MODULE I

MEASURES OF RELATIONSHIP

Measures of relationships between sex and performance height and mass (weight) and similar comparisons of variables especially when they are more than one. These are the areas where correlation and regression operates. They help in determining the extent and nature of relationship between variables.

As educationist, we may be interested in the relationship between the Biology test and the grade point average (GPP). Or we may be interested in the relationship between intelligence scores and performance of science test. Correlation co-efficient are used in establishing such relationships. The symbol for correlation co-efficient is r.

The value of the correlation co-efficient r, vary between 1.00 and -1.00. A correlation of +1.00 represents a perfect positive relationship between two set of test scores under consideration. This means that individuals obtaining high scores on one test tend to obtain high scores on the second test. The converse is also true, that is, individuals scoring low on one score tend to score low on a second score.

A correlation of -1.00, on the other hand, represents a perfect negative relationship between the two set of test scores. In this case individuals scoring low on one test tend to score high on the second test.

A correlation co-efficient of zero represents a complete absent of a relationship.

It should be pointed out that nature does not structured event into perfect relationship (be it positive or negative) with each other. In reality, the correlation we observe between the psychological measures such as performance in Arithmetic and Spelling depart from perfect. They are such that may be high but lower that +1.00/

There are many types of correlation co-efficient.

For instance, we have:

1. Pearson Product-Moment Correlation
2. Spearman Rank Order Correlation or simply called Spearman Rho, and Kendall Rank's Correlation
3. Multiple Correlation
4. Partial Correlation
5. Point-Biserial Correlation
6. Biserial Correlation
7. Phi Correlation
8. Tetrachoric Correlation

The decision of which one to employ with a specific test scores depends upon such factors as:

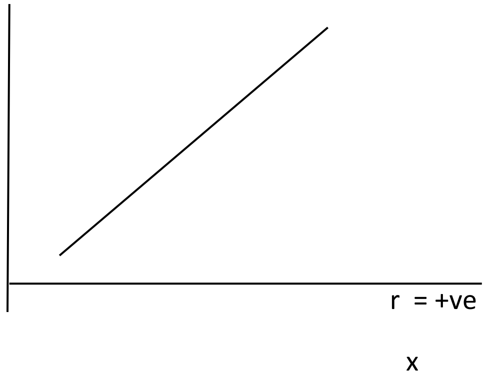
a. the type of scale of measurement in which each of the scores is expressed;
b. the nature of the underlying distribution (continuous or discrete);
c. the characteristics of the scores (Linear or non-linear).

For our present discussion, only the first two correlational types will be treated.

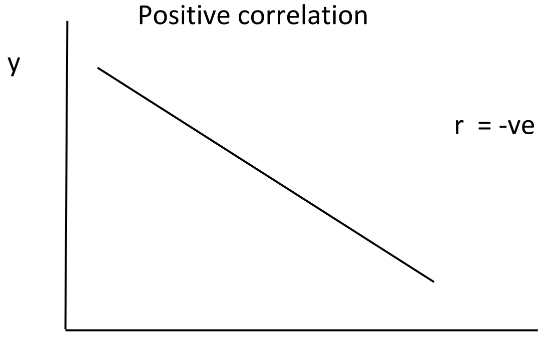
Types of Correlation

Correlation measures the degree of association (relationship between two variables.

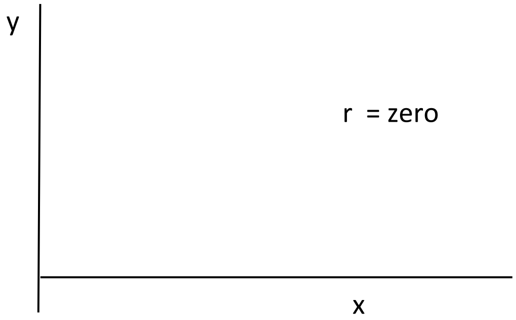
Therefore two variables are said to be correlated or related when change one variable result as change of the other variable. The degree of correlation (r) between two variables x & y is expressed by a real number when lies between -1 & +1 (-1 < r < +1); simple correlation is basically classified according to the values of its coefficient. Hence, there are



Positive correlation



Negative correlation



No correlation

PEARSON PRODUCT MOMENT CORRELATION

There are a lot of computational processes employed in the calculation of Pearson ‘r’. The following two are more commonly employed.

1. The mean deviation formula
2. The raw score formula

The mean deviation formula employs a fundamental expression in the form of:

$$r = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{{}^{NS} X \, {}^{S} Y}$$

The X and Y are variables, e.g. Mathematics and Physics scores. It should be noted in the expression above that (X-X) and (Y-Y) are deviation scores similar to the one encountered while finding the standard deviation.

Each of the deviation expressions measures the departure of individual X and Y scores about their means X and Y respectively. If the deviations scores are expressed as simple variables such that (X-X) is represented by X, and (Y-Y) is by Y, then the above formula can be stated as:

$$r = \frac{\sum XY}{{}^{NS} x \, {}^{S} y}$$

where,

r = Correlation between X scores and Y scores

x = Deviation of X score from the mean of X scores

Y = Deviation of a corresponding Y from the mean of the Y scores

XY = The sum of all the products of deviations, such X deviation times its corresponding Y deviation.

S_x and S_y = standard deviations of y and Y

This procedure for calculating the Pearson ‘r’ is appropriate and effective if in the same calculation we have separate need for the standard deviation of each of the test scores we are considering. If the standard deviation are not specially required, then the formula above could be unnecessarily cumbersome to use. We may use the shorter method that omits the computation of S_x and S_y.

Method 1: The Mean Deviation Method

$$r = \sum \frac{XY}{\sqrt{(\sum X^2)(\sum Y^2)}}$$

where,

X^2 = sum of squared deviation of X scores

Y^2 = sum of squared deviation of Y scores

XY = sum of all products of deviations,

each X deviation times its corresponding Y deviation.

Let us now use the above formula to illustrate the calculation of ‘r’ by deviation method

Table

Calculation of Pearson Product Moment

Correlation Coefficient between Two Set of scores

	X	Y	(X-X)	(Y-Y)	X^2	Y^2	XY
Giyath	13	11	4.6	3.1	21.16	9.61	14.26
Solape	12	14	3.6	6.1	12.96	37.21	21.96
Moronkola	11	12	2.6	4.1	6.76	16.81	10.66
Anita	9	9	.6	1.1	.36	1.21	0.66
Binta	8	6	-0.4	-1.9	.16	3.61	0.76

Yegoa	8	7	-0.4	-.9	.16	.81	0.36
Daniel	7	7	-1.4	-.9	1.96	.81	1.26
Lilian	7	5	-1.4	-2.9	1.96	8.11	4.26
Mercy	5	6	-3.4	-1.9	11.56	3.61	6.46
Sandra	4	2	-4.4	-5.9	19.36	34.81	25.96
	∑X=84	∑Y=79	0	0	76.40	116.9	86.40
	$\bar{X}$=8.4	$\bar{Y}$=7.9			$\sum X^2$	$\sum Y^2$	∑XY

$$r = \sum \frac{\sum XY}{\sqrt{(\sum X^2)(\sum Y^2)}}$$

$$r = \frac{86.40}{\sqrt{(76.4)(116.9)}}$$

$$r = \frac{86.40}{94.5048}$$

= 0.9142

The steps involved in the computation of the Pearson Product Moment Correlation Coefficient of the above data can be presented thus:

Step 1: List in parallel columns the paired X and Y scores noting that corresponding scores are together

Step 2: Determine the mean of the two set of scores.

Step 3: Determine for every pair of scores the deviations X and Y and check that the sum of them equals zero.

Step 4: Square individual deviation of the X and Y columns to obtain X^2 and Y^2

respectively.

Step 5: Find the sum of the squares of each of the two set of deviation scores to obtain $\sum X^2$ and $\sum Y^2$

Step 6: Find the cross-product of each of pair of the deviation scores to obtain XY.

Step 7: Sum the cross product to obtain $\sum XY$

Step 8: Substitute the values of step 5 and step 7 in the formula as illustrated in the worked example.

Method II

The Raw of Data Procedure

The first procedure discussed does not make use of original data in calculating Pearson 'r'. This makes the method a little difficult to handle by beginners. The raw data procedure aims at reducing the amount of work to be done. This is done by using the formula below:

$$r = \frac{N\sum XY - (\sum X)(\sum Y)}{(N\sum X^2 - (X)^2)(N\sum y^2 - (\sum y)^2)}$$

where,

X and Y are the original scores on the individuals with respect to the variable under consideration.

N is the number of paired cases.

The steps to follow are:

Step 1: List in parallel columns the paired X and Y scores

Step 2: Find the sum of scores of X and Y columns to obtain X and Y.

Step 3: Find the square of the individual X and Y scores to obtain X^2 and Y^2.

Step 4: Sum square of the individual X and Y columns above to obtain $\sum Y^2$.

Step 5: Determine the cross product of each paired X and Y to obtain XY

Step 6: Sum of the cross product above to obtain $\sum XY$.

Step 7: Substitute the values obtained in step 1 to step 6 above in the formula. Noting that the N stands for the paired cross product.

The next table shows computational procedures for Pearson'r' from the original scores.

Table 2: Computational Procedures for Pearson 'r' from the original scores.

X	Y	X^2	Y^2	XY
13	11	169	121	143
12	14	144	196	168
11	12	121	144	132
9	9	81	81	81
8	6	64	36	48
8	7	64	49	56

7	7	49	49	49
7	5	49	25	35
5	6	25	36	30
4	2	16	4	8
∑X84	∑Y=79	∑X^2	∑Y^2	∑XY
X=8.4	Y=7.9	782	741	750

$$r = \frac{N\sum XY - (\sum X)(\sum Y)}{(N\sum X^2 - (\sum X)^2)(N\sum y^2 - (\sum y)^2)}$$

$$= \frac{10\,x750 - 6636}{\sqrt{(7820 - 7056)(7410 - 6241)}}$$

$$r = \frac{864}{\sqrt{(764)(1164)}}$$

$$= 0.91$$

The answer obtained in this process is the same with the one we get through the first method used.

It should be noted that the use of Pearson r is justified by the assumption that the relationship between the traits considered is Linear and that the scores on the traits are normally distributed. The scores must be at the interval or ratio level of measurement.

MODULE II

SPEARMAN RANK ORDER CORRELATION

This type of correlation is also referred to as Correlation by Ranks: Rank correlation or Spearman Rho.

Instead of actual values or scores, it is possible to attach ranks to variables e.g. X and Y. the idea was introduced in 1906 by C. Spearman. He used the algebraic formulae for the sum of natural numbers together with the sum of their squares.

Spearman coefficient is represented by the formula

$$^{r}\text{rho} = .1 - \quad 6\frac{6\sum d^{2}}{n(n^{2}-1)}$$

where

d = difference in the rank, i.e. d = rank X, rank Y

n = the number of ranked pairs

The Spearman r, ^{r}rho is appropriate when one of the scores whose relationship you want to fight with another one is at ordinal level of measurement (rank) and the other one may be ordinal or higher.

In a situation where a class teacher intends to find out whether social withdrawal among the student is related to performance in Mathematics, the teacher therefore need to obtain the scores of a number of pupils in a Mathematics test and thereafter rank the student who have the highest in social relation to the lowest.

The resulting measurement constitute of course an ordinal scale. The Spearman r is preferable. Now following the ranking of student on the social withdrawal trait, you must rank them on Mathematics score. The data obtained from the students are presented in the table.

Table 3

Withdrawal Rank	Mathematics Rank	D	D^2
1	3	-2	4
2	2	0	0
3	5	-2	4
4	1	3	9
5	5	0	0
6	5	1	1
7	9.5	-2.5	6.25
8	7	1	1
9	8	1	1
10	9.5	.5	0.25
		$\sum D$	$\sum D^2$
		0	26.5

$$^{r}\text{rho} = 1 - \frac{6\sum D^{2}}{N(n^{2}0-1)}$$

$$^{r}\text{rho} = 1 - \frac{6x26.5}{10(100-1)}$$

$$= 1 – 0.160606$$

$$= .839$$

$$= .84$$

The step in the procedure can be summarized as

Step 1: List the ranking of the first variable

Step 2: Determine the ranks of the second variable and pair them with the first accordingly.

Step 3: Find the differences in ranks to obtain D.

Step 4: Sum of the differences to obtain ∑D which must equal zero

Step 5: Find the square of differences to obtain D^2

Step 6: Sum the square of difference to arrive at $∑D^2$.

Step 7: Substitute the values of $∑D^2$ into the formula.

It is essential to note that correlation coefficient is very important because it has wide application in testing and psychology. For a number of reasons it is very important, for instance, correlation is a method used in establishing the reliability of tests. When the consistency with which a measuring instrument measures that it purports to measure is to be determined, correlation coefficient are usually calculated. This may be in finding the relationship between two parallel forms of the same test or the relationship of the scores on the same administered at two different points in time.

Second, correlation coefficient underline all forms of prediction, while it should be noted that more correlation should not be noted that more correlation should not be interpreted in terms of causal relationship, prediction of future occurrence of events derive directly from some sort of correlation. We might want to use a scholastic aptitude test as a predictor of grade point average, thus, correlation of test with the grades would give an indication of the test usefulness as a predictor.

Exercise

Name of Contestants arranged alphabetically	Ranking in Contest	
	Poise	Beauty
	X	Y

	Bunmi	1	3
1.	Beatrice	5	5
	Cornelia	2	1
	Funke	4	4
	Folake	3	2

Find the difference between each contestant rank on X and on Y.

Prepare another table to include d and d^2

2. Calculate Spearman rho for beauty contest

MODULE III SCATTER DIAGRAM

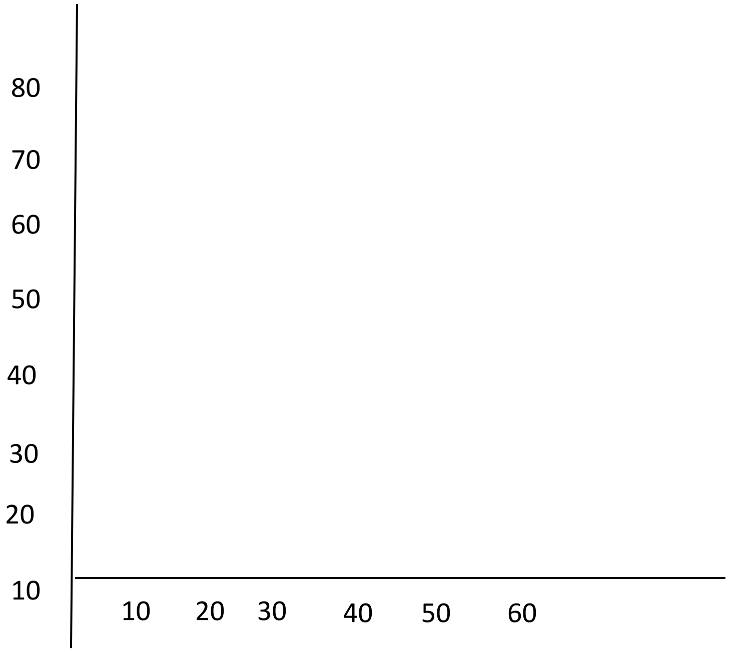
The scatter Diagram

A graphical presentation of bivariate data on a two-axis coordinate graphics known as scatter diagram. Here the bivariate data are plotted on a rectangular coordinate system in order to see the exiting relationship between the two variables under study. The following bivariate data will be used as illustration.

Draw the scatter diagram for the following data

Score X in %	30	40	35	40	20	25	50
Scores Y in %	50	70	65	68	40	60	80

Solution



MODULE IV

REGRESSION ANALYSIS

Regression shows a relationship between the average values of two variables. This regression is very helpful in estimating and predicting the average value of one variable for a given value of the other variable. The estimate or prediction may be made with the help of variable y. The best average value of one variable associated with the given value of the other variable may also estimated or predicted by means of an equation and the equation is known as Regression equation.

Types of Regression

(1) Simple regression: The regression analysis confirmed to the study of only two variables at a time is called simple regression

(2) Multiple regression: The regression analysis for studying more than two variables at a time is known as multiple regression

Methods of obtaining or fitting regression line

We have two methods of fitting the regression line. These are

(1) Graphical and (11) Algebraic

(1) Graphical Method: The following steps are to be taken

(i) Draw the scatter diagram for the data

(ii) Look at two points that a straight line will pass through on the diagram (x,y)

(iii) Estimate constant a and b from the graph

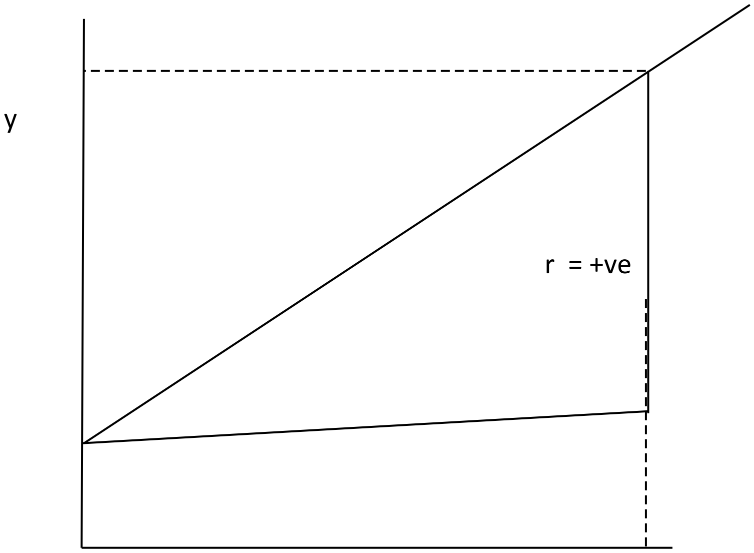
a = Intercept on the y – axis

b = slope or gradient

(iv) Regression line y = a + bx

X	0	1	2	3	4	5	6

Y	2	1	3	2	4	3	5



$$\text{Slope (b)} = \frac{\text{vertical length} \quad x}{\text{Horizontal length}} = \frac{5 - 2}{6 - 0} = \frac{3}{6} = 0.5$$

a = 2

y = a + bx ➔ y = 2 + 0.5x

(b) Algebric Method: in the algebraic method, we use the normal equation when is derived by the least squares method the said normal equations are

$na + bEx = Ey$ .... (1)

$aEx + bEx^2 = Exy$ .... (2)

which are used to fit the regression line if y on x as

$y = abx$

it should be noted that when the equation (1) and (2) are solved simultaneously, we have the following estimated of a & b

b = nExy – ExEy --- (3)

a = y - b x

Example 1: Use the table below to calculate or fit the regression line of y on x

X	Y	Xy	X^2
1	3	3	1
2	5	10	4
3	7	21	9
4	9	36	16
5	11	125	55

b = (5 x 125) – (15x 35) = 625 – 525 = 2

$(5 X 55) – (15)^2$ 275 -225

$$a = \frac{35}{5} - 2\left(\frac{15}{5}\right) = 7 - 2 x 3 = 1$$

y = 1 + 2x is the income and expenditure (in ₦) of a man for 5 months. Find the least square regression line of y in x.

X	Y

0	1.0
1	1.8
2	3.3
3	4.5
4	6.3

Solution

X	Y	Xy	X^2
0	1.0	O	0
1	1.8	1.8	1
2	3.3	6.6	4
3	4.5	13.5	9
4	6.3	25.2	16
10	16.9	47.1	30

Y = a + bx

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{(5 \times 47.1) - (10 \times 16.9)}{(5 \times 30) - (10)^2}$$

$$= \frac{235.5-169}{150-100} = \frac{66.5}{50} = 1.33$$

$$a = y - b\,x = \frac{16.9}{5} - 1.33\left(\frac{10}{5}\right)$$

$$= 3.38 - 2.66 = 0.72$$

$$a = 0.72$$

$y = a + bx$ ➔ $y = 0.72 + 1.33x$

Exercise

The table below shows the income and saving of a student

X(000)	7	5	10	15	13	14	8	11
Y(000)	1	1	4	6	5	2	2	4

(a) Construct a scatter diagram

(b) Find the least square regression line of y on x

(c) How much is student likely save if the income is ₦20,000

MODULE V

CHI SQUARE

This is another distribution of considerable theoretical and practical importance. When researchers are making their guesses in an experimental situation, they may want to compare observed with theoretical frequencies. Information obtained empirically be directed observation in an experiment are referred to as observed frequencies. While the theoretical frequencies are generated on the basis of some hypothesis, which is different from the data obtained. The researcher however may want to find out whether the difference between the observed and theoretical frequencies are significant. It is the result that will then give him/her the opportunity to reject or accept his hypothesis.

It should be explained further that, on the basis of probability, certain expectations exist from change especially for previously established distributions. And in actual fact, things do not work as expected. Instead there is actual happenings as opposed to theoretical predictions. For example, in a market research survey of two types of ball pen among the university undergraduates, Eleganza ball pen and Bic ball pen were distributed to random sample of 100 undergraduates at University of Ibadan. After about 6 weeks, the students were requested to indicate which of the ball pen they prefer. The result shows that 60 prefer Bic ball pen while 40 prefer Eleganza ball pen.

The probability is that we expect the undergraduate to come out with 50:50 response (that is, a change of 50:50 is expected from the result).

TABLE 1 R x C CONTIGENCY TABLE

ROW(R)	1	2	3	4	……..	C
	n11	n12	n13	n14		n1c
2	n21	n22	n23	n24		n2c
3	n31	n32	n33	n34		n3c
4	n41	n42	n43	n44		n4c
. .	. .	. .	. .	. .		. .
r	nr1	nr2	nr3	nr4		nrc
total	n.1	n.2	n.3	n.4		n.c

The total frequency in each row or column is called marginal frequency. This marginal frequency fro R_1 is n_1 and that if column j called n_j.

Note that $P_{ij} = n_{ij}$

$$P(R_i) = \sum_j^c \frac{n_{ij}}{n} = \frac{n_i}{n} = P_{i\neg}$$

$$P(C_i) = \sum_i^r \frac{n_{ij}}{n} = \frac{n_i}{n} = P_{i\neg j}$$

$$n = \sum_{i}^{r} n_{i} = \sum_{j}^{c} n_{ij}$$

$$\sum_{i}^{r} P_{i} = \sum_{j}^{c} P_{ij} = 1$$

Given that $P_{ij} = P(R_i \text{ n } C_i)$ under the null hypothesis, if the row and column classifications are independent, then

$$P_{ij} = P(R_i)\,P(C_i)$$

$$= \left(\frac{n_i}{n}\right)\left(\frac{n_j}{n}\right)$$

$$= P_i P_{ij}$$

Therefore, the expected frequency for each cell is obtained as

$$E_{ij} = nP_{ij}^{2}$$

$$= n\left(\frac{n_i}{n}\right)\left(\frac{n_j}{n}\right)$$

$$= \frac{(n_i)\,(n_j)}{n}$$

$$= \frac{\text{row i total x column j total}}{\text{Grand total}}$$

For large n_i the statistics

$$X^2 = \sum_{i=1}^{r}\sum_{i=j}^{c}\frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

$$= \quad \sum_{i=1}^{r}\sum_{i=j}^{c}\frac{(0_{ij}-E_{ij})^{2}}{E_{ij}} \qquad \text{where } n_{ij} = 0$$

Is approximately distributed as chi-square with (r-1) (c-1) degree of freedom if the hypothesis is true. Hence, we would reject the null hypothesis of independence if the calculated value of the test statistic is greater than the tabulated value $X^2(X^2_{\alpha,(r-1),(c-1)})$.

Note: That the chi-square as expressed above, can be written in any of the following equivalent form

$$X^2 = \sum_{i=1}^{r}\sum_{i=j}^{c}\frac{(n_{ij}-nP_{ij})^{2}}{nP_{ij}}$$

Or $$X^2 = \sum_{i=1}^{r}\sum_{i=j}^{c}\frac{\left(n_{ij}-\frac{n_{i.}n_{.j}}{n}\right)^{2}}{\frac{n_{i.}n_{.j}}{n}}$$

The Chi square test or test of goodness of fit is the statistics used to find out if there is a marked difference between the observed (O) or actual frequencies and the expected (E). If there is a marked difference between both the Chi square, i.e. X^2 test will yield a numerical value large enough to be interpreted as statistically significant.

The formula is for Chi-square is given as:

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

Where, O represents Observed frequency

E represents Expected frequency

X^2 = The calculated value of chi square

X^2 table means the table value of Chi-square.

We must point out that if X^2 calculated $\sum X^2$ table, then there is significant deviation. This means that if the calculated chi-square is equal to or greater than “table value” of chi-square, there is significant deviation.

Computation Process

Example 1

A psychological skill was introduced to one or two groups of people who were suffering from a complaint (INSOMNIA). The numbers cured in each group are given in the table below. Test if the psychological skill has helped in curing the complaint (INSOMIA).

To answer this question, you need to draw a 2 x 2 contingency table.

Step 1:

Group	Cured	Not Cured

I	19	6
II	11	14

Step 2: Find the row totals, column total and gland total.

Group	Cured	Not Cured	Total
I	19	6	25
II	11	14	25
	30	20	50

This is obtained through the following process:

Row total = 19 + 6 = 25

Row total = 11 + 14 = 25

Column total = 19 + 11 = 30

Column total = 6 + 14 = 20

Grand total = 30 + 20 = 50 or 25 + 25 = 50

Step 3. Work out the expected frequency (B) for each of the cell separately using this formula

$$h = \frac{\text{row total x column total}}{\text{grand total}}$$

	Cured	Not Cured	
Group I	C = 15^A	O = 6^B	25
Group II	O=11^C	O = 14^D	25
	30	20	50

For Cell A, B = $\frac{rxc}{\sum}$

For Cell E = $\frac{25\ x30}{50} = 15$

For Cell B, E = $\frac{25\ x20}{50} = 10$

For Cell C, E = $\frac{25\ x30}{50} = 15$

For Cell D, E = $\frac{25\ x30}{50} = 10$

Therefore the table will now read thus

	Cured	Not Cured	

Group I	O = 19
	E = 15
		O = 6
		E = 10
			25

Group II	O = 11
	E = 15
		O = 14
		E = 10
			25

	30	20	50

Step 4: Calculate the X^2 by working out the difference between O and E for each cell.

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

$$X^2 = \frac{(19 - 15)^2}{15} + \frac{(6 - 10)^2}{10} + \frac{(11 - 16)^2}{15} + \frac{(14 - 10)^2}{10}$$

X^2 = 1.06 + 1.6 + 1.06 + 1.6 = 5.32

X^2 = 5.32

The degree of freedom (df) for any statistic is the number of component in its calculation that are free to vary.

df = (R-1) (C-1)

= (2-1) (2-1)

df = 1

We observed from X^2 table the following X^2 at 5%(1) = 3.84.

Since 5.32 > 3.84, the null hypothesis that the number cured is independent of the psychological skills is related at the 5 percent level of significance. Thus we conclude that the psychological skills have helped in caring:

Plant (INSOMNIA)

Example 2

A random number of persons in two groups, were asked to test if they could tell the difference between two brands of butter, the results are given in the table below.

Sex	Responses	
	Could tell	Couldn’t tell
Male	5	12
Female	20	13

Step 1: Let us find the row totals column totals and grand total.

Sex	Responses		
	Could tell	Couldn’t tell	
Male	5(a)	12(b)	17
Female	20(c)	13(d)	33
	25	25	50

Step II: Work out the expected frequency (E) for each cell separately using this formula

$$E = \frac{\text{Row total x Column total}}{\text{Grand total}}$$

For Cell (a) E = $\frac{17 x 25}{50} = 8.5$

For Cell (b) E = $\frac{17 x 25}{50} = 8.5$

For Cell (c) E = $\frac{33 x 25}{50} = 16.5$

For Cell (d), E = $\frac{33 x 25}{50} = 16.5$

Sex	Responses	

	Could tell	Couldn’t tell	

Male	O = 5
	E = 8.5
		O = 12
		E = 8.5
			17

Female	O = 20
	E = 16.5
		O = 13
		E = 16.5
			33

	25	25	50

Step 2: Calculate the X^2 by working out the difference between (O) and (E) for each.

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

$$X^2 = \frac{(5-8.5)^2}{8.5} + \frac{(12-3.5)^2}{8.5} + \frac{(20-16)^2}{16.5} + \frac{(13-16.5)^2}{16.5}$$

$X^2 = 1.441 + 1.441 + 0.7424 + 0.7424$

$X^2 = 4.5668$. df = (2-1) (2-1) = 1

= 4.77 df = 1

We observed from the X^2 table the following X^2 at 5%(1) = 3.84

Since 4.37 > 3.84, the null hypothesis that the difference in the butter is independent of sex is rejected at the 5 percent level of significance.

Example 1

To determine whether the age of a driver age 21 years or older has any effect on the number of motor accidents he is involved in a survey which was conducted and the following information was obtained.

Test the hypothesis that the number of accidents is independent of the age of the driver at 5%.

Age of driver and number of accident

Number of accident	21-30	31-40	41-50	51-60	61-70	Total
0	148	221	186	120	72	747
1	44	30	21	36	20	151
2	19	13	10	4	3	49
More than 2	4	5	2	1	3	15
Total	215	269	219	161	98	962

The expected value for each cell is obtained

$$E_{11} = \frac{215 \text{ x } 747}{962} = 166.95$$

$$E_{12} = \frac{269 \text{ x } 747}{962} = 208.88$$

$$E_{13} = \frac{219 \text{ x } 747}{962} = 170.06$$

$$E_{14} = \frac{15 \times 98}{962} = 1.5$$

The corresponding table for expected value is given below

Number of accident	21-30	31-40	41-50	51-60	61-70	Total
0	167	208.9	170	125	76.1	747
1	33.7	42.2	34.4	25.3	15.4	151
2	11	13	10	4	3	49
More than 2	4	5	2	1	3	15
Total	215	269	219	161	98	962.1

The null and alternative hypothesis are stated as

Ho: The age of a driver age 21 years or older has no effect on the number of motor accidents he is involved in

H_1: The age actually has effect on the number of accidents

The chi-square statistics is used and computed as follows

$$X^2 = \sum_{i=1}^{r}\sum_{j=j}^{c}\frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

$$= \frac{(148-167)^2}{167} + \frac{(221-208.9)^2}{208.9} + \frac{(120-125)^2}{125} + \frac{(72-76.1)^2}{76.1} + \frac{(44-33.7)^2}{33.1} +$$

$$\frac{(30-42.2)^2}{42.2} + \frac{(21-34.4)^2}{34.4} + \frac{(36-25.3)^2}{25.3} + \frac{(20-15.4)^2}{15.4} + \frac{(19-11)^2}{11} + \frac{(13-13.7)^2}{13.7}$$

$$+\frac{(10-11.2)^2}{11.2}+\frac{(4-8.2)^2}{8.2}+\frac{(3-5)^2}{5}+\frac{(4-3.4)^2}{3.4}+\frac{(5-4.2)^2}{4.2}+\frac{(1-2.5)^2}{2.5}+\frac{(3-1.5)^2}{1.5}$$

= 21.6 + 0.0054 + 1.51 + 0.2 + 0.22 + 3.15 + 3.53 + 5.22 + 4.53 + 1.37 + 5.82 +

0.036 + 0.13 + 2.15 + 0.8 + 1.06 + 1.52 + 0.58 + 0.961 + 1.5

= 36. 4524

Decision Rule: We Reject Ho if $X^2 > X^2_{0.05\ (4-1)(5-1)}$

Decision: since $X^2 = 36.45 > X^2_{0.095..1,2} =$ 21.0 we reject Ho and conclude that the age of a driver aged 21 years or older has significant effect on the number of accident he is involved in. this conclusion simply means that the number of accidents a driver involved is depends on his age.

2 x 2 CONTIGENCY TABLE

In this case, we have two variables or categories of interest classified in only two rows and two columns as follow

Row (R)	Columns (c)		
	1	2	Total
1	N_{11}	N_{12}	N_1
2	N_{21}	N_{22}	$N_{2.}$
Total	n.1	n.2	N

Equivalently, the chi-square under this situation is computed as

$$X^2 \quad = \quad \sum_{i=1}^{r}\sum_{j=j}^{c}\frac{(n_{ij}-E_{ij})^2}{E_{ij}}$$

EXERCISE

The below shows the performance of students in Mathematics and Account

Mathematics	Recover	Did not Recover	total
60 – 79	15	20	35
40 – 59	10	12	22
20 – 39	5	9	14

Test the hypothesis that performance in English language is independent of performance in Mathematics at 5% level of significance

2. A survey to investigate the relationship between exposure to unprotected sex and HIV infection and the information below was obtained

	HIV infection	NO HIV
Unprotected sex	70	200
No Sex	15	20

Is HIV infection independent of exposure to sex? Take α = 1%

Exercises

1. Five hundred individuals are classified according to sex and whether or not they are short-sighted, the results are given in the table below. Test if short-sightedness is independent of sex.

	Male	Female
Not short-sighted	233	246
Short-sighted	7	14

2. The table below gives the percentages of two sample size 100 of Ghanaian and Nigerian in regard to their Mathematical abilities.

Nationality	Good	Fair	Bad
Ghanaian	32	38	30
Nigerian	53	28	10

Do these figures indicate a significant (i.e. 1 percent level) difference between nationalities

The computations for c can be arranged as shown in table below:

MODULE VI

ESTIMATION AND TEST OF HYPOTHESIS

INTRODUCTION

Hypothesis can be defined as a conjectural statement, a theoretical composition or an assertion about some characteristic of a population

HYPOTHESIS TESTING: This is the inductive process of drawing sample from the population in order to ascertain or test where the certain or statement about the population could be upheld or removed by the observation data.

NULL AND ALTERNATIVE HYPOTHESIS

This null hypothesis is denoted by Ho is a statement that claims the existence of no disagreement with the condition in the population or interest.

ALTERNATIVE HYPOTHESIS: The alternative hypothesis of any statement that contradicts the null hypothesis it is denoted by H_1 in the rejection of the null hypothesis leads to the acceptance of the alternative hypothesis.

TYPES 1 AND TYPES 2 ERROR: Types 1 error committed when a time hypothesis is rejected. That is rejected Ho when is should be accepted. Type 1 error is therefore the probability or the risk of rejecting a true null hypothesis.' It is denoted by alpha (*a*)

Hence α = P (Rejected Ho/Ho is true)

TYPE II ERROR: Is committed when Ho is accepted when it should be rejected. That is accepting a false Ho. Type II error then is the probability or the risk: of accepting a false Ho, It is denoted by beta (β)

Hence β = P (Accepting Ho/Ho is false)

These may be represented as

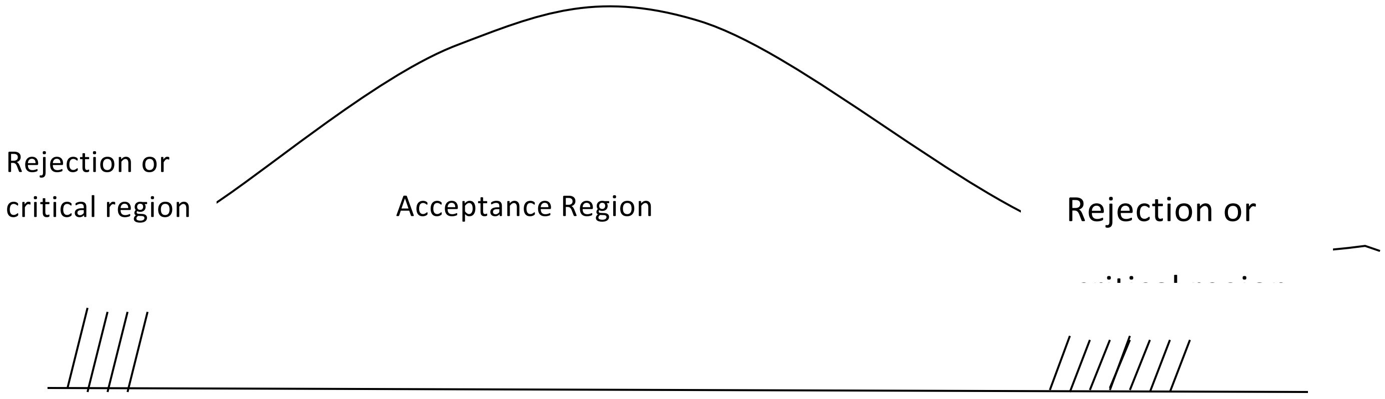
Decision

State of nature	Reject Ho	Accept Ho
True Ho	Type I error.	No error
False H0	No error	Type 11 error

In the hypothesis testing one is always faced with the risk of making one types of error or the other. Also in typing to reduce the risk of type I we automatically increase the risk of type II error and vice verse.

Level of Significant: This is the error margin fixed by the researchable prior to the research. That is, it is the pre-set probability of rejecting a true null hypothesis and it is denoted by alpha (α). Simply put the significant level as the probability of type I error. It is conventional to fix this at 5% 1% or 0.1%.

Critical or Rejection Region: the tail ends of the normal distribution curve in which the extreme value arc found and referred to as the critical region or the rejected region,



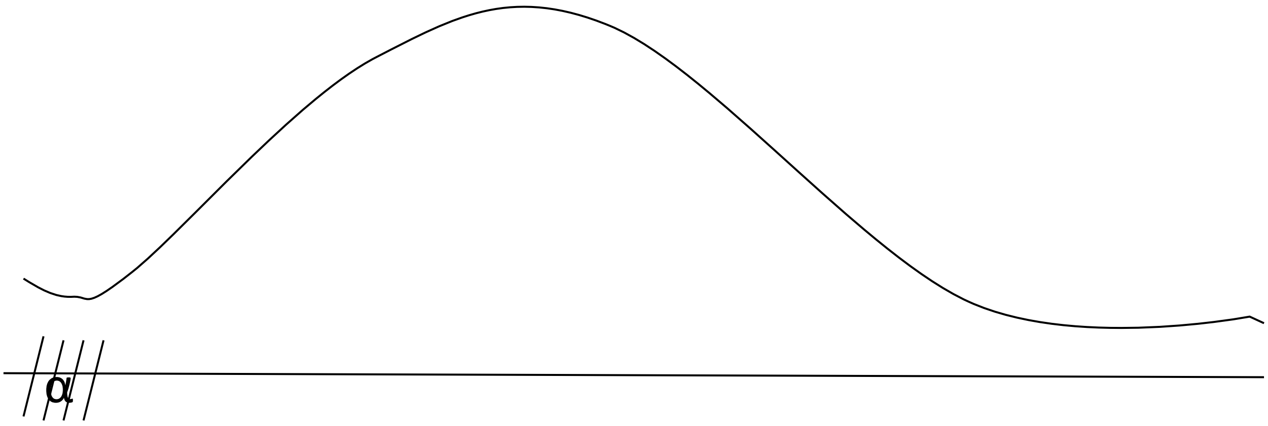
One tailed and two tailed test: the critical region may be located on both tails of the normal curve .as above or it may be concentrated on one tail of the distribution, the level of the significant equals the total area enclosed within the critical regions distribution on both tails of the distribution is referred to as two tailed or two-way test, while test in which the critical region is located only on one tail of the distribution is called one tailed or one-way test.

One-way test is used when the alternative hypothesis clearly specifies the direction in which the variation will occur.

Example

H_1: $\mu < a$

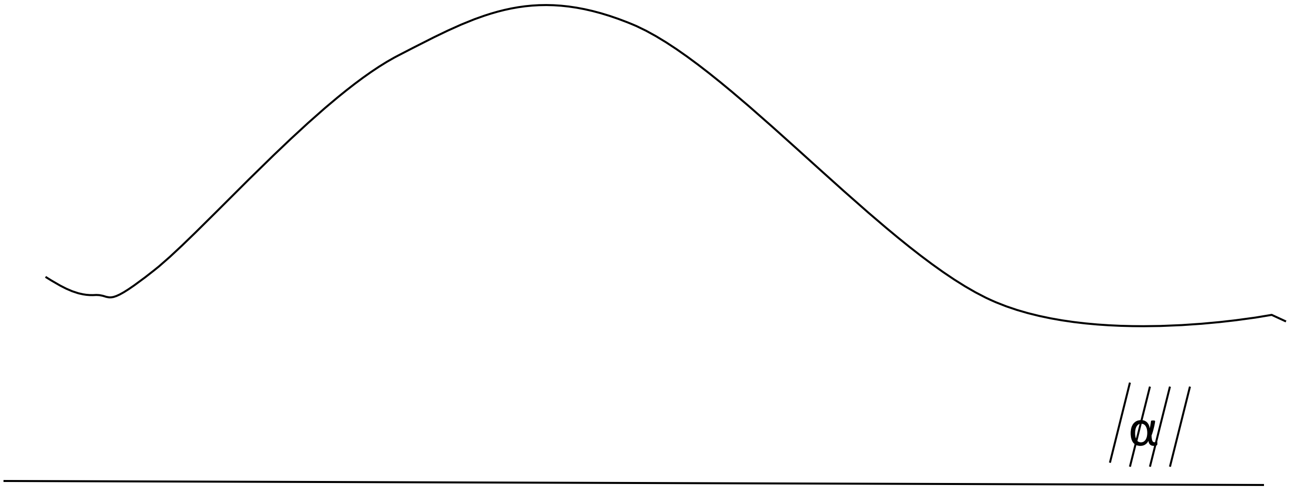
Hence the critical region is located at the lower tail end distribution as



Also if H_1 is

H_1: μ > b

Then, the critical region will be located at the upper tail end of the distribution as



Note that the size of the shaded portion is determined by the significance level (α).

Two tail tests is used when alternative hypothesis does not specify the direction of the variation.

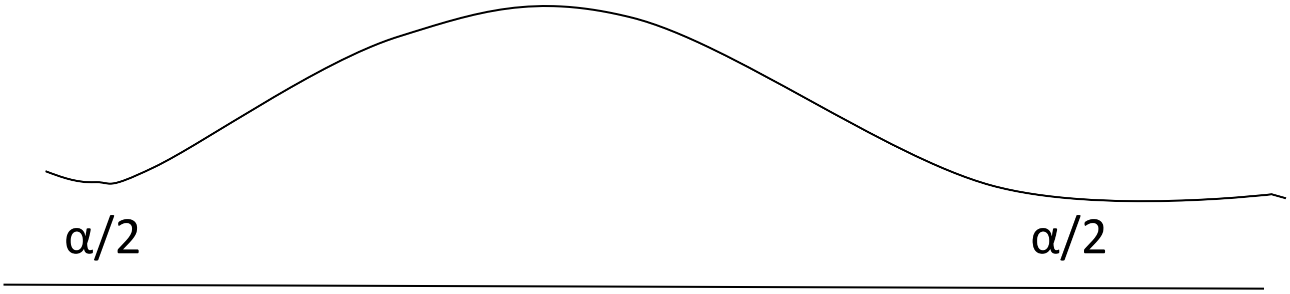
Example

If H_1 is give as

H_1: μ ≠ a

The above means that

μ < a or μ > a



Exercise

I. Differentiate between the following

(i) Null and alternative hypothesis

(ii) Type I and II error

(iii) Acceptance and rejection region

(iv) One tailed and two tail test

MODULE VII

STEP IN HYPOTHESIS TESTING

1. Formulate the null and alternative hypothesis
2. Specify the level of significant to be used
3. Select the test statistics
4. Establish the critical value taking into consideration if it one or two tail test
5. Set the decision rule
6. Determine the value of the test statistic
7. Making the decision

TEST INVOLVING ONE POPULATION MEAN

We may wish to test the null hypothesis

Ho: $\mu = a$

Against any of the alternative

(i) H_1: $\mu < \alpha$ (ii) H_1: $\mu > a$ (iii) H_1: $\mu \neq a$

The choice of the best statistic depends on

- Whether the sample is dream from a normal distribution population with known variance
- Whether the samples are drawn from a non – normally distributed population but the sample size is significantly large.

Case 1

When the population means and variance is known and the population is normally distribution, we use

$$Z = \frac{x- \mu}{\sigma}$$

$$Z = \frac{x- \mu}{\sigma/\sqrt{n}} = \frac{(x-\mu)\sqrt{n}}{\sigma}$$

Case II

When the sample is drawn from a non=normally distributed population and the sample size is large enough (i.e greater or equal to 30) we also employ the standard normal test (z-test) as in case I

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

Where

$$\sqrt{\Sigma\frac{(x-x)^2}{n-1}}$$

Is distributed standard normal we use Z – test as in case I

Case III

When the sample are drawn from a normally distribution with unknown variable and the sample size is small (i.e less than 30), we

Where

$$S = \sqrt{\Sigma \frac{(x\text{-}x)^2}{n\text{-}1}}$$

Decision Rule

Given α level of significance, the decision rule or the rejection point condition base on the alternative hypothesis is

Alternative Hypothesis

$H\neg_1: \mu > a$	$Z > Z\alpha\ (Z_1 - \alpha)$
$H\neg_1: \mu < a$	$Z < -Z\alpha$ or $Z <-Z_1 - \alpha$
$H\neg_1: \mu \neq a$	$\lvert Z \rvert > Z_{\alpha/2}$ or $\lvert Z \rvert > Z_{1-}{}^{\alpha}\!/_{2}$

For a one tail test

Accept Ho if $\lvert Z \rvert < Z_{1-\alpha}$ or $\lvert Z \rvert < Z_\alpha$

Rejected Ho if $\lvert Z \rvert > Z_{1-\alpha}$ or $\lvert Z \rvert > Z_\alpha$

For a two tail test

Accept Ho if $\lvert Z \rvert < Z_{1-}{}^{\alpha}\!/_{2}$ or $\lvert Z \rvert < Z^{\alpha}\!/_{2}$

Rejected Ho if |Z| > $Z_{1-}{}^{\alpha}\!/_{2}$ or |Z| > $Z^{\alpha}\!/_{2}$

Note: when sampling is done with replacement from a finite population or when sample is selected from an infinite population then the test statistic is

$$Z = \frac{x- \mu}{\sigma/\sqrt{n}}$$

When sampling is done without replacement from finite population then

$$Z = \frac{x- \mu}{\sqrt{\frac{\sigma^2}{n}\left(\frac{N-n}{n-1}\right)}}$$

When $\left(\frac{N-n}{n-1}\right)$ is called finite population correction factor

Example

It is known from past experiment that the mean breaking strength of a species of rock is 1800N and variance 1000N in a geological survey, a random sample of 50 pieces of the rock was tested and a mean strength of 1850N was obtained.

Test the hypothesis that the population means strength of the rock is 1800N at 5% level of significance.

Solution

Ho:μ = 1800N

H_1: μ > 1800N

$$Z = \frac{x- \mu}{\sigma/\sqrt{n}} = \frac{1850-1800}{\sqrt{\frac{1000}{50}}}$$

$= 11.18$

Decision Rule:

Rejected Ho if $|Z| > Z_{1-\alpha}$ otherwise do not rejected Ho

$Z_{1-0.05} = Z_{0.95} = 1.645$

Decision

We rejected Ho and concluded that the mean strength of the rock is greater than 1800N

Example

Using the same example

Test Ho: $\mu = 1800N$

Test H_1: $\mu = 1800N$

At 5% level of significant

$$Z = \frac{x- \mu}{\sigma/\sqrt{n}} = \frac{1850-1800}{\sqrt{\frac{100}{50}}}$$

$= 11.18$

Decision rule

We rejected Ho and concluded that the population breaking strength of the rock is not equal to 1800N

Example

A population of 100 items has means 200 and variance 49. A random sample of 25 items selected without replacement from this population gave a means of 185. At 1% level of significant has the population mean of the item decrease

Solution

Ho: $\mu = 200$

H_1: $\mu < 200$

$$Z = \frac{x - \mu}{\sqrt{\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)}} = \frac{185-200}{\sqrt{\frac{49}{25}\left(\frac{100-25}{100-1}\right)}}$$

$$\frac{-15}{\sqrt{1.4848}} = \frac{-15}{1.2185}$$

$$= 12.31$$

Example

The mean of a normal population is 17.0 and a sample of size 25 drawn from the population gave a sample mean 18.1 and a sample variance 16. Verify whether the population means is greater than 17.0 at 5%

Solution

Ho: $\mu = 17.0$

H_1: $\mu > 17.0$

Since σ^2 is not known and $n < 30$, we use the test statistic t

$$t = \frac{x- \mu}{\sigma/\sqrt{n}} = \frac{17\text{-}18.1}{\sqrt{\frac{16}{25}}} = 1.375$$

Decision Rule

Rejected Ho: if $t > t_{n=1}\ \alpha$

$t_{n\text{-}1}, \alpha = t_{24,\ 0.05} = 1.711$

Decision

Since 1.375 < 1.711 we accept Ho and conclude that mean of the population is 17

MODULE VIII

TEST INVOLVING POPULATION PROPORTIONS

In the most experiment in which there is only two possible and mutually exclusive outcome test the proportion may be required. Given the population proportion π_o and the sample proportion P may wish to test the hypothesis that

Ho: $\pi_o = \pi$

H_1: $\pi_o \neq \pi$

We employ the z-test if n > 30

$$Z = \frac{P- \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$$

If sampling was done with replacement or that sample was drawn from an infinite population, if however, sampling was done without replacement from a finite population.

Then

$$Z = \frac{P- \pi}{\sqrt{\frac{\pi(1-\pi)}{n}\left(\frac{N-n}{N-1}\right)}}$$

Or

$$Ƶ = \frac{P- \pi}{\sqrt{\pi(1-\pi)}}$$

Example

The manufacturer of a potent medicine claimed that the medicine was 90% effective. In sample of 200 people the medicine provided relief for 160 persons. Determine if the claim is legitimate at 1% level of significance use two tail test.

Solution:

Ho: $\pi_o = 0.90$

H_1: $\pi_o \neq 0.90$

$$P = \frac{160}{200} = 0.8$$

$$Z = \frac{P - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} = \frac{0.8-0.90}{\sqrt{\frac{0.9(1-0.9)}{200}}}$$

$$= \frac{-0.10}{\sqrt{0.00045}} = \frac{-0.10}{0.0212} = -4.717$$

MODULE IX

TEST THE DIFFERENT BETWEEN TWO MEANS

The standard normal score z is also used here and two independent sample are involving

The test statistic

$$-Z = \frac{x_1 - x_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n}}}$$

Note that n_1 and n_2 must be greater than 30. Also under the null hypothesis, the two populations means are assumed to be equal. Hence $\mu_1 = \mu_2$ and $\mu_1 - \mu_2 = 0$ and the test statistic becomes.

$$Z \quad = \quad \frac{x_1 - x_2}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n}}}$$

Note that we can replace r_2 in the above formula with s_2 = sample variance provided n is large from the above, it is not only assumed under Ho that the population means are equal but that the two sample value are obtained from the same population. This means that also the variance are equal ($\sigma_1^2 = \sigma_2^2$)

To obtain the common variance, we pool the two variance as

$$\sigma_p^2 \quad = \quad \frac{(n-1)\sigma_1^2 + (n_2 - 1)\sigma_2^2}{n_1 + n_2 - 2}$$

Then

$$Z \quad = \quad \frac{x_1 - x_2}{\sigma_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Example 5

The mean weekly wages for a sample of 30 employees in company A is ₦14.00. Also in company B, a random sample of 40 works has a means wages of ₦ 270 with a sample standard deviation of ₦ 10. Assuming the population variances are equal, test the hypothesis of no difference between the mean weekly wages of the two companies at 5% level of significance.

Solution

Ho: $\mu_1 = \mu_2$

H_1: $\mu_1 \neq \mu_2$

$$Z \quad = \quad \frac{x_1 - x_2}{\sqrt{\frac{S_1}{n_1} + \frac{S_2^2}{n_2}}} \quad = \quad \frac{280-270}{\sqrt{\frac{142}{30} + \frac{102}{30}}}$$

$$= \quad \frac{10}{\sqrt{6.53 + 2.5}} \quad = \quad \frac{10}{3.005}$$

$$Z_{1-\alpha/2} = \quad Z_{0.975} \quad = \quad 1.96$$

Decision Rule:

Reject Ho if Z > $Z_{1-\alpha}$ otherwise accept H_0

Decision: since z computed is greater than *z* tabulate, we reject the null hypothesis and conclude that there is significant difference in the mean wages of the two companies.

MODULE X

TESTING THE DIFFERENCE BETWEEN TWO PROPORTION

Given two proportion from normally distributed populations, the test statistic for testing the significance of the difference between them is given as.

$$Z = \frac{\rho_1 - \rho_2(\pi_1 - \pi_2)}{\sqrt{\frac{\rho_1 q_1}{n_1} + \frac{\rho_2 q_2}{n_2}}}$$

However, under the null hypothesis $\pi_1 = \pi_2$ and as such

$$Z = \frac{\rho_1 - \rho_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$\text{Where } \rho = \frac{(n_1 - 1)\,\rho_1 + (n_2 - 1)\,\rho_2}{n_1 + n_2 - 1}$$

and

$$q = 1 - \rho$$

Example 6

A sample of 300 voters from zone A and 200 voters from zone B showed that 56% and 48% respectively were in favor of a given candidate. At a level of significance of 0.05, test the hypothesis that there is difference between the districts

Solution

$H_0 : p_1 = p_2$

$H_1 : p_1 \neq p_2$

$$Z = \frac{\rho_1 - \rho_2}{\sqrt{pq\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

$$\rho = \frac{(n_1 - 1)\,\rho_1 + (n_2 - 1)\,\rho_2}{n_1 + n_2 - 2}$$

$$= \frac{(300-1)(0.56)+(200-1)(0.48)}{300+200-2}$$

$$= \frac{167.44 + 95.52}{498}$$

$$= 0.528$$

Hence,

$$Z = \frac{0.56-0.48}{\sqrt{(0.528)(0.472)\left(\frac{1}{300}+\frac{1}{200}\right)}}$$

$$= \frac{0.08}{0.6456}$$

$$= 1.75$$

$$Z_{1-\alpha/2} = Z_{0.975} = 1.96$$

Decision rule: We reject H_0 if Z is greater the $Z_{1-}{}^{\alpha}\!/_{2}$, otherwise accept H_o

Decision: Since Z = 1.75 is less than Z_{tab} = 1.96 we cannot reject $H_{0,}$ we therefore conclude that there is no significant difference between the preference of the voters.

Exercise:

1. Outline the steps in hypothesis testing
2. A test of the tensile strength of 6 flat metal sheets manufactured by a rolling mill showed a mean breaking strength of 7750N and a standard deviation of 145N, whereas the manufacturer claimed a mean tensile strength of 8000N, as a building contractor, can you support the manufacturer's claim at 0.01 level of significance?
3. The mean breaking strength of a cord manufactured by a company is 300N and standard deviation 24N. it believed that by a newly developed process the mean breaking strength can be increased and a random sample of 64 new cords was tested giving a sample mean of 305N. test the hypothesis 40N = 300N us Hi : N = B10N. calculate:

(i) Type I error

(ii) Type II error

2. Outlines the steps in hypothesis testing