**Inferential Statistics (2)**

**Contents**

**1.0 Introduction 1**

1.1 The Objective 1

 1.2 Inferential Statistics 1

**2.0 Definition of a Hypothesis 2**

 2.1 Null and Alternative Hypotheses 2

 2.2 Test of Hypothesis 2

**3.0 How to choose an appropriate statistical technique 3**

 3.1 Important test statistics and their applications 4

**4. 0 Analysis of Data 6**

 4.1 Univariate Analysis (a) z-test (b) t-test 6

 4.2 Bivariate Analysis: (a) Chi-Square (b) One Way ANOVA 10

 4.3 Limitations of Hypothesis Testing 11

**5.0 Measures of Association 12**

 5.1 Correlation Research 12

 5.2 Multiple Regression 13

 5.3 Important Tests of Significance & their Importance 20

 5.4 How do you evaluate a correlational study? 22

**6. 0 Qualitative Analysis 23**

 6.1 A Case Study 24

 6.2 Case Study and Triangulation 28

**7.0 Discussion of Results and Suggestions 29**

 7.1 How do you discuss the results? 29

**8.0 References** **30**

**1.0 INTRODUCTION**

After the collection of data from the participants, the next step is to analyse the data in accordance with the requirements of whether the research approach is quantitative or qualitative. For quantitative research, the purpose is to prove a theory which is defined in the research topic. On the contrary, the purpose of a qualitative research is to develop a theory from the data gathered. Clearly, the type of data required for quantitative research and the qualitative research are different based on the intention of the research study. This conceptual difference between the two research approaches has often been confused among the students when reviewing the literature and in the writing of the research methodology.

Therefore, it is necessary to be clear of the research approach before collecting the data in primary research. It is to ensure that a quality standard on the data is maintained. Whatever methods used in the collection of data such as observation, interview or questionnaire, are applied to secure a quality standard on the data. The data are coded and prepared for processing before they are analysed in accordance with the requirements for quantitative approach or qualitative approach. It is imperative that the appropriate data are gathered for the purpose of analysis and findings are in alignment with the research topic.

**1.1 The Objective**

The purpose of this writing is to guide the students in the analysis of the collected data to reach the conclusion of the findings. Often students overlooked the purpose of their research as stipulated in their research topics.

It has been encountered that a quantitative approach from the beginning turned out being presented in qualitative form in the process of interpreting the quantitative data; whether it is a mixed method is not clear. There are situations, in which data are not presented in the form of verbatim in qualitative approach. There is apparent confusion in the analysis of data.

This chapter seeks to provide the fundamentals in the process of analysis of the data and the findings in line with the purpose of the research. It provides the knowledge and understanding of some common statistical approaches for the analysis of quantitative data and the description statistics, both of which are referred to as ‘inferential statistics’. The inferential statistics is to enable the researcher to prove a theory that he has established in the research topic. Qualitative approach, however, is different as it is relating to a central phenomenon and data are presented in the form of verbatim from which categories are established to develop a theory.

**1.2 Inferential Statistics**

The purpose of a quantitative research is to prove a theory. A theory is about the relationship between the variables i.e., the independent and dependence variables. This assumption of a relationship is known as a hypothesis. The concept to test a theory is then based on the test of the hypothesis and it is known as “hypothesis testing”.

It is not possible to test a population but a sample is taken from it and the sample data is used to infer the results to the population. The purpose of the hypothesis testing is to determine whether the results of a study indicate a real relationship between variables, or if the results show that the difference is due to natural change or sampling error.

**2.0 DEFINITION OF A HYPOTHESIS**

A hypothesis is an educated guess that a researcher makes based on information available to him. He would have obtained the information from his own experience or literature review.

The hypothesis so developed will be tested using an appropriate statistical analysis procedure to determine if it can be accepted or rejected.

**2.1 The Null and Alternative Hypotheses**

The traditional practice to test the significance of hypotheses, is the use of two kinds of hypotheses. The first one is called the null hypothesis (H0) where a statement is made that no differences exist between the value of a parameter (population) and the value of statistics (a sample drawn from the population).

The second hypothesis is called the alternative hypothesis (HA) that there is a difference between the value of a parameter (population) and the value of statistics (a sample drawn from the population). The alternative can be stated in a non-directional form e.g., there is a change (increased or decreased), or in a directional form greater (right tail) or lesser (left trail).

**2.2 Test of hypotheses**

To test a hypothesis, it is assumed that there is a relationship between the independent variable and dependent variable and this is termed the alternative hypothesis, HA. This HA is challenged by the null hypothesis, Ho, which stipulates that there is no such relationship. In the test it is either the null hypothesis or the alternative hypothesis being accepted.

However, in the testing of hypothesis two types of errors can occur i.e. (1) reject the null hypothesis when it is true (Type 1 error) and (2) accept the null hypothesis when it is false (Type II error).

The probability of rejecting the null hypothesis when it is true is often termed the significance level.

If the null hypothesis is rejected at the significance level of 0.05 it *implies that the null hypothesis is rejected in 5% of the cases when it is correct*. This leaves 95% of the time the null hypothesis is rejected when it is false (not true).

 **Hypothesis Testing Procedure**

 State H0 and HA

 Choose the relevant statistical test

 Choose the desired level of significance,

 determine the degree of freedom

 Compute the test value

 Obtain the critical value

Is the calculated No

test value > Do not reject H0

 critical value?

 Yes

 Reject H0

 Formula for:

 1. Variance (S2) = ⅀ (X – X)2/n

 2. Standard deviation (S) = ⅀ (X – X)2 /n

 3. Standard error (SE) = S n- 1

**3.0 How to choose an appropriate statistical technique**

It is based on 4 criteria:

1. The type of question to be answered

 (a) if the researcher is interested in the mean value of a sample he can choose the one sample t- test.

 (b) If he is interested to compare whether there is a difference in the mean score of male students versus female students, then he can use the independent sample t-test.

 NB: It is important that the researcher choose the technique to be used i.e., quantitative or qualitative before choosing the research design, which determines what type of data to be collected (to answer the research questions).

2. Number of variables

 The number of variables involved in the analysis determines the statistical technique to

 be used.

1. Univariate (one variable) only dealing with the mean of total scores of students.
2. Bivariate (2 variables) e.g., the scores of male and female marks & gender.
3. Multivariate (more than 2 variables) e.g., dealing with several independent

 variables.

3. Scale of measurement

 This refers to the 4 types of scale of measurement to collect data:

1. Nominal data
2. Ordinal data
3. Interval data
4. Ratio data

The type of scale of measurement determines the type of statistical analysis. For

 hypothesis testing about mean values, the measurement must be the interval or ratio

 level. If it is to know whether the student fails or passed the exam, then the nominal

 scale data is used.

4. Parametric versus non-parametric hypothesis testing

 Parametric tests involve interval or ratio scale of measurement.

 Nonparametric tests involve nominal or ordinal scale of measurement.

 Parametric tests are more powerful than the non-parametric tests. However, the use of

 parametric tests requires certain assumptions to be met. For example, in ANOVA,

 three assumptions must be satisfied:

 (i) Observations are independent

 (ii) The sample data have a normal distribution.

 (iii) Scores in different groups have homogenous variance.

 Similarly, a regression model assumes the absence of collinearity, autocorrelation, random residuals, linearity, etc.

 If the research does not know how the data is distributed, the non-parametric tests should be used. Under such a situation, it is difficult to ensure the population distribution to be normal sampling distribution. The non-parametric tests are also known as “distribution free tests”.

**3.1 Important Test Statistics and their Applications**

 Four important parametric tests:

1. z-test

2. t-test

3. X2 test

4. F-test

**1. z-test**

It is based on the normal probability distribution and is generally used for:

1. comparing the mean of a sample to some hypothesized mean for the population in case of large sample or when population variance is known.
2. judging the significance of difference between means of independent samples in case of large samples or when the population variance is known.
3. comparing the sample population to a theoretical value of population proportions or judging the difference in proportions of two independent samples when n happens to be large.

The relevant test statistic, z is worked out and compared with its probable value from the z table at a specified level of significance.

The standard score is determined by the formula:

 z = X – x SD

**2. t-test**

This test is used when the sample size is small (n<30) and the population variance is not known. In this case, the variance of the sample (standard error) is used. It is an appropriate test for judging the significance of a sample mean or for judging the significance of difference between the means of two samples. It is also used to judge the significance of coefficients of simple and partial correlations.

The relevant test statistic ‘t’ is calculated from the sample data and then compared with its probable values on t-distribution table at specified level of significance for concerning degree of freedom for accepting or rejecting the null hypothesis.

 t = X - µ SE

 SE = SD √n

 E.g. A supermarket chain is to add a new product, at least 100 units per week to be sold. The new product is tested in 10 randomly selected stores for a limited time. The product will be introduced at a national scale if 100 units are sold per store.

The null and alternative hypotheses are:

 Ho: X ≤ 100

 HA: X > 100

Assume at significance level = 0.05.

The mean = 109.4 and Standard deviation = 14.90.

Standard error = SD √n = 14.90 √10 – 1 = 4.55

Calculating t = X - µ SE = 109.4 – 100 4.55 = 2.07

Critical value from the t-distribution table = 1.833 at significance level 0.05. The calculated value is higher than the critical value, the Null hypothesis is rejected and the relationship between the variables influenced the dependent variable.

**3. X2-test**

It is based on Chi-square distribution and as a parametric test is used for comparing a sample variance to a theoretical population variance.

**4. F-test**

It is based on F-distribution and is used to compare the variance of the two independent samples. It is also used in ANOVA for testing the significance of more than two sample means at one and the same time. It is also used for judging the significance of multiple correlation coefficients.

Test statistics F is calculated and compared with its probable value (from the F-ratio tables for different degrees of freedom for greater and smaller variances at specified level of significance, for accepting or rejecting the null hypothesis.

**4.0 Analysis of Data**

**4.1. Univariate Analysis**

This analysis is used when there is a single sample and we wish to test the hypothesis that comes from a specified population. Is there a difference between observed frequencies and expected frequencies? Or is there a difference between observed and expected proportions?

**Testing difference between sample and population means**

**(a) z test**

When the sample size is large (n >30), the test statistic z is calculated as follows:

 X - µ

 z =

 α/√n

 Where,

 X is the sample mean

 µ is the population mean

 α is the population standard deviation and

 n is the size of the sample.

If α is unknown, and sample size is large (n >30), the sample standard deviation S is used.

 S = ⅀(X – X)2

 \ n – 1

**(b) t-test**

t-test is used if sample size is small (n < 30) and population variance is unknown. It is calculated as follows:

 X - µ

 t = with degree of freedom(d/f/) = n - 1

 S/√n

**Calculating confidence interval using the t-distribution**

The formula to find the confidence interval is as follows:

 µ = X ± t(SE)

 Upper limit µ = X + t(S/√n)

 Lower limit µ = X – t(S/√n)

 Where,

 µ = population mean

 X = sample mean

 t = critical value of t at a specified confidence level

 SE = standard error of the mean

 S = sample standard deviation

 n = sample size

Table of data from the research study

|  |  |  |  |
| --- | --- | --- | --- |
| n | X | (X – X) | (X – X)2 |
| 1 | 65 | -3 | 9 |
| 2 | 75 | 7 | 40 |
| 3 | 84 | 16 | 256 |
| 4 | 62 | -6 | 36 |
| 5 | 53 | -15 | 225 |
| 6 | 74 | 6 | 36 |
| 7 | 48 | -20 | 400 |
| 8 | 94 | 26 | 676 |
| 9 | 59 | -9 | 81 |
| 10 | 66 | -2 | 4 |

 ⅀X = 680 ⅀(X – X)2 = 1772

 X = ⅀X/n = 680/10 = 68

 ⅀(X – X)2 1772

 S = = = 14.032

 \ n – 1 \ 9

 SE = S/√n = 14.032/ √10 = 14.032/3.1623 = 4.437

 The critical value of t for 9 df at 0.05 level of significance is 2.262.

Next calculate the upper and lower limit as follows:

 Upper limit µ = X + t(S/ √n) = 68 + 2.262(4.437) = 78

 Lower limit µ = X - t(S/√n) = 68 - 2.262(4.437) = 58

Based on the calculation above, *95% level of confidence that the population mean will be between 58 and 78 marks i.e., 58 ≤ µ ≤ 78.*

Univariate Hypothesis test using the t-distribution

To test whether the average mark 68 from the sample is significantly different from the population mean say 65.

**Six steps hypothesis testing procedure**

 1. State the hypothesis

 H0: µ = 65 marks

 HA:  µ ≠ 65 marks

2. Choose the statistical test

 X - µ

 t = with degree of freedom (d.f.) = (n – 1)

 S √n

3. Select the desired level of significance

 Let level of significance is set at 5%

 α = 0.05; d.f. = n – 1 = 10 – 1 = 9

4. Compare the calculated value

 68 – 65

 t = = 0.68

 14.032 √10

5. Obtain the critical value based on the test in step 2

 As HA is two-sided, we shall determine the rejection region applying two-tailed test at 5% level of significance for 9 df which comes to as under, using t-distribution with a critical value of 2.262

6. Draw the conclusion

 The observed value of t is 0.68 which is lower than the critical value (2.262) and thus we

 accept H0 and conclude that the sample mean is consistent with the population mean.

**4.2. Bivariate Analysis**

Bivariate analysis is where the analysis involves two variables at a time. It includes a group of techniques called the test of difference and another group called test of relationship.

In the test of ***differences***, the researcher is interested to see if the score of particular variable differs by a certain nominal variable e.g., gender or race. In the case of test of ***relationship***, the researcher wishes to know if there exist any association between the two variables of interest such as the marks obtained by students in BRM and total hours spent on self-study.

***a) Measure of differences***

* Chi-Square test for goodness of fit is used to find whether the two attributes are associated.

The procedures for the hypothesis testing pp. 223-226, Mukesh Kumar et al.

* t-test for comparing two means

This is used when we want to compare an interval level or ratio level data between two groups for some criterion or characteristics. When we want to know whether parameters of two populations are alike or difference E.g., whether the average marks scored in the BRM course differ between male and female students or if the average marks scored in BRM are different before and after the remedial classes.

Independent sample t-test

t-test is commonly used to test the differences between the means of two independent samples. The null hypothesis for testing of difference between means is generally stated as H0 = µ1 = µ2, where µ1 is the population mean of one population and µ2 is the population mean of the second population, assuming both the populations follow the normal distribution with an approximately equal variance.

The formula for the calculation of the t-statistics is as follows:

 X1 – X2

t  =

⅀(X1 - X1)2 +⅀(X2 - X2)2 x 1/n1 + 1/n2

 n1 + n2 - 2

An alternate equation is:

 X1 – X2

 t =

 (n1 – 1) S12 + (n2 – 1) S22 x 1/n1 + 1/n2

 n + n - 2

 with df = (n + n -2)

 Where,

 X1 = mean of group 1

 X2 = mean of group 2

 S21 = sample variance of group 1

 S22 = sample variance of group 2

 n1 = sample size of group 1 and

 n2 = sample size of group 2

 Assuming the data about the two groups are given as:

 Male Female

 marks X = 71 X = 65

 Variance S1 = 17.34 S2 = 10.95

 Sample size n1 = 5 n2 = 5

1. State the hypothesis

 H0 = µ1 = µ2

2. Choose the alternative formula

 X1 – X2

 t =

 (n1 – 1) S1 + (n2 – 1) S2 x 1/n1 + 1/n2

 n + n - 2

 71 - 65

 = = 0.65

 (5 – 1)(17.34)2 + (5 – 1) (10.95)2  1/5 + 1/5

 5 + 5 – 2

3. Obtain the critical value from the t-distribution table

 As HA is one-sided, we shall apply a one-tailed test at the right side of the distribution

 curve, at 5% level of significance at 8 df the critical value is 1.860. Since, the calculated

 value is less than the critical value, the null hypothesis is accepted; the average marks do

 not differ between male and female students.

**(b) More than Two Independent Variables and one Dependent Variable**

 **ONE-WAY** **ANOVA**

This is the most suitable analysis when we have more than 2 independent variables and a dependent variable which is in either interval or ratio scale.

 The technique of ANOVA involves the following steps:

 1. Say there are k samples. Calculate the mean of each sample: X1, X2, X3 ....., Xk

 2. Calculate the mean of the sample

 X1 + X2 + X3 + ........ + Xk

 X = ⅀Xk k =

 K

 3. Calculate the sum of squares for the variance between the sample (SS between)

 SSbetween = nk ⅀( Xk - X )2

 = n1 (X1 – X)2 + n2 (X2 – X)2 + n3(X3 – X)2 ....+ nk (Xk – X)2

 4. Calculate the variance of mean square by dividing SS between by degree of freedom

 between.

 MSbetween = SSbetween (k – 1) = ⅀nk ( Xk - X)2 (k – 1)

 where (k – 1) is the degree of freedom between samples.

 5. Calculate the sum of squares for the variance within the sample (SSwithin)

 SSwithin  = ⅀(Xk1 – Xk)2

 = ( Xki – X)2 + (X2 – X)2 + (X3 – X)2 + .... + (Xk – X)2

 6. Calculate the variance mean within the sample (MSwithin).

MSwithin = SSwithin (n – k)

 Note: The SStotal = SSbetween + SSwithin

df : (n – 1) (k – 1) (n – k)

 7. Calculate the F-value in an ANOVA

 F-ratio = MSbetween MSwithin

 8. Find the critical value of F at the selected level of significance for the degree of freedom and then determine whether the outcome is due to the influence of the attributes or sampling error or due to chance.

**Limitations of Hypothesis Testing**

1. The statistical tests are not decision making but the tests are only useful aids for decision-making.
2. The statistical test does not explain as to why the difference exists. It simply indicates whether the difference is due to chance or because of other reasons.
3. Results of the significance tests are based on probabilities and as such cannot be expressed with certainty. If the test shows a difference that is statistically significance, then it simply means that the difference is probably not due to chance.

Statistical inference based on significance tests cannot be said to be entirely correct evidence concerning the truth of the hypothesis particularly so in case of small sample size. It is necessary that this sample be a representative of the population.

(The t-test provides the finding whether there is a significant relationship between the variables. If there is a no difference between the calculated value of t and the critical value from the t distribution table, then the independent values have not influenced the dependent variable but being influenced by chance or sampling error. However, if there is a significant difference between the calculated value and the critical value then the difference is due to the influence of the independent variables on the dependent variable and this is not due to change or sampling error. Therefore, this significance test or hypothesis testing, gives an indication that there is an influence or no influence by the independent variables on the dependent variable.

NB:

* What has been found in the sample is applied to the population. This is because the null hypothesis assumed that there is no significant difference between the sample value and the population value (the critical value).
* The alternative hypothesis, on the contrary, assumed that there is a difference between the population value and the sample value. This occurs when the calculated value of the sample is higher than the critical value of the population (obtained from the t-table). The probability that here is a positive influence of the independent variables on the dependent variable, beyond the chance or sampling error.
* The significance level of 5% only tells us that there is 5% of the cases which are correct are being rejected.
* However, the degree or strength of the influence is not indicated i.e., there is not measurement of the extent of the influence on the dependent variable. This aspect of the weakness in the hypothesis testing is remedied by the use of correlation statistics.)

**5.0 Measures of Association**

The hypothesis testing can only provide an indication whether there is a relationship between the variables but did not measure the strength of the relationship. As a consequence, we have to refer to the measurement of association. Two of the most popular techniques for measuring the relationship (association) are (a) the correlation and (b) regression analysis.

**5.1 Correlation Research**

Correlation statistical test is used to describe and measure the degree of association (relationship) between 2 or more variables or sets of scores (as seen in Chi-squares). A correlation is a statistical test to find out the tendency or pattern for 2 or more variables or 2 sets of data to vary consistently.

In a set of 2 variables, changes in one variable causes the other variable to change i.e., they co-vary. It implies that a change in the independent variable can enable us to predict a change in the dependent variable. This kind of situation is illustrated in the linear relationship known as the *Pearson product-movement correlation* and it is indicated by an “r” for its notation.

 ∑(x- x) (y – y)

 r =

 ∑(x – x)2 ∑(y – y)2

The combined effect of the independent variables on the dependent variable is determined by:

 Explained variation

 r2 =

 Total Variation

The r2 is known as the *coefficient of determination* and is found by squaring the value of r from the correlation coefficient, r.

This implies that if the squared correlation coefficient between X and Y is r2 = 50%, then 50% of the variation in Y is accounted for by variation in X.

 Next we calculate the statistical t of Pearson’s correlation coefficient, r using the formula:

 r

 t =

 (1 – r2) (n – 2)

This calculated t value is then compared with the critical value in the t-distribution table based on the selected significance level of 0.05 and the degree of freedom.

If the calculated value is higher than the critical value, then a positive relationship exists between the 2 variables. However, if the calculated value is lower than the critical value, then we can conclude that there is no relationship between the 2 variables.

**Characteristics of Correlational Design**

1. There is a scatterplot of variables x and y.

2. A correlation Matrix – the correlation coefficient of each variable is presented in a table.

|  |  |
| --- | --- |
| Variables |  Correlation Coefficient |
| 1. Dependent variable2. Independent variable 13. Independent variable 24. Independent variable 35. Independent variable 4 | 1 | 2 | 3 | 4 |
|  - √ √ √ √ |  - √ √ √ |  √ √ |  √ |

3. Association between scores

 Understanding the direction of the association, the form of the distribution, the degree of

 association and its strength.

 Positive Linear Negative Linear No correlation

 x

 x xx xx x x

 x x x xxx x x

 xx x xxx x x

 xx xxx x x x

 Curvilinear Curvilinear

 x x

 x x x

 x x x x

 Figure 1: Types of association between two variables

4. The value of “r” can range from -1 to +1. A -1 means a perfect negative relationship exists while a +1 means a perfect positive relationship exists. When r is zero there is no correlation.

 Perfect negative Relationship = 0 Perfect positive

 relationship relationship

 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

 - 1 0 +1

 Figure 2 Continuum of correlation

**5.2 Multiple Regression**

There are two types:

**1. Partial Correlation**

This is related to 3,4 or 5 variables as predictors of outcomes. They include the *mediating variable (intervening variable)* that stands between the independent variable and dependent variable and influences between them.

 Independent variable Mediating variable Dependent variable

Partial correlation is used to determine the amount of variance brought about by the intervening variable. This effect can be removed by a partial correlation statistical analysis. The effect is it brings down the value of r2 than before this removal.

**2. Multiple Regression (Multiple Correlation)**

Correlation is used to predict scores. For example, if there are multiple variables on an outcome, what would be the future outcome. Then multiple regression analysis is used.

We start by understanding the regression line. It is a line of “best fit” for all of the points on the plot and it is calculated by drawing a line that minimizes the squared distance of the points from the line. (See Figure 1)

 y

 x

 x x x

 x x x x x

 x

 Figure 1: The Regression Line

The equation that expresses the regression line is:

 y = b(x) + c

 Where:

 y = predicted score

 x = actual score

 b = slope of the regression line

 c = the intercept/constant, the value of the predicted score, when x = 0.

Multiple regression is a statistical procedure for examining the combined relationship of multiple independent variables with a single dependent variable.

In regression, the variation in the dependent variable is explained by the variance of each independent variable, as well as the combined effect of all the independent variables designated by R2. If the squared correlation coefficient between X and Y is R2 = 50%, then 50% of the variation in Y is accounted for by variation in X.

The equation for multiple regression is:

 y (predicted) = b1(x1) + b2(x2) + c

 Where,

 y = the predicted score

 b1 = a constant for the slope of x1 (b2 for x2)

 c = the intercept

Suppose that:

1. Time-on-task
2. Motivation
3. prior achievement in the subject area
4. Peer friends

 are predicted to influence student learning.

 The result

 Sample regression table: Student Learning (Outcome variable)\_\_\_\_\_\_\_\_\_\_\_

 Predictable variables Beta (correlation coefficient)

1. Time-on-task .11
2. Motivation .18
3. prior achievement in the subject area .20
4. Peer friends .05

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

 R = .38

 R2 = .14

 P < 0.05

 N = 90

From the table :

* Prior achievement explains the most variance, followed by motivation & time-on-task.
* R represents the effect of the combination of variables of .38
* R2 explains that 14% of the total variation is due to the independent variables on the dependent variable. Other factors contribute (86%) to the learning of students.

A correlation coefficient shows correlation between two variables, and not that the variables are causally related. It gives two indications:

 (i) It tells the magnitude of the relationship; and

 (ii) It also points to an inverse relationship between the variables where one variable (x) increases, and the other variable (Y) decreases.

However, in correlation coefficient (r), both variables are not distinguished to be independent or dependent variables; they are related and are interdependent to each other. As a consequence, correlation coefficient cannot be applied to forecasting a variable with the help of another variable as seen in regression analysis.

An example of an analysis of number of hours worked in manufacturing industries with unemployment rate

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| No | UnemploymentRate (X) | No. of hoursWorked (Y) |  Xi - X | (Xi –X)2 |  Yi - Y | (Yi –Y)2 | (Xi – X) (Yi –Y) |
| 1 | 5.5 | 39.6 | .51 | .2601 | -.71 | .5041 | -.3621 |
| 2 | 4.4 | 40.7 | -.59 | .3481 | .39 | .1521 | -.2301 |
| 3 | 4.1 | 40.4 | -.89 | .7921 | .09 | .0081 | -.0801 |
| 4 | 4.3 | 39.8 | -.69 | .4761 | -.51 | .2601 | .3519 |
| 5 | 6.8 | 39.2 | 1.81 | 3.2761 | -1.11 | 1.2321 | -2.0091 |
| 6 | 5.5 | 40.3 | .51 | .2601 | -.01 | .0001 | -.0051 |
| 7 | 5.5 | 39.7 | .51 | .2601 | -.61 | .3721 | -.3111 |
| 8 | 6.7 | 39.8 | 1.71 | 2.9241 | -.51 | .2601 | -.8721 |
| 9 | 5.5 | 40.4 | .51 | .2601 | .09 | .0081 | .0459 |
| 10 | 5.7 | 40.5 | .71 | .5041 | .19 | .0361 | .1349 |
| 11 | 5.2 | 40.7 | .21 | .0441 | .39 | .1521 | .0819 |
| 12 | 4.5 | 41.2 | -.49 | .2401 | .89 | .7921 | -.4361 |
| 13 | 3.8 | 41.3 | -1.19 | 1.4161 | .99 | .9801 | -1.1781 |
| 14 | 3.8 | 40.6 | -1.19 | 1.4161 | .29 | .0841 | -.3451 |
| 15 | 3.6 | 40.7 | -1.39 | 1.9321 | .39 | .1521 | -.5421 |
| 16 | 3.5 | 40.6 | -1.49 | 2.2201 | .29 | .0841 | -.4321 |
| 17 | 4.9 | 39.8 | -.09 | .0081 | -.51 | .2601 | .0459 |
| 18 | 5.9 | 39.9 | .91 | .8281 | -.41 | .1681 | -.3731 |
| 19 | 5.6 | 40.6 | .61 | .3721 | .29 | .0841 | .1769 |

 X = 4.99

 Y = 40.31

 ∑ (Xi – X)2 = 17.8379

 ∑ (Yi - Y)2 = 5.5899

 ∑(Xi – X)(Yi - Y) = 6.3389

 ∑(Xi – X)(Yi – Y) 6.3389 6.3389

 r = = = = 0.635

 √∑(Xi – X)2 (Yi – Y)2 √ (17.8379) (5.5899) √ 99.712

The correlation value of 0.635 shows that the variables are related.

The t value:

 r 0.635

 t = = = 0.0235

 (1 – r2) (n – 2) (1- 0.6352) (19 – 2)

At the significance level of 5% and with 17 degrees of freedom the critical value is 1.74 (two-sided from t-distribution table). Since the calculated value of t is lower than the critical value, there is no relationship between the two variables.

The coefficient of determination is denoted by R2 which is (0.635)2 = 0.40. This indicates that 40% variance score can be explained by the variance of number of hours of work.

**For rank variable, the Spearman’s Rank Correlation** is applied.

 6⅀(X – Y)2

 rs = 1 -

 n(n2 – 1)

 and

 n – 2

 t = rs

√1-rs2

 The degree of freedom is n - 2, The computed t value is compared with the t-value in

 the table and a decision is made.

Estimation of Model Parameters

In a simple linear regression equation, the main interest is on the bo and b1. and for the purpose we calculatebo and b1.

In a study, the researcher has taken a random sample of 20 households from a town. For each household, the monthly income and expenditure data is collected through the survey. The table 1 below reports data on expenditure and income (both measured in hundreds) for the 20 households in the sample and the scatter plots of expenditure and income data along with the linear regression line are shown in Figure 2.

 Table 1: the Data

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  No | Expenditure (RM’00) Y | Income(RM’00)X |  No | Expenditure(RM’00)Y | Income(RM’00)X |
| 1 | 18 | 24 | 11 | 42 | 73 |
| 2 | 36 | 65 | 12 | 50 | 91 |
| 3 | 30 | 43 | 13 | 38 | 62 |
| 4 | 55 | 98 | 14 | 22 | 27 |
| 5 | 15 | 17 | 15 | 28 | 31 |
| 6 | 53 | 121 | 16 | 24 | 29 |
| 7 | 33 | 45 | 17 | 31 | 38 |
| 8 | 29 | 32 | 18 | 60 | 102 |
| 9 | 33 | 46 | 19 | 58 | 84 |
| 10 | 30 | 53 | 20 | 32 | 33 |

 Let the linear regression line to be estimated from the sample be:

 ŷ = bo +b1X1

Where,

 ŷ is the estimator of Y1

 bo is the estimator of β0

 b1 is the estimator of β1

 Thus the hypothesized (population) regression equation can be expressed as

 Y = ŷ + e

 Y = b0 + b1 + e

 Where, e is the residual and is an estimate of the error in the population (µ).

Estimate the linear regression line from the plots of dependent variable Y against independent variable X. as shown in Figure 2.

 70 \* \*

 \*

 60 \* \*

 \* \* \* \*

 50 \*

 \* \* \*

 40 \* \*

 \*

 30 \*

 \*

 20 \*

 \*

 10

 0

 0 20 40 60 80 100 120 140

 Figure 2: The Regression Line

The scatter points show the actual observation on monthly household expenditure (Y) for a given value of monthly household income (X). the linear regression line passes through mean of these scatter points such that some of the observation lie above the line and some other observations lie below the line. However, if the regression passes through the mean values of the scatter points, the sum of estimated positive errors is off set with sum of negative estimated errors, which in turn results in the sim of estimated error as zero, ⅀e = 0.

The method of ordinary least squares (OLS) provides us with unique estimates of β0 and β0 that provides the minimum possible value of ⅀ei2. The minimization of ⅀ei2 with respect to b0 and b1 results in two more equations.

 ⅀Yi  = nb0 + b1⅀Xi

 ⅀X Y = b0 ⅀Xi + b1 ⅀Xi2

 Where, n is the sample size.

By solving the two equations, we get:

 n⅀XY - ⅀X ⅀Y

 b1 =

 n ⅀X2 – (⅀X)2

 and,

 b0 = Y – b X

 Where, ⅀ Y ⅀ X

 Y = and X =

 n n

 b0 and b1 are the regression coefficients.

Next, we have to find the reliability of the estimates of b0 and b1. This is measured by the standard error.

 αu2

 (i) Var (b1) =

 ⅀ (X – X )2  where α is the standard deviation

 α

 SE = √ Var (b1) =

 √⅀ (X – X )2

 αu2 ⅀X2

 (ii) Var (b0) =

 n ⅀ (X – X )2

 ⅀X2

 SE(b0) = Var (b0) = x α

 √ n ⅀ (X – X2)

 Where αu2 is the constant variance of µ and is estimated as:

 ⅀ei2

 Se2 =

 n – 2

Where Se2 [called *Mean Square Error* (MSE)] is the estimator of true but unknown αu2 and n-2 is the degree of freedom. The square root of MSE, denoted by SEe is the ***standard error of estimate.***

 ⅀ei2 ⅀ (Yi - Ŷi)2 ⅀Yi2 - b0 ⅀Yi – b1 ⅀XiYi

 SEe = = =

 √  n- 2 √ n – 2 √ n – 2

The standard error of estimate means that for any given value of independent variable X, the dependent variable tends to be distributed about the predicted value Yi, with a *standard deviation equal to the standard error of estimate.*

Once SEe is calculated, the standard error of estimators b0 and b1 can be calculated.

 SEe SEe

 SEb1 = =

√⅀(X – X)2 √⅀X2 – n(X)2

⅀Xi 2 ⅀Xi2

SEb0  = x SEe = x SEe

 √ n⅀(X – X)2 √ n⅀Xi2 - (nX)2

Next do the t-test of the significance of independent variable

 b1 - β1

 t b1  =

 SEb1

Refer to the value in the t-table at 1% level of significance for (n – 2) degree of freedom to see whether the null hypothesis is rejected and alternative hypothesis accepted.

How good is the prediction?

Measure how well the regression model predicts the variation of total variation in term of the dependent variable on the independent variables. This is called the *coefficient of determination (R2)* as a ratio of ESS/TSS.

 ESS

 R2 =

 TSS

TSS = total sum square which measures the total variations of actual Y values about their (⅀Yi – Y)2

ESS = explained sum of squares which measures the variation of the estimated Y values about their mean (the mean of predicted values of Y is the same as the mean of actual values of Y).

 (⅀Ŷi – Y)2

RSS = unexplained or residual sum of squares which measures the variation of Y values about the regression line. (⅀Yi - Ŷi)2

The limits: 0 ≤ R2 ≤ 1

 ⅀(X – X)2 ⅀X2 – n(X)2

 R2 = b12 = b12

⅀(Y – Y)2 ⅀Y2 - n (Y)2

If the calculated value is .885, it means 88.5% of the dependent variable is determined by the independent variables.

**5.3 Important tests of significance and their interpretations in a multiple linear regression**

(a) Tests of R2

It involves testing the significance of the overall regression model as well as specific partial regression coefficients. The null and alternative hypotheses for the overall test are that coefficient of multiple determination in the population is zero.

 H0: R2pop =0

 H1: R2pop ≠ 0

This is equivalent to stating:

 H0: βj = 0; j = 1, 2, ...., k

 H1: at leastone of the βj is nonzero

The overall test can be conducted using an F-statistic:

 R2/k

 F =

 (1 – R2)/(n-k-1)

 with k and (n-k-1) degrees of freedom.

 Where k = number of independent variables and n = sample size.

(b) Test of Regression Coefficients

To determine which specific coefficients (bj) can be done in the same manner as the test for b1 in the simple linear regression model by using t-tests.

 bj - βj

 tbj =

 SEbj

with (n-k-1) degree of freedom, where k = number of independent variables and n = sample size.

(c) Tests of increments in the population of variance accounted for by additional variable

One might be interested to know if adding more explanatory (independent) variables in the regression model can help to explain variation in dependent variable. This can be inferred by assessing the difference in R2 between model 1 (with large number of independent variables) and model 2 (with small number of independent variables). Whether the difference between the R2 obtained from the regression model is statistically significant or not can be tested using F-test.

 R12 - R22  d1

 F = x

 1 – R12 d2 - d1

Where R12  and R22 are the total variance explained by model 1 and model 2 respectively and d1 and d2 are the degrees of freedom in model 1 and model 2 respectively.

(d) Evaluating the importance of independent variables

Which of the independent variables has the greatest influence on the dependent variable?

Need to consider 2approaches:

1. Look at the t-values for the partial regression coefficients (bj). The one with the largest t-value can be interpreted as the one that is least likely to have a zero value for the parameter.
2. Examine the size of the partial regression coefficients. The larger the size of coefficient, the greater is the influence of the corresponding independent variable on the dependent variable.

If independent variables are measured in different units, we can assess their strength of influence on dependent variables by looking at standardized regression coefficients, which can be obtained by multiplying the unstandardised regression coefficient (bj) by the ratio of the standard deviations of the corresponding independent variable to the dependent variable.

 Sxj

Standardized bj = bj

Sy

Standardized bj can be compared with each other irrespective of the units of measurement of independent variables. The larger the value of standardized bj, the stronger its impact on the dependent variable.

**5.4 How do you evaluate a correlational study? (Check list)**

1. Is the size of the sample adequate for hypothesis testing? (N > 30)
2. Does the researcher adequately display the results in matrices/graphs?
3. Is there an interpretation about the direction and magnitude of the association between 2 variables?
4. Is the researcher concerned about the form of the relationship so that an appropriate statistic is chosen for analysis?
5. Has the research identified the predictor and the criterion variables?
6. If a visual model of the relationship is advanced, does the researcher indicate the direction of the relationship among variables? Or the predicted direction based on observed data?
7. Are the statistical procedures clearly identified?

**6.0 QUALITATIVE ANAYSIS**

Data analysis is the process of bringing order, structure and meaning to the mass of collected data. Marshall and Rossman (1999) advocated that the act of analysing qualitative data should occur at the same time as the data is collected and continue to do so after the collection of data. This assists the direction for the collection of data and data analysis that follows.

Glaser (1992) defined qualitative analysis as any kind of analysis that produces findings or concepts and hypotheses that are not arrived at by statistical methods. The data for qualitative and quantitative approaches is not the same. However, similarities are found between qualitative and quantitation data analysis. For example, observations can be categorised and categories can be assigned numbers.

|  |  |  |
| --- | --- | --- |
| No | Qualitative Data | Quantitative Data |
| 1 | Based on meanings expressed through words. | Based on meaning derived from numbers. |
| 2 | Collection results in non-standardized data requiring classification into categories. | Collection results in numerical and standardized data. |
| 3 | Analysis conducted through the use of conceptualization. | Analysis conducted through the use of diagrams and statistics. |

The general activities involved in qualitative analysis are:

1. Categorisation

The first activity is the classification of the data into meaningful categories. These categories are in fact the codes or labels to arrange the data.

1. Unitising Data

The next activity is to put the relevant data into the appropriate category or categories. This is a selective process which has the effect of reducing and rearranging the data gathered into a more manageable and comprehensive form such as matrices, charts, graphs and networks. Using these analytical techniques may enable the researcher to recognise emergent patterns in the data gathered that will provide the researcher with an indication about how to further the data collection.

1. Recognising Relationships and Developing Categories

The researcher looks for themes and patterns or relationships in the data set and explanations for the research questions and objectives that form the focus of the research study.

1. Developing and Testing Hypotheses

From the patterns developed from the data gathered, the researcher comes out with relationships between the categories and from which hypotheses are developed. From the hypotheses, the researcher is able to draw valid and well-grounded conclusions.

Interpretation plays an important role in especially qualitative research because it is needed to arrive at understanding. Furthermore, ethical practices play an important role in qualitative research.

**Validity of Qualitative Research**

It is imperative that the analysis of a qualitative research must have ***descriptive validity*** in terms of the degree the actual description holds true. It further needs to have ***interpretive validity*** in ensuring how good is the interpretation of the data received or observed. In addition, there must be ***theoretical validity*** in the adequacy of suggested theory on explanation in terms of the amount of research done to support and reinforce the finding of the research theory. Thus, adequate data must be gathered to support these three forms of validity.

As with regard to generalizable validity, the findings of qualitative research may not apply to other settings by virtue of differences of the situation and environment.

**How do you report the results?**

When researchers conclude the statistical testing, they represent the results in tables and figures and reporting the results in a discussion.

* A table is a summary of quantitative data organised into rows and columns.
* One table is for one statistical test.
* Notes are added to qualify, explain or provide additional information in the tables.
* A figure is a summary of quantitative information presented as a c hart, graph or picture that shows relations among scores or variables.

Present Results

* The researcher needs to describe in detail the results of the statistical tests. He presents detailed information about the specific results of the descriptive and inferential statistical analyses. This process requires explaining the central results of each statistical rest and presenting this information using language acceptable to quantitative researchers.
* For each statistical test result, the researcher summaries the findings n one or two sentences. At a minimum, sufficient information necessary for reporting results of each test should have:
* Report whether the hypothesis test was significant or not;
* Provide important information about the statistical test, given the statistics.
* Include language typically used in reporting statistical results.

**6.1 An example of a qualitative research study: A CASE STUDY**

What is a case study?

A case study is often used in descriptive or exploratory research (Yin, 1994; Bonoma, 1985; Gauri, 1983). In business studies, case study is particularly useful when the phenomenon under investigation is difficult to study outside its natural setting and also when the concepts and variables under study are difficult to quantify. Often it is due to the fact that there are too many variables to be considered; which makes experiment or survey methods inappropriate to use (Bonoma, 1985, Yin, 1994).

Many phenomena cannot be understood if removed from their social context. In these cases qualitative approaches are alternative methods to scientific investigation.

A case study involves a qualitative approach. It is based on a process model. It is fundamentally a description of a situation. It involves data collection through multiple sources such as verbal reports, personal interview and observation as primary data sources.

In addition, case methods involve data collection through sources such as financial reports, archives and budget and operating statements including market and competition reports.

However, case study is not applicable to all types of research. It is the research problem and the objective that decide whether the case method is suitable or not. It is, in fact, useful for theory development and testing. The main feature is the intensity of the study of the object, individual, group, organisation, culture, incident or situation. We need to have sufficient information to characterize, to explain the unique features of the case, as well as to point out the characteristics that are common in several cases. Finally, this approach relies on integrative powers of research; the ability to study an object with many dimensions and then to draw an integrative interpretation (Selltiz et al, 1976).

When to use a case study?

* When “how” or “why” questions are to be answered.
* Where existing theory is inadequate; or complementary to incremental theory building/development
* For comparative study e.g., to study a number of organisations with respect to a set of variables (known as comparative case studies).

Preparing for a case study

It is suggested to involve 4 stages of development (Bonoma, q985):

1. Drift stage - learning the area of research; concepts and terminology in the field in order to get a feel of the phenomenon and how it operates.
2. Design stage – decide on the strategy to collect data to get answers to the research questions plus starting to develop explanations of the observation so far. Here, the researcher assesses and refines major areas of the research project as suggested by the drift stage.
3. Prediction stage – researcher has a good understanding of the factors on which case information may be grouped and can proceed with further case construction and analysis. Proceed on compiling more cases with the purpose of drawing conclusions and developing some tentative explanations. Any generalization needs to ensure it is valid to particular circumstances, settings, industries or firms.
4. Disconfirmation stage (final) – refers to further testing/analysis of the results suggested by the prediction stage

 **Disconfirmation Model/theory Creation \***

 **Prediction \***

 **Design \***

 **Drift \***

 Early Middle Late

 **A process model for case research**

**Prior to Case Study**

1. Clarify research questions such as “how”, “why” and “where” before deciding the case study design.
2. For exploratory study, no need to have propositions (hypotheses).
3. For descriptive or causal studies, need to have a theoretical base and clearly stated propositions.
4. For data analysis, there should be a link between data and propositions. This is pattern building.

If we can find a systematic or unsystematic pattern, we can accept or reject our propositions. Statistical tests are not necessary to establish a pattern. The pattern has to be sufficiently systematic to accept certain propositions.

How to select the cases

* It is important to decide the target population which is to be used for the investigation. It includes those firms, individuals, groups or elements that will be represented in the study.
* Next atage is to assess the accessible population to which we have access.
* Select one or a few cases, objects or firms for study.
* Time available, financial resources for travelling and other practical issues are import for the study or even the size of the firm selected for the study.
* If it is to study a specific and complex issue, select a bigger firm; because ut gas experienced complex problems and has the expertise in-house that can provide us with in-depth information on the particular issue.
* The cases should correspond to the theoretical framework and the variables we are studying.
* If we are studying behaviour of industrial buyers we have to select firms that are dealing with industrial marketing and purchasing. Once we have selected a firm, we should select a manager who is involved in the process of marketing and purchasing.
* In bigger firm, it is important to select the right department, section or individual. It is not the question of interview but the most important is interviewing the “right” person from an organisation: right from the point of view of our research questions and variables.

How many cases should we study?

There is no upper or lower limit. Often one is enough. Note: it is the research problems and the research objectives that influence the number and choice of cases to be studied.

Conducting a case study

* Need special skills and caution.
* Data collection is crucial in a case study – it is very demanding.
* The researcher has to collect the data personally.
* He must fully understand the research problem and purpose of the study.
* He needs to have the ability to ask relevant and probing questions as well as to listen and interpret the answers given
* Able to read between the lines and understand what is said but also what is meant.
* Not to let biasness influence the interpretation and this can be helped by using multiple data sources.

Types of case study design

There are 4 types of research design:

 1. Single case design, holistic for critical case and for testing a theory; unique

 2. Single case design embedded situation or for revelatory.

 3. Multiple case design, holistic

 4. Multiple case design, embedded for other studies that are not critical or revelatory

Generally, a single case design is used for deductive approach while a multiple case design is for inductive (qualitative) approach.

Analysing case studies

 Data

 Collection Data

 Display

 Data

 Reduction Conclusions:

 Drawing/verifying

With qualitative methods, the data collection and analysis are often done simultaneously and sometimes the research problem is even formulated /reformulated at the same time. This often leads to new questions and new data collection and there is no definitive phase of data analysis. This is shown in the diagram above.

To analyse data we often have to code them so that they can be broken down, conceptualized, put together and presented on an understandable manner.

For qualitative studies, there are theories to be developed. This involves the process of coding of data. The process requires care, rigour, creative and persistent.

Data collection and data analysis are closely interconnected, so that while collecting data, we are doing the data analysis, coding and the type of data to be collected. One way to analyse the data collected is to look for commonalities and differences as in multiple cases.

In fact, in the drawing of conclusions, the cases that display contrast or an extreme situation are most useful.

**6.2 Case Studies and Triangulation**

Triangulation refers to the combination of methodologies in the study of the same phenomenon. Through triangulation we can improve the accuracy of judgements and results, by collecting data from different methods on the same subject matter of our study. It enhances the validity of our findings.

In fact, in a case study, triangulation is practised because data are obtained for many sources of information such as observations, intervene and questionnaire. *The main advantage of triangulation is that it produces a more complete, holistic and contextual portrait of the object under study.*

It is quite useful to use qualitative methods in a pilot study to build hypotheses or propositions, and then to use quantitative methods to test these hypotheses. This can be called a *two-step study*. It leads us to a better understanding or to new questions that can be answered by later research.

Some possible problems with triangulation:

1. Sometime difficult to judge if the results from different methods are consisten or not.
2. Different methods come up with contradictory results.
3. A researcher prefers one method over another.

**7.0 Discussion of Results and Suggestions**

**7.1 How do you discuss the results?**

After reporting and explaining the detailed results, researchers conclude a study by summarizing key findings, developing explanations for results, suggesting limitations in the research and making recommendations for future inquiries.

* **Summarize the major results**
* A summary is a statement that reviews the major conclusions to each of the research questions or hypotheses. It represents general conclusions. General conclusions state overall whether the hypothesis was rejected or whether the research question was supported or not supported.
* Implications are those suggestions for the importance of the study for different audiences. They are reflections of the importance of the study stated in chapter 1, the Introduction.
* **Explain why the results occurred**
* Explain why their results turned out the way they did. Go back to the predictions made from a theory or conceptual framework that guided the development of research questions or hypotheses. It may include discussing the existing literature and indicating how the results either confirmed or disconfirmed prior studies.
* **Advance limitations**
* Limitations are potential weaknesses or problems with the study identified by the researcher. They an be about inadequate measures of variables, loss or lack of participants, small sample sizes, error in measurement and other factors typically related to data collection and analysis. These limitations are useful to others potential researchers. Limitations also help readers judge to what extent the findings can or cannot be generalised to other people and situations.
* **Suggest Future Research**
* Future research directions are suggestions made by the researcher about additional studies that need to be conducted based on the results of the present research.

**8.0 References**

1. Mukesh Kumar, Salim Abdul Talib & T. Ramayah (2017). Business Research Methods.

 Oxford Fajar Sdn Bhd, Shah Alam. pp. 240-261

2. Pervez Ghauri, Kjell Gronhaug & Roger Strange (2020), Research Methods in Business Studies – A Practical Guide,(5th Edition) Prentice Hall, UK.

3. Prahlad Mishra (2018). Business Research Methods, Oxford University Press, India, pp. 362-432.

4. L.R. Gay, G.E. Mills & P. Airasian (2020). Educational Research Competencies for Analysis and Applications (8th Edition), Pearson Prentice Hall, New Jersey, pp. 337- 375).

5. Pamelo S. Schindler (2019). Business Research Methods, 13th Edition, McGraw Hill, New York.

6. John W. Creswell (2020). Educational Research, Planning, Conducting and Evaluating Quantitative and Qualitative Research, 6th Edition, Pearson, Singapore.