→ A variable whose value can be determine by performing a random experiment.

(1) **Random Variable** :- A random variable is a real valued function defined on the sample space of a random experiment. This means that the values of a random variable are determined by chance. For example:- consider the experiment of throwing three coins. The sample space of this experiment is

$$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TT$$

Let X denotes the number of heads in the sample point of the above space.

Here $X(HHH)=3$, $X(HHT)=2$, $X(HTH)=2$, $X(HTT)$
$X(THH)=2$, $X(THT)=1$, $X(TTH)=1$ and $X(TTT)=0$.

So X takes values 0, 1, 2 and 3 So X is a random variable or a stochastic variable

Definition :- A random variable is a real valued function X defined on the sample space S of a random experiment which associates a unique real number X.

A random variable is said to be discrete if it can assume only finite or countably infinite number of values and it is said to be continuous if it can assume any real value i.e. uncountable infinite number of values.

Scanned by CamScanner

Probability Distribution of a Random Variable

Let X be a random variable defined on the sample space S of a random experiment. Let X assumes values $x_1, x_2, x_3, \ldots, x_n$, probabilities of the values x_i of the variable X, denoted by $P(X=x_i)$ or $P(x_i)$ is defined as $P(X=x_i) =$ Sum of probabilities of all the simple events which are mapped to x_i by the ~~var~~ variable X i.e. $P(X=x_i) = \sum_{X(\omega)=x_i} P(\omega)$

eg: Consider the experiment of 'throwing three coins'. Let X be the random variable defined on the sample space

$$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$

of this experiment as $X = 0, 1, 2, 3$.

$P(X=0) = P(TTT) = \frac{1}{8}$

$P(X=1) = P(HTT, THT, TTH) = \frac{3}{8}$

$P(X=2) = P(HTH, THH, HHT) = \frac{3}{8}$

$P(X=3) = P(HHH) = \frac{1}{8}$

Now $P(X=0) + P(X=1) + P(X=2) + P(X=3) = 1$

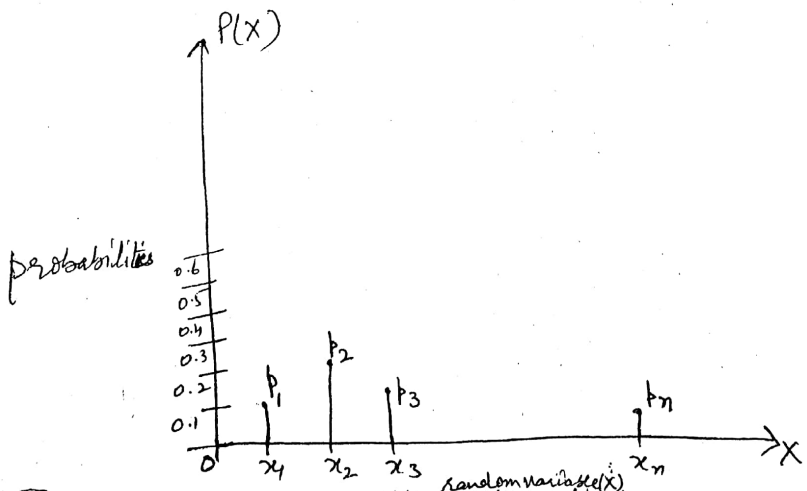
Scanned by CamScanner

The following Table is said to be describe the Probability distribution of the random variable X

X	0	1	2	3
P(X)	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

In general if $p_1, p_2, \ldots, p_n$ are respective probabilities of the values $x_1, x_2, \ldots, x_n$ of a random variable X, then $p_1 + p_2 + \ldots + p_n = 1$ and the table is called the probability distribution of X

X	x_1	x_2	x_3	———	x_n
P(X)	p_1	p_2	p_3	———	p_n



Graph of probability distribution of a random variable X

Unit-IV

Tests of Hypothesis

Any statement or assumption about the population parameter is known as hypothesis. Hypothesis testing is a statistical technique through which we can test any assumption about the population. Test of the hypothesis are also known as tests of the significance. We can find that whether the difference between the actual value and the hypothesis value is significant or not. eg: we want to test the assumption that the average age of P.G students in India is 22 yes. We take a sample of say 1000 students and note down their age. Suppose that the sample average age is 23 yrs. So, there is a difference of one year. With the help of tests of the hypothesis we shall decide that whether the difference of one year between actual (sample) value and assumed (hypothesis) value is large or small or we can say significant (rejected) or insignificant (accepted or confident)

Scanned by CamScanner

The Procedure of testing hypothesis involves the following steps :-

1) Set up a hypothesis
2) Set up a suitable significance level. or To choose the level of significance.
3) setting a test criterion
4) choice of test statistics
5) Making decisions.

1) Set up a hypothesis :- The first step in the testing of the hypothesis is to set up the hypothesis. There are two types of hypothesis
(a) Null hypothesis (b) Alternative hypothesis.

1) Null hypothesis :- Setting up of the null hypothesis is the most important step in the testing of hypothesis. It is denoted by the symbol H_0. As the name itself implies, null means nothing. Therefore null hypothesis implies that there is no significant difference between the hypothesis value and the actual value. When we want to test the significance of single population mean (μ) or single population proportion (Probability) (P) or single

Scanned by CamScanner

single population S.D (σ), the given value of the population is considered as true in the null hypothesis.

1) eg. suppose the mean marks of 40 students of a class are 55. We want to test whether the average score in the entire college is 52. We set up the following null hypothesis.

$$H_0 : \mu = 52$$

2) Similarly, suppose a teacher claims that 80% of the students taught by him pass in the examination and we want to test him claim by using sample data. We set up the null hypothesis. $H_0 : P = 80\%$ or 0.8

3) suppose when we want to test the significance of the difference between two populations means $\mu_1 - \mu_2$ or two population proportions $(P_1 - P_2)$ or two population S.D $(\sigma_1 - \sigma_2)$, we set up the null hypothesis that there is no difference between them i.e they are equal.

so our null hypothesis is

Scanned by CamScanner

$H_0: \mu_1 - \mu_2 = 0$ or $H_0: \mu_1 = \mu_2$

$H_0: P_1 - P_2 = 0$ or $H_0: P_1 = P_2$

$H_0: \sigma_1 - \sigma_2 = 0$ or $H_0: \sigma_1 = \sigma_2$

eg: Suppose in a college 40% students pass in paper A and 55% students pass in paper B, we want to test the proportion of the students who pass in two papers in the entire university is same or not. so null hypothesis is

$H_0: P_1 = P_2$

4) The null hypothesis should be set in such a way that we are able to get the unique expected value under it.

eg: Suppose a coin is tossed 100 times and we get head 55 times. We want to test whether the coin is unbaised we set up following ~~to~~ null hypothesis

H_0: coin is unbaised.

eg: Suppose in above example we are asked to test whether the coin is baised in favour of heads. then there can be many expected values of the number of the heads in the null hypothesis. So we can not get unique value under null.

Scanned by CamScanner

hypothesis. because unbaised means

$p, q = \frac{1}{2}$

p = prob of successes $= \frac{1}{2}$

$q = 1 - p = \frac{1}{2}$

ie we get a unique answer both have same value $= \frac{1}{2}$

but in case of baised we get result

$p > \frac{1}{2}$ or $p < \frac{1}{2}$

in above example coin is baised in favour of heads ie $p > \frac{1}{2}$ so we get many enputed values not a unique value

so we write null hypothesis as

H_0: coin is unbaised.

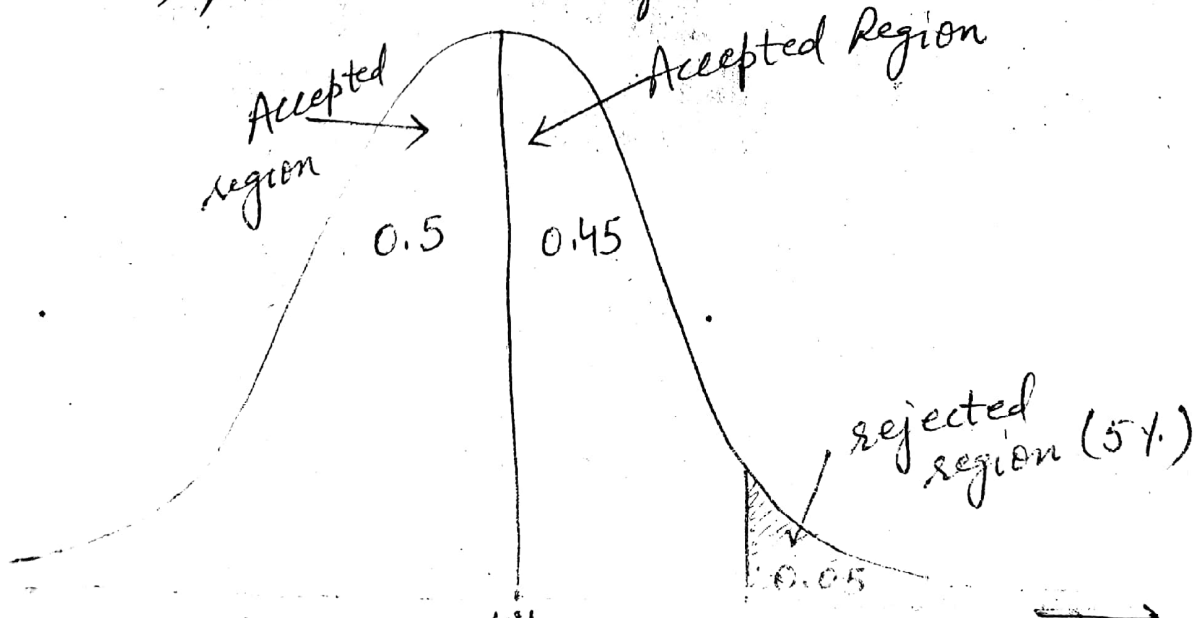
(b) Alternative hypothesis :- As the name itself implies, it is an alternative to the null hypothesis. It is denoted by H_1 or H_a. when the null hypothesis is ~~generally~~ rejected, the alternative hypothesis is accepted. The null hypothesis is generally simple hypothesis whereas the alternative hypothesis is the composite hypothesis. The concept of the alternative hypothesis is important because it also determines the tail of the test. There are two types of the

tests :

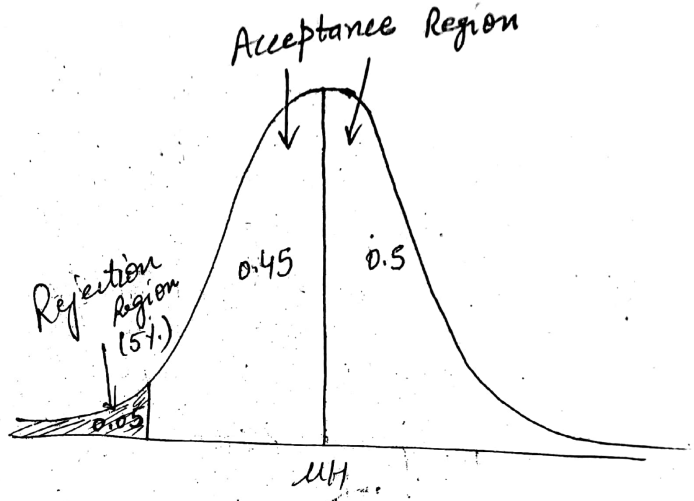
(1) One tail test

(2) Two tail Test.

(1) **One tail Test** :- In one tail test the rejection region is on one side. If it is on the left side then test is said to be left tailed test. If the rejection region is on the right side then the test is called right tailed test. The form of the alternative hypothesis (i.e. whether it is one tailed or two tailed) depends upon the Problem in the hand. If the nature of the question is such that it gives any direction (increase or decrease) it is case of one tail test.



Acception and rejection region one tailed right test.



Acceptance and Rejection Region. (One-tailed left Test)

eg. If we are asked to test whether the coin is ~~bai~~ biased in favour of the heads we set the following hypothesis

H_0: coin is unbiased i.e. $P=0.5$

H_1: coin is biased in favour of heads i.e. $P>0.5$

(One tail test – right tail test)

eg) If we are asked to test whether coaching improves the average marks of the students, we set of the following null hypothesis

$H_0: \mu_1 = \mu_2$

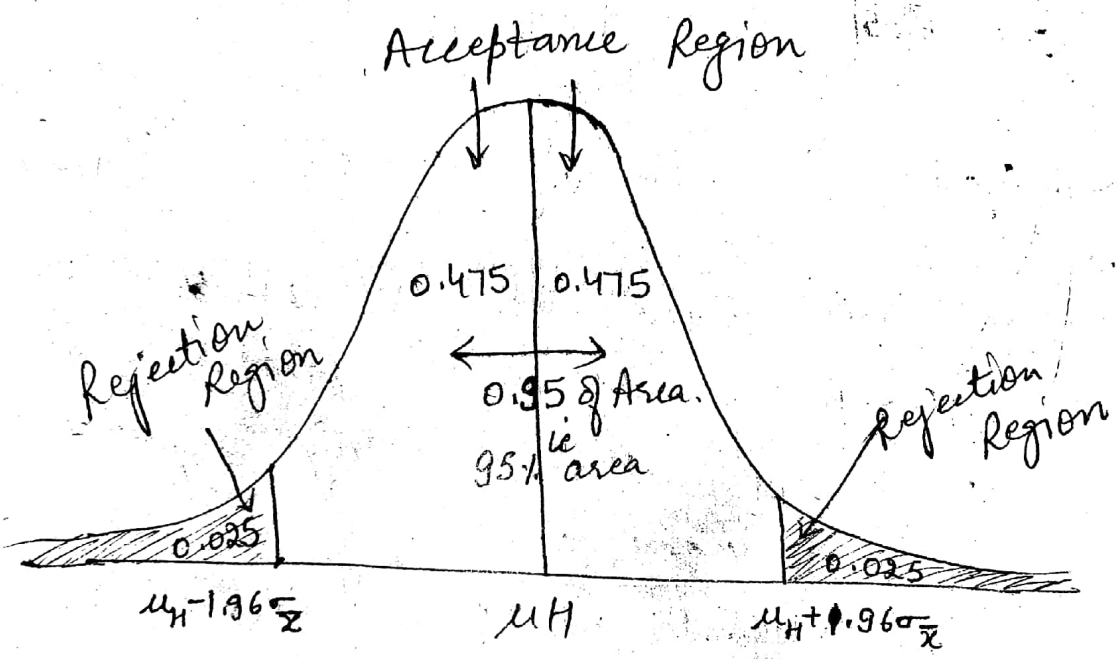
$H_1: \mu_1 < \mu_2$ [i.e. marks after coaching are more than before coaching] (One tail test)

Scanned by CamScanner

(2) <u>Two tail Test</u> :- In case of the two tailed test, the rejection region is equally divided into two parts. ie half on the left side and half on the right side of the test statistic.

Let $\alpha = 5\%$ of level of significance

ie 95% confidence interval



eg:- Suppose a coin is tossed 100 times and we get head 55 times. We want to test the coin is unbiased.

H_0: Coin is unbiased i.e. $P = 0.5$

H_1: coin is biased i.e. $P \neq 0.5$ (two tail test)

Scanned by CamScanner

Suppose we have data relating to average marks of the students before coaching and after coaching and we want to test whether coaching makes any difference in their marks: we set up the following ~~null~~ hypothesis.

$H_0 : \mu_1 = \mu_2$ (i.e. there is no difference)

$H_1 : \mu_1 \neq \mu_2$. (i.e there is difference) (two tail test).

2) Set up a suitable significance level

It is generally denoted by α. It can be 1%, 2%, 5%, 10% etc. Generally, it is taken as 5%, 1%. If in question if significance level is not given then we take $\alpha = 5\%$. In any statistical making these are four types of the decisions

1) The hypothesis is true but our test rejects it (Type I error)

Let α = Prob (Type I error)

2) The hypothesis is false but our test accepts it (type of II error)

Scanned by CamScanner

Type II error is denoted by β.

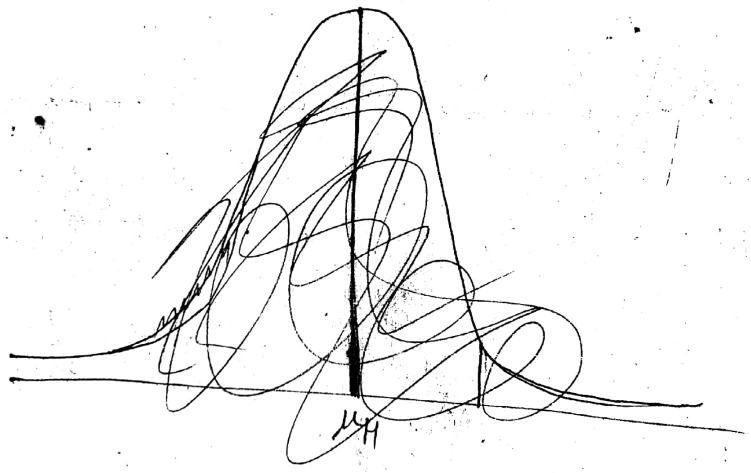
$$\beta = 1 - \alpha.$$

(3.) The hypothesis is true and our test accepts it (correct decision)

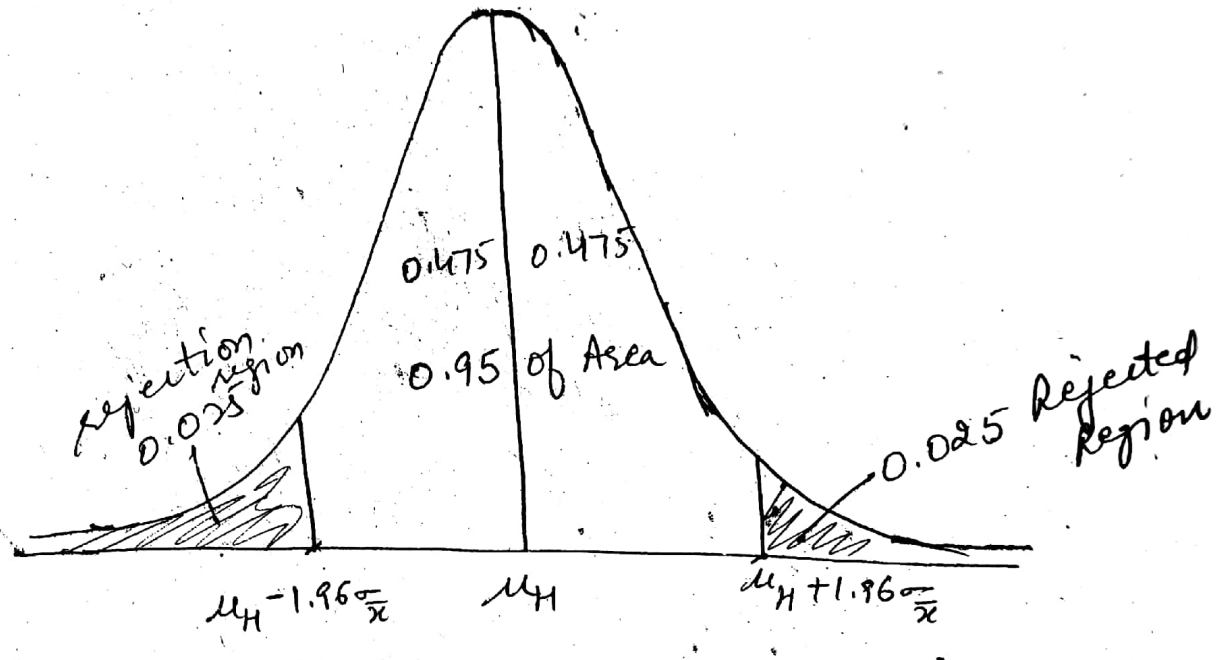
4) The hypothesis is false and our test ~~accepts~~ ~~it~~ rejects it (Correct decision)

If $\alpha = 5\%$ level of significance (Rejected Region)

$\Rightarrow$ 95% confidence internal (accepted Region)



Scanned by CamScanner



3) Setting a test criterion :- The third step in hypothesis testing procedure is to construct a test ~~test~~ criterion. This involves selecting an appropriate probability distribution for the particular test. Some probability distributions that are commonly used in testing procedures are t, f, χ^2.

4) Choice of Test statistics :- Having taken the first three steps, we have completely designed a statistical test.

Scanned by CamScanner

~~The test statistics is difference b/w sample value (actual value) and population value and this difference divided by standard error~~

The test statistics is the ratio of difference b/w sample value (actual value) and population value to standard error

i.e. $\frac{\text{sample value} - \text{pop. value}}{S.E}$

5) Making decision :- The calculated value is compared with critical value at given level of significance for a given tail of test. If the calculated value is more than the critical value the null hypothesis is rejected. In other words it implies that the difference b/w the sample mean and population mean is significant. Then we take decision from confidence internal i.e. 95%, 99% or 100% confidence internal.

Scanned by CamScanner

96c

Non Parametric Test

Simran Sahni

We know that t-test, F-test all are based on assumption that population follows N.D. All these tests are called parametric tests. We may have a situation in which it is not possible to make assumption about population or population parameter. So, this has led to development of a group of alternative test which are not based on any distribution (not based on any assumption) are known as non-parametric test or distribution free test, while using non parametric test it does not matter whether the population follows N.D or B.D or any other distribution.

Reasons for use of Non-Parametric Tests :-

1) These are distribution free tests i.e. no assumption about population parameter is required.

2) These are easy to compute as to parametric test.

(3) These can be used for all types of data both qualitative as well as quantitative

Assumptions for Non Parametric Tests

These tests do not make any assumption about population however they are based on following assumption

1) Sample observations are independent.
2) Variable under study is continuous.
3) The probability density function is continuous.

Advantages of Non-Parametric Tests

1) These are simple to understand and easy to calculate.

(2) No assumptions are made in any form about population.

(3) The data may involve quantitative or qualitative measurement

(4) Many non-parametric tests can be used when sample size is small.

Scanned by CamScanner

Limitations of Non - Parametric Test

1) With the help of Non Parametric Test we can test only the hypothesis but we can't make any estimation about population parameter.

2) When the assumption of parametric tests are fulfilled they are more reliable. In other words parameter test give better results than non parametric test.

3) Non parametric Tests don't use the entire information which is given. So, there is a wastage of information.

The Non parametric Tests are

1) Chi- Square Test
2) Run test
3) U-test etc

Chi Square distribution :-

It is most important & widely used distribution in statistics. The square of standard Normal variate is called chi-square variate with some degree of freedom

Scanned by CamScanner

ie if $X \sim N.D[\mu, \sigma]$ then $z = \frac{X-\mu}{\sigma}$ is called standard Normal variate

further $z^2 = \left(\frac{X-\mu}{\sigma}\right)^2$ is called a chi-square variate with one degree of freedom.

similarly, if $z_1, z_2, z_3, \text{——}, z_n$ be n standard Normal variates then $z_1^2 + z_2^2 + z_3^2 + \text{—} + z_n^2$ is called chi-square variate with n degrees of freedom.

ie $\sum_{i=1}^{n} z_i^2 = \chi_n^2$ ie chi-square variate with n dif.

Probability density function of chi-square distribution

If χ^2 is a random variable following χ^2 distribution with v d.f. then its p.d.f is given by

$$F(\chi^2) = \frac{e^{-\frac{\chi^2}{2}} (\chi^2)^{\frac{v}{2}-1}}{2^{\frac{v}{2}} \lfloor \frac{v}{2}-1}; \quad v \geq 1, \quad 0 < \chi^2 < \infty$$

Scanned by CamScanner

Properties of chi square distribution

1) It is a continuous probability distribution as the chi square variates ranges from 0 to ∞.

2) The total area under a chi square curve is equal to 1.

3) Mean of χ^2 dist. is v ie degree of freedom

4) Variance of χ^2- dist. is $2v$

5) v is the only parameter of χ^2-dist.

6) Additive Property of χ^2-dist.

The sum of independent χ^2 variates is also be a χ^2 variates.

ie if $\chi_1^2, \chi_2^2, \ldots, \chi_K^2$ be K-χ^2 variates with $v_1, v_2, v_3, \ldots, v_K$ d.f. Then

$S = \chi_1^2 + \chi_2^2 + \chi_3^2 + \ldots + \chi_K^2$ will also be a χ^2 variate with $v_1 + v_2 + v_3 + \ldots + v_K$ d.f.

7) χ^2 dist. Tends to N.D as $v \to \infty$

Scanned by CamScanner

χ^2-Test

This test was given by Karl Pearson's. It is a non-Parametric or distribution free test. Further, it is most powerful and most useful non Parametric test, with the help of χ^2-test without making any assumption about population parameter we can compare not only one observed value with expected value but many values can be compared at a time the following formula is used

$$\chi^2 = \Sigma \left[\frac{(O-E)^2}{E} \right]$$

where O means observed frequencies
E = expected frequencies

<u>decision</u>:- The calculated value of χ^2 is compared with the Table value from χ^2 Table. from χ^2 table at a given level of significance & for (n-1) d.f. If the calculated value is more than the Table value then hypothesis is rejected.

Scanned by CamScanner

Conditions for the use of χ^2-test

1) The sum of the observed and expected frequencies should be equal

i.e. $\sum O = \sum E$

2) No individual observed or expected frequency should be less than 5

eg:
0
7
5
4
6
3
7

Where 4, 3 < 5 then we use technique pooling
In case the frequencies are less than 5 we use pooling technique with its preceding or succeeding frequencies so as to have a total greater than equal to 5 The d.f. are reduced accordingly.

in example.
0
7
5]+9
4
6
3]+10
7

Scanned by CamScanner

Initially $n=6$, $v=n-1=5$

after pooling $n=4$, $v=n-1=3$

ie when we apply pooling technique 2 times then d.f. also reduced 2 times.

(3) n = no. of observations should be sufficiently large.

Applications of χ^2-test / Uses of χ^2-test

1) χ^2-test as a test of hypothesis.
2) χ^2-test as a test of uniformity or equality
3) χ^2-test as a test of goodness of a fit.
4) χ^2-test as a test of independence or association b/w attributes
5) χ^2-test as a test of population variance.

1) χ^2-test as a test of hypothesis

With the help of χ^2-test, we can test any hypothesis concerning the population whatever information about the population is

Scanned by CamScanner

given. we consider it as true under the null hypothesis & expected frequencies are calculated then we apply the formula

$$\chi^2 = \Sigma \left[\frac{(O-E)^2}{E}\right]$$

we have to compare this calculated value with the table value to have a decision.

2 A dice is thrown 180 times with the following results

No turned up	1	2	3	4	5	6	Total
frequencies	25	35	40	22	32	26	180

Test the hypothesis that die is unbiased.

Sol n = no. of observation = 6
N = total frequency = 180

we set up the hypothesis that dice is unbiased. On the basis of this hyp. we expect that each of these six numbers are equally probable.

so the expected frequencies of getting each of number 1,2,3,4,5,6 is $\frac{180}{6} = 30$

Scanned by CamScanner

O	E	O − E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
25	30	−5	25	0.8333
35	30	5	25	0.833
40	30	10	100	3.333
22	30	−8	64	2.133
32	30	2	4	0.133
26	30	−4	16	0.533
			234	7.798

$$\therefore \chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 7.8$$

<u>Decision</u>:- Table value of χ^2 at $\alpha = 5\%$ level of significance & for $\nu = n-1 = 6-1 = 5$ d.f. i.e. $\chi^2_{(0.05,\,5)} = 11.07$

Calculated value is less than table value

$\therefore$ hyp is accepted.

i.e. dice is unbiased.

Q 200 digits were choosen at random from a set of table. The frequencies of digits were

digits	0	1	2	3	4	5	6	7	8	9
freq	18	19	23	21	16	25	22	20	27	15

Scanned by CamScanner

Use χ^2-test to access the correctness of hypothesis that digits were distributed in equal numbers.

Sol ~~[illegible]~~

~~Case~~

~~Case~~ We set ~~up~~ the hypothesis that digits were equally distributed in numbers

$n = 10 < 30$

$\therefore$ ~~we~~ we use small sample test. we test the hypothesis so we χ^2-test.

$N = 200$.

Under the null hyp. we expect that each digit must occur an equal no. of times

$\therefore$ Expected freq. of each of these digits 0 to 9 is $\frac{200}{10} = 20$

O	E	O − E	$(O-E)^2$	$(O-E)^2/E$
18	20	−2	4	0.2
19	20	−1	1	0.05
23	20	3	9	0.45
21	20	1	1	0.05
16	20	−4	16	0.8
25	20	5	25	1.25
22	20	2	4	0.2
20	20	0	0	0
21	20	1	1	0.05
15	20	−5	25	1.25
				4.3

$\therefore \chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right]$

$= 4.3$

$\chi^2_{(0.05,9)} = 16.919$

Scanned by CamScanner

As calculated value of χ^2 is less that the value ∴ null hyp is accepted.

---

2) χ^2-test as a test of uniformity or equality

With the help of χ^2-test we can test whether there is a uniformity or equality in the data or not. ie we can test whether all the items in the data are equal or not.

We set up the hypothesis that there is a uniformity in the data ie all the observations are equal. Under this hypothesis we calculate expected frequency and apply the formula

$$\chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right]$$

Q A survey of 320 families with 5 children each has the following distribution.

Boys:	5	4	3	2	1	0
Girls:	0	1	2	3	4	5
Families	14	56	110	86	40	12

Scanned by CamScanner

the result consistent with hypothesis (7) that males & females birth are equally probable

we have $n = 5$, $N = 320$.

$P(\text{male birth}) = p = \frac{1}{2}$

$P(\text{female birth}) = q = 1 - p = \frac{1}{2}$.

we set up the hypothesis that male birth and female birth are equal

By Binomial distribution we know that prob. of getting r mail births in a family of 5 is

$$P(X=r) = {}^5C_r\, p^r\, q^{5-r}$$

$$= {}^5C_r \left(\frac{1}{2}\right)^r \left(\frac{1}{2}\right)^{5-r}$$

$$= {}^5C_r \left(\frac{1}{2}\right)^{\cancel{r}+5-\cancel{r}}$$

$$= {}^5C_r \left(\frac{1}{2}\right)^5$$

$\therefore$ Expected frequencies of r male birth in total of 320 families is

$$f(X=r) = N \times P(X=r)$$

$$= 320 \times {}^5C_r \left(\frac{1}{2}\right)^5$$

Scanned by CamScanner

$f(X=5) = 320 \times {}^5C_5 \left(\frac{1}{2}\right)^5 = 10$

$f(X=4) = 320 \times {}^5C_4 \left(\frac{1}{2}\right)^5 = 50$

$f(X=3) = 320 \times {}^5C_3 \left(\frac{1}{2}\right)^5 = 320 \times \frac{5 \times 4 \times 3}{\lfloor 3} \left(\frac{1}{2}\right)^5$

$= \overset{10}{\cancel{320}} \times \frac{5 \times \overset{2}{\cancel{4}} \times \cancel{3}}{\cancel{6}\,\cancel{2}} \left(\frac{1}{\cancel{32}}\right)^5$

$= 100.$

$f(X=2) = 320 \times {}^5C_2 \left(\frac{1}{2}\right)^5 = \overset{10}{\cancel{320}} \times \frac{5 \times 4}{\lfloor 2} \times \frac{1}{\cancel{32}}$

$= \overset{10}{\cancel{320}} \times \frac{5 \times \cancel{4}^{\,2}}{\cancel{2}} \times \frac{1}{\cancel{32}}$

$= 100$

$f(X=1) = 320 \times {}^5C_1 \left(\frac{1}{2}\right)^5 = \overset{10}{\cancel{320}} \times 5 \left(\frac{1}{\cancel{32}}\right)$

$= 50.$

$f(X=0) = 320 \times {}^5C_0 \left(\frac{1}{2}\right)^5 = \overset{10}{\cancel{320}} \times 1 \times \frac{1}{\cancel{32}} = 10$

O	E	O−E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
14	10	4	16	1.6
56	50	6	36	0.72
110	100	10	100	1
88	100	−12	144	1.44
40	50	−10	100	2
12	10	2	4	0.4
				7.16

$\Sigma\left[\frac{(O-E)^2}{E}\right] = 7.16$

Scanned by CamScanner

$\therefore$ Now $\Sigma\left[\frac{(O-E)^2}{E}\right] = 7.16$

& $\chi^2_{(0.05,4)} = 9.488$

As calculated value of χ^2 < Table value

$\therefore$ Null hyp. is accepted.

(5). χ^2-test as a test of goodness of fit

With the help of χ^2-test, we can test that any of theoratic distribution like B.D, P.D or N.D is good fit for any set of data or not.

We set up the hypothesis that the particular theoratical distribution is goodfit for given set of data. We find the expected frequencies under this hypothesis.

Next we calculate the calculated value by applying the formula.

$$\chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right]$$

this value compare with table value.

Scanned by CamScanner

Decision :- The calculated value of χ^2 is comp[illegible] with table value at given level of signif[illegible] & for appropriate d.f. If calculated value is more than the table value the null hyp. is rejected.

In case of B.D d.f is $n-1$

P.D " " $n-2$

N.D " " $n-3$.

Assignment

Q 4 coins were tossed 160 times and following results were observed.

No. of heads	0	1	2	3	4
frequency	17	52	54	31	6

Under the assumption that coins are balanced (unbiased). Test the goodness of fit of B.D.

Hint $p = \frac{1}{2}$, $q = \frac{1}{2}$

$P(X=r) = {}^4C_r \, p^r q^{4-r} = {}^4C_r \left(\frac{1}{2}\right)^r \left(\frac{1}{2}\right)^{4-r}$

$f(X=r) = N \times P(X=r)$

$= 160 \times {}^4C_r \left(\frac{1}{2}\right)^r \left(\frac{1}{2}\right)^{4-r}$

Scanned by CamScanner

$= 160 \times {}^4C_r \left(\frac{1}{2}\right)^{\cancel{r+4-r}}$

$f(x=r) = 160 \times {}^4C_r \left(\frac{1}{2}\right)^4$

$f(x=0) = 160 \times {}^4C_0 \left(\frac{1}{2}\right)^4 = 10$

$f(x=1) = 160 \times {}^4C_1 \left(\frac{1}{2}\right)^4 = 40$

$f(x=2) = 160 \times {}^4C_2 \left(\frac{1}{2}\right)^4 = 60$

$f(x=3) = 160 \times {}^4C_3 \left(\frac{1}{2}\right)^4 = 40$

$f(x=4) = 160 \times {}^4C_4 \left(\frac{1}{2}\right)^4 = 10.$

O	E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
17	10		
52	40		
54	60		
31	40		
6	10		

$\underline{12.725}$

$\therefore \chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 12.725$

and $\chi^2_{(0.05,3)} = 7.815$

Scanned by CamScanner

Q The following set of observation be regarded as those of random sample from a P.D.

X	0	1	2	3	4
freq.	122	60	15	2	1

Sol we set up the hypothesis that there is no difference b/w observed frequencies and the frequencies observed by fitting P.D. i.e. P.D is goodfit for given data. By P.D we know that the expected freq. of getting a successes.

$$f(X=r) = N \times P(X=r)$$

$$= N \times \frac{e^{-m} m^r}{\lfloor r}, \quad m = \text{mean of P.D}$$

i.e. $m = \text{mean} = \frac{\sum f_i x_i}{\sum f_i}$

$$= \frac{0\times122 + 1\times60 + 2\times15 + 3\times2 + 4\times1}{122+60+15+2+1}$$

$$= 0.5$$

Also $N = \sum f_i = 200$

Scanned by CamScanner

Expected frequency $= f(X=r) = N \times \frac{e^{-m} m^r}{\lfloor r}$ (10)

$= 200 \times \frac{e^{-0.5} (0.5)^r}{\lfloor r}$

~~f(X=0) = 200~~ $= \frac{200 \times (0.61)(0.5)^r}{\lfloor r}$

$f(X=0) = 200 \times \frac{(0.61)(0.5)^0}{\lfloor 0} = 122$

$f(X=1) = 200 \times \frac{(0.61)(0.5)^1}{\lfloor 1} = 61$

$f(X=2) = 200 \times \frac{(0.61)(0.5)^2}{\lfloor 2} = 15.25 \simeq 15$

$f(X=3) = 200 \times \frac{(0.61)(0.5)^3}{\lfloor 3} = 2.54 \simeq 2$

$f(X=4) = 200 \times \frac{(0.61)(0.5)^4}{\lfloor 4} = 0.317 \simeq 0$

O	E
122	122
60	61
15	15
2	2
1	0
200	200

+18 17

After Pooling

O	E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
122	122	0^2	0
60	61	$(-1)^2$	0.016
18	17	$(1)^2$	0.059
200	200		0.075

After pooling number of items = 3.

$\chi^2 = \sum [(O-E)^2/E] = 0.075$

Now $\chi^2_{(0.05,1)} = 3.84$

$\therefore$ Calculated value of χ^2 is less than the table value. Therefore the hypothesis is true. Hence the fit is good.

(4) χ^2-test as a test of independence or associations b/w attributes.

With the help of χ^2-test we can test whether two or more than two attributes are associated or independent.

egr (1) We may want to know whether education or employment are associated.

(2) Whether post graduation or marriage are associated.

(3) Whether weight and height are associated.

So, we set up the hypothesis that the given attributes are not associated i.e. they are independent.

We find expected frequencies under this hypothesis and apply the formula

$$\chi^2 = \sum \left[\frac{(O-E)^2}{E}\right]$$

Decision:- The calculated value of χ^2 is compare with table value at given level of significance & for $(c-1) \times (r-1)$ d.f's. where

c refers to number of columns &

r refers to number of rows

If calculated values is more than the table value then hyp. is rejected.

eg. Suppose we have data relating to two attributes education & employment

Let X stands for education

& Y stands for employment

Y↓ X→	educated (A)	uneducated (A^c)	Total
Employed (B)	a	b	a+b
unemployed (B^c)	c	d	c+d
Total	a+c	b+d	N

Scanned by CamScanner

By def. Expected freq ie $f(X=r) = N \times P(r$

$$E(a) \text{ or } f(X=a) = \frac{(a+c)(a+b)}{N}$$

$$E(b) \text{ or } f(X=b) = \frac{(a+b)(b+d)}{N}$$

$$E(c) \text{ or } f(X=c) = \frac{(c+d)(a+c)}{N}$$

$$E(d) \text{ or } f(X=d) = \frac{(c+d)(b+d)}{N}$$

<u>Note</u>:- In the special case of 2x2 table. We can calculate the value of χ^2 directly without calculating the expected frequency by using formula.

$$\chi^2 = \frac{N(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

Q With the help of following data find out whether there is any relationship b/w smoking & drinking.

Y ↓ \ X →	Smoking	Not smoking
drinking	74	26
Not		

Scanned by CamScanner

Y ↓ \ X →	Smoking	NA smoking	Total
Drinking	74	26	100
NA drinking	10	30	40
Total	84	56	N = 140

We set up the hypothesis that there is no association b/w smoking & drinking i.e They are independent.

Ist Method

It is 2×2 table

$$\therefore \chi^2 = \frac{140\,[(74)(30) - (10)(26)]^2}{100 \times 40 \times 84 \times 56}$$

$$= 28.59$$

IInd Method

$$E(74) = \frac{100 \times 84}{140} = 60$$

$$E(26) = \frac{100 \times 56}{140} = 40$$

$$E(10) = \frac{40 \times 84}{140} = 24$$

$$E(30) = \frac{40 \times 56}{140} = 16$$

Scanned by CamScanner

O	E	$(O-E)^2$	$(O-E)^2/E$
74	60	196	3.267
26	40	196	4.9
10	24	196	8.167
30	16	196	12.25
			28.584

Therefore $\chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 28.584$

Now table value of χ^2 at $\alpha = 5\%$ for $(2-1)\times(2-1) = 1$ d.f. is 3.841

As calculated value of χ^2 is greater than its table value.

Therefore hyp is rejected

∴ They are not independent.

Q The following table gives the no. of good and bad parts produce by each of three shifts in a factory.

Shifts	Good	Bad	Total
Morning	900	130	1030
Evening	700	170	870
Night	400	200	600

Scanned by CamScanner

there is any association b/w shifts and quality of product produced?

(13)

Sol. We set up the hyp. that shifts and quality of product are independent. there is no association b/w them.

It is 3X2 table

$$E(900) = \frac{1030 \times 2000}{2500} = 824$$

$$E(130) = \frac{1030 \times 500}{2500} = 206$$

$$E(700) = \frac{870 \times 2000}{2500} = 696$$

$$E(170) = \frac{870 \times 500}{2500} = 174$$

$$E(400) = \frac{600 \times 2000}{2500} = 480$$

$$E(200) = \frac{600 \times 500}{2500} = 120.$$

O	E	$(O-E)^2$	$(O-E)^2/E$
900	824	5776	7.009
130	206	5776	28.039
700	696	16	0.023
170	174	16	0.092
400	480	6400	13.333
200	120	6400	53.333

Scanned by CamScanner

$\therefore \chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 101.829$

Now the table value of χ^2 at $\alpha = 5\%$ for $(2-1)\times(3-1) = 2$ d.f. is 5.991

As calculated value of $\chi^2 > 5.991$

$\therefore$ hyp is rejected.

$\therefore$ They are not independent.

Note Yates correction

eg: Shifts	Good	Bad	Total
Morning	7] 10	6] 16	13
Evening	3	10	13
	10	16	28

$d.f = (r-1)\times(s-1) = (1-1)\times(1-1) = 0.$

~~d.f~~

and d.f. cannot be zero.

so we cannot make pooling so we use Yates correction.

Scanned by CamScanner

It has two ways

(14)

In other words we can say in χ^2-test when any expected frequency is less than 5 we generally used the technique of pooling this reduces the d.f. accordingly however in case of 2x2 table, if any expected frequency is less than 5, we can't use the ~~the~~ technique of pooling. It is because of reason that d.f. is already one which can't be reduced further as we know that d.f. can never be less than 1.

So, the solution of this problem is reduced by Prof. Yates. It is well known correction factor known as Yates correction.

It can be applied in two ways

1) We change the basic formula of χ^2

$$\chi^2 = \frac{N(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)}$$ to

$$\chi^2 = \frac{N\left(|ad-bc| - \frac{N}{2}\right)^2}{(a+c)(b+d)(a+b)(c+d)}$$

Scanned by CamScanner

(2) We add 0.5 to the frequency w[illegible] is less than 5 & adjust all the frequencies so as to have same total and apply the formula

$$\chi^2 = \frac{N(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)}$$

This calculated value of χ^2 is compared with table value as discussed above,

Assignment

Q An experiments the following results were obtain

Shifts	Good	Bad	Total
Morning	2	10	12
Evening	6	6	12
Total	8	16	24

examine the effect of shift in controlling the quality of Product.

Scanned by CamScanner

a long period of time the grades are given by a group of instructors in a particular course have average.

12% A's, 18% B's, 40% C's, 18% D's & 12% E's

A new instructor gives 22 A's, 34 B's, 66 C's, 16 D's & 12 E's. During two semester's determine at 1% level of significance whether the new instructor is following the way pattern set of others.

Sol: we set up the hyp. that new instructor is not following the grade pattern set by others i.e they are independent.

Now Expected freq. of A's i.e $E(A) = 150 \times 12\%$

$= \frac{\cancel{12}\,6}{\cancel{100}\,\cancel{8}} \times \cancel{150}\,3 = 18$

$E(B) = 18\%$ of $150 = \frac{18}{100} \times 150 = 27$

$E(C) = 40\%$ of $150 = \frac{40}{100} \times 150 = 60$

$E(D) = 18\%$ of $150 = \frac{18}{100} \times 150 = 27$

$E(E) = 12\%$ of $150 = \frac{12}{100} \times 150 = 18$

Scanned by CamScanner

O	E	$(O-E)^2$	$(O-E)^2/E$
22	18	16	0.889
34	27	49	1.815
66	60	36	0.6
16	27	121	4.487
12	18	36	2
		258	9.785

$$\therefore \chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 9.785$$

Decision

~~Now~~ Now the table value of χ^2 at $\alpha = 1\%$ level of significance & for $v = 5-1=4$ d.f. is 9.488

Now the calculated value of χ^2 is greater than the table value

$\therefore$ hyp. is rejected.

---

(5) χ^2-test as a test of Population variance

With the help of χ^2-test we can test whether the pop. has a given value of variance for small samples

Scanned by CamScanner

Then we want to test if a random sample of $x_1, x_2, x_3, \rightarrow x_n$ has been drawn from a normal population with mean μ & a specified variance σ^2 then the statistic is

$$\chi^2 = \frac{\Sigma (x_i - \bar{x})^2}{\sigma^2}$$

$$= \frac{n s^2}{\sigma^2}$$

Where S is S.D of the sample follows chi-square distribution with $(n-1)$ d.f.

By comparing the calculated value of χ^2 with the table value for $(n-1)$ d.f. at certain level of significance. We may accept or reject the hypothesis.

Eg. Weights in Kg. of 10 students are given below:

38 40 45 53 47 43 55 48 52 49

Can we say that variance of the distribution of weights of all students from which the above sample of 10 students was drawn is equal to 20 sqrd. Kg? You are given

the following table values:

Degree of Freedom	$\chi^2_{0.05}$	$\chi^2_{0.01}$
9	16.92	21.67
10	18.31	23.21

Sol Applying χ^2 test, we set up the hyp. that the population variance is 20 square kg.

X	$X-\bar{X}$	$(X-\bar{X})^2$
38	−9	81
40	−7	49
45	−2	4
53	+6	36
47	0	0
43	−4	16
55	8	64
48	1	1
52	5	25
49	2	4
470	0	280

$$\bar{X} = \frac{\Sigma X}{N} = \frac{470}{10} = 47$$

$$\sigma^2 = \frac{\Sigma(X-\bar{X})^2}{n} = \frac{280}{20} = 14$$

Scanned by CamScanner

Calculated value of χ^2 is less than the table value. therefore hypothesis is true ie population variance is 20 square kg.

Q) A random sample size 25 from a population gives the sample standard deviation 8.5. Test the hypothesis that the population standard Deviation is 10.

Sol) we set the hypothesis that the population standard deviation is 10.

we are given $n = 25$, $S = 8.5$

& $\sigma = 10$

$\sigma^2 = 100$.

Now $\chi^2 = \frac{nS^2}{\sigma^2} = \frac{n(8.5)^2}{100} = \frac{25(8.5)^2}{100}$

$= \frac{1806.25}{100}$

$= 18.0625$

Decision

$\nu = n-1 = 25-1 = 24$

$\nu = 24$, $\chi^2_{(0.05, 24)} = 36.415$

$\therefore$ The calculated value of χ^2 is less than the table value $\therefore$ the hyp. is accepted

Therefore the population [illegible] is 10.

Scanned by CamScanner

Simran Sahni

F-test or Variance Ratio Test.

The F-test is named in honour of the great statistician R.A. Fisher. This test is based on F-distribution. With the help of F-test we can test whether population variances are equal or not. The following procedure is used.

We set up the hypothesis.

$H_0: \sigma_1^2 = \sigma_2^2$

$H_1: \sigma_1^2 > \sigma_2^2$

or $H_1: \sigma_2^2 > \sigma_1^2$

In other words the F-test is always be a one tail test. Further we can say that it will always be a right tail test by writting the alternative hypothesis as

$$\sigma_1^2 > \sigma_2^2 \quad \text{or} \quad \sigma_2^2 > \sigma_1^2$$

For finding the value F, the greater variance is always taken in numenator.

ie if $S_1^2 > S_2^2$ then alternative hypothesis will be $\sigma_1^2 > \sigma_2^2$ and F is defined as.

Scanned by CamScanner

$$S_1^2 = \frac{\Sigma(x_1-\bar{x}_1)^2}{n_1-1}, \quad S_2^2 = \frac{\Sigma(x_2-\bar{x}_2)^2}{n_2-1}$$

further if $S_2^2 > S_1^2$ then the alternative hypothesis will be $H_1: \sigma_2^2 > \sigma_1^2$ & F is defined as

$$F = \frac{S_2^2}{S_1^2} \quad \frac{\text{Large estimate of variance}}{\text{Small estimate of variance}}$$

<u>Decision</u> :- The calculated value of F is compared with table value at v_1 and v_2 d.f. where v_1 is the d.f. of numerator and v_2 is the d.f. of denominator.

ie if $S_1^2 > S_2^2$ then $F = \frac{S_1^2}{S_2^2}$

$\therefore v_1 = n_1 - 1, \quad v_2 = n_2 - 1$

and further if $S_2^2 > S_1^2$ then $F = \frac{S_2^2}{S_1^2}$

$\therefore v_1 = n_2 - 1, \quad v_2 = n_1 - 1$

If the calculated value is more than the table value then the null hypothesis is rejected.

Scanned by CamScanner

distribution

Since f-test is based on the ratio of two variances, it is also known as Variance Ratio Test.

Ratio of two variances follows a distribution called the f-distribution named after the famous statistician R.A. Fisher.

Assumptions of f-test

→ Normality :- The values in each group are normally distributed.

→ Homogeneity :- The variance within each group should be equal for all groups ($\sigma_1^2 = \sigma_2^2 = \ldots \sigma_n^2$). This assumption is needed in order to combine or pool the variances within the groups into a single 'within groups' source of variation.

→ Independence of error :- It states that the error should be independent for each value.

Scanned by CamScanner

F- distribution :- The variance ratio test or F-test is based on F-distribution. The F-statistic is defined as ratio of two ~~ide~~ independent χ^2- variates divided by their respective d.f.

$$F = \frac{\frac{\chi^2_{v_1}}{v_1}}{\frac{\chi^2_{v_2}}{v_2}}$$; where $\chi^2_{v_1}$, $\chi^2_{v_2}$ are two independent χ^2- variates with v_1, v_2 d.f.

Further we know that

$$\chi^2 = \frac{nS^2}{\sigma^2}$$

$\therefore$ we can have, $\chi^2_{v_1} = \frac{n_1 S_1^2}{\sigma_1^2}$, $v_1 = n_1 - 1$

& $\chi^2_{v_2} = \frac{n_2 S_2^2}{\sigma_2^2}$, $v_2 = n_2 - 1$

In case of F- distribution, it is assume that the population variances are equal i.e. $\sigma_1^2 = \sigma_2^2$.

$$\therefore F = \frac{S_1^2}{S_2^2}$$

Scanned by CamScanner

Properties of F-dist.

1) It is based on the d.f. v_1, v_2 of two independent χ^2-variates.

2) The mean of F-distribution is $\frac{v_2}{v_2-2}$, $v_2 > 2$.

ie mean of F-dist. is independent of v_1 and is numarically greater than 1

ie. $\boxed{\text{Mean} > 1}$

) For large values of v_1, v_2 the variance of F-dist. is approximately is given by

$$\text{var}(F) \cong 2\left(\frac{1}{v_1} + \frac{1}{v_2}\right)$$

) Mode of F-dist is $\text{Mode} = \frac{v_2(v_1-2)}{v_1(v_2+2)}$.

) If $v_1 = 1$ Then F-dist becomes equal to t^2 distribution with v_2 d.f.

ie. $F = t^2_{v_2}$, when $v_1 = 1$

Scanned by CamScanner

~~[illegible]~~ -

⑥ Inter relation b/w t, z, χ^2, F

① $$F = \frac{\frac{\chi^2_{v_1}}{v_1}}{\frac{\chi^2_{v_2}}{v_2}}$$

② One taking $v_1 = 1$ in ①, we get

$$F = t^2_{v_2} = \left(\frac{z}{\sqrt{\frac{\chi^2_{v_2}}{v_2}}}\right)^2$$

③ $$t = \frac{z}{\sqrt{\frac{\chi^2_v}{v}}}$$

④ $$\chi^2_v = \sum_{i=1}^{v} z_i^2$$

⑤ As $v \to \infty$, t distribution becomes N.D and χ^2-dist. becomes N.D.

---

eg: Two samples of size 9 and 8. Give the sum of the squares of the deviations from their respective means 160 and 91 resp. Can they be regarded as drawn from the two normal populations with same variance.

Sol given $n_1 = 9$, $n_2 = 8$ & $\Sigma(x_1 - \bar{x}_1)^2 = 160$ &

$\Sigma(x_2 - \bar{x}_2)^2 = 91$

Scanned by CamScanner

$H_0: \sigma_1^2 = \sigma_2^2$.

Now $S_1^2 = \dfrac{\Sigma(x_1-\bar{x}_1)^2}{n_1-1} = \dfrac{160}{8} = 20$

and $S_2^2 = \dfrac{\Sigma(x_2-\bar{x}_2)^2}{n_2-1} = \dfrac{91}{7} = 13$.

As $S_1^2 > S_2^2$ $\therefore H_1: \sigma_1^2 > \sigma_2^2$
at $\alpha = 5\%$ level of significance
i.e Test statistic is

$$F = \frac{S_1^2}{S_2^2} = \frac{20}{13} = 1.5$$

<u>Decision</u>: Here $v_1 = n_1 - 1 = 8$
$v_2 = n_2 - 1 = 7$

Table value of F at $\alpha = 5\%$ for $v_1 = 8$ and $v_2 = 7$ i.e. $F_{0.05}(8,7) = 3.7257$.

$\therefore$ The calculated value of $F <$ table value.

$\therefore$ Null hyp. is accepted

i.e. Two samples can be regarded as drawn from two normal population with same variance.

Scanned by CamScanner

Q Two random samples were drawn from normal populations and their values are:

A	66	67	75	76	82	84	88	90	92
B	64	66	74	78	82	85	87	92	93 95

Test whether the two populations have the same variance at 5% level of significance ($F = 3.36$) at 5% level for $v_1 = 10$ and $v_2 = 8$.

Sol

$H_0: \sigma_1^2 = \sigma_2^2$.

$A(X_1)$	$B(X_2)$	$X_1 - 80$ $X_1 - \bar{X}_1$	$(X_1 - \bar{X}_1)^2$	$X_2 - 83$ $X_2 - \bar{X}_2$	$(X_2 - \bar{X}_2)^2$
66	64	−14	196	−19	361
67	66	−13	169	−17	289
75	74	−5	25	−9	81
76	78	−4	16	−5	25
82	82	2	4	−1	1
84	85	4	16	2	4
88	87	8	64	4	16
90	92	10	100	9	81
92	93	12	144	10	100
	95			12	144
720	97		734	14	196
	913				1298

$\bar{X}_1 = \frac{\Sigma X_1}{n_1} = \frac{720}{9} = 80$

$\bar{X}_2 = \frac{\Sigma X_2}{n_2} = \frac{913}{11} = 83$

$s_1^2 = \frac{\Sigma x_1^2}{n_1 - 1} = \frac{734}{8} = 91.75$

$s_2^2 = \frac{\Sigma x_2^2}{n_2 - 1} = \frac{1298}{11-1} = 129.8$

$F = \frac{99.75}{129.8} = 0.768$

$F = 3.36$

Analysis of Variance (ANOVA)

The analysis of variance ~~is~~ frequently referred to by the contraction ANOVA is a statistical technique specially designed to test whether the means of more than two quantitative populations are equal.

The ~~anlay~~ analysis of variance technique, developed by R.A. Fisher in 1920's, is apable of fruitful application to a diversity of Practical Problems. Basically, it consists of classifying and cross-classifying statistical results and testing whether the means of a specified classification differ significantly.

Analysis of variance thus enables us to analysis the total variation of our data into components which may be attributed to various "sources" or "causes" of variation.

In "Sampling and Test of significance" the t-test of the difference of means was discussed. This test is an adequate procedure for

Scanned by CamScanner

$\therefore S_1^2 = \frac{\Sigma(X_1-\bar{X}_1)^2}{n_1-1} = \frac{734}{9-1} = \frac{734}{8} = 91.$

$S_2^2 = \frac{\Sigma(X_2-\bar{X}_2)^2}{n_2-1} = \frac{1298}{11-1} = 129.8$

As. $S_2^2 > S_1^2$ $\therefore H_1: \sigma_2^2 > \sigma_1^2$

Let 5% level of significance

The test statistic is

$$F = \frac{S_2^2}{S_1^2} = \frac{129.8}{91.75} = 1.418$$

<u>Decision</u> Here $v_1 = n_2 - 1 = 10.$

& $v_2 = n_1 - 1 = 8$

$\therefore F_{0.05}(10, 8) = 3.36$

ie The table value of F at $\alpha = 5\%$ level of significance for $v_1 = 10$, $v_2 = 8$ is 3.36.

The calculated value of $F = 1.418 < 3.36$

$\therefore$ Null hyp is accepted.

$\therefore$ it may be calculated that the two populations have the same variance.

Scanned by CamScanner

g the null hypothesis when we have means only two samples to consider. However, in a situation where we have three or more samples to consider at a time an alternative procedure is needed for testing the hypothesis that all samples could likely be drawn from the same population.

Analysis of variance ~~originated~~ originated in agrarian research and its language is thus loaded with such agricultural terms as blocks (referring to land) and treatments (referring to populations or samples) which are differentiated in terms of varieties of seeds, fertilizers or cultivation methods.

Analysis of variance test discussed here is not intended to serve the ultimate purpose of testing for the significance of the difference between two sample variances, rather its purpose is to test for the significance of the differences among sample means. They do this via the mechanism of F-test

Scanned by CamScanner

for testing for the significance of the difference between two variances, but the t is so designed that the variances being compared are different only if the means under consideration are not homogeneous. In this way, significant values of F indicate that the means are significantly different from one another.

Assumptions in analysis of variance

1) Normality:- The values in each group are normally distributed.

2) Homogeneity:- The variance within each group should be equal for all groups. This assumption is needed in order to combine or pool the variances within the groups into a single "within groups" source of variation.

3) Independence of error:- It states that the error should be independent for each value.

Scanned by CamScanner

Techniques of Analysis of variance

1). One-way classification
2) Two-way classification

) One-way classification

In one way classification the data are classified according to only one criterion. The null hypothesis is

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots \mu_K$$

$$H_1 : \mu_1 \neq \mu_2 \neq \mu_3 \neq \cdots \neq \mu_K$$

that is the arithmatic means of populations from which the K samples were randomly drawn are equal to one another. The steps in carrying out the analysis are

I) Calculate variance between the samples

The variance between samples measures the differences between the sample means of each group and the overall mean weighted by the number of observations in each group.

Thus the steps in calculating variance between the samples are

1) first find the sum of all items of various samples, it is denoted by T.

2) Next find correction factor $= \frac{T^2}{N}$

where N = number of ~~terms~~ overall terms of various samples.

3) find Total sum of squares ~~eg~~ suppose we have four samples X_1, X_2, X_3, X_4 then

Total sum of squares $= \Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2 + \Sigma X_4^2 - \frac{T^2}{N}$

4) Now find sum of squares between the samples

$$= \frac{(\Sigma X_1)^2}{N_1} + \frac{(\Sigma X_2)^2}{N_2} + \frac{(\Sigma X_3)^2}{N_3} + \frac{(\Sigma X_4)^2}{N_4} - \frac{T^2}{N}$$

5) Now we find degree of freedom i.e. $v = K-1$,

K = number of samples.

Suppose we have ~~two~~ four samples then

$v = 4-1 = 3$.

6) Now we find Mean square $= \frac{\text{sum of squares}}{\text{degree of freedom}}$

This values becomes variance between samples.

Scanned by CamScanner

Calculate variance within the samples.

The variance within samples measures variability around the mean of each group.

The following steps are used in calculating variance within the samples.

Sum of squares within samples = Total sum of squares − Sum of squares between samples

where Total sum of squares already calculated in (I) step.

Next we find degree of freedom $v = N - K$,

where N = number of terms in various samples,

K = number of samples.

Next we find Mean square = $\frac{\text{sum of squares within samples}}{\text{degree of freedom}}$

This value becomes the value of variance within samples.

I Calculate the ratio F

$$F = \frac{\text{Variance between samples}}{\text{Variance within samples}}$$

Compare the calculated value of F with table value of F for the degree of

If the calculated value of F is greater than table value, it is concluded that the diffe in sample means is significant. Or we can say samples do not come from the same population.

Analysis of Variance (ANOVA) Table: One way Classification

Source of variation	SS (Sum of squares)	v = Degree of freedom	MS (Mean square)	Variance Ratio of F
Between Samples	SSC	$v_1 = K-1$	$MSC = SSC/v_1$	MSC/MSE
Within Samples	SSE	$v_2 = N-K$	$MSE = SSE/v_2$	
Total	SST	$N-1$		

SST = Total sum of squares of variations

SSC = Sum of squares between samples (columns)

SSE = sum of squares within samples (Rows)

MSC = Mean sum of squares between samples

MSE = Mean sum of squares within samples.

Scanned by CamScanner

Example:- To assess the significance of possible variation in performance in a certain test [Page - 1013/eg5] [S.P. Gupta]

Schools

A	B	C	D
8	12	18	13
10	11	12	9
12	9	16	12
8	14	6	16
7	4	8	15

Sol:

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$

$H_1: \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$

Let $\alpha = 5\%$ level of significance.

where $n = 20 < 30$ ∴ we use small sample Test, Sample more than two we use ANOVA

Sample I		Sample II		Sample III		Sample IV	
X_1	X_1^2	X_2	X_2^2	X_3	X_3^2	X_4	X_4^2
8	64	12	144	18	324	13	169
10	100	11	121	12	144	9	81
12	144	9	81	16	256	12	144
8	64	14	196	6	36	16	256
7	49	4	16	8	64	15	225
$\Sigma X_1 = 45$	421	50	558	60	824	65	875

The sum of all the items of various samples

$$T = \Sigma X_1 + \Sigma X_2 + \Sigma X_3 + \Sigma X_4$$

$$= 45 + 50 + 60 + 65$$

$$= 220.$$

Scanned by CamScanner

correction factor $= \frac{T^2}{N}$

$$= \frac{(220)^2}{20} = \frac{18400}{20} = 2420$$

Total sum of squares

$$= \Sigma x_1^2 + \Sigma x_2^2 + \Sigma x_3^2 + \Sigma x_4^2 - \frac{T^2}{N}$$

$$= 421 + 558 + 824 + 875 - 2420$$

$$= 2678 - 2420$$

$$= 258$$

Sum of squares between the samples (SSC)

$$= \frac{(\Sigma x_1)^2}{N_1} + \frac{(\Sigma x_2)^2}{N_2} + \frac{(\Sigma x_3)^2}{N_3} + \frac{(\Sigma x_4)^2}{N_4} - \frac{T^2}{N}$$

$$= \frac{(45)^2}{5} + \frac{(50)^2}{5} + \frac{(60)^2}{5} + \frac{(65)^2}{5} - 2420$$

$$= \frac{2025}{5} + \frac{2500}{5} + \frac{3600}{5} + \frac{4225}{5} - 2420$$

$$= \frac{12350}{5} - 2420$$

$$= 2470 - 2420$$

$$= 50.$$

Sum of squares within the samples (SSE)

= Total sum of squares − sum of squares between the sample

ANOVA TABLE

Source of variation	Sum of squares	Degree of freedom	Mean square
between samples	SSC = 50	$v_1 = k-1 = 4-1 = 3$	16.7
within samples	SSE = 208	$v_2 = N-k = 20-4 = 16$	13.0
Total	258	19	

$$F = \frac{\text{variance between samples}}{\text{variance within samples.}} = \frac{16.7}{13} = 1.285$$

The table value of F for $v_1 = 3$, $v_2 = 16$ at 5% level of significance = 3.24. The calculated value = 1.285 is less than the table value 3.24 so difference in mean values of the samples ~~are~~ is not significance. ie samples could ~~b~~ have come from the same universe.

Analysis of Variance in two-way classification Model

In a one-factor analysis of variance we use one-way classification (one way classification). This treatment constitute different levels of a single factor which is controlled in the experiment. There are many situations in which the

response variable of interest may be affected by more than one factor.

In a two-way classification the data are classified according to two different criteria or factors. The procedure for analysis of variance is somewhat different than the one followed while dealing with problems of one-way classification. In a two way classification the analysis of variance table takes the following form:

Source of Variation	Sum of squares	Degree of freedom	Mean sum of squares	Ratio of F
~~Between Samples~~ Between Columns	SSC	$c-1$	$MSC = SSC/(c-1)$	If $MSC > MSE$ MSC/MSE
Between Rows	SSR	$r-1$	$MSR = SSR/(r-1)$	If $MSR > MSE$ MSR/MSE
Residual or error	SSE	$(c-1)(r-1)$	$MSE = SSE/(r-1)(c-1)$	
Total	SST	$n-1$		

SSC = Sum of squares between columns

SSR = Sum of ~~error~~ squares between rows

SSE = Sum of squares due to error

SST = Total sum of squares

Scanned by CamScanner

Sum of squares for the source 'Residual' is obtained by subtracting from the total sum of squares the sum of squares between columns and rows i.e. $SSE = SST - [SSC + SSR]$

The total number of degrees of freedom $= n-1$ or $cr-1$

where c refers to number of columns

r refers to number of rows.

The F values are calculated as follows:

If $MSC > MSE$ then

$$F_{(v_1, v_2)} = \frac{MSC}{MSE}$$

where $v_1 = c-1$ and $v_2 = (c-1)(r-1)$

and if $MSC < MSE$ then

$$F_{(v_1, v_2)} = \frac{MSE}{MSC}, \text{ where } v_1 = (c-1)(r-1), \; v_2 = c-1$$

~~where $v_1 = c-1$ and $v_2 = (c-1)(r-1)$~~

and similarly if $MSR > MSE$ then

$$F_{(v_1, v_2)} = \frac{MSR}{MSE}$$

where $v_1 = r-1$ and $v_2 = (c-1)(r-1)$

& if $MSR < MSE$ then

$$F_{(v_1, v_2)} = \frac{MSE}{MSR}$$

where $v_1 = (c-1)(r-1)$, $v_2 = r-1$

Scanned by CamScanner

The calculated values of F are compared with the table values. If calculated value of F is greater than the table value at pre-assigned level of significance, the null hypothesis is rejected, otherwise accepted.

It would be clear from above that in problems involving two-way classification, 'Residual' is the measuring rod for testing significance. It represents the magnitude of variation due to forces called 'chance'.

~~Example 2: Lokayal (S.P. Gupta)~~

Coding of Data

The original measurement can be coded to simplify calculations without the need for any subsequent adjustments of the results.

Coding refers to the addition, multiplication, subtraction or division of data by a constant.

Scanned by CamScanner

mple 7:- 1020 Page (S.P. Gupta)

The above data are classified according to criteria or factors (i) salesman (ii) seasons. In order to simplify calculations we code the data by subtracting 30 from each figure. The data in the coded form are given below

Seasons	Salesmen				Seasons Total
	A	B	C	D	
Summer	6	6	-9	5	8
Winter	-2	-1	1	2	0
Monsoon	-4	-2	-1	-1	-8
Salesmen's Total	0	3	-9	6	Grand Total T=0

Correction factor $= \frac{T^2}{N} = \frac{0^2}{12} = 0$

N = number of items = 12.

Sum of squares between salesmen (columns)

$$= \frac{(0)^2}{3} + \frac{(3)^2}{3} + \frac{(-9)^2}{3} + \frac{(6)^2}{3} - \frac{T^2}{N}$$

$$= 0 + 3 + 27 + 12 - 0 = 42$$

$v = df = 4-1 = 3$

Scanned by CamScanner

Sum of squares between seasons (Rows).

$$= \frac{(8)^2}{4} + \frac{(0)^2}{4} + \frac{(-8)^2}{4} - \frac{T^2}{N}$$

$$= 16 + 0 + 16 - 0$$

$$= 32$$

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

Total sum of squares

$$= (6)^2 + (-2)^2 + (-4)^2 + (6)^2 + (-1)^2 + (-2)^2 + (-9)^2 + (1)^2 + (-1)^2 + (.5)^2 + (2)^2 + (-1)^2 - \frac{T^2}{N}$$

$$= 210 - 0$$

$$= 210.$$

$v = d.f = 12 - 1 = 11$

The above information will be presented in the following table of Analysis of Variance:

Source of Variation	Sum of squares	d.f.	Mean squares
Between columns (Salesmen)	42	3	14
Between rows (Seasons)	32	2	16
Residual	136	6	22.67

Scanned by CamScanner

st. compare the salesmen variance estimate with residual variance estimate,

$$F = \frac{22.67}{14} = 1.619$$

The Table value of F for $v_1 = 6$, $v_2 = 3$ at 5% level of significance is 8.94

The calculated value is less than the table value so we conclude that the sales of different salesmen do not differ significantly.

Now let us compare the season variance estimate with the residual variance estimate thus

$$F = \frac{22.67}{16} = 1.417$$

The critical value of F for $v_1 = 6$ and $v_2 = 2$ at 5% level of significance is 19.33

The calculated value is less than the table value and hence there is no significant difference in the seasons as far as the [illegible]

Scanned by CamScanner

Thus, the test shows that the salesmen [illegible] seasons are alike so far as the [illegible] are concerned.'

Scanned by CamScanner

26c

Q. What are the two types of errors in testing the hypothesis?

Ans: The main aim of sampling a theory is to draw conclusions about the population parameter on the basis of sample results. In doing this we may commit two types of errors.

1) Type I error
2) Type II error

1) Type I error :- when H_0 is true, we may reject it $P(\text{Reject } H_0 \text{ when it is true}) = P(\text{Reject } H_0/H_1) = \alpha$

α is called the size of the type I error also referred to as producer's risk.

2) Type II error :- when H_0 is wrong, we may accept it.

$P(\text{Accept } H_0 \text{ when it is wrong}) = P(\text{Accept } H_0/H_1) = \beta$

β is called the size of the type II error, also referred to as consumer's risk.

Scanned by CamScanner

Skewness :- The term "skewness" refers to lack of symmetry. ie when a distribution is not symmetrical. is called a skewed distribution.

Definitions

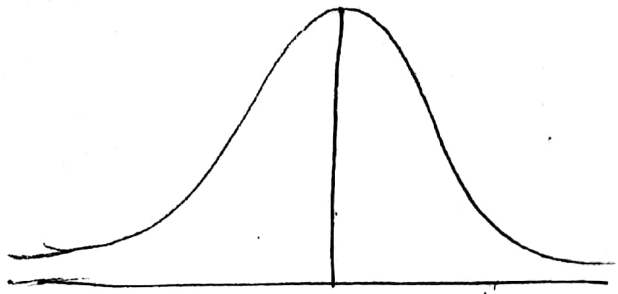
1) According to Morris Hamburg, "Skewness refers to the asymmetry or lack of symmetry in the shape of a frequency distribution."

2) According to Simpson & Kafka, "Measures of skewness tell us the direction and the extent of skewness. In symmetrical distribution the mean, median and mode are identical. The more the mean moves away from the mode, the larger asymmetry or skewness."

We may have Positively skewed distribution ~~and~~ or Negatively skewed distribution.

1) Symmetrical distribution :- When the values of mean, median and mode are identical. The ~~spre~~ spread of the frequencies is the same on both sides of the centre point of the curve.

Scanned by CamScanner



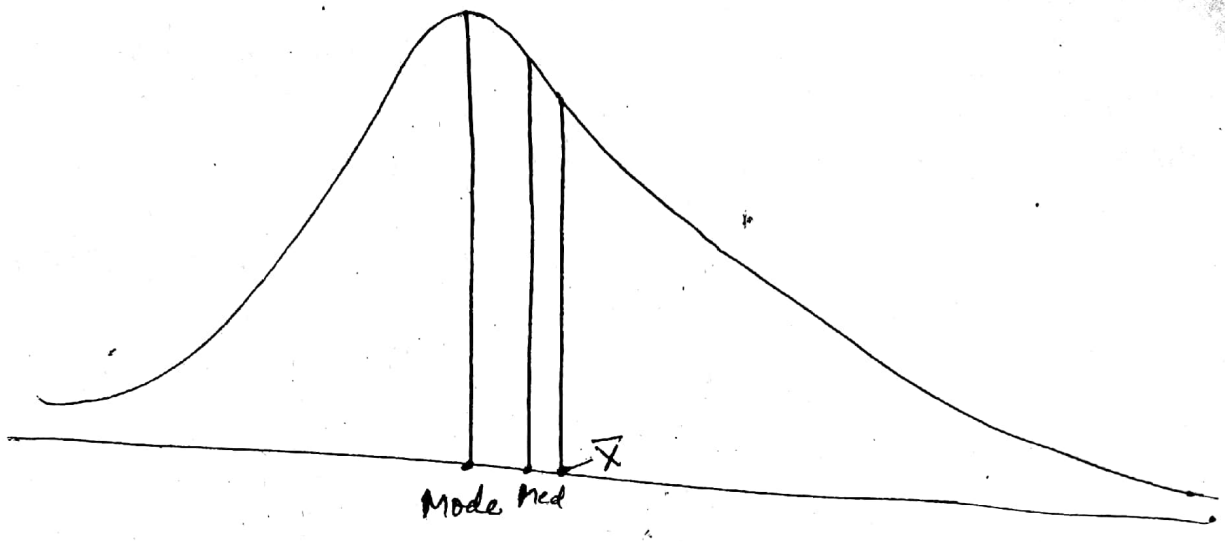
Mean = Median = Mode.

2) Asymmetrical Distribution :- A distribution which is not symmetrical is called a skewed distribution or Asymmetrical Distribution. which is of two type Positively skewed Dist. & Negatively skewed Dist.

(a) Positively skewed Distribution :- In this distribution the value of mean is maximum and that of mode least, the median lies in between the two. In the +vely skewed dist. the frequencies are spread out over a greater range of values on the high value end of the curve (the right hand side) than they are on the low value end.

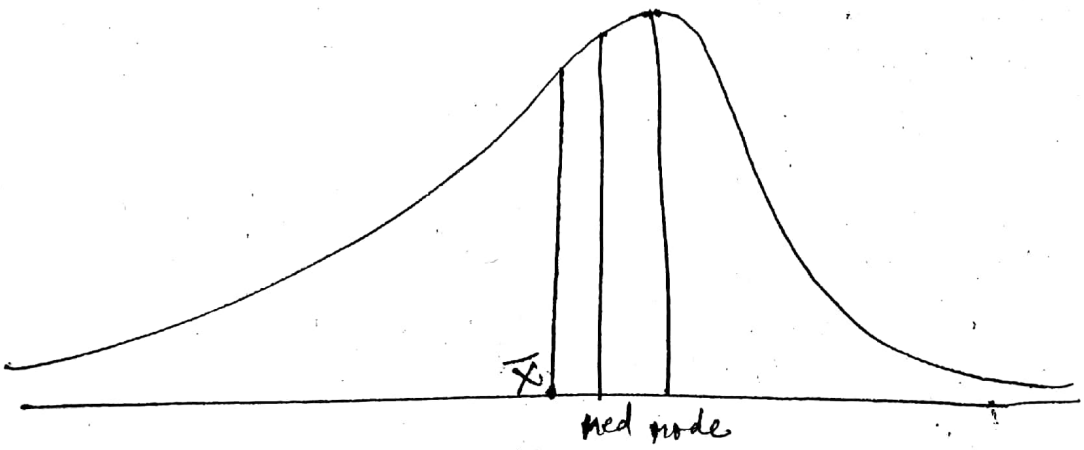
~~(b) Negatively skewed distribution :-~~

Scanned by CamScanner



Positively skewed distribution

(b) <u>Negatively</u> <u>skewed</u> <u>distribution</u> :- In this distribution the value of mode is maximum and that of mean least, the median lies in between the two. In this distribution the position is reversed ie the excess tail is on the left-hand side.



Negatively skewed distribution.

Scanned by CamScanner

It should be noted that in symmetrical distributions the internal between the mean and the median is approximately one-third of the internal between the mean and the mode.

Tests of skewness :-

Skewness is present if

1) The values of mean, median and mode donot coincide.
2) When the data are plotted on a graph they do not give the normal bell-shaped form.
3) The sum of the positive deviations from median is not equal to the sum of the negative deviations.
) Frequencies are not equally distributed at points of equal deviation from the mode.

Moments :- " Moments is a familiar mechanical term which refers to the measure of a force with respect to its tendency to provide rotation. The strength of the tendency depends on the amount of force and the

Scanned by CamScanner

distance from the origin of the point at which the force is exerted. If the number of forces $f_1, f_2, \ldots, f_n$ at distances $x_1, x_2, x_3 \ldots, x_n$ are applied, the moment of first force about the origin is $f_1 x_1$, the moment of the second force about the origin is $f_2 x_2$. etc. These moments are additive so that Σfx is the total moment about the origin. If Total moment is divided by the total forces then the formula is

$\frac{\Sigma fx}{\Sigma f = N}$ is ~~the~~ called 'moment'. These are the moments about the origin.

and the moments about mean is called 'central moment' are denoted by Greek letter μ thus μ_1 = first moment about mean

μ_2 = second moment about mean

$$\mu_1 = \frac{\Sigma f(x-\bar{x})}{N}$$

$$\mu_2 = \frac{\Sigma f(x-\bar{x})^2}{N}$$

$$\mu_3 = \frac{\Sigma f(x-\bar{x})^3}{N}$$ etc.

Scanned by CamScanner

and $\mu_4 = 0$. ie μ_4 always equal to zero. (4)

Two important constants of a distribution are calculated from μ_2, μ_3, μ_4. They are

1) $\beta_1 = \dfrac{\mu_3^2}{\mu_2^3}$

2) $\beta_2 = \dfrac{\mu_4}{\mu_2^2}$

β_1 measures skewness.

β_2 measures Kurtosis.

<u>Kurtosis</u>:- It is the degree of Peakedness of a distribution, usually taken relative to a normal distribution. Thus measures of Kurtosis tell us the extent to which a distribution is more Peaked or flat-topped than the normal curve. Kurtosis in Greek means "bulginess". In statistics kurtosis refers to the degree of flatness or

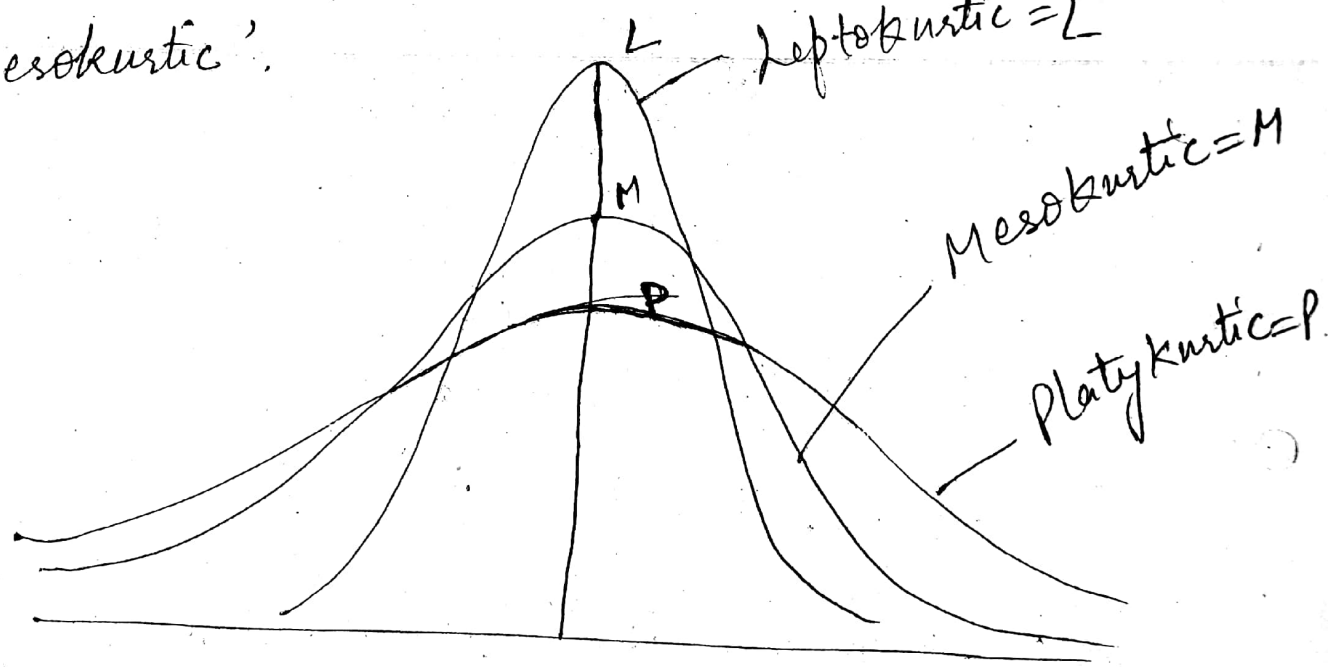
Scanned by CamScanner

peakedness in the region about the mode of a frequency curve.

1) If a curve is more peaked than the normal curve, it is called 'leptokurtic'

2) If a curve is more flat-topped than the normal curve, it is called 'platykurtic'

3) The normal curve itself is known as 'mesokurtic'.



<u>Definition</u>: According to Spiegel, "Kurtosis is the degree of peakedness of a distribution usually taken relative to a normal distribution."

Scanned by CamScanner

Roll No. Total No. of Pages : 02

Total No. of Questions : 09

B.Sc.(BT) (2009 & Onwards) (Sem.–1)

BIOSTATISTICS

Subject Code : BSBTM-109

Paper ID : [F0239]

Time : 3 Hrs. Max. Marks : 60

INSTRUCTION TO CANDIDATES :

1. SECTION-A is COMPULSORY consisting of TEN questions carrying TWO marks each
2. SECTION-B contains FIVE questions carrying FIVE marks each and students has to attempt any FOUR questions.
3. SECTION-C contains THREE questions carrying TEN marks each and students has to attempt any TWO questions.

SECTION - A

1. *Write briefly :*

a. Define population and sample.

b. Define quantitative and qualitative variables

c. Define descriptive statistics.

d. What is cumulative frequency curve?

e. Explain histogram

f. Define permutation and combination

g. What are Type-I and type-II errors?

h. Define standard deviation.

i. What is significance of using F test?

j. Define with example what is percentile

SECTION - B

2. Enlist various sources of data. Explain any one in detail

3. The following frequency data was observed for the adult cocaine users

Lifetime frequency of cocaine use	Male	Female	Total
1-19 times (A)	32	07	39
20-99 times (B)	18	20	38
100+ times (C)	25	09	34
Total	75	36	111

Suppose we pick a subject at random from the 111 subjects and find that he is a male. What is the probability that this male will be one who has used cocaine 100 times or more during his life time?

4. Explain with example what Poisson distribution is and where it is used

5. What are applications of Analysis of variance?

6. Explain the steps involved in hypothesis testing procedure

SECTION - C

7. Explain in detail Binomial distribution with examples

8. Calculate the chi-square value for the following data.

Color	Red	Green	Yellow
Observed frequency	12	16	20
Expected frequency	16	8	25

9. Compare characteristics of mean, mode and median with examples

Scanned by CamScanner

Roll No. [ ]

Total No. of Pages: 02
Total No. of Questions: 09

B.Sc (Biotechnology) (Sem.-1st)

BIOSTATISTICS

Subject Code: BSBT-109
Paper ID: [F0239]

Time: 3 Hrs. Max. Marks: 60

INSTRUCTIONS TO CANDIDATE:

1. Section – A, is Compulsory.
2. Attempt any four questions from Section-B.
3. Attempt any two questions from Section-C.

Section –A (10x2=20)

Q.1

(a) What is inferential Biostatistics?

(b) IF $\log_{10}2=0.3010$ and $\log_{10}3=0.4771$ then find the value of $\log_3 512$.

(c) If $\log 27=1.1431$ then find the value of $\log 9$.

(d) Explain the classification of data.

(e) What is Histogram?

(f) What is Ogive curve?

(g) If the mean of [illegible] 28, x, 42, 78 and 104 is 62, then find the mean of 48, 62, [illegible] and x?

(h) What is median?

(i) What is Hypothesis?

(j) What do you mean by independence of random variable?

Section –B (4x5=20)

Q.2. What are the various Aims of Bio stat?

Q.3 Test Score for a class of 20 students are as follows:

93,84,97,98,100,78,86,100,85,92,72,55,91,90,75,94,83,60,81,95

Solve the following

(a) Complete the table

Test Score	Frequency
91-100	
81-90	
71-80	
61-70	
51-60	

(b) Find the model interval?

(c) Find the interval that contain median?

Q.4. The marks obtained by a student in different subjects:

100, 85,55,95,75,100.

For this data, how many marks are within one standard deviation of the mean?

Q.5. What is F-Test? What are various benefits of the F-Test?

Q.6 Calculate (i) Quartile Deviation (Q.D)

(ii) Mean deviation (M.D) from mean for the following data.

Marks	No. of students
0-10	6
10-20	5
20-30	8
30-40	15
40-50	7
50-60	6
60-70	3

Section –C (2x10=20)

Q.7 State and prove baye's Theorem?

Q.8. Ten coins are thrown simuhaneously. Find the probability of getting at least seven heads.

Q.9 What is Chi-square test? What are various benefits and limitations of Chi-Square test?

---:END:---

Scanned by CamScanner

Roll No.....

Total No.of Pages:02

BSC(BT)-2014

BIOSTATISTICS

Paper Code (BSBTM-109)

Paper Id. [F0239]

Time allowed: 3 Hrs

Max. Marks: 60

Note: There are three sections A, B and C and each section carry 20 marks

Section A

1. Write short notes on the following 10 x 2 = 20

a) Sample
b) Confidence limits
c) Cluster sampling
d) Mean deviation
e) Skewness
f) Test of significance
g) Contingency table
h) Probability
i) Correlation
j) Deciles

Section B

Answer any four questions 4 x 5 = 20

2. Write notes on binomial and Poisson distribution.
3. Briefly describe the probability laws.
4. Briefly describe how do you present quantitative data?

5. Emergency hospital admissions in USA with heart disease are found to follow Poisson distribution. An investigation showed that there are four admissions per day (mean). Find the probability that exactly three emergency admissions will occur in a given day?

6. Velocities of height (cm) in 2 groups of infants with calorie protein adequacies in their dietaries are 5.5 and 3.5 with SD of 1.5and 2.0. Can we conclude that velocities of height are higher in those of infants than the other? (n = 10 for each group)

P.T.O

Section C

Answer any two questions 2 x 10 = 20

7. Following data gives the protein adequacy and calorie adequacy of pre-school children. Can we conclude that with calorie-protein adequacies are associated in preschool dietaries? Apply χ^2 test.

Details		Protein adequacy		Total
		+	--	
Calorie adequacy	+	90	10	100
	--	66	34	100
Total		156	44	200

8. Calculate the correlation coefficient and coefficient of determination between birth and death rates observed for five countries given below.

Details	India	USA	Japan	China	Sweden
Birth Rate (X)	30	15	11	41	13
Death Rate (Y)	11	9	7	15	10

9. a) Describe the application of computers in Biostatistics.
b) Explain the importance of ANOVA.

----------END----------

Scanned by CamScanner

Roll No. ..... Total No. of Pages : 2

Total No. of Questions : 09

B.Sc (BT) (Sem.–1)

BIOSTATISTICS

Subject Code : BSBTM-09 (2009-2011 Batch)

Paper ID : [F0239]

Time : 3 Hrs. Max. Marks : 60

INSTRUCTION TO CANDIDATES :

1. SECTION-A is COMPULSORY.
2. Attempt any FOUR questions from SECTION-B.
3. Attempt any TWO questions from SECTION-C.

SECTION-A (10 × 2 = 20 Marks)

1. Write short notes on :
   (a) Frequency distribution
   (b) Kurtosis
   (c) Coefficient of variation
   (d) Quartiles and deciles
   (e) Standard error
   (f) Normal distribution
   (g) Hypothesis testing
   (h) Geometric Mean
   (i) Mean deviation
   (j) Binomial distribution

SECTION-B (4 × 5 = 20 Marks)

2. What is paired t test and how does it differ from unpaired test?
3. What is probability? Explain different probability laws.
4. What is test of significance? Briefly describe the procedure and steps in utilization of any test of significance.

CHANDIGARH COLLEGE OF PHARMACY * LANDRAN (MOHALI) * MAY 19 LIBRARY

5. Emergency hospital admissions in USA with heart disease are found to follow Poisson distribution. An investigation showed that there are four admissions per day (mean). Find the probability that exactly three emergency admissions will occur in a given day.
6. The body surface area of 15 children are 196, 101, 184, 227, 253, 185, 217, 122, 336, 148, 114, 135, 233, 198 and 109. Calculate the mean, variance, standard variation and standard error.

SECTION-C (2 × 10 = 20 Marks)

7. An experiment was conducted to test the effect of 4 different training methods on increasing mental ability of children. Test whether the four methods of training differ significantly with each other by using one way ANOVA.

Training 1	Training 2	Training 3	Training 4
15	12	10	14
20	17	15	20
16	11	12	16
19	18	14	18
19	16	14	13

8. Calculate the correlation coefficient and coefficient of determination between birth and death rates observed for five countries given below

Details	India	USA	Japan	China	Sweden
Birth Rate (X)	30	15	11	41	13
Death Rate (Y)	11	9	7	15	10

9. What is ANOVA? Briefly describe when will you use one way analysis of variance and describe the least significant difference (LSD) test procedure

Punjab Technical University

Scanned by CamScanner