n\sum_{j=1}^{T}\tau_j^2 + n(T-1)\frac{\sigma^2}{n}$, as $\sum_{j=1}^{T}\tau_j^2$ is a constant and errors are random.

Hence $E(\text{Treatment Mean Square}) = \dfrac{E(\text{Treatment } SS)}{T-1}$

($T-1 = d.f.$ for treatment)

$$= \sigma^2 + \frac{n}{T-1}\sum_{j=1}^{T}\tau_j^2$$

(b) **T-treatments are random.** Here τ_j are random variables and are NID (o, σ_τ^2), where σ_τ^2 denotes the variance among the treatments. τ_j average to zero when averaged over all possible values, but for the T values, they usually will not average zero. Hence

$$E(\text{Treatment } MS) = \sigma^2 + \frac{n(T-1)}{T-1}\sigma_\tau^2$$

$$= \sigma^2 + n\sigma_\tau^2.$$

(ii) **Residual or Error Sum of Squares is given by**

$$\text{Error } SS = \sum_{i=1}^{n}\sum_{j=1}^{T}(X_{ij} - \bar{X}_{.j})^2, \text{ where}$$

$$X_{ij} - \bar{X}_{.j} = (\mu + \tau_j + \varepsilon_{ij}) - (\mu + \tau_j + \bar{\varepsilon}_{.j})$$

$$= \varepsilon_{ij} - \bar{\varepsilon}_{.j}$$

Squaring and summing over both i and j, we get

$$\sum_{i=1}^{n}\sum_{j=1}^{T}(X_{ij}-\bar{X}_{\cdot j})^2 = \sum_{i=1}^{n}\sum_{j=1}^{T}(\varepsilon_{ij}-\bar{\varepsilon}_{\cdot j})^2$$

Taking the expected value, we get

$$E(\text{Error } SS) = E\left[\sum_{i=1}^{n}\sum_{j=1}^{T}(\varepsilon_{ij}-\bar{\varepsilon}_{\cdot j})^2\right]$$

$$= \sum_{j=1}^{T} E\left[\sum_{i=1}^{n}(\varepsilon_{ij}-\bar{\varepsilon}_{\cdot j})^2\right]$$

$$= \sum_{j=1}^{T}(n-1)\,\sigma^2 = T(n-1)\,\sigma^2.$$

Hence $E(\text{Error Mean Square}) = \dfrac{E(\text{Error } SS)}{T(n-1)} = \sigma^2$

[$T(n-1)$ = *d.f.* for Error]

Q.20.35. Given: Number of samples = 5, Total SS = 280, Between Subgrade Soil SS = 120, and Type of subgrade soil = 4.

(i) The analysis of variance table is set up below:

Source of variation	*d.f.*	*SS*	*MS*	*F*-ratio
Between Subgrade Soil	3	120	40	4
Error	16	160	10	--
Total	19	280	--	--

(ii) A fixed-effects model for the problem is

$$Y_{ij} = \mu + \tau_j + \varepsilon_{ij}, \text{ where}$$

Y_{ij} is the ith observation on jth treatment (Types of subgrade soil),

μ is the general effect for the whole experiment,

τ_j is the effect of the jth treatment and

ε_{ij} is the random effect present in the ith observation on jth treatment.

The error term ε_{ij} is normally and independently distributed with mean zero and constant variance.

$\Sigma\tau_j = 0$ is also assumed.

(iii) $F = 4.0$, which is greater than $F_{0.05}(3,16) = 3.24$; so reject H_0.

(iv) One set of orthogonal contrasts for this problem is

$C_1 = T_{\cdot 1} - T_{\cdot 2}$,

$C_2 = T_{\cdot 1} + T_{\cdot 2} - 2T_{\cdot 3}$,

$C_3 = T_{\cdot 1} + T_{\cdot 2} + T_{\cdot 3} - 3T_{\cdot 4}$

(v) When the null hypothesis of equal means is rejected by *F*-test after the ANOVA, we can test the significance of differences between means of k samples (or treatments) by using the ordinary two-sample t-test on every pair of the $k(k-1)/2$ possible pairs of $\bar{X}_i$ and $\bar{X}_j$ $(i \neq j)$ at significance level α. But this procedure involves a large number of decisions. Accordingly, several tests, called the *Multiple Comparison Tests*, have been developed for this purpose.

❖❖❖❖❖❖❖❖❖❖

STATISTICAL INFERENCE IN REGRESSIOIN AND CORRELATION

Q.21.3. The $100(1-\alpha)\%$ confidence interval for the population regresson co-efficient, β is established as

$$-t_{\alpha/2,(\upsilon)} < \frac{b-\beta}{s_b} < t_{\alpha/2,(\upsilon)}$$

or $\quad b \pm t_{\alpha/2,(n-2)}\, s_b,$

where $t_{\alpha/2,(n-2)}$ is the value of the t-distribution with $(n-2)$ *d.f.* leaving an area equal to $\alpha/2$ to the right, and

$$s_b^2 = \frac{\sum(Y-\hat{Y})^2}{(n-2)\sum(X-\bar{X})^2},$$

Given: $n=20$, $\sum X=40$, $\sum Y=60$, $\sum X^2=95$, $\sum Y^2=97$ and $\sum XY=150$.

Now $\quad b = \dfrac{n\sum XY-(\sum X)(\sum Y)}{n\sum X^2-(\sum X)^2} = \dfrac{(20)(150)-(40)(60)}{(20)(95)-(40)^2}$

$$= \frac{3000-2400}{1900-1600} = \frac{600}{300} = 2,$$

$$a = \frac{1}{n}[\sum Y - b\sum X] = \frac{1}{20}[60-2(40)] = -1,$$

$$\sum(Y-\hat{Y})^2 = \sum Y^2 - a\sum Y - b\sum XY$$

$$= 297-(-1)(60)-(2)(150) = 297+60-300 = 57, \text{and}$$

$$\sum(X-\bar{X})^2 = \sum X^2 - \frac{(\sum X)^2}{n} = 95-80 = 15.$$

$$\therefore \quad s_b^2 = \frac{\sum(Y-\hat{Y})^2}{(n-2)\sum(X-\bar{X})^2} = \frac{57}{(18)(15)} = \frac{57}{270}$$

$$= 0.2111, \text{ so that } s_b = \sqrt{0.2111} = 0.46$$

Suppose $\alpha=0.05$. Then $100(1-0.05)\%$, *i.e.* 95% confidence interval for the population regression coefficient β would be

$$b \pm t_{.025(18)}\, s_b$$

Substituting the values, we get

$$2 \pm (2.101)(0.46) \text{ } i.e. \text{ } 2 \pm 0.97 \quad (\because t_{.025(18)} = 2.101)$$

Hence the required 95% confidence interval for β is

1.03 to 2.09

Q.21.4. (i) The slope of the regression line is given by

$$b = \frac{n\sum XY-(\sum X)(\sum Y)}{n\sum X^2-(\sum X)^2}$$

$$= \frac{(100)(1124828)-(6826)(16440)}{(100)(466540)-(6826)^2} = \frac{263360}{59724} = 4.4096$$

(ii) The 95% confidence interval for α is given by

$$a \pm zs_a, \qquad (z \text{ is used as sample size is large})$$

where $\quad s_a = s_{Y.X}\sqrt{\dfrac{1}{n}+\dfrac{\bar{X}^2}{\sum(X-\bar{X})^2}}$

Now $\quad a = \bar{Y}-b\bar{X} = 164.40-(4.4096)(68.26) = -136.5993$

$$\sum(X-\bar{X})^2 = \sum X^2 - \frac{(\sum X)^2}{n} = 466540 - \frac{(6826)^2}{100} = 597.24$$

$$\sum(Y-\hat{Y})^2 = \sum Y^2 - a\sum Y - b\sum XY.$$

$$= 2766596-(-136.5993)(16440)-(4.4096)(1124828)$$

$$= 52246.9432$$

$$\therefore \quad s_{Y.X} = \sqrt{\frac{52246.9432}{100-2}} = 23.0897, \text{ and}$$

$$s_a = (23.0897)\sqrt{\frac{1}{100} + \frac{(68.26)^2}{597.24}}$$

$$= (23.0897)\,(2.7949) = 64.5334$$

Substituting the values, we get

$$-136.5993 \pm (1.96)\,(64.5334)$$

i.e., -136.5993 ± 126.4855 *i.e.*, -263.08 to -10.11.

Note: If we write the equation of the regression line as

$\hat{Y} = \bar{Y} + b(X-\bar{X})$, (estimate of $\mu_{Y.X} = A + B(X-\bar{X})$ (see Dixon and Massey)

then the $(1-\alpha)100\%$ confidence interval for A would be

$$\bar{Y} \pm t_{\alpha/2}\frac{s_{Y.X}}{\sqrt{n}}$$

i.e., $$164.40 \pm (1.96)\frac{23.0897}{\sqrt{100}}$$

i.e., 164.40 ± 4.5256 *i.e.*, 159.87 to 168.93.

The 95% confidence interval for β is given by

$$b \pm t_{\alpha/2,(n-2)}\, s_b,$$

where $s_b = \dfrac{s_{Y.X}}{\sqrt{\Sigma(X-\bar{X})^2}} = \dfrac{23.0897}{\sqrt{597.24}} = 0.9448.$

Substituting the values, we get

$$4.4096 \pm (1.96)\,(0.9448)$$

i.e., 4.4096 ± 1.858 or 2.56 to 6.26.

Q.21.5. (a) The estimated regression line is

$\hat{Y} = a + bX$, where

$$b = \frac{n\Sigma XY - \Sigma X \Sigma Y}{n\Sigma X^2 - (\Sigma X)^2} = \frac{(100)\,(542735) - (4421)(12130)}{(100)\,(208349) - (4421)^2}$$

$$= \frac{646770}{1289659} = 0.5015, \text{ and}$$

$$a = \bar{Y} - b\bar{X} = 121.30 - (0.5015)\,(44.21) = 99.1287.$$

Thus $\hat{Y} = 99.13 + 0.502X.$

OR The estimate of $\mu_{Y.X} = A + B(X-\bar{X})$ would be

$\hat{Y} = 121.3 + 0.502(X-44.21)$ (Dixon and Massey)

(b) (i) The 95% confidence interval for α is

$$a \pm t_{\alpha/2,(n-2)}\, s_a,$$

where $s_a = s_{Y.X}\sqrt{\dfrac{1}{n} + \dfrac{\bar{X}^2}{\Sigma(X-\bar{X})^2}}$

Now $\Sigma(X-\bar{X})^2 = \Sigma X^2 - \dfrac{(\Sigma X)^2}{n}$

$= 208349 - (4421)^2/100 = 12896.59$

$\Sigma(Y-\hat{Y})^2 = \Sigma Y^2 - a\Sigma Y - b\Sigma XY$

$= 1498976 - (99.1287)(12130) - (0.5015)(542735)$

$= 24363.2665$, and

$$s_{Y.X} = \sqrt{\frac{\Sigma(Y-\hat{Y})^2}{n-2}} = \sqrt{\frac{24363.2665}{98}} = 15.7672$$

Thus $$s_a = (15.7672)\sqrt{\frac{1}{100} + \frac{(44.21)^2}{12896.59}}$$

$$= (15.7672)\,(0.4019) = 6.3368$$

Substituting the values, we get the 95% *C.I.* for α as

$$99.13 \pm (1.96)\,(6.3368) \qquad (n \text{ is large})$$

i.e., 99.13 ± 12.42 *i.e.*, 86.71 to 111.55

Note: For the estimate of A, in the equation $\mu_{Y.X} = A + B(X-\bar{X})$, these limits for A become 118.16 to 124.44.

The 95% confidence interval for β is

$$b \pm t_{\alpha/2(n-2)}\, s_b,$$

where $s_b = \dfrac{s_{Y.X}}{\sqrt{\sum(X-\bar{X})^2}} = \dfrac{15.7672}{\sqrt{12896.59}} = 0.1388$

Thus the 95% *C.I.* for β is

$0.5015 \pm (1.96)\ (0.1388)$ (n is large)

or 0.5015 ± 0.272 or 0.23 to 0.78.

(ii) The predicted value of Y when $X = 50$ is

$$\hat{Y} = 99.13 + 0.502(50) = 124.23.$$

The 90% confidence interval for Y when $X = 50$ is given by

$$\hat{Y} \pm t_{\alpha/2} \cdot s_{Y.X} \sqrt{1 + \frac{1}{n} + \frac{(X_0-\bar{X})^2}{\sum(X-\bar{X})^2}}.$$

Substituting the values, we get

$$124.23 \pm (1.66)\ (15.77) \sqrt{1 + \frac{1}{100} + \frac{(50-44.21)^2}{12896.59}}$$

or $124.23 \pm (1.66)\ (15.77)\ (1.0063)$

or 124.23 ± 26.34 or 97.89 to 150.57.

Q.21.6. (i) The 90% confidence interval for β is given by

$$b \pm t_{\alpha/2,(n-2)}\, s_b,$$

where $s_b^2 = \dfrac{\sum(Y-\hat{Y})^2}{(n-2)\sum(X-\bar{X})^2}$

Now $b = \dfrac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2} = \dfrac{(10)\ (10074) - (311)\ (310.1)}{(10)\ (10100) - (311)^2}$

$= \dfrac{4298.9}{4279} = 1.0047;$

$a = \bar{Y} - b\bar{X} = 31.01 - (1.0047)\ (31.1) = -0.2362;$

$\sum(X-\bar{X})^2 = \sum X^2 - (\sum X)^2/n = 10100 - (311)^2/10 = 427.9;$

$\sum(Y-\hat{Y})^2 = \sum Y^2 - a\sum Y - b\sum XY$

$= 10055.09 - (-0.2362)(310.1) - (1.0047)(10074)$

$= 6.9879.$

$\therefore \quad s_b = \sqrt{\dfrac{6.9878}{8(427.9)}} = \sqrt{0.0020} = 0.0447$ and $t_{0.05,(8)} = 1.86.$

Substituting the values, we get the 90% *C.I.* for β as

$1.0047 \pm (1.86)\ (0.0447)$

or 1.0047 ± 0.0831 or 0.92 to 1.09.

The 90% confidence interval for $\mu_{Y.X=30}$ is given by

$$\hat{Y} \pm t_{\alpha/2,(n-2)}\, s_{\hat{Y}},$$

where $s_{\hat{Y}} = s_{Y.X} \sqrt{\dfrac{1}{n} + \dfrac{(X_0-\bar{X})^2}{\sum(X-\bar{X})^2}},$

When $X = 30$, $\hat{Y} = -0.2362 + (1.0047)\ (30) = 29.90.$

$$s_{Y.X} = \sqrt{\frac{\sum(Y-\hat{Y})^2}{n-2}} = \sqrt{\frac{6.9878}{8}} = 0.9346$$

Thus $s_{\hat{Y}} = (0.9346) \sqrt{\dfrac{1}{10} + \dfrac{(30-31.1)^2}{427.2}}$

$= (0.9346)\ (0.3206) = 0.2996.$

$t_{0.05,(8)} = 1.86.$

Substitution gives 29.90 ± 0.56 or 29.34 to 30.46 as the desired interval.

(ii) We state our hypotheses as

$$H_0 : \beta = 1 \text{ and } H_1 : \beta \neq 1$$

Let α be 0.05.

The test-statistic under H_0 is

$$t = \frac{(b - \beta_0)}{s_b} = \frac{1.0047 - 1}{0.0447} = 0.11$$

The critical region is $|t| \geq t_{0.025(8)} = 2.306.$

Since the computed value of $t = 0.11$ does not fall in the critical region, we therefore accept H_0.

Q.21.7. The estimated regression line is

$$\hat{Y} = a + bX, \text{ where}$$

$$b = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{n\Sigma X^2 - (\Sigma X)^2} = \frac{(12)(109380) - (766)(1700)}{(12)(49068) - (766)^2}$$

$$= \frac{1312560 - 1302200}{588816 - 586756} = \frac{10363}{2064} = 5.02, \text{ and}$$

$$a = \bar{Y} - b\bar{X} = \frac{1700}{12} - (5.02)\left(\frac{766}{12}\right)$$

$$= 141.67 - 320.44 = -178.77$$

Thus $\hat{Y} = -178.77 + 5.02X$ is the equation of the fitted regression line.

(i) The hypotheses are stated as

$$H_0 : \beta = 0 \text{ and } H_1 : \beta \neq 0$$

(ii) The significance level is set at $\alpha = 0.05$

(iii) The test-statistic under H_0 is

$$t = \frac{b}{s_b}, \text{ where } s_b^2 = \frac{\Sigma(Y - \hat{Y})^2}{(n-2)\,\Sigma(X - \bar{X})^2}, \text{ and}$$

$$d.f. = n - 2.$$

(iv) Computations.

Now $\Sigma(Y - \hat{Y})^2 = \Sigma Y^2 - a\Sigma Y - b\Sigma XY$

$$= 246100 - (-178.77)(1700) - (5.02)(109380)$$

$$= 246100 + 303909 - 549087.6 = 921.4,$$

$$\Sigma(X - \bar{X})^2 = \Sigma X^2 - (\Sigma X)^2 / 12 = 49068 - (766)^2/12$$

$$= 49068 - 48896.33 = 171.63$$

$\therefore$ $s_b^2 = \dfrac{921.4}{(10)(171.67)} = 0.5367$, so that $s_b = \sqrt{0.5367} = 0.73$, and

$$t = \frac{5.02}{0.73} = 6.88$$

(v) The critical region is $|t| \geq t_{0.025,(10)} = 2.23$

(vi) **Conclusion.** Since the calculated value of t falls in the critical region, we therefore reject H_0.

Q.21.8. The estimated regression line is $\hat{Y} = a + bX$, where

$$b = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{n\Sigma X^2 - (\Sigma X)^2} = \frac{(20)(150) - (40)(60)}{(20)(95) - (40)^2}$$

$$= \frac{3000 - 2400}{1900 - 1600} = \frac{600}{300} = 2, \text{ and}$$

$$a = \bar{Y} - b\bar{X} = 3 - (2)(2) = -1$$

Thus $\hat{Y} = 2X - 1$.

(i) The hypotheses are stated as

$H_0 : \beta = 0$, *i.e.* the two variables are independent, and

$H_1 : \beta \neq 0$.

(ii) We use a significance level of $\alpha = 0.05$.

(iii) The test-statistic under H_0 is

$$t = \frac{b}{s_b}, \text{ where}$$

$s_b^2 = \dfrac{\Sigma(Y - \hat{Y})^2}{(n-2)\,\Sigma(X - \bar{X})^2}$ and t has Student's distribution with $(n-2)$ *d.f.*

(iv) Computations.

Now $\Sigma(X - \bar{X})^2 = \Sigma X^2 - (\Sigma X)^2/n = 95 - 80 = 15,$

$\Sigma(Y - \hat{Y})^2 = \Sigma Y^2 - a\Sigma Y - b\Sigma XY$

$= 297 - (-1)(60) - (2)(150) = 57,$

$$s_b^2 = \frac{57}{(18)(15)} = \frac{57}{270} = 0.2111, \text{ so that}$$

$$s_b = \sqrt{0.2111} = 0.46$$

$$t = \frac{2}{0.46} = 4.35$$

(v) The critical region is $|t| \geq t_{0.025,(18)} = 2.10$

(vi) **Conclusion.** As the computed value of $t = 4.35$ falls in the critical region, we therefore reject H_0, and may conclude that the variables are linearly related.

Q.21.9. (i) We are required to test the hypothesis

$$H_0 : \beta = 0 \text{ against } H_1 : \beta \neq 0$$

(ii) We use a significance level of $\alpha = 0.05$, and a two-sided test.

(iii) The test-statistic would be

$$t = \frac{b}{s_b}, \text{ where } s_b^2 = \frac{\Sigma(Y_i - \hat{Y})^2}{(n-2)\,\Sigma(X_i - \bar{X})^2}$$

Assuming that the distribution of Y_i for each X, is normal with the same mean and the same variance, the statistic conforms to the Student's t-distribution with $(n-2)$ *d.f.*

(iv) Computation.

	X_i	Y_i	X_iY_i	X_i^2	Y_i^2
	60	130	7800	3600	16900
	62	135	8370	3844	18225
	65	158	10270	4225	24964
	70	170	11900	4900	28900
	72	185	13320	5184	34225
Σ	329	778	51660	21753	123214

Now $$b = \frac{n\Sigma X_iY_i - \Sigma X_i \Sigma Y_i}{n\Sigma X_i^2 - (\Sigma X_i)^2} = \frac{(5)\,(51660) - (329)\,(778)}{(5)\,(21753) - (329)^2}$$

$$= \frac{258300 - 255962}{108765 - 108241} = \frac{2338}{524} = 4.46,$$

$$a = \frac{1}{n}\,[\Sigma Y_i - b\Sigma X_i] = \frac{1}{5}[778 - (4.46)\,(329)]$$

$$= \frac{1}{5}[778 - 1467.34] = \frac{1}{5}(-689.34) = -137.87.$$

And $$\Sigma(Y_i - \hat{Y})^2 = \Sigma Y_i^2 - a\Sigma Y_i - b\Sigma X_iY_i$$

$$= 123214 - (-137.87)\,(778) - (4.46)\,(51660)$$

$$= 123214 + 107262.86 - 230403.6 = 73.26,$$

$$\Sigma(X_i - \bar{X})^2 = \Sigma X_i^2 - \frac{(\Sigma X_i)^2}{n} = 21753 - \frac{(329)^2}{5}$$

$$= 21753 - 21648.2 = 104.8, \text{ and}$$

$$s_b^2 = \frac{73.26}{(3)\,(104.8)} = \frac{73.26}{314.4} = 0.2330, \text{ so that}$$

$$s_b = \sqrt{0.2330} = 0.483.$$

$$\therefore \quad t = \frac{b}{s_b} = \frac{4.46}{0.483} = 9.23$$

(v) The critical region is $|t| \geq t_{.025,(3)} = 3.18$

(vi) **Conclusion.** Since the computed value of t falls in the critical region, so we reject H_0. There is sufficient evidence to indicate that the heights are linearly related to the weights.

(b) Here we are required to test the hypothesis $H_0 : \beta = 6$ against $H_1 : \beta \neq 6$. The test-statistic would be

$$t = \frac{b - \beta_0}{s_b}, \text{ where } s_b^2 = \frac{\Sigma(Y - \hat{Y})^2}{(n-2)\,\Sigma(X - \bar{X})^2}, \text{ etc.}$$

$$\therefore \quad t = \frac{4.46 - 6.00}{0.483} = \frac{-1.54}{0.483} = -3.188$$

We reject the hypothesis $H_0 : \beta = 6$ as the computed value of t falls in the critical region.

(c) The regression equation (or predicting equation) of Y on X is $\hat{Y} = -137.87 + 4.46X$.

When $X = 66$, $\hat{Y} = -137.87 + 4.46\,(66) = 156.49$

Thus the weight of an individual who is 66 inches in height, is 156.5.

The 95% "prediction interval" for $\hat{Y}$ is given by

$$\hat{Y} \pm t_{\alpha/2,(n-2)}\, s_{\hat{y}} \text{ where } s^2_{\hat{y}} = s^2_e \left[1 + \frac{1}{n} + \frac{(X-\bar{X})^2}{\sum(X-\bar{X})^2}\right]$$

Now $$s^2_{\hat{y}} = \frac{73.26}{3}\left[1 + \frac{1}{5} + \frac{(66-65.8)^2}{104.8}\right]$$

$$= 24.42\,[1 + 0.2 + 0.0004]$$

$$= 24.42\,(1.2004) = 29.3138, \text{ so that } s_{\hat{y}} = 5.414.$$

Substituting the value of $t_{.025,(3)} = 3.182$, we get

$156.49 \pm (3.182)\,(5.414)$, *i.e.* 156.49 ± 17.23

Hence the 95% prediction interval is 139.26 to 173.72.

Q.21.10. Calculation for the fitting of a regression line $Y = a + bX$.

X	Y	X^2	Y^2	XY
65	85	4225	7225	5525
50	74	2500	5476	3700
55	76	3025	5776	4180
65	90	4225	8100	5850
55	85	3025	7225	4675
70	87	4900	7569	6090
65	94	4225	8836	6110
70	98	4900	9604	6860
55	81	3025	6561	4455
70	91	4900	8281	6370
50	76	2500	5776	3800
55	74	3025	5476	4070
Σ 725	1011	44,475	85905	61685

The two normal equations are

$$\Sigma Y = na + b\Sigma X \text{ and } \Sigma XY = a\Sigma X + b\Sigma X^2$$

Solving for a and b, we get

$$a = \frac{1}{n}[\Sigma Y - b\Sigma X] \text{ and } b = \frac{n\Sigma XY - \Sigma X \Sigma Y}{n\Sigma X^2 - (\Sigma X)^2}.$$

Substituting the values, we get

$$b = \frac{(12)\,(61685) - (725)\,(1011)}{(12)\,(44475) - (725)^2} = \frac{740220 - 732975}{533700 - 525625}$$

$$= \frac{7245}{8075} = 0.897, \text{ and}$$

$$a = \frac{1}{12}[1011 - (0.897)\,(725)] = \frac{1}{12}[1011 - 650.325]$$

$$= \frac{1}{12}(360.675) = 30.056$$

Hence the required equation of the regression line is

$$\hat{Y} = 30.056 + 0.897X$$

Next, we are required to test the hypothesis

(a) (i) $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$

(ii) The level of significance is $\alpha = 0.01$, and we use a two-sided test.

(iii) The test-statistic under H_0 would be

$$t = \frac{b}{s_b}, \text{ where } s^2_b = \frac{\Sigma(Y_i - \hat{Y})^2}{(n-2)\,\Sigma(X_i - \bar{X})^2}.$$

Assuming that the distribution of Y_i for each X, is normal with the same mean and the same variance, the statistic t conforms to the Student's t-distribution with $(n-2)$ *d.f.*

(iv) Computation.

$$\Sigma(Y_i - \hat{Y})^2 = \Sigma Y_i^2 - a\Sigma Y_i - b\Sigma X_i Y_i$$

$$= 85905 - (30.056)\,(1011) - (0.897)\,(61685)$$

$$= 85905 - 30386.616 - 5533.445 = 186.939$$

$$\Sigma(X_i - \bar{X})^2 = \Sigma X_i^2 - \frac{(\Sigma X_i)^2}{n} = 44475 - \frac{(725)^2}{12}$$

$$= 44475 - 43802.083 = 672.917, \text{ and}$$

$$s_b^2 = \frac{186.939}{(10)\,(672.917)} = 0.027780, \text{ so that}$$

$$s_b = \sqrt{0.027780} = 0.167$$

$$\therefore \quad t = \frac{0.897}{0.167} = 5.37$$

(v) The critical region is $|t| \geq t_{.005,(10)} = 3.169$

(vi) The computed value of t falls in the critical region. We reject H_0. The data provide an evidence to conclude that the regression coefficient, β, is almost certainly different from zero.

(b) (i) The hypotheses are stated as

$$H_0 : \alpha = 32 \text{ and } H_1 : \alpha \neq 32$$

(ii) The level of significance is $\alpha = 0.01$

(iii) The test-statistic is

$$t = \frac{a - \alpha}{s_a}, \text{ where}$$

$$s_a = s_{Y.X}\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\Sigma(X - \bar{X})^2}},$$

and the statistic t has Student's t-distribution with $(n-2)$ *d.f.*

(iv) Computations. We have

$$a = 30.056 \quad \Sigma(X - \bar{X})^2 = 672.917,$$

$$s_{Y.X} = \sqrt{\frac{\Sigma(Y - \hat{Y})^2}{n-2}} = \sqrt{\frac{186.939}{10}} = 4.3236,$$

Now $$s_a = 4.3236\sqrt{\frac{1}{12} + \frac{(68.42)^2}{672.917}}$$

$$= 4.3236\sqrt{0.0833 + 6.9567} = 11.47, \text{ and}$$

$$\therefore \quad t = \frac{30.056 - 32}{11.47} = \frac{-1.944}{11.47} = -0.169$$

(v) The critical region is $|t| \geq t_{0.005,(10)} = 2.764$

(vi) **Conclusion.** Since the computed value of t does not fall in the critical region, we therefore accept H_0.

Q.21.11. (b) (i) Estimates of β_1 and β_2.

For Set A:

$$b_1 = \frac{n\Sigma XY - (\Sigma X)\,(\Sigma Y)}{n\Sigma X^2 - (\Sigma X)^2} = \frac{(4)\,(76) - (8)\,(37)}{(4)\,(18) - (8)^2}$$

$$= \frac{304 - 296}{72 - 64} = 1.$$

For Set B:

$$b_2 = \frac{n\Sigma XY - (\Sigma X)\,(\Sigma Y)}{n\Sigma X^2 - (\Sigma X)^2} = \frac{(4)\,(179) - (15)\,(47)}{(4)\,(59) - (15)^2}$$

$$= \frac{716 - 705}{236 - 225} = 1.$$

(ii) The 95% confidence interval for $\beta_1 - \beta_2$ is given by

$$(b_1 - b_2) \pm t_{0.025,(4)}\, s_{Y.X.p}\sqrt{\left[\frac{1}{\Sigma(X_1 - \bar{X}_1)^2} + \frac{1}{\Sigma(X_2 - \bar{X}_2)^2}\right]}$$

where $\Sigma(X_1 - \bar{X}_1)^2 = \Sigma X_1^2 - (\Sigma X_1)^2 / n = 18 - (64)/4 = 2,$

$$\Sigma(X_2 - \bar{X}_2)^2 = \Sigma X_2^2 - (\Sigma X_2)^2 / n = 59 - (225)/4 = 2.85,$$

$$a_1 = \bar{Y}_1 - b_1\bar{X}_1 = 9.15 - (1)\,(2) = 7.15,$$

$$a_2 = \bar{Y}_2 - b_2\bar{X}_2 = 11.75 - (1)\,(3.75) = 8,$$

$$\Sigma(Y_1 - \hat{Y}_1)^2 = \Sigma Y_1^2 - a_1\Sigma Y_1 - b_1\Sigma X_1 Y_1$$

$$= 349 - (7.15)\,(37) - (1)\,(76)$$

$$= 349 - 264.55 - 76 = 8.45,$$

$$\sum(Y_2-\hat{Y}_2)^2 = \sum Y_2^2 - a_2\sum Y_2 - b_2\sum X_2Y_2$$

$$= 557 - (8)\,(47) - (1)\,(179)$$

$$= 557 - 376 - 179 = 2, \text{ and}$$

$$s^2_{Y.X.P} = \frac{\sum(Y_1-\hat{Y}_1)^2 + \sum(Y_2-\hat{Y}_2)^2}{n_1 + n_2 - 4} = \frac{8.45 + 2}{4} = 2.6125,$$

so that $s_{Y.X.P} = \sqrt{2.6125} = 1.6163$

Substituting the values, we get

$$(1-1) \pm (2.776)\,(1.6163)\sqrt{\frac{1}{2} + \frac{1}{2.85}}$$

i.e. $0 \pm (2.776)\,(1.4909)$ or 0 ± 4.14.

Hence the 95% C.I. is $-4.14 < \beta_1 - \beta_2 < 4.14$

To test $H_0 : \beta_1 = \beta_2$ against $H_1 : \beta_1 \neq \beta_2$ at the 0.05 level of significance, we find

$$t = \frac{b_1 - b_2}{s_{Y.X.P.}\left[\frac{1}{\sum(X_1-\bar{X}_1)^2} + \frac{1}{\sum(X_2-\bar{X}_2)^2}\right]^{1/2}}$$

$$= \frac{1-1}{1.1616\left[\frac{1}{2} + \frac{1}{2.85}\right]^{1/2}} = 0,$$

which does not fall in the critical region ($|t| \geq t_{0.025,(4)} = 2.776$), we therefore accept the null hypothesis $H_0 : \beta_1 = \beta_2$.

Q.21.12. (i) Estimates of β_1, β_2 and β_3.

For b_1 and b_2, see Q.21.11 (i) above.

Now $$b_3 = \frac{(4)\,(47) - (7)\,(26)}{(4)\,(15) - (7)^2} = \frac{188 - 182}{60 - 49} = 0.5455.$$

(ii) To test $H_0 : \beta_1 = \beta_2$, we found in Q.21.11 (ii), $b_1 - b_2 = 0$, $t = 0$, so we accept H_0.

To test $H_0 : \beta_1 = \beta_3$, we find

$$\sum(X_3-\bar{X}_3)^2 = \sum X_3^2 - (\sum X_3)^2 / n = 15 - (7)^2/4 = 2.75,$$

$$a_3 = \bar{Y}_3 - b_3\bar{X}_3 = 6.5 - (0.5455)\,(1.75) = 5.5454,$$

$$\sum(Y_3-\hat{Y}_3)^2 = \sum Y_3^2 - a_3\sum Y_3 - b_3\sum X_3Y_3$$

$$= 174 - (5.5454)\,(26) - (0.5455)\,(47)$$

$$= 174 - 144.1804 - 25.6385 = 4.18,$$

$$s^2_{Y.X.P} = \frac{\sum(Y_1-\hat{Y}_1)^2 + \sum(Y_3-\hat{Y}_3)^2}{n_1 + n_3 - 4} = \frac{8.45 + 4.18}{4}$$

$= 3.1575$, so that $s_{Y.X.P} = \sqrt{3.1575} = 1.7769$.

$$\therefore \quad t = \frac{b_1 - b_3}{s_{Y.X.P.}\left[\frac{1}{\sum(X_1-\bar{X}_1)^2} + \frac{1}{\sum(X_3-\bar{X}_3)^2}\right]^{1/2}}$$

$$= \frac{1 - 0.5455}{1.7769\left[\frac{1}{2} + \frac{1}{2.75}\right]^{1/2}} = \frac{0.4545}{(1.7769)\,(0.9293)}$$

$$= \frac{0.4545}{1.6513} = 0.275, \text{ so accept } H_0.$$

Similarly, for $H_0 = \beta_2 = \beta_3$, we find

$t = 0.43$ and accept H_0.

Q.21.13. Using all eleven values, we find that the estimated regression line is $\hat{Y} = 7.41 + 1.786X$ & $\sum(X-\bar{X})^2 = 28$.

To test the hypothesis, we proceed as below.

(i) We state our hypotheses as

H_0 : The regression is linear, and

H_1 : The regression is not linear.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic under H_0 is

$$F = \frac{\chi_1^2 / (k-2)}{\chi_2^2 / (n-k)}, \text{ where}$$

$$\chi_1^2 = \sum \frac{Y_{i.}^2}{n_i} - \frac{(\sum Y_{ij})^2}{n} - b^2 \sum (X - \bar{X})^2, \text{ and}$$

$$\chi_2^2 = \sum Y_{ij}^2 - \sum \frac{Y_{i.}^2}{n_i};$$

and which has an F-distribution with $v_1 = k-2$ and $v_2 = n-k$ *d.f.*

(iv) Computations. To find the necessary calculations, we rearrange the data as follows:

X	Y-values	$Y_{i.}$
2	4, 3, 8	15
3	18, 22	40
4	24,	24
5	24, 18	42
6	13,10,16	39
	$\sum Y_{ij}$	160

Now $\chi_1^2 = \sum \frac{Y_{i.}^2}{n_i} - \frac{(\sum Y_{ij})^2}{n} - b^2 \sum (X - \bar{X})^2$

$= 2840 - 2327.27 - (1.786)^2 (28)$

$= 2840 - 2416.58 = 423.42$, and

$\chi_2^2 = 2898 - 2840 = 58.$

$\therefore \quad F = \frac{(423.42) / (5-2)}{58 / (11-5)} = \frac{423.42}{3} \times \frac{6}{58} = 14.60$

(v) The critical region is $F > F_{0.05,(3,6)} = 4.76$.

(vi) **Conclusion.** Since the computed value of $F = 14.60$ falls in the critical region, we therefore reject H_0 and conclude that the regression is not linear.

Q.21.14. Using all the sixteen values, we find that the estimated regression line is

$\hat{Y} = 2.5X - 5$, and $\sum (X - \bar{X})^2 = \sum X^2 - (\sum X)^2 / n$

$= 1104 - 1024 = 80.$

To test the hypothesis, we proceed as below:

(i) We state our hypotheses as

H_0 : the regression is linear, and

H_1 : the regression is not linear.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic under H_0 is

$$F = \frac{\chi_1^2 / (k-2)}{\chi_2^2 / (n-k)},$$

where $\chi_1^2 = \sum \frac{Y_{i.}^2}{n_i} - \frac{(\sum Y_{ij})^2}{n} - b^2 \sum (X - \bar{X})^2$, and

$$\chi_2^2 = \sum Y_{ij}^2 - \sum \frac{Y_{i.}^2}{n_i}, \text{ and}$$

the ratio F has an F-distribution with $v_1 = k-2$ and $v_2 = n-k$ degrees of freedom.

(iv) Computations. The necessary computations are given below:

X	Y-values				$Y_{i.}$
5	2,	8,	10,	12	32
7	9,	13,	14,	16	52
9	9,	14,	16,	21	60
11	18,	19,	23,	36	96
				$\sum Y_{ij}$	240

Now $\chi_1^2 = \frac{(32)^2}{4} + \frac{(52)^2}{4} + \frac{(60)^2}{4} + \frac{(96)^2}{4} - \frac{(240)^2}{16} - (2.5)^2 (80)$

$= 4136 - 3600 - 500 = 36$, and

$$\chi_2^2 = (2)^2 + (8)^2 + \ldots + (36)^2 - 4136 = 362$$

$$\therefore \quad F = \frac{36/(4-2)}{362/(16-4)} = \frac{432}{724} = 0.60.$$

(v) The critical region is $F \geq F_{0.05,(2,12)} = 3.89$.

(vi) **Conclusion.** Since the computed value of $F = 0.60$ does not fall in the critical region, we therefore accept H_0 and conclude that there is sufficient evidence to indicate that the regression is linear.

Q.21.15. Using all the eighteen values, we find that

$$\Sigma X = 675, \Sigma Y = 488, \Sigma XY = 25005 \text{ and } \Sigma X^2 = 37125.$$

$$\therefore \quad b = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{n\Sigma X^2 - (\Sigma X)^2} = \frac{(18)(25005) - (675)(488)}{(18)(37125) - (675)^2}$$

$$= \frac{120690}{212625} = 0.568, \text{ and}$$

$$a = \bar{Y} - b\bar{X} = \frac{488}{18} - (0.568)\left(\frac{675}{18}\right)$$

$$= 27.111 - (0.568)(37.5) = 5.811.$$

Thus the estimated regression line is

$$\hat{Y} = 5.811 + 0.568X$$

To test the hypothesis, we proceed as below:

(i) We state our hypotheses as

H_0 : the regression is linear, and

H_1 : the regression is not linear.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic under H_0 is

$$F = \frac{\chi_1^2/(k-2)}{\chi_2^2/(n-k)},$$

which has an F-distribution with $v_1 = k-2$, $v_2 = n-k$ degrees of freedom.

(iv) Computations.

X_i	Y-values			Y_i
0	8,	6,	8	22
15	12,	10,	14	36
30	25,	21,	24	70
45	31,	33,	28	92
60	44,	39,	42	125
75	48,	51,	44	143
			ΣY_{ij}	488

Now, $\Sigma(X-\bar{X})^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{n}$

$$= 37125 - \frac{(1675)^2}{18} = 11812.5.$$

$$\chi_1^2 = \frac{(22)^2}{3} + \frac{(36)^2}{3} + \ldots + \frac{(143)^2}{3} - \frac{(488)^2}{18} - (0.568)^2(11812.5)$$

$= 17072.667 - 13230.222 - 3810.996 = 31.449$, and

$$\chi_2^2 = (8)^2 + (6)^2 + \ldots + (44)^2 - 17072.667$$

$= 17142 - 17072.667 = 69.333$

$$\therefore \quad F = \frac{\chi_1^2/(k-2)}{\chi_2^2/(n-k)} = \frac{31.449}{6-2} \times \frac{18-6}{69.333} = 1.36.$$

(v) The critical region is $F \geq F_{0.05,(4,12)} = 3.26$

(vi) **Conclusion.** Since the computed value of F does not fall in the critical region, we therefore accept H_0 and conclude that the regression is linear.

Q.21.16. (b) To find the confidence interval for ρ, we first find the confidence interval for μ_z. Thus 99% confidence interval for μ_z would be

$$z_f \pm 2.58\left(\frac{1}{\sqrt{n-3}}\right), \text{ where}$$

$$z_f = 1.1513 \log (1.2/0.8) = (1.1513)\,(0.1761) = 0.203$$

Substituting the values, we get

$$0.203 \pm 2.58\,(0.040096) \text{ or } 0.203 \pm 0.103$$

$$\therefore \quad 0.100 < \mu_z < 0.306$$

Substituting $1.1513 \log \dfrac{1+\rho}{1-\rho}$ for μ_z and simplifying, we find that the 99% confidence interval for ρ is 0.100 to 0.297.

Q.21.17. (a) To find the confidence interval for ρ, we first find the confidence interval for $\mu_z = 1.1513 \log \dfrac{1+\rho}{1-\rho}$.

The 90% confidence interval for μ_z would be

$$z_f \pm \frac{1.645}{\sqrt{n-3}}$$

$$\text{Now} \quad z_f = 1.1513 \log \frac{1+r}{1-r} = 1.1513 \log \frac{1+0.59}{1-0.59}$$

$$= (1.1513)\,(0.5886) = 0.6777, \text{ and } n-3 = 20$$

Substituting, we get

$$0.6777 \pm \frac{1.645}{\sqrt{20}}$$

$$i.e \quad 0.6777 \pm (1.645)\,(0.2236), \textit{ i.e. } 0.6777 \pm 0.3678$$

$$\therefore \quad 0.3099 < \mu_z < 1.0455$$

Using Fisher's z-table, we find values, ρ_L and ρ_U, that correspond to Fisher's z-values equal to 0.3099 and 1.0455 respectively. Thus $P_L = 0.30$ and $P_U = 0.78$.

Hence the approximate 90% confidence interval for ρ is (0.30, 0.78).

(b) First we calculate the correlation coefficient.

$$r = \frac{\dfrac{\Sigma XY}{n} - \left(\dfrac{\Sigma X}{n}\right)\left(\dfrac{\Sigma Y}{n}\right)}{\sqrt{\dfrac{\Sigma X^2}{n} - \left(\dfrac{\Sigma X}{n}\right)^2}\sqrt{\dfrac{\Sigma Y^2}{n} - \left(\dfrac{\Sigma Y}{n}\right)^2}}$$

$$= \frac{\dfrac{90113.83}{243} - \left(\dfrac{5442.2}{243}\right)\left(\dfrac{4019.6}{243}\right)}{\sqrt{\dfrac{122155.04}{243} - \left(\dfrac{5442.2}{243}\right)^2}\sqrt{\dfrac{66588.92}{243} - \left(\dfrac{4019.6}{243}\right)^2}}$$

$$= \frac{370.8388 - 370.4630}{\sqrt{(1.120)\,(0.405)}} = \frac{0.3758}{0.6735} = 0.558.$$

Now z_f for $r = 0.558$ from Fisher's table is 0.630. Therefore the 95% confidence interval for μ_z is $z_f \pm 1.96\left(\dfrac{1}{\sqrt{n-3}}\right)$.

Substituting the values, we get

$$0.630 \pm (1.96)\left(\frac{1}{\sqrt{240}}\right)$$

$$i.e. \quad 0.630 \pm 0.127 \text{ or } 0.503 \text{ to } 0.757$$

Thus $0.503 < \mu_z < 0.757$

Using Fisher's z-table, we find values ρ_L and ρ_U, that correspond to Fisher's z-values equal to 0.503 and 0.757 respectively. Thus $\rho_L = 0.465$ and $\rho_U = 0.640$.

Hence the approximate 95% confidence interval for ρ is (0.465, 0.640).

Q.21.18. (a) The 95% confidence interval for

$$\mu_z = 1.1513 \log \frac{1+\rho}{1-\rho}, \text{ is given by}$$

$$z_f \pm \frac{1.96}{\sqrt{n-3}}, \text{ where}$$

$z_f = 1.1513 \log \frac{1.781}{0.269} = 0.9309$, $z_{0.025} = 1.96$ and $n = 25$.

Now $0.9309 \pm \frac{1.96}{\sqrt{22}}$ or 0.9309 ± 0.4179 gives (0.513, 1.3488)

From Fisher's z-table $\rho_L = 0.47$ and $\rho_U = 0.87$ approximately.

Hence the approximate 95% C.I. for ρ is (0.47, 0.87).

(b) The 98% confidence interval for $\mu_z = 1.1513 \log \frac{1+\rho}{1-\rho}$, is given by

$$z_f \pm \frac{2.326}{\sqrt{n-3}}, \text{ where}$$

$z_f = 1.1513 \log \frac{1+0.92}{1-0.92} = 1.5890$, $n = 20$ and $z_{0.01} = 2.326$.

Now $1.5890 \pm \frac{2.326}{\sqrt{17}}$ or 1.5890 ± 0.564 gives (1.025, 2.153)

From Fisher's table, $\rho_L = 0.771$ and $\rho_U = 0.973$ approximately.

Hence the approximate 98% confidence interval for ρ is (0.771, 0.973).

Q.21.19. (a) (i) The hypotheses would be stated as

$$H_0 : \rho = 0.7 \text{ and } H_1 : \rho \neq 0.7$$

(ii) We use a level of significance of $\alpha = 0.05$, and a two-sided test.

(iii) The test-statistic would be

$$Z = \frac{Z_f - \mu_z}{1/\sqrt{n-3}}, \text{ where}$$

$$Z_f = 1.1513 \log \frac{1+r}{1-r}, \text{ and}$$

$$\mu_z = 1.1513 \log \frac{1+\rho}{1-\rho}$$

The statistic Z is approximately standard normal.

(iv) Computations.

Now $Z_f = 1.1513 \log \frac{1.6}{0.4}$

$= (1.1513)\,(0.6021) = 0.6932$, and

$\mu_z = 1.1513 \log \frac{1 + 0.7}{1 - 0.7}$

$= (1.1513)\,(0.7533) = 0.8673$ so that

$$z = \frac{0.6932 - 0.8673}{\frac{1}{\sqrt{50-3}}} = \frac{-0.1741}{\frac{1}{6.856}} = -1.19$$

(v) The critical region is $Z < -1.96$ and $Z > 1.96$.

(vi) The computed value of Z does not fall in the critical region and we cannot reject H_0.

(b) (i) The hypotheses would be stated as

$$H_0 : \rho = 0.4 \text{ and } H_1 : \rho \neq 0.4.$$

(ii) We use a level of significance of $\alpha = 0.05$, and a two-sided test.

(iii) The test-statistic would be

$$Z = \frac{Z_f - \mu_z}{1/\sqrt{n-3}},$$

where $Z_f = 1.1513 \log \frac{1+r}{1-r}$ and $\mu_z = 1.1513 \log \frac{1+\rho}{1-\rho}$

and the Z-statistic is a S.N.V.

(iv) Computations. Now

$$Z_f = 1.1513 \log \frac{1.6}{0.4} = 1.1513\ (0.6021) = 0.6932, \text{ and}$$

$$\mu_z = 1.1513 \log \frac{1.4}{0.6} = 1.1513\ (0.3680) = 0.4237, \text{ so that}$$

$$z = \frac{0.6932 - 0.4237}{1/\sqrt{39-3}} = (0.2695)\ (6) = 1.62$$

(v) The critical region is $Z < -1.96$ and $Z > 1.96$.

(vi) The computed value of Z does not fall in the critical region. We cannot reject H_0. We may conclude that ρ would be 0.4.

Q.21.20. (a) (i) We state the hypotheses as

$$H_0 : \rho_1 = \rho_2 \text{ and } H_1 : \rho_1 \neq \rho_2.$$

(ii) We use a level of significance of $\alpha = 0.05$, and a two sided test.

(iii) The test-statistic would be

$$Z = \frac{Z_{f_1} - Z_{f_2}}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

where $Z_{f_1} = 1.1513 \log \frac{1 + r_1}{1 - r_1}$, $Z_{f_2} = 1.1513 \log \frac{1 + r_2}{1 - r_2}$.

and Z-statistic under H_0 is a S.N.V.

(iv) Computations.

$$Z_{f_1} = 1.1513 \log \frac{1.55}{0.45} = 1.1513\ (0.5371) = 0.62$$

$$Z_{f_2} = 1.1513 \log \frac{1.75}{0.25} = 1.1513\ (0.8451) = 0.97 \text{ so that}$$

$$z = \frac{0.62 - 0.97}{\sqrt{\frac{1}{25} + \frac{1}{16}}} = \frac{-0.35}{\sqrt{0.04 + 0.0625}} = \frac{-0.35}{0.32} = -1.09.$$

(v) The critical region is $Z < -1.96$ and $Z > 1.96$.

(vi) The computed value of Z lies in the acceptance region, so we do not reject our hypothesis, *i.e.* the given values of r are consistent with the hypothesis that the samples are drawn from the same population.

(b) (i) The hypotheses would be stated as

$$H_0 : \rho_1 = \rho_2 \text{ and } H_1 : \rho_1 \neq \rho_2.$$

(ii) We use a level of significance of $\alpha = 0.05$, and a two-sided test.

(iii) The test-statistic would be

$$Z = \frac{Z_{f_1} - Z_{f_2}}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}, \text{ where}$$

$$Z_{f_1} = 1.1513 \log \frac{1 + r_1}{1 - r_1} \text{ and } Z_{f_2} = 1.1513 \log \frac{1 + r_2}{1 - r_2}$$

The Z-statistic under H_0 is a S.N.V.

(iv) Computations. Now

$$Z_{f_1} = 1.1513 \log \frac{1+0.72}{1-0.72} = (1.1513)\ (0.7884) = 0.91, \text{ and}$$

$$Z_{f_2} = 1.1513 \log \frac{1+0.84}{1-0.84} = (1.1513)(1.0607) = 1.22, \text{ so that}$$

$$z = \frac{0.91 - 1.22}{\sqrt{\frac{1}{64} + \frac{1}{36}}} = \frac{-0.31}{0.208} = -1.49$$

(v) The critical region is $Z < -1.96$ and $Z > 1.96$.

(vi) **Conclusion.** Since the computed value of Z does not fall in the critical region, we therefore cannot reject H_0. The samples may be considered as coming from populations having equal coefficients of correlation.

Q.21.21. (b) (i) We are required to test the hypotheses

$$H_0 : \rho = 0 \text{ and } H_1 : \rho \neq 0.$$

(ii) We set the level of significance at $\alpha = 0.05$, and use a two-sided test.

(iii) The test-statistic under H_0 would be

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}.$$

This statistic has a t-distribution with $(n-2)$ *d.f.* when the population is normally distributed.

(iv) Computation.

$$t = \frac{0.45\sqrt{28-2}}{\sqrt{1-(0.45)^2}} = \frac{(0.45)\,(5.099)}{\sqrt{0.7975}} = \frac{2.2946}{0.893} = 2.57$$

(v) The critical region is $|t| \geq t_{.025(26)} = 2.056$

(vi) Since the computed value of t falls in the critical region, so we reject H_0 is favour of H_1.

Q.21.22 (a) (i) The hypotheses are stated as

$$H_0 : \rho = 0.85 \text{ and } H_1 : \rho \neq 0.85.$$

(ii) We choose the significance level at $\alpha = 0.05$.

(iii) The test-statistic is

$$Z = \frac{Z_f - Z_z}{1/\sqrt{n-3}},$$

where $Z_f = 1.1513 \log \dfrac{1+r}{1-r}$ and

$\mu_z = 1.1513 \log \dfrac{1+\rho_0}{1-\rho_0}$. The variable Z is approximately standard normal.

(iv) Computations. Here $n = 10$, $r = 0.7$.

Now $Z_f = 1.1513 \log \dfrac{1+0.7}{1-0.7} = 1.1513 \log \dfrac{1.7}{0.3} = 0.87$, and

$$\mu_z = 1.1513 \log \frac{1+0.85}{1-0.85} = 1.1513 \log \frac{1.85}{0.15} = 1.26$$

$$\therefore \quad z = \frac{0.87 - 1.26}{1/\sqrt{10-3}} = (-0.39)\,(2.646) = -1.03$$

(v) The critical region is $|Z| \geq 1.96$.

(vi) Since the computed value $z = -1.02$ does not fall in the critical region, so we accept H_0 and conclude that the correlation coefficient in the population might be 0.85.

(b) (i) The hypotheses are stated as

$$H_0 : \rho = 0 \text{ and } H_1 : \rho \neq 0$$

(ii) The significance level is chosen at $\alpha = 0.05$.

(iii) The test-statistic to use is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

which, if H_0 is true, has a t-distribution with $(n-2)$ *d.f.*

(iv) **Computations.** Substituting the values, we get

$$t = \frac{0.7\sqrt{10-2}}{\sqrt{1-0.49}} = \frac{1.98}{0.71} = 2.79.$$

(v) The critical region is $|t| \geq t_{0.025,(8)} = 2.306$.

(vi) **Conclusion.** Since the computed value falls in the critical region, so we reject H_0 and may conclude that the value of r is significant.

(c) (i) The hypotheses are stated as

$$H_0 : \rho_1 = \rho_2 \text{ and } H_1 : \rho_1 \neq \rho_2$$

(ii) The significance level is at $\alpha = 0.05$.

(iii) The test-statistic under H_0 is

$$Z = \frac{Z_{f_1} - Z_{f_2}}{\sqrt{\dfrac{1}{n_1-3} + \dfrac{1}{n_2-3}}},$$

where $Z_{f_1} = 1.1513 \log \frac{1+r_1}{1-r_1}$ and $Z_{f_2} = 1.1513 \log \frac{1+r_2}{1-r_2}$

The variable Z is approximately standard normal.

(iv) The critical region is $|Z| \geq 1.96$.

(v) Computations.

$$Z_{f_1} = 1.1513 \log \frac{1+0.7}{1-0.7} = 1.1513 \log \frac{1.7}{0.3} = 0.87, \text{ and}$$

$$Z_{f_2} = 1.1513 \log \frac{1+0.9}{1-0.9} = 1.1513 \log \frac{1.9}{0.1} = 1.472.$$

$$z = \frac{0.87 - 1.472}{\sqrt{\frac{1}{7} + \frac{1}{9}}} = \frac{-0.602}{0.504} = -1.19$$

(vi) **Conclusion.** Since the computed value $z = -1.19$ does not fall in the critical region, so we accept H_0.

Q.21.23. (a) (i) The hypotheses are stated as

$$H_0 : \rho = 0 \text{ and } H_1 : \rho \neq 0$$

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic to use is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}},$$

which, if H_0 is true, has a t-distribution with $(n-2)$ *d.f.*

(iv) **Computations.** Substituting the values, we get

$$t = \frac{0.55\sqrt{20-2}}{\sqrt{1-(0.55)^2}} = \frac{(0.55)\ (4.2426)}{\sqrt{0.6975}}$$

$$= \frac{2.3334}{0.8352} = 2.79.$$

(v) The critical region is $|t| \geq t_{0.025,(18)} = 2.10$

(vi) **Conclusion.** Since the computed value falls in the critical region, so we reject H_0.

(b) (i) The hypotheses are stated as

$$H_0 : \rho = 0 \text{ and } H_1 : \rho \neq 0.$$

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic under H_0 is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}},$$

which has a t-distribution with $(n-2)$ *d.f.*

(iv) **Computations.** Substituting the values, we get

$$t = \frac{0.30\sqrt{27-2}}{\sqrt{1-0.09}} = \frac{1.5}{0.95} = 1.58.$$

(v) The critical region is $|t| \geq t_{0.025,(25)} = 2.06$

(vi) **Conclusion.** Since the computed value of $t = 1.56$ does not fall in the critical region, we therefore accept H_0 and may conclude that the coefficient of correlation in the population is zero.

Q.21.24. (a) (i) We state our hypotheses as

$$H_0 : \rho = 0 \text{ and } H_1 : \rho \neq 0.$$

(ii) The significance level is set at $\alpha = 0.01$.

(iii) The test-statistic under H_0 is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}},$$

which has a t-distribution with $(n-2)$ *d.f.*

(iv) **Computations.** Substituting the values, we get

$$t = \frac{0.32\sqrt{12-2}}{\sqrt{1-(0.32)^2}} = \frac{(0.32)\ (3.1623)}{0.9474} = 1.07$$

(v) The critical region is $|t| \geq t_{0.005,(10)} = 3.169$

(vi) **Conclusion.** Since the computed value does not fall in the critical region, we therefore accept H_0.

(b) (i) We state the hypotheses as

$$H_0 : \rho = 0 \text{ and } H_1 : \rho \neq 0$$

(ii) We use a significance level of $\alpha = 0.05$

(iii) The test-statistic under H_0 is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}},$$

which has Student's distribution with $(n-2)$ *d.f.* when the population is normally distributed.

(iv) The critical region is $|t| \geq t_{0.025,(17)} = 2.11$.

(v) **Computations.** Here $n = 19$ and $r = 0.65$.

$$\therefore \quad t = \frac{0.65\sqrt{19-2}}{\sqrt{1-(0.65)^2}} = \frac{(0.65)\,(4.123)}{0.76} = 3.53.$$

(vi) **Conclusion.** Since the computed value of t falls in the critical region, so we reject H_0 and conclude that the correlation coefficient in the population does not appear to be zero.

(c) (i) The hypotheses are stated as

$$H_0 : \rho = 0 \text{ and } H_1 : \rho \neq 0.$$

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic to use is

$$F = \frac{r^2\,(n-2)}{1-r^2}$$

which, if H_0 is true, has an F-distribution with $v_1 = 1$ and $v_2 = n-2$ *d.f.*

(iv) Computations. Here $n = 20$ and $r = 0.78$.

$$\therefore \quad F = \frac{(0.78)^2\,(18)}{1-(0.78)^2} = \frac{10.9512}{0.3916} = 27.97.$$

(v) The critical region is $F \geq F_{0.05(1,18)} = 4.41$.

(vi) **Conclusion.** Since the computed value of $F = 27.97$ falls in the critical region, we therefore reject H_0.

Q.21.25. (a) The test statistic for the significance of r is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}},$$

Substituting the value of n, we obtain

$$t = \frac{r\sqrt{38-2}}{\sqrt{1-r^2}} = \frac{6r}{\sqrt{1-r^2}}$$

(i) For a significant value of r at the 0.05 level, t should be $> t_{0.025,(36)} = 2.03$.

That is $\dfrac{6r}{\sqrt{1-r^2}} \geq 2.03$

or $\quad 36r^2 \geq (2.03)^2\,(1-r^2)$

or $\quad 40.1209r^2 \geq 4.1209$ or $r^2 \geq 0.1027$

$\therefore \quad |r| \geq 0.32$.

Hence the required least value (significant at $\alpha = 0.05$) of $r = 0.32$.

(ii) For a significant value of r at $\alpha = 0.01$, t should be $\geq t_{0.05,(36)} = 2.72$.

Thus $\dfrac{6r}{\sqrt{1-r^2}} \geq 2.72$ or $36r^2 \geq (2.72)^2\,(1-r^2)$

or $\quad 43.3984r^2 \geq 7.3984$ or $r^2 \geq 0.1705$

$\therefore \quad |r| \geq 0.41$.

Hence the required least value significant at $\alpha = 0.01$ of $r = 0.41$.

(b) For a significant value of r at the 5% level, t should be $\geq t_{0.025,(25)} = 2.06$, *i.e.*,

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \geq t_{0.025(25)}$$

or $\frac{r\sqrt{27-2}}{\sqrt{1-r^2}} \geq 2.06$ or $\frac{5r}{\sqrt{1-r^2}} \geq 2.06.$

or $25r^2 \geq (2.06)^2 (1-r^2)$ or $29.2436r^2 \geq 4.243$

or $r^2 \geq 0.1451$

$\therefore$ $|r| \geq 0.38.$

Hence the required last value of $r = 0.38$.

(c) To find n, the number of pairs of observations $\frac{0.42\sqrt{n-2}}{\sqrt{1-(0.42)^2}}$ should be greater than 2.72.

i.e. $\frac{0.42\sqrt{n-2}}{\sqrt{0.8236}} \geq 2.72$

i.e. $(0.42)^2 (n-2) \geq (2.72)^2 (0.8236)$

i.e. $0.1764n \geq 6.0933 + 0.3528$

i.e. $n \geq 36.54.$

Hence the required number of pairs of observations that must be included in the sample to get the given results = 37.

Q.21.26. (b) The values of r_i are first converted to Z_{f_i} values by the relation $Z_{f_i} = 1.1513 \log \frac{1+r_i}{1-r_i}$. Thus

$$Z_{f_1} = 1.1513 \log \frac{1+0.3}{1-0.3} = 1.1513 \log 1.8571 = 0.3097.$$

$$Z_{f_2} = 1.1513 \log \frac{1+0.4}{1-0.4} = 1.1513 \log 2.3333 = 0.3197.$$

$$Z_{f_3} = 1.1513 \log \frac{1+0.49}{1-0.49} = 1.1513 \log 2.9216 = 0.5360.$$

The other computations are given below:

r_i	Z_{f_i}	n_i-3	$(n_i-3)\, Z_{f_i}$	$(n_i-3)\, Z_{f_i}^2$
0.30	0.3097	7	2.1679	0.6714
0.40	0.3197	12	3.8364	1.2265
0.49	0.5360	17	9.1120	4.8840
Σ	--	36	15.1163	6.7819

$$\therefore \quad \bar{Z} = \frac{\sum (n_i-3)\, Z_{f_i}}{\sum (n_i-3)} = \frac{15.1163}{36} = 0.4199$$

The combined estimate of ρ corresponding to $\bar{Z} = 0.4199$ from tables is found as $\rho = 0.397$.

To test the homogeneity; we proceed as follows:

(i) We state our null hypotheses as

$H_0 : \rho_1 = \rho_2 = \rho_3$, *i.e.* the population correlation coefficients are equal, and

H_1 : Not all correlation coefficients are equal.

(ii) We use a level of significance of $\alpha = 0.05$.

(iii) The test-statistic would be

$$u = \sum (n_i-3)\, (Z_{f_i} - \bar{Z})^2$$

which is distributed approximately as a chi-square distribution with $(k-1)$ *d.f.*

(iv) Computation of test statistic.

$$u = \sum (n_i - 3)\, (Z_{f_i} - \bar{Z})^2$$

$$= \sum (n_i - 3)\, Z_{f_i}^2 - \frac{[\sum (n_i-3)\, Z_{f_i}]^2}{\sum (n_i-3)}$$

$$= 6.7819 - \frac{(15.1163)^2}{36} = 6.7819 - 6.3473 = 0.4346.$$

(v) The critical region is $u \geq \chi^2_{0.05,(2)} = 5.99.$

(vi) **Conclusion.** Since the computed value of u does not fall in the critical region, so we do not reject H_0. The correlation coefficients are regarded as homogeneous.

Q.21.27. The values of r_i are first coverted to Z_{f_i} values by the relations.

$$Z_{f_i} = 1.1513 \log \frac{1 + r_i}{1 - r_i} \quad \textbf{(Fisher's } z\textbf{-transformation)}$$

Thus

$$Z_{f_1} = 1.1513 \log \frac{1+0.41}{1-0.41} = 1.1513 \log 2.3898 = 1.1513 \times 0.3784 = 0.4357,$$

$$Z_{f_2} = 1.1513 \log \frac{1+0.60}{1-0.60} = 1.1513 \log 4.0000 = 1.1513 \times 0.6021 = 0.6932,$$

$$Z_{f_3} = 1.1513 \log \frac{1+0.51}{1-0.51} = 1.1513 \log 3.0816 = 1.1513 \times 0.4888 = 0.5628,$$

$$Z_{f_4} = 1.1513 \log \frac{1+0.48}{1-0.48} = 1.1513 \log 2.8462 = 1.1513 \times 0.4542 = 0.5229.$$

The other computations are given below:

r_i	Z_{f_i}	$n_i - 3$	$(n_i-3)\, Z_{f_i}$	$(n_i-3)\, Z_{f_i}^2$
0.41	0.4357	17	7.4069	3.2272
0.60	0.6932	27	18.7164	12.9742
0.51	0.5628	37	20.8236	11.7195
0.48	0.5229	47	24.5763	12.8509
Σ	--	128	71.5232	40.7718

$$\therefore \quad \bar{Z} = \frac{\Sigma (n_i - 3)\, Z_{f_i}}{\Sigma (n_i - 3)} = \frac{71.5232}{128} = 0.5588$$

The combined estimate of population correlation, that corresponds to 0.5588 from Fisher's z-values table is found as $\rho = 0.508$.

To test their homogeneity, we proceed as follows:

(i) We state our null hypothesis as

H_0 : The samples come from the same bivariate normal population (or that the correlation coefficients are equal).

(ii) We use a level of significance of $\alpha = 0.05$.

(iii) The test-statistic is

$$u = \Sigma (n_i - 3)\, (Z_{f_i} - \bar{Z})^2$$

which is distributed approximately as a χ^2-distribution with $(k-1)$ *d.f.*

(iv) Computations.

$$u = \Sigma (n_i - 3)\, (Z_{f_i} - \bar{Z})^2$$

$$= \Sigma (n_i - 3)\, Z_{f_i}^2 - \frac{[\Sigma (n_i - 3)\, Z_{f_i}]^2}{\Sigma (n_i - 3)}$$

$$= 40.7718 - \frac{(71.5232)^2}{128} = 40.7718 - 39.9654 = 0.8064$$

(v) The critical region is $u \geq \chi^2_{0.05,(3)} = 7.82$.

(vi) **Conclusion.** Since the computed value of u does not fall in the critical region, so we accept H_0. We are reasonably confident that the samples come from the same bivariate pouplation.

Q.21.28. (a) (i) The hypotheses would be stated as

$H_0 : \rho_{12.34} = 0$, *i.e.* a partial correlation of order two in the population is zero, and

$H_1 : \rho_{12.34} \neq 0$.

(ii) We use a significance level of $\alpha = 0.05$, and a two-sided test.

(iii) The test-statistic under H_0 would be

$$t = \frac{r_{12.34} \sqrt{n-k-2}}{\sqrt{1 - r_{12.34}^2}}$$

which has the Student's t-distribution with $(n-k-2)$ $d.f.$ Here $k=2$, the order of the partial correlation.

(iv) Computation.

$$t = \frac{0.5\sqrt{20-2-2}}{\sqrt{1-(0.5)^2}} = \frac{(0.5)\,(4)}{0.866} = 2.31.$$

and $d.f. = 20-2-2 = 16.$

(v) The critical region is $|t| \geq t_{.025(16)} = 2.12.$

(vi) Since the computed value of t falls in the critical region, so we reject H_0 and conclude that it is significantly different from zero.

(b) (i) The hypotheses would be stated as

$H_0 : \rho_{12.34} = 0$, *i.e.* a partial correlation of order two in the population is zero, and

$H_1 : \rho_{12.34} \neq 0.$

(ii) The significance level is set at $\alpha = 0.05.$

(iii) The test-statistic under H_0 would be

$$t = \frac{r_{12.34}\sqrt{n-k-2}}{\sqrt{1-r^2_{12.34}}},$$

which has Student's t-distribution with $(n-k-2)$ $d.f.$

(iv) Computations.

$$t = \frac{0.48\sqrt{25-2-2}}{\sqrt{1-(0.48)^2}} = \frac{0.48\sqrt{21}}{\sqrt{0.7626}} = \frac{2.1984}{0.8773} = 2.51, \text{ and}$$

$d.f. = 25-2-2 = 21.$

(v) The critical region is $|t| \geq t_{0.025,(21)} = 2.08$

(vi) **Conclusion.** Since the computed value of t falls in the critical region, we therefore reject H_0. We conclude that this is not consistent with the hypothesis that the corresponding partial correlation in the population is zero.

Q.21.29. (a) (i) The hypotheses are stated as

$$H_0 : \rho_{12.34} = 0 \text{ and } \rho_{12.34} \neq 0$$

(ii) We use a significance level of $\alpha = 0.05.$

(iii) The test-statistic under H_0 is

$$t = \frac{r_{12.34}\sqrt{n-k-2}}{\sqrt{1-r^2_{12.34}}}$$

which has a t-distribution with $(n-k-2)$ $d.f.$

(iv) Computations.

$$t = \frac{0.51\sqrt{20-2-2}}{\sqrt{1-(0.51)^2}} = \frac{2.04}{0.8602} = 2.37, \text{ and}$$

$d.f. = 20-2-2 = 16.$

(v) The critical region is $|t| \geq t_{0.025(16)} = 2.12$

(vi) **Conclusion.** Since the computed value of t falls in the critical region, we therefore reject H_0.

Use of F-test:

$H_0 : \rho_{12.34} = 0$ and $H_1 : \rho_{12.34} \neq 0$

$$F = \frac{r^2_{12.34}\,(n-k-2)}{1-r^2_{12.34}} = \frac{(0.51)^2\,(16)}{1-(0.51)^2} = \frac{4.1616}{0.7399} = 5.62\ (=t^2)$$

which exceeds $F_{0.05(1,16)} = 4.49$. Hence we reject H_0.

(b) The values of partial correlation coefficients are first converted to z-values by Fisher's z-transformation. Thus

$$Z_{f_1} = 1.1513 \log \frac{1+0.38}{1-0.38} = 1.1513 \log 2.2258 = 1.1513 \times 0.3476 = 0.4002,$$

$$Z_{f_2} = 1.1513 \log \frac{1+0.54}{1-0.54} = 1.1513 \log 3.3478 = 1.1513 \times 0.5247 = 0.6041,$$

$$Z_{f_3} = 1.1513 \log \frac{1+0.60}{1-0.60} = 1.1513 \log 4.0000 = 1.1513 \times 0.6021 = 0.6932,$$

$$Z_{f_4} = 1.1513 \log \frac{1+0.50}{1-0.50} = 1.1513 \log 3.0000 = 1.1513 \times 0.4771 = 0.5493,$$

$$Z_{f_5} = 1.1513 \log \frac{1+0.42}{1-0.42} = 1.1513 \log 2.4483 = 1.1513 \times 0.3888 = 0.4476.$$

Calculation for $\bar{Z}$ etc.

Partial correlation	Z_{f_i}	n_i-p-3	$(n_i-p-3)Z_{f_i}$	$(n_i-p-3)Z_{f_i}^2$
0.38	0.4002	18	7.2036	2.8829
0.54	0.6041	23	13.8483	8.3658
0.60	0.6932	27	18.7164	12.9742
0.50	0.5493	32	17.5776	9.6554
0.42	0.4476	36	16.6136	7.4362
Σ	--	136	73.4595	41.3145

$$\therefore \quad \bar{Z} = \frac{\Sigma (n_i-p-3) Z_{f_i}}{\Sigma (n_i-p-3)} = \frac{73.4595}{136} = 0.5401,$$

where p denotes the number of variables being held constant.

The pooled estimate that corresponds to $\bar{Z} = 0.5401$ from tables, is 0.493.

To test their homogeneity, we proceed as follows:

(i) We state our null hypothesis as

H_0 : The partial correlation coefficients are homogeneous.

(ii) We use a level of significance of $\alpha = 0.05$.

(iii) The test-statistic is

$$u = \Sigma(n_i-p-3)\,(Z_{f_i}-\bar{Z})^2$$

which is approximated as a chi-square distribution with *d.f.* one less than the number of correlation co-efficients.

(iv) Computations.

$$u = \Sigma(n_i-p-3)\,(Z_{f_i}-\bar{Z})^2$$

$$= \Sigma(n_i-p-3)\,Z_{f_i}^2 - \frac{[\Sigma\,(n_i-p-3)\,Z_{f_i}]^2}{\Sigma\,(n_i-p-3)}$$

$$= 41.3145 - \frac{(73.4595)^2}{136} = 41.3145 - 39.6787 = 1.6358.$$

(v) The critical region is $u \geq \chi^2_{0.05,(4)} = 9.49$.

(vi) We accept H_0 as the computed value of u does not fall in the critical region. Hence the data do not provide evidence to indicate a difference in the correlation coefficients.

Q.21.30. (a) We get (see solution to Q.11.5. Part-I).

$$r_{12} = -0.89,\ r_{23} = 0.96,\ r_{31} = -0.97.$$

$$\text{Now} \quad r_{13.2} = \frac{r_{13} - r_{12}\,r_{23}}{\sqrt{(1-r_{12}^2)\,(1-r_{23}^2)}} = \frac{(-0.97) - (-0.89)(0.96)}{\sqrt{1-(0.89)^2}\sqrt{1-(0.96)^2}}$$

$$= \frac{-0.97 + 0.8544}{\sqrt{(0.2079)\,(0.0784)}} = \frac{-0.1156}{0.1277} = -0.905;$$

$$R_{1.23}^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}\,r_{23}\,r_{31}}{1-r_{23}^2}$$

$$= \frac{(-0.89)^2 + (-0.97)^2 - 2(-0.89)(0.96)(-0.97)}{1-(0.96)^2}$$

$$= \frac{0.7921 + 0.9409 - 1.6575}{1-0.9216} = \frac{0.0755}{0.0784} = 0.9630$$

$$\therefore \quad R_{1.23} = \sqrt{0.9630} = 0.98.$$

(b) To test each one of these correlation coefficients for a significance at 5% level, we use the F-test.

(i) The value of F-ratio for $r_{12} = -0.89$, is

$$F = \frac{r_{12}^2\,(n-2)}{1-r_{12}^2} = \frac{(-0.89)^2\,(6-2)}{1-(-0.89)^2}$$

$$= \frac{3.1684}{0.2079} = 15.24; \qquad (t = 3.90)$$

(ii) The value of F-ratio for $r_{13.2} = -0.905$, i

$$F = \frac{r_{12.3}^2 (n-k-2)}{1 - r_{12.3}^2} = \frac{(-0.905)^2 (6-1-2)}{1 - (-0.905)^2}$$

$$= \frac{0.8190 \times 3}{1 - 0.8190} = \frac{2.457}{0.181} = 13.57; \qquad (t = 3.68)$$

(iii) The value of F-ratio for $R_{1.23} = 0.98$ is

$$F = \frac{R_{1.23}^2 (n-p-1)}{(1 - R_{1.23}^2) p} = \frac{(0.963)^2 (6-1-2)}{(1 - 0.963)\ (2)}$$

$$= \frac{2.889}{0.074} = 39.04$$

The critical regions are (i) $F \geq F_{0.05,(1,4)} = 7.71$

(ii) $F \geq F_{0.05,(1,3)} = 10.13$

(iii) $F \geq F_{0.05,(2,3)} = 9.55$

Conclusion. Since the computed value of F in each case falls in the critical region so the correlation coefficients are significant.

9.21.31.(a) (i) The hypotheses are stated as

H_0: The multiple correlation coefficient in the population is zero, *i.e.* $R_{1.23} = 0$, and

$H_1 : R_{1.23} \neq 0.$

(ii) The significance level to use is $\alpha = 0.05$.

(iii) The test-statistic under H_0, is

$$F = \frac{R_{1.23}^2 (n-p-1)}{(1 - R_{1.23}^2)\ (p)},$$

which has an F-distribution with $v_1 = p$ and $v_2 = n-p-1$ *d.f.* Here $p = 2$.

(iv) Computations. Substituting the values, we get

$$F = \frac{(0.35)^2 (20-2-1)}{[1-(0.35)^2]\ (2)} = \frac{2.0825}{1.7550} = 1.19.$$

(v) The critical region is $F \geq F_{0.05(2,17)} = 6.11$

(vi) **Conclusion.** Since the computed value of F does not fall in the critical region, we therefore accept H_0, and may conclude that the multiple correlation coefficient in the population does not differ from zero significantly.

(b) (i) We state the hypotheses as

$H_0 : R_{1.234} = 0$, *i.e.* the multiple correlation coefficient in the population is zero, and

$H_1 : R_{1.234} \neq 0.$

(ii) We use a level of significance of $\alpha = 0.05$.

(iii) The test-statistic under H_0, is

$$F = \frac{R_{1.234}^2 (n-p-1)}{(1 - R_{1.234}^2)\ (p)},$$

which has an F-distribution with $v_1 = p$ and $v_2 = n-p-1$ *d.f.*

(iv) Computations. Here $n = 25$, $R_{1.234} = 0.4$, $p = 3$.

$$\therefore \quad F = \frac{(0.4)^2 (25-3-1)}{(1-0.16)\ (3)} = \frac{3.36}{2.52} = 1.33, \text{ and}$$

$v_1 = 3, \quad v_2 = 21.$

(v) The critical region is $F \geq F_{0.05(3,21)} = 3.07$

(vi) **Conclusion.** Since the computed value of $F = 1.33$ does not fall in the critical region, so we do not reject H_0. Thus the given value of $R_{1.234}$ does not differ significantly from zero.

Q.21.32. (b) Total $SS = \sum(Y-\bar{Y})^2 = 1539$

Regression $SS = b^2\sum(X-\bar{X})^2$, where $b = \dfrac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2}$

$$= \left(\frac{1532}{1561}\right)^2 \times 1561 = \frac{(1532)^2}{1561} = \frac{2347024}{1561} = 1503.5$$

$\therefore$ Residual SS = Total SS − Regression SS

$= 1539 - 1503.5 = 35.5.$

Thus the Linear regression analysis is

Source of variation	*d.f.*	Sum of Squares	*MS*	*F*
Regression	1	1503.5	1503.5	338.6
Residual	8	35.5	4.44	--
Total	9	1539.0	--	--

(i) We are required to test the hypotheses

$H_0 : \beta = 1$ against $H_1 : \beta \neq 1.$

(ii) We use a significance level of $\alpha = 0.01.$

(iii) The test-statistic is $t = \dfrac{b-\beta_0}{s_b}$,

where $s_b^2 = \dfrac{\sum(Y-\hat{Y})^2}{(n-2)\sum(X-\bar{X})^2}$.

This statistic has a Student's distribution with $(n-2)$ *d.f.*

(iv) The necessary computations are given below:

$$b = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2} = \frac{1532}{1561} = 0.98$$

$$s_b^2 = \frac{35.5}{8\times1561} = 0.002843, \text{ so that } s_b = \sqrt{0.002843} = 0.053$$

$\therefore$ $$t = \frac{0.98-1}{0.053} = \frac{-0.02}{0.053} = -0.38.$$

(v) The critical region is $|t| \geq t_{0.005,(8)} = 3.36.$

(vi) We accept the hypothesis as the computed value of t is less than the corresponding critical value of t at the 0.01 level of significance.

Q.21.33. (b) From Q. 21.9, we find that

$\sum X = 329$, $\sum Y = 778$, $\sum XY = 51660$,

$\sum Y^2 = 123214$, $b = 4.46$, and $a = -137.87.$

Now Residual $SS = \sum(Y-\hat{Y})^2 = \sum Y^2 - a\sum Y - b\sum XY$

$= 123314-(-137.87)(778)-(4.46)(51660)$

$= 73.26$; and

Total $$SS = \sum(Y-\bar{Y})^2 = \sum Y^2 - \frac{(\sum Y)^2}{n}$$

$$= 123214 - (778)^2/5 = 2157.2$$

$\therefore$ Regression $SS = 2157.2 - 73.26 = 2083.94.$

The analysis of variance table for testing the regression is

Source of variation	*d.f.*	*SS*	*MS*	*F*
Regression	1	2083.9	2083.94	85.34
Residual	3	73.2	24.42	--
Total	4	2157.20	--	--

Q.21.34. (i) Here $b_1 = \dfrac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})} = \dfrac{-800}{100} = -8$, and

$b_0 = \bar{Y} - b\bar{X} = 4 - (-8)(6) = 90$

Hence $\hat{Y} = 90 - 8X.$

(ii) $\sum(Y-\bar{Y})^2 = \sum(Y-\hat{Y})^2 + \sum(\hat{Y}-\bar{Y})^2$, *i.e.*,

Total SS = SS about regression line + SS of regression line about mean.

Now $\sum(Y-\bar{Y})^2 = 10{,}000.$

$$\Sigma(\hat{Y}-\bar{Y})^2 = b_1^2 \Sigma(X-\bar{X})^2 = \left[\frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\Sigma(X-\bar{X})^2}\right]^2 \Sigma(X-\bar{X})^2$$

$$= \left(\frac{-800}{100}\right)^2 \times 100 = 6,400.$$

$$\therefore \quad \Sigma(Y-\hat{Y})^2 = 10,000 - 6,400 = 3,600.$$

We may put them in the analysis of variance table as below:

Source of variation	*d.f.*	*SS*	*MS*	*F*
Regression	1	6,400	6,400	64
Residual	36	3,600	100	--
Total	37	10,000	--	--

(iii) We wish to test the hypothesis

$H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$.

The significance level is set at $\alpha = 0.05$.

The test-statistic is $t = \dfrac{b_1 - \beta_1}{s_b}$,

where $s_b^2 = \dfrac{\Sigma(Y-\hat{Y})^2}{(n-2)\,\Sigma(X-\bar{X})^2}$, and which under the usual assumptions, conform to t-distribution with $n-2=36$ degrees of freedom.

Computations. $s_b^2 = \dfrac{3600}{36 \times 100} = 1$, so that $s_b = 1$.

$$\therefore \quad t = \frac{-8-0}{1} = -8$$

The critical region is $|t| \geq t_{0.025,(36)} = 2.03$

Since the computed value of $t = -8$ falls in the critical region, we therefore reject H_0.

(iv) Adjusted $Y = Y - b_1 (X - \bar{X})$

$$= 36 - (-8)(8-6) = 52$$

(v) Here b_1 is an estimate of β_1, the population linear regression co-efficient. The estimate b_1 is obtained from the sample observations.

Q.21.35. To test the hypothesis $H_0 : \beta_1 = \beta_2 = 0$ in the multiple regression $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, we use F-distribution.

For this purpose, the necessary calculations are to be shown in an analysis of variance table.

The estimated multiple regression is obtained (see solution of question 11.5) as

$$\hat{Y} = 61.40 - 3.65X_1 + 2.54X_2.$$

Also $\Sigma Y = 300$, $\Sigma Y^2 = 19008$, and $R_{Y.12} = 0.9927$.

Now Total $SS = \Sigma(Y-\bar{Y})^2 = \Sigma Y^2 - \dfrac{(\Sigma Y)^2}{n}$

$$= 19,008 - (300)^2 / 6 = 4,008.$$

Regression $SS = \Sigma(\hat{Y}-\bar{Y})^2 = R^2 \Sigma(Y-\bar{Y})^2$

$$= (0.9927)^2 (4008) = 3949.6964,$$

Residual $SS = \Sigma(Y-\hat{Y})^2 = (1-R)^2 \Sigma(Y-\bar{Y})^2$

$$= (0.01455)(4,008) = 58.3164.$$

The analysis of variance table would be as below:

Source of variation	*d.f.*	*SS*	*MS*	*F*
Regression	2	3949.7	1974.85	101.64
Residual	3	58.3	19.43	--
Total	5	4008.0	--	--

The critical region is $F > F_{0.05(2,3)} = 9.55$.

Conclusion. Since the calculated value of F exceeds the table value, we therefore reject $H_0 : \beta_1 = \beta_2 = 0$.

Q.21.36. The equation of the estimated multiple linear regression is $\hat{Y} = a + b_1 X_1 + b_2 X_2$, where a, b_1 and b_2 are the least-squares estimates of α, β_1 and β_2 respectively.

The sums needed to calculate a, b_1 and b_2 are found to be

$$\Sigma Y = 64,\ \Sigma X_1 = 15,\ \Sigma X_2 = 5,\ \Sigma X_1^2 = 49,$$

$$\Sigma X_2^2 = 11,\ \Sigma Y^2 = 894,\ \Sigma X_1 X_2 = 16,\ \Sigma X_1 Y = 203,$$

$$\Sigma X_2 Y = 82,\ \text{and}\ n = 5.$$

Substituting these sums in the normal equations, we get

$$5a + 15b_1 + 5b_2 = 64$$

$$15a + 49b_1 + 16b_2 = 203$$

$$5a + 16b_1 + 11b_2 = 82$$

Solving them simultaneously, we get

$$a = 3.88,\ b_1 = 2.09 \text{ and } b_2 = 2.65$$

as the least squares estimates of the parameters.

To test the overall significance of the regression coefficients, we proceed as below:

(i) We state the hypotheses as

$H_0 : \beta_1 = \beta_2 = 0$, *i.e.* none of the regression coefficients is significant, and

H_1 : At least one of the β_1 and β_2 is non-zero.

(ii) We use a significance level of $\alpha = 0.05$.

(iii) The test-statistic to use is

$$F = \frac{MSR}{MSE},$$

which, if H_0 is true, has an F-distribution with $v_1 = 2$ and $v_2 = n-3$ *d.f.* assuming that population is normally distributed.

(iv) Computations. To set up the ANOVA-table, we find the necessary sums of squares as below:

$$\text{Total } SS = \Sigma(Y-\bar{Y})^2 = \Sigma Y^2 - (\Sigma Y)^2 / n$$

$$= 894 - (64)^2 / 5 = 74.8.$$

$$\text{Regression } SS = \Sigma(\hat{Y}-\bar{Y})^2$$

$$= a\Sigma Y + b_1 \Sigma X_1 Y + b_2 \Sigma X_2 Y - (\Sigma Y)^2/n$$

$$= (3.88)(64) + (2.09)(203) + (2.65)(82) - (64)^2/5$$

$$= 248.32 + 424.27 + 217.3 - 819.2$$

$$= 70.69$$

$$\text{Residual } SS = \text{Total } SS - \text{Regression } SS$$

$$= 74.8 - 70.69 = 4.11$$

The ANOVA-table is

Source of variation	*d.f.*	*SS*	*MS*	*F*
Regression	2	70.69	35.345	17.20
Error	2	4.11	2.055	--
Total	4	74.80	--	--

(v) The critical region is $F > F_{0.05(2,2)} = 19.00$

(vi) **Conclusion.** Since the computed value of $F = 17.20$ does not fall in the critical region, so we accept H_0.

❖❖❖❖❖❖❖❖❖❖

CHAPTER 22

THE ANALYSIS OF COVARIANCE

Q.22.1. Definition of Covariance Analysis. Covariance analysis may be defined as a statistical procedure by means of which we remove the effect of one or more variables (called covariates) on the dependent variable before assessing the effect of the treatments on the dependent variable. It thus consists of the combined application of linear regressior and the anlaysis of variance techniques. The covariance effects are partitioned into their components and a test of the existance of regression is made. Finding evidence of any regression, its effect is removed before tests are made on the significance of the treatments. It is a useful technique to compare treatments in the presence of covariates which can neither be eliminated nor controlled.

Covariance Model for one-way classification.

The appropriate covariance model (fixed effects) is

$$Y_{ij} = \mu + \tau_j + \beta (X_{ij} - \bar{X}.) + \varepsilon_{ij}, \qquad \begin{Bmatrix} i=1,2,\ldots,m \\ j=1,2,\ldots,k \end{Bmatrix}$$

where Y_{ij} = yield of the ith observation of the jth treatment,

μ = the true mean yield,

τ_j = the effect of the jth treatment,

β = the true linear regression between Y and X over all the data,

$X_{ij} - \bar{X}..$ = the deviation of the ijth covariate from the mean of the covariate, and

ε_{ij} = random error component, assumed to be normally and independently distributed with zero mean and common variance.

The model is clearly a combination of the regression model

$$Y_{ij} = \mu + \beta (X_{ij} - \bar{X}..) + \varepsilon_{ij}$$

and a one-way analysis of variance model

$$Y_{ij} = \mu + \tau_j + \varepsilon_{ij}$$

where all the letters have their usual meanings.

Let us consider the regression model in the deviation form

$$y = bx + e$$

wher $y = Y_{ij} - \bar{Y}..$, $x = X_{ij} - \bar{X}..$ and b is the sample slope. Now the sum of the squares of errors of estimate is

$$\begin{aligned} \Sigma e^2 &= \Sigma (y - bx)^2 \\ &= \Sigma y^2 - 2b\Sigma xy + b^2 \Sigma x^2. \end{aligned}$$

The method of least squares yields

$$b = \frac{\Sigma xy}{\Sigma x^2}$$

$$\therefore \quad \Sigma e^2 = \Sigma y^2 - 2\left(\frac{\Sigma xy}{\Sigma x^2}\right)\Sigma xy + \left(\frac{\Sigma xy}{\Sigma x^2}\right)^2 \Sigma x^2$$

$$= \Sigma y^2 - 2\frac{(\Sigma xy)^2}{\Sigma x^2} + \frac{(\Sigma xy)^2}{\Sigma x^2}$$

$$= \Sigma y^2 - \frac{(\Sigma xy)^2}{\Sigma x^2}$$

Thus the term $\frac{(\Sigma xy)^2}{\Sigma x^2}$ which is actually the sum of squares due to linear regression, is the amount of reduction in the sum of squares of the Y variable due to its linear regression on X. Subtracting this term from the sum of squares for the dependent variable Y, we get the *adjusted* sumof squares for Y. This is how the analysis of variance is adjusted to provide for the removal of effect, if any, of the covariate.

Assumptions: The assumptions made in a covariance analysis are:

(i) The populations are normally distributed with homogeneous variances.

(ii) The samples are random and independent.

(iii) The regression is linear and the slope is not zero.

(iv) The treatment effects and regression effects are additive.

(v) The covariate values are not affected by the treatments. If treatments affect the covariates, the covariance analysis is inappropriate.

(vi) The residuals ε_{ij} are normally and independently distributed with zero mean and homogeneous variances.

Q.22.8. Computation for the analysis of covariance.

	A			B				
	X	Y	XY	X	Y	XY		
	8(64)	17(289)	136	9(81)	19(361)	171		
	7(49)	15(225)	105	8(64)	20(400)	160		
	6(36)	15(225)	90	8(64)	17(289)	136		
	7(49)	18(324)	126	10(100)	20(400)	200	$T..$	$T..'$
Total	28	65	457	35	76	667	63	141
$(\text{Total})^2$	784	4225	--	1225	5776	--	2009	10001
Sum of squares	198	1063	--	309	1450	--	507	2513

The various sums of squares and products are then computed as below:

Source	Σx^2	Σy^2	Σxy
(i) Correction Factor	$\frac{T_{..}^2}{n} = \frac{(63)^2}{8} = 496.125$	$\frac{T_{..}'^2}{n} = \frac{(141)^2}{8} = 2485.125$	$\frac{T_{..}T_{..}'}{n} = \frac{(63)(141)}{8} = 1110.375$
(ii) Total SS	$\Sigma\Sigma X_{ij}^2 - C.F.$ $= 507 - 496.125$ $= 10.875\ (=S_{xx})$	$\Sigma\Sigma Y_{ij}^2 - C.F.$ $= 2513 - 2485.125$ $= 27.875\ (=S_{yy})$	$\Sigma\Sigma X_{ij}Y_{ij} - C.F.$ $= (457+667) - 1110.375$ $= 13.625\ (=S_{xy})$
(iii) Treatment SS	$\frac{\Sigma_j T_j^2}{m} - C.F.$ $= \frac{2009}{4} - 496.125$ $= 6.125\ (=T_{xx})$	$\frac{\Sigma_j T_j'^2}{m} - C.F.$ $= \frac{10001}{4} - 2485.125$ $= 15.125\ (=T_{yy})$	$\frac{\Sigma_j T_j T_j'}{m} - C.F.$ $= \frac{(28)(65)+(35)(76)}{4} - C.F.$ $= 9.625\ (=T_{xy})$
(iv) Error SS	E_{xx} = By subtraction	E_{yy} = By subtraction	E_{xy} = By subtraction

Hence the analysis of covariance table is:

S.V.	d.f	Sum of Squares and Products Σx^2	Σy^2	Σxy	Adjusted Results $SS(=\Sigma y^2 - \frac{(\Sigma xy)^2}{\Sigma x^2})$	d.f	MS	F
Between Treatments	1	6.125	15.125	9.625	--	--		
Within (Error)	6	4.750	12.750	4.000	9.382	5	$s_e^2 = 1.876$	
Total (T+E)	7	10.875	27.875	13.625	10.805	6		
Treatments adjusted					1.423	1	$s_t^2 = 1.423$	0.76

For Error, the regression coefficient, $b_{yx} = \frac{\Sigma xy}{\Sigma x^2} = \frac{4.000}{4.750} = 0.842$

To test the hypothesis $H_0 : \beta = 0$, we compute the value of F by the relation

$$F = \frac{E_{xy}^2 / E_{xx}}{E_{yy} - E_{xy}^2/E_{xx}} \times (n-k-1) = \frac{(\Sigma xy)^2 / \Sigma x^2}{\Sigma y^2 - (\Sigma xy)^2/\Sigma x^2} \times (n-k-1)$$

$$\therefore \quad F = \frac{(4.000)^2 / 4.75}{12.75 - (4)^2 / 4.75} \times 5 = \frac{(3.368)(5)}{9.382} = 1.79.$$

The computed value of F (in case of regression co-efficient) is less than $F_{.05(1,5)} = 6.61$. It is not significant. Hence there is no evidence of any linear regression between the variables X and Y.

Q.22.9. (i) We state our null hypotheses as

$H_0 : \mu_A = \mu_B = \mu_C = \mu_D$ before adjustment for quantity of food, and

$H_0' : \mu_A = \mu_B = \mu_C = \mu_D$ after adjustmnet for quantity of food.

(ii) We use a significance level of $\alpha = 0.05$.

(iii) The test-statistic is

$F = \frac{MS\ (\text{Effect})}{MS\ (\text{Error})}$, which under H_0, has an F-distribution provided the usual assumptions are satisfied.

(iv) Computations. Subtracting 100 from all values of X_{ij} and 80 from all values of Y_{ij}, we give the necessary calculations as follows:

	A			B			C			D		
	X	Y	XY	X	Y	XY	X	Y	XY	X	Y	XY
	-4(16)	18(324)	-72	9(81)	-16(256)	-144	79(6241)	-9(81)	-711	27(729)	-8(64)	-216
	8(64)	22(484)	176	25(625)	6(36)	150	32(1024)	4(16)	128	0(0)	-26(676)	0
	-6 (36)	22 (484)	-132	-15 (225)	-29 (841)	435	63 (3969)	-9 (81)	-567	51 (2601)	29 (841)	1479
	28 (784)	28 (784)	784	-18 (324)	-8 (64)	144	43 (1849)	-18 (324)	-774	16 (256)	13 (169)	208
Total	26	90	756	1	-47	585	217	-32	-1924	94	8	1471
$(\text{Total})^2$	676	8100	--	1	2209	--	47089	1024	--	8836	64	--
Sum of squares	900	2076	--	1255	1197	--	13083	502	--	3586	1750	--

Now $T_{..} = 338$, $\sum_j T_{.j}^2 = 56602$, $\sum_i \sum_j X_{ij}^2 = 18824$.

$T'_{..} = 19$, $\sum_j T'^2_j = 11397$, $\sum_i \sum_j Y_{ij}^2 = 5525$ and $\sum_i \sum_j X_{ij} Y_{ij} = 888$.

Next, we compute the sums of squares and products as below:

Source	Σx^2	Σy^2	Σxy
(i) Correction Factor	$\frac{T_{..}^2}{n} = \frac{(338)^2}{16} = 7140.25$	$\frac{T'^2_{..}}{n} = \frac{(19)^2}{16} = 22.56$	$\frac{T_{..} T'_{..}}{n} = \frac{(338)(19)}{16} = 401.38$
(ii) Total SS	$S_{xx} = \Sigma\Sigma X_{ij}^2 - C.F. = 18824 - 7140.25 = 11683.75$	$S_{yy} = \Sigma\Sigma Y_{ij}^2 - C.F. = 5525 - 22.56 = 5502.44$	$S_{xy} = \Sigma\Sigma X_{ij} Y_{ij} - C.F. = 888 - 401.38 = 486.62$
(iii) Treatment SS	$T_{xx} = \sum_j \frac{T_j^2}{m} - C.F. = \frac{56602}{4} - 7140.25 = 7010.25$	$T_{yy} = \sum_j \frac{T'^2_j}{m} - C.F. = \frac{11397}{4} - 22.56 = 2826.69$	$T_{xy} = \sum_i \frac{T_j T'_j}{m} - C.F. = \frac{(26)(90)+..+(94)(8)}{4} - C.F. = \frac{-3899}{4} - 401.38 = -1376.13$
(iv) Error SS	$E_{xx} = S_{xx} - T_{xx} = 4673.50$	$E_{yy} = S_{yy} - T_{yy} = 2675.75$	$E_{xy} = S_{xy} - T_{xy} = 1862.75$

These results are then put in the analysis of covriance table, carrying out other necessary computations.

Source of Variation	d.f.	Sum of squares and products			Regression Co-efficient
		Σx^2	Σy^2	Σxy	
Treatments	3	7010.25	2826.69	-1376.13	
Error (E)	12	4673.50	2675.75	1862.75	$b = 0.3986$
Total	15	11683.75	5502.44	486.62	$b' = 0.0416$

Source of Variation	Regression		Adjusted		
	d.f.	SS	d.f	SS	MS
Total (T+E)	1	20.27	14	5502.44-20.27 = 5482.17	---
Error (E)	1	742.45	11	2675.75-742.45 = 1933.30	$s_e^2 = 175.75$
Treatment adjusted			3	5481.17-1933.30 = 3548.87	$s_t^2 = 1182.96$

(a) To test the hypothesis H_0; we compute the value of F-ratio by using the unadjusted y-values as

$$F_1 = \frac{2826.69 / 3}{2675.75 / 12} = \frac{924.23}{222.98} = 4.14.$$

and $\upsilon_1 = 3$, $\upsilon_2 = 12$.

(v) The critical region is $F_1 > F_{.05,(3,12)} = 3.49$.

(vi) As the computed value of F_1 falls in the critical region, so we reject H_0 and conclude that the four rations A, B, C and D produced significantly different gains among the rats.

(b) To test the hypothesis H_0', we compute the F-ratio by using the adjusted y-values as

$$F_2 = \frac{1182.96}{175.75} = 6.73, \text{ and } \upsilon_1 = 3, \upsilon_2 = 11.$$

(v) The critical region is $F_2 > F_{.05,(3,11)} = 3.59$.

(vi) Since the computed value of F_2 falls in the critical region, so we reject H_0 and conclude that the rations produced significantly different gains adjusted for quantity of food.

Q.22.10. Computation for an analysis of covariance of X on T.

	1			2			3			4		
	X	T	XT	X	T	XT	X	T	XT	X	T	XT
	3.2	4.0	12.80	2.4	4.7	11.28	2.6	3.7	9.62	2.0	3.4	6.80
	3.3	3.6	11.88	3.0	6.4	19.20	3.5	6.8	23.80	3.8	9.4	35.72
	2.3	1.3	2.99	2.2	3.8	8.36	4.1	8.8	36.08	2.8	6.3	17.64
	4.4	9.2	40.48	2.5	4.9	12.25	2.5	4.5	11.25	2.3	4.7	10.81
	–	–	–	3.6	8.7	31.32	–	–	–	2.5	5.1	12.75
	–	–	–	1.2	2.0	2.40	–	–	–	–	–	–
Total	13.2	18.1	68.15	14.9	30.5	84.81	12.7	23.8	80.75	13.4	28.9	83.72
$(\text{Total})^2$	174.24	327.61	–	222.01	930.25	–	161.29	566.44	–	179.56	835.21	–

$\sum_i \sum_j X_{ij} = T_{..} = 54.2$, $\sum_i \sum_j X_{ij}^2 = (3.2)^2 + (3.3)^2 + \ldots + (2.5)^2 = 165.92$,

$\sum_i \sum_j T_{ij} = T'_{..} = 101.3$, $\sum_i \sum_j T_{ij}^2 = (4.0)^2 + (3.6)^2 + \ldots + (5.1)^2 = 641.81$,

$\sum_i \sum_j X_{ij} T_{ij} = (68.15) + (84.81) + (80.75) + (83.72) = 317.43$,

The "Total SS and Products" are computed as

$$S_{xx} = \sum_i \sum_j X_{ij}^2 - \frac{T_{..}^2}{n} = 165.92 - \frac{(54.2)^2}{19}$$

$$= 165.92 - 154.61 = 11.31,$$

$$S_{tt} = \sum_i \sum_j T_{ij}^2 - \frac{(T'_{..})^2}{n} = 641.81 - \frac{(101.3)^2}{19}$$

$$= 641.81 - 540.09 = 101.72 \text{ and}$$

$$S_{xt} = \sum_i \sum_j X_{ij} T_{ij} - \frac{(T_{..})\,(T'_{..})}{n} = 317.43 - \frac{(54.2)\,(101.3)}{19}$$

$$= 317.43 - 288.97 = 28.46.$$

"Treatment SS and Products" are computed as

$$T_{xx} = \sum_j \frac{T_{.j}^2}{m_j} - \frac{T_{..}^2}{n}$$

$$= \frac{174.24}{4} + \frac{222.01}{6} + \frac{161.29}{4} + \frac{179.56}{5} - \frac{(54.2)^2}{19}$$

$$= 156.79 - 154.61 = 2.18,$$

$$T_{tt} = \sum_j \frac{T'^2_j}{m_j} - \frac{(T'_{..})^2}{n}$$

$$= \frac{327.61}{4} + \frac{930.25}{6} + \frac{566.44}{4} + \frac{835.21}{5} - \frac{(101.3)^2}{19}$$

$$= 645.59 - 540.09 = 5.50, \text{ and}$$

$$T_{xt} = \sum_j \frac{(T_{.j})\,(T'_j)}{m_j} - \frac{(T_{..})\,(T'_{..})}{n}$$

$$= \frac{(13.2)\,(18.1)}{4} + \ldots + \frac{(13.4)\,(28.9)}{5} - \frac{(54.2)\,(101.3)}{19}$$

$$= 59.73 + 75.74 + 75.56 + 77.45 - 288.97 = -0.49.$$

"Error SS and Products" are obtained by subtraction as

$E_{xx} = S_{xx} - T_{xx} = 11.31 - 2.18 = 9.13$,

$E_{tt} = S_{tt} - T_{tt} = 101.72 - 5.50 = 96.22$, and

$E_{xt} = S_{xt} - T_{xt} = 28.46 - (-0.49) = 28.95$.

Hence the analysis of covariance table is:

S.V.	d.f	Sum of Squares and Products			Adjusted for T			
		Σx^2	Σt^2	Σxt	$SS(=\Sigma x^2 - \frac{(\Sigma xt)^2}{\Sigma x^2})$	d.f	MS	F
Treatments	3	2.18	5.50	−0.49	--	--	--	--
Error	15	9.13	96.22	28.95	9.13−8.71 = 0.42	14	0.03	--
Total (T+E)	18	11.31	101.72	28.46	11.31−7.96 = 3.35	17	--	--
Treatments adjusted					3.35−0.42 = 2.93	3	0.98	32.67

For error, the regression co-efficient, $b_{xt} = \frac{\Sigma xt}{\Sigma t^2} = \frac{28.95}{96.22} = 0.3009$.

To test the hypothesis $H_0 : B = 0$, we compute the statistic F by the relation

$$F = \frac{E_{xt}^2 / E_{tt}}{E_{xx} - E_{xt}^2 / E_{tt}} \times (n-k-1).$$

Thus $$F = \frac{(28.95)^2 / 96.22}{9.13 - (28.95)^2 / 96.22} \times (19-4-1).$$

$$= \frac{(8.71)\,(14)}{0.42} = 290.33, \text{ and } \nu_1 = 1,\ \nu_2 = 14.$$

The 5-percent F for 1,14 *d.f.* from tables = 4.60. Hence this result is highly significant and we certainly reject the hypothesis $H_0 : B = 0$. Hence adjustment on the mean values of X becomes necessary.

To compare mean values of X, we compute the adjusted means, *i.e.* means which are free from the influence of regression, by the relation

$$\text{corrected } \bar{X}_{i\cdot} = \bar{X}_{i\cdot} - b_{xt}\,(\bar{T}_{i\cdot} - \bar{T}_{\cdot\cdot})$$

Thus we get the adjusted means as follows:

Group	$\bar{T}_{i\cdot}$	$\bar{T}_{i\cdot}-\bar{T}_{\cdot\cdot}$	$b(\bar{T}_{i\cdot}-\bar{T}_{\cdot\cdot})$	$\bar{X}_{i\cdot}$	Corrected means $\bar{X}_{i\cdot}-b\,(\bar{T}_{i\cdot}-\bar{T}_{\cdot\cdot})$
1	4.52	−0.81	−0.24	3.30	3.54
2	5.08	−0.25	−0.08	2.48	2.56
3	5.95	+0.62	+0.19	3.18	2.99
4	5.78	+0.45	+0.14	2.68	2.54

For comparison, we calculate the standard error for the difference between two adjusted treatment means by the formula

$$s = s_e \sqrt{\frac{1}{m_1} + \frac{1}{m_2} + \frac{(\bar{T}_{1\cdot}-\bar{T}_{2\cdot})^2}{E_{tt}}}$$

Thus to compare treatment means $\bar{X}_{1\cdot}$ and $\bar{X}_{2\cdot}$, we require

$$s = \sqrt{0.03\,[\frac{1}{4} + \frac{1}{6} + \frac{(4.52-5.08)^2}{96.22}}$$

$$= \sqrt{0.0126} = 0.11$$

Now, $\bar{X}_{1\cdot}' - \bar{X}_{2\cdot}' = 3.54 - 2.56 = 0.98$, which is obviously significant.

Similarly other adjusted means are compared. A separate calculation for the standard error is required for each comparison.

Q.22.11. Computations for an analysis of covariance of Y on X.

Group	I			II			III				
	X	Y	XY	X	Y	XY	X	Y	XY		
	14	10	140	11	5	55	7	5	35		
	9	6	54	9	2	18	6	4	24		
	11	8	88	8	6	48	2	1	2		
	12	6	72	10	5	50	10	7	70		
	10	9	90	10	4	40	7	9	63	$T_{\cdot\cdot}$	$T_{\cdot\cdot}'$
Total	56	39	444	48	22	211	32	26	194	36	86
$(\text{Total})^2$	3136	1521	--	2304	484	--	1024	676	--	6464	2681

The "*Total SS and Products*" are calculated as below:

$$S_{xx} = \sum_i \sum_j X_{ij}^2 - \frac{T_{\cdot\cdot}^2}{n} = (14)^2 + (9)^2 + \ldots + (7)^2 - \frac{(136)^2}{15}$$

$$= 1346 - 1233.07 = 112.93,$$

$$S_{yy} = \sum_i \sum_j Y_{ij}^2 - \frac{(T_{\cdot\cdot}')^2}{n} = (10)^2 + (6)^2 + \ldots + (9)^2 - \frac{(87)^2}{15}$$

$$= 595 - 504.6 = 90.4, \text{ and}$$

$$S_{xy} = \sum_i \sum_j X_{ij} Y_{ij} - \frac{(T_{\cdot\cdot})\,(T_{\cdot\cdot}')}{n} = 444 + 211 + 194 - \frac{(136)(87)}{15}$$

$$= 849 - 788.8 = 60.2.$$

The "*Between Groups (Treatments) SS and Products*" are computed as

$$T_{xx} = \sum_j \frac{T_j^2}{m} - \frac{(T..)^2}{n} = \frac{6464}{5} - \frac{(136)^2}{15}$$

$$= 1292.8 - 1233.07 = 59.73.$$

$$T_{yy} = \sum_j \frac{(T'_{j.})^2}{m} - \frac{(T'_{..})^2}{n} = \frac{2681}{5} - \frac{(87)^2}{15}$$

$$= 536.2 - 504.6 = 31.6, \text{ and}$$

$$T_{xy} = \sum_j \frac{(T_{.j})\,(T'_j)}{m} - \frac{(T..)\,(T'_{..})}{n}$$

$$= \frac{(56\times39)+(48\times22)+(32\times26)}{5} - \frac{(136)\,(87)}{15}$$

$$= \frac{4072}{5} - 788.88 = 25.6$$

The *"Within Groups or Error SS and Products"* are obtained by subtraction.

$\therefore$ $E_{xx} = S_{xx} - T_{xx} = 112.93 - 59.73 = 53.20,$

$E_{yy} = S_{yy} - T_{yy} = 90.4 - 31.6 = 58.8,$ and

$E_{xy} = S_{xy} - T_{xy} = 60.2 - 25.6 = 34.6$

Adjustment on Total *SS* for *Y*-values.

$$\text{Adusted } \Sigma y^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = 90.4 - \frac{(60.2)^2}{112.93}$$

$$= 90.4 - 32.09 = 58.31.$$

Adjustment on Error *SS* for *Y*-values.

$$\text{Adusted } \Sigma y^2 = E_{yy} - \frac{E_{xy}^2}{E_{xx}} = 58.8 - \frac{(34.6)^2}{53.2}$$

$$= 58.8 - 22.50 = 36.30.$$

Hence the Analysis of Covariance table is

S.V.	d.f	Sum of Squares and Products			Adjusted			
		Σx^2	Σy^2	Σxy	Σy^2	d.f	MS	F
Between Groups	2	59.73	31.6	25.6	--	--	--	
Within Groups	12	53.20	58.8	34.6	36.30	11	3.30	
Total	14	112.93	90.4	60.2	58.31	13	--	--
Treatments adjusted					22.01	2	11.005	3.33

To test the hypothesis H_0 : there is no difference between the subsequent scores, we calculate the value of F by using adjusted *y*-values. Thus

$$F = \frac{11,005}{3.30} = 3.33, \text{ and } v_1 = 2,\ v_2 = 11.$$

The 5-per cent F for 2, 11 *d.f.* from tables = 3.98.

Since the computed value of F does not fall in the critical region so we do not reject H_0. Hence the difference between the subsequent scores is not significant.

Q.22.12. (b) For Error, the regression co-efficient is

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2} = \frac{-650}{3317} = -0.196.$$

Adjustment on Total *SS* for *y*-values.

$$\text{Adjusted } \Sigma y^2 = \Sigma y^2 - \frac{(\Sigma xy)^2}{\Sigma x^2} = (112+516) - \frac{(-1182)^2}{8001}$$

$$= 628 - 174.62 = 453.38.$$

Adjustment on Error *SS* for *y*-values.

$$\text{Adjusted } \Sigma y^2 = 516 - \frac{(-650)^2}{3317} = 516 - 127.37 = 388.63.$$

Hence the analysis of covariance table is

S.V.	d.f	Sum of Squares and Products			Adjusted		
		Σx^2	Σy^2	Σxy	Σy^2	d.f	MS
Tratment	9	4684	112	-532	--	--	--
Error	27	3317	516	-650	388.63	26	14.95
Total (T+E)	36	8001	628	-1182	453.38	35	--
Treatments adjusted					64.75	9	7.19

To test the hypthesis $H_0 : B = 0$, the F-ratio would be

$$F = \frac{(\Sigma xy)^2 / \Sigma x^2}{\Sigma y^2 - (\Sigma xy)^2 / \Sigma x^2} \times (n-k-1),$$

$$= \frac{(-650)^2 / 3317}{516 - (-650)^2 / 3317} \times 26$$

$$= \frac{(127.37)\ (26)}{388.63} = 8.52$$

The 5-per cent F for 1, 26 *d.f.* from table = 4.22, which is less than the calculated value. Hence we reject the hypothesis $H_0 : B = 0$.

To test the hypothesis H_0 : there is no significant difference between treatment means, we compute the F-ratio by using the adjusted y-values. Thus

$$F = \frac{7.19}{14.95} = 0.48.$$

The 5-per cent F for 9, 26 *d.f.* from tables = 2.28.

Comparing the values of F, we find that it is not significant. Hence we cannot reject the hypothesis of no difference between treatments.

Q.22.14. To set up an analysis of covariance table, the following calculations are made.

Blocks	1			2			3			4				
	X	Y	XY	X	Y	XY	X	Y	XY	X	Y	XY	B_i	B_i'
1	28	21	588	22	17	374	27	18	486	34	25	850	111	81
2	24	21	504	22	16	352	26	20	520	21	15	315	93	72
3	20	16	320	18	16	288	19	14	266	21	16	336	78	62
4	25	21	525	23	17	391	31	19	589	27	21	567	106	78
T_j, T_j'	97	79	1937	85	66	1405	103	71	1861	103	77	2068	388	293
$T_j^2, T_j'^2$	9409	6241	--	7225	4356	--	10609	5041	--	10609	5929	--	37852	21567

(i) Total SS and Products.

$$S_{xx} = \sum_i \sum_j X_{ij}^2 - \frac{(T..)^2}{n} = (28)^2 + (24)^2 + \ldots + (27)^2 - \frac{(388)^2}{16}$$

$$= 9700 - 9409 = 291,$$

$$S_{yy} = \sum_i \sum_j Y_{ij}^2 - \frac{(T_{..}')^2}{n} = (21)^2 + (21)^2 + \ldots + (21)^2 - \frac{(293)^2}{16}$$

$$= 5497 - 5365.56 = 131.44, \text{ and}$$

$$S_{xy} = \sum_i \sum_j X_{ij} Y_{ij} - \frac{(T..)\ (T_{..}')}{n} = 1937 + \ldots + 2068 - \frac{(388)(293)}{16}$$

$$= 7271 - 7105.25 = 165.75$$

(ii) Treatment SS and Products

$$T_{xx} = \sum_j \frac{T_j^2}{r} - \frac{(T..)^2}{n} = \frac{37852}{4} - 9409$$

$$= 9463 - 9409 = 54$$

$$T_{yy} = \sum_j \frac{T_j'^2}{r} - \frac{(T'..)^2}{n} = \frac{21567}{4} - \frac{(293)^2}{16}$$

$$= 5391.75 - 5365.56 = 26.19, \text{ and}$$

$$T_{xy} = \sum_j \frac{(T_{.j})\,(T'_{.j})}{r} - \frac{(T..)\,(T'..)}{n}$$

$$= \frac{(97\times79) + \ldots + (103\times77)}{4} - C.F.$$

$$= \frac{28517}{4} - \frac{(388)\,(293)}{16} = 7129.25 - 7105.25 = 24.00$$

(iii) Block SS and Products

$$B_{xx} = \sum_i \frac{B_{i.}^2}{c} - \frac{(T..)^2}{n} = \frac{(111)^2+\ldots+(106)^2}{4} - \frac{(388)^2}{16}$$

$$= \frac{38290}{4} - 9409 = 9572.5 - 9409 = 163.5$$

$$B_{yy} = \sum_i \frac{B'^2_{i.}}{c} - \frac{(T'..)^2}{n} = \frac{(81)^2 + \ldots + (78)^2}{4} - \frac{(293)^2}{16}$$

$$= \frac{21673}{4} - \frac{(293)^2}{16}$$

$$= 5418.25 - 5365.56 = 52.69, \text{ and}$$

$$B_{xy} = \sum_i \frac{(B_{i.})\,(B'_{i.})}{c} - \frac{(T..)\,(T'..)}{n}$$

$$= \frac{(111\times81) + \ldots + (106\times78)}{4} - \frac{(388)\,(293)}{16}$$

$$= \frac{28791}{4} - 7105.25 = 7197.75 - 7105.25 = 92.50$$

(iv) *Error SS and Products* are obtained by subtracton.

$E_{xx} = S_{xx} - T_{xx} - B_{xx} = 291 - 54 - 163.5 = 73.5,$

$E_{yy} = S_{yy} - T_{yy} - B_{yy} = 131.44 - 26.19 - 52.69 = 52.56,$ and

$E_{xy} = S_{xy} - T_{xy} - B_{xy} = 165.75 - 24.00 - 92.50 = 49.25.$

(v) Adjusted *SS* for Error.

$$\text{Adjusted } \Sigma y^2 = \Sigma y^2 - \frac{(\Sigma xy)^2}{\Sigma x^2} = 52.56 - \frac{(49.25)^2}{73.5}$$

$$= 52.56 - 33.00 = 19.56.$$

Adjusted SS for Treatment plus Error

$$\text{Adjusted } \Sigma y^2 = \Sigma y^2 - \frac{(\Sigma xy)^2}{\Sigma x^2} = 78.75 - \frac{(73.25)^2}{127.5}$$

$$= 78.75 - 42.08 = 36.67.$$

Hence the Analysis of Covariance for these results is:

S.V.	d.f	Sum of Squares and Products			Adjusted for Covariate x		
		Σx^2	Σy^2	Σxy	Σy^2	d.f	MS
Total	15	291	131.44	165.75	--	--	--
Blocks	3	163.5	52.69	92.50	--	--	--
Treatments	3	54.0	26.19	24.00	--	--	--
Error	9	73.5	52.56	49.25	19.56	8	2.445
T+E	12	127.5	78.75	73.25	36.37	11	--
Treatments adjusted					17.11	3	5.703

The first problem is to find out if there is evidence of any regression between the variables X and Y. This implies the testing of the hypothesis $H_0 : B = 0$, for which we compute the value of F-ratio as

$$F = \frac{(\Sigma xy)^2 / \Sigma x^2}{\Sigma y^2 - (\Sigma xy)^2 / \Sigma x^2} \times (n-k-1)$$

$$= \frac{(49.25)^2 / 73.5}{52.56 - (49.25)^2 / 73.5} \times 8 = \frac{(33.00)\,(8)}{19.56} = 13.50$$

The 5-percent value of F for 1,8 *d.f.* from tables = 5.32. The computed value of F is greater than critical value. We therefore, reject H_0. There is sufficient evidence of regression between X and Y. Hence adjustment on the mean values of Y becomes necessary.

The next problem is to test the hypothesis of no difference in adjusted Y-values. For this, we compute the F-ratio as

$$F = \frac{\text{Adjusted Treatment } MS}{\text{Adjusted Error } MS} = \frac{5.703}{2.445}$$

$$= 2.33, \text{ with } v_1 = 3, v_2 = 8.$$

The computed value of F is less than the corresponding critical value of $F_{.05(3,8)} = 4.07$. Hence we do not reject the hypothesis of no difference among the treatment means for Y after ajusting for the covariate X.

Q.22.15. (a) For Error, the regression co-efficient,

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2} = \frac{682.20}{28665.1} = 0.0238.$$

To test the hypothesis $H_0 : B = 0$, the F-ratio would be

$$F = \frac{(\Sigma xy)^2 / \Sigma x^2}{\Sigma y^2 - (\Sigma xy)^2 / \Sigma x^2} \times (n-k-1) \text{ with } v_1 = 1, v_2 = n-k-1.$$

$$\therefore \quad F = \frac{(682.20)^2 / 28665.1}{23.23 - (682.20)^2 / 28665.1} \times 29$$

$$= \frac{(16.2357)\,(29)}{6.99} = \frac{470.84}{6.99} = 67.36$$

The 5-percent F for 1,29 *d.f.* is 4.18. The computed value of F is far greater than table value. The hypothesis $H_0 : B = 0$ is certainly rejected.

(b) To test the hypothesis H_0: there are no differences among the treatment means for Y adjusted for variation attributed to X, we use the F-ratio given by the relation

$$F = \frac{\text{Adjusted Treatment } MS}{\text{Adjusted Error } MS}, \text{ with } v_1 = k-1, v_2 = n-k-1$$

Adjusted SS for Error is

$$\text{Adjusted } \Sigma y^2 = \Sigma y^2 - \frac{(\Sigma xy)^2}{\Sigma x^2}$$

$$= 23.23 - \frac{(682.20)^2}{28665.1}$$

$$= 23.23 - 16.24 = 6.99, \text{ and}$$

Adjusted SS for Treatment plus Error is

$$\text{Adjusted } \Sigma y^2 = (T_{yy} + E_{yy}) - \frac{(T_{xy} + E_{xy})^2}{T_{xx} + E_{xx}}$$

$$= (112.86 + 23.23) - \frac{(3598.05 + 682.20)^2}{116020.3 + 28665.1}$$

$$= 136.09 - 126.62 = 9.47.$$

The analysis of covariance table then becomes

S.V.	d.f	Sum of Squares and Products			Adjusted		
		Σx^2	Σy^2	Σxy	Σy^2	d.f	MS
T + E	36	144685.4	136.09	4280.25	9.47	35	--
E	30	28665.1	23.23	682.20	6.99	29	0.241
Treatments adjusted					2.48	6	0.413

Thus $F = \frac{0.413}{0.241} = 1.71$ with $v_1 = 6$, $v_2 = 29$ *d.f.*

This value is not significant at the 5-percent level as it is less than $F_{.05,(6,29)} = 2.43$. Hence we do not reject the hypothesis of no differences among the treatment means for Y after adjusting for the variation attributed to X.

CHAPTER 23

EXPERIMENTAL DESIGNS

Q.23.4. Computations for the standard error of a feeding treatment. To simplify the arithmetic, we choose our origin at Y=100. The computations are then given below:

	Level of Feeding			Total
	Subnormal	Normal	Supernormal	
	18 (324)	42 (1764)	62 (3844)	
	22 (484)	29 (841)	73 (5329)	
	21 (441)	34 (1156)	68 (4624)	
	26 (676)	32 (1024)	83 (6889)	
	9 (81)	35 (1225)	72 (5184)	
$T_{.j}$	96	172	358	626
T_j^2	9216	29584	128164	166964
$\sum_i Y_{ij}^2$	2006	6010	25870	33886

$$\text{Total } SS = \sum_i \sum_j Y_{ij}^2 - \frac{T_{..}^2}{n}$$

$$= 33886 - \frac{(626)^2}{15} = 33886 - 26125.07 = 7760.93$$

$$\text{Between Levels } SS = \sum_j \frac{T_j^2}{r} - \frac{T_{..}^2}{n}$$

$$= \frac{166964}{5} - \frac{(626)^2}{15} = 33392.8 - 26125.07$$

$$= 7267.73, \text{ and}$$

$$\text{Within Levels } SS = \text{Total } SS - \text{Between Levels } SS$$

$$= 7760.93 - 7267.73 = 493.20$$

The analysis of variance table is:

Source of Variation	$d.f$	Sum of Squares	Mean Square
Between Levels	2	7267.73	3633.865
Within Levels	12	493.20	$s_e^2 = 41.100$
Total	14	7760.93	--

Now the standard error of a feeding treatment is

$$S.E. = \sqrt{\frac{\text{error mean square}}{\text{number of observations in treatment mean}}}$$

$$= \sqrt{\frac{41.1}{5}} = \sqrt{8.22} = 2.87.$$

Q.23.5. (i) We state our null hypothesis as

H_0 : There is no difference in the yielding capabilities of the varieties of wheat.

(ii) We use a significance level of $\alpha = 0.05$.

(iii) The test-statistic is $F = \frac{MS \text{ (Varieties)}}{MS \text{ (Error)}}$, which under H_0 has an F-distribution provided the assumptions of normality, homogeneity and independence are satisfied.

(iv) Computations.

	Varieties							Total
	1	2	3	4	5	6	7	
	13	15	14	14	17	15	16	
	11	11	10	10	15	9	12	
	10	13	12	15	14	13	13	
	16	18	13	17	19	14	15	
	12	12	11	10	12	10	11	
$T_{.j}$	62	69	60	66	77	61	67	462
T_j^2	3844	4761	3600	4356	5929	3721	4489	30700
$\sum_i Y_{ij}^2$	790	983	730	910	1215	771	915	6314

$$\text{Total } SS = \sum_i \sum_j Y_{ij}^2 - \frac{T_{..}^2}{n}$$

$$= 6314 - \frac{(462)^2}{35} = 6314 - 6098.4 = 215.6$$

$$\text{Between Varieties } SS = \sum_j \frac{T_j^2}{r} - \frac{T_{..}^2}{n}$$

$$= \frac{30700}{5} - \frac{(462)^2}{35} = 6140 - 6098.4$$

$$= 41.6$$

Error or Within Varieties SS = Total SS − Between Varieties SS

$$= 215.6 - 41.6 = 174.0.$$

The analysis of variance table is:

Source of Variation	$d.f$	Sum of Squares	Mean Square	F
Between Varieties	6	41.6	6.933	1.12
Within Varieties	28	174.0	6.214	--
Total	34	215.6	--	

(v) The critical region is $F > F_{.05;(6,28)} = 2.45$.

(vi) Since the computed value of F does not fall in the critical region, so we do not reject H_0. Hence we may conclude that the experiment as a whole does not indicate significant variation in the yielding capabilities of the varieties of wheat.

Q.23.6. (i) The null hypothesis would be stated as

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ or in words, that there is no difference in the effects of storage conditions on the moisture content of white pine lumber.

(ii) We use a significance level of $\alpha = 0.05$.

(iii) The test-statistic is $F = \dfrac{MS \text{ for conditions}}{MS \text{ for error}}$, which under H_0 has an F-distribution with $v_1 = 4$ and $v_2 = 9$ $d.f.$

(iv) Computations.

	Storage Conditions					Total
	1	2	3	4	5	
	7.3(53.29) 8.3(68.89) 7.6(57.76) 8.4(70.56) 8.3(68.89)	5.4(29.16) 7.4(54.76) 7.1(50.41) -- --	8.1(65.61) 6.4(40.96) -- -- --	7.9(62.41) 9.5(90.25) 10.5(110.25) -- --	7.1(50.41) -- -- -- --	
$T_{\cdot j}$	39.9	19.9	14.5	27.9	7.1	109.3
T_j^2	1592.01	396.01	210.25	778.41	50.41	--
r_j	5	3	2	3	1	14
$\sum_i Y_{ij}^2$	319.39	134.73	106.57	262.91	50.41	873.61

$$\text{Total } SS = \sum_i \sum_j Y_{ij}^2 - \frac{T_{..}^2}{n}$$

$$= 873.61 - \frac{(109.3)^2}{14} = 873.61 - 853.32 = 20.29$$

$$\text{Between Condition } SS = \sum_j \frac{T_j^2}{r} - \frac{T_{..}^2}{n}$$

$$= \frac{1592.01}{5} + \frac{396.01}{3} + \ldots + \frac{50.41}{1} - \frac{(109.3)^2}{14}$$

$$= 865.40 - 853.32 = 12.08$$

Within Condition SS = Total SS − Between Conditions SS

$$= 20.29 - 12.08 = 8.21$$

The analysis of variance table is:

Source of Variation	$d.f$	Sum of Squares	Mean Square	F
Between conditions	4	12.08	3.02	3.32
Within conditions	9	8.21	0.91	--
Total	13	20.29	--	

(v) The critical region is $F > F_{.05;(4,9)} = 3.63$

(vi) Since the computed value of F does not fall in the critical region, so we do not reject H_0.

Q.23.7. To perform the analysis of variance, the following calculations are carried out:

$$\bar{Y}_{..} = \frac{n_1\bar{Y}_1 + n_2\bar{Y}_2 + n_3\bar{Y}_3}{n_1 + n_2 + n_3}$$

$$= \frac{(10)(90) + (10)(86) + (11)(83)}{10 + 10 + 11} = \frac{2673}{31} = 86.23;$$

$$\text{Within } SS = (n_1-1)s_1^2 + (n_2-1)s_2^2 + (n_3-1)s_3^2$$

$$= (9)(4.37) + (9)(3.76) + (10)(4.21) = 115.27; \text{ and}$$

$$\text{Between } SS = \sum n_j (\bar{Y}_{.j} - \bar{Y}_{..})^2$$

$$= [10(90-86.23)^2 + 10(86-86.23)^2 + 11(83-86.23)^2]$$

$$= 142.129 + 0.529 + 114.762 = 257.42.$$

Thus the analysis of variance table becomes

Source of Variation	$d.f$	Sum of Squares	Mean Square	F
Between Formats	2	257.42	128.71	31.24
Within Formats	28	115.27	4.12	--
Total	30	372.69	--	

Significance level for the hypothesis that mean reading comprehension is the same for all formats is

$$P(F_{(2,28)} \geq 31.24) \simeq 0.001.$$

Q.23.8. (b) (i) Completing the ANOVA-Table, we get

Source of Variation	$d.f$	Sum of Squares	Mean Square	F-ratio
Treatments	3	28.2	**9.4**	**4.14**
Blocks	5	69.0	13.80	--
Error	15	34.1	**2.27**	--

(ii) $F_{0.01}(3,15) = 5.42$. Since the computed value of $F = 4.14$ is not greater than Table value (5.42), we therefore conclude that the data do not provide sufficient evidence to indicate a difference among the treatment means at $\alpha = 0.01$.

However, the treatment means differ significantly at $\alpha = 0.05$.

(iii) The 90% confidence interval for the difference between two population means $\mu_A - \mu_B$, is

$$(\bar{y}_A - \bar{y}_B) \pm t_{\alpha/2(v)} \sqrt{\frac{2(MSE)}{r}}.$$

Substituting the values, we get

$$(9.7-12.1) \pm 1.753 \sqrt{\frac{2(2.27)}{6}} \quad (\because t_{0.05,(15)} = 1.753)$$

i.e. $-2.4 \pm (1.753)(0.87)$

i.e. -2.4 ± 1.53 or $(-3.93, -0.87)$

Q.23.9. (b) (i) We state the null hypotheses as

$H_0 : \mu_{.1} = \mu_{.2} = \mu_{.3}$, *i.e.* the average yield of the three varieties is the same,

$H_0' : \mu_{1.} = \mu_{2.} = \mu_{3.} = \mu_{4.}$, *i.e.* there is no difference among the replications.

(ii) We use a significance level of $\alpha = 0.05$.

(iii) The test-statistic is $F = \frac{MS\text{ (Effect)}}{MS\text{ (Error)}}$, which under H_0 has an F-distribution with appropriate degrees of freedom.

(iv) Computations.

Rep-licates	Varieties			$B_{i.}$	$B_{i.}^2$
	A	B	C		
1	32.1 (1030.41)	31.7 (1004.89)	34.2 (1169.64)	98.0	9604.00
2	17.0 (289.00)	32.7 (1069.29)	30.7 (942.49)	80.4	6464.16
3	40.8 (1664.64)	25.3 (640.09)	48.2 (2323.24)	114.3	13064.49
4	26.8 (718.24)	47.9 (2294.41)	59.6 (3552.16)	134.3	18036.49
$T_{.j}$	116.7	137.6	172.7	427.0	47169.14
$T_{.j}^2$	13618.89	18933.76	29825.29	62377.94	--
$\sum_i Y_{ij}^2$	3702.29	5008.68	7987.53	16698.50	--

$$\text{Total } SS = \sum_i \sum_j Y_{ij}^2 - \frac{T_{..}^2}{n} = 16698.50 - \frac{(427.0)^2}{12}$$

$$= 16698.50 - 15194.08 = 1504.42,$$

$$\text{Varieties } SS = \sum_j \frac{T_j^2}{r} - \frac{T_{..}^2}{n} = \frac{62377.94}{4} - \frac{(427.0)^2}{12}$$

$$= 15594.48 - 15194.08 = 400.40,$$

$$\text{Replicates } SS = \sum_i \frac{B_{i.}^2}{c} - \frac{T_{..}^2}{n} = \frac{47169.14}{3} - \frac{(427.0)^2}{12}$$

$$= 15723.05 - 15194.08 = 528.97$$

Error SS = Total SS − Varieties SS − Replicates SS

$$= 1504.42 - 400.40 - 528.97 = 575.05$$

These results are displayed in the following analysis of variance table:

Source of Variation	$d.f$	SS	MS	F
Between Varieties	2	400.40	200.20	$F_1 = 2.09$
Between Replicates	3	528.97	176.32	$F_2 = 1.87$
Error	6	575.05	95.84	--
Total	11	1504.42	--	

(v) The critical regions are $F_1 > F_{0.05;(2,6)} = 5.14$ and

$F_2 > F_{0.05;(3,6)} = 4.76$

(vi) The computed values of F do not fall in the critical region. We do not reject the null hypotheses.

Q.23.10. (a) (i) We state our hypotheses as

H_0' : $\mu_A = \mu_B = \mu_C = \mu_D$,

H_0'' : $\mu_{1.} = \mu_{2.} = \mu_{3.} = \mu_{4.}$, and

H_1' : At least two varietal means are different,

H_1'' : At least two block means are different.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic to use is

$$F = \frac{\text{Effect Mean Square}}{\text{Error Mean Square}}$$

(iv) Computations.

Blocks	Varieties of wheat				$T_{i.}$	$T_{i.}^2$
	A	B	C	D		
I	17	27	15	25	84	7056
II	28	26	16	22	92	8486
III	22	25	11	14	72	5184
IV	17	18	18	19	72	5184
$T_{.j}$	84	96	60	80	320	25888
T_j^2	7056	9216	3600	6400	26272	

$$\text{Total } SS = \sum_i \sum_j Y_{ij}^2 - \frac{T_{..}^2}{cr}$$

$$= (17)^2 + (28)^2 + \ldots + (19)^2 - \frac{(320)^2}{16}$$

$$= 6792 - 6400 = 392,$$

$$\text{Varieties } SS = \sum \frac{T_j^2}{r} - \frac{T_{..}^2}{cr} = \frac{26272}{4} - 6400 = 168;$$

$$\text{Blocks } SS = \sum \frac{T_{i.}^2}{c} - \frac{T_{..}^2}{cr} = \frac{25888}{4} - 6400 = 72, \text{ and}$$

Error SS = Total SS − (Varieties SS + Blocks SS)

$$= 392 - (168 + 72) = 152$$

The *ANOVA*-Table is:

Source of Variation	*d.f*	*SS*	*MS*	*F*
Varieties	3	168	56	$F_1 = 3.32$
Blocks	3	72	24	$F_2 = 1.42$
Error	9	152	16.89	--
Total	15	392	--	

(v) The critical regions are $F_j \geq F_{0.05(3,9)} = 6.99$

(vi) **Conclusion.** Since the computed values of F in both cases do not fall in the critical regions, we therefore accept both the null hypotheses.

(b) If no blocking had been done, then the analysis of variance would become.

S.V.	*d.f*	*SS*	*MS*	*F*
Varieties	3	168	56	3.00
Error	12	224	18.67	--
Total	15	392	--	

The critical region would be $F > F_{0.05(3,12)} = 5.95$.

Since the computed value of $F=3.00$ does not fall in the critical region, we therefore accept the hypothesis of no difference in the yields of four varieties.

Q.23.11 (a) To perform the analysis of variance, we first arrange the data as below:

Blocks	Methods (Treatments)			
	A	*B*	*C*	$B_{i.}$
1	813	647	713	2173
2	795	759	814	2368
3	705	598	652	1955
4	774	559	617	1950
5	687	580	539	1806
6	581	480	437	1498
$T_{.j}$	4355	3623	3772	11750

$$\text{Total } SS = \sum_i \sum_j Y_{ij}^2 - \frac{T_{..}^2}{cr}$$

$$= (813)^2 + (795)^2 + \ldots + (437)^2 - \frac{(11750)^2}{18}$$

$$= 7888448 - 7670138.9 = 218309.1$$

$$\text{Treatments } SS = \sum \frac{T_j^2}{r} - \frac{T_{..}^2}{cr}$$

$$= \frac{(4355)^2 + (3623)^2 + (3772)^2}{6} - \frac{(11750)^2}{18}$$

$$= 7720023 - 7670138.9 = 49884.1$$

$$\text{Blocks } SS = \sum \frac{B_{i.}^2}{c} - \frac{T_{..}^2}{cr}$$

$$= \frac{(2173)^2 + (2368)^2 + \ldots + (1498)^2}{3} - \frac{(11750)^2}{18}$$

$$= 7819839.3 - 7670138.9 = 149700.4$$

These results are presented in the following *ANOVA*-Table:

Source of Variation	*d.f*	*SS*	*MS*	*F*
Treatments	2	49884.1	24942.1	13.3
Blocks	5	149700.4	29940.1	
Error	10	18724.6	1872.5	--
Total	17	218309.1	--	

The critical region is $F \geq F_{0.05(2,10)} = 4.10$

Since the computed value of $F=13.3$ falls in the critical region, we therefore conclude that there is a strong indication of real treatment differences.

(b) If the effect of the block had been ignored, the results form a one-way experiment and the analysis of variance table would become

S.V.	d.f	SS	MS	F
Treatments	2	49884.1	24942.1	2.22
Error	15	168425.0	11288.0	--
Total	17	218309.1	--	

The treatments show no significant differences in this case as the computed value of $F=2.22$ does not fall in the critical region which is $F \geq F_{0.05(2,15)} = 3.68$.

Q.23.12. (a) The statistical model for a randomized block design with b blocks and p-treatments is

$$Y_{ij} = \mu + B_i + \tau_j + e_{ij}, \qquad i = 1, 2, ..., b,\ j = 1, 2, ..., p.$$

where B_i represents the block effects, τ_j represents the treatment effects and e_{ij} represents the random error which is a normally and independently distributed effect.

The appropriate analysis of variance table for this design is

S.V.	d.f	SS	MS
Between Blocks (B_i)	$(b-1)$	$\sum_i \frac{B_{i.}^2}{p} - \frac{T_{..}^2}{bp} = Q_1$	$s_b^2 = Q_1/(b-1)$
Between Treatments (T_j)	$(p-1)$	$\sum_j \frac{T_j^2}{b} - \frac{T_{..}^2}{bp} = Q_2$	$s_t^2 = Q_2/(p-1)$
Error (e_{ij})	$(b-1)(p-1)$	By Subtraction $= Q_3$	$s_e^2 = Q_3/(b-1)\times(p-1)$
Total	$bp-1$	$\sum_i \sum_j Y_{ij}^2 - \frac{T_{..}^2}{bp}$	--

The test-statistics to test the significance of the block effects and treatment effects are:

$F_1 = \frac{s_b^2}{s_e^2}$ and $F_2 = \frac{s_t^2}{s_e^2}$, which under H_0 have the F-distributions with appropriate degrees of freedom.

(b) Computations for the given data:

Blocks	Treatments 1	2	3	4	$B_{i.}$	$B_{i.}^2$
I	23	19	25	23	90	8100
II	20	17	24	21	82	6724
III	24	20	29	27	100	10000
IV	22	21	24	18	85	7225
$T_{.j}$	89	77	102	89	357	32049
T_j^2	7921	5929	10404	7921	32175	--
$\sum_i Y_{ij}^2$	1989	1491	2618	2023	8121	--

$$\text{Total } SS = \sum_i \sum_j Y_{ij}^2 - \frac{T_{..}^2}{bp} = 8121 - \frac{(357)^2}{16}$$

$$= 8121 - 7965.56 = 155.44,$$

$$\text{Block } SS = \sum_i \frac{B_{i.}^2}{b} - \frac{T_{..}^2}{bp} = \frac{32049}{4} - \frac{(357)^2}{16}$$

$$= 8012.55 - 7985.56 = 46.69, \text{and}$$

$$\text{Treatment } SS = \sum_j \frac{T_j^2}{b} - \frac{T_{..}^2}{bp} = \frac{32175}{4} - \frac{(357)^2}{16}$$

$$= 8043.75 - 7965.56 = 78.19, \text{and}$$

Error SS = Total SS − (Block SS + Treatment SS)

$$= 155.44 - (46.69 + 78.19) = 30.56.$$

Hence the analysis of variance table is:

Source of Variation	d.f	SS	MS	F
Blocks	3	46.69	15.563	--
Treatments	3	78.19	26.063	7.67
Error	9	30.56	3.396	--
Total	15	155.44	--	--

To test the hypothesis of no difference among the treatment means, *i.e.* $H_0 : \mu_{\cdot 1} = \mu_{\cdot 2} = \mu_{\cdot 3} = \mu_{\cdot 4}$, F value is 7.67, which is significantly larger than the corresponding critical $F = 3.86$ at the 5-percent level. Hence we reject H_0 and conclude that the treatment means are significantly different.

Q.23.13 (a) For an analysis of variance, the data are arranged as below:

Blocks	Fertilizers				
	f_1	f_2	f_3	f_4	$B_{i\cdot}$
1	42.7	39.3	48.5	32.8	163.3
2	50.0	38.0	50.9	40.2	179.1
3	51.9	46.3	53.5	51.1	202.8
$T_{\cdot j}$	144.6	123.6	152.9	124.1	545.2

Now the computations:

$$\text{Total } SS = \sum_i \sum_j Y_{ij}^2 - \frac{T_{\cdot\cdot}^2}{cr}$$

$$= (42.7)^2 + (50.0)^2 + \ldots + (51.1)^2 - \frac{(545.2)^2}{12}$$

$$= 25257.48 - 24770.25 = 487.23$$

$$\text{Fertilizers } SS = \sum \frac{T_j^2}{r} - \frac{T_{\cdot\cdot}^2}{cr}$$

$$= \frac{(144.6)^2 + (123.6)^2 + (152.9)^2 + (124.1)^2}{3} - \frac{(545.2)^2}{12}$$

$$= 24988.45 - 24770.25 = 218.20;$$

$$\text{Blocks } SS = \sum \frac{B_{i\cdot}^2}{c} - \frac{T_{\cdot\cdot}^2}{cr}$$

$$= \frac{(163.3)^2 + (179.1)^2 + (202.8)^2}{4} - \frac{(545.2)^2}{12}$$

$$= 24967.89 - 24770.25 = 197.64.$$

The *ANOVA*-Table is:

Source of Variation	*d.f*	*SS*	*MS*	*F*
Fertilizers	3	218.20	72.73	6.11
Blocks	2	197.64	98.82	8.30
Error	6	71.39	11.90	--
Total	11	487.23	--	

(b) Let Q_1 denote the contrast between (f_1, f_3) and (f_2, f_4).

Then $Q_1 = T_1 + T_3 - T_2 - T_4$, where T_i denotes the respective total.

$$= 144.6 + 152.9 - 123.6 - 124.1 = 49.8, \text{ and}$$

$$SSQ_1 = \frac{Q_1^2}{\sum r_j c_j^2} = \frac{(49.8)^2}{3[(1)^2 + (1)2 + (-1)^2]} = 206.67.$$

Again, let Q_2 denote the contrast between f_1 and f_3. Then

$$Q_2 = T_1 - T_3 = 144.6 - 152.9 = -8.3, \text{ and}$$

$$SSQ_2 = \frac{(-8.3)^2}{3\,[(1)^2 + (-1)^2]} = 11.48$$

These results are put in the *ANOVA*-Table as:

S.V.	*d.f*	*SS*	*MS*	*F*	
(f_1, f_3) vs. (f_2, f_4)	1	206.67	206.67	17.37	Significant
(f_1) vs. (f_3)	1	11.48	11.48	0.96	Not-significant
Error	6	71.39	11.90		

Q.23.14. Computations of the analysis of variance and the *F*-ratio

$$\text{Total } SS = \sum_i \sum_j Y_{ij}^2 - C.F. = 425{,}314.$$

$$\text{Block } SS = \sum_i \frac{B_{i\cdot}^2}{c} - C.F.$$

$$= \frac{(1487)^2 + (1134)^2 + (1892)^2 + (614)^2}{4} - C.F.$$

$$= \frac{7453785}{4} - C.F. = 1863446 - 1642883 = 220563,$$

$$\text{Treatment } SS = \sum_j \frac{B_j^2}{r} - C.F.$$

$$= \frac{(819)^2 + (904)^2 + (1840)^2 + (1563)^2}{4} - C.F.$$

$$= \frac{7319673}{4} - C.F = 1829918 - 1642883 = 187035.$$

Error SS = Total SS − (Block SS + Treatment SS)

= 425,314 − (220,563 + 187,035) = 17,716

Thus the analysis of variance table is:

Source of Variation	*d.f*	*SS*	*MS*
Blocks	3	220,563	73,521
Treatments	3	187,035	62,345
Error	9	17,716	1,968
Total	15	425,314	--

The F-ratio for treatment $= \dfrac{MS \text{ for Treatments}}{MS \text{ for Error}}$

$$= \frac{62,345}{1968} = 31.68$$

Q.23.15 (a) Completion of the analysis

Source of Variation	*d.f*	*SS*	*MS*	*F*
Blocks	9	0.4074	0.0453	--
Treatments	3	1.1986	0.3995	17.29
Error	27	0.6247	0.0231	--
Total	39	2.2309	--	

(b) The standard error of a treatment mean is estimated as

$$S.E. = \sqrt{\frac{\text{error mean square}}{\text{no. of observations in treatment mean}}}$$

$$= \sqrt{\frac{0.0231}{10}} = \sqrt{0.00231} = 0.048, \text{ and}$$

the standard error for the difference between 2 treatment mean is

$$S.E. = \sqrt{s_e^2\left(\frac{1}{r_1} + \frac{1}{r_2}\right)} = \sqrt{\frac{2s_e^2}{r}}, \text{ when } r_1 = r_2 = r.$$

$$= \sqrt{\frac{2(0.0231)}{10}} = \sqrt{0.00462} = 0.068$$

(c) To test the difference between two treatment means selected at random, we calculate the *least significant difference* (*LSD*) by

$$LSD = t_{0.025(27)}\sqrt{\frac{2s_e^2}{r}} = 2.052\sqrt{\frac{2(0.0231)}{10}}$$

$$= 2.052 \times 0.068 = 0.140$$

Arranging the treatment means in ascending order of magnitude and drawing a line under set for means that are not significantly different, we get

T_2	T_3	T_1	T_4
1.195	1.325	1.464	1.662

The pair of means representing different populations are immediately observed from this presentation.

(d) The efficiency of this design relative to a completely randomized design is estimated by the relation

$$\text{Relative Efficiency} = \frac{(r-1)\,s_b^2 + r(k-1)\,s_e^2}{(rk-1)\,s_e^2}$$

where r = no. of blocks and k = no. of treatments.

Hence the required $R.E.$ $= \dfrac{(10-1)(0.0453) + 10(4-1)(0.0231)}{(10\times4-1)(0.0231)}$

$$= \frac{1.1007}{0.9009} = 1.22 = 122\%$$

Q.23.17. (b) Computations for the missing value in a Randomized Complete Block Design.

Age	Tests					$T_{i.}$
	A	B	C	D	E	
10	5	6	6	7	4	28
11	x	7	7	8	5	$27+x$
12	7	7	7	9	5	35
$T_{.j}$	$12+x$	20	20	20	14	$90+x$

Now Error SS = Total SS − Treatment SS − Block SS

$$= \sum_i \sum_j Y_{ij}^2 - \frac{T_{..}^2}{n} - \left(\sum_j \frac{T_j^2}{r} - \frac{T_{..}^2}{n}\right) - \left(\sum_i \frac{T_{i.}^2}{c} - \frac{T_{..}^2}{n}\right)$$

$$= \sum_i \sum_j Y_{ij}^2 - \sum_j \frac{T_j^2}{r} - \sum_i \frac{T_{i.}^2}{c} + \frac{T_{..}^2}{n}$$

$$\text{Thus Error } SS = (5)^2 + (x)^2 + (7)^2 + \ldots + (5)^2 - \frac{(12+x)^2 + (20)^2 + \ldots + (14)^2}{3}$$

$$- \frac{(28)^2 + (27+x)^2 + (35)^2}{5} + \frac{(90+x)^2}{15}$$

To find the best estimate of x that will minimize the Error SS, we differentiate it *w.r.t.* x and equate it to zero. Thus we get

$$\frac{d(\text{Error } SS)}{dx} = 2x - \frac{2(12+x)}{3} - \frac{2(27+x)}{5} + \frac{2(90+x)}{15} = 0$$

($\because$ Derivatives of constant terms are zero)

or $15x - 5(12+x) - 3(27+x) + (90+x) = 0$

or $15x - 5x - 3x + x = 60 + 81 - 90$

$8x = 51$

$x = 6.375 = 6$ approximately.

For the augmented data, we prepare the *ANOVA*-Table as follows:

Age	Tests					$T_{i.}$	$T_{i.}^2$
	A	B	C	D	E		
10	5	6	6	7	4	28	784
11	6	7	7	8	5	33	1089
12	7	7	7	9	5	35	1225
$T_{.j}$	18	20	20	24	14	96	3098
T_j^2	324	400	400	576	196	1896	--
$\sum_i Y_{ij}^2$	110	134	134	194	66	638	--

$$\text{Total } SS = \sum_i \sum_j Y_{ij}^2 - \frac{T_{..}^2}{cr}$$

$$= 638 - \frac{(96)^2}{15} = 638 - 614.4 = 23.6,$$

$$\text{Treatment } SS = \sum_j \frac{T_j^2}{r} - \frac{T_{..}^2}{cr}$$

$$= \frac{1896}{3} - \frac{(96)^2}{15} = 632 - 614.4 = 17.6,$$

$$\text{Block (Age) } SS = \sum_i \frac{T_{i.}^2}{r} - \frac{T_{..}^2}{n}$$

$$= \frac{3098}{5} - \frac{(96)^2}{15} = 619.6 - 614.4 = 5.2, \text{ and}$$

Error SS = Total SS − Treatment SS − Block SS

$$= 23.6 - 17.6 - 5.2 = 0.8$$

Thus the analysis of variance table for this design with one missing observation is:

Source of Variation	*d.f*	*SS*	*MS*	*F*
Treatments (Tests)	4	17.6	4.4	38.60
Blocks (Ages)	2	5.2	2.6	--
Error	7	0.8	0.114	--
Total	13	23.6	--	--

To test the significance of tests (treatments), we obtained $F=38.60$ which is significantly larger than the corresponding critical $F=4.12$ at the 5-percent level for (4 and 7) *d.f.*

Q.23.18. Estimation of missing values in the Randomized Block Design by Iterative calculations. The data with the missing observations are given below:

Blocks	Treatments						Block Total ($B_{i\cdot}$)
	1	2	3	4	5	6	
I	18.5	15.7	16.2	14.1	13.0	13.6	91.1
II	11.7	y_{22}	12.9	y_{24}	16.9	12.5	$54.0+y_{22}+y_{24}$
III	15.4	16.6	15.5	20.3	18.4	21.6	107.8
IV	y_{41}	18.6	12.7	15.7	16.5	18.0	$81.5+y_{41}$
Treatment Total ($T_{\cdot j}$)	$45.6+y_{41}$	$50.9+y_{22}$	57.3	$50.1+y_{24}$	64.8	65.7	$334.4+y_{41}+y_{22}+y_{24}$

There are three missing observations, y_{41}, y_{22} and y_{24}. We assign some reasonable values (say, the mean of the block mean and the treatment mean relating to the missing value) to two of the three missing observations, say, y_{22} and y_{24} and compute y_{41} by the formula

$$y_{ij} = \frac{kT_{\cdot j} + r B_{i\cdot} - G}{(k-1)(r-1)},$$ where the symbols have their usual meaning.

Now let $y_{22} = \frac{\bar{y}_{\cdot 2}+\bar{y}_{2\cdot}}{2} = \frac{(50.9/3)+(54.0/4)}{2} = 15.2$, and

$$y_{24} = \frac{\bar{y}_{\cdot 4}+\bar{y}_{2\cdot}}{2} = \frac{(50.1/3)+(54.0/4)}{2} = 15.1$$

Then the value of y_{41} is obtained as follows:

First Cycle.

$$y_{41} = \frac{kT_{\cdot 1} + rB_{4\cdot} - G}{(k-1)(r-1)} = \frac{6(45.6) - 4(81.5) - 364.7}{(6-1)(4-1)}$$

$$= \frac{234.9}{15} = 15.7.$$

(Here $T_{\cdot 1}=45.6$, $B_{4\cdot}=81.5$ and $G = 334.4 + 15.2 + 15.1 = 364.7$)

Using this value of y_{41} and the assigned value of y_{24}, we calculate the value of y_{22} as

$$y_{22} = \frac{kT_{\cdot 2} + rB_{2\cdot} - G}{(k-1)(r-1)} = \frac{6(50.9) + 4(69.1) - 365.2}{(6-1)(4-1)}$$

$$= \frac{216.6}{15} = 14.4$$

(Here $B_{2\cdot}=54.0+15.1=69.1$ and $G=334.4+15.1+15.7 = 365.2$)

Using the calculated values of y_{41} and y_{22}, we calculate the value of y_{24} as

$$y_{24} = \frac{kT_{\cdot 4} + rB_{2\cdot} - G}{(k-1)(r-1)} = \frac{6(50.1) + 4(68.4) - 364.4}{(6-1)(4-1)}$$

$$= \frac{209.7}{15} = 14.0$$

(Here $B_{2\cdot}=54.0+14.4=68.4$ and $G=334.4+15.7+14.4 = 364.5$)

Second Cycle.

Using the calculated values of y_{22} and y_{24}, we recalculate y_{41} as

$$y_{41} = \frac{6(45.6) + 4(81.5) - 362.8}{(6-1)(4-1)} = \frac{236.8}{15} = 15.8$$

$\{G = 334.4 + 14.4 + 14.0 = 362.8\}$

Now using the calculated value of y_{41} and y_{24}, we recalculate y_{22} as

$$y_{22} = \frac{6(50.9) + 4(68.0) - 364.2}{(6-1)(4-1)} = \frac{213.2}{15} = 14.2$$

$\{B_{2.} = 54.0+14.0=68.0,\ G=334.4+14.0+15.8=364.2\}$

Again using the estimated values of y_{41} and y_{22}, we recalculate y_{24} as

$$y_{24} = \frac{6(50.1) + 4(68.2) - 364.4}{(6-1)(4-1)} = \frac{209.0}{15} = 13.9$$

$\{G=334.4+14.2+15.8=364.4\}$

Third Cycle.

Continuing the procedure, we find that the values remain the same. Hence the final estimates are $y_{41}=15.8$, $y_{22}=14.2$ and $y_{24}=13.9$. Inserting these values, we perform the analysis of variance as follows:

Blocks	Treatments						$(B_{i.})$
	1	2	3	4	5	6	
I	18.5	15.7	16.2	14.1	13.0	13.6	91.1
II	11.7	14.2	12.9	13.9	16.9	12.5	82.1
III	15.4	16.6	15.5	20.3	18.4	21.6	107.8
IV	15.8	18.6	12.7	15.7	16.5	18.0	97.3
$T_{.j}$	61.4	65.1	57.3	64.0	64.8	65.7	378.3
$\sum_i Y_{ij}^2$	965.94	1069.65	830.39	1050.60	1065.42	1131.77	6113.77

$$\text{Total } SS = \sum_i\sum_j Y_{ij}^2 - \frac{T_{..}^2}{n} = 6113.77 - \frac{(378.3)^2}{24}$$

$$= 6113.77 - 5962.95 = 15.82,$$

$$\text{Treatment } SS = \sum_j \frac{T_j^2}{r} - \frac{T_{..}^2}{n}$$

$$= \frac{(61.4)^2+(65.1)^2+\ldots+(65.7)^2}{4} - \frac{(378.3)^2}{24}$$

$$= \frac{23902.79}{4} - 5962.95 = 5975.70 - 5962.95 = 12.75,$$

$$\text{Blocks } SS = \sum_i \frac{B_{i.}^2}{r} - \frac{T_{..}^2}{n} = \frac{(91.1)^2+\ldots+(97.3)^2}{6} - \frac{(378.3)^2}{24}$$

$$= \frac{36127.75}{6} - 5962.95 = 6021.29 - 5962.95 = 58.34, \text{and}$$

Error SS = Total SS − (Treatments SS + Block SS) = 79.73.

Hence the analysis of variance table for the augmented data is

Source of Variation	$d.f$	SS	MS	F
Blocks	3	58.34	19.45	2.93
Treatments	5	12.75	2.55	0.38
Error	12	79.73	6.64	--
Total	20	150.82	--	--

Q.23.19. (b) In order to apply the covariance analysis technique, we insert zero for the missing observation and pair the observation with the covariate composed of zeros and a one as shown in the table:

Blocks	TREATMENTS								$T_{i\cdot}$	$T'_{i\cdot}$	$T'^2_{i\cdot}$
	I		II		III		IV				
	X	Y	X	Y	X	Y	X	Y			
1	0	4.4	0	6.8	0	6.3	0	6.4	0	23.9	571.21
2	0	4.0	1	0	0	4.9	0	7.3	1	16.2	262.44
3	0	4.5	0	7.0	0	5.9	0	6.7	0	24.1	580.81
4	0	3.1	0	6.4	0	7.1	0	6.7	0	23.3	542.89
$T_{\cdot j}, T'_{\cdot j}$	0	16.0	1	20.2	0	24.2	0	27.1	1	87.5	1957.35
$T'^2_{\cdot j}$	256.00		408.04		585.64		734.41		---	---->	1984.09

(i) Total SS and Products.

$$S_{xx} = \sum_i\sum_j X_{ij}^2 - \frac{T_{..}^2}{n} = 1 - \frac{(1)^2}{16} = 1 - 0.0625 = 0.9375,$$

$$S_{yy} = \sum_i\sum_j Y_{ij}^2 - \frac{T'^2_{..}}{n} = (4.4)^2 + (4.0)^2 + \ldots + (6.7)^2 - \frac{(87.5)^2}{16}$$

$$= 534.37 - 478.52 = 55.85,$$

$$S_{xy} = \sum_i \sum_j X_{ij} Y_{ij} - \frac{(T..)\,(T..')}{n} = 0 - \frac{(1)\,(87.5)}{16} = -5.47$$

(ii) Treatments SS and Products.

$$T_{xx} = \sum_j \frac{T_j^2}{r} - \frac{T_{..}^2}{n} = \frac{1}{4} - \frac{1}{16} = 0.1875,$$

$$T_{yy} = \sum_j \frac{T_j'^2}{r} - \frac{T_{..}'^2}{n} = \frac{1984.09}{4} - \frac{(87.5)^2}{16}$$

$$= 496.02 - 478.52 = 17.50,$$

$$T_{xy} = \sum_j \frac{(T_{.j})\,(T'_{.j})}{r} - \frac{(T..)\,(T'..)}{n} = \frac{(1)\,(20.2)}{4} - \frac{(1)\,(87.5)}{16}$$

$$= 5.05 - 5.47 = -0.42$$

(iii) Blocks SS and Products.

$$B_{xx} = \sum_i \frac{T_{i.}^2}{c} - \frac{T_{..}^2}{n} = \frac{1}{4} - \frac{1}{16} = 0.1875,$$

$$B_{yy} = \sum_i \frac{(T'_{i.})^2}{c} - \frac{(T'_{..})^2}{n} = \frac{1957.35}{4} - \frac{(87.5)^2}{16}$$

$$= 489.34 - 478.52 = 10.82,$$

$$B_{xy} = \sum_i \frac{(T_{i.})(T'_{i.})}{c} - \frac{(T..)(T'_{..})}{n} = \frac{(1)\,(16.2)}{4} - \frac{(1)\,(87.5)}{16}$$

$$= 4.05 - 5.47 = -1.42$$

(iv) Error *SS* and Products

$E_{xx} = S_{xx} - T_{xx} - B_{xx} = 0.9375 - 0.1875 - 0.1875 = 0.5625,$

$E_{yy} = S_{yy} - T_{yy} - B_{yy} = 55.85 - 17.50 - 10.82 = 27.53,$ and

$E_{xy} = S_{xy} - T_{xy} - B_{xy} = -5.47 - (-0.42) - (-1.42) = -3.63.$

The analysis of covariance table is then set up as below:

Source of Variation	Sum of Squares and Products				Regression Coefficient
	d.f.	$\sum x^2$	$\sum y^2$	$\sum xy$	
Blocks	3	0.1875	10.82	−1.42	
Treatments	3	0.1875	17.50	−0.42	
Error	9	0.5625	27.53	−3.63	$b = -6.453$
Total	15	0.9375	55.85	−5.47	

Hence the estimate of the missing observation $= -b = 6.5$.

Now computations for analysis of variance of the augmented data proceed as follows:

Blocks	Treatments				$T_{i.}$	$T_{i.}^2$
	I	II	III	IV		
1	4.4	6.8	6.3	6.4	23.9	571.21
2	4.0	6.5	4.9	7.3	22.7	515.29
3	4.5	7.0	5.9	6.7	24.1	580.81
4	3.1	6.4	7.1	6.7	23.3	542.89
$T_{.j}$	16.0	26.7	24.2	27.1	94.0	2210.20
T_j^2	256.00	712.89	585.64	734.41	2288.94	--
$\sum_i Y_{ij}^2$	65.22	178.45	148.92	184.03	576.62	--

$$\text{Total } SS = \sum_i \sum_j Y_{ij}^2 - \frac{T_{..}^2}{n} = 576.62 - \frac{(94)^2}{16}$$

$$= 576.62 - 552.25 = 24.37.$$

$$\text{Treatment } SS = \sum_j \frac{T_j^2}{r} - \frac{T_{..}^2}{n} = \frac{2288.94}{4} - \frac{(94)^2}{16}$$

$$= 572.24 - 552.25 = 19.99,$$

$$\text{Blocks } SS = \sum_i \frac{T_{i.}^2}{c} - \frac{T_{..}^2}{n} = \frac{2210.20}{4} - \frac{(94)^2}{16}$$

$$= 552.55 - 552.25 = 0.30, \text{ and}$$

$$\text{Error } SS = \text{Total } SS - (\text{Treatments } SS + \text{Block } SS)$$

$$= 24.37 - (19.99 + 0.30) = 4.08.$$

The analysis of variance table for the augmented data is

Source of Variation	*d.f*	*SS*	*MS*
Blocks	3	0.30	--
Treatments	3	19.99	6.66
Error	8	4.08	0.51
Total	14	24.37	--

Q.23.22. (b) Computations for the anlaysis of variance in a Latin Square Design.

Rows	Columns				R_i	R_i^2
	1	2	3	4		
1	2.3 V_1	3.0 V_2	3.3 V_3	2.5 V_4	11.1	123.21
2	3.1 V_2	4.1 V_3	2.4 V_4	2.4 V_1	12.0	144.00
3	4.3 V_3	2.5 V_4	2.1 V_1	2.9 V_2	11.8	139.24
4	2.6 V_4	2.0 V_1	2.4 V_2	4.4 V_3	11.4	129.96
C_j	12.3	11.6	10.2	12.2	46.3	536.41
C_j^2	151.29	134.56	104.04	148.84	538.73	--
$\sum_i Y_{ij}^2$	40.15	36.06	26.82	39.78	142.81	--

$$\text{Now, Total } SS = \sum_i \sum_j Y_{ij(h)}^2 - \frac{G^2}{k^2} = 142.81 - \frac{(46.3)^2}{16}$$

$$= 142.81 - 133.98 = 8.83$$

$$\text{Rows } SS = \sum_i \frac{R_i^2}{k} - \frac{G^2}{k^2} = \frac{536.41}{4} - \frac{(46.3)^2}{16} = 134.10 - 133.98 = 0.12$$

$$\text{Columns } SS = \sum_j \frac{C_j^2}{k} - \frac{G^2}{k^2} = \frac{538.73}{4} - \frac{(46.3)^2}{16}$$

$$= 134.68 - 133.98 = 0.70$$

For treatment totals, squares and means, we construct another table as follows:

Variety	T_h	T_h^2	Means
V_1	8.8	77.44	2.20
V_2	11.4	129.96	2.85
V_3	16.1	259.21	4.02
V_4	10.0	100.00	2.50
Total	46.3	566.61	---

$$\therefore \text{ Treatment } SS = \sum_h \frac{T_h^2}{k} - \frac{G^2}{k^2} = \frac{566.61}{4} - \frac{(46.3)^2}{16}$$

$$= 141.65 - 133.98 = 7.67, \text{ and}$$

$$\text{Error } SS = \text{Total } SS - (\text{Rows } SS + \text{Columns } SS + \text{Treatment } SS)$$

$$= 8.83 - (0.12 + 0.70 + 7.67) = 0.34.$$

These results are arranged in the following Analysis of Variance table:

Source of Variation	*d.f*	*SS*	*MS*	*F*
Rows	3	0.12	0.04	--
Columns	3	0.70	0.23	--
Treatments	3	7.67	2.557	44.86
Error	6	0.84	0.057	--
Total	15	8.83	--	--

The varieties are significantly different at the 5-per cent level.

Q.23.23. (i) The null hypotheses would be stated as

H_0 : There is no difference among the districts,

H_0' : There is no variation among the heights, and

H_0'' : There is no difference between the sensitivity of the test.

(ii) We use a significance level of $\alpha = 0.05$.

(iii) The test-statistic is $F = \dfrac{MS\ (\text{Effect})}{MS\ (\text{Error})}$, which under H_0, has an F-distribution provided the assumptions of normality, homogeneity and independence are satisfied.

(iv) Computations.

Heights	Districts				R_i	R_i^2
	1	2	3	4		
1	A 8	B 5.3	C 4.1	D 5	22.4	501.76
2	D 6.8	A 4.9	B 4.1	C 3.2	19.0	361.00
3	B 6.3	C 4.7	D 4.0	A 5	20.0	400.00
4	C 5.7	D 3.3	A 4.0	B 4.2	17.2	295.84
C_j	26.8	18.2	16.2	17.4	78.6	1558.60
C_j^2	718.24	331.24	262.44	302.76	1614.68	--
$\sum_i Y_{ij(h)}^2$	182.42	85.08	65.62	77.88	411.00	--

Test	A	B	C	D	Total
T_h	21.9	19.9	17.7	19.1	78.6
T_h^2	479.61	396.01	313.29	364.81	1553.72

Now, Total $SS\ (S_{yy}) = \sum_i \sum_j Y_{ij(h)}^2 - \dfrac{G^2}{k^2} = 411.00 - \dfrac{(78.6)^2}{16}$

$= 411.00 - 386.1225 = 24.8775,$

Rows $SS\ (R_{yy}) = \sum_i \dfrac{R_i^2}{k} - \dfrac{G^2}{k^2} = \dfrac{1558.60}{4} - \dfrac{(78.6)^2}{16}$

$= 389.65 - 386.1225 = 3.5275,$

Columns $SS\ (C_{yy}) = \sum_j \dfrac{C_j^2}{k} - \dfrac{G^2}{k^2} = \dfrac{1614.68}{4} - \dfrac{(78.6)^2}{16}$

$= 403.67 - 386.1225 = 17.5475,$

Tests $SS\ (T_{yy}) = \sum_h \dfrac{T_h^2}{k} - \dfrac{G^2}{k^2} = \dfrac{1553.72}{4} - \dfrac{(78.6)^2}{16}$

$= 388.43 - 386.1225 = 2.3075,$ and

Error $SS\ (E_{yy}) = S_{yy} - (R_{yy} + C_{yy} + T_{yy})$

$= 24.8775 - (3.5275 + 17.5475 + 2.3075)$

$= 1.4950$

These results are arranged in the following analysis of variance table:

Source of Variation	$d.f.$	Sum of Squares	Mean Square	F	5 percent critical F
Districts	3	17.5475	5.8492	23.47	4.76
Heights (Rows)	3	3.5275	1.1758	4.72	4.76
Tests	3	2.3075	0.7692	3.09	4.76
Error	6	1.4950	0.2490	--	--
Total	15	24.8775	--	--	--

(v) The critical region is $F_{\cdot} > F_{.05(3,6)} = 4.76$.

(vi) Comparing the computed valus of F with the corresponding critical F, we find that

(a) The variation among the districts is significant.

(b) The variation among the heights is not significant, and

(c) There is no difference between the sensitivity of the test.

Q.23.24. Computations for analysis of variance in a Latin Square Design. To simplify the computational work, we subtract 100 from all the observations and carry out the arithmetic as below:

Runs	Positions 1	2	3	4	R_i	R_i^2
1	A 50	B 45	D 30	C 33	158	24964
2	D 30	C 72	A 70	B 27	199	39601
3	C 33	D 32	B 15	A 70	150	22500
4	B -2	A 71	C 32	D 20	121	14641
C_j	111	220	147	150	628	101706
C_j^2	12321	48400	21609	22500	104830	--
$\sum_i Y_{ij(h)}^2$	4493	13274	7049	7118	31934	--

Grades of Leather (Treatments)

Treatment	A	B	C	D	Total
T_h	261	85	170	112	628
T_h^2	68121	7225	28900	12544	116790

$$\text{Now, Total } SS\ (S_{yy}) = \sum_i \sum_j Y_{ij(h)}^2 - \frac{G^2}{k^2} = 31934 - \frac{(628)^2}{16}$$

$$= 31934 - 24649 = 7285,$$

$$\text{Position } SS\ (C_{yy}) = \sum_j \frac{C_j^2}{k} - \frac{G^2}{k^2} = \frac{104830}{4} - \frac{(628)^2}{16}$$

$$= 26207.5 - 24649 = 1558.5,$$

$$\text{Runs } SS\ (R_{yy}) = \sum_i \frac{R_i^2}{k} - \frac{G^2}{k^2} = \frac{101706}{4} - \frac{(628)^2}{16}$$

$$= 25426.5 - 24649 = 777.5,$$

$$\text{Treatments } SS\ (T_{yy}) = \sum_h \frac{T_h^2}{k} - \frac{G^2}{k^2} = \frac{116790}{4} - \frac{(628)^2}{16}$$

$$= 29197.5 - 24649 = 4548.5, \text{ and}$$

$$\text{Error } SS\ (E_{yy}) = S_{yy} - (C_{yy} + R_{yy} + T_{yy})$$

$$= 7285 - (1558.5 + 777.5 + 4548.5) = 400.5$$

An analysis of variance on these data gives the results in table-below:

ANOVA-Table

Source of Variation	*d.f.*	Sum of Squares	Mean Square	*F*
Positions	3	1558.5	519.50	--
Runs	3	777.5	259.17	--
Treatments	3	4548.5	1516.17	22.71
Error	6	400.5	66.75	--
Total	15	7285.0	--	--

For grades of leather (treatments), we found $F = 22.71$, which is significantly larger than the corresponding 5-per cent critical $F = 4.76$. Hence there is evidence to conclude that the grades of leather are significantly different.

Q.23.25. We are required to find which machine should be adjusted and by how much. The computations proceed as follows:

Source of Variation	*d.f.*	Mean Square	*F*
Rows (Operators)	3	136	--
Columns (Time Periods)	3	30	--
Machines (Treatments)	3	1649	20.36
Error	6	$s_e^2 = 81$	--

For machines, the F value is 20.36, which is significantly larger than the corresponding critical $F = 4.76$ at the 5-per cent level. This means that adjustment on machines is necessary. Hence we calculate the *least significant difference* (*LSD*) to

determine which machine is to be adjusted to produce more uniform product.

$$\text{Now,} \quad LSD = t_{0.025,(6)} \sqrt{\frac{2\,(MSE)}{k}}$$

$$= 2.447 \sqrt{\frac{2\,(81)}{4}} = 2.447 \times 6.364 = 15.57$$

Arranging the treatment (machines) means in ascending order of magnitudue, we get

T_3	T_4	T_2	T_1
40.00	43.00	44.75	83.00

Treatment No. I is significantly different from the other three treatments. This implies that machine No. I should be adjusted to produce more uniform product. The maximum allowable range within which all the four machines can be declared producing almost uniform product would be the smallest mean plus LSD, *i.e.* 40.00+15.57=55.57. Hence machine I is to be adjusted at least by 83.00–55.67=27.43 units.

Q.23.26. (b) The hypotheses are stated as

(i) $H_0 : \mu_A = \mu_B = \mu_C = \mu_D = \mu_E$, and
H_1 : At least two treatment effects differ.

(ii) The level of significance is set at $\alpha = 0.05$

(iii) The test statistic under H_0 is

$$F = \frac{MS\ (\text{Effect})}{MS\ (\text{Error})}, \quad \text{which has an } F\text{-distribution.}$$

(iv) Computations.

Rows	Columns					R_i	R_i^2
	I	II	III	IV	V		
I	52.5	46.3	44.1	48.1	40.9	231.9	53777.61
II	44.2	42.9	51.3	49.3	32.6	220.3	48532.09
III	49.1	47.3	38.1	41.0	47.2	222.7	49595.29
IV	43.2	42.5	67.2	55.1	45.3	253.3	64160.89
V	47.0	43.2	46.7	46.0	43.2	226.1	51121.21
C_j	236	222.2	247.4	239.5	209.2	1154.3	26187.09
C_j^2	55696	49372.84	61206.76	57360.25	43764.64	267400.49	--
ΣY_{ij}^2	11195.94	9893.88	12724.84	11577.11	8881.74	54273.51	--

$$C.F. = \frac{G^2}{n} = \frac{(1154.3)^2}{25} = 53296.3396.$$

$$\text{Total } SS = \sum_i \sum_j Y_{ij}^2 - C.F. = 54273.51 - 53296.3396 = 977.1704$$

$$\text{Between Columns } SS = \sum_j \frac{C_j^2}{k} - C.F.$$

$$= \frac{267400.49}{5} - 53296.3396 = 183.7584$$

$$\text{Between Rows } SS = \sum_i \frac{R_i^2}{k} - C.F.$$

$$= \frac{267187.09}{5} - 53296.3396 = 141.0784$$

Yields are rearranged to get the total for treatments as below:

	Treatments					Total
	A	B	C	D	E	
	52.5	40.9	48.1	44.1	46.3	
	51.3	42.9	32.6	44.2	49.3	
	47.3	49.1	38.1	41.0	47.2	
	45.3	55.1	43.2	42.5	67.2	
	46.0	46.7	43.2	43.2	47.0	
T_h	242.4	234.7	205.2	215.0	257.0	1154.3
T_h^2	58757.76	55084.09	42107.04	46225.00	66049.00	268222.89

$$\text{Between Treatments } SS = \sum \frac{T_h^2}{k} - C.F.$$

$$= \frac{268222.89}{5} - 53296.3396 = 348.2384$$

$$\therefore \text{ Error } SS = TSS - (\text{Columns } SS + \text{Rows } SS + \text{Treat } SS)$$

$$= 977.1704 - (183.7584 + 141.0784 + 348.2384)$$

$$= 340.0952$$

The ANOVA-TABLE is

Source of Variation	*d.f.*	*SS*	*MS*	*F*-ratio
Between Columns	4	183.7584	45.9396	--
Between Rows	4	141.0784	35.2696	--
Between Treatments	4	348.2384	87.0596	$F=3.44$
Error	12	304.0952	25.3413	--
Total	24	977.1704	--	--

(v) The critical region is $F \geq F_{0.05(4,12)} = 3.26$

(vi) Conclusion. Since the computed value of F falls in the critical region, we therefore reject H_0 and may conclude that there is a significant difference among treatment means.

Q.23.27. (i) We state our null hypotheses as

(a) H_0 : all effects on yield due to variation of soil in Columns are zero,

(b) H_0' : all effects on yield due to variation of soil in Rows are zero,

(c) H_0'' : all effects on yield due to change in Treatments, are zero,

(ii) We use a significance level of $\alpha = 0.05$.

(iii) The test-statistic is $F = \dfrac{MS\ (\text{Effect})}{MS\ (\text{Error})}$, which under H_0, has an F-distribution with $\upsilon_1 = k-1$, $\upsilon_2 = (k-1)(k-2)$ degrees of freedom.

(iv) Computations. Taking the origin at Y = 10, the computations proceed as follows:

Rows	Columns 1	2	3	4	5	R_i	R_i^2
1	−2.6 A (6.76)	−1.1 D (1.21)	−4.2 E (17.64)	2.0 B (4.00)	4.3 C (18.49)	−1.6	2.56
2	1.8 C (3.24)	−3.5 B (12.25)	−1.3 A (1.69)	−2.4 E (5.76)	−2.1 D (4.41)	−7.5	56.25
3	0.1 D (0.01)	7.9 C (62.41)	−1.0 B (1.00)	−1.5 A (2.25)	−2.9 B (8.41)	2.6	6.76
4	−1.2 E (1.44)	0.1 A (0.01)	5.7 A (32.49)	1.1 D (1.21)	−2.6 A (6.76)	3.1	9.61
5	1.8 B (3.24)	−1.2 E (1.44)	4.3 D (18.49)	8.4 C (70.56)	0.1 A (0.01)	13.4	179.56
C_j	−0.1	2.2	3.5	7.6	−3.2	10.0	254.74
C_j^2	0.01	4.84	12.25	57.76	10.24	85.10	--
$\sum_i Y_{ij(h)}^2$	14.69	77.32	71.31	83.78	38.08	285.18	

Treatment	A	B	C	D	E	Total
T_h	−5.2	−3.3	28.1	2.3	−11.9	10.0
T_h^2	27.04	10.89	789.61	5.29	141.61	974.44

Now, Total SS $(S_{yy}) = \sum_i \sum_j Y_{ij(h)}^2 - \dfrac{G^2}{k^2} = 285.18 - \dfrac{(10)^2}{25}$

$= 285.18 - 4.00 = 281.18,$

Columns SS $(C_{yy}) = \sum_j \dfrac{C_j^2}{k} - \dfrac{G^2}{k^2} = \dfrac{85.10}{5} - \dfrac{(10.0)^2}{25}$

$= 17.02 - 4.00 = 13.02,$

Rows SS $(R_{yy}) = \sum_i \dfrac{R_i^2}{k} - \dfrac{G^2}{k^2} = \dfrac{254.74}{5} - \dfrac{(10.0)^2}{25}$

$$= 50.95 - 4.00 = 46.95,$$

$$\text{Treatments } SS\ (T_{yy}) = \sum_h \frac{T_h^2}{k} - \frac{G^2}{k^2} = \frac{974.44}{5} - \frac{(10.0)^2}{25}$$

$$= 194.89 - 4.00 = 190.89, \text{ and}$$

$$\text{Error } SS\ (E_{yy}) = S_{yy} - (C_{yy} + R_{yy} + T_{yy})$$

$$= 281.18 - (13.02 + 46.95 + 190.89) = 30.32$$

Thus the analysis of variance table is

Source of Variation	*d.f.*	Sum of Squares	Mean Square	*F*
Columns	4	13.02	3.26	$F_1 = 1.29$
Rows	4	46.95	11.74	$F_2 = 4.64$
Treatments	4	190.89	47.72	$F_3 = 18.86$
Error	12	30.32	2.53	--
Total	24	281.18	--	--

(v) The critical region is $F > F_{.05,(4,12)} = 3.26$

(vi) Comparing the computed values of F with the corresponding critical F at the 5-percent level, we find that

(a) $F_1 < 3.26$. We accept H_0 and conclude that the variation of soil in columns is not significant.

(b) F_2 falls in the critical region and thus H_0' is rejected. The rejection of this H_0' says that there are significant differences in soil in rows.

(c) F_3 falls in the critical region, so we reject H_0''. The rejection of this H_0'' implies effects on yield due to change in treatments are significantly different.

Q.23.28. (i) The null hypotheses woud be stated as

(a) H_0: all the responses under different *periods* are the same,

(b) H_0': all the responses of different *monkeys* are the same, and

(c) H_0'': all the responses under different *conditions* (treatment) are the same.

(ii) We use a significance level of $\alpha = 0.05$.

(iii) The test-statistic is $F = \dfrac{MS\ (\text{Effect})}{MS\ (\text{Error})}$, which under H_0, has an F-distribution with $v_1 = 4$, $v_2 = 12$ degrees of freedom provided the assumptions of additivity, normality, homogeneity, and independence are satisfied.

(iv) Computations.

Monkeys	Periods 1	2	3	4	5	R_i
1	194 B	369 D	344 C	380 A	693 E	1980
2	202 D	142 B	200 A	356 E	473 C	1373
3	355 C	301 A	439 E	338 B	528 D	1961
4	515 E	590 C	552 B	677 D	546 A	2880
5	184 A	421 E	355 D	284 C	366 B	1610
C_j	1450	1823	1890	2035	2606	9804
$\sum_i Y_{ij(h)}^2$	503546	772267	781786	924365	1414834	4396798

Treatments (Conditions)

Condition	A	B	C	D	E	Total
T_h	1611	1592	2046	2131	2424	9804
Mean	322.2	318.4	409.2	426.2	484.8	--

$$\text{Total } SS\ (S_{yy}) = \sum_i \sum_j Y_{ij(h)}^2 - \frac{G^2}{k^2} = 4396798 - \frac{(9804)^2}{25}$$

$$= 4396798 - 3844736.64 = 552061.36,$$

$$\text{Periods } SS\ (C_{yy}) = \sum_j \frac{C_j^2}{k} - \frac{G^2}{k^2} = \frac{(1450)^2 + \ldots + (2606)^2}{5} - \frac{(9804)^2}{25}$$

$$= 3986078 - 3844736.64 = 141341.36,$$

$$\text{Monkeys } SS\ (R_{yy}) = \sum_i \frac{R_i^2}{k} - \frac{G^2}{k^2} = \frac{(1980)^2 + \ldots + (1610)^2}{5} - \frac{(9804)^2}{25}$$

$$= 4107510 - 3844736.64 = 262773.36,$$

$$\text{Conditions } SS\ (T_{yy}) = \sum_h \frac{T_h^2}{k} - \frac{G^2}{k^2} = \frac{(1611)^2 + \ldots + (2424)^2}{5} - \frac{(9804)^2}{25}$$

$$= 3946567.6 - 3844736.64 = 101830.96$$

and $$\text{Error } SS\ (E_{yy}) = S_{yy} - (C_{yy} + R_{yy} + T_{yy})$$

$$= 552061.36 - 505945.68 = 46115.68$$

Thus the analysis of variance table is

Source of Variation	*d.f.*	Sum of Squares	Mean Square	*F*
Periods	4	141341.36	35335.34	9.19
Monkeys	4	262773.36	65693.34	17.09
Conditions	4	101830.96	25457.74	6.62
Error	12	46115.68	3842.97	--
Total	24	552061.36	--	--

(v) The critical region is $F > F_{.05,(4,12)} = 3.26$.

(vi) Since all the computed values of F fall in the critical region, so we reject our null hypotheses. There is sufficient evidence to indicate that all the responses under different periods, of different monkeys and under different conditions (treatments) are significantly different.

We are further required to test the hypothesis

$$H_0 : T_E - T_D = 0 \text{ against } H_1 : T_E - T_D > 0,$$

where T_E denotes the condition E and T_D, the condition D.

The component of the sum of squares corresponding to the *comparison*, $C = T_E - T_D$ is

$$SSC = \frac{(T_E - T_D)^2}{k\,[(1)^2 + (-1)^2]} = \frac{(2424 - 2131)^2}{2 \times 5} = \frac{(293)^2}{10} = 8584.9$$

Then the variance ratio $F = \dfrac{MS \text{ for comparison}}{MS \text{ for error}}$

$$= \frac{8584.9}{3842.97} = 2.23.$$

The computed value of $F = 2.23$ does not exceed the corresponding critical value of $F_{0.05(1,12)} = 4.75$, we therefore do not reject the hypothesis $H_0 : T_E - T_D = 0$. Hence the data fail to support the conjecture that condition E would produce a greater response on the average than condition D.

Q.23.29. (a) Completely Randomized Design.

We wish to test the yielding abilities of 6 varieties of wheat and an area of land sufficient for 36 plots is available for experimentation. We would place each variety on 6 plots *completely at random*. The total degrees of freedom for the design would be sub-divided as below:

Source of Variation	*d.f.*	*F*-ratio
Between Varieties	5	$\rightarrow F_1$
Within Varieties (Error)	30	
Total	35	--

To test the hypothesis of no difference in the yielding abilities of the varieties, we would compute the statistic

$$F_1 = \frac{\text{Mean Square for Between Varieties}}{\text{Mean Square for Within Varieties}}$$

and would compare with the critical tabulated value of F for 5, 30 *d.f.* at the stated level of significance. The hypothesis would be reject when F_1 exceeds $F_{\alpha(5,30)}$.

(b) Randomized Complete Block Design

We have 6 varieties and an area of land sufficient for 36 plots is available for experimentation. We would group the plots into 6 blocks, each block containing 6 plots. The direction of blocks would depend on the fertility trend and they would be made in such a manner that one block would contain the most fertile group of plots, the second block would contain the next most fertile plot and so on. We would then assign at random the

six varieties to plots within each block, making a new randomization in each block.

The total degrees of freedom for the design would be partitioned as follows:

Source of Variation	*d.f.*	*F*-ratio
Between Blocks	5	
Between Varieties	5	$\rightarrow F_2$
Error	25	
Total	35	--

To test the hypothesis H_0 : There is no difference in the yielding abilities of 6 varieties of wheat, the *F*-statistic would be computed by the relation

$$F_2 = \frac{\text{Mean Square for Varieties}}{\text{Mean Square for Errors}}$$

We would reject the hypothesis H_0 if F_2 exceeds the critical value of *F* for 5, 25 *d.f.* at the stated level of significance.

(c) Latin Square Design

We have 6 varieties of wheat and 6 plots are available for experimentation. We would set up 6 blocks of 6 plots each, in two mutually perpendicular directions, known as *rows* and *columns*. The varieties would be applied at random with the restriction that the varieties appear once and only once in each direction. The total degrees of freedom for this design would be subdivided as below:

Source of Variation	*d.f.*	*F*-ratio
Rows	5	
Columns	5	
Varieties	5	$\rightarrow F_3$
Error	20	
Total	35	--

To test the hypothesis of no difference in the yielding abilities of the 6 varieties of wheat, the *F*-statistic would be

$$F_3 = \frac{\text{Mean Square for Varieties}}{\text{Mean Square for Error}}$$

Q.23.31. (a) Let y represent the missing value, and let R, C and T be the totals of the row, column and treatment, which contain the missing value. If G represents the grand total of the r^2-1 actual values in a $r \times r$ Latin Square, then the error sum of squares (SSE) in terms of the unknowns is obtained as

$$SSE = y^2 - \frac{(R+y)^2}{r} - \frac{(C+y)^2}{r} - \frac{(T+y)^2}{r} + \frac{2(G+y)^2}{r^2}$$

+ terms not involving y.

To minimize the value of SSE, we equate $\frac{\partial (SSE)}{\partial y}$ to zero, getting the equation

$$y - \frac{(R+y)}{r} - \frac{(C+y)}{r} - \frac{(T+y)}{r} + \frac{2(G+y)}{r^2} = 0$$

or $$y(r^2-3r+2) = r(R + C + T) - 2G$$

$$\therefore \quad y = \frac{r(R + C + T) - 2G}{(r-1)(r-2)}$$

(b) The data with missing observation? is given below:

A 7.4	D ?	E 5.8	B 12.0	C 14.3	39.5=R
C 11.8	B 6.5	A 8.7	E 7.6	D 7.9	42.5
D 10.1	C 17.9	B 9.0	A 8.5	E 7.1	52.6
E 8.8	A 10.1	C 15.7	D 11.1	B 7.4	53.1
B 11.8	E 8.8	D 14.3	C 18.4	A 10.1	63.4
49.9	43.3=C	53.5	57.6	46.8	251.1=G

$$T = 7.9 + 10.1 + 11.1 + 14.3 = 43.4$$

Substituting these values in the formula

$y = \frac{r(R+C+T)-2G}{(r-1)(r-2)}$, we get the missing observation (y) as

$$y = \frac{5(39.6 + 43.3 + 43.4) - 2(251.1)}{(5-1)(5-2)}$$

$$= \frac{5(126.2) - 502.2}{4 \times 3} = \frac{128.8}{12} = 10.7$$

Q.23.32. (a) Two Latin Squares are said to be *orthogonal* if each letter of one square occurs exactly once with every letter of other square when they are superimposed. A set of Latin squares such that any two of them are orthogonal, is known as an orthogonal set. A set containing precisely $k-1$ orthogonal Latin Squares, where k is the size of the square, is called a *complete set*. The complete set of orthogonal Latin squares of side 4 contains three orthogonal Latin squares.

The following two Latin squares (one with Latin letters and the other with Greek letters).

A	B	C	D
B	A	D	C
C	D	A	B
D	C	B	A

and

α	β	γ	δ
δ	γ	β	α
β	α	δ	γ
γ	δ	α	β

are orthogonal. A third Latin square (with numerals) orthogonal to these two, is

1	2	3	4
3	4	1	2
4	3	2	1
2	1	4	3

which completes the set.

Q.23.33 Construction of a 5×5 Graeco-Latin Square deisgn

Let us first construct a 5×5 Latin Square with Latin letters by rotation as follows:

	5×5 LS with Latin letters				
Slip 1 →	A	B	C	D	E
	B	C	D	E	A
	C	D	E	A	B
	D	E	A	B	C
	E	A	B	C	D

To construct a Graeco-Latin square, a second *LS* with Greek letters is to be superimposed on the Latin letters such that each Greek letter occurs once in each row, once in each column and once with each Latin letter. These requirements are satisfied by the following *LS* with Greek letters, obtained by slipping 2 steps.

	5×5 LS with Greek letters				
Slip 2 →	α	β	γ	δ	ε
	γ	δ	ε	α	β
	ε	α	β	γ	δ
	β	γ	δ	ε	α
	δ	ε	α	β	γ

Combining these two, we get the following 5×5 Graeco-Latin square:

Aα	Bβ	Cγ	Dδ	Eε
Bγ	Cδ	Dε	Eα	Aβ
Cε	Dα	Eβ	Aγ	Bδ
Dβ	Eγ	Aδ	Bε	Cα
Eδ	Aε	Bα	Cβ	Dγ

The analysis of variance appropriate to this design is given below:

Source of Variation	*d.f.*	Sum of Squares	Mean Square	Expected Mean Square
Rows	4	$\sum_i \frac{R_i^2}{5} - \frac{G^2}{25} = R_{yy}$	$s_r^2 = \frac{R_{yy}}{4}$	$\sigma^2 + \frac{5}{4}\sum_i \rho_i^2$
Columns	4	$\sum_j \frac{C_j^2}{5} - \frac{G^2}{25} = C_{yy}$	$s_c^2 = \frac{C_{yy}}{4}$	$\sigma^2 + \frac{5}{4}\sum_j \gamma_j^2$
Greek letters	4	$\sum_k \frac{Q_k^2}{5} - \frac{G^2}{25} = Q_{yy}$	$s_q^2 = \frac{Q_{yy}}{4}$	$\sigma^2 + \frac{5}{4}\sum_k \delta_k^2$
Latin letters (Treatments)	4	$\sum_h \frac{T_h^2}{5} - \frac{G^2}{25} = T_{yy}$	$s_t^2 = \frac{T_{yy}}{4}$	$\sigma^2 + \frac{5}{4}\sum_h \tau_h^2$
Error	8	By subtraction = E_{yy}	$s_e^2 = \frac{R_{yy}}{4}$	σ^2
Total	24	$\sum_i \sum_j Y_{ij}^2 - \frac{G^2}{25} = S_{yy}$	--	--

where R_i, C_j, Q_k, and T_h are the totals of the ith row, jth column, kth Greek letter and hth Latin letter (Treatment), and G stands for the grand total. The assumption necessary for the analaysis of the 5×5 Graeco-Latin square design given above is that the observations may be represented by the linear statistical model

$$Y_{ij(kh)} = \mu + \rho_i + \gamma_j + \delta_k + \tau_h + \varepsilon_{ij(kh)}, \quad i,j,k,h = 1,2,..,5.$$

where $\sum_i \rho_i = \sum_j \gamma_j = \sum_k \delta_k = \sum_h \tau_h = 0,$

and μ, ρ_i, γ_j, δ_k, τ_h and $\varepsilon_{ij(kh)}$ denote the general mean, the effect of the ith row, the effect of the jth column, the effect of the kth Greek letter, the effect of the hth Latin letter and experimental error respectively, and where $\varepsilon_{ij(kh)}$ are independently and normally distributed with zero mean and common variance σ^2.

A Graeco-Latin square design is used when we desire to control three sources of variation, one source of variation being controlled by rows, the other source by columns and the third source by Greek letters. This design is least expensive because of savings in numbers of observations. The *G-LS* design has an occassional application as it is unsatisfactory from the point of view of randomization test and it has manifold requirements.

Q.23.34. The recorded data are:

Rows	Columns				R_i
	1	2	3	4	
1	$C\beta$ (16)	$B\gamma$ (12)	$D\delta$ (17)	$A\alpha$ (11)	56
2	$B\alpha$ (15)	$C\delta$ (14)	$A\gamma$ (15)	$D\beta$ (14)	58
3	$A\delta$ (12)	$D\alpha$ (6)	$B\beta$ (14)	$C\gamma$ (13)	45
4	$D\gamma$ (9)	$A\beta$ (9)	$C\alpha$ (8)	$B\delta$ (9)	35
C_j	52	41	54	47	194

Now, Total $SS = \sum_i \sum_j Y_{ij}^2 - \frac{G^2}{k^2} = (16)^2 + (12)^2 + \ldots + (9)^2 - \frac{(194)^2}{16}$

$= 2504 - 2352.25 = 151.75,$

$$\text{Columns } SS = \sum_j \frac{C_i^2}{k} - \frac{G^2}{k^2}$$

$$= \frac{(52)^2+(41)^2+(54)^2+(47)^2}{4} - \frac{(194)^2}{16}$$

$$= 2377.5 - 2352.25 = 25.25,$$

$$\text{Rows } SS = \sum_i \frac{R_i^2}{k} - \frac{G^2}{k^2}$$

$$= \frac{(56)^2+(58)^2+(45)^2+(35)^2}{4} - \frac{(194)^2}{16}$$

$$= 2437.5 - 2352.25 = 85.25.$$

Summary for acryloid concentrations (Latin letters) and acetone concentrations (Greek letters). (Total for a Latin letter or a Greek letter is obtained by adding all the observed values of the letter concerned in the square).

Latin Letters	A	B	C	D	Total
Total (T_h)	47	50	51	46	194
Greek Letters	α	β	γ	δ	--
Total (Q_l)	40	53	49	52	194

$$\text{Latin Letters } SS = \sum_h \frac{T_h^2}{k} - \frac{G^2}{k^2}$$

$$= \frac{(47)^2+(50)^2+(51)^2+(46)^2}{4} - \frac{(194)^2}{16}$$

$$= 2356.5 - 2352.25 = 4.25,$$

$$\text{Greek Letters } SS = \sum_l \frac{Q_l^2}{k} - \frac{G^2}{k^2}$$

$$= \frac{(40)^2+(53)^2+(49)^2+(52)^2}{4} - \frac{(194)^2}{16}$$

$$= 2378.5 - 2352.25 = 26.25.$$

Error SS is obtained by difference.

Hence the analysis of variance table is

Source of Variation	d.f.	Sum of Squares	Mean Square	Computed F
Columns	3	25.25	8.42	2.35
Rows	3	85.25	28.42	7.94
Latin Letters	3	4.25	1.42	0.40
Greek Letters	3	26.25	8.75	2.44
Error	3	10.75	3.58	--
Total	15	151.75	--	--

The critical region is $F \geq F_{0.05(3,3)} = 9.28$.

None of the treatments differs significantly.

Q.23.37. Computations proceed as follows:

	Treatments				$T_{..}$
	a_0b_0	a_0b_1	a_1b_0	a_1b_1	
	6	12	26	21	
	14	14	17	16	
	8	13	21	20	
	9	11	30	17	
	7	13	27	21	
$T_{.j}$	44	63	121	95	323
$T_{.j}^2$	1936	3969	14641	9025	29571
$\sum_i Y_{ij}^2$	426	799	3035	1827	6087

$$\text{Total } SS = \sum_i \sum_j Y_{ij}^2 - \frac{T_{..}^2}{n} = 6087 - \frac{(323)^2}{20}$$

$$= 6087 - 5216.45 = 870.55,$$

$$\text{Treatment } SS = \sum_j \frac{T_{.j}^2}{r} - \frac{T_{..}^2}{n} = \frac{29571}{5} - 5216.45$$

$$= 5914.2 - 5216.45 = 697.75, \text{ and}$$

$$\text{Error } SS = \text{Total } SS - \text{Treatment } SS = 870.55 - 697.75 = 172.80$$

To partition the treatment sum of squares into *main effects* and *interaction components*, we first find the effect-totals as

$$[A] = -[a_0b_0] + [a_1b_0] - [a_0b_1] + [a_1b_1]$$

$$= -44 + 121 - 63 + 95 = 109,$$

$$[B] = -[a_0b_0] - [a_1b_0] + [a_0b_1] + [a_1b_1]$$

$$= -44 - 121 + 63 + 95 = -7, \text{ and}$$

$$[AB] = [a_0b_0] - [a_1b_0] - [a_0b_1] + [a_1b_1]$$

$$= 44 - 63 - 121 + 95 = -45.$$

$$\text{Then } SS \text{ for } A = \frac{[A]^2}{4r} = \frac{(109)^2}{20} = 594.05,$$

$$SS \text{ for } B = \frac{[B]^2}{4r} = \frac{(-7)^2}{20} = 2.45, \text{ and}$$

$$SS \text{ for } AB = \frac{[AB]^2}{4r} = \frac{(-45)^2}{20} = 101.25.$$

These results are presented in the following *ANOVA*-Table:

Source of Variation	d.f.	Sum of Squares	Mean Square	F
Treatments	3	697.75	232.58	21.54
A	1	594.05	594.05	55.00
B	1	2.45	2.45	0.23
AB	1	101.25	101.25	9.38
Error	16	172.80	10.80	--
Total	19	870.55	--	--

To test the hypothesis of no difference amog the treatment combinations, the computed F is significantly larger than the corresponding 5-per cent critical $F=3.24$. Hence we reject it and proceed to test the other hypotheses which would be stated as

(a) $H_0 : A_i = 0,\ i = 0, 1$ and H_1 : Not all A_i are zero.

The computed value of $F=55.00$ is significantly larger than the corresponding 5-per cent critical $F=4.49$. We reject H_0 and say that there are significant differences between the two levels of factor A.

(b) $H_0 : B_j = 0,\ j = 0, 1$ and H_1 : Not all B_j are zero.

Here the computed value of F is less than the corresponding 5-per cent critical F. We do not reject H_0 and conclude that the differences between the two levels of factor B are not significant.

(c) $H_0 : (AB)_{ij} = 0,\ i = 0, 1,\ j = 0, 1$ and H_1 : Not all $(AB)_{ij}$ are zero.

The computed value of $F=9.38$ is larger than the corresponding 5-percent critical $F=4.49$. We therefore reject H_0 and conclude that the factor A and B do interact; they are not independent of one another.

Q.23.38. Computations proceed as follows:

Rows	Columns				R_i	R_i^2
	1	2	3	4		
1	p 19	np 31	(1) 15	n 18	83	6889
2	np 27	(1) 10	n 21	p 21	79	6241
3	n 23	p 17	np 30	(1) 13	83	6889
4	(1) 12	n 20	p 15	np 26	73	5329
C_j	81	78	81	78	318	25348.
C_j^2	6561	6084	6561	6084	25290	--
$\sum_i Y_{ij}^2$	1763	1750	1791	1610	6914	--

Treatment	(1)	n	p	np	Total
T_h	50	82	72	114	318
T_h^2	2500	6724	5184	12996	27404

Total SS (S_{yy}) $= \sum_i \sum_j Y_{ij(h)}^2 - \frac{G^2}{k^2} = 6914 - \frac{(318)^2}{16}$

$= 6914 - 6320.25 = 593.75,$

Columns SS (C_{yy}) $= \sum_j \frac{C_j^2}{k} - \frac{G^2}{k^2} = \frac{25290}{4} - \frac{(318)^2}{16}$

$= 6322.5 - 6320.25 = 2.25,$

Rows SS (R_{yy}) $= \sum_i \frac{R_i^2}{k} - \frac{G^2}{k^2} = \frac{25348}{4} - \frac{(318)^2}{16}$

$= 6337 - 6320.25 = 16.75,$

Treatments SS (T_{yy}) $= \sum_h \frac{T_h^2}{k} - \frac{G^2}{k^2} = \frac{27404}{4} - \frac{(318)^2}{16}$

$= 6851 - 6320.25 = 530.75,$ and

Error SS (E_{yy}) $= S_{yy} - (C_{yy} + R_{yy} + T_{yy})$

$= 593.75 - (2.25 + 16.75 + 530.75) = 44.00$

To compute the sums of squares for Main Effects and Interaction, we first find the effect-totals as

$[N] = -(1) + n - p + np = -50 + 82 - 72 + 114 = 74,$

$[P] = -(1) - n + p + np = -50 - 82 + 72 + 114 = 54,$ and

$[NP] = (1) - n - p + np = 50 - 82 - 72 + 114 = 10.$

Then SS for $N = \frac{[N]^2}{4k} = \frac{(74)^2}{16} = 342.25,$

SS for $P = \frac{[P]^2}{4k} = \frac{(54)^2}{16} = 182.25,$

$$SS \text{ for } NP = \frac{[NP]^2}{4k} = \frac{(10)^2}{16} = 6.25.$$

These results are presented in the following *ANOVA*-Table:

Source of Variation	*d.f.*	Sum of Squares	Mean Square	*F*	5-per cent critical *F*
Columns	3	2.25	0.75	--	--
Rows	3	16.75	5.58	--	--
Treatments	3	530.75	176.92	24.14	4.76
N	1	342.25	342.25	46.69	5.99
P	1	182.25	182.25	24.86	5.99
NP	1	6.25	6.25	0.85	5.99
Error	6	44.00	7.33	--	--
Total	15	593.75	--	--	--

To test the hypothesis H_0 : there is no difference among treatment combinations, we obtained $F=24.14$, which is significantly larger than the corresponding 5-per cent critical $F=4.76$ for (3,6) *d.f.* Hence we reject the hypothesis of no differences among the treatment combinations.

It is also obvious from the *ANOVA*-table that

(i) there are significant differences between the two levels of the factor *N*,

(ii) there are significant differneces between the two levels of the factor *P*, and

(iii) the factors *N* and *P* do not interact, they are independent of each other.

Q.23.39. This is a 2^3-factorial experiment, involving three factors a, b and c, each at two levels in all combinations. The experimental design used is the Randomized Complete Block design.

In additon to the usual analysis of variance for a RCB design, we have to partition the treatment sum of squares into *main effects* and *interaction* components. The levels of the main

effects and interactions are determined from the following combination of treatments.

Effect	Combination of Treatments							
	(1)	*a*	*b*	*ab*	*c*	*ac*	*bc*	*abc*
Total	+	+	+	+	+	+	+	+
A	–	+	–	+	–	+	–	+
B	–	–	+	+	–	–	+	+
AB	+	–	–	+	+	–	–	+
C	–	–	–	–	+	+	+	+
AC	+	–	+	–	–	+	–	+
BC	+	+	–	–	–	–	+	+
ABC	–	+	+	–	+	–	–	+

where the totals of each column are added or subtracted as indicated. The sum of squares for any effect in the analysis of variance is $\frac{(\text{effect})^2}{8r}$, where r denotes the number of replications.

Computations.

Replication	Treatment Combination								T_i
	(1)	*a*	*b*	*ab*	*c*	*ac*	*bc*	*abc*	
1	13	10	12	14	9	11	8	7	84
2	14	11	13	15	10	12	9	8	92
$T_{.j}$	27	21	25	29	19	23	17	15	176
$T_{.j}^2$	729	441	625	841	361	529	289	225	4040
$\sum_i Y_{ij}^2$	365	221	313	421	181	265	145	113	2024

$$\text{Total } SS = \sum_i \sum_j Y_{ij}^2 - \frac{T_{..}^2}{n} = 2024 - \frac{(176)^2}{16}$$

$$= 2024 - 1936 = 88,$$

$$\text{Treatment } SS = \sum_j \frac{T_j^2}{r} - \frac{T_{..}^2}{n} = \frac{4040}{2} - \frac{(176)^2}{16}$$

$$= 2020 - 1936 = 84,$$

$$\text{Block } SS = \sum_i \frac{T_{i.}^2}{c} - \frac{T_{..}^2}{n} = \frac{(84)^2 + (92)^2}{8} - \frac{(176)^2}{16}$$

$$= 1940 - 1936 = 4, \text{ and}$$

$$\text{Error } SS = \text{Total } SS - (\text{Treatment } SS + \text{Block } SS)$$

$$= 88 - (84 + 4) = 0$$

To compute the sums of squares for Main Effects and Interactions, we first determine the effect-totals as

$$[A] = -(1) + a - b + ab - c + ac - bc + abc$$

$$= -27 + 21 - 25 + 29 - 19 + 23 - 17 + 15 = 0,$$

$$[B] = -(1) - a + b + ab - c - ac + bc + abc$$

$$= -27 - 21 + 25 + 29 - 19 - 23 + 17 + 15 = -4,$$

$$[AB] = (1) - a - b + ab + c - ac - bc + abc$$

$$= 27 - 21 - 25 + 29 + 19 - 23 - 17 + 15 = 4.$$

Similarly $[C] = -28$, $[AC] = 4$, $[BC] = -16$ and $[ABC] = -16$

Now SS for $A = \frac{[A]^2}{8r} = \frac{(0)^2}{16} = 0,$

SS for $B = \frac{[B]^2}{8r} = \frac{(-4)^2}{16} = 1,$

SS for $AB = \frac{[AB]^2}{8r} = \frac{(0)^2}{16} = 1.$

SS for $C = 49$, SS for $AC = 1$, SS for $BC = 16$ and SS for $ABC = 16$.

In this design, the error sum of squares turns out to be zero, we may consider the Interaction mean squares of little importance, so that their pooled sum may be used to provide an estimate of error mean square.

Q.23.40. Computations for Analysis of Variance.

Replication	Treatments								$T_{i.}$
	(1)	n	p	k	np	pk	nk	npk	
1	24	25	24	29	24	22	30	32	210
2	30	31	31	39	27	25	34	23	240
3	23	33	27	31	23	24	29	37	227
4	28	26	24	36	27	29	36	34	240
$T_{.j}$	105	115	106	135	101	100	129	126	917
T_j^2	11025	13225	11236	18225	10201	10000	16641	15876	106429
$\sum_i Y_{ij}^2$	2789	3351	2842	4619	2563	2526	4193	4078	26961

$$\text{Total } SS = \sum_i \sum_j Y_{ij}^2 - \frac{T_{..}^2}{n} = 26961 - \frac{(917)^2}{32}$$

$$= 26961 - 26277.78 = 683.22,$$

$$\text{Replication } SS = \sum_i \frac{T_{i.}^2}{c} - \frac{T_{..}^2}{n} = \frac{(210)^2 + \ldots + (240)^2}{8} - \frac{(917)^2}{32}$$

$$= 26353.62 - 26277.78 = 75.84,$$

$$\text{Treatment } SS = \sum_j \frac{T_j^2}{r} - \frac{T_{..}^2}{n} = \frac{106429}{4} - \frac{(917)^2}{32}$$

$$= 26607.25 - 26277.78 = 329.47, \text{ and}$$

$$\text{Error } SS = \text{Total } SS - (\text{Replication } SS + \text{Treatment } SS)$$

$$= 683.22 - (75.84 + 329.47) = 277.91$$

To compute the sums of squares for Main Effects and Interactions, we first determine the effect-totals as

$$[N] = -(1) + n - p + np - k + nk - pk + npk$$

$$= -105 + 115 - 106 + 101 - 135 + 129 - 100 + 126 = 25$$

$[P] = -(1) - n + p + np - k - nk + pk + npk$

$= -105 - 115 + 106 + 101 - 135 + 129 + 100 + 126 = -51,$

$[K] = -(1) - n - p - np + k + nk + pk + npk$

$= -105 - 115 - 106 - 101 + 135 + 129 + 100 + 126 = 63.$

$[NP] = (1) - n - p + np + k - nk - pk + npk$

$= 105 - 115 - 106 + 101 + 135 - 129 - 100 + 126 = 17,$

$[NK] = (1) - n + p - np - k + nk - pk + npk$

$= 105 - 115 + 106 - 101 - 135 + 129 - 100 + 126 = 15,$

$[PK] = (1) + n - p - np - k - nk + pk + npk$

$= 105 + 115 - 106 - 101 - 135 - 129 + 100 + 126 = -25,$ and

$[NPK] = -(1) + n + p - np + k - nk - pk + npk$

$= -105 + 115 + 106 - 101 + 135 - 129 - 100 + 126 = 47.$

Then SS for $N = \frac{[N]^2}{8r} = \frac{(25)^2}{32} = 19.53,$

SS for $P = \frac{[P]^2}{8r} = \frac{(-51)^2}{32} = 81.28,$

SS for $K = \frac{[K]^2}{8r} = \frac{(63)^2}{32} = 124.03,$

SS for $NP = \frac{[NP]^2}{8r} = \frac{(17)^2}{32} = 9.03,$

SS for $NK = \frac{[NK]^2}{8r} = \frac{(15)^2}{32} = 7.03,$

SS for $PK = \frac{[PK]^2}{8r} = \frac{(-25)^2}{32} = 19.53,$ and

SS for $NPK = \frac{[NPK]^2}{8r} = \frac{(47)^2}{32} = 69.03.$

These results are presented in the following ANOVA-Table:

Source of Variation	*d.f.*	Sum of Squares	Mean Square	*F*
Replication	3	75.84	25.28	--
Treatments	7	329.47	47.07	3.56
N	1	19.53	19.53	1.48
P	1	81.28	81.28	6.14
K	1	124.03	124.03	9.37
NP	1	9.03	9.03	0.68
NK	1	7.03	7.03	0.53
PK	1	19.53	19.53	1.48
NPK	1	69.03	69.03	5.22
Error	21	277.91	13.23	--
Total	31	683.22	--	--

The 5-per cent F for (7,21) *d.f.* from tables = 2.49, and the 5-per cent F for (1,21) *d.f.* from tables = 4.32.

Comparing the computed values of F with the correspnding critical values of F, we find that there is a significant difference in the treatments; the main effects P and K also differ significantly, while the main effect N is insignificant. All the first order Interactions are not significant.

Q.23.41. (i) Computation of the sums of squares for all factorial effects by Contrast Method.

Effect	Treatments Combination								Total
	(1)	*a*	*b*	*ab*	*c*	*ac*	*bc*	*abc*	
Total	+41	+51	+57	+67	+63	+54	+76	+73	482
A	-41	+51	-57	+67	-63	+54	-76	+73	8
B	-41	-51	+57	+67	-63	-54	+76	+73	64
AB	+41	-51	-57	+67	+63	-54	-76	+73	6
C	-41	-51	-57	-67	+63	+54	+76	+73	50
AC	+41	-51	+57	-67	-63	+54	-76	+73	-32
BC	+41	+51	-57	-67	-63	-54	+76	+73	0
ABC	-41	+51	+57	-67	+63	-54	-76	+73	6

Now $SSA = \frac{[A]^2}{8r} = \frac{(8)^2}{24} = 2.6667; \quad (\because r=3)$

$SSB = \frac{[B]^2}{8r} = \frac{(64)^2}{24} = 170.6667;$

$SSC = \frac{[C]^2}{8r} = \frac{(50)^2}{24} = 104.1667;$

$SS(AB) = \frac{[AB]^2}{8r} = \frac{[6]^2}{24} = 1.5000;$

$SS(AC) = \frac{[AC]^2}{8r} = \frac{[-32]^2}{24} = 42.6667;$

$SS(BC) = \frac{[BC]^2}{8r} \quad \frac{[0]^2}{24} = 0;$

$SS(ABC) = \frac{[ABC]^2}{8r} = \frac{[6]^2}{24} = 1.5000.$

(ii) Computation of the sums of squares for all factorial effects by *Yates' Method*.

The totals for all effects are obtained by Yates' method as follows:

Treatment Combination	(1)	(2)	(3)	Effect
(1) 41	41+51=92	92+124=216	216+266=482	Total
a 51	57+67=124	117+149=266	20+(−12)=8	A
b 57	63+54=117	10+10=20	32+32=64	B
ab 67	76+73=149	−9+(−3)=−12	0+6=6	AB
c 63	51−41=10	124−92=32	266−216=50	C
ac 54	67−57=10	149−117=32	−12−20=−32	AC
bc 76	54−63=−9	10−10=0	32−32=0	BC
abc 73	73−76=−3	−3−(−9)=6	6−0=6	ABC

The sums of squares have been computed in (i) above.

Q.23.42. (b) Computations for analysis of covariance in a Latin Square design.

Rows	COLUMNS 1			2			3			4			R_i	R'_i
	tr	Y	X	tr.	Y	X	tr.	Y	X	tr.	Y	X		
1	A	88	46	C	71	44	D	80	45	B	75	44	314	179
2	B	74	45	A	100	47	C	84	48	D	74	46	332	186
3	C	73	47	D	84	46	B	75	47	A	75	47	307	187
4	D	86	46	B	73	45	A	86	48	C	85	46	330	185
C_j	321		184	328		182	325		188	309		183	1283	737
C'_j														

Total SS and Products.

$$S_{yy} = \sum_i \sum_j Y^2_{ij(h)} - \frac{T^2_{..}}{k^2} = (88)^2+(74)^2+\ldots+(85)^2 - \frac{(1283)^2}{16}$$

$$= 103799 - 102880.56 = 918.44$$

$$S_{xx} = \sum_i \sum_j X^2_{ij(h)} - \frac{(T'_{..})^2}{k^2} = (46)^2+(45)^2+\ldots+(46)^2 - \frac{(737)^2}{16}$$

$$= 33971 - 33948.06 = 22.94, \text{ and}$$

$$S_{xy} = \sum_i \sum_j Y_{ij}X_{ij} - \frac{(T_{..})\,(T'_{..})}{k^2}$$

$$= (88\times46)+(74\times45)+\ldots+(85\times46) - \frac{(1283)\,(737)}{16}$$

$$= 59162 - 59098.19 = 63.81.$$

Columns SS and Products.

$$C_{yy} = \sum_j \frac{C^2_{.j}}{k} - \frac{(T_{..})^2}{k^2}$$

$$= \frac{(321)^2+(328)^2+(325)^2+(309)^2}{4} - \frac{(1283)^2}{16}$$

$$= 102932.75 - 102880.56 = 52.19$$

$$C_{xx} = \sum_j \frac{(C'_j)^2}{k} - \frac{(T'..)^2}{k^2}$$

$$= \frac{(184)^2+(182)^2+(188)^2+(183)^2}{4} - \frac{(737)^2}{16}$$

$$= 33953.25 - 33948.06 = 5.19, \text{ and}$$

$$C_{xy} = \sum_j \frac{(C_{.j})\,(C'_j)}{k} - \frac{(T..)\,(T'..)}{k^2}$$

$$= \frac{(321\times184)+\ldots+(309\times183)}{4} - \frac{(1283)(737)}{16}$$

$$= 59101.75 - 59098.19 = 3.56.$$

Rows SS and Products.

$$R_{yy} = \sum_i \frac{R_{i.}^2}{k} - \frac{T_{..}^2}{k^2}$$

$$= \frac{(314)^2+(332)^2+(307)^2+(330)^2}{4} - \frac{(1283)^2}{16}$$

$$= 102992.25 - 102880.56 = 111.69,$$

$$R_{xx} = \sum_i \frac{(R'_{i.})^2}{k} - \frac{(T'..)^2}{k^2}$$

$$= \frac{(179)^2+(186)^2+(187)^2+(185)^2}{4} - \frac{(737)^2}{16}$$

$$= 33957.75 - 33948.06 = 9.69, \text{ and}$$

$$R_{xy} = \sum_i \frac{(R_{i.})\,(R'_{i.})}{k} - \frac{(T..)\,(T'..)}{k^2}$$

$$= \frac{(314\times179)+\ldots+(330\times185)}{4} - \frac{(1283)(737)}{16}$$

$$= 59104.25 - 59098.19 = 6.06.$$

To compute the sum of squares and products for treatments, we construct the following table:

Treatment	$T_h(Y)$	T'_h	T_h^2	$(T'_h)^2$	$(T_h)\,(T'_h)$
A	349	188	121801	35344	65612
B	297	181	88209	32761	53757
C	313	185	97969	34225	57905
D	324	183	104976	33489	59292
Total	1283	737	412955	135819	236566

$$T_{yy} = \sum_h \frac{T_h^2}{k} - \frac{T_{..}^2}{k^2} = \frac{412955}{4} - \frac{(1283)^2}{16}$$

$$= 103238.75 - 102880.56 = 358.19,$$

$$T_{xx} = \sum_h \frac{(T'_h)^2}{k} - \frac{(T'..)^2}{k^2} = \frac{135819}{4} - \frac{(737)^2}{16}$$

$$= 33954.75 - 33948.06 = 6.69, \text{ and}$$

$$T_{xy} = \sum_h \frac{(T_h)(T'_h)}{k} - \frac{(T..)(T'..)}{k^2}$$

$$= \frac{236566}{4} - \frac{(1283)\,(737)}{16} = 59141.50 - 59098.19 = 43.31$$

Error SS and Products are obtained by subtraction. Thus

$$E_{yy} = S_{yy} - (C_{yy} + R_{yy} + T_{yy})$$

$$= 918.44 - (52.19 + 111.69 + 358.19) = 396.37,$$

$$E_{xx} = S_{xx} - (C_{xx} + R_{xx} + T_{xx})$$

$$= 22.94 - (5.19 + 9.69 + 6.69) = 1.37, \text{ and}$$

$$E_{xy} = S_{xy} - (C_{xy} + R_{xy} + T_{xy})$$

$$= 63.81 - (3.56 + 6.06 + 43.31) = 10.88.$$

Adjusted SS for Error is

$$\text{Adjusted } \Sigma y^2 = \Sigma y^2 - \frac{(\Sigma xy)^2}{\Sigma x^2} = 396.37 - \frac{(10.88)^2}{1.37}$$

$$= 396.37 - 86.40 = 309.97, \text{ and}$$

Adjusted SS for Treatment plus Error is

$$\text{Adjusted } \Sigma y^2 = \Sigma y^2 - \frac{(\Sigma xy)^2}{\Sigma x^2} = 754.56 - \frac{(54.19)^2}{8.06}$$

$$= 754.56 - 364.34 = 309.22.$$

Hence the analysis of covariance table for these results is

Source of Variation		Sum of Squares and Products			Adjusted for X		
	d.f.	Σx^2	Σy^2	Σxy	Σy^2	*d.f.*	*MS*
Total	15	22.94	918.44	63.81	--	--	--
Columns	3	5.19	52.19	3.56	--	--	--
Rows	3	9.69	111.69	6.06	--	--	--
Treatments (T)	3	6.69	358.19	43.31	--	--	--
Error (E)	6	1.37	396.37	10.88	309.97	5	61.99
$T + E$	9	8.06	754.56	54.19	390.22	8	--
Treatments adjusted					80.25	3	26.75

For error, the regression co-efficient is,

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2} = \frac{10.88}{1.37} = 7.9416.$$

Now $\bar{X}_{..} = 46.0625$, $\bar{X}_{i.}$ and $\bar{Y}_{i.}$ denote the means of the ith treatment of X and Y. The corrected means of Y-values, which are free from the influence of regression, are computed below:

Treatment	$\bar{X}_{i.}$	$\bar{X}_{i.}-\bar{X}_{..}$	$b(\bar{X}_{i.}-\bar{X}_{..})$	$\bar{Y}_{i.}$	Corrected means $\bar{Y}_{i.}-b(\bar{X}_{i.}-\bar{X}_{..})$
A	47.00	0.9375	7.45	87.25	79.80
B	45.25	-0.8125	-6.45	74.25	80.70
C	46.25	0.1875	1.49	78.25	76.76
D	45.75	-0.3125	-2.48	81.00	83.48

❖❖❖❖❖❖❖❖❖

CHAPTER 24

NON-PARAMETRIC TESTS

Q.24.3. (b) Using the Sign Test

(i) We set up our hypotheses as

H_0 : Median = 55 and H_1 : Median < 55.

(ii) Let us set the significance level at $\alpha = 0.05$.

(iii) The test-statistic to be used is X, the number of times the less frequent sign occurs.

(iv) Computations. Subtracting 55, the hypothesized value of the median from each observation and writing down the signs, we get

$$-, -, -, -, +, +, -, -, +, -, +, +, +, +, -, -, -, +, -, +, -, -, -, -,$$

Now, $n=24$ and $X=9$, the number of plus signs (less frequent). As n, the number of pluses and minuses exceeds 10, the test statistic becomes

$$Z = \frac{(X+1/2) - n/2}{\sqrt{n/4}} = \frac{9.5 - 12}{\sqrt{24/4}} = -1.02$$

(v) The critical region is $Z < -z_{0.05} = -1.645$

(vi) **Conclusion.** Since the computed value of $z = -1.02$ does not fall in the critical region, we therefore accept H_0.

Using the Wilcoxon Signed-Rank Test.

(i) We state our hypotheses as

H_0 : Median = 55 and H_1 : Median < 55.

(ii) Let us set the significance level at $\alpha = 0.05$.

(iii) The test-statistic to be used is T, the smaller sum of ranks with the sign ignored.

(iv) Computations.

X	48, 51, 49, 53, 61, 59, 45, 52, 65, 47, 58, 57,
$X-55$	−7, −4, −6, −2, +6, +4, −10, −3, +10, −8, +3, +2
Rank of $\lvert X-55\rvert$	16, 11.5, 14, 5.5, 14, 11.5, 22, 9, 17.5, 9, 5.5
Signed (+) Rank (−)	+14, +11.5, +22, +9, +5.5 −16, −11.5, −14, −5.5, −22, −9, −17.5,

X	65, 56, 45, 49, 54, 63, 46, 57, 54, 53, 52, 45	
$X-55$	+10, +1, −10, −6, −1, +8, −9, +2, −1, −2, −3, −10	
Rank of $\lvert X-55\rvert$	22, 2, 22, 14, 2, 17.5, 19, 5.5, 2, 5.5, 9, 22	
Signed (+) Rank (−)	+22, +2, +17.5, +5.5, −22, −14, −2, −19, −2, −5.5, −9, −22	$109=T$ -191

(v) Looking in the table for the Wilcoxon Signed-Rank Test, we find that for $n=24$ at $\alpha=0.05$ for *one-tailed test*, the critical value of $T=91$; at or below which lies the critical region.

(vi) **Conclusion.** Since the computed value of $T=109$ does not fall in the critical region, we therefore accept H_0.

Q.24.4. Using the Sign Test.

(i) We state our hypotheses as

$H_0: \mu = 1.8$ hours, and $H_1: \mu \neq 1.8$ hours.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic to be used is X, the number of times the less frequent sign occurs.

(iv) Computations. Subtracting 1.8, the hypothesized value of μ from each observation and writing down the signs, we get

−, +, −, −, +, −, 0, −, +, −, − .

Now $n=10$ as zero is ignored and $X=3$, the number of plus signs (less frequent). Under H_0, X has a binomial distribution with $p=1/2$ and $n=8$.

$$\therefore \quad P(X \leq 3) = \sum_{x=0}^{3} \binom{10}{x} \left(\frac{1}{2}\right)^{10} = 0.1719$$

(v) Critical region. For a two-tailed test, to reject H_0, the computed probability should be less than 0.025.

(vi) **Conclusion.** Since the computed probability is more than $\alpha/2=0.025$, we therefore accept H_0 and conclude that the average operating time is not significantly different from 1.8 hours.

Using the Wilcoxon Signed-Rank Test.

(i) We state our hypotheses as

$H_0: \mu = 1.8$ hours, and $H_1: \mu \neq 1.8$ hours.

(ii) The significance level is set at $\alpha=0.05$.

(iii) The test-statistic to be used is T, the smaller sum of ranks with the sign ignored.

(iv) Computations.

Observation X	Difference $X-1.8$	Rank of $\lvert X-1.8\rvert$	Signed rank (+)	Signed rank (−)
1.5	−0.3	5.5		−5.5
2.2	0.4	7	7	
0.9	−0.9	10		−10
1.3	−0.5	8		−8
2.0	0.2	3	3	
1.6	−0.2	3		−3
1.8	0			
1.5	−0.3	5.5		−5.5
2.0	0.2	3	3	
1.2	−0.6	9		−9
1.7	−0.1	1		−1
			$13=T$	−42

(v) Looking in the table for the Wilcoxon signed-rank test, we find that for $n=10$ (zero is ignored) at $\alpha=0.05$ for two-tailed test, the critical value of $T=8$, at or below which lies the critical region.

(vi) **Conclusion.** Since the computed value of $T=13$ is larger than the critical value, so the null hypothesis that the average operating time is 1.8 hours, is accepted.

Q.24.5. Using the Sign Test.

(i) We state our hypotheses as

H_0 : Median = 2 and H_1 : Median $\neq 2$.

(ii) Let us set the significance level at $\alpha = 0.05$.

(iii) The test-statistic to be used is X, the number of times the less frequent sign occurs.

(iv) Computations. Subtracting 2, the hypothesized value of the median from each observation and writing down the signs, we get

$$+, +, +, +, -, +, +, -.$$

Now $n=8$ and $X=2$, the number of minus signs (less frequent). Under H_0, X has a binomial distribution with $p=1/2$ and $n=8$.

$$\therefore \quad P(X \leq 2) = \sum_{x=0}^{2} \binom{8}{x} \left(\frac{1}{2}\right)^8 = 0.1445.$$

(v) Critical region. For a two-tailed test, the computed value for rejecting H_0 should be less than 0.025.

(vi) **Conclusion.** Since the computed probability is more than $\alpha/2=0.025$, we therefore accept H_0 and conclude that the population median equals 2.

Using the Wilcoxon Signed-Rank Test.

(i) We state our hypotheses as

H_0 : Median = 2, and H_1 : Median $\neq 2$

(ii) Let us set the significance level at $\alpha=0.05$.

(iii) The test-statistic to be used is T, the smaller sum of ranks with the sign ignored.

(iv) Computations.

Observation X	Difference $X-2$	Rank of $\lvert X-2\rvert$	Signed rank (+)	Signed rank (−)
2.55	0.55	3	3	
4.62	2.62	8	8	
2.93	0.93	4	4	
2.46	0.46	2	2	
1.95	−0.05	1		−1
4.55	2.55	7	7	
3.11	1.11	6	6	
0.90	−1.10	5		−5
			30	$-6=T$

(v) Looking in the table for the Wilcoxon signed-rank test, we find that for $n=8$ at $\alpha=0.05$ for two-tailed test, the critical value of $T=3$, at or below which lies the critical region.

(vi) **Conclusion.** Since the computed value of $T=6$ is larger than the critical value, we therefore accept H_0.

Q.24.6. (b) Using the Sign Test.

(i) We state the hypotheses as

H_0 : The two populations are identical regarding the fruit-producing abilities (or the two populations have equal medians); and

H_1 : The two populations are not identical regarding the fruit-producing abilities.

(ii) The level of significance is set at $\alpha = 0.05$.

(iii) We may use the Chi-square statistic given by

$$\chi^2 = \frac{(n_1 - n_2)^2}{n_1 + n_2},$$

where n_1 and n_2 denote the number of plus and minus signs.

(iv) Computations. Subtracting the yield of variety B from that of variety A and writing down a plus sign for each positive difference, and a minus sign for each negative difference, we get

$$+, -, +, +, +, -, +, +, -, +$$

Now $n_1 = 7$, the number of positive signs, and

$n_2 = 3$, the nuber of negative signs.

$$\therefore \quad \chi^2 = \frac{(7-3)^2}{7+3} = 1.6$$

(v) The critical region is $\chi^2 > \chi^2_{0.05,(1)} = 3.84$.

(vi) **Conclusion.** Since the computed value of $\chi^2 = 1.6$ does not fall in the critical region, we therefore accept H_0.

Using the Wilcoxon Signed-Rank Test. (Matched Pairs)

(i) We state our hypotheses as

H_0: The two populations have equal medians, *i.e.*, $M_1 = M_2$; and

H_1: $M_1 \neq M_2$.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic is T, the smaller sum of ranks with the sign ignored.

(iv) Computations are shown as follows:

Location	Variety		Difference	Rank of	Signed rank	
	A.	B	$A-B$	$\lvert A-B \rvert$	(+)	(−)
1	3.03	2.28	0.75	8	8	
2	3.10	3.68	−0.58	7		−7
3	2.35	2.17	0.18	3.5	3.5	
4	3.86	3.56	0.30	5	5	
5	3.91	3.73	0.18	3.5	3.5	
6	1.72	1.85	−0.13	2		−2
7	2.65	1.48	1.17	10	10	
8	2.30	1.86	0.44	6	6	
9	2.70	2.76	−0.06	1		−1
10	3.60	2.68	0.92	9	9	
					45	$-10 = T$

(v) Looking in the table for the Wilcoxon signed-rank test, we find that for $n = 10$ at $\alpha = 0.05$ for two-tailed test, the critical value of $T = 8$, at or below which lies the critical region.

(vi) **Conclusion.** Since the computed value of $T = 10$ does not lie in the critical region, we therefore accept our null hypothesis.

Q.24.7. Applying the Sign Test.

(i) Let μ_1 and μ_2 represent the mean score of natural sciences and that of the social sciences respectively. Then our hypotheses would be stated as

$$H_0 : \mu_1 - \mu_2 = 5 \text{ and } H_1 : \mu_1 - \mu_2 > 5.$$

(ii) The significance level is set at $\alpha = 0.01$.

(iii) The test-statistic to be used is X, the number of times the less frequent sign occurs.

(iv) Computations. Subtracting 5 from the difference $(X_i - Y_i)$ and writing down a plus sign for each positive difference and a minus sign for each negative difference, we get

+, +, +, +, +, −, +, +, +, −, +, +, +, +, +, −, +, +, +, +.

Now $n=20$ and $X=3$, the number of minus signs (less frequent). Under H_0, X has a binomail distribution with $p=1/2$ and $n=20$.

$$\therefore \quad P(X\leq 3) = \sum_{x=0}^{3}\binom{20}{x}\left(\frac{1}{2}\right)^{20} = 0.0013.$$

(v) Critical region. For a one-tailed test, to reject H_0, the probability should be less than 0.01.

(vi) **Conclusion.** Since the computed probability (=0.0013) is less than $\alpha=0.01$, we therefore reject H_0 and conclude that natural sciences test scores (X) are more than five points higher than the social sciences test scores (Y).

Applying the Wilcoxon Signed-Rank Test.

(i) We state the hypotheses as

$H_0: (\mu_1-\mu_2) - 5 = 0$ and $H_1: (\mu_1-\mu_2) - 5 > 0$.

(ii) The significance level is set at $\alpha = 0.01$.

(iii) The test-statistic is T, the smaller sum of ranks with the sign ignored.

(iv) Computations. The differences $(X_i-Y_i)-5$ are given below

6, 4, 9, 5, 3, −1, 7, 5, 10, −5, 5, 5, 5, 8, 6, −8, 6, 2, 3, 5.

The respective ranks of $|(X_i-Y_i)-5|$ are

14, 5, 19, 9, 3.5, 1, 16, 9, 20, 9, 9, 9, 9, 17.5, 14, 17.5, 14, 2, 3.5, 9.

T = the smaller sum of ranks, which in this case is the sum of ranks corresponding to negative differences.

$\therefore \quad T = 1 + 9 + 17.5 = 27.5$

(v) Looking in the table for Wilcoxon signed-rank test, we find that for $n=20$ at $\alpha=0.01$ for one-tailed test, the critical value of $T = 43$, at or below which lies the critical region.

(vi) **Conclusion.** Since the observed value of $T=27.5$ lies in the critical region, we therefore reject H_0.

Q.24.9. (i) We state our hypotheses as

H_0: The two samples come from identical populations, or the medians of the two populations are identical;

H_1: The two samples do not come from identical populations.

(ii) We choose the significance level at $\alpha = 0.05$.

(iii) The test-statistic to use is R, the sum of the ranks of sample I.

(iv) Computations. Arranging the observations of combined samples in order of increasing magnitude and underlining the observations from sample I, we get

80, 80, 84, 84, 85, 87, 89, 89, 90, 92.

Assigning the ranks to these observations (underlining the ranks of the observations from sample *I*), we get

1.5, 1.5, 3.5, 3.5, 5, 6, 7.5, 7.5, 9, 10.

Adding the underlined ranks, we get $R=26$.

(v) Looking in the table of critical values of R for the Wilcoxon Rank sum test, we find for $n_1=5$ and $n_2=5$, the critical region which consists of all value ≤ 17 and all values ≥ 38 (two-tailed test, upper pair).

(vi) **Conclusion.** Since the computed value of $R=26$ does not fall in the critical region, we therefore accept H_0, and conclude that the difference between two paints is not significant.

Q.24.10. (i) We state our hypotheses as

$$H_0: M_1 = M_2 \text{ and } H_1: M_1 < M_2$$

(ii) The significance level is set at $\alpha = 0.05$.

(iii) As both n_1 and n_2 exceed 10, the test-statistic to use is

$$Z = \frac{R-\mu_R}{\sigma_R},$$

where R is the sum of ranks of smaller sample. (Sample *I*). The statistic Z, if H_0 is true, is approximately standard normal.

(iv) Computations. Arranging the observations of the combined samples in order of increasing magnitude and underlining the observations from sample 1, we get

25, 25, 26, 26, 30, 32, 33, 34, 34, 35, 35, 35, 37, 38, 40, 42, 42, 43, 43, 44, 44, 46, 46, 47, 47, 47, 48, 48, 49.

Assigning the ranks to these observations and underlining the ranks of the observations from sample 1, we get

1.5, 1.5, 3.5, 3.5, 5, 6, 7, 8.5, 8.5, 11, 11, 11, 13, 14, 15, 16.5, 16.5, 18.5, 18.5, 20.5, 20.5, 22.5, 22.5, 25, 25,25,27.5,27.5,29.

Adding the underlined ranks, we get $R = 153$.

Now $n_1 = 13$ and $n_2 = 16$, therefore

$$\mu_R = \frac{n_1(n_1+n_2+1)}{2} = \frac{(13)(30)}{2} = 195, \text{ and}$$

$$\sigma_R = \sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}} = \sqrt{\frac{(13)(16)(30)}{12}} = 22.8$$

Thus $$z = \frac{R-\mu_R}{\sigma_R} = \frac{153-195}{22.8} = -1.84.$$

(v) The critical region is $Z < -z_{0.05} = -1.645$.

(vi) **Conclusion.** Since the computed value of $z = -1.84$ falls in the critical region, we therefore reject H_0.

Q.24.11. (a) (i) We state our hypotheses as

H_0 : There is no difference in length of life of the two types of tyres; and

H_1 : There is a significant difference in length of life of the two types of tyres.

(ii) We choose the significance level of $\alpha = 0.05$.

(iii) The test-statistic to use is the Mann-Whitney U, (the smaller of the two values).

(iv) Computations. Arranging the observations of the combined samples (Types) in order of increasing magnitude and underlining the observations from Type A we get

19, 23, 24, 27, 29, 30, 31, 35, 37, 40.

Replacing these values with their ranks (with the ranks of observations from Type A underlined), we get

1, 2, 3, 4, 5, 6, 7, 8, 9, 10.

Totalling the ranks of the two types (samples) separately, we get

$$R_1 = 2 + 4 + 5 + 6 = 17, \text{ and}$$

$$R_2 = 1 + 3 + 7 + 8 + 9 + 10 = 38$$

Now $$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 = (4)(6) + \frac{4(4+1)}{2} - 17 = 17, \text{ and}$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 = (4)(6) + \frac{6(6+1)}{2} - 38 = 7.$$

The smaller of the two values for U_1 and U_2 is taken as U-statistic, *i.e.* $U = 7$.

(v) Looking in Table of critical values of U for the Mann-Whitney Test, we find for $n_1 = 4$, $n_2 = 6$ and $\alpha = 0.05$ the critical region consists of all values of $U \leq 2$ and of all values of $U \geq 22$.

(vi) **Conclusion.** Since the computed value of $U = 7$ does not fall in the critical region, we therefore accept H_0 and conclude that three is no difference in length of life of the two types of tyres.

(b) (i) We state our hypotheses as

H_0 : The scores for the two groups do not differ, or the two groups come from identical populations

H_1 : The scores for the two groups differ significantly

(ii) We choose the significance level of $\alpha = 0.05$.

(iii) The test statistic to use is the Mann-Whitney U, which becomes

$$Z = \frac{U-\mu_u}{\sigma_u},$$

as n_1 and n_2 are both greater than 8, and Z is approximately standard normal.

(iv) Computations. Arranging the observations of the combined groups and underlining the observations from Group I, we get

25, 26, 27, 28, 29, 30, 31, 33, 34, 36, 37, 39, 41, 42, 43, 44, 46, 48, 49, 51, 53.

The corresponding ranks are

1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21.

Now $R_1 = 1 + 4 + 5 + 7 + 8 + 9 + 10 + 14 + 16 + 19 + 20 = 113,$

$R_2 = 2 + 3 + 6 + 11 + 12 + 13 + 15 + 17 + 18 + 21 = 118,$

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 = (11)(10) + \frac{11(11+1)}{2} - 113 = 63,$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 = (11)(10) + \frac{10(10+1)}{2} - 118 = 47$$

Thus $U = \min [63, 47] = 47.$

The mean and standard deviation of the sampling distribution of U are

$$\mu_U = \frac{n_1 n_2}{2} = \frac{11 + 10}{2} = 55,$$

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} = \sqrt{\frac{(11)(10)\ (11+10+1)}{12}} = 14.2$$

$$\therefore \quad z = \frac{U - \mu_U}{\sigma_U} = \frac{47 - 55}{14.2} = -0.56$$

(v) The critical region is $|Z| \geq z_{0.025} = 1.96.$

(vi) **Conclusion.** Since the computed value of Z does not fall in the critical region, so we accept H_0 and conclude that the scores for the two groups do not differ significantly.

Q.24.12. (i) We state our hypotheses as

$H_0 : M_m = M_w$ and $H_1 : M_w < M_m$.

(ii) We choose the significance level of $\alpha = 0.05$.

(iii) The test-statistic is the Mann-Whitney U, which becomes

$$Z = \frac{U - \mu_U}{\sigma_U},$$

as n_1 and n_2 are both greater than 8; and Z is approximately standard normal.

(iv) Computations. Arranging the observations (ages) of the combined samples in order of increasing magnitude and underlining the ages of women, we get

25, 25, 26, 26, 30, 32, 33, 34, 34, 35, 35, 35, 37, 38, 40, 42, 42, 43, 43, 44, 44, 46, 46, 47, 47, 47, 48, 48, 49.

Replacing these values with their ranks (with the ranks of ages of women underlined), we obtain

1.5, 1.5, 3.5, 3.5, 5, 6, 7, 8.5, 8.5, 11, 11, 11, 13, 14, 15, 16.5, 16.5, 18.5, 18.5, 20.5, 20.5, 22.5, 22.5, 25, 25, 25, 27.5, 27.5, 29.

Totalling the ranks of the two samples separately, we get

$R_1 = 1.5 + 1.5 + 3.5 + 3.5 + 7 + 11 + 13 + 14 + 15 + 16.5 + 18.5 + 20.5 + 27.5 = 153,$ and

$R_2 = 5 + 6 + 8.5 + 8.5 + 11 + 11 + 16.5 + 18.5 + 20.5 + 22.5 + 22.5 + 25 + 25 + 25 + 27.5 + 29 = 282.$

Now the value of the test-statistic U is the smaller of

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 = (13)(16) + \frac{13 \times 14}{2} - 153 = 146,$$

or $$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 = (13)(16) + \frac{16 \times 17}{2} - 282 = 62.$$

Thus $U = 62.$

The mean and standard deviation of the sampling distribution of U are

$$\mu_U = \frac{n_1 n_2}{2} = \frac{(13)\ (16)}{2} = 104, \text{ and}$$

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} = \sqrt{\frac{(13)(16)(13+16+1)}{12}} = 22.8$$

$$\therefore \quad z = \frac{62 - 104}{22.8} = \frac{-42}{22.8} = -1.84$$

(v) The critical region is $Z < -z_{0.05} = -1.645$.

(vi) **Conclusion.** Since the computed value of $z = -1.84$ falls in the critical region, we therefore reject H_0.

Q.24.13. (i) We state our hypotheses as

H_0 : The mean scores of students in arithmetic computation in two types of school are equal.

H_1 : The mean scores of students are not equal.

(ii) The significance level is set at $\alpha = 0.05$

(iii) The test-statistic is U, which becomes

$$Z = \frac{U - \mu_U}{\sigma_U},$$

as n_1 and n_2 are both greater than 8, and Z is approximately standard normal.

(iv) Computations. Regarding all observations in each group as ties, we determine their average ranks. We compute the sum of ranks for the residential students by multiplying each average rank by the number of observations in that rank. These computations are shown below:

Marks	0–9	10–19	20–29	30–39	40–49	50–59	Total
Residential	1	3	10	63	38	5	120
Non-residential	4	7	25	37	13	4	90
Total	5	10	35	100	51	9	210
Cumultative Total	5	15	50	150	201	210	--
Average Rank	$\frac{5+1}{2}$ +0 = 3	$\frac{10+1}{2}$ + 5 = 10.5	$\frac{35+1}{2}$ + 15 = 33	$\frac{100+1}{2}$ +50 = 100.5	$\frac{51+1}{2}$ + 100 = 176	$\frac{9+1}{2}$ + 201 = 206	--
Total of ranks for residentials	3 × 1 = 3	10.5×3 = 31.5	33×10 = 330	100.5×63 = 6331.5	176×38 = 6688	206×5 = 1030	R_1 = 14414

Now $U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 = (120)(90) + \frac{120(120+1)}{2} - 14414$

$= 10800 + 7260 - 14414 = 3636$, and

$U_2 = n_1 n_2 - U_1 = 10800 - 3646 = 7154,$

$\therefore \quad U = \min [U_1, U_2] = 3646.$

The mean and standard deviation of the sampling distribution of U are

$$\mu_U = \frac{n_1 n_2}{2} = \frac{(120)\,(90)}{2} = 5400, \text{ and}$$

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} = \sqrt{\frac{(10800)\,(211)}{12}}$$

$$= \sqrt{189900} = 435.775$$

$$\therefore \quad z = \frac{U - \mu_U}{\sigma_U} = \frac{3646 - 5400}{435.775} = -4.025$$

(v) The critical region is $|Z| \geq z_{0.025} = 1.96$

(vi) **Conclusion.** Since the computed value of Z falls in the critical region, so we reject H_0. We conclude that the mean scores of students in two types of school are not the same.

Q.24.14. (i) We state our hypotheses as

H_0 : There is no significant difference in achievements of the two groups, and

H_1 : There is a significant difference in achievements of the two groups.

(ii) We use a significance level of $\alpha = 0.05$

(iii) The test-statistic is U, which becomes

$$Z = \frac{U - \mu_U}{\sigma_U},$$

as n_1 and n_2 are both greater than 8, and Z is approximately standard normal.

(iv) Computations. Regarding all observations in each category or group as ties, we determine their average ranks. We compute the sum of ranks by multiplying each average rank by the corresponding number of observations in that rank. These computations are shown as follows:

Scores	40–49	50–59	60–69	70–79	80–89	90–99	Total
Teacher instructed	21	40	55	38	10	2	166
Machine-instructed	18	35	42	46	19	4	164
Total	39	75	97	84	29	6	330
Cumultative Total	39	114	211	295	324	330	--
Average Rank	$\frac{39+1}{2}$ + 0 = 20	$\frac{75+1}{2}$ + 39 = 77	$\frac{97+1}{2}$ + 114 = 163	$\frac{84+1}{2}$ + 211 = 253.5	$\frac{29+1}{2}$ + 295 = 310	$\frac{6+1}{2}$ + 324 = 327.5	--
Total of ranks for Teacher-instructed	20×21 = 420	77×40 = 3080	163×55 = 8965	253.5×38 = 9633	310×10 = 3100	327.5×2 = 655	R_1 = 25853

Now $U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$

$= (166)(164) + \frac{(166)(166+1)}{2} - 25853$

$= 27224 + 13861 - 25853 = 15232$, and

$U_2 = n_1 n_2 - U_1 = 27224 - 15232 = 11992$

$\therefore \quad U = \min [U_1, U_2] = 11992$

The mean and standard deviation of the sampling distribution of U are

$$\mu_U = \frac{n_1 n_2}{2} = \frac{(166)(164)}{2} = 13612, \text{ and}$$

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}} = \sqrt{\frac{(166)(164)(331)}{12}}$$

$$= \sqrt{750928.6667} = 866.56$$

$$\therefore \quad z = \frac{U - \mu_U}{\sigma_U} = \frac{11992 - 13612}{866.56} = -1.87$$

(v) The critical region is $|Z| \geq z_{0.025} = 1.96$

(vi) **Conclusion.** Since the computed value of $z = -1.87$ does not fall in the critical region, so we cannot reject H_0. We conclude that there is no significant difference in achievements of the two groups.

Q.24.15. (b) (i) We state our hypotheses as

$H_0 : M_m = M_w$ and $H_1 : M_w$ and M_m are different.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic to use is the Chi-square statistic with 1 degree of freedom.

(iv) Computations. Arranging the combined ages in order of increasing magnitude and locating the median of the combined data, we find that median = 40. Counting the ages that are above or below 40 in each class, we get the following information:

	Men	Women	Total
Above median	10	4	14
Below median	6	8	14
Total	16	12	28

Now $\chi^2 = \frac{n(bc-ad)^2}{(a+c)(b+d)(a+b)(c+d)}$

$= \frac{28[10 \times 8 - 4 \times 6]^2}{16 \times 12 \times 14 \times 14} = \frac{87808}{37632} = 2.33.$

(v) The critical region is $\chi^2 \geq \chi^2_{0.05,(1)} = 3.84$.

(vi) **Conclusion.** Since the calculated value of $\chi^2 = 2.33$ does not fall in the critical region, we therefore do not reject H_0.

Q.24.16. (i) We state our hypotheses as

H_0 : The two samples are drawn from populations with the same median, *i.e.* $M_1 = M_2$; and

H_1 : The populations have different medians.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic to use is the Chi-square statistic with 1 degree of freedom.

(iv) Computations. Arranging the combined observations in order of increasing magnitude and locating the median of the combined data, we find that median = 74.5. Counting the

observations that are above or below 74.5 in each sample, we find the following information:

	Sample 1	Sample 2	Total
Above median	9	11	20
Below median	11	9	20
Total	20	20	40

Now $\chi^2 = \dfrac{n(bc-ad)^2}{(a+c)\,(b+d)\,(a+b)\,(c+d)}$

$= \dfrac{40\,(9\times 9 - 11\times 11)^2}{20\times 20\times 20\times 20} = 0.4.$

(v) The critical region is $\chi^2 \geq \chi^2_{0.05,(1)} = 3.84$.

(vi) **Conclusion.** Since the calculated value of $\chi^2 = 0.4$ does not fall in the critical region, we therefore do not reject H_0 and conclude that the two samples are drawn from populations with the same median.

Q.24.17. (i) We state our hypotheses as

H_0: All the four samples come from the identical populations or $M_1=M_2=M_3=M_4$; and

H_1: At least two population medians differ.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic to use is the chi-square statistic with 3 *d.f.*

(iv) Computations. Arranging the combined observations in order of increasing magnitude and locating the median of the combined data, we find that median = 34. Counting the observations that are above or below 34 in each sample, we find the following information:

	Sample 1	Sample 2	Sample 3	Sample 4	Total
Above median	3 (3.75)	3 (3.75)	6 (3.75)	3 (3.75)	15
Below median	4 (3.25)	4 (3.25)	1 (3.25)	4 (3.25)	13
Total	7	7	7	7	28

The expected frequencies are calculated and shown in brackets in the contingency table.

$$\therefore\quad \chi^2 = \sum_i\sum_j \frac{(o_i - e_i)^2}{e_i} = \frac{(3-3.75)^2}{3.75} + \frac{(4-3.25)^2}{3.25} + \ldots + \frac{(4-3.25)^2}{3.25}$$

$$= 3.877$$

(v) The critical region is $\chi^2 \geq \chi^2_{0.05,(3)} = 7.82$

(vi) **Conclusion.** Since the calculated value does not fall in the critical region, we therefore accept H_0.

Q.24.18. (a) (i) We state our hypotheses as

H_0: There is randomness in the professor's arrangement of T and F answers, and

H_1: There is nonrandomness in the arrangement of answers.

(ii) Let us set the significance level at $\alpha = 0.05$.

(iii) The test-statistic is n_r, the number of runs in the sample.

(iv) Computations. The arrangement of T and F answers gives $n_1 = 10$, $n_2 = 10$ and $n_r = 16$ runs.

(v) Looking in the table of *critical values of n_r in the runs test* for $n_1=10$, $n_2=10$ and $\alpha=0.05$, the critical region is found to consist of all values of $n_r \leq 6$ and all values of $n_r \geq 16$.

(vi) **Conclusion.** Since the observed value of $n_r=16$ falls in the critical region, we therefore reject H_0 and conclude that evidence exists to indicate nonrandomness in the professor's arrangement of T and F answers.

(b) (i) We state our hypotheses as

H_0: The series of boys and girls selected is a random sample, and

H_1: The series is not a random sample.

(ii) Let us set the significance level at $\alpha = 0.05$.

(iii) The test-statistic is n_r, the number of runs in the sample. The distribution of n_r is approximately normal for large samples.

(iv) Computations. The series of boys and girls selected give $n_1 = 19$ boys, $n_2 = 22$ girls and $n_r = 27$ runs.

As it is a large sample, we therefore compute

$$\mu_r = \frac{2n_1n_2}{n_1 + n_2} + 1 = \frac{2(19)(22)}{19 + 22} + 1 = \frac{836}{41} + 1 = 21.39,$$

$$\sigma_r = \sqrt{\frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}}$$

$$= \sqrt{\frac{2(19)(22)[2(19)(22) - 19 - 22]}{(19+22)^2 (19+22-1)}} = \sqrt{\frac{(836)(765)}{(1681)(40)}}$$

$$= \sqrt{9.884} = 3.144.$$

Thus $z = \dfrac{n_r - \mu_r}{\sigma_r} = \dfrac{27 - 21.39}{3.144} = 1.78.$

(v) The critical region is $|Z| \geq 1.96$.

(vi) **Conclusion.** Since the computed value of $z = 1.78$ does not fall in the critical region, we therefore accept the hypothesis of a random sample.

Q.24.19. (a) (i) We state ou hypotheses as

H_0 : The given sequence of measurements is random, and

H_1 : The given sequence of measurements is nonrandom.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic is n_r, the number of runs in the given sequence.

(iv) Computations. Arranging the measurements in order of increasing magnitude, we find that median = 3.9.

Replacing each observation by a plus or a minus sign depending on whether the measurement is above or below the median (and ignoring a measurement equal to median) we get the following sequence

$-, +, -, -, -, -, +, -, +, +, -, +, +$

This gives $n_1 = 7$, $n_2 = 6$ and $n_r = 8$ runs.

(v) The critical region for $n_1 = 7$, $n_2 = 6$ and $\alpha = 0.05$, consists of all values of $n_r \leq 3$ and all values of $n_r \geq 12$. (See table for critical values of n_r in the runs test).

(vi) **Conclusion.** Since the observed value of $n_r = 8$ does not fall in the critical region, we therefore accept H_0 and conclude that the sequence of measurements is random.

(b) (i) We state our hypotheses as

H_0: The two samples are drawn from populations having identical distributions, and

H_1: The two samples are drawn from populations having different distributions.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic is n_r, the number of runs in two samples. (In other words, the Wald-Wolfowitz runs test is to be used).

(iv) Computations. Arranging the observations of the two samples in one sequence according to their magnitude, we get

<u>25</u>, <u>25</u>, <u>26</u>, <u>26</u>, 30, 32, <u>33</u>, 34, 34, <u>35</u>, 35, 35, <u>37</u>, <u>38</u>, <u>40</u>, <u>42</u>, 42, <u>43</u>, 43, <u>44</u>, 44, 46, 46, 47, 47, 47, <u>48</u>, 48, 49.

Replacing each observation of sample 1 (observations underlined) by the letter A and each observation of sample 2 by the letter B, we get the following sequence of A's and B's:

AAAABBABBABBAAAABABABBBBBBABB

This gives $n_1 = 13$, $n_2 = 16$ and $n_r = 14$ runs.

As both n_1 and n_2 exceed 10, we therefore compute

$$\mu_r = \frac{2n_1n_2}{n_1+n_2} + 1 = \frac{2(13)(16)}{13+16} + 1 = \frac{416}{29} + 1 = 15.34.$$

$$\sigma_r = \sqrt{\frac{2n_1n_2\,(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2\,(n_1+n_2-1)}}$$

$$= \sqrt{\frac{2(13)(16)[2(13)(16)-13-16]}{(13+16)^2(13+16-1)}} = \sqrt{\frac{(416)(387)}{(841)(28)}}$$

$$= \sqrt{6.8368} = 2.61.$$

Thus $z = \dfrac{n_r - \mu_r}{\sigma_r} = \dfrac{14 - 15.34}{2.61} = -0.51.$

(v) The critical region is $|Z| \geq 1.96$.

(vi) **Conclusion.** Since the computed value of $z = -0.51$ does not fall in the critical region, we therefore accept H_0.

Q.24.20. (b) (i) We state our hypotheses as

H_0: The population distribution is normal with mean 0.5 and variance 1; and

H_1: The population distribution is not normal.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic is the Kolmogorov-Smirnov one-sample D statistic.

(iv) Computations. Let $S_{20}(X_i)$ denote the *cumulative* relative frequency (probability) of 20 sample observations (*ordered*) and let $F_0(X)$ denote the *cumulative* probability distribution under H_0, where H_0 is that X is normally distributed with mean = 0.5 and variance = 1. This implies that the standard normal variable is $Z = \dfrac{X_i - 0.5}{\sqrt{1}} = X_i - 0.5$. Thus the figures in the $F_0(X_i)$ column are taken off the standard normal table as $P(Z \leq X_i - 0.5)$. The necessary computations appear in the following table:

Sample Value (X_i)	$S_{20}(X_i)$	$F_0(X_i)$	$\lvert S_{20}(X_i) - F_0(X_i)\rvert$
−1.26	1/20 = 0.05	0.04	0.01
−1.06	2/20 = 0.10	0.06	0.04
−0.95	3/20 = 0.15	0.07	0.08
−0.74	4/20 = 0.20	0.11	0.09
−0.56	5/20 = 0.25	0.14	0.11
−0.49	6/20 = 0.30	0.16	0.14
−0.48	7/20 = 0.35	0.16	0.19
−0.15	8/20 = 0.40	0.26	0.14
−0.10	9/20 = 0.45	0.27	0.18
0.15	10/20 = 0.50	0.36	0.14
0.24	11/20 = 0.55	0.40	0.15
0.32	12/20 = 0.60	0.43	0.17
0.36	13/20 = 0.65	0.44	0.21
0.55	14/20 = 0.70	0.52	0.18
0.56	15/20 = 0.75	0.52	0.23
0.70	16/20 = 0.80	0.58	0.22
0.82	17/20 = 0.85	0.63	0.22
0.92	18/20 = 0.90	0.66	0.24
1.74	19/20 = 0.95	0.89	0.06
1.86	20/20 = 1.00	0.91	0.09

Now $D = \max |S_{20}(X_i) - F_0(X_i)| = 0.24.$

(v) The critical value of D for $n = 20$ and $\alpha = 0.05$ from the tables of critical values for Kolmogorov-Smirnov-one sample test is 0.29.

(vi) **Conclusion.** Since the computed value of $D = 0.24$ does not fall in the critical region, we therefore accept H_0 and conclude that the sample comes from a normal distribution with mean 0.5 and variance 1.

Q.24.21. (i) We state our hypotheses as

H_0: The population distribution is normal with mean = 85 grams and σ = 15 grams; and

H_1: The population distribution is not normal.

(ii) We choose the significance level of $\alpha = 0.05$.

(iii) The test-statistic is the Kolmogorov-Smirnov one sample D-statistic.

(iv) Computations. Let $S_{36}(X_i)$ denote the *cumulative relative frequency* (probability) of the kidney weights in grams of 36 days (*ordered*) and let $F_0(X_i)$ denote the *cumulative* probability distribution under H_0, where H_0 is that X is normally distributed with mean = 85 and σ = 15. This implies that the standard normal variable is $Z = (X_i - 85)/15$. Thus the figures in the $F_0(X_i)$ column are taken off the standard normal table as $P[Z \leq (X_i - 85)/15]$.

The necessary computations appear in the following table:

Sample value (X_i)	f	$S_{36}(X_i)$	z_i [$(X_i - 85/15]$	$F_0(X_i)$	$\lvert S_{36}(X_i) - F_0(X_i)\rvert$
58	1	1/36 = 0.0278	−1.80	0.0359	0.0081
59	1	2/36 = 0.0556	−1.73	0.0418	0.0138
67	1	3/36 = 0.0833	−1.20	0.1151	0.0318
68	3	6/36 = 0.1667	−1.13	0.1292	0.0375
70	4	10/36 = 0.2778	−1.00	0.1587	0.1191
74	1	11/36 = 0.3056	−0.73	0.2327	0.0729
75	1	12/36 = 0.3333	−0.67	0.2514	0.0819
76	1	13/36 = 0.3611	−0.60	0.2743	0.0868
78	1	14/36 = 0.3889	−0.47	0.3192	0.0697
80	2	16/36 = 0.4444	−0.33	0.3707	0.0737
82	2	18/36 = 0.5000	−0.20	0.4207	0.0793
83	1	19/36 = 0.5278	−0.13	0.4483	0.0795
84	3	22/36 = 0.6111	−0.07	0.4721	0.1392
86	1	23/36 = 0.6389	+0.07	0.5279	0.1112
88	1	24/36 = 0.6667	0.13	0.5517	0.1152
90	4	28/36 = 0.7778	0.33	0.6293	0.1485 = D
92	1	29/36 = 0.8056	0.47	0.6808	0.1248
93	1	30/36 = 0.8333	0.53	0.7010	0.1323
94	1	31/36 = 0.8611	0.60	0.7257	0.1354
97	1	32/36 = 0.8889	0.80	0.7881	0.1008
98	1	33/36 = 0.9167	0.87	0.8078	0.1089
104	1	34/36 = 0.9444	1.27	0.8980	0.0464
110	1	35/36 = 0.9722	1.67	0.9515	0.0207
112	1	36/36 = 1.0000	$+\infty$	1.0000	0
Σ	36	--	--	--	--

Now $D = \max \lvert S_{36}(X) - F_0(X)\rvert = 0.1485$

(v) The critical value of D for $n = 36$ and $\alpha = 0.05$ from the tables of critical values for Kolmogorov-Smirnov-one sample test is 0.221.

(vi) **Conclusion.** Since the computed value of D does not exceed the table values, so we accept H_0 and conclude that the data have come from a normally distributed population with mean = 85 and σ = 15.

Q.24.22. (i) We state our hypotheses as

H_0: The two samples come from populations having identical distributions, or $F_1(X) = F_2(X)$, and

H_1: The two samples come from different population distributions, *i.e.* $F_1(X) \neq F_2(X)$.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic to use is the Kolmogorov-Smirnov-two-sample D statistic.

(iv) Computations. We arrange all the observations together in increasing order of magnitude. We compute the sample cumulative relative frequencies at each sample value and find the differences between them at each listed point. The ordered sample values, corresponding values of $S_{13}(X_1)$ and $S_{16}(X_2)$ and differences $S_{13}(X_1) - S_{16}(X_2)$ are given as follows:

Ordered Observations				
X_1	X_2	$S_{13}(X_1)$	$S_{16}(X_2)$	$\lvert S_{13}(X_1) - S_{16}(X_2)\rvert$
25,25	--	2/13	0	$\lvert 2/13 - 0\rvert = 2/13 = 32/208$
26,26	--	4/13	0	$\lvert 4/13 - 0\rvert = 4/13 = 64/208$
--	30	4/13	1/16	$\lvert 4/13 - 1/16\rvert = 51/208$
--	32	4/13	2/16	$\lvert 4/13 - 2/16\rvert = 38/208$
33	--	5/13	2/16	= 54/208
--	34,34	5/13	4/16	= 28/208
35	35,35	6/13	6/16	= 18/208
37	--	7/13	6/16	= 24/208
38	--	8/13	6/16	= 50/208
40	--	9/13	6/16	= 66/208
42	42	10/13	7/16	= 69/208
43	43	11/13	8/16	= 72/208
44	44	12/13	9/16	= 75/208 (= D)
--	46,46	12/13	11/16	= 49/208
--	47,47,47	12/13	14/16	= 10/208
48	48	13/13	15/16	= 13/208
--	49	13/13	16/16	= 0

Now $D = \max |S_{13}(X_1) - S_{16}(X_2)| = \frac{75}{208} = 0.36$; and

$n_1 = 13$ and $n_2 = 16$.

(v) The critical value of D for $n_1 = 13$, $n_2 = 16$ and $\alpha = 0.05$ is

$$1.2239\sqrt{\frac{n_1 + n_2}{n_1 n_2}} = 1.2239\sqrt{\frac{29}{208}}$$

$$= (1.2239)\,(0.37339) = 0.46$$

(vi) **Conclusion.** Since the computed value of D does not exceed the table value, so we accept H_0.

(b) (i) The hypotheses are stated as

H_0 : There is no difference between the two types of schools or equivalently, $F_1(X) = F_2(X)$,

H_1 : The two types of schools differ.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) As both sample sizes n_1 and n_2 are greater than 40, for a one-tailed test, the test statistic to use is

$$\chi^2 = 4D^2\left[\frac{n_1 n_2}{n_1 + n_2}\right]$$

which has approximately a chi-square distribution with 2 *d.f.*

(iv) The necessary computations are shown below:

Marks	Cumulative *RF*		Cumulative *RF*		
	f_1	$S(X_1)$	f_2	$S(X_2)$	$\lvert S(X_1)-S(X_2)\rvert$
0–9	1	1/120 = 0.0083	4	4/90 = 0.0444	0.0361
10–19	3	4/120 = 0.0333	7	11/90 = 0.1222	0.0889
20–29	10	14/120 = 0.1167	25	36/90 = 0.4000	0.2833 = *D*
30–39	68	77/120 = 0.6417	37	73/90 = 0.8111	0.1694
40–49	38	115/120 = 0.9583	13	86/90 = 0.9556	0.0027
50–59	5	120/120 = 1.0000	4	90/90 = 1.000	0
Σ	120	--	90	--	--

Now $D = \max |S(X_1) - S(X_2)| = 0.2833$

$$\therefore \quad \chi^2 = 4(0.2833)^2\left[\frac{120 \times 90}{210}\right]$$

$$= (0.3210)\,(51.42857) = 16.51.$$

(v) The critical region is $\chi^2 \geq \chi^2_{0.05,(2)} = 5.99$

(vi) **Conclusion.** Since the computed value of χ^2 falls in the critical region, we therefore reject H_0. We conclude that the two types of schools differ significantly.

Q.24.23. (i) The hypotheses are stated as

H_0 : The four groups are taken from populations with the same means, *i.e.* $\mu_1 = \mu_2 = \mu_3 = \mu_4$; and

H_1 : At least two of the four means are not equal.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic to use is the Kruskal-Wallis H statistic.

(iv) Computations. Replacing each observation with its corresponding rank, we get

Group *A*	6, 20, 11, 10, 25, 19, 12, 5	$108 = R_1$
Group *B*	3, 9, 2, 4, 1	$19 = R_2$
Group *C*	22, 13, 7, 21, 23, 24	$110 = R_3$
Group *D*	8, 14, 16, 15, 18, 17	$88 = R_4$

Now $H = \frac{12}{n(n+1)}\sum_{i=1}^{4}\frac{R_i^2}{n_i} - 3(n+1)$

$$= \frac{12}{25(25+1)}\left[\frac{(108)^2}{8} + \frac{(19)^2}{5} + \frac{(110)^2}{6} + \frac{(88)^2}{6}\right] - 3(25+1)$$

$$= \frac{12}{650}[1458 + 72.2 + 2016.67 + 1290.67] - 78$$

$$= \frac{12}{650}(4837.54) - 78 = 89.32 - 78 = 11.32$$

(v) The critical region is $H > \chi^2_{0.05,(3)} = 7.82$

(vi) **Conclusion.** Since the computed value of H falls in the critical region, so we reject H_0 and conclude that the four groups come from populations having different means.

Q.24.24. (i) The hypotheses are stated as

H_0 : The mean tensile strengths are the same for all the methods, and

H_1 : The mean tensile strengths are not the same for all the methods.

(ii) The significance level is chosen at $\alpha = 0.05$.

(iii) The test-statistic to use is the Kruskal-Wallis H statistic.

(iv) Computations. We arrange the observations of the combined methods in increasing order of magnitude and assign ranks, which are shown in the following table:

Methods	Ranks	R_i
A	10.5, 22.5, 21, 18.5, 25, 22.5, 16	136
B	32.5, 26, 30, 30, 32.5, 28, 27, 30	236
C	24, 14, 13, 10.5, 18.5, 18.5, 18.5, 15	132
D	5.5, 7, 3.5, 8, 5.5, 2, 1, 10.5, 3.5, 10.5	57

Now $$H = \frac{12}{n(n+1)} \sum \frac{R_i^2}{n_i} - 3(n+1)$$

$$= \frac{12}{33(33+1)} \left[\frac{(136)^2}{7} + \frac{(236)^2}{8} + \frac{(132)^2}{8} + \frac{(57)^2}{10}\right] - 3(33+1)$$

$$= \frac{12}{1122} [2642.29 + 6962 + 2178 + 324.9] - 102$$

$$= 129.49 - 102 = 27.49$$

(v) The critical region is $H > \chi^2_{0.05,(3)} = 7.82$

(vi) **Conclusion.** Since the computed value of $H = 27.49$ falls in the critical region, we therefore reject H_0. We conclude that at least one of the methods gives a mean tensile strength that differs from at least one of the other methods.

❖❖❖❖❖❖❖❖❖❖

APPENDIX - A

VITAL STATISTICS

Q.A.7. (a) (i) Vital Events. There are some factors which cause changes in the size and composition of human population, e.g. births add and deaths take away some members of the population. Such factors are called *Vital Events* and they include births, deaths, marriages, divorces, sickness, adoptions, legitimation, etc.

(ii) **Sources of Vital Data.** Some of the important sources of vital data are given below:

(a) **Population Census Reports** which contain data relating to age, sex, marital status, occupation, housing conditions, etc.

(b) **Birth and Death Returns** compiled by Health Departments.

(c) **Demography Yearbook** which contains statistics on population, births and deaths.

(d) **Migration Records,**

(e) **Epidemiological Reports** which contain data on infectious diseases, etc.

(iii) **Vital Index** is defined as the ratio of births to deaths in a specified time. That is

$$\text{Vital Index} = \frac{\text{number of births}}{\text{number of deaths}} \times 100$$

If the vital index is greater than 100, it means that the population is increasing. If it is less than 100, it implies that the population is decreasing. The vital index will not be affected when deaths and births are under recorded.

(b) Calculation of Age-specific Death Rates.

Age-specific death rates are computed by the formula

$$A.S.D.R = \frac{d_i}{p_i} \times 1{,}000$$

where d_i denotes the number of deaths in the ith age-group during a given year, and

p_i denote the mid-year population figure of the same age-group.

Applying this formula, we get

Age (Completed years)	1964 estimated population. Both sexes in 1000 (p_i)	1964 estimated deaths in 100 (d_i)	Age-specific death rates = $(d_i/p_i)\times 1000$
All ages	42,390	8,274	19.52
Under 1	1,363	3,175	232.94
1 – 4	5,578	1,841	33.00
5 – 9	6,595	401	6.08
10 – 19	8,121	255	3.14
20 – 29	6,818	329	4.83
30 – 39	5,150	309	6.00
40 – 49	3,686	316	8.57
50 – 59	2,403	277	11.53
60 & over	2,679	1,370	51.14

Q.A.8. The age-specific death rates are given below:

Age-group (years)	Age-specific death rates per 1000 persons per year	
	Males	Females
10 – 14	13.20	16.09
15 – 19	18.90	20.95
20 – 29	9.28	11.12
30 – 39	14.16	16.86
40 – 49	20.98	23.20
50 – 59	32.96	38.38
60 & over	38.25	45.46

Q.A.9. (b) Computation of Specific Mortality Rates.

Age in years	Population as of Census data (p_i)	Deaths from Cancer (d_i)	Specific Mortality Rates $= (d_i/p_i)$ 100,000
20 – 29	725,369	79	10.89
30 – 39	700,213	238	33.99
40 – 49	609,616	660	108.26
50 – 59	548,400	1,582	288.48
60 – 69	402,054	2,342	582.51
70 & over	287,280	3,402	1184.21

Q.A.11. (b) Calculation of the Crude and Standardized Death Rates of 1964 population, using 1941 population as Standard.

(i) By Direct Method.

Age-group (years)	1941		1964			Expected deaths $= (d_i/p_i)\times P_i$
	Population (P_i)	Deaths (D_i)	Population (p_i)	Specific death rate	Deaths (d_i)	
0 – 9	10,000	220	15,000	20	300	200
10 – 19	11,000	132	12,000	11	132	121
20 – 49	7,000	105	8,000	14	112	98
50 & over	2,000	90	5,000	42	210	84
Total	30,000	547	40,000	--	754	503

$$\therefore \text{ Crude Death Rate} = \frac{\text{Total deaths in 1964}}{\text{Total Population in 1964}} \times 1{,}000$$

$$= \frac{754}{40{,}000} \times 1{,}000 = 18.85, \text{ and}$$

$$\text{S.D.R.(Direct)} = \frac{\text{Expected deaths in standard population}}{\text{Total standard population}} \times 1{,}000$$

$$= \frac{\sum (d_i/p_i) \times P_i}{\sum P_i} \times 1{,}000 = \frac{503}{30{,}000} \times 1{,}000 = 16.77$$

(ii) **By Indirect Method**, *i.e.* by computing the expected deaths by the formula: expected death $= \frac{D_i}{P_i} \times p_i$.

Age-group (years)	1941		1964	Expected deaths $= (D_i/P_i) \times p_i$
	Population (P_i)	Deaths (D_i)	Population (p_i)	
0 – 9	10,000	220	15,000	330
10 – 19	11,000	132	12,000	144
20 – 49	7,000	105	8,000	120
50 & over	2,000	90	5,000	225
Total	30,000	547	40,000	819

C.D.R. of standard population $= \frac{547}{30,000} \times 1,000 = 18.23$

$\therefore$ S.D.R. (Indirect) $= \frac{\text{Deaths in actual population}}{\text{Expected deaths}} \times$ C.D.R. of standard population.

$= \frac{547}{819} \times 18.23 = \frac{13745.42}{819} = 16.78$

Hence the crude death rate is 18.85 and the standardized death rate by *direct* method is 16.77, while by *indirect* method is 16.78.

Q.A.12 Computation of the *crude* and the *standardized* death rates by the direct method.

Age-group (years)	Standard		Local		Expected deaths $= (d_i/p_i) \times P_i$
	Population (P_i)	Deaths (D_i)	Population (p_i)	Deaths (d_i)	
0 – 9	600	18	400	16	24
10 – 19	1,000	5	1,500	6	4
20 – 59	3,000	24	2,400	24	30
60 & over	400	20	700	21	12
Total	5,000	67	5,000	67	70

Hence Crdue death rate (Local) $= \frac{\sum d_i}{\sum p_i} \times 1,000$

$= \frac{67}{5,000} \times 1,000 = 13.4$ and

Standardized death rate $= \frac{\sum (d_i/p_i) \times P_i}{\sum P_i} \times 1,000$

$= \frac{70}{5,000} \times 1,000 = 14.0$

Q.A.13. (b) Calculation of Crude Death Rate and Standardized Death Rate.

Age (years)	Standard Population ('000)		District A				Expected Death	
			Population		Deaths			
	males (P_{im})	females (P_{if})	males (p_{im})	females (p_{if})	males (d_{im})	females (d_{if})	$\frac{d_{im}}{p_{im}} \times P_{im}$	$\frac{d_{if}}{p_{if}} \times P_{if}$
0 - 4	59	55	2,110	2,010	30	27	838.9	738.8
5 - 14	109	102	3,340	3,230	6	8	195.8	252.6
15-34	177	180	7,320	7,310	16	20	386.9	492.5
35-59	121	122	7,960	8,750	70	57	1,064.1	794.7
60&over	34	41	3,240	4,280	196	230	2,056.8	2,203.3
Total	500	500	23,970	25,580	318	342	4,542.5	4,481.5

Now, C.D.R. $= \frac{\sum d_{im} + \sum d_{if}}{\sum p_{im} + \sum p_{if}} \times 1,000$

$= \frac{318 + 342}{23970 + 25580} \times 1000 = \frac{660}{49550} \times 1000 = 13.32$, and

S.D.R $= \frac{\sum \frac{d_{im}}{p_{im}} \times P_{im} + \sum \frac{d_{if}}{p_{if}} \times P_{if}}{\sum P_{im} + \sum P_{if}} \times 1,000$

$= \frac{4,542.5 + 4,481.5}{500,000 + 500,000} \times 1,000 = \frac{9024.0}{1,000,000} \times 1,000 = 9.024$

Q.A.14 Calculation of the crude accident rate per cent for each factory and the standardized accident rate per cent for factory II.

Age-group (years)	Factory I		Factory II		Expected
	No. of employees (P_i)	No. of accidents (D_i)	No. of employees (p_i)	No. of accidents (d_i)	No. of accidents $= (d_i/p_i) \times P_i$
Under 21	330	28	400	38	31.35
21 – 29	570	40	720	67	53.04
30 – 39	710	45	810	60	52.59
40 – 49	780	55	390	34	68.00
50 – 59	690	54	250	25	69.00
60&over	250	25	80	11	34.38
Total	3330	247	2650	235	308.36

Thus (i) crude accident rate for

$$\text{Factory I} = \frac{\text{Number of accidents}}{\text{Number of employees}} \times 100$$

$$= \frac{247}{3330} \times 100 = 7.42,$$

$$\text{Factory II} = \frac{235}{2650} \times 100 = 8.87.$$

(ii) Standardized accident rate for factory II

$$= \frac{\text{Expected number of accidents}}{\text{Total number of employees}} \times 100$$

$$= \frac{\sum \frac{d_i}{p_i} \times P_i}{\sum P_i} \times 100$$

$$= \frac{308.36}{3330} \times 100 = 9.26.$$

Hence the crude accident rate per cent for Factory I is 7.42, for Factory II is 8.87 and the standardized accident rate per cent for Factory II is 9.26.

Q.A.17. Computation of Age-specific fertility rates, T.F.R., GRR and NRR.

Age group (1)	Mid year Female Population (000) (2)	Registered Births (3)	Female Babies (4)	Age-specific fertility rate per 1000 women (5)	Age-specific fertility rates for daughters only (6)	Probability of survival (7)	Expected Survivors of female births per woman (8)
15 – 19	1424	27,639	13,405	19.41	0.0094	0.9645	0.0091
20 – 24	1531	226,817	110,006	148.15	0.0719	0.9607	0.0691
25 – 29	1653	280,506	136,045	169.70	0.0823	0.9554	0.0786
30 – 34	1658	194,526	94,345	117.33	0.0569	0.9489	0.0540
35 – 39	1741	113,966	55,274	65.46	0.0317	0.9416	0.0298
40 – 44	1669	32,363	15,696	19.39	0.0094	0.9324	0.0088
45 – 49	1561	2,215	1,074	1.42	0.0007	0.9201	0.0006
Total	11,237	878,032	425,845	540.86	0.2623	--	0.2500
Multiply by 5 for single age figures				2704	1.3115	--	1.25

Hence T.F.R. $= 5 \sum$ (age-specific birth rates)

$= 2{,}704$ per 1,000 women

G.R.R. $= 5 \sum$ (age-specific fertility rates for daughters only)

$= 1.3115$ per woman, and

N.R.R. $= 5 \sum$ (expected survivors of female births)

$= 1.25$ per woman.

Note: Female births were calculated by taking the sex-ratio of 106.18 per cent into account.

Q.A.18. (b) Calculation of the Net Reproduction Rate.

Age group (years)	Female Population (P_{if})	Female Births (b_{if})	Survival rate	Mean fertility rate (b_{if}/P_{if})	Female offspring of Survivors
15 – 19	87	4	0.850	0.0460	0.03910
20 – 24	63	11	0.800	0.1746	0.13968
25 – 29	55	8	0.700	0.1455	0.10185
30 – 34	41	6	0.650	0.1463	0.09510
35 – 39	33	4	0.600	0.1212	0.07272
40 – 44	36	2	0.500	0.0556	0.02780
Total	315	35	--	--	0.47625

Hence N.R.R. = Class-Interval × Σ (Mean fertility rate × Survival factor)

= 5 × 0.47625 = 2.38125

Hence the required net reproduction rate is 2.38.

Q.A.19 Computation of the Gross and Net Reproduction Rates.

Age group (years)	Female Population ('000) (P_{if})	Female live Births (b_{if})	Survival rate	Mean fertility rate (b_{if}/P_{if})	Female offspring of Survivors
15 - 19	1399	15,133	0.9694	0.010817	0.010486
20 - 24	1422	94,155	0.9668	0.066213	0.064015
25 - 29	1521	102,676	0.9632	0.067506	0.065022
30 - 34	1756	72,490	0.9584	0.041281	0.039564
35 - 39	1451	31,402	0.9515	0.021642	0.020601
40 - 44	1689	10,640	0.9424	0.006300	0.005937
45 - 49	1667	700	0.9279	0.000420	0.000390
Total	--	--	--	0.214179	0.206015

Hence G.R.R. = Class-Interval × $\Sigma \frac{\text{Female births}}{\text{Female population}}$

= 5 × 0.214179 = 1.071, and

N.R.R. = 5 × 0.206015 = 1.030.

Q.A.20. (b) A sex-ratio = 105.2% implies that there are 1052 males for 1000 females.

Calculation of Gross and Net Reproduction Rates.

Age group (years) (1)	Female Population (2)	Registered Births (3)	Female Babies (4)	Age-specific fertility rates for daughters only (5)	Probability of survival (6)	Expected Survivors of female births per woman (7)
15 - 19	8981	1835	894	0.0905	0.634	0.0631
20 - 24	5875	2616	1275	0.2170	0.602	0.1306
25 - 29	3613	2563	1249	0.3457	0.568	0.1964
30 - 34	3380	1062	518	0.1533	0.530	0.0812
35 - 39	3345	558	272	0.0813	0.488	0.0397
40 - 44	3248	37	18	0.0055	0.444	0.0024
Total	28,442	8,671	4,226	0.9023	--	0.5134
Multiply by 5 for single age figures				4.5115	--	2.5670

Hence G.R.R. = 5 Σ (age-specific fertility rate for daughters only)

= 4.512 per woman, and

N.R.R. = 5Σ (expected survivors of female births)

= 2.567 per woman.