Muna H. AL-Saeed

The Normal Distribution المنحنى الطبيعي الاعتدالي

Population &Sample -:المجتمع والعينة

المجتمع :- عبارة عن جمع القيم او المفردات التي يمكن ان ياخذها المتغير ويمكن ان يكون مجتمعا محدودا او غير محدود.

العينة :- جزء من المجتمع وهي مجموعة من المشاهدات اختيرت بطريقة ما من المجتمع(أي العينة تكون جزء من المجتمع)

Normal Distribution المنحنى الطبيعي

يمثل الشكل المنحنى الطبيعي ان المنحنى الطبيعي القياسي يشبه شكل الناقوسمع انحدار الاطراف نحو خط الاساس, ان العمود المقام على مركز المنحنى يمثل وسط التوزيع حيث انه يشطر التوزيع الممثل بهذا المنحنى الى جزئين متساويين فانه لذلك يمثل الوسيط .

وسيط المجتمع يطلق عليه (μ) وسط العينة عليه يطلق $(\bar{X})$. اذن(μ)هو واحد من معالم المجتمع وهذا من خلاله يمكن ان نحدد توزيع الجماعة.

Variance -: قياسات التشتت

مقدار تشتت البيانات عن الوسط ويسمى التباين σ^2 وهو واحد ايضا من معالم الجماعة او المجتمع.

وهناك قياس اخر هو الانحراف المعياري SD او σ ويرمز للتباين بالنسبة للعينة الرمز S^2 .

مواصفات اخذ عينة من مجتمع ما:-

1- توخذ بشكل غير متحيز أي عشوائي.
2- ان تكون قليلة الى الحد الذي لايمكن ان تمثل المجتمع ولاكبيرة بحيث تفقد الهدف من اخذ العينة.

*لاحظ جداول العشوائية

هناك نوعين من البيانات

1- بيانات وصفية او بيانات عد مثاله بيانات جمع لون الشعر وتسمى مقياس بيانات العد.
2- بيانات القياس مثل قياس الطول , الوزن , الزمن...........الخ

اذن في حالة الاحصاء يجب وضع شيئين اساسين هما

1- طبيعة بيانات القياس.

2- تحديد الاختبارات المناسبة.

هناك نوعين من الاختبارات:-

1- اختبارات معلمية : - Paramatric test

2- اختبارات لا معلمية:- Nonparamatric test

الاختبارات المعلمية :- هو ان تختبرالبيانات مع ربط μ و σ بمعالم الجماعة أي تعتمد على معالم المجتمع.

الاختبارات اللامعلمية :- هي اختبارات تقوم به او بها دون الحاجة الى تحديد معالم الجماعة ودائما يعرف احصائيا بان الاختبارات المعلمية تكون نتئجها اقوى من الاختبارات اللامعلمية

Measures of central Tendency:-

1-The Arithmetic Mean
2-The Geometric Mean
3-Harmonic Mean
4-The Quadratic Mean
5-The Median
6-The Mode

الوسط الحسابي:- الوسط الحسابي لقيم ما هو القيمة الناتجة من قسمة مجموع تلك القيم على عددها ويرمز لها بالرمز $\bar{Y}$

طرق حسابه: بيانات مبوبة
بيانات غير مبوبة

بيانات غير مبوبة :-

$$\bar{y} = \frac{\sum y_i}{n}$$

Example:-The data represent weight of 5 calves(kg)in animal field of Basrah city. 400, 380,450,350,520

$$\bar{y} = \frac{\sum y_i}{n} = \frac{400+380+450+350+520}{5} = \frac{2100}{5} = 420\text{Kg.}$$

بيانات مبوبة

$$\bar{y} = \frac{\sum f_i y_i}{\sum f_i}$$

Example2:The data represent degree of student of second class in biostatic.

y_i	f_i	y_i	$f_i y_i$
31-40	1	35.5	35.5
41-50	2	45.5	91
51-60	5	55.5	277.5
61-70	15	65.5	982.5
71-80	25	75.5	1887.5
81-90	20	85.5	1710
91-100	12	95.5	1146
	80		61300

$$\bar{y} = \frac{\sum f_i y_i}{\sum f_i} = \frac{61300}{80} = 76.62$$

Statistics

Statistics is concerned with scientific methods for collecting, organizing, summarizing, presenting, and analyzing data as well as with drawing valid conclusions and making reasonable decisions on the basis of such analysis.

In a narrower sense, the term statistics is used to denote the data themselves or numbers derived from the data, such as averages. Thus we speak of employment statistics, accident statistics, etc.

Population and Sample; Inductive and Descriptive Statistics

In collecting data concerning the characteristics of a group of individuals or objects, such as the heights and weights of students in a university or the numbers of defective and non defective bolts produced in a factory on a given day, it is often impossible or impractical to observe the entire group, especially if it is large. Instead of examining the entire group, called the population, or universe,one examines a small part of the group, called a sample.

A population can be finite or infinite. For example, the population consisting of all bolts produced in a factory on a given day is finite, whereas the population consisting of all possible outcomes (heads,tails) in successive tosses of a coin is infinite.

If a sample is representative of a population, important conclusions about the population can often be inferred from analysis of the sample. The phase of statistics dealing with conditions under which such inference is valid is called inductive statistics, or statistical inference. Because such inference cannot be absolutely certain, the language of probability is often used in stating conclusions.

The phase of statistics that seeks only to describe and analyze a given group without drawing any conclusions or inferences about a larger group is called descriptive, or deductive, statistics.

Statistics

Describing a distribution:-

A series of measurments or count is called a distribution. Table1 shows such a distribution, purposely kept very simple so the principles involved can be shown without getting all tangled up in arithmetic at this stage of the game.

X $(X-\bar{\bar{X}})=$	x	x^2	
10 (10-5)=	+5	25	N=8
8 (8-5) =	+3	9	$\sum X \quad 40$
6 (6-5) =	+1	1	
5 (5-5) =	0	0	$\bar{X} = 5$
5 (5-5) =	0	0	$\sum x = 0$
3 (3-5) =	-2	4	$\sum x^2 = 64$
2 (2-5) =	-3	9	$S^2 = 8$
1 (1-5) =	-4	16	$S = 2.82$

First, note that there are 8numbers, or members, in the distribution shown in Table 1. This is the N number of the distribution, therefore N=8.
Also note that upper-case X is used to denote the members of the distribution.
If we sum all the X in the distribution, we obtain $\sum X = 40$. The Greek letter sigma $(\sum)$ is used to denote " the sum of" and is one of the most frequently used symbols in statistics.

$$\sum_{i=1}^{n} X = 40$$

The mean is symbolized by $\bar{X}$, $\quad \bar{X} = \frac{\sum X}{N}$,

$$S^2 = \frac{\sum x^2}{N} \quad \text{or} \quad S^2 = \frac{\sum (X-\bar{X})^2}{N} \quad \text{or} \quad S^2 = \frac{\sum X^2 - (\sum X)^2/N}{N}$$

$$S^2 = \frac{N\sum X^2 - (\sum X)^2}{N^2}$$

$$S = \sqrt{S^2} = \sqrt{8} = 2.82$$

-The Normal Distribution

1- Measurement Data

There are numerous characteristics found in the biotic and physical environments which can be measured in some fasion. Variables such as height, weight, length, temperature, and hemoglobin content are some familiar source of measurement, or continuous data.

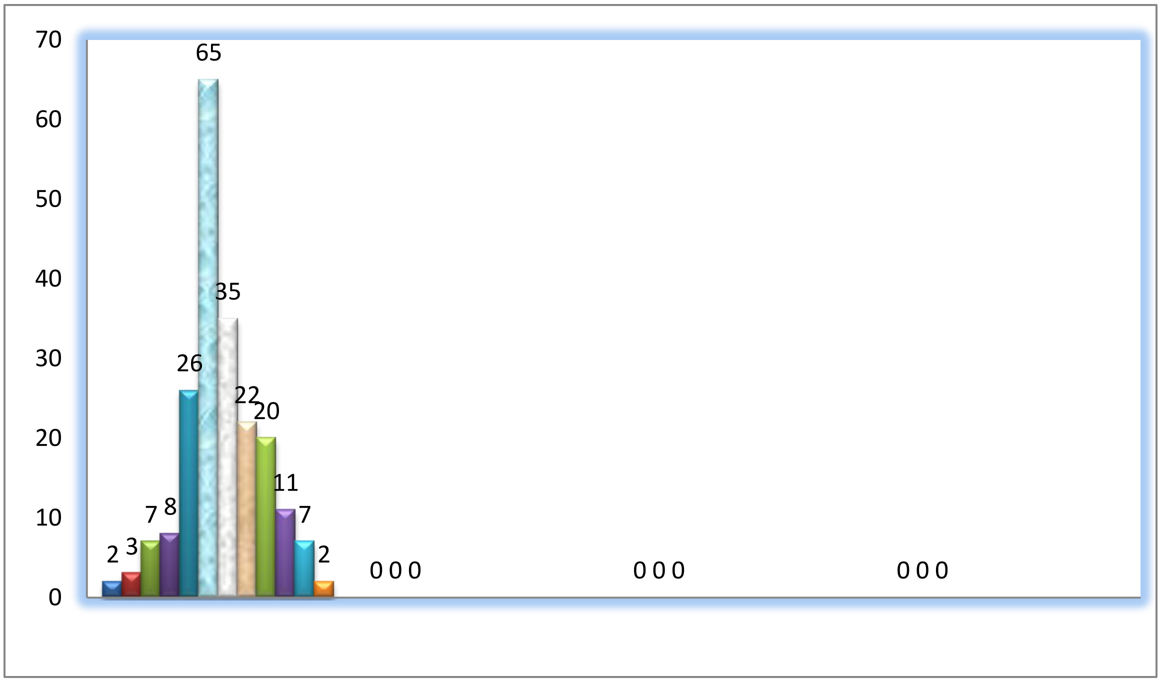
These data are continuous because they may assume any value at all on a given continum existing between two limits. Suppose that a large population of birds has bills ranging in length from 9 to 15mm. It would then be theoretically possible to find birds in this population with bill lengths matching every conceivable measurement between the two extremes. It is as though the bill of one bird suddenly started to grow, like Pinocchio's nose, from 9mm to 15mm without missing any possible measurement along the way!

In practice, measurements are limited by the precision of the measuring instruments and by human error. Even the most sophisticated analytical balance cannot supply the exact weight of an object, and even the most precise instruments are subject to error in use and interpretation. We therefore take measuremnts to the nearest centimeter, millimeter, milligram, microgram, or whatever is dictated by the nature and precision of the instrument, admitting that the reported measurement will almost certainly deviate from the true value. Actually, when a measurement is reported as being 50mm, this implies that the true value lies somewhere between 49.50 mm and 50.50mm.

In practice, continuous data is therefore to some extent discrete, or discontinuous. Also, variables such as pulse rate or the number of fin rays on a fish are discrete in nature, but under certain conditions may be treated as measurement data.

-Normally Distributed Data

Shows a histogram based on measurements obtaied from 120 lima beans drawn one at a time a half-bushel container. Each bean was measured to the nearest millimeter along its longest axis. The histogram is constructed by plotting length



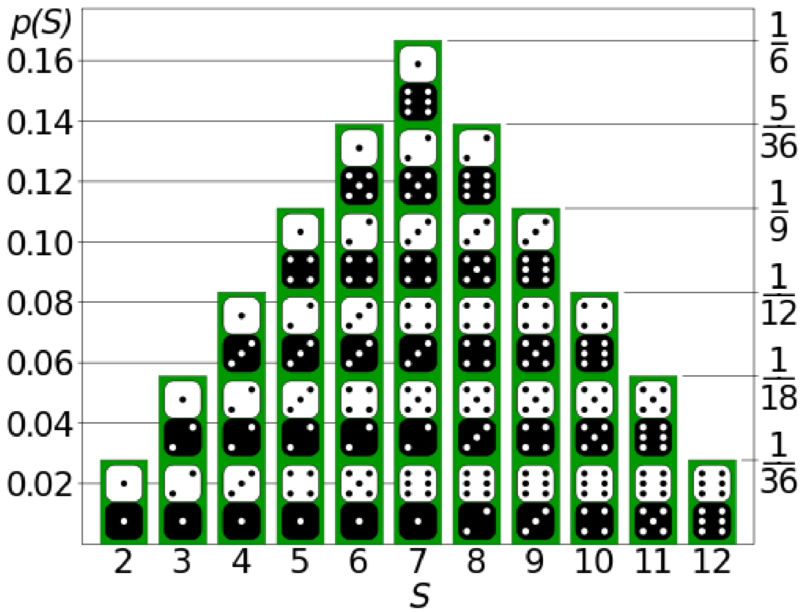
The area of each bar therefore represents the frequency of cases found between the real upper and lower limits of the associated measurement. Thus since 65cases are associated with a length of 20mm, this meants that 65 cases have measurement values between 19.50mm and 20.50mm.
It is important to note that even with this relatively small sample, the majority of cases tend to cluster around the central portion of the histogram.

Probability:-

The term probability is one commonly used in ordinary conversation.
In statistics, we attempt to express probability in quantitative term. The basis for this quantitative expression may be built into the situation, as with a coin. In that event it is called a priori prabability, i.e established "before the fact" Or we may formulate a quantitative tement based entirely on past experience; in this case we are dealing with empirical probability.
We assume, on the basis of meiotic division, that one-half the sperm produced by the human male carries the *X*-chromosome and one-half carries the *Y*-chromosome.

To define probability distributions for the simplest cases, one needs to distinguish between **discrete** and **continuous** random variables. In the discrete case, one can easily assign a probability to each possible value: for example, when throwing a die, each of the six values *1* to *6* has the probability 1/6. In contrast, when a random variable takes values from a continuum, probabilities can be nonzero only if they refer to intervals: in quality control one might demand that the probability of a "500 g" package containing between 490 g and 510 g should be no less than 98%.

If the random variable is real-valued (or more generally, if a total order is defined for its possible values), the cumulative distribution function (CDF) gives the probability that the random variable is no larger than a given value; in the real-valued case, the CDF is the integral of the probability density function (pdf) provided that this



The Chi-Square Test:-

Observed and theoretical frequencies:-

As we have already many times, the results obtained in sample do not always agree exactly with the theortical results expected according to the rules of probability. For example, although theoretical considerations lead us to expect 50 heads and 50 tails when we toss a fair coin 100 times, it is rare that these results are obtained χSuppose that in a particular sample a set of possible events E_1, E_2, E_3,……,E_k ar observed to oocur with frequencies $o_1,o_2,o_3,\ldots\ldots..o_k$, called observed frequencies and that according to probability rules they are expected to occure with frequencies $e_1,e_2,e_3,\ldots\ldots\ldots\ldots e_k$,called expected , or theoretical, frequencies. Often we wish to know whether the observed frequencies differ significantly from the expected frequencies.

Definition of χ^2:-

Ameasure of discrepancy existing between the observed and expected frequencies is supplied by the statistic χ^2 (reach Chi-square) given by

$$\chi^2 = \frac{(o_1-e_1)^2}{e_1} + \frac{(o_2-e_2)^2}{e_2} + \frac{(o_3-e_3)^2}{e_3} + \cdots\cdots\cdots + \frac{(o_k-e_k)^2}{e_k} = \sum_{j=1}^{k} \frac{(o_j-e_j)^2}{e_j}$$

where if the total frequency is N

$$\sum o = \sum e = N$$

An

Event	E_1	E_2	E_3	…….	E_k
Observed frequency	o_1	o_2	o_3	…….	o_k
Expected frequency	e_1	e_2	e_3	……..	e_k

expression equivalent to formula is

$$\chi^2 = \sum \frac{o^2}{e} - N$$

If $\chi^2 = 0$, the observed and theoretical frequencies agree, while if $\chi^2 > 0$, they do not agree exactly. The larger the value of χ^2, the greater is discrepancy between the observed and expected.

The sampling distribution of χ^2 is approximated very closely by the Chi-square distribution.

The expected frequencies are at least equal to 5. The approximation improves for larger values.

The number of degrees of freedom df, is given by df=k-1 if the expected frequencies can be computed without having the popultion parameters from sample statistics.

Significances tests:-

In practice, expected frequencies are computed on the basis of a hypothesis H_o . If under this hypothesis the computed value of χ^2 given by equation.

A review of this problem shows that, based on 100 tosses of an unbiased coin, the probability of obtaining 65 or more heads by chance alone was only 0.0019. we there fore rejected the H_o (the null hypothesis) of (no difference) between 65 heads and the expected 50 heads, and concluded that coin was not fair, but biased in favor of heads.

It should be obvious that the size of obtained chi-sqquare value will be determined by the magnitude of the differences between the observed and expected frequencies. Large differences will produce a large value of chi-square; small differences will produce smaller chi-square values, and if no differences at all exist, then X^2=0.

In our present example, we will need to slightly modify the basic formula chi-square is based on a discrete, not a continuous variable. We therefore need to correct for continuity by subtracting 0.50 from the absolute difference between each observed and expected frequency combination. This is called the Yates correction factor and is to be used with $2\times 2\ tales\ only$.

<u>**Example 1:-**</u> **Tosses 100 times of the coin the observed 65heads and 35 tail.**

The expected number, if the coin are fair, are determined from the distribution of X^2

Expected Head = $\frac{1}{2}\times 100$=50

Expected Tail = $\frac{1}{2}\times 100$=50

Now we shall use chi-square to determine whether 65 heads and 35 tails is a significant departure from the 50 heads and 50 tails that one would expect from 100 tosses of an honest coin. By now, we are much too sophisticated about these things to really expect exactly 50 heads and 50 tails every time a coin is tossed 100 times, but based on the probability value of $\frac{1}{2}$ that is attached to a fair coin, we

should certainly be suspicious if the deviation of the observed frequencies from the expected frequencies is unusually large!
H_o: The observed frequency not differ(=) the expected frequency.
H_A: The observed frequency differ ($\neq$)the expected frequency.

$$X^2 = \sum \frac{(O-E)^2}{E}$$

$$X^2 = \sum \frac{[(O-E)\ -0.5]^2}{E}$$

	Heads	Tails	
Observed	65	35	100
Expected	50	50	100
	115	85	

$$= \frac{[(65-50)-0.5]^2}{50} + \frac{[(35-50)-0.5]^2}{50}$$

$$= \frac{(14.5)^2}{50} + \frac{(14.5)^2}{50} = 8.40$$

K=2
DF=k-1
DF=2-1=1
The obtained value=8.40
The tabulated value in 0.01 level=6.635
The tabulated value in 0.05 level=7.8**79**
We therefore rejected the null hypothesis(H_o) of no difference between observed and expected frequency.
Accepted alternative hypothesis(H_A) of difference between observed and expected frequency.

<u>Example 2</u>:-There is a genetic model which assumes that black coat color in mice is inherited as a simple dominant trait, and that brown color is inherited as a recessive trait. Across between pairs of heterozygous black mice produced an F_2 generation consisting of 220 black mice and 60 brown mice.

1- According to our genetic model, across between heterozygous black mice would produce offspring as follows:

Bb× Bb

F_2 BB, Bb, Bb, bb

H_o: The observed frequency not differ(=) the expected frequency.
H_A: The observed frequency differ ($\neq$)the expected frequency.
Which represent a phenotype ratio of 3 black mice to 1 brown.

2- Now, if the total of 280 offspring occurred in exactly a 3:1 phenotype ratio, as expected from our genetic model, we would have $\frac{3}{4}\times 280$ and $\frac{1}{4}\times 280$, or 210 black and 70 brown mice.

Expected Black = $\frac{3}{4}\times 280 = 210$

Expected Brown = $\frac{1}{4}\times 280=70$

3Our next step involves setting up a chi-squre table as follows:

	Black	Brown	
Observed	220	60	280
Expected	210	70	280

$$X^2 = \sum \frac{(O-E)^2}{E}$$

$$X^2 = \sum \frac{[(O-E)\ -0.5]^2}{E}$$

$$= \frac{[(220-210)-0.5]^2}{210} + \frac{[(60-70)-0.5]^2}{70}$$

$$= \frac{(9.5)^2}{210} + \frac{(9.5)^2}{70} = 0.42 + 1.28=1.70$$

K=2
DF=k-1
DF=2-1=1
The obtained value=1.70
The tabulated value in 0.01 level=6.635
The tabulated value in 0.05 level=7.8**79**
We therefore accepted the null hypothesis(H_o) of no difference between observed and expected frequency.
rejected alternative hypothesis(H_A) of difference between observed and expected frequency.

<u>Example 3</u>:- Suppose that two dihyrids are crossed in a situation where complete dominance is assumed. It is further assumed that no linkage or other complicating factor are present. We therefore have the genetic model.

AaBb × AaBb

F_2 9A-B-, 3A-bb, 3aaB-, 1aabb

Which is the classic 9:3:3:1 phenotype ratio. Now, suppose that the actual F_2 generation shows frequencies of 85 A-B-, 28A-bb,35aaB-,and 12 aabb.

H_o: The observed frequency not differ(=) the expected frequency.
H_A: The observed frequency differ(≠) the expected frequency

Expected A-B- = $\frac{9}{16} \times 160 = 90$

Expected A-bb = $\frac{3}{16} \times 160=30$

Expected aaB- = $\frac{3}{16} \times 160 = 30$

Expected aabb= $\frac{1}{16} \times 160=10$

	A-B-	A-bb	aaB-	aabb	
Observed	85	28	35	12	160
Expected	90	30	30	10	160

$$X^2 = \sum \frac{(O-E)^2}{E}$$

$$X^2 = \frac{(85-90)^2}{90} + \frac{(28-30)^2}{30} + \frac{(35-30)^2}{30} + \frac{(12-10)^2}{10}$$
$$= \frac{25}{90} + \frac{4}{30} + \frac{25}{30} + \frac{4}{10} =1.63$$

K=4
DF=k-1
DF=4-1=3
The obtained value=1.63
The tabulated value in 0.05 level=7.815
We therefore accepted the null hypothesis(H_o) of no difference between observed frequency and expected frequency.
rejected alternative hypothesis(H_A) of difference between observed frequency and expected frequency.

We have failed to provide statistical evidence that the assumed genetic model is not operating as expected.

<u>Example 4</u>:-In fowls, the creeper gene(producing deformed legs) is dominant over the gene for normal leg development. A series of crosses between heterozygous creepers(Cc) produce a phynotype ratio of 164 creepers to 76 normal birds.

Cc × Cc

F_2 CC, Cc, Cc, cc

1-Simple inspection of the obtained phenotype ratio reveals an obvious deviation from the 3:1 ratio expected from a Cc× Cc cross.

H_o: The observed frequency not differ (=) the expected frequency.
H_A: The observed frequency differ(≠) the expected frequency

Expected Creeper = $\frac{3}{4} \times 240$ = 180

Expected Normal = $\frac{1}{4} \times 240$=60

	Creeper	Normal	
Observed	164	76	240
Expected	180	60	240

$\boldsymbol{X^2} = \sum \frac{(\boldsymbol{O-E})^2}{\boldsymbol{E}}$

$\boldsymbol{X^2} = \sum \frac{[(\boldsymbol{O-E}) \ -\mathbf{0.5}]^2}{\boldsymbol{E}}$

$= \frac{[(\mathbf{164-180})-\mathbf{0.5}]^{\,2}}{\mathbf{180}} + \frac{[(\mathbf{76-60})-\mathbf{0.5}]^{\,2}}{\mathbf{60}}$

$= \frac{(\mathbf{15.5})^{\,2}}{\mathbf{180}} + \frac{(\mathbf{15.5})^{\,2}}{\mathbf{60}} =$ 1.33 + 4.00=5.33

K=2
DF=k-1
DF=2-1=1
The obtained value=5.33
The tabulated value in 0.01 level=6.635
The tabulated value in 0.05 level=7.8**79**

We therefore accepted the null hypothesis(H_o) of no difference between observed and expected frequency.
rejected alternative hypothesis(H_A) of difference between observed and expected frequency.

<u>Example 5</u>:- Suppose that we wish to know if an association exist between the factors sex and hair color. We proceed to check the first 50 menand the first 50 women who come down the street, noting in each case whether the individual is blonde or brunette.

1- As the first step, we wil organize the data in the form of the following table:

Hair color

Sex	Blonde	Brunette	
Men	20	30	50
Women	24	26	50
	44	56	100

Expected Men of Blonde = $\frac{50}{100} \times 44$ = 22,
Expected Women of Blonde= $\frac{50}{100} \times 44$ = 22

Expected Men of Brunette = $\frac{50}{100} \times 56$ =28
Expected Women of Brunette = $\frac{50}{100} \times 56$ =28

$$X^2 = \sum \frac{(O-E)^2}{E}$$

$$X^2 = \sum \frac{[(O-E)\ -0.5]^2}{E}$$

$$= \frac{[(20-22)-0.5]^2}{22} + \frac{[(30-28)-0.5]^2}{28} + \frac{[(24-22)-0.5]^2}{22} + \frac{[(26-28)-0.5]^2}{28}$$

= 0.102+ 0.080 + 0.102 + 0.080= 0.364

K=2
DF=k-1
DF=2-1=1
The obtained value=0.364
The tabulated value in 0.01 level=6.635
The tabulated value in 0.05 level=7.879
We therefore accepted the null hypothesis(H_o) of no difference between observed and expected frequency.
rejected alternative hypothesis(H_A) of difference between observed and expected frequency.
We have therefore not shown that the observed frequencies and those computed on the basis of a "no relationship" hypothesis are significantly different, and we are led to the conclusion that hair color is not associated with sex. In other words, the two factor appear to be independent.

<u>Example 6</u>:- A therapeutic drug was tested against aplacebo in terms of three subjectively evaluated patient categories(1) much improved, (2) slightly improved, and (3) not improved. A total of 120 patients were assigned to the drug group and 90 other patients were given the placebo. All were judged to be in approximately the same initial condition. Physician evaluation was then made without knowing which treatment the patient received. The resulting data were organized in the following 2×3 table:

	Much improved	Slightly improved	Not improved	
Drug	60	32	28	120
Placebo	28	17	45	90
	88	49	73	210

Expected Drug of Much improved = $\frac{120}{210} \times 88$ = 50.28
Expected Placebo of improved = $\frac{90}{210} \times 88$ = 37.71

Expected Drug of Slightly improved = $\frac{120}{210} \times 49$ =28.00
Expected Placebo of Slightly improved = $\frac{90}{210} \times 49$ =21.00

Expected Drug of not improved = $\frac{120}{210} \times 73$ = 41.71

Expected Placebo of notimproved = $\frac{90}{210} \times 73$ = 31.28

	Much improved	Slightly improved	Not improved	
Drug	60/50.28	32/28	28/41.71	120
Placebo	28/37.71	17/21	45/31.28	90
	88	49	73	210

$$\boldsymbol{X^2} = \sum \frac{(O-E)^2}{E}$$

$$\boldsymbol{X^2} = \frac{(60-50.28)^2}{50.28} + \frac{(32-28)^2}{28} + \frac{(28-41.71)^2}{41.71} + \frac{(28-37.71)^2}{37.71} + \frac{(17-21)^2}{21} + \frac{(45-31.28)^2}{31.28}$$

$$= \frac{94.48}{50.28} + \frac{16}{28} + \frac{187.96}{41.71} + \frac{94.28}{37.71} + \frac{16}{21} + \frac{188.24}{31.28}$$

= 1.88+0.57+4.51+2.50+0.76+6.02

$\boldsymbol{X^2}$= 16.24

K=3

DF=k-1

DF=3-1=2

The obtained value=16.24

The tabulated value in 0.01 level=9.210

The tabulated value in 0.05 level=5.991

We therefore rejected the null hypothesis(H_o) of no difference between observed and expected frequency.

accepted alternative hypothesis(H_A) of difference between observed and expected frequency.

We therefore have statistical evidence that a significant difference in degree of improvement does indeed exist between the placebo group and the drug group.

Example 7:- The Chi-Square Test for Independence evaluates the relationship between two variables. It is a nonparametric test that is performed on categorical(nominal or ordinal) data.

500 elementary school boys and girls are asked which is their favorite color: blue, green, or pink?

	Blue	Green	Pink	
Boys	**100/64.8**	**150/97.2**	**20/108**	**270**
Girls	**20/55.2**	**30/82.8**	**180/92**	**230**
	120	**180**	**200**	**N=500**

H_o: For the population of elementary school students, gender and favorite color are not related

$H_{A:}$: For the population of elementary school students, gender and favorite color are related

Expected Boys of Blue = $\frac{270}{500} \times 120$ =64.8

Expected Girls of Blue = $\frac{230}{500} \times 120$ = 55.2

Expected Boys of Green = $\frac{270}{500} \times 180$ = 97.2

Expected Girls of Green = $\frac{230}{500} \times 180$ =82.8

Expected Boys of Pink = $\frac{270}{500} \times 200$ = 108

Expected Girls of Pink = $\frac{230}{500} \times 200$ = 92

$$X^2 = \sum \frac{(O-E)^2}{E}$$

$$X^2 = \frac{(100-64.8)^2}{64.8} + \frac{(150-97.2)^2}{97.2} + \frac{(20-108)^2}{108} + \frac{(20-55.2)^2}{55.2} + \frac{(30-82.8)^2}{82.8} + \frac{(180-92)^2}{92}$$

$$= \frac{1.239.04}{64.8} + \frac{2.787.84}{97.2} + \frac{7.744}{108} + \frac{1.239.04}{55.2} + \frac{2.787.84}{55.2} + \frac{7.744}{92}$$

= 19.12+28.68+71.703+22.44+50.50+84.17
X^2= 226.11

K=3
DF=k-1
DF=3-1=2
The obtained value=226.11
The tabulated value in 0.01 level=9.210
The tabulated value in 0.05 level=5.991
We therefore rejected the null hypothesis(H_o) of no difference between observed and expected frequency.
accepted alternative hypothesis(H_A) of difference between observed and expected frequency.
State Conclusion
In the population, there is a relationship between gender and favorite color.

One Sample Hypothesis:-

From the experimental data and from the null hypothesis, we assume that mean of this sampling distribution is $\mu\, or\, 15.80\, g/100ml$. Our sample mean, $\overline{X}$is 16.50g/100ml, and therefore fall on distribution some where to the right of μ. Now we need to determine just how for to the right or how far out it is This is require that we compute the standrd error of our hypothetical sampling distribution.
In computing the standard error, use the procedure developed . Since we are working with a fairly large sample(64), we can use S as a good estimate of the known6.It may be assumed that S^2 has been computed the unbiased of σ^2.Thus we have

$$S_{\overline{x}} = \frac{S}{\sqrt{N}} = \frac{2.00}{\sqrt{64}} = \frac{2}{8} = 0.25$$

Our next step is again a familiarane. In order to place our sample statistic, on the sampling distribution, we calculate a Z-value in the usual manner. Thus

$$Z = \frac{\overline{x}-\mu}{S_{\overline{x}}} = \frac{16.50-15.80}{0.25} = 2.80$$

Having obtained 2.80 at the Z, or standard score, we now turn to table.

$$Z= \frac{\bar{X}-M}{\sigma_{\bar{x}}}$$

$$t= \frac{\bar{X}-M}{S_{\bar{x}}}$$

<u>One sample hypothesis:-</u>

Example1:- Mean of age of horse = 22 years, new breed of horse may be change in the mean of age of horse and become 24.23 years for 25 horses and standard deviation = 4.25 years.

H_o: μ = 22years
H_A: $\mu \neq$ 22years

μ = 22years
$\bar{x}$= 24.23 years
n= 25 horse
S=4.24

$$t= \frac{\bar{x}-\mu}{S_{\bar{x}}}$$

$$S_{\bar{x}} = \frac{S}{\sqrt{n}} = \frac{4.25}{\sqrt{25}} = 0.85$$

$$t= \frac{\bar{X}-M}{S_{\bar{x}}} = \frac{24.23-22}{0.85} = 2.62 \text{ (obtained value)}$$

(v)df =n-1=25-1=24

$t_{0.05(2),24=}$2.064 tabulated value
we therefore rejected H_o and accepted the H_A
If /t/≥t(2),v

<u>Two Sample Hypothesis:-</u>

This hypothesis is represent the ways for compared between two samples for investigation if the fined of variance between two groups. This distribution can use F distribution.

Example1:-The following data of the weight of two types of fish.
Type1 number of fish=11

Type 2 number of fish=8

Type1	Type2
41	52
34	57
33	62
36	55
40	64
25	57
31	56
37	55
34	n=8
30	v=7
38	
n=11	
v=10	
SS_1=218.73	
S^2= 21.87	

H_o: $\sigma^2{}_1 = \sigma^2{}_2$

H_A: $\sigma^2{}_1 \neq \sigma^2{}_2$

$F = \frac{S_1^2}{S_2^2} = \frac{21.87}{15.36}$ =1.42 obtained value

$F_{0.05\,(2),10,7}$ =3.64 tabulated value

∴ obtained value $(calculated) <$ tabulated value

∴ rejected H_o and accepted H_A

Example2:- The following data of clotting time in people measurment by minute received different two types of drugs.

Group received drug B	Group received drug G
8.8	**9.9**
8.4	**9.0**

7.9		11.1	
8.7		9.6	$S_p^2 = \frac{SS1+SS2}{V1+V2} = 0.519$
9.1		6.7	
9.6	$n_1=6$	10.4	$n_2=7$
	$v_1=5$	9.5	$v_2=6$
	$x_1=8.75$		$\overline{x_2}=9.74$
	$SS_1=1.69$		$SS_2=4.019$

$H_o: M_1=M_2$

$H_A: M_1 \neq M_2$

$$t = \frac{\bar{X}1 - \bar{X}2}{S_{\bar{x}-\bar{x}}}$$

$$S_p^2 = \frac{SS1+SS2}{V1+V2} = \quad S_p^2 = \text{pooled variance}$$

$$S^2_{\bar{x}-\bar{x}} = \frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}$$

$$S_{\bar{x}-\bar{x}} = \sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}$$

$$= \sqrt{\frac{0.5193}{6} + \frac{0.5193}{7}}$$

$$=0.40$$

$$t = \frac{8.75 - 9.74}{0.40}$$

$= 2.475$

$t_{0.05(2),11} = 2.201$

note$=v_1+v_2$ =DF

$\therefore$ **reject H_o and accepted H_A**

Example3:-Data of number two types of bird groupwild=9 and captive=7

$F=\frac{S_2}{S_1}=\frac{2.5}{0.48}=5.21$

$H_o: \sigma^2{}_1 \geq \sigma^2{}_2$

$H_A: \sigma^2{}_1 < \sigma^2{}_2$

Captive	**Wild**
10	9
11	8
12	11
11	12
10	10
11	13
11	11
	10
	12

$SS_1 = 2.86$ $n_1=7$ $n_2=9$

$S_1^2 = 0.48$ $v_1=6$ $v_2=8$

$SS_2 = 20$, $S_1^2=2.5$

Drug group	*Control group*
$N_D = 37$ $\bar{X}_D=290$ $S^2{}_D= 196$	$N_c = 40$ $\bar{X}_c=287$ $S^2{}_c= 144$

$F_{0.05(1)8,6}=4.15$ Tabulated value

F (obtained value)>F (Tabulated value)

∴ reject H_o and accepted H_A

لرر Acommon problem in biological research involves differences between sample mean.

Example3:-Suppose that we wish to test a newly developed drug inorder to determine whether it is effective in changing the heart rate of the norway rat.

-Following the random procedures, we would divide our sample into two groups. One group would be given the drug and the other group would sever as a control. Inorder to make certain that as many variable as possible are kept constant, we will administer aplacebo(harmless inter substance)to the control group. In this way, we will reduce any difference due to the injection procedure itself. Following an appropriate time interval, the heart rate of each rat is measured, and we establish the following data.

1-We state the null hypothesis as H_o: $\mu_D - \mu_c = 0$. Note that we do not say $\bar{X}_D - \bar{X}_c$=0, Since we intersted in generalizing our conculsion to all Norway rats.
2-We set a level of significance at 0.05. This now becomes our present maxium level of rejection.
3- We assume the sampling distribution shown. In this case we are hypothesizing a sampling distribution based on the statistic, $\bar{X}_D - \bar{X}_c$

$$S_{\bar{x}D-\bar{x}\,c} = \sqrt{\frac{S_D^{\,2}}{N_D} + \frac{S_C^{\,2}}{N_C}} = \sqrt{\frac{196}{37} + \frac{144}{40}} = \sqrt{8.90} = 2.98$$

- We make the decision concerning rejection of the null hypothesis. Since we are interested only in the question as to whether our drug will change heart rate, and we are not at this time concerned with the direction of change we will use a two tailed test. It may be seen that a Z-value of $\mp$ 1.96 or more is necessary to reject at the 0.05 level. We therefore fail to reject the hypothesis $\mu_D - \mu_c = 0$, since our obtained Z-value is 1.01

Example4:-The mean level of prothrombin in normal population is known to be approximately 20mg/100ml of plasma. A sample of 40 patients showing a vitamin k deficiency has a mean prothrombin level of 18.50mg/100ml. The sample standard deviation is 4 mg/100ml. Is the sample mean significantly lower than the population mean?

1. $H_o: \mu - \bar{x} = 0$, $H_A: \mu - \bar{x} \neq 0$
2. $S_{\bar{x}} = \frac{S}{\sqrt{n}} = \frac{4}{\sqrt{40}} = 0.632$
3. $Z = \frac{\mu - \bar{x}}{S_{\bar{x}}} = \frac{20-18.50}{0.632} = 2.37$
4. Reject the null hypothesis be yound the 0.05 level.

Example5:-Asample of male was drawn from each two geographically isolated populations of Rana pipiens, and their body lengths were measured to the nearest millimeter. From the data below, determine whether there is statistically significant difference beetween the males of the two populations terms of body length.

$\bar{X}_1$ =74 | **$\bar{X}_2$=78**
$S^2 1$=**225** | $S^2 2$ =**169**
N_1 = | **N_2 =56**

Drug group	*Placebo group*

1. H_o: $\boldsymbol{\mu_1 - \mu_2} = 0$
H_o: $\boldsymbol{\mu_1 - \mu_2} \neq 0$

2. $\boldsymbol{S_{\bar{x}1-\overline{x\,2}}} = \sqrt{\frac{S_1{}^2}{N_1} + \frac{S_2{}^2}{N_2}} = \sqrt{\frac{225}{42} + \frac{169}{56}} = \sqrt{\mathbf{8.38}}$=**2.89**
3. $Z = \frac{\bar{X} - \bar{X}}{S_{\bar{x}-\bar{x}}} = \frac{78-74}{2.89} = \frac{4}{2.89} = 1.38$
4.Fail to reject $H_o$null hypothesis at 0.05 level ,and accept H_A , using a two sampling hypothesis.

Example6:-A drug which was delieved to hasten blood clotting time was tested **by comparing a drug group with a plcebo group.**
Analyze the following data in order to determine whether the mean clotting time of the drug group is significantly lower than the meanclotting of the placebo group. The clotting time is given in minutes.

Muna H. AL-Saeed

$\bar{X}_D$=6.30 $S^2{}_{\mathrm{D}}$= 10.24 $N_D = 64$	$\bar{X}_p$=7.45 $S^2{}_{\mathrm{p}}$= 12.96 $N_p = 64$

1. H_o: $\boldsymbol{\mu_p - \mu_D} = 0$

H_o: $\boldsymbol{\mu_p - \mu_D} \neq 0$

2. $\mathbf{S}_{\bar{x}-\bar{x}} = \sqrt{\frac{S^2}{N_p} + \frac{\mathbf{S}^2}{N_D}} = \sqrt{\frac{12.96}{\mathbf{64}} + \frac{10.24}{\mathbf{64}}} = \sqrt{0.}362 = 0.602$

3. $Z = \frac{\bar{X}-\bar{X}}{\mathbf{S}_{\bar{x}-\bar{x}}} = \frac{7.45}{0.602} = 1.91$

4. Rject $H_o$null hypothesis at 0.05 level ,and accept H_A .

Muna H. AL-Saeed

In 200 tosses of a coin, 115 heads and 85 tails were observed. Test the hypothesis that the coin is fair, using Appendix IV and significance levels of (*a*) 0.05 and (*b*) 0.01. Test the hypothesis by computing the *p*-value and (*c*) comparing it to levels 0.05 and 0.01.

SOLUTION

The observed frequencies of heads and tails are $o_1 = 115$ and $o_2 = 85$, respectively, and the expected frequencies of heads and tails (if the coin is fair) are $e_1 = 100$ and $e_2 = 100$, respectively. Thus

$$\chi^2 = \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2} = \frac{(115 - 100)^2}{100} + \frac{(85 - 100)^2}{100} = 4.50$$

Since the number of categories, or classes (heads, tails), is $k = 2$, $\nu = k - 1 = 2 - 1 = 1$.

(*a*) The critical value $\chi^2_{.95}$ for 1 degree of freedom is 3.84. Thus, since $4.50 > 3.84$, we reject the hypothesis that the coin is fair at the 0.05 significance level.

(*b*) The critical value $\chi^2_{.99}$ for 1 degree of freedom is 6.63. Thus, since $4.50 < 6.63$, we cannot reject the hypothesis that the coin is fair at the 0.02 significance level.

$$\chi^2\,(\text{corrected}) = \frac{(|o_1 - e_1| - 0.5)^2}{e_1} + \frac{(|o_2 - e_2| - 0.5)^2}{e_2} = \frac{(|115 - 100| - 0.5)^2}{100} + \frac{(|85 - 100| - 0.5)^2}{100}$$
$$= \frac{(14.5)^2}{100} + \frac{(14.5)^2}{100} = 4.205$$

Since $4.205 > 3.84$ and $4.205 < 6.63$, the conclusions reached in Problem 12.1 are valid. For a comparison with previous methods, see Problem 12.3.

12.4 Table 12.8 shows the observed and expected frequencies in tossing a die 120 times.

(*a*) Test the hypothesis that the die is fair using a 0.05 significance level by calculating χ^2 and giving the 0.05 critical value and comparing the computed test statistic with the critical value.

(*b*) Compute the *p*-value and compare it with 0.05 to test the hypothesis.

Table 12.8

Die face	1	2	3	4	5	6
Observed frequency	25	17	15	23	24	16
Expected frequency	20	20	20	20	20	20

SOLUTION

$$\chi^2 = \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2} + \frac{(o_3 - e_3)^2}{e_3} + \frac{(o_4 - e_4)^2}{e_4} + \frac{(o_5 - e_5)^2}{e_5} + \frac{(o_6 - e_6)^2}{e_6}$$
$$= \frac{(25 - 20)^2}{20} + \frac{(17 - 20)^2}{20} + \frac{(15 - 20)^2}{20} + \frac{(23 - 20)^2}{20} + \frac{(24 - 20)^2}{20} + \frac{(16 - 20)^2}{20} = 5.00$$

(*a*) The 0.05 critical value is given by the EXCEL expression =CHIINV(0.05, 5) or 11.0705. The computed value of the test statistic is 5.00. Since the computed test statistic is not in the 0.05 critical region, do not reject the null that the die is fair.

(*b*) The *p*-value is given by the EXCEL expression =CHIDIST(5.00, 5) or 0.4159. Since the *p*-value is not less than 0.05, do not reject the null that the die is fair.

In his experiments with peas, Gregor Mendel observed that 315 were round and yellow, 108 were round and green, 101 were wrinkled and yellow, and 32 were wrinkled and green. According to his theory of heredity, the numbers should be in the proportion 9 : 3 : 3 : 1. Is there any evidence to doubt his theory at the (*a*) 0.01 and (*b*) 0.05 significance levels?

SOLUTION

The total number of peas is $315 + 108 + 101 + 32 = 556$. Since the expected numbers are in the proportion 9 : 3 : 3 : 1 (and $9 + 3 + 3 + 1 = 16$), we would expect

$$\tfrac{9}{16}(556) = 312.75 \text{ round and yellow} \qquad \tfrac{3}{16}(556) = 104.25 \text{ wrinkled and yellow}$$

$$\tfrac{3}{16}(556) = 104.25 \text{ round and green} \qquad \tfrac{1}{16}(556) = 34.75 \text{ wrinkled and green}$$

Thus
$$\chi^2 = \frac{(315 - 312.75)^2}{312.75} + \frac{(108 - 104.25)^2}{104.25} + \frac{(101 - 104.25)^2}{104.25} + \frac{(32 - 34.75)^2}{34.75} = 0.470$$

Since there are four categories, $k = 4$ and the number of degrees of freedom is $\nu = 4 - 1 = 3$.

(*a*) For $\nu = 3$, $\chi^2_{.99} = 11.3$, and thus we cannot reject the theory at the 0.01 level.

(*b*) For $\nu = 3$, $\chi^2_{.95} = 7.81$, and thus we cannot reject the theory at the 0.05 level.

We conclude that the theory and experiment are in agreement.

Note that for 3 degrees of freedom, $\chi^2_{.05} = 0.352$ and $\chi^2 = 0.470 > 0.352$. Thus, although the agreement is good, the results obtained are subject to a reasonable amount of sampling error.

Another method

Using Yates' correction, we find

$$\chi^2 = \frac{(|3 - 6| - 0.5)^2}{6} + \frac{(|9 - 6| - 0.5)^2}{6} = \frac{(2.5)^2}{6} + \frac{(2.5)^2}{6} = 2.1$$

which leads to the same conclusion given above. This is to be expected, of course, since Yates' correction always *reduces* the value of χ^2.

It should be noted that if the χ^2 approximation is used despite the fact that the frequencies are too small, we would obtain

$$\chi^2 = \frac{(2 - 3)^2}{3} + \frac{(5 - 3)^2}{3} + \frac{(4 - 3)^2}{3} + \frac{(1 - 3)^2}{3} = 3.33$$

Since for $\nu = 4 - 1 = 3$, $\chi^2_{.95} = 7.81$, we would arrive at the same conclusions as above. Unfortunately, the χ^2 approximation for small frequencies is poor; hence, when it is not advisable to combine frequencies, we must resort to the exact probability methods of Chapter 6.

In 360 tosses of a pair of dice, 74 sevens and 24 elevens are observed. Using the 0.05 significance level, test the hypothesis that the dice are fair.

SOLUTION

A pair of dice can fall 36 ways. A seven can occur in 6 ways, an eleven in 2 ways. Then $\Pr\{\text{seven}\} = \frac{6}{36} = \frac{1}{6}$ and $\Pr\{\text{eleven}\} = \frac{2}{36} = \frac{1}{18}$. Thus in 360 tosses we would expect $\frac{1}{6}(360) = 60$ sevens and $\frac{1}{18}(360) = 20$ elevens, so that

$$\chi^2 = \frac{(74-60)^2}{60} + \frac{(24-20)^2}{20} = 4.07$$

For $\nu = 2 - 1 = 1$, $\chi^2_{.95} = 3.84$. Thus, since $4.07 > 3.84$, we would be inclined to reject the hypothesis that the dice are fair. Using Yates' correction, however, we find

$$\chi^2\,(\text{corrected}) = \frac{(|74-60| - 0.5)^2}{60} + \frac{(|24-20| - 0.5)^2}{20} = \frac{(13.5)^2}{60} + \frac{(3.5)^2}{20} = 3.65$$

Thus on the basis of the corrected χ^2 we could not reject the hypothesis at the 0.05 level.

In general, for large samples such as we have here, results using Yates' correction prove to be more reliable than uncorrected results. However, since even the corrected value of χ^2 lies so close to the critical value, we are hesitant about making decisions one way or the other. In such cases it is perhaps best to increase the sample size by taking more observations if we are interested especially in the 0.05 level for some reason; otherwise, we could reject the hypothesis at some other level (such as 0.10) if this is satisfactory.

Example:-The F_2 generation resulting from crosses between heterozygous red owls contained 16 red and 8 grey owls. Are these results consistent with the genetic theory that is dominant over grey 3:1

H_o: The observed frequency not differ (=) the expected frequency.
H_A: The observed frequency differ($\neq$) the expected frequency

Expected Red owls= $\frac{3}{4} \times 24 = 18$

Expected Grey owls = $\frac{1}{4} \times 24 = 6$

	Red owls	Grey owls	
Observed	16	8	24
Expected	18	6	24

$X^2 = \sum \frac{(O-E)^2}{E}$

$X^2 = \sum \frac{[(O-E)\ -0.5]^2}{E}$

$$= \frac{[(16-18)-0.5]^2}{18} + \frac{[(8-6)-0.5]^2}{6}$$

$$= \frac{(1.5)^2}{18} + \frac{(1.5)^2}{6} = \frac{2.25}{18} + \frac{2.25}{6} = \frac{9.0}{18} = 0.5$$

K=2
DF=k-1
DF=2-1=1
The obtained value=0.5
The tabulated value in 0.01 level=6.635
The tabulated value in 0.05 level=7.8**79**
We therefore accepted the null hypothesis(H_o) of no difference between observed and expected frequency.
rejected alternative hypothesis(H_A) of difference between observed and expected frequency.
X^2 is not significant; therefore is no reason to assume that observed frequency donot fit the 3:1 model

The Sampling Distribution:-

Our sample statistic in this case is $\bar{X}_E - \bar{X}_W$, so we construct a hypothetical sampling distribution of "differences between means" as shown by as usual, this is based on an infinitude of samples drawn from the populations of interest. Since our null hypothesis implies a prameter of zero difference, the sampling distribution is assumed to be Zero.

-**Level of significant:-**

We shall present the level of significance at 0.05. Since we are interested only in whether a difference exists, and not the direction of difference, a two sample test is appropriate.

1- The first step involves the calculation of a pooled variance. This procedure takes into consideration all information that is obtaied from both samples. The pooled variance, S_p^2, is computed by application of the familiar formula for variance. Thus

$$S_p^2 = \frac{\sum\sum(\boldsymbol{X_E}-\bar{\boldsymbol{X}}_E)^2+\sum(\boldsymbol{X_W}-\bar{\boldsymbol{X}}_W)^2}{\boldsymbol{N_{E+}N_W-2}}$$

$$= \frac{\mathbf{2.39+2.74}}{\mathbf{14+18-2}} = \frac{\mathbf{5.13}}{\mathbf{30}} = 0.17$$

Note that in the above formula we are using the sums of squared deviations from two means, one from each each sample. Since this involved the calculation of two separate means, or two parameter estimate we lose one degree of freedom for each parameter estimated. This accounts for the expression, $N_{E+}N_W - 2$.

2- Next, we use the pooled variance, S_p^2 to compute the standard error of differences between means. Again, we utilize an algebraic extension of the basic formula for the standard error of mean. Thus

$$\boldsymbol{S}_{\bar{\boldsymbol{X}}_E-\bar{\boldsymbol{X}}_W} = \sqrt{\frac{S_p^2}{\boldsymbol{N_E}}+\frac{S_p^2}{\boldsymbol{N_W}}} = \sqrt{\frac{0.17}{\mathbf{14}}+\frac{0.17}{\mathbf{18}}} = \sqrt{0.021} = 0.15$$

3- Now,having obtained the standard error, we proceed as usual to place our statistic, $\bar{\boldsymbol{X}}_E - \bar{\boldsymbol{X}}_W$, on the sampling distribution. This done by computing the t-value as

$$t = \frac{\bar{X}_E-\bar{X}_W}{S_{\bar{X}_E-\bar{X}_W}} = \frac{8.57-8.40}{0.15} = 1.13$$

The Decision

Recalling that the number of freedom associated with this design is N_1+N_2-2, we enter at $14+18-2$, or 30 degree of freedom.

We look for the criticalbvalue in the 0.05 colum for a two sample test and we find the value of 2.04.

Example 1:- Agroup of mice were placed in a series of stress situations which elicited a fear response. After a period of time under these conditions, the mice were compared to those of a control group which had not been put under stress. Analyze the following data to determine whether a significant difference in a drenal gland weight exists between the two groups. The adrenal weight is

Experimenal	Control	Experimenal	Control
3.8	4.2	3.9	3.6
6.8	4.8	5.9	2.4
8.0	4.8	6.0	3.2
3.6	2.3	5.7	4.9
3.9	6.5	5.6	
4.5	4.9	4.5	

(1) $\sum \boldsymbol{X_E^2} = 20.80$
$\sum \boldsymbol{X_c^2} = 15.18$

(2) $\boldsymbol{S_p^2} = \frac{\boldsymbol{X_E^2+X_c^2}}{\boldsymbol{N_{1+}N_2-2}} = \frac{\mathbf{20.80+15.18}}{\mathbf{20}} = 1.80$

(3) $S_{\bar{X}_{E-}\ \bar{X}_C} = \sqrt{\frac{S_p^2}{N_E}+\frac{S_p^2}{N_C}} = \sqrt{\frac{\mathbf{1.80}}{12}+\frac{\mathbf{1.80}}{10}} = \sqrt{0.33} = 0.575$

(4) $t = \frac{\bar{X}_{E-}\ \bar{X}_C}{S_{\bar{X}_{E-}\ \bar{X}_C}} = \frac{5.2-4.2}{0.575} = -1.74$

(5) With 20 degree of freedom, t is not significant at 0.05 level.

Example2 :- Two different methods were used to determine the concetration Of prothrombin in plasma. Both determinations were made on the same subject, using 8 subjects in all. On the basis of the following data, where prothrombin is expressed in milligrams/100ml, determine whether a significant difference exists between the two methods.

Subject	**Methods I**	**Methods II**
1	17	18
2	17	17
3	18	20
4	21	24
5	22	23
6	17	15
7	23	25
8	23	22

	D	**D-$\bar{D}$**	**(D-D)^2**
1	+1	0.25	0.063
2	0	−0.75	0.563
3	+2	1.25	1.563
4	+3	2.25	5.063
5	+1	0.25	0.063
6	−2	−2.75	7.563
7	+2	1.25	1.563
8	$\underline{-1}$	−1.75	$\underline{3.063}$

$\bar{D}$=0.75 $\qquad$ 19.504=$\sum(D-\bar{D})^2$

(2) $S_D=\sqrt{\frac{19.504}{7}}=1.67$

(3) $S_{\bar{D}}=\frac{1.67}{\sqrt{8}}=0.592$

(4) t=$\frac{0.75}{0.592}=1.27$

(5) with 7 degrees of freedom, t is not significant at the level 0.05.

Example3:- A sample of 14 measurements yields mean of 50 and a standard deviation of 5. Estimate the population mean with 99% confidence.

(1) $S_x = \frac{S}{\sqrt{N}} = \frac{5}{\sqrt{14}} = \frac{5}{3.74} = 1.34$

(2) $\mu = \bar{X} \mp S_x\,(t_{0.01}) = 50 \mp 1.34(3.01) = 50 \mp 4.03 = 45.97 - 54.03$

Example:- An investigator testes a drug which he has reason to believe will increase hemoglobin content in gram/100ml. The hemoglobin content of 8 subjects is measured before and after administration of the drug. Analyze the following data in terms of the effectivness of the drug.

Subject	Before	After
1	10	12
2	9	11
3	11	13
4	12	14
5	8	9
6	7	10
7	12	12
8	10	14

	D	$(\bar{D} - D)$	$(\bar{D} - D)^2$
1	+2	0	0
2	+2	0	0
3	+2	0	0
4	+2	0	0
5	+1	-1	1
6	+3	1	1
7	0	-2	4
8	+4	2	4

D=2 $\qquad$ $10 = \sum(\bar{D} - D)^2$

(2) $S_D = \sqrt{\frac{\sum(D-D)^2}{N-1}} = \sqrt{\frac{10}{7}} = 1.20$

(3) $S_{\bar{D}} = \frac{S_D}{\sqrt{N}} = \frac{1.20}{\sqrt{8}} = 0.425$

(4) t= $\frac{\bar{D}-0}{S_{\bar{D}}} = \frac{2}{0.425} = 4.71$

(5) Assuming a one-sample hypothesis, tis significant beyond the 0.01level.

Example 4:- Two varieties of peas were compared in terms of ascorbic acid content measured in milligrams/100g. From the following data, which were derived from 10 samples drawn each variety, determine whether asignificant difference in ascorbic acid content exists between the two varieties.

VarietyI	VarietyII
39	42
40	39
34	41
32	36
29	28
28	34
26	27
21	25
19	31
22	22

(1) $\sum X_1^2 = 478$

$\sum X_2^2 = 438$

(2) $S_p^2 = \frac{478+438}{18} = 50.89$

(3) $S_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{50.89}{10} + \frac{50.89}{10}} = 3.19$

(4) t= $\frac{\bar{X}_1-\bar{X}_2}{S_{\bar{X}_1-\bar{X}_2}} = \frac{29.00-32.50}{3.19} = -1.10$

<u>Multi Sample Hypothesis</u>:-
ANOVA Test:-
Analysis of Variance:-
Asingle factor analysis of variance of for random model experimental design.

<u>Example1:-</u> A laboratory employs a certain technique for determining the phosphorus content of hay.
Do phosphorus determination differ with technician performing analysis?
Each of four randomly selected technicians was given five samples from the same bath of hay. The result of 20 phosphorus determinations (in mg phosphorus/g of hay). Are shown.

H_o:-Determinations of phosphorus content do not differ among technicians.
H_A:- Determinations of phosphorus content do differ among technicians.

α=0.05

Technician

	1	2	3	4
	34	**37**	**34**	**36**
	36	**36**	**37**	**34**
	34	**35**	**35**	**37**
	35	**37**	**37**	**34**
	34	**37**	**36**	**35**
Group sums:	$\sum X_1$ **=173**	$\sum X_2$**=182**	$\sum X_3$ **=179**	$\sum X_4$**=176**

$\sum\sum Xij=\sum X_1+\sum X_2+\sum X_3+\sum X_4$ = 173+ 182 +179 +176 =710
$\sum\sum Xij^2$= $(34)^2 + (36)^2 + (34)^2 +(35)^2 +(34)^2 +(37)^2 + (36)^2 + (35)^2 +(37)^2 +(37)^2 + (34)^2 + (37)^2 + (35)^2 +(37)^2 +(36)^2 +(36)^2 + (34)^2 + (37)^2 +(34)^2 +(35)^2$ =25234
N=20

$C=\frac{(\sum\sum Xij)^2}{N} = \frac{(710)^2}{20}$ = 25205.00

Total SS $=\sum\sum Xij^2 - C$
= 25234 – 25205
= 29.00

Group SS = $\frac{\sum X_i^2}{K_i} - C = \frac{(173)^2}{5} + \frac{(182)^2}{5} + \frac{(179)^2}{5} + \frac{(176)^2}{5}$ – 25205 =

= 25214.00– 25205
=9
Error SS = Total SS – Group SS
= 29 – 9
= 20.00

Source of variation	SS	DF	MS
Total	29	19	
Group (Technicians)	9	3	3.00
Error	20	16	1.25

Total DF = N-1=20-1= 19
Group DF = K-1= 4-1=3
Error DF = Total DF – Group DF =19 – 3= 16

Group MS = $\frac{SS}{DF} = \frac{9}{3}$ **=** 3.00

Error MS =$\frac{SS}{DF} = \frac{20}{16}$ **=**1.25

F =$\frac{MS\ group}{MS\ error} = \frac{3}{1.25}$ **=** 2.40

$F_{0.05\ (1),3,16}$ **=**3.24

Do not reject H_o $0.10 < P < 0.25$

<u>Example2:-</u> The Table contains three groups of measurements-A,B, and C- together with the data derived from these measurements. Note that (k) is used to symbolize the number of measurements in a specific group, (N) represents the total number of all three groups combined, and $\bar{X}$ denotes the grand mean, or mean all X.

A	B	C
2	9	10
3	6	6
1	8	9
5	7	7
4	5	8
K_A=5	K_B=5	K_C=5

Group sums: $\sum X_A = 15$ $\sum X_B = 35$ $\sum X_C = 40$

$\bar{X}_A = 3$ $\bar{X}_B = 7$ $\bar{X}_C = 8$

$\bar{\boldsymbol{X}}=6$

N=15

$\sum\sum Xij = \sum \boldsymbol{X_A} + \sum \boldsymbol{X_B} + \sum \boldsymbol{X_C}$

$= 15 + 35 + 40$

$\sum\sum Xij = 90$

$\sum\sum Xij^2 = (2)^2 + (3)^2 + (1)^2 + (5)^2 + (4)^2 + (9)^2 + (6)^2 + (8)^2 + (7)^2 + (5)^2 + (10)^2 + (6)^2 + (9)^2 + (7)^2 + (8)^2$

$= 4 + 9 + 1 + 25 + 16 + 81 + 36 + 64 + 49 + 25 + 100 + 36 + 81 + 49 + 64$

$=640$

$$C = \frac{(\sum\sum Xij)^2}{N} = \frac{(90)^2}{15} = \frac{8100}{15} = 540$$

Total SS $= \sum\sum Xij^2 - C$

$= 640 - 540$

$= 100$

Group SS$= \frac{(\sum Xij)^2}{K_i} - C =$

$$= \frac{(\sum Xij)^2}{K_i} - \frac{(\sum\sum Xij)^2}{N} = \frac{(\sum \boldsymbol{X_A})^2}{K_A} + \frac{(\sum \boldsymbol{X_B})^2}{K_B} + \frac{(\sum \boldsymbol{X_C})^2}{K_C} - \frac{(\sum\sum Xij)^2}{N}$$

$= \frac{(15)^2}{5} + \frac{(35)^2}{5} + \frac{(40)^2}{5} - 540$

$= \frac{225}{5} + \frac{1225}{5} + \frac{1600}{5} - 540$

$= 45 + 245 + 320 - 540$

$= 610 - 540$

$= 70$

Error SS= Total SS – Group SS

= 100 – 70

= 30

Source of variation	**SS**	**DF**	**MS**
Total	100	14	
Group	70	2	35
Error	30	12	2.5

Total DF = N–1=15–1= 14

Group DF = K–1= 3–1=2

Error DF = Total DF – Group DF =14 – 2= 12

Group MS = $\frac{SS}{DF} = \frac{70}{2} = 35$

Error MS = $\frac{SS}{DF} = \frac{30}{12} = 1.25$

F = $\frac{MS\ group}{MS\ error} = \frac{35}{2.5} = 14$

$F_{0.05\,(1),2,12} = 3.88$

∴ **reject** H_o $0.10 < P < 0.25$ **and accepted** H_A

<u>Example 3:-</u> Eight samples were taken from each of 3 locations in ariver in order to determine whether a significant difference in total nitrogen content existed among the locations. Analyze the data below for significance. The total nitrogen content is expressed in milligrams/100g.

Location

A	B	C
222	326	263
300	275	260
262	218	299
264	207	221
200	272	198
211	268	211
267	308	266
326	229	319
K_A=8	K_B=8	K_C=8
$\bar{X}_A$ =256.50	$\bar{X}_B$ =262.88	$\bar{X}_C$ =254.63
$\sum X_A$ =2052	$\sum X_B$ =2103	$\sum X_C$ =2037

Group sums: $\sum X_A$ =2052 $\sum X_B$ =2103 $\sum X_C$ =2037

$\bar{\boldsymbol{X}}$= 258

N=24

$\sum\sum Xij = \sum \boldsymbol{X_A} + \sum \boldsymbol{X_B} + \sum \boldsymbol{X_C}$

= 2052 + 2103 + 2037

$\sum\sum Xij$= 6192

$\sum\sum Xij^2$ = $(222)^2 + (300)^2 + (262)^2 + (264)^2 + (200)^2 + (211)^2 + (267)^2 + (326)^2$
$+ (326)^2 + (275)^2 + (218)^2 + (207)^2 + (272)^2 + (268)^2 + (308)^2 + (229)^2$
$+ (263)^2 + (260)^2 + (299)^2 + (221)^2 + (198)^2 + (211)^2 + (266)^2 + (319)^2$
= 1.636.350

$$C = \frac{(\sum\sum Xij)^2}{\boldsymbol{N}} = \frac{\boldsymbol{(6192)^2}}{\boldsymbol{24}} = \frac{\boldsymbol{38.340.864}}{\boldsymbol{24}} = 1.597.536$$

Total SS = $\sum\sum Xij^2 - \text{C}$
= 1.636.350− 1.597.536
= 38.814

Group SS= $\frac{(\sum Xij)^2}{\boldsymbol{K_i}} - \text{C} =$

$$= \frac{(\sum Xij)^2}{K_i} - \frac{(\sum\sum Xij)^2}{N} = \frac{(\sum X_A)^2}{K_A} + \frac{(\sum X_B)^2}{K_B} + \frac{(\sum X_C)^2}{K_C} - \frac{(\sum\sum Xij)^2}{N}$$

$$= \frac{(2052)^2}{8} + \frac{(2103)^2}{8} + \frac{(2037)^2}{8} - 1597536$$

= 526338 + 552826 + 518671 – 1597536

= 1597835 – 1597536

= 299

Error SS= Total SS – Group SS

= 39214 – 299

= 38.915

Source of variation	**SS**	**DF**	**MS**
Total	39214	23	
Group	299	2	149.50
Error	38.915	21	17682.86

Total DF = N–1=24–1= 23

Group DF = K–1= 3–1=2

Error DF = Total DF – Group DF =23 – 2= 21

Group MS = $\frac{SS}{DF} = \frac{299}{2}$ = 149.50

Error MS =$\frac{SS}{DF}$ =$\frac{38.915}{21}$ =17682.86

F =$\frac{MS\ group}{MS\ error} = \frac{149.50}{17682.86}$ = 0.084

$F_{0.05\,(1),2,21}$ =3.88

Do not reject H_o $0.10 < P < 0.25$

<u>Example 4:-</u> Two different baits were tested for significant difference in terms of consumption by wildrats. The following datawere obtained from 5 location per bait, and are expressed in percentage consumption. Analyze these data for significance, applying the arc sine transformation for percentage data.

(1) Perform arc sine trans formation by use of Table

A	B
18.4	**22.8**
22.8	**26.6**
20.3	**23.6**
26.6	**30.0**
22.0	**26.6**

(2) $K_A=5$ $K_B=5$

$\bar{X}_A = 22.02$ $\bar{X}_B = 25.92$

Group sums: $\sum X_A = 110.1$ $\sum X_B = 129.6$

$\bar{\boldsymbol{X}}= 23.97$

N=10

$\sum\sum Xij = \sum \boldsymbol{X_A} + \sum \boldsymbol{X_B}$

$= 110.1 + 129.6$

$\sum\sum Xij = 239.7$

$\sum\sum Xij^2 = (18.4)^2 + (22.8)^2 + (20.3)^2 + (26.6)^2 + (22.0)^2 + (22.8)^2 + (26.6)^2 + (23.6)^2 + (30.0)^2 + (26.6)^2$

$= 5854.0$

$$C = \frac{(\sum\sum Xij)^2}{N} = \frac{(239.7)^2}{10} = \frac{57456}{10} = 5745.6$$

Total SS $= \sum\sum Xij^2 - C$

$= 5854 - 5745.6$

$= 108.4$

Group SS$= \frac{(\sum Xij)^2}{K_i} - C =$

$$= \frac{(\sum Xij)^2}{K_i} - \frac{(\sum\sum Xij)^2}{N} = \frac{(\sum X_A)^2}{K_A} + \frac{(\sum X_B)^2}{K_B} - \frac{(\sum\sum Xij)^2}{N}$$

$$= \frac{(\mathbf{110.1})^2}{\mathbf{5}} + \frac{(129.6\,)^2}{\mathbf{5}} - \frac{(\,239.7)^2}{\mathbf{10}}$$

=2424.4 + 3359.2−5745.6

=5783.6−5745.6

=38.0

Error SS= Total SS – Group SS

= 108.4 – 38.0

= 70.4

Source of variation	**SS**	**DF**	**MS**
Total	108.4	23	
Group	38.0	1	38
Error	70.4	8	8.8

Total DF = N–1=10–1= 9

Group DF = K–1= 2–1=1

Error DF = Total DF – Group DF =9 – 1= 8

Group MS = $\frac{\boldsymbol{SS}}{\boldsymbol{DF}} = \frac{\mathbf{38}}{\mathbf{1}}$ = 38

Error MS =$\frac{\boldsymbol{SS}}{\boldsymbol{DF}}$ =$\frac{\mathbf{70.4}}{\mathbf{8}}$ = 8.8

F =$\frac{\boldsymbol{MS\ group}}{\boldsymbol{MS\ error}} = \frac{38}{8.8}$ = 4.32

$\mathbf{F}_{0.05\,(\mathbf{1}),1,8}$ =3.88

∴ **reject** $\mathbf{H_o}$ $0.10 < P < 0.25$

<u>Example 5:-</u> Three clinical methods of determination of hemoglobin content were tested for significant difference in results. Six subjects were used, each subject constituting a single block. Analyze the data below, where the figures represent grams/100ml.

(1) Obtain block and treatment total

	Block						
Method	**A**	**B**	**C**	**D**	**E**	**F**	**Treatment**
1	**14**	**12**	**16**	**15**	**10**	**11**	**78**
2	**18**	**16**	**17**	**19**	**12**	**13**	**95**
3	**15**	**14**	**12**	**14**	**12**	**9**	**76**
Block	**47**	**42**	**45**	**48**	**34**	**33**	**249**

$\sum\sum Xij = 249$

$\sum\sum Xij^2 = 3571$

Total SS = $\sum\sum Xij^2 - \mathrm{C}$

$$= 3571 - \frac{(\sum\sum Xij)^2}{N}$$

$$= 3571 - \frac{(249)^2}{18}$$

$$= 3571 - 3444 = 127$$

Group SS=$\frac{(\sum Xij)^2}{K_i}$ – C =

$$= \frac{(\sum Xij)^2}{K_i} - \frac{(\sum\sum Xij)^2}{N} = \frac{(\sum X_A)^2}{K_A} + \frac{(\sum X_B)^2}{K_B} + \frac{(\sum X_C)^2}{K_C} + \frac{(\sum X_D)^2}{K_D} + \frac{(\sum X_E)^2}{K_E}$$

$$+\frac{(\sum X_F)^2}{K_F} - \frac{(\sum\sum Xij)^2}{N}$$

$$= \frac{(47)^2}{3} + \frac{(42)^2}{3} + \frac{(45)^2}{3} + \frac{(48)^2}{3} + \frac{(34)^2}{3} + \frac{(33)^2}{3} - \frac{(249)^2}{18}$$

$$= 736.3 + 588 + 675 + 768 + 385.3 + 363 - 3444.5$$

=3515.6−3444.5

=71.1

Method SS = $\frac{(78)^2+(95)^2+(76)^2}{6} - \frac{(249)^2}{18}$ = 36

= 3480.83 **-** 3444.5

Error SS= Total SS – Group SS

= 127 – (71.1+36)

= 127 – 107.1

= 19.9

Source of variation	**SS**	**DF**	**MS**
Total	127	17	
Group	71.1	5	
Method	36	2	18
Error	19.9	10	1.99

Total DF = N–1=18–1= 17

Group DF = K–1= 6–1=5

Method DF = $K_m - 1$=3-1=2

Error DF = Total DF – (Group DF + Method DF) =17 – (5+2)= 10

Method MS = $\frac{SS}{DF} = \frac{36}{2}$ = 18

Error MS = $\frac{SS}{DF} = \frac{19.9}{10}$ = 1.99

F = $\frac{MS\ method}{MS\ error} = \frac{18}{1.99}$ = 9.04

$F_{0.05\ (1),2,10}$ =3.88

∴ **reject** H_o $0.10 < P < 0.25$ **and accepted** H_A

Example 6:- The table contain Four groups of weights of 19 Fish exposed to experiment of effect of feed.

$Feed_1$	**$Feed_2$**	**$Feed_3$**	**$Feed_4$**
60.8	**68.7**	**102.6**	**87.9**
57.6	**67.7**	**102.1**	**84.2**
65.0	**74.0**	**100.2**	**83.1**
58.6	**66.3**	**96.5**	**85.7**
61.7	**69.8**		**90.3**
$\sum X_1$ =**301.1**	$\sum X_2$=**346.5**	$\sum X_3$ =**401.4**	$\sum X_4$=**431.2**

K_1 = 5 **K_2 =5** **K_3 =4** **K_4 =5**

$\boldsymbol{\bar{X}_1} = 60.62$ $\boldsymbol{\bar{X}_2} = 69.3$ $\boldsymbol{\bar{X}_3} = 100.35$ $\boldsymbol{\bar{X}_4} = 86.24$

H_o: $\boldsymbol{\mu_1}$=$\boldsymbol{\mu_2}$=$\boldsymbol{\mu_3}$=$\boldsymbol{\mu_4}$

H_A: $\boldsymbol{\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4}$

N=$\sum_{i=i}^{k} \boldsymbol{ni}$

$\sum\sum Xij$=$\sum X_1$+$\sum X_2$ +$\sum X_3 + \sum X_4$ = 301.1 + 346.5 + 401.4 + 431.2 =1480.2

$\sum\sum Xij^2$= $(60.8)^2 + (57.6)^2 + (65.0)^2 +(58.6)^2 +(61.7)^2 +(68.7)^2 + (67.7)^2 + (74.0)^2 +(66.3)^2 +(69.8)^2 + (102.6)^2 + (102.1)^2 + (100.2)^2 +(96.5)^2 +(87.9)^2 +(84.2)^2 + (83.1)^2 + (85.7)^2 +(90.3)^2$ =119981.9

N=19

C=$\frac{(\sum\sum Xij)^2}{N} = \frac{(1480.2)^2}{19}$ = 115315.370

Total Sum of Square(SS) =$\sum\sum Xij^2$ – C

= 119981.9 – 115315.370

= 4666.53

Group Sum of Square = $\frac{\sum X_i^2}{K_i}$ – C =$\frac{(\mathbf{301.1})^2}{5} + \frac{(\mathbf{346.5})^2}{5} + \frac{(\mathbf{401.4})^2}{4} + \frac{(\mathbf{431.2})^2}{5}$ – 115315.370

= 18132.24 + 24012.45+ 40280.49 + 37186.688– 115315.370

=119611.868 – 115315.370

= 4296.498

Error Sum of Square (SS)= Total SS – Group SS

= 4666.53 – 4296.498

$= \quad 370.032$

Source of variation	SS	DF	MS
Total	4666.53	18	
Group	4296.498	3	1432.166
Error	370.032	15	24.66

Total DF = N-1=19-1= 18
Group DF = K-1= 4-1=3
Error DF = Total DF – Group DF =18 – 3= 15

Group MS = $\frac{SS}{DF} = \frac{4296.498}{3}$ = 1432.166
Error MS = $\frac{SS}{DF} = \frac{370}{15}$ = 24.66

F = $\frac{MS\ group}{MS\ error} = \frac{1432.166}{24.66}$ = 58.07

$F_{0.05\,(1),3,15} = 3.24$

reject H_o and accepted H_A