List of Experiment

1. Graphical Representation of Data.
2. Measures of Central Tendency (Ungrouped data) with Calculation of Quartiles, Deciles & Percentiles.
3. Measures of Dispersion (Ungrouped Data).
4. Measure of Dispersion (Grouped Data).
5. Moment, Measures of Skewness & Kurtosis (Ungrouped Data).
6. Moments, Measures of Skewness & Kurtosis (Grouped Data).
7. Correlation & Regression Analysis.
8. Application of One Sample t – test.
9. Application of Two Sample Fisher's t – test.
10. Chi – Square test of Goodness of Fit.
11. Chi – Square test of independent of Attributes for 2 X 2 contingency table.
12. Analysis of Variance One Way Classification.
13. Analysis of Variance Two Way Classification.
14. Selection of Random Sample Using Simple Random Sampling.

Experiment No.1

Aim:To perform Graphical Representation of Datain MS Office

Procedure:

1. The first step in creating a graph using Microsoft Excel is entering the data. The data should be in two adjacent columns with the x data in the left column. The columns should be labeled in row one in order to identify what the numbers in the spreadsheet represent

2. Position the cursor on the first X value (i.e., at the top of the column containing the x values, or "Moles of Mg" values), hold down the left mouse button and drag the mouse cursor to the bottom Y value (i.e., at the bottom of the column containing the y values, or "Volume of HCl" values). All of the X-Y values should now be highlighted

3. Click on **Insert** at the top left of the toolbar.

4. Click on **Chart**

5. Click on the box labeled **XY (Scatter)**.

6. Click on **Next >**.

7. Click on the X-Y pattern without lines (Format Option 1).

8. Click on **Next >**; a reduced version of your graph will appear.

9. Click on **Next >**.

10. Click in the rectangular box labeled "Chart Title" and type in a title for the graph (e.g., "Volume of HCl vs. Moles Mg).

11. Click separately on the boxes labeled "Category (X)" and "Value (Y)" and type a label for the X axis (e.g., Moles Mg) and the Y axis (e.g., Volume HCl (mL)).

12. Click on **As New Sheet**. This will instruct the program to plot the data on a separate sheet labeled "Chart1".

13. Click on **Next >**.

14. Click on **Finish**. At this point you will have created an X-Y plot of the data.

Plotting a Best Fit Line

After creating a chart in Microsoft Excel, a best fit line can be found as follows:

1. Be sure you are on the worksheet which contains the chart you wish to work with.

2. Move the mouse cursor to any data pointand press the left mouse button. All of the data points should now be highlighted. Now, while the mouse cursor is still on any one of the highlighted data points, press the right mouse button, and click on Add Trendline from the menu that appears.

3. From within the "Trendline" window, click on the box with the type of fit you want (e.g., Linear).

4. Click on **Options** at the top of the "Trendline" window.

5. Click in the checkbox next to "Display Equation on Chart" and the checkbox next to "Display R-squared Value on Chart". **Do not click on the checkbox next to** "**Set Intercept = 0**".

6. Click **OK**. A line, an equation, and an R-squared value should appear on the graph.

Printing a Graph

To print the graph:

1. Click on **File** in the left-hand corner of the toolbar, and then click on **Page Setup...**.

2. Click on "Header/Footer" at the top of the "Page Setup" window.

3. Click on **Custom Header...**.

4. Click on the box labeled "Left Section" and type in (on two separate lines) your name and your section number.

5. Click on **OK**.

6. Click on **Print...**.

7. Click on **OK**.

8. Retrieve your printout from the printer.

Experiment: 2

Aim: To Perform Measure of Central Tendency

Procedure: Take a sample data.

Calculate mean, median, mode and quartiles.

Marks	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70	70 – 80
No. of Students	2	18	30	45	35	20	6	4

Marks	No. of St. f	c.f	Mid Value	dx'= (X – A) /C	f dx'
0 – 10	2	2	5	-3	-6
10 – 20	18	20	15	-2	-36 -72
20 – 30	30	50	25	-1	-30
30 – 40	45	95	35	0	0
40 – 50	35	130	45	+1	35
50 – 60	20	150	55	+2	40 109
60 – 70	6	156	65	+3	18
70 - 80	4	160	75	+4	16

N= 160 , $\sum fdx' = 37$

$$\overline{X} = A + \frac{\sum fdX'}{N} \times C = 35 + \frac{37}{160} \times 10 = 37.3$$

M = Size of the (N/2)th items = 80th items falls in 30 – 40

$$M = L + \frac{N/2 - cf}{f} \times c = 30 + \frac{80 - 50}{45} \times 10 = 36.67$$

By inspection mode lies in 30 – 40

$$Z = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times c = 30 + \frac{45 - 30}{90 - 30 - 35} \times 10 = 36$$

Q_1 = Size of the (N/4) th items

= (160/4) th items i.e 40th items falls in 20 – 30

$$Q_1 = L_1 + \frac{N/4 - cf}{f} \times c = 20 + \frac{40-20}{30} \times 10 = 26.67$$

$$Q_3 = L_1 + \frac{3(N/4) - cf}{f} \times C = 40 + \frac{120-95}{35} \times 10 = 47.143$$

Experiment: 3&4

Aim: To Perform Measure of Dispersion (Grouped Data)

Procedure: Take a sample Grouped data.

Age in years	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70
No. of Persons	2	4	8	10	12	4

Age in Years	No. of persons f	Mid Value(X)	A=35,dx'=(X-A)/10	C=10 fdx'	fdx'^2
10 – 20	2	15	-2	-4	8
20 – 30	4	25	-1	-4	4
30 – 40	8	35	0	0	0
40 – 50	10	45	+1	10	10
50 – 60	12	55	+2	24	48
60 – 70	4	65	+3	12	36
	N=40			$\sum fdx' = 38$	$\sum fdx'^2 = 106$

$$\sigma = \sqrt{\frac{\sum fdx'^2}{N} - (\frac{\sum fdx'}{N})^2} \times C = \sqrt{\frac{106}{40} - (\frac{38}{40})^2} \times 10$$

= 13.219

Again $\overline{X} = A + \frac{\sum fdx'}{N} \times C = 13.216$

Coefficient of S.D= $\frac{\sigma}{\overline{X}} = \frac{13.219}{44.5} = 0.297$

Experiment: 5

Aim: To Perform Measure of Skewness (Ungrouped Data).

Procedure: Take a sample data.

X	4.5	5.5	6.5	7.5	8.5	9.5	10.5	11.5
f	35	40	48	100	125	87	43	22

Calculation:

X	f	dx = (X – 7.5)	fdx	fdx^2
4.5	35	-3	-105	315
5.5	40	-2	-80	160
6.5	48	-1	-48	48
7.5	100	0	0	0
8.5	125	1	125	125
9.5	87	2	174	348
10.5	43	3	129	387
11.5	22	4	88	352
	N=500		$\sum fdx = 283$	$\sum fdx^2 = 1{,}735$

Coefficient of S.K = $\dfrac{\overline{X} - Mode}{\sigma}$

$$\overline{X} = A + \frac{\sum fdx}{N} = 7.5 + \frac{283}{500} = 8.066$$

$$\sigma = \sqrt{\frac{\sum fdx^2}{N} - (\frac{\sum fdx}{N})^2} = \sqrt{\frac{1{,}735}{50} - (\frac{283}{500})^2} = 1.78$$

Coefficient of S.K = $\dfrac{\overline{X} - Mode}{\sigma}$

= - .244

Experiment: 6

Aim: To Perform Measure of Skewness (Grouped Data).

Procedure: Take a sample data.

Age in Years	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60
No. of Persons	18	20	30	22	10

Table:

Age in Years	No. of Persons	MV(X)	dx=(X – 35)/10	fdx	fdx^2
10 – 20	18	15	-2	-36	72
20 – 30	20	25	-1	-20	20
30 – 40	30	35	0	0	0
40 – 50	22	45	1	22	22
50 - 60	10	55	2	20	40

Coeff. of S.K= $\frac{\overline{X} - Mode}{\sigma}$

$$\overline{X} = A + \frac{\sum fdx}{N} \times i = 35 + \frac{-14}{100} \times 10 = 33.6$$

Mode = 35.56

$$\sigma = 12.33$$

Coeff. of S.K= $\frac{\overline{X} - Mode}{\sigma} = \frac{33.6 - 35.56}{12.33} = -0.159$

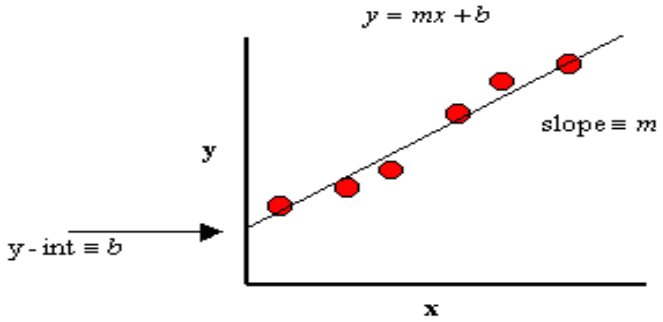
Experiment No.7

Aim: To Perform Linear Regression

Procedure: Take a sample data

X	Y
1.0	2.6
2.3	2.8
3.1	3.1
4.8	4.7
5.6	5.1
6.3	5.3

Say we have a set of data (X, Y) shown at the left. If we have reason to believe that there exists a linear relationship between the variables **x** and **y**, we can plot the data and draw a "best-fit" straight line through the data. Of course, this relationship is governed by the familiar equation Y= mX + C. We can then find the slope, *m*, and y-intercept, *b***,** for the data which are shown in figure below.

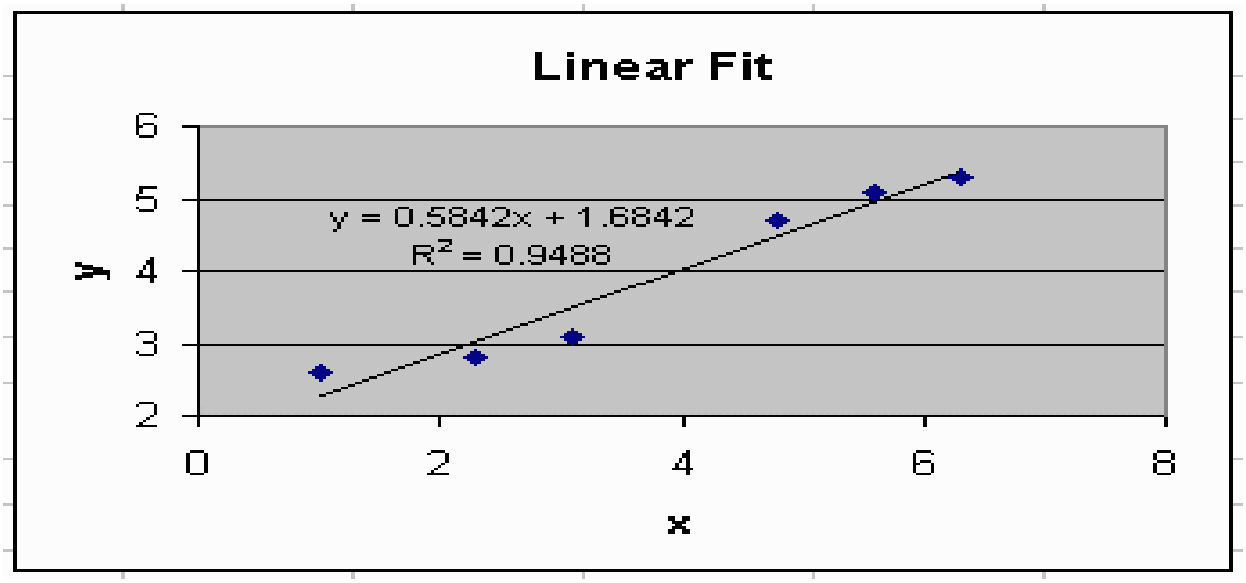


Let's enter the above data into an Excel spread sheet, plot the data, create a trend line and display its slope, y-intercept and R-squared value. Recall that the R-squared value is the square of the correlation coefficient. (Most statistical texts show the correlation coefficient as "*r*", but Excel shows the coefficient as "*R*". Whether you write is as *r* or *R*, the correlation coefficient gives us a measure of the reliability of the linear relationship between the *x* and *y* values. (Values close to 1 indicate excellent linear reliability.)

A11 = =COUNT(B3:B8)

	A	B	C	D	E	F	G	H	I	J
1										
2		x	y	xy	x^2	y^2				
3		1.0	2.6	2.6	1.0	6.8				
4		2.3	2.8	6.44	5.3	7.8				
5		3.1	3.1	9.61	9.6	9.6				
6		4.8	4.7	22.56	23.0	22.09				
7		5.6	5.1	28.56	31.4	26.0				
8		6.3	5.3	33.39	39.7	28.1				
9										
10	n	Σ x	Σ y	Σ (xy)	Σ (x^2)	Σ (y^2)				
11	6	23.1	23.6	103.16	110.0	100.4				
12										
13		$(\Sigma x)^2$	$(\Sigma y)^2$							
14		533.61	556.96							
15										
16	slope, m =	0.5842		=(A11*D11-B11*C11)/(A11*E11-B14)						
17	y-int, b =	1.6842		=(C11-C16*B11)/A11						
18	r =	0.9741		=(A11*D11-B11*C11)/SQRT((A11*E11-B14)*(A11*F11-C14))						

Enter your data as we did in columns B and C. The reason for this is strictly cosmetic as you will soon see.



Linear regression equations.

If we expect a set of data to have a linear correlation, it is not necessary for us to plot the data in order to determine the constants *m* (slope) and *b* (y-intercept) of the equation Y= mX + C. Instead, we can apply a statistical treatment known as linear regressionto the data and determine these constants.

Given a set of data (X, Y) with *n* data points, the slope, y-intercept and correlation coefficient, *r*, can be determined using the following:

$$m = \frac{n\sum(xy)-\sum x \sum y}{n\sum(x^2)-(\sum x)^2}$$

$$C = \frac{\sum \quad y - m\sum x}{n}$$

$$r = \sum \frac{n\sum(xy) - \sum x \sum y}{\sqrt{[n\sum(x^2) - (\sum x)^2][n\sum(y^2) - (\sum y)^2]}}$$

We can use the technique of correlationto test the statistical significance of the association. In other cases we use regression analysis to describe the relationship precisely by means of an equation that has predictive value. We deal separately with these two types of analysis - correlation and regression - because they have different roles.

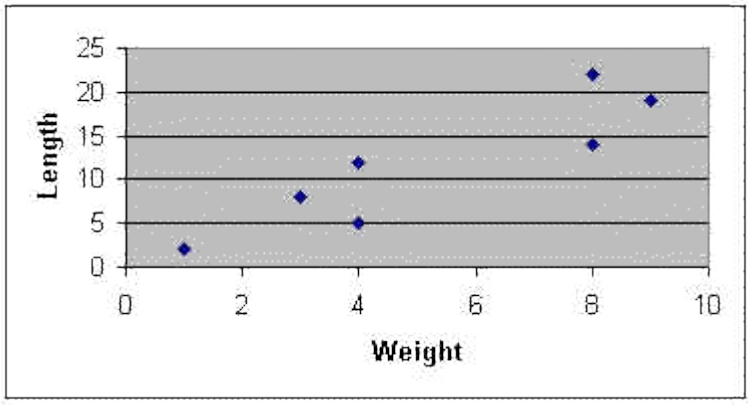
Correlation

Suppose that we took 7 mice and measured their body weight and their length from nose to tail. We obtained the following results and want to know if there is any relationship between the measured variables. [To keep the calculations simple, we will use small numbers.

Mouse	Units of weight (x)	Units of length (y)
1	1	2
2	4	5
3	3	8
4	4	12
5	8	14
6	9	19
7	8	22

Procedure

(1) Plot the results on graph paper. This is the essential first step, because only then can we see what the relationship might be - is it linear, logarithmic, sigmoid, etc?



In our case the relationship seems to be linear, so we will continue on that assumption. If it does not seem to be linear we might need to transform the data.

(2) Set out a table as follows and calculate $\sum x$, $\sum y$, $\sum x^2$, $\sum y^2$, $\sum xy$, $\bar{x}$,and$\bar{y}$(mean of x and*y*).

	Weight (x)	Length (y)	x^2	y^2	xy
Mouse 1	1	2	1	4	2
Mouse 2	4	5	16	25	20
Mouse 3	3	8	9	64	24
Mouse 4	4	12	16	144	28
Mouse 5	8	14	64	196	112
Mouse 6	9	19	81	361	152
Mouse 7	8	22	64	484	176
Total	$\sum x = 17$	$\sum y = 82$	$\sum x^2 = 251$	$\sum y^2 = 1278$	$\sum xy = 553$
Mean	$\bar{x}$= 5.286	$\bar{y}$=11.714			

(3) Calculate $\sum dx^2 = \sum x^2 - \frac{(\sum x)^2}{n}$ = 55.429 in our case.

(4) Calculate $\sum dy^2 = \sum y^2 - \frac{(\sum y)^2}{n}$= 317.429 in our case.

(5). $$\sum dxdy = \sum xy - \frac{\sum x \sum y}{n}$$

(6). $$r = \frac{\sum dxdy}{\sqrt{\sum dx^2 \sum dy^2}} = 0.9014$$

Experiment No.8&9

Aim: To Perform One Sample and Two Sample t – test

Procedure:

One – Tailed and Two – Tailed Test

θ_0 = Population Parameters

$\theta =$ Corresponding Sample Statistic

If $H : \theta = \theta_0$ then the alternative hypothesis (H_1) which is complementary to θ_0 can be any one of the following:

(i) $H_1 : \theta \neq \theta_0$, i.e $\theta > \theta_0$ or $\theta < \theta_0$

(ii) $H_1 : \theta > \theta_0$

(iii) $H_1 : \theta < \theta_0$

Table:

Nature of Test	LOS 1%	2%	5%	10%
Two Tailed	$\lvert z_\alpha \rvert$=2.58	$\lvert z_\alpha \rvert$=2.33	$\lvert z_\alpha \rvert$=1.96	$\lvert z_\alpha \rvert$1.645
Right Tailed	z_α=2.33	z_α=2.055	z_α=1.645	z_α=1.28
Left Tailed	z_α=-2.33	z_α=-2.055	z_α=-1.645	z_α=-1.28

Learning Objectives

• Understand the steps of hypothesis testing.

• Know when to use a z-test and when to use a t-test to test a single-sample hypothesis.

• Compare the t-distribution to the Normal distribution.

• Understand degrees of freed

Steps of Hypothesis Testing

Recall that hypothesis testing is a form of statistical inference. Previously, we inferred about a population by

calculating a confidence interval. We estimated the true mean of a population from a sample mean and created a confidence interval provided a margin of error for our estimate. In this chapter, we also use sample data to help us make decisions about what is true about a population.

When conducting a hypothesis test, we are asking ourselves whether the information in the sample is consistent, or inconsistent, with the null hypothesis about the population. We follow a series of four basic steps:

1. State the null and alternative hypotheses.

2. Select the appropriate significance level and check the test assumptions.

3. Analyze the data and compute the test statistic.

4. Interpret the result

If we reject the null hypothesis we are saying that the difference between the observed sample mean and the hypothesized population mean is too great to be attributed to chance. When we fail to reject the null hypothesis, we are saying that the difference between the observed sample mean and the hypothesized population mean is probable if the null hypothesis is true. Essentially, we are willing to attribute this difference to sampling error.

Conducting a Hypothesis Test on One Sample Mean When the Population Parameters are Known

Although this is rarely the case, we can use our familiar z-statistic to conduct a hypothesis test on a single sample mean. In short, we find the z-statistic of our sample mean in the sampling distribution and determine if that z-score falls within the critical (rejection) region or not. This test is only appropriate when you know the true mean and standard deviation of the population.

Example A

The school nurse thinks the average height of 7th graders has increased. The average height of a 7th grader five years ago was 145 cm with a standard deviation of 20 cm. She takes a random sample of 200 students and finds that the average height of her sample is 147 cm. Are 7th graders now taller than they were before? Conduct a single-tailed hypothesis test using a .05 significance level to evaluate the null and alternative hypotheses.

First, we develop our null and alternative hypotheses:

$$H_0 : \mu \leq 145$$

$$H_\alpha : \mu > 145$$

Choose a α = .05. The critical value for this one tailed test is z=1.64. This is a one-tailed test, and a z-score of 1.64

cuts off 5% in the single tail. Any test statistic greater than 1.64 will be in the rejection region.

Next, we calculate the test statistic for the sample of 7th graders.

$$Z = \frac{147 - 145}{\frac{20}{\sqrt{200}}} \approx 1.414$$

The calculated z score of 1.414 is smaller than 1.64 and thus does not fall in the critical region. Our decision is to fail to reject the null hypothesis and conclude that the probability of obtaining a sample mean equal to 147 is likely to have been due to chance.

T-Test for One Sample Mean:

So when do we use the t-distribution and when do we use the normal distribution? It's simple: When we know the population standard deviation we use the normal distribution. When we don't know the population standard deviation (so we need to use our sample standard deviation), we use the t-distribution.

We use the Student's t -distribution in hypothesis testing the same way that we use the normal distribution. Each row in the t distribution table (see above) represents a different t distribution. Each distribution is associated with a unique number of degrees of freedom (the number of observations minus one). The column headings in the table represent the portion of the area in the tails of the distribution –we use the numbers in the table just as we used the z scores.

In calculating the t - test statistic, we use the formula:

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

where:

t is the test statistic and has n – 1 degrees of freedom.

$\bar{x}$ is the sample mean

μ_o is the population mean under the null hypothesis.

s is the sample standard deviation

n is the sample size

$\frac{s}{\sqrt{n}}$ is the estimated standard error

Assumptions of the single sample t-test:

• A random sample is used.

• The random sample is made up of independent observations

• The population distribution must be nearly normal, or the size of the sample is large.

Example:

The high school athletic director is asked if football players are doing as well academically as the other student athletes. We know from a previous study that the average GPA for the student athletes is 3.10. After an initiative to help improve the GPA of student athletes, the athletic director randomly samples 20 football players and finds that the average GPA of the sample is 3.18 with a sample standard deviation of 0.54. Is there a significant improvement? Use a 0.05 significance level.

• Hypothesis Step 1: Cleary state the null and alternative hypotheses.

$$H_0 : \mu = 3.10$$

$$H_\alpha : \mu \neq 3.10$$

• Hypothesis Step 2: Identify the appropriate significance level and confirm the test assumptions.

We were told that we should use a 0.05 significance level. We assume that each football player is independently tested –that their GPA is not related to another football player's GPA. Without the data, we have to assume that the sample is nearly normal (the sample is an indication of the population shape). The size of the sample also helps here, as we have 20 players. So, we can conclude that the assumptions for the single sample T-test have been met.

• Hypothesis Step 3: Analyze the data

We use our t-test formula:

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{3.18 - 3.10}{\frac{0.54}{\sqrt{20}}} = 0.66$$

We know that we have 20 observations, so our degrees of freedom for this test are 19. Nineteen degrees of freedom at the 0.05 significance level gives us a critical value of $\pm$ 2.093.

• Hypothesis Step 4: Interpret your results

Since our calculated t-test value is lower than our t-critical value, we fail to reject the Null Hypothesis.

Therefore, the average GPA of the sample of football players is not significantly different from the average GPA of student athletes. Therefore, we can conclude that the difference between the sample mean and the hypothesized value is not sufficient to attribute it to anything other than sampling error. Thus, the athletic director can conclude that the mean academic performance of football players does not differ from the mean performance of other student athletes.

Experiment No.10&11

Aim: (i) To Perform Chi – Square Test

Procedure:

The chi-square test is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories. Do the number of individuals or objects that fall in each category differ significantly from the number you would expect? Is this difference between the expected and observed due to sampling error, or is it arealdifference?

Chi-Square Test Requirements

1. Quantitative data.

2. One or more categories.

3. Independent observations.

4.Adequate sample size (at least 10).

5. Simple random sample.

6. Data in frequency form.

7. All observations must be used.

Expected Frequencies:

When you find the value for chi square, you determine whether the observed frequencies differ significantly from the expected frequencies. You find the expected frequencies for chi square in three ways:

. You hypothesize that all the frequencies equal each category. For example, you might expect that half of the entering freshmen class of 200 at Tech College will be identified as women and half as men. You figure the expectedfrequency by dividing the number in the sample by the number of categories. In this exam pie, where there are 200 entering freshmen and two categories, male and female, you divide your sample of 200 by 2, the number of categories, to get 100 (expected frequencies) in each category.

2. You determine the expected frequencieson the basis of some prior knowledge. Let's use the Tech College example again, but this time pretend we have prior knowledge of the frequencies of men and women in each category from last year's entering class, when 60% of the freshmen were men and 40% were women. This year you might expect that 60% of the total would be men and 40% would be women. You find the expected frequencies by multiplying the sample size by each of the hypothesized population proportions. If the freshmen total were 200, you would expect 120 to be men (60% x 200) and 80 to be women (40% x 200).

Now let's take a situation, find the expected frequencies, and use the chi-square test to solve the problem. Situation Thai, the manager of a car dealership, did not want to stock cars that were bought less frequently because of their unpopular color. The five colors that he ordered were red, yellow, green, blue, and white. According to Thai, the expected frequencies or number of customers choosing each color should follow the percentages of last year. She felt 20% would choose yellow, 30% would choose red, 10% would choose green, 10% would choose blue, and 30% would choose white. She now took a random sample of 150 customers and asked them their color preferences. The results of this poll are shown in Table 1 under the column labeled observed frequencies."

Category Color	Observed Frequencies	Expected Frequencies
Yellow	35	30
Red	50	45
Green	30	15
Blue	10	15
White	25	45

PROCEDURE for above Example

We are now ready to use our formula for $X^2$and find out if there is a significant difference between the Observed and expected frequencies for the customers in choosing cars. We will set up a worksheet; then you will follow the directions to form the columns and solve the formula.

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

where O is the Observed Frequency in each category

E is the Expected Frequency in the corresponding category is sum of

df is the "degree of freedom"(n-1)

$X^2$is Chi Square

We are now ready to use our formula for $X^2$and find out if thereisa significant difference between the observed and expected frequencies for the customers in choosing cars. We will set up a worksheet; then you will follow the directions to form the columns and solve the given formula.

Category Color	Observed Frequencies	Expected Frequencies	O – E	$(O - E)^2$	$(O - E)^2 /E$
Yellow	35	30	5	25	0.83
Red	50	45	5	25	0.56
Green	30	15	15	225	15
Blue	10	15	-5	25	1.67
White	25	45	-20	400	8.89

$X^2 = 26.95$

2. After calculating the Chi Square value, find the "Degrees of Freedom.

3. Find the table value for Chi Square. Begin by finding the df found in step 2 along the left hand side of the table. Run your fingers across the proper row until you reach the predetermined level of significance (.05) at the column heading on the top of the table. The table value for Chi Square in the correct box of4 df and P=.05 level of significance is 9.49.

4. If the calculated chi-square value for the set of data you are analyzing (26.95) is equal to or greater than the table value (9.49), rejectthe null hypothesis. There IS a significant difference between the data sets that cannot be due to chance alone.If the number you calculate is LESS than the number you find on the table, than you can probably say that any differences are due to chance alone.

The steps in using the chi-square test may be summarized as follows:

Chi-Square Test Summary

I. Write the observed frequencies in column Os

2. Figure the expected frequencies and write them in column E.

3. Use the formula to find the chi-square value:

4. Find the df. (N-1)

5. Find the table value (consult the Chi Square Table.)

6. If yourchi-square value is equal to or greater thanthe table value, reject the null hypothesis: differences in your data are not due to chance alone

For example, the reason observed frequencies in a fruit fly genetic breeding lab did not match expected frequencies could be due to such influences as:

•Mate selection (certain flies may prefer certain mates)

•Too small of a sample size was used

•Incorrect identification of male or female flies

•The wrong genetic cross was sent from the lab

•The flies were mixed in the bottle (carrying unexpected alleles)

Experiment No.12

Aim: (i) To Perform Analysis of Variance One Way Classification.

Procedures for a One-Way ANOVA

we will compute a one-way ANOVA for data from three independent groups. The raw data for the 16 subjects are listed below. Note that this is a between-subjects design, so different people appear in each group.

The Raw Data

Here are the raw data from the three groups (6 people in Group 1, and 5 each in Groups 2 and 3).

Group 1	Group 2	Group 3
3	4	9
1	3	7
3	5	8
2	5	11
4	4	9
3		

The first step in the computation is to add the scores in each column and compute the sum of the squared scores for each column. We also count the number of scores in each column, compute the mean by dividing the sum by the number of scores, and compute the sum of squares.

We defined the sum of squares in Chapter 5 as the average squared deviation from the mean. However, there is an easier computational formula for the SS, which you learned in the section of this website that showed you how to compute the variance. Listed below is both the definitional formula (first part) and the computational formula (second part).

$$SS = \sum (X - \overline{X})^2$$

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

Summary Statistics

We have done the computations described above for each of the three groups and organized them in three columns. We have also included a fourth column to put the total scores, total sum of X^2, and total sample size, all of which will also be needed for the computations. In this way, we have all of the values that we will need for the computation of a one-way ANOVA at our fingertips.

	Group 1	Group 2	Group 3	Totals
Sum of X	16	21	44	81
Sum of X^2	48	91	396	535
n	6	5	5	16
Mean	2.67	4.20	8.80	
SS	5.33	2.80	8.80	

Compute the SSs for the ANOVA

The formulas for computing the three sums of squares (between, within, and Total) are shown below, with the numbers plugged in. The notation may look complicated, but all of the needed values can be found in the summary table that we just prepared. The only real new terminology uses summation notation in which we sum across the groups (*i*, which refers to the group number, goes from 1 to *k*, which is the number of groups). We use this notation because we can have any number of groups in a design like this and we want a formula that will describe what we should do regardless of how many groups we have.

Most students find it easier to understand the notation by looking at the formula and see where the numbers for the formula can be found in the summary table above. We have used more parentheses than actually needed algebraically to specify what needs to be done. The rule is that you always do things inside a parenthesis before you do things outside of the parenthesis. If you remember that simple rule, you will not have to remember the more complicated algebraic rules about what computations should be done first.

$$SS_b = \{\sum_{i=1}^{K} \frac{((\sum X)_i)^2}{n_i}\} - \frac{((\sum X)_T)^2}{N}$$

$$= \{\frac{16^2}{6} + \frac{21^2}{5} + \frac{44^2}{5}\} = 108.00$$

$$SS_W = \sum_{i=1}^{k} SS_i = 5.33 + 2.80 + 8.80 = 16.93$$

$$SS_T = (\sum X^2)_T - \frac{((\sum X)_T)^2}{N} = 535 - \frac{81^2}{16} = 124.94$$

Double check the computation of the SSs by seeing if they add up. $SS_T = SS_b + SS_w = 108.00 + 16.93 = 124.93$

Fill in the Summary Table

The df_b is equal to the number of groups (*k*) minus 1. The df_w is equal to the total number of participants minus the number of groups (*N* - *k*). The df_T is equal to the total number of participants (*N*) minus 1. Note that the df_T is equal to the df_b plus the df_w in the same way that the SS_T is equal to the sum of the SS_b and SS_w.

The MSs are computed by dividing the SSs by their respective dfs, and the *F* is computed by dividing the MS_b by the MS_w. All of these values have been inserted into the standard summary table for ANOVA below.

Source	df	SS	MS	F
Between	2	108.00	54.00	41.46
Within	13	16.93	.30	
Total	15	124.94		

The final step is to compare the value of the *F* computed in this analysis with the critical value of *F* in the *F* Table. You look up the critical value by using the degrees of freedom. In our case, the df_b is 2 and the df_w is 13. The critical value of *F* for an alpha of .05 is 3.80. Since our obtained *F* exceeds this value, we reject the null hypothesis and conclude that there is a significant difference between the groups.