Module-V

Statistical Techniques-III

Application in Engineering:

The application of the *t* distribution to the following four types of problem will now be considered.

1. The calculation of a confidence interval for a sample mean.
2. The mean and standard deviation of a sample are calculated and a value is postulated for the mean of the population. How significantly does the sample mean differ from the postulated population mean?
3. The means and standard deviations of two samples are calculated. Could both samples have been taken from the same population?
4. Paired observations are made on two samples (or in succession on one sample). What is the significance of the difference between the means of the two sets of observations?

chi-square test for categorical variables determines whether there is a difference in the population proportions between two or more groups. In the medical literature, the Chi-square is used most commonly to compare the incidence (or proportion) of a characteristic in one group to the incidence (or proportion) of a characteristic in other group(s). ANOVA technique is intended to analyse variability in data in order to infer the inequality among population means. The application data were analysed using computer program MATLAB that performs these calculations.

Course Outcome

After completion of this course, students will be able to learn

CO-5	Understand the statistical method of data samples , hypothesis testing and applying the study of control chart and their properties.	**BL-2,3**

Syllabus

Sampling, Testing of Hypothesis and Statistical Quality Control: Introduction, Sampling Theory (Small and Large), Hypothesis, Null Hypothesis, Alternative Hypothesis, Testing a Hypothesis, Level of Significance, Confidence Limits, Test of Significance of Difference Means, T-test, F-test, and Chi-square test, One way Analysis of Variance (ANOVA), Statistical Quality Control SQC, Control Charts, Control Chart for Variables ($\bar{X}$ and R Charts) Control Charts for Variables (p, np and C Charts).

Content

5.1 Introduction

Population or Universe

An aggregate of objects (animate or inanimate) under the study is called population or universe. Thus it is a collection of individuals or of their attributes or of results of operations which can be numerically specified.

There are different types of population in statics-

(1) **Finite population-** An universe containing finite number of individuals or member is called finite population.
Example: Universe of weights of students in a particular class, books in college library.

(2) **Infinite population-** It consists of infinite number of elementary or units.
Example: Universe of pressures of various points in the atmosphere or stars.

(3) **Real or true or existent population--** The universe of concrete objects is an existent universe.
Example: Employ of central govt. at certain date and time.

(4) **Hypothetical population-** The collection of all possible ways in which specified event can happen is called hypothetical universe.
Example: Possible outcomes of rolling a die n times

Sampling

A finite subset of universe is called a sample. The number of individuals in a sample is called sample size. The process of selecting a sample from a universe is called sampling.

Example: In a shop, we assess the quality of sugar, rice or any other commodity by taking only a handful of it from the bag and then decide whether to purchase it or not.

5.2 Test of Significance

Population mean and variance are denoted by μ and σ^2 while sample mean and variance are represented by $\bar{x}$ and s^2. For applying test of significance we first set up a hypothesis which is the definite statement about the population.

Null Hypothesis:-It is a definite statement about population parameter and denoted by H_0. A null hypothesis is tested for possible rejection under the assumption that it is true.

Alternative hypothesis:- Any hypothesis which complimentary to the null hypothesis (H_0) is called an alternative hypothesis. It is denoted by H_1.

For example, If we want to test the null hypothesis that the population has a specified mean μ_0 then we have

$$H_0: \mu = \mu_0$$

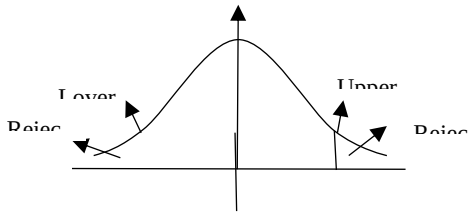
Then alternative hypothesis will be

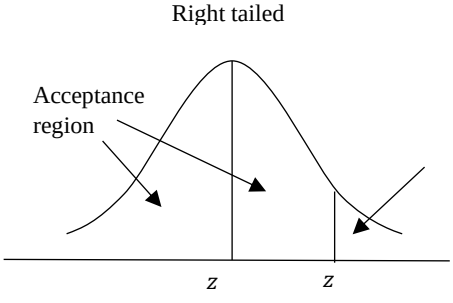
(i) $H_1: \mu \neq \mu_0$ (two tailed alternative hypothesis)
(ii) $H_1: \mu > \mu_0$ (right tailed alternative hypothesis)
(iii) $H_1: \mu < \mu_0$ (left tailed alternative hypothesis)

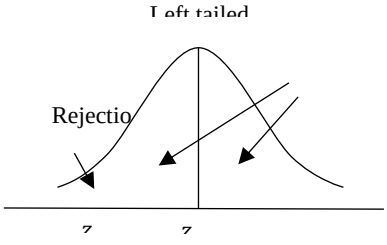
Level of Significance

The probability level below which we reject the hypothesis is known as level of significance. It is denoted by α. The level of significance usually employed in testing of hypothesis are 5% and 1%.

(Two tailed test)







Level of significance			
	1% (0.01)	5% (0.05)	10% (0.1)
Two tailed test	$\|z_\alpha\| = 2.58$	$\|z_\alpha\| = 1.96$	$\|z_\alpha\| = 1.645$
Right tailed test	$z_\alpha = 2.33$	$z_\alpha = 1.645$	$z_\alpha = 1.28$
Left tailed test	$z_\alpha = -2.33$	$z_\alpha = -1.645$	$z_\alpha = -1.28$

5.3 Student's t-Distribution (t- Test)

t- distribution is used when sample size ≤ 30. t- statics is defined as

$$t = \frac{\bar{x}-\mu}{S/\sqrt{n}}, \text{ where } S = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

$\bar{x}$ is the mean of the sample, μ is population mean, S is standard deviation of population and n is sample size.

If the standard deviation of the sample $'s'$ is given then t-statics is defined as

$$t = \frac{\bar{x}-\mu}{s/\sqrt{(n-1)}}$$

Application of t-distribution:

1. To test if the sample mean $(\bar{x})$ differ significantly from the hypothetical value μ of the population mean.
2. To test significance difference between two sample means.
3. To test the significance of observed partial and multiple correlation coefficients.

Test I: To test whether the mean of a sample drawn from normal population deviates significantly from a stated value when variance of population is unknown.

Working rule

- H_0 – There is no significant difference between sample mean $\bar{x}$ and population mean μ.
- Calculate t-statics:

$$t = \frac{\bar{x}-\mu}{S/\sqrt{n}}, \text{ where } S = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

with degree of freedom $(n-1)$.

- If calculated t value is such that $|t| < t_\alpha$, the null hypothesis is accepted and if $|t| > t_\alpha$, the null hypothesis is rejected.

Note:

1. 95% confidence limits (level of significance 5%) are $\bar{x} \pm t_{0.05} S/\sqrt{n}$
2. 99% confidence limits (level of significance 1%) are $\bar{x} \pm t_{0.01} S/\sqrt{n}$

Example 1: A random sample size 16 and 53 as mean. The sum of the squares of the deviation from the mean is 135. Can this sample be regarded as taken from the population having 56 as mean? Obtain 95% and 99% confidence limits of the mean of the population.

Solution: Null hypothesis H_0: there is no significant difference between the sample mean and hypothetical population mean i.e. $\mu = 56$.

Alternative hypothesis H_1: $\mu \neq 56$ (two tailed test).

We know that

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

Given $\bar{x} = 53, \mu = 56, n = 16, \sum(x-\bar{x})^2 = 135$.

Now we have $S = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} = \sqrt{\frac{135}{15}} = 3$

$$\Rightarrow t = \frac{\bar{x}-\mu}{\frac{S}{\sqrt{n}}} = \frac{53-56}{\frac{3}{\sqrt{16}}} = -4$$

$$\Rightarrow |t| = 4$$

$$d.f. = (n-1) = 16-1 = 15,\ t_{0.05} = 2.13$$

$\because |t| = 4 > t_{0.05} = 2.13$, the null hypothesis is rejected. Hence the sample mean has not come from a population having 56 as mean.

95% confidence limits of the population mean

$$= \bar{x} \pm \frac{S}{\sqrt{n}} t_{0.05} = 53 \pm \frac{3}{\sqrt{16}}(2.13) = 51.4025, 54.5975$$

99% confidence limits of the population mean

$$= \bar{x} \pm \frac{S}{\sqrt{n}} t_{0.01} = 53 \pm \frac{3}{\sqrt{16}}(2.95) = 50.7875, 55.2125$$

Example 2: The lifetime of the electric bulbs for a random sample of 10 from a large consignment gave the following data:

Item	1	2	3	4	5	6	7	8	9	10
Life in '000 hrs	4.2	4.6	3.9	4.1	5.2	3.8	3.9	4.3	4.4	5.6

Can we accept the hypothesis that the average lifetime of bulb is 4000 hrs ?

Solution: Null hypothesis H_0: There is no significant difference in the sample mean and population mean i.e. $\mu = 4000\ hrs$.

Alternative hypothesis $H_1: \mu \neq 4000\ hrs$ (two tailed test)

Now we have

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

The calculation table is

x	4.2	4.6	3.9	4.1	5.2	3.8	3.9	4.3	4.4	5.6
$x - \bar{x}$	-0.2	0.2	-0.5	-0.3	0.8	-0.6	-0.5	-0.1	0	1.2
$(x - \bar{x})^2$	0.04	0.04	0.25	0.09	0.64	0.36	0.25	0.01	0	1.44

$$\bar{x} = \frac{\sum x}{n} = \frac{44}{10} = 4.4, \quad \sum (x - \bar{x})^2 = 3.12$$

$S = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}} = 0.589,$

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}} = \frac{4.4 - 4}{\left(\frac{0.589}{\sqrt{10}}\right)} = 2.123$$

Degree of freedom = $n - 1 = 10 - 1 = 9$, $t_{0.05} = 2.26$

Since the calculated value of t is less than the tabulated value of t at 5% level of significance.

$\therefore$ Null hypothesis is accepted. i.e. the average lifetime of bulbs could be 4000 hrs.

Example 3: A shampoo manufacturing company was distributed a particular brand of shampoo through a large number of retail shops. Before a heavy advertisement. Campaign, the mean sales per shampoo was 140 dozens.

After the campaign a sample of 26 shampoo was taken and the mean sales figure was found to be 147 dozens with standard dozens with standard deviation 16. Can you consider the advertisement effective ? (Given $t_{0.05,25} = 1.708$)

Solution: Null hypothesis H_0: There is no difference in between sample means and population means, i.e. the advertisement is not effective i.e. $\mu = 140$.

Given $n = 26$, mean $\bar{x} = 147, S.D. = 16,$

and degree of freedom $= n - 1 = 26 - 1 = 25.$

We know that

$$t = \frac{\bar{x}-\mu}{s/\sqrt{(n-1)}}, \qquad \text{(S.D. is known)}$$

$$= \frac{(147 - 140)\sqrt{25}}{16} = 2.19.$$

The tabulated value of t at 5% level of significance for $d.f. = 25$ is 1.708. i.e. $t_{0.05,25} = 1.708.$

Hence calculated value of $|t| = 2.19 >$ tabulated value of $t_{0.05,25} = 1.708.$

∴ Null hypothesis is rejected. The advertisement is effective.

Test II : t −test for difference of means of two small samples (from a normal population)

Let two samples $x_1, x_2, \dots, x_{n_1}, y_1, y_2, \dots, y_{n_2}$ of size n_1, n_2 have been drawn from two normal populations with mean μ_1 and μ_2 respectively under the assumption that the population variance are equal ($\sigma_1 = \sigma_2 = \sigma$).

Let $\bar{x}, \bar{y}$ be their means of two samples, the test statistic is given by

$$t = \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$d.f. = n_1 + n_2 - 2$$

- If two sample's standard deviations s_1, s_2 are given then we have $S^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$.
- If s_1, s_2 are not given then $S^2 = \frac{\sum(x_1 - \overline{x_1})^2 + \sum(x_2 - \overline{x_2})^2}{n_1 + n_2 - 2}$.

Example 4: Two samples of sodium vapour bulbs were tested for length of life and the following results were got:

	Size	Sample mean	Sample S.D.
Type I	8	1234 hrs	36 hrs
Type II	7	1036 hrs	40 hrs

Is the difference in the means significance to generalize that Type I is superior to Type II regarding length of life?

Solution: Null hypothesis: H_0: $\mu_1 = \mu_2$ i. e. two types of bulbs have same lifetime.

Alternative hypothesis $H_1: \mu_1 > \mu_2$ i. e. Type I is superior than Type II.

Hence we use right tailed test.

We know that

$$t = \frac{\overline{x_1} - \overline{x_2}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \qquad (1)$$

where

$$S^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} = \frac{8(36)^2 + 7(40)^2}{8+7-2} = 1659.076$$

$$\therefore S = 40.7317$$

From eq. (1)

$$t = \frac{1234 - 1036}{40.7317\sqrt{\frac{1}{8} + \frac{1}{7}}} = 18.1480$$

$$d.f. = n_1 + n_2 - 2 = 13$$

$\therefore t_{0.05}$ at 13 $d.f.$ is 1.77.

Since calculated value of $|t| >$ tabulated value at 5% level of significance.

$\therefore H_0$ is rejected. Hence Type I is superior than Type II.

$\Longrightarrow$ There are discriminate between two horses at 5% level of significance.

Example 6: The height of 6 randomly chosen sailors in inches are 63, 65, 68, 69, 71, and 72. Those of 9 randomly chosen soldiers are 61, 62, 65, 66, 69, 70, 71, 72 and 73. Test whether the sailors are on the average taller than soldiers.

Solutions: Let x_1 and x_2 be the two samples denoting the heights of sailors and soldiers.

$$n_1 = 6, \qquad n_2 = 9$$

Null hypothesis $H_0: \mu_1 = \mu_2$ i.e. the mean of the population are the same.

Alternative hypothesis $H_1: \mu_1 > \mu_2$ (one tailed test)

We know that $t = \frac{\bar{x}_1 - \overline{x_2}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$, where $S^2 = \frac{\sum(x_1 - \overline{x_1})^2 + \sum(x_2 - \overline{x_2})^2}{n_1 + n_2 - 2}$

$$\overline{x_1} = \frac{\sum x_1}{n_1} = 68, \qquad \overline{x_2} = \frac{\sum x_2}{n_2} = 67.6$$

Calculation of two sample means-

x_1	$x_1 - \overline{x_1}$	$(x_1 - \overline{x_1})^2$	x_2	$x_2 - \overline{x_2}$	$(x_2 - \overline{x_2})^2$
63	-5	25	61	-6.66	44.36
65	-3	9	62	-5.66	32.035
68	0	0	65	-2.66	7.0756
69	1	1	66	1.66	2.7556
71	3	9	69	1.34	1.7956

72	4	16	70 71 72 73	2.34 3.34 4.34 5.34	5.4756 11.1556 18.8356 28.5156
		$\sum(x_1 - \overline{x_1})^2 = 60$			$\sum(x_2 - \overline{x_2})^2 = 152.0002$

From (1) eq. $S = 4.038$

$$\Rightarrow t = 0.1569.$$

The value of t at 5% level of significance for 13 $d.f.$ is 1.77.

Since $t_{cal} < t_{0.05} = 1.77$, the null hypothesis H_0 accepted. The sailors are not on the average taller than the soldier.

Practice Questions

1. The manufacturer of certain make of LED bulb claims that his bulbs have a mean life of 20 months. A random sample of 7 such bulbs in months 19, 21, 25, 16, 17, 14, 21. Can you regard the producer's claims to be validate 1% level of significance?
2. A sample of 20 items has mean 42 units and S.D. 5 units. Test the hypothesis that it is a random sample from a normal population with mean 45 units.
3. The 9 items of a sample have the following values:
   45, 47, 50, 52, 48, 47, 49, 53, 51.
   Does the mean of these values differ significantly from the assumed mean 47.5 ?
4. Ten individuals are chosen at random from a normal population of students and their marks are found to be 63, 63, 66, 67, 68, 69, 70, 70, 71, 71. In the light of these data, discuss the suggestion that mean marks of population of students is 66.
5. Samples of sizes 10 and 14 were taken from two normal populations with S.D. 3.5 and 5.2 the sample mean were found to be 20.3 and 18.6. Test whether the means of two populations are the same at 5% level.
6. The mean life of 10 electric motors was found to be 1450 hrs with S.D. 423 hrs. A second sample of 17 motors chosen from a different batch showed a mean life of 1280 hrs with a S.D. of 398 hrs. Is there a significance difference between mean of two samples?
7. The marks obtained by a group of 9 regular course students and another group of 11 part time course students in the a test are given bellow:

Regular	56	62	63	54	60	51	67	69	58		
Part time	62	70	71	62	60	56	75	64	72	68	66

   Examine whether the marks obtained by regular students and part time students differ significantly at 5% and 1% level of significance.
8. The average number of articles produced by two machines per day are 200 and 250 with standard deviation 20 and 25 respectively on the basis of records of 25 days production. Can you regard both the machines equally efficient at 5% level of signif

5.4 Snedecor's Variance Ratio Test or F-test

This test is known as Fisher's F-test or simply F-test. F-test refers to a test of hypothesis concerning two variances derived from two samples.

Let n_1 and n_2 be sizes of two samples with variances s_1^2 and s_2^2. Then we define the variance ratio F as

$$F = \frac{s_1^2}{s_2^2}, \quad s_1^2 > s_2^2 \quad \{\text{F-statics is always greater than 1}\}$$

where, $s_1^2 = \frac{\sum(x_1 - \bar{x}_1)^2}{n_1 - 1}$, $s_2^2 = \frac{\sum(x_2 - \bar{x}_2)^2}{n_2 - 1}$

Degree of freedom $v_1 = n_1 - 1$, $v_2 = n_2 - 1$.

Assumptions-

1. The populations for each sample must be normally distributed.
2. The sample must be random and independent.
3. The ratio of σ_1^2 to σ_2^2 should be equal to 1 or greater than 1. That is why we take the larger variance in the numerator of ratio.

Applications

F- test is used to test

1. Whether two independent samples have been drawn from the normal populations with same variance σ^2.
2. Whether the two independent estimates of the population variance are homogeneous or not.

Working process

- Null hypothesis H_0: $\sigma_1^2 = \sigma_2^2$
- Alternative hypothesis: $\sigma_1^2 \neq \sigma_2^2$
- Calculation of F-statics :

$$F = \frac{s_1^2}{s_2^2}, \text{ where } s_1 > s_2 \text{ or } F = \frac{s_2^2}{s_1^2}, \text{ where } s_2 > s_1$$

- Find F_{tab} at α% level of significance with d.f. v_1 and v_2.
- Decision-

  If $F_{cal} < F_{tab}$, accept H_0.

  If $F_{cal} \not< F_{tab}$, reject H_0.

Example 1: The random samples are drawn from two populations and the following results we obtained:

Sample x	20	16	26	27	23	22	18	24	25	19		
Sample y	27	33	42	35	32	34	38	28	41	43	39	37

Find variance of two populations and test whether two samples have same variance (Given $F_{0.05}$ for 11 and 9 d.f. is 3.112)

Solution: Null hypothesis H_0: Let $\sigma_1^2 = \sigma_2^2$ i.e. two sample have same variance.

Alternative hypothesis H_1: $\sigma_1^2 \neq \sigma_2^2$.

Calculation of F −statitistic:

x	$x - \bar{x}$	$(x - \bar{x})^2$	y	$y - \bar{y}$	$(y - \bar{y})^2$
20	-2	4	27	-8	64

16	-6	36	33	-2	4
26	4	16	42	7	49
27	5	25	35	0	0
23	1	1	32	-3	9
22	0	0	34	-1	1
18	-4	16	38	3	9
24	2	4	28	-7	49
25	3	9	41	6	36
19	-3	9	43	8	64
			30	-5	25
			37	2	4
220	0	120	420	0	314

$n_1 = 10, n_2 = 12,$

Degree of freedom $v_1 = n_1 - 1 = 9, v_2 = n_2 - 1 = 11$

$\bar{x} = \frac{\sum x}{n_1}, \quad \bar{y} = \frac{\sum y}{n_2}$

$$s_1^2 = \frac{\sum (x - \bar{x})^2}{n_1 - 1} = \frac{120}{9} = 13.3$$

$$s_2^2 = \frac{\sum (y - \bar{y})^2}{n_2 - 1} = \frac{314}{11} = 28.5$$

Hence,

$$F = \frac{s_2^2}{s_1^2} = \frac{28.5}{13.3} = 2.14, \qquad \because \ s_2 > s_1$$

The tabulated value of F at 5% level of significance for the d.f. 11 and 9 is 3.112 i.e. $F_{0.05} = 3.112$.

The calculated value of $F = 2.14 <$ tabulated value of $F_{0.05} = 3.112$

$\therefore$ The null hypothesis is accepted. The two samples have same variance.

Example 2: Two random samples drawn from 2 normal populations are as follows:

A	17	27	18	25	27	29	13	17
B	16	16	20	27	26	25	21	

Test whether the samples are drawn from the same normal population.

Solution: To test if two independent samples have been drawn from same population we have to test

(i) Equality of means by applying t −test
(ii) Equality of population variance by applying F −test.

F-test :

Null hypothesis H_0: $\sigma_1^2 = \sigma_2^2$ i.e. the population variance do not differ significantly.

Alternative hypothesis H_1: $\sigma_1^2 \neq \sigma_2^2$

Calculation for S_1^2 and S_2^2

x_1	$x_1 - \overline{x_1}$	$(x_1 - \overline{x_1})^2$	x_2	$x_2 - \overline{x_2}$	$(x_2 - \overline{x_2})^2$
17	-4.625	21.29	16	-2.714	7.365
27	5.735	28.89	16	-2.714	7.365
18	-3.625	13.14	20	1.286	1.653
25	3.375	11.39	27	8.286	68.657
27	5.735	28.89	26	7.286	53.085
29	7.735	54.39	25	6.286	39.513
13	-8.625	74.39	21	2.286	5.226
17	-4.625	21.39			
		$\sum(x_1 - \overline{x_1})^2 =$ 253.87			$\sum(x_2 - \overline{x_2})^2 =$ 182.859

$$n_1 = 8, \ \ n_2 = 7$$

$\overline{x_1} = 21.625, \quad \overline{x_2} = 18.714$

$$s_1^2 = \frac{\sum(x_1 - \overline{x_1})^2}{n_1 - 1} = \frac{253.87}{7} = 36.267$$

$$s_2^2 = \frac{\sum(x_2 - \overline{x_2})^2}{n_2 - 1} = \frac{182.859}{6} = 30.47$$

$$F = \frac{s_1^2}{s_2^2} = 1.190.$$

The table value of F for $v_1 = 7$ and $v_2 = 6$ d.f. at 5% level is 4.21.

Since $F_{cal} < F_{tab}$, $\therefore$ null hypothesis is accepted. Hence the variability in two populations is same.

t-test:

Null hypothesis H_0: $\mu_1 = \mu_2$ i.e. the population means are equal.

Alternative hypothesis H_1: $\mu_1 \neq \mu_2$

We know that

$$t = \frac{\bar{x}_1 - \overline{x_2}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

where,

$$S^2 = \frac{\sum(x_1 - \overline{x_1})^2 + \sum(x_2 - \overline{x_2})^2}{n_1 + n_2 - 2} = \frac{253.87 + 182.859}{8 + 7 - 2} = 33.594$$

$$\Longrightarrow S = 5.796$$

$$\therefore t = 0.9704$$

Degree of freedom $n_1 + n_2 - 2 = 13$.

The tabulated value of t at 5% level of significance for 13 d.f. is 2.16. The calculated value of t less than tabulated value. H_0 is accepted. There is no significant difference between population mean i.e. $\mu_1 = \mu_2$.

Therefore two sample have been drawn from the same population.

Practice Questions

1. In laboratory experiment, two sample gave the following results:

Sample	Size	Sample mean	Sum of squares of deviation from mean
I	10	15	90
II	12	14	108

Test whether the sample come from same normal population.

2. Daily wages in Rupees of skilled workers in two cities are as follows:

	Size of sample of workers	S.D. of wages in the sample
City A	16	25
City B	13	32

Test whether the sample come from same normal population.

3. Two independent samples of size 8 and 9 had the following values of the variables:

Sample I	20	30	23	25	21	22	23	24	
Sample II	30	31	32	34	35	29	28	27	26

Do the estimates of the population variance differ significantly?

4. The standard deviation calculated from two random samples of sizes 9 and 13 are 2:1 and 1:8 respectively. Can the samples be regarded as drawn from normal populations with the same standard deviation?

5.5 Chi-Square χ^2 Test

The Chi-square test is very powerful test for testing the significance of the discrepancy between actual (or observed) frequencies and theoretical (or expected) frequencies. If O_i $(i = 1, 2, \ldots . n)$ is the set of observed frequencies and E_i $(i = 1,2, \ldots n)$ is the corresponding set of expected frequencies, then χ^2 is defined as

$$\chi^2 = \sum_{i=1}^{n} \left[\frac{(O_i - E_i)^2}{E_i} \right],$$

where $\sum O_i = \sum E_i = N$ (total frequency)

Application of Chi-square test

(1) **Test of independents of attributes –** With the help of chi-square test we can find out whether two or more attributes associated or not.

(2) **Test of goodness of fit-** On several occasions the decision makers need to understand whether an actual sample distribution matches with known probability distribution. Such as poison, binomial or normal.

(3) Compare a number of frequency distributions.
(4) Test for specified standard deviation i.e. it may be used to test of population variance.

Conditions for applying χ^2 Test

(1) Each cell should contain at least 5 observation.
(2) The members of sample should be independent.
(3) Constrains on the cell frequencies should be linear.
(4) Total frequencies N should be reasonably large, say greater than 50.

Working Rule

Step 1- Consider the null hypothesis and alternative hypothesis.

Step 2- Calculate the expected frequency E_i corresponding to each cell.

Step 3- Calculate χ^2 by the formula

$$\chi^2 = \sum_{i=1}^{n} \left[\frac{(O_i - E_i)^2}{E_i}\right]$$

and calculate dof.

Step 4- See the value of χ^2 from the table i.e. value of χ^2 at α% level of significance and for dof v, as calculated in step 3.

Step 5- Decision:

(i) If calculated value of $\chi^2 <$ tabulated value of $\chi^2_{\alpha,v}$. Then accept the null hypothesis.

(ii) If calculated value of $\chi^2 \not<$ tabulated value of $\chi^2_{\alpha,v}$. Then reject the null hypothesis i.e. accept the alternative hypothesis.

χ^2 Test as a test of goodness of fit

Example 1: A die is thrown 276 times and the results of these throws are given bellow:

No. appeared on the die	1	2	3	4	5	6
Frequency	40	32	29	59	57	59

Test whether the die is biased or not.

Solution: Null hypothesis H_0: Die is unbiased.

The expected frequency for each digit is $\frac{276}{6} = 46$.

O_i	40	32	29	59	57	59
E_i	46	46	46	46	46	46
$(O_i - E_i)^2$	36	196	289	169	121	169

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{980}{46} = 21.30$$

The tabulated value of χ^2 at 5% level of significance for $(6-1) = 5$ dof is 11.07.

Since calculated value of $\chi^2 >$ the tabulated value.

$\therefore H_0$ is reject. That is, die is not unbiased or die is biased.

Example 2: In experiment on pea breading, the following frequencies of seeds were obtained:

Round and yellow	Wrinkled and yellow	Round and green	Wrinkled and green	Total
315	101	108	32	556

Theory predicts that the frequencies should be in proportions 9:3::3:1.
Examine the correspondence between theory and experiment.

Solution: Null hypothesis H_0: There is correspondence between theory and experiment or there is no significance difference between observed and theoretical frequency.

Calculation of expected frequencies-

The given frequencies in proportions 9:3:3:1

Total sum of proportions =9+3+3+1=16.

Expected frequency for round and yellow seed is $E_1 = \frac{9}{16} \times 556 = 312.75$

Expected frequency for wrinkled and yellow seed $E_2 = \frac{3}{16} \times 556 = 104.25$

Expected frequency for round and green seed $E_3 = \frac{3}{16} \times 556 = 104.25$

Expected frequency for wrinkled and green seed is $E_4 = \frac{1}{16} \times 556 = 34.75$

To calculate the value of χ^2:

Observed frequency O_i	315	101	108	32
Expected frequency E_i	312.75	104.25	104.25	34.75
$\frac{(O_i - E_i)^2}{E_i}$	0.016187	0.101319	0.134892	0.217626

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 0.470024$$

The tabular value of χ^2 at 5% level of significance for $n - 1 = 3\ d.f.$ is 7.815 i.e. $\chi^2_{0.05} = 7.815$

Since calculated value of $\chi^2 <$ the tabulated value.

$\therefore H_0$ accepted. i.e. experimental result support the theory.

Example 3: When the first proof of 392 pages of a book of 1200 pages were read, the distribution of printing mistakes were found to be as follows:

Mistakes per page (x)	Observed frequency O_i	Expected frequency (E_i)	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
0	275	242.1	1082.41	4.471
1	72	116.7	1998.09	17.121
2	30	28.1	3.61	0.128
3	7	4.5		
4	5	0.5		
5	2	0.1	98.01	19.217
6	1	0		
Total	392	392		40.937

No. of mistakes in a page (x)	0	1	2	3	4	5	6
No. of pages (f)	275	72	30	7	5	2	1

Fit a poisson distribution to the above data and test the goodness of fit.

Solution: Null hypothesis- Poisson distribution is a good fit to the data

Mean $(\lambda) = \frac{\sum fx}{\sum f} = \frac{189}{392} = 0.4821$

The frequency of x mistakes per page is given by the poisson law as follow:

$$N(x) = NP(x)$$

$$= \frac{392[e^{-0.4821}(0.4821)^x]}{x!}$$

$$= \frac{242.05(0.4821)^x}{x!}$$

Expected frequencies are

$$N(0) = 242.05, N(1) = 116.69, N(2) = 28.13, N(3) = 4.52, N(4) = 0.54, N(5) = 0.052, N(6) = 0.0042$$

χ^2 table is as follows:

$$\chi^2_{\text{cal}} = \sum \frac{(O_i - E_i)^2}{E_i} = 40.937$$

$$d.f. = 7 - 1 - 1 - 3 = 2$$

$$\chi^2_{\text{tab},0.05} = 5.991$$

Since $\chi^2_{\text{cal}} > \chi^2_{\text{tab}}$, null hypothesis is rejected at 5% level of significance.

∴ Poisson distribution is not a good fit to the given data.

Example 4: The demand for a particular spare part in a factory was found to vary from day to day. In a sample study, the following information was obtained

Days	Mon	Tue	Wed	Thurs	Fri	Sat
No. of parts demanded	1124	1125	1110	1120	1126	1115

Test the hypothesis that the no. of parts demanded does not depend on the day of the week.

Solution: Let the no. of parts demanded are uniformly distributed i.e. no. of parts demanded does not depend on the day of the week.

Expected no. of parts on each day of week = $\frac{1124+\cdots+1115}{6} = 1120$

Days	No. of parts	E_i	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
Mon	1124	1120	16	0.014285
Tue	1125	1120	25	0.22321
Wed	1110	1120	100	0.08928
Thurs	1120	1120	0	0
Fri	1126	1120	36	0.03214
Sat	1115	1120	25	0.02232
				0.180346

$$\Rightarrow \chi^2_{\text{cal}} = \sum \frac{(O_i - E_i)^2}{E_i} = 0.18034$$

$$d.f. = 6 - 1 = 5$$

Tabulated value of χ^2 for 5 $d.f.$ at 5% level of significance is 11.07.

Since $\chi^2_{\text{cal}} < \chi^2_{\text{tab}}$.

$\Rightarrow H_0$, the null hypothesis is accepted.

χ^2 test as a test of independent-

With the help of χ^2 test we can find whether or not, two attributes are associated. We take the null hypothesis that there is no association between the attributes under study, i.e.

H_0: Given attributes are independent.

Let us consider two attributes A and B, A divided into r classes $A_1, A_2, \ldots, A_r$ and B divided into s classes $B_1, B_2, \ldots, B_s$. Then contingency table $r \times s$ is given as follows:

A / B	A_1	A_2	A_3	$\ldots A_r$	Total
B_1	(A_1B_1)	(A_2B_1)	(A_3B_1)	$\ldots (A_rB_1)$	(B_1)
B_2	(A_1B_2)	(A_2B_2)	(A_3B_2)	$\ldots (A_rB_2)$	(B_2)
B_3	(A_1B_3)	(A_2B_3)	(A_3B_3)	$\ldots (A_rB_3)$	(B_3)
…..	…..	…..	…..	…..	…..
…..	…..	…..	…..	…..	…..
B_s	(A_1B_s)	(A_2B_s)	(A_3B_s)	$\ldots (A_rB_s)$	(B_s)
Total	(A_1)	(A_1)	(A_1)	$\ldots(A_1)$	N

The expected frequency

$$(A_iB_j)_o = \frac{(A_i)(B_j)}{N}$$

i.e. expected frequency in each cell is

$$= \frac{\text{Product of total column and row total}}{\text{whole total}}$$

Hence,

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\left[\frac{\left[(A_iB_j)-(A_iB_j)_o\right]^2}{(A_iB_j)_o}\right]$$

Degree of freedom $(r-1)(s-1)$

Example 5: From the following table regarding the color eyes of father and son, test if color of son's eye is associated with that of father.

		Eye color of son	
		Not Light	Light
Eye color of father	Not Light	230	148
	Light	151	471

Solution: Null Hypothesis H_0- The color of son's eyes is not associated with the color of father's eyes, i.e. they are independent.
Given observed frequencies are

	Not light	Light	Total
Not light	230	148	378
Light	151	471	622
Total	381	619	1000

Expected frequencies are

	Not light	Light	Total
Not light	$\frac{378\times 381}{1000} = 144$	$\frac{378\times 619}{1000} = 234$	378
Light	$\frac{381\times 622}{1000} = 237$	$\frac{619\times 622}{1000} = 385$	622
Total	381	619	1000

Calculation of χ^2 statics:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\left[\frac{\left[(A_iB_j)-(A_iB_j)_o\right]^2}{(A_iB_j)_o}\right]$$

$$= \frac{(230-144)^2}{144} + \frac{(148-234)^2}{234} + \frac{(151-237)^2}{237} + \frac{(471-385)^2}{385} = 133.29$$

Tabulated value of χ^2at 5% level, for 1 d.f. is 3.841

Since $\chi^2_{\text{cal}} > \chi^2_{\text{tab}}$

$\therefore H_0$ is rejected. i.e. they are not independent, the color eye of son's eye is associated with that of father.

Example 6: To test the effectiveness of inoculation against cholera, the following table was obtained:

	Attacked	Not attacked	Total
Inoculated	30	160	190
Not inoculated	140	460	600
Total	170	620	790

(The figures represent the no. of persons)

Use χ^2-test to defend or refute the statement that the inoculation prevents attack from choler

Solutions: Null hypothesis H_0 −The inoculation does not prevent attack from cholera.

Expected frequencies are

	Attacked	Not attacked
Inoculated	$\frac{190 \times 170}{790} = 40.886$	$\frac{190 \times 620}{790} = 149.11$
Not inoculated	$\frac{600 \times 170}{790} = 129.11$	$\frac{600 \times 620}{790} = 470.89$

$$\chi^2_{\text{cal}} = \frac{(30-40.886)^2}{40.886} + \frac{(160-149.11)^2}{149.11} + \frac{(140-129.11)^2}{129.11} + \frac{(460-470.89)^2}{470.89} = 4.863$$

Tabulated value of χ^2 at 5% level of significance for 1 d.f. is 3.841.

Since $\chi^2_{\text{cal}} > \chi^2_{\text{tab}}$ at 5% level of significance.

$\therefore H_0$ is rejected. Hence inoculation prevents attack from cholera.

Example 7: The following table gives the no. of good and bad parts produced by each of three shift in a factory.

	Good parts	Bad parts	Total
Day shift	960	40	1000
Evening shift	940	50	990
Night shift	950	45	995
Total	2850	135	2985

Test whether or not the production of bad parts is independent of the shift on which they were produced.

Solution: Null hypothesis H_0 : Production of bad part is independent of the shift.

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{3}\frac{\left[(A_iB_j)_0-(A_iB_j)\right]^2}{(A_iB_j)_0}$$

(1)

Now,

$$(A_1B_1)_0 = \frac{(A_1)(B_1)}{N} = \frac{2850\times 1000}{2985} = 954.77$$

$$(A_1B_2)_0 = \frac{(A_1)(B_2)}{N} = \frac{2850\times 990}{2985} = 945.226$$

$$(A_1B_3)_0 = \frac{(A_1)(B_3)}{N} = \frac{2850\times 995}{2985} = 950$$

$$(A_2B_1)_0 = \frac{(A_2)(B_1)}{N} = \frac{135\times 1000}{2985} = 45.27$$

$$(A_2B_2)_0 = \frac{(A_2)(B_2)}{N} = \frac{135\times 990}{2985} = 44.773$$

$$(A_2B_3)_0 = \frac{(A_2)(B_3)}{N} = \frac{135\times 995}{2985} = 45.$$

To calculate the value of χ^2

Class	O_i	E_i	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
(A_1B_1)	960	954.77	27.3529	0.02864
(A_1B_2)	940	945.226	27.3110	0.02889
(A_1B_3)	950	950	0	0
(A_2B_1)	40	45.27	27.7729	0.61349
(A_2B_2)	50	44.773	27.3215	0.61022
(A_2B_3)	45	45	0	0
				1.28126

$$\chi^2_{\text{cal}} = 1.28126.$$

$$d.f. = (r-1)(s-1) = (2-1)(3-1) = 2$$

∴ The tabulated value of χ^2 at 5% level of significance for 2 d.f. is 5.991.

Since 5% H_0 is

	Yes	No	Total
Yes	56	31	87
No	18	6	24
Total	74	37	111

$\chi^2_{\text{cal}} < \chi^2_{\text{tab}}$ at level of significance.∴ accepted. The production of bad part is independent of the shift on which they were produced.

Example 8: In a sample survey of public opinion, answers to questions (i) Do you drink? (ii) Are you in favor of local option on sale of liquor? are tabulated below:

Can you infer whether or not the local option on the sale of liquor is dependent on individual drink ? (Given that the value of $\chi^2$for 1 dof at 5% level of significance is 3.841)

Solution: Null hypothesis H_0: The option on sale of liquor is not dependent with individual drinking.

Calculation of expected frequencies:

$$E_{11} = \frac{74 \times 87}{111} = 58$$

$$E_{12} = \frac{37 \times 87}{111} = 29$$

$$E_{21} = \frac{74 \times 24}{111} = 16$$

$$E_{21} = \frac{37 \times 24}{111} = 8$$

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

We have

O_{ij}	E_{ij}	$\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$
56	58	4/58
31	29	4/29
18	16	4/6
6	8	4/8
Total		0.957

Hence the value of $\chi^2 = 0.957$

Also the degree of freedom $\nu = (2-1)(2-1) = 1$

The tabulated value of χ^2 at 5% level of significance for 1 d.f. is 3.841.

$\because \chi^2_{\text{cal}} < \chi^2_{\text{tab}} \Longrightarrow H_0$ is accepted.

$\Rightarrow$ Sale of liquor is not dependent or not associated with the individual drinking.

Practice Questions

1. Out of 300 customers, we find 88 prefer brown color, 65 prefer grey, 52 prefer red, 40 prefer blue and 55 prefer white. Test the hypothesis that all color are equally popular.
2. 2. A survey of 320 families with 5 children shows the following distribution:

No. of boys	5	4	3	2	1	0
No. of girls	0	1	2	3	4	5
No. of families	14	56	110	88	40	12

Is this result consistent with the hypothesis that the male and female birth are equally probable? (Given $\chi^2_{0.05} = 10.07$).

3. Consider the hypothetical experiment on the effect of smoking on divorce to find is there is any relationship between them

	Divorced	Not divorced	Total
Smoking	73	12	85
Not smoking	43	39	82
Total	116	51	167

4. The following table shows the distribution of digit in numbers chosen at random from the telephone directory:

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

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

5. The following data is collected on two characters:

	Smokers	Non smokers
Literate	83	57
Illiterate	45	68

Based on this information can you say that there is no relation between habit of smoking and literacy?

6. By using χ^2, find out whether there is any association between income level and type of schooling:

Income	Public School	Govt. School
Low	200	400
High	1000	400

Z test

Test of significance for single Mean:

To test whether the difference between population mean is significant or not.

Let $X_1, X_2 \ldots\ldots X_n$ be a random sample of size N with mean μ and variance σ^2. The standard error of mean of a random sample of size n from a population with variance σ^2 is $\frac{\sigma}{\sqrt{n}}$

The test statistic is $Z = \frac{x-\mu}{\frac{\sigma}{\sqrt{n}}}$ where σ is standard deviation

- Question 1. A random sample of 900 members has a mean 3.4 cms. Can it be reasonably regarded as a sample from a large population of mean 3.2
- cms and S.D 2.3 cms.

Sol: Null Hypothesis: Assume that the sample is drawn from a large population with mean 3.2 and S. D 2.3

$$Z = \frac{x-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{3.4-3.2}{2.3/\sqrt{900}} = 0.261$$

$$|Z| = 0.261 < 1.96$$

Null hypothesis is accepted.

Question 2 : Intelligence tests were given to two groups of boys and girls

Girls	75	8	60
Boys	73	10	100

Examine if the difference between means scores is significant

Solution: Null Hypothesis : There is no significant difference between mean scores.

Test statistic: under the null hypothesis:

$$z = \frac{\overline{x_1}-\overline{x_2}}{\sqrt{\frac{{s_1}^2}{n_1}+\frac{{s_2}^2}{n_2}}} = \frac{75-73}{\sqrt{\frac{64}{60}+\frac{100}{100}}} = 1.3912$$

As the calculated value of $|Z| < 1.96$ Null hypothesis is accepted i.e there is no significant difference between mean scores.

Test of Significance for Difference of Means of Two Large Samples

Let $\overline{x_1}$ be the mean of a sample of size $n_1$from a population with mean μ_1, and variance σ_1^2.

Let $\overline{x_2}$ be the mean of an independent sample of size n_2 from another population with mean μ_2 and variance σ_2^2. The test statistic is given by $z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

Under the null hypothesis that the samples are drawn from the same population where $\sigma_1 = \sigma_2 = \sigma$ i.e. $\mu_1 = \mu_2$the test statistic is given by $z = \frac{\overline{x_1} - \overline{x_2}}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

Note 1. If $\sigma_1, \sigma_2$are not known and $\sigma_1 \neq \sigma_2$ the test statistic in this case is $z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

Note 2. If σ is not known and $\sigma_1 = \sigma_2$. we use $\sigma^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}$ to calculate σ.

Example 1. The average income of persons was ₹ 210 with a S.D. of ₹ 10 in sample 100 people of a city. For another sample of 150 persons, the average income was ₹ 220 with S.D. of ₹ 12 . The S.D. of incomes of the people of the city was ₹ 11. Test whether there is any significant difference between the average incomes of the localities.

Sol.

Here $n_1 = 100, n_2 = 150, \overline{x_1} = 210, \overline{x_2} = 220, s_1 = 10, s_2 = 12.$

Null hypothesis: The difference is not significant.i.e. there is no difference between the incomes of the localities. $H_0: \overline{x_1} = \overline{x_2}$

Alternative hypothesis

$H_1: \overline{x_1} = \overline{x_2}$ (two tailed test)

Test statistic:

Under H_0, $z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{210 - 220}{\sqrt{\frac{10^2}{100} + \frac{12^2}{150}}} = -7.1428 \therefore |z| = 7.1428$

Conclusion: As the calculated value of $|z| > 1.96$, the significant value of z at 5% level of significance,H_0 is rejected i.e., there is significant difference between the average income of the localities.

Exercise:

1. Two random samples of sizes 1000and 2000 Farms gave an average yield of 2000kg and 2050kg respectively . The variance of wheat farms in the country may be taken as 100 kg. Examine whether the two samples differ significantly in yield.

Answer Highly Significantly.

2. The means of two large samples of 1000 and 2000 members are 168.75 cms and 170 cms respectively. Can the samples be regarded as drawn from the same population of standard deviation 6.25 cms?.
3. In a survey of buying habbits , 400 women shoppers are chosen at random in supermarket A. Their average weekly food expenditure is Rs. 250 with a S.D of Rs 40. For 500 women shoppers chosen at supermarket B, the average weekly food expenditure is Rs. 220 with a S.D of Rs. 45. Test at 5% level of significance whether the average food expenditure of the two groups are equal.

   Answer: Highly Significant
4. The yield of wheat in a random sample of 1000 farms in a certain area has a S.D of 192 kg.Another random sample of 1000 farms gives a S.D of 224 kg. Are the S.D's significantly different?

   Answer: z=4.851. The S.D are significantly different

ANOVA

Definition: ANOVA is a statistical technique which can be used to make comparisons among more than two groups. Let us suppose we want to test the effectiveness of various doses of a drug on blood pressure . We take a large number of individuals and divide in to the number of groups depending upon thee number of doses. For example if we want to test four different doses we make four groups (one control group and four experimental group). Each experimental group is meant for a given dose of the drug.

One way ANOVA is used to see the effects of an independent variable on a given dependent Variable .

In ANOVA we compare the variations existing in the observed value of the dependent variable in the classified groups. We calculate the variance of the total population, variance between the groups and the variance within the groups.

ANOVA Table: The technique of analysis of variance is referred to as ANOVA. A table showing the source of variation , the sum of squares, degree of freedom, mean square and the formula for F ratio is known as ANOVA table.

Assumptions: Following are the assumptions for the study of analysis of variance

(i) Each of the sample is a simple random sample.

(ii) Population, from which the samples are selected are normally distributed

(iii) Each of the sample is independent of the other sample.

One way classification:

(a) **NULL Hypothesis:** the means of the populations from which p samples are drawn are equal to one another.

The various sum of squares involved in the computation of F statistic are:

(i) **SUM of the squares of variations amongst the columns(SSC):**It is the sum of squares of deviation between the columns.

Variance amongst columns=$MSC = \frac{SSC}{c-1}$, where c is the number of columns. (ii) Sum of squares of Variations within columns (SSE): It is the sum of the square of variation between individual items and the columns means.

Mean of square of columns errors=$MSE = \frac{SSE}{c(r-1)}$, where c is the number of columns and r is the number of rows. (iii) **Total sum of squares of variation** (SST): SST=SSC+SSE

Total Variance= $\frac{SST}{n-1}$

Test Statistic: F statistic= $\frac{MSC}{MSE}$ with $(c-1)$and $c(r-1)$degree of freedom.

Source of variation	Sum of square	Degree of Freedom	Mean of sum square	Variance ratio of F
Between samples	SSC	c-1	$MSC = \frac{SSC}{c-1}$	F statistic= $\frac{MSC}{MSE}$
Within samples	SSE	c(r-1)	$MSE = \frac{SSE}{c(r-1)}$	
Total	SST	cr-1	-	-

Question1: It is desired to compare three hospitals with regards to the no. of deaths per month. A sample of death records were selected from the records of each hospital and number of deaths was a given below . From these data suggest a difference in the no. of the deaths per month among three hospitals

A	B	C
3	6	7
4	3	3
3	3	4
5	4	6
0	4	5

Null Hypothesis: There is no difference in the no. of deaths.

Total no of sample space

$\sum y_A = 3 + 4 + 3 + 5 + 0 = 15$

$\sum y_B = 6 + 3 + 3 + 4 + 4 = 20$

$\sum y_A = 7 + 3 + 4 + 5 + 5 = 25$

Grand total=$\sum y_A + \sum y_B + \sum y_C = 60$

There are hospitals =3

Each hospital have 5 samples, so total sample =

$5 \times 3 = 15$

Correction factor =(C.F.)=$\frac{G.T.^2}{n} = \frac{(60)^2}{15} = 240$

Sum of square of sample

$\sum {y_A}^2 = 9 + 16 + 9 + 25 + 0 = 59$

$\sum {y_B}^2 = 36 + 9 + 9 + 16 + 16 = 86$

$\sum {y_C}^2 = 49 + 9 + 16 + 36 + 25 + 25 = 135$

Total sum: $\sum {y_A}^2 + \sum {y_B}^2 + \sum {y_C}^2 - 240 = 40$

Sum of squares b/w samples=$\frac{(\sum y_A)^2}{n_1} + \frac{(\sum y_B)^2}{n_2} + \frac{(\sum y_C)^2}{n_3} - 240$

$= \frac{(15)^2}{5} + \frac{(20)^2}{5} + \frac{(25)^2}{5} - 240 = 10$

Sum of square within samples:

=Total sum of square – sum of squares b/w samples

$= 40 - 10$

$= 30$

Degree of freedom

I. For total sum of square $= n - 1 = 15 - 1 = 14$

II. For hospital $= k - 1 = 3 - 1 = 2$
III. For error $= n - k = 15 - 3 = 12$

ANOVA Table

Note: ANOVA table: A table showing the source of variance , the sum of squares , degree of freedom , mean square (variance) & the formula for F-ratio is known as ANOVA Table

Source of variation	Sum of square	Degree of Freedom	Mean of sum square	Variance ratio of F
b/w samples	10	2	5	F(2,12)=5/2.5=2
Within samples	30	12	2.5	
Total	40	14	-	-

$F_{tabular} = 3.89$
Conclusion: $F_{calculated} < F_{tabular}$ $so\ null\ hypothesis\ is\ accepted$

Question.2: A manufacturing company purchased three new machines of different makes and wishes to determine whether one of them is faster than others in producing a certain output , five hourly production figures are observed at random from each machine and results are given below

Observation	A_1	A_2	A_3
1	25	31	24
2	30	39	30
3	36	30	28
4	38	42	25
5	31	35	28

Solution:
Null Hypothesis: There is no significant difference in speed of machine

$$\sum y_{A_1} = 25 + 30 + 36 + 30 + 31 = 160$$
$$\sum y_{A_2} = 31 + 39 + 38 + 42 + 35 = 185$$
$$\sum y_{A_3} = 24 + 30 + 28 + 25 + 28 = 135$$

Grand total=$\sum y_A + \sum y_B + \sum y_C = 480$

Correction factor =(C.F.)=$\frac{G.T.^2}{n} = \frac{(480)^2}{15} = 15360$
$n = no.of\ machines \times observation = 3 \times 5 = 15$
Sum of square of sample

$$\sum {y_{A_1}}^2 = 625 + 900 + 1256 + 1444 + 961 = 5226$$
$$\sum {y_{A_2}}^2 = 961 + 1521 + 1444 + 1764 + 1225 = 6915$$
$$\sum {y_{A_3}}^2 = 576 + 900 + 784 + 625 + 784 = 3669$$

Total sum of squares: $\sum {y_{A_1}}^2 + \sum {y_{A_2}}^2 + \sum {y_{A_3}}^2 - C.F. = 15810 - 15360 = 450$

Sum of squares b/w samples=$\frac{(\sum y_{A_1})^2}{n_1}+\frac{(\sum y_{A_2})^2}{n_2}+\frac{(\sum y_{A_3})^2}{n_3}-C.F.$

$=\frac{(160)^2}{5}+\frac{(185)^2}{5}+\frac{(135)^2}{5}-15360=250$

Sum of square within samples:

=Total sum of square – sum of squares b/w samples

$=450-250$

$=20$

Degree of freedom

I. For total sum of square $=n-1=15-1=14$

II. For hospital $=k-1=3-1=2$

III. For error $=n-k=15-3=12$

ANOVA Table

Note: ANOVA table: A table showing the source of variance , the sum of squares , degree of freedom , mean square (variance) & the formula for F-ratio is known as ANOVA Table

Source of variation	Sum of square	Degree of Freedom	Mean of sum square	Variance ratio of F
b/w samples	250	2	$\frac{250}{2}=125$	F(2,12)=125/16.67=7.49
Within samples	200	12	$\frac{200}{12}=16.67$	
Total	450	14	-	-

$F_{tabular}=3.89$

Conclusion: $F_{calculated}>F_{tabular}$ *so null hypothesis is NOT accepted*

Questions:

1.Below are given the yield per kg for four varieties of tablets. Prepare ANOVA table and test that varieties differ Significantly.

A	B	C	D
20	25	24	23
19	23	20	20
21	21	22	20

Answer: No

2. To test the significance of the variations of the retail prices in the commodity in three principal cities: Mumbai, Bangalore and Chennai , the four shops were chosen at random in each city and prices observed in rupees were as follows:

Mumbai	16	8	12	14
Bangalore	14	10	10	16
Chennai	4	10	8	8

3. Do the data indicate that the prices in the three cities are significantly different? The following table gives the yields on 15 samples plots under three varieties of seeds:

Variety I:	20	21	23	16	20
Variety II:	18	20	17	15	25
Variety III:	25	28	22	28	32

Show that the seed varieties show variations more than could be covered by sampling variation

CHAPTER5
STATISTICAL QUALITY CONTROL

DEFINITION:

when statistical techniques are employed to control, improve and maintain quality or to solve quality problems. Building an information system to Satisfy the concept of prevention and control and improving upon product quality requires statistical thinking.

Advantages of Statistical Quality Control:

(1) Efficiency
(2) Reducing of scrap
(3) Easy detection of faults
(4) Increased output and reduced wasted machine and man hours
(5) Creates quality awareness in employees.

Control Chart

A control chart is a graphical representation of the collected information. In other words, control chart is a device which satisfies the state of statistical control or a device for attaining quality control or is a device for attaining quality control or is a device to judge whether the statistical control has attained.

Types of Control Chart:

There are many types of control chart designed for different control situations . Most commonly used control chart are

(1) Control Chart for variables: They are useful to measure quality characteristics and to control fully automatic process. It includes $\bar{X}$ and R charts for $\bar{X}$ and σ.
(2) Control Chart for Attributes: It includes P chart for fraction defective . A fraction defective control chart discloses erratic fluctuations in the quality of inspection which , may result in improvement in inspection practice and inspection standards.

Construction of Control chart FOR variables :

A random sample of size n is taken during a manufacturing process over a period of time and quality measurements $x_1, x_2, x_3 \ldots\ldots\ldots x_n$ are has been observed .

Sample mean $\bar{x} = \frac{x_1+x_2+\cdots\ldots\ldots+x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Sample Range $R = x_{max} - x_{min}$

If the process is found stable, k consecutive samples are selected and for each sample, $\bar{x}$ and R are calculated . Then we find $\bar{\bar{x}}$ and $\bar{R}$ as

$$\bar{R} = \frac{R_1+R_2+\cdots\ldots\ldots+R_k}{k} = \frac{1}{k}\sum_{i=1}^{k} R_I$$

For $\bar{X}$ – chart

Central line=$\bar{\bar{x}}$ when tolerance limits are not given

μ, when tolerance limits are given

$$\mu = \frac{1}{2}[\text{LCL+UCL}]$$

Now, LCL (for $\bar{X}$ -chart) =$\bar{\bar{x}} - A_2\ \bar{R}$

UCL (for $\bar{X}$ -chart) =$\bar{\bar{x}} + A_2\ \bar{R}$

A_2 depends on sample size n and can be found from the following table

For R chart Central line (CL)= $\bar{R}$

LCL (for R chart) =$D_3\bar{R}$

UCL (for R chart) =$D_4\bar{R}$

Where D_3 and D_4 depend on sample size and are found from the following table

Control Chart for Attributes

Following Control chart will be considered as control chart for Attributes

P chart (ii) np chart (iii) C chart

$$\bar{p} = \frac{No\ of\ defectives\ found\ \ in\ any\ inpection}{Total\ no.of\ articles\ atually\ inspected}$$

Control limits on p chart The control limits for p chart will be

CL= $\bar{p}$

Upper and lower limits for $p-$ chart are

$$UCL_p = \bar{p} + 3\sigma_p = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

$$LCL_p = \bar{p} - 3\sigma_p = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

Control limits for np chart : the actual number of defectives called np chart is used

Upper and lower limits for $np-$ chart are

$CL = n\bar{p}$ where $\bar{p} = \frac{\sum np}{\sum n}$

$$UCL_{np} = n\bar{p} + 3\sigma_{np} = n\bar{p} + 3\sqrt{n\bar{p}(1-\bar{p})}$$

$$LCL_{np} = n\bar{p} - 3\sqrt{n\bar{p}(1-\bar{p})}$$

Question1 The following are the mean heights and ranges of lengths of a finished product from 10 samples each of size 5. The specification limits are 200±5 cm. Construct $\overline{X}$ and R chart and examine whether the process is under control and state your recommendations.

Sample No.	1	2	3	4	5	6	7	8	9	10
Mean ($\overline{X}$)	201	198	202	200	203	204	199	196	199	201
Range (R)	5	0	7	3	3	7	2	8	5	6

Assume for n=5, $A_2 = 0.58$, $D_4 = 2.11$ *and* $D_3 = 0$

Solution : CL=200 ; UCL (for $\overline{X}$ -chart) =$\bar{\bar{x}} + A_2$ $\bar{R}$=$200 + (0.58 \times 4.6) = 202.668$

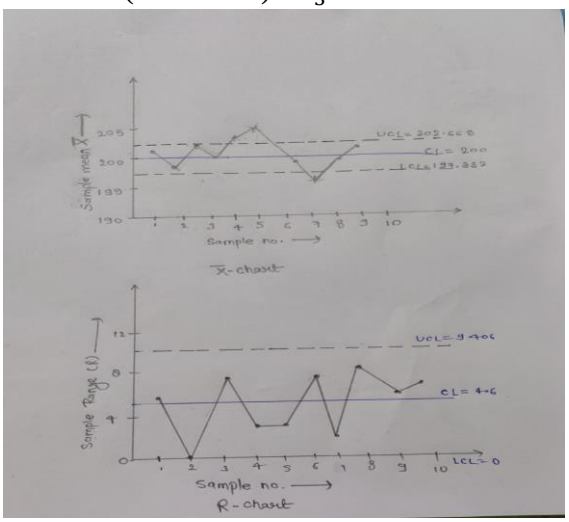
LCL (for $\overline{X}$ -chart) =$\bar{\bar{x}} - A_2$ $\bar{R}$=$200 - (0.58 \times 4.6) = 197.332$

Control limit for R chart:

(CL)= $\bar{R}$=4.6

UCL (for R chart) =$D_4\bar{R}$=$2.11 \times 4.6 = 9.706$

LCL (for R chart) =$D_3\bar{R}$=$0 \times 4.6 = 0$



it is observed that all points lie within the control limits on R chart .Hence the process is under control. But in Xchart, points corresponding to sample 5,6 and8 lie outside the control limits hence the process is not in statistical control.

Question2 . In a blade manufacturing factory , 1000 blades are examined daily . Draw the np chart for the following table and examine whether the process is under control?

Date	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

No. of defective blades	9	10	12	8	7	15	10	12	10	8	7	13	14	15	16

Solution: n=1000 $\sum np$= total number of defectives =165

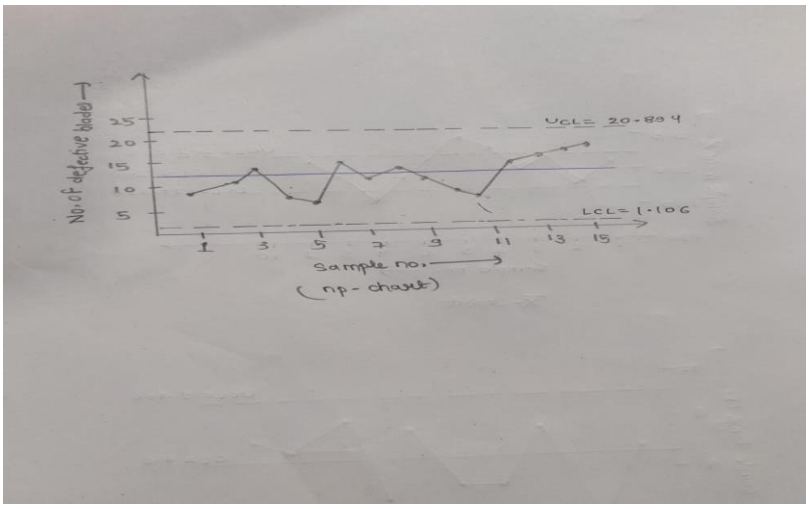
$$\sum n = total\ number\ of\ inspectd = 1000 \times 15$$

$$\bar{p} = \frac{\sum np}{\sum n} = \frac{166}{1000 \times 15} = 0.011$$

Control limits are $CL = n\bar{p}$ =$1000 \times 0.011 = 11$

$$UCL_{np} = n\bar{p} + 3\sigma_{np} = n\bar{p} + 3\sqrt{n\bar{p}(1-\bar{p})} = 11 + 3\sqrt{11(1-0.011)}$$
$$= 20.894$$

$LCL_{np} = n\bar{p} - 3\sqrt{n\bar{p}(1-\bar{p})}$== $11 - 3\sqrt{11(1-0.011)} = 1.106$



Since all points lie within the control limits , the process is under control.

Question3 . In a manufacturing process, the number of defectives found in the inspection of 20 lots of 100 samples is given below

Lot No.	No. of defectives	Lot no.	No.of defectives
1	5	11	7
2	4	12	6
3	3	13	3
4	5	14	5
5	4	15	4
6	6	16	2
7	9	17	8
8	15	18	7
9	11	19	6
10	6	20	4

Determine the control limits of p- chart and state the process is in control.

Solution :

$$\bar{p} = \frac{No\ of\ defectives\ found\ \ in\ any\ inpection}{Total\ no.of\ articles\ atually\ inspected} = \frac{120}{20\times100} = 0.66$$

$$UCL_p = \bar{p} + 3\sigma_p = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.06 + 3\sqrt{\frac{0.06\ (1-0.06)}{100}} = 0.13095$$

$$LCL_p = \bar{p} - 3\sigma_p = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.06 - 3\sqrt{\frac{0.06\ (1-0.06)}{100}} = -0.01095$$

Since the fraction defective can not be negative .

$$LCL_p = 0$$

Question 4: If the fraction defective of a large sample of a product is 0.1537, calculate the control limits Given that sub- group size is 2000.

Solution : Average fraction defective $\bar{p} = 0.1537$

Sub group size is 2000

n=2000

Central line $UCL_{np} = n\bar{p} + 3\sigma_{np} = n\bar{p} + 3\sqrt{n\bar{p}(1-\bar{p})}$

$=307.4 + 3\sqrt{307.4(1-0.1537)}$ = 355.787742

$$LCL_{np} = n\bar{p} - 3\sigma_{np} = n\bar{p} - 3\sqrt{n\bar{p}(1-\bar{p})}$$

$=307.4 - 3\sqrt{307.4(1-0.1537)}$ = 259.01225

Links

Population and sampling

https://www.youtube.com/watch?v=VPM84_yfx5Q

https://www.youtube.com/watch?v=iUutXUIwAvw&t=60s

Testing of hypothesis

https://www.youtube.com/watch?v=7kb-a7n2bcQ&t=16s

https://www.youtube.com/watch?v=1bywo64eIC0&t=344s

t-test

https://www.youtube.com/watch?v=UW1tUJFmmm8

https://www.youtube.com/watch?v=fF0tFUQv-c8

https://www.youtube.com/watch?v=gePqxb9Vxuo

F-test

https://www.youtube.com/watch?v=h5Glm738j84

https://www.youtube.com/watch?v=7eTO7faJqSg

https://www.youtube.com/watch?v=PmGHMzZ1u3Q

Chi Square test

https://www.youtube.com/watch?v=3aRIwDDMc88

https://www.youtube.com/watch?v=dXB3cUGnaxQ

https://www.youtube.com/watch?v=hpWdDmgsIRE

Z Test

https://www.oreilly.cyom/library/view/common-statistical-methods/9781607642282/ch06.html

Statistical Quality Control

http://www.au.af.mil/au/awc/awcgate/navy/bpi_manual/mod10-control.pdf

ANOVA

https://www.youtube.com/watch?v=TKTWIyC3LOQ