CHAPTER III

STATISTICAL AND ECONOMICAL THEORY

3.1 Mann-Whitney Test

How the central locations of two population distributions could be compared when a random sample of matched pairs was available. In this chapter, it is introduced a test for the same problem when *independent random samples* are taken from the two populations.

Table 3.1 shows the numbers of hours per week students claim to spend studying for introduction finance and accounting courses. The data are from independent random samples of ten finance students and twelve accounting students.

$$\text{The distribution of } Z = \frac{T - u_T}{\sigma_T} \qquad (1)$$

When the null hypothesis that the distribution of the paired differences is centered on 0 is true and the decision rule for testing his hypothesis against the alternative that the center of this distribution is less than 0 at the significant level concerning.

3.2 Tests Concerning Means

To consider the problem of testing the hypothesis that the mean μ of a population, with known variance σ^2, equals a specified value μ_0 against the two-sided alternative that the mean is not equal to μ_0, that is test [2];

$$H_0: \mu = \mu_0,$$
$$H_1: \mu = \mu_0, \qquad (2)$$

An appropriate statistic that based on decision criterion is the random variable X. The sampling distribution of X is approximately normally distributed with mean $\mu_{\boxtimes} = \mu$ and variance σ^2 / n, where μ and σ^2 are the mean and variance of the population

from which is selected random samples of size n. By using a significance level of α, it is possible to find two critical values, X_1 and X_2, such that the interval $X_1 \leq X \leq X_2$ defines the acceptance region and the two tails of the distribution, $X < X_1$ and $X > X_2$, constitute the critical region.

The critical region can be given in terms of z value by means of the transformation [2];

$$Z = \frac{X - \mu}{\sigma / \sqrt{n}} \tag{3}$$

Hence, for the α level of significance, the critical values of the random variable Z, corresponding to X_1 and X_2, are shown in Figure 1 to be,

$$-Z_{\alpha/2} = \frac{X_1 - \mu}{\sigma / \sqrt{n}} \tag{4}$$

$$Z_{\alpha/2} = \frac{X_2 - \mu}{\sigma / \sqrt{n}} \tag{5}$$

From the population is selected by a random sample of size n and compute the sample mean X. If X falls in the acceptance region, $X_1 \leq X \leq X_2$, then

$$Z = \frac{X - \mu_0}{\sigma / \sqrt{n}} \tag{6}$$

is fell in the region $-Z_{\alpha/2} < Z < Z_{\alpha/2}$ and it can conclude that $\mu = \mu_0$; otherwise the result reject H_0 and accept the alternative hypothesis that $\mu = \mu_0$. The critical region is usually state in terms of Z rather than X.

The two-tailed test procedure just described is equivalent to finding a $(1-\alpha)$ 100% confidence interval for μ and accepting H_0 if μ_0 lies in the interval. If μ_0 lies outside the interval, it rejects H_0 in favor of the alternative hypothesis H_1. Consequently when one makes inferences about the mean μ from a population with known variance σ^2 whenever it be by the construction of a confidence interval or through the testing of the statistical hypothesis, the same value, $Z = (X - \mu) / (\sigma / \sqrt{n})$, is employed [2].

In general, if one uses an appropriate Z or t value to construct a confidence interval for a population mean $\mu = \mu_0$ or $\mu_1 - \mu_2 = d_0$ against an appropriate alternative. The assumption is made, relative to the use of a given statistic apply to the tests described here. This means that all samples are selected either from approximately normal populations or are of size $n \geq 30$, in which we can refer to the Central Limit Theorem to justify using a normal test statistic.

It can be listed the values of the statistics used to test specified hypothesis H_0 concerning means and give the corresponding critical region for one and two tailed alternative hypothesis H_1.

Table 3.1 Number of hours per week spent studying for introductory finance and accounting courses.

Finance	**(Rank)**	**Accounting**	**(Rank)**
10	(10)	13	(17.5)
6	(2)	17	(22)
8	(4.5)	14	(19)
10	(10)	12	(15.5)
12	(15.5)	10	(10)
13	(17.5)	9	(7)
11	(13)	15	(20)
9	(7)	16	(21)
5	(1)	11	(13)
11	(13)	8	(4.5)
		9	(7)
		7	(3)
	Rank sum 93.5		Rank sum 159.5

These ranks are shown in the Table 3.1 Now, if the null hypothesis were true, it would expect the average rank for the finance students is 9.35, while that for the accounting students is 13.29. As usual, when testing hypothesis, it want to know likely a discrepancy of this magnitude would be if the null hypothesis were true.

It note that is it is not necessary to calculate both rank sums, for if it know one, it can deduce the other. In the example, for instance, the ranks must sum to the sum of the integers 1 through 22 that is to 253. Thus any test of our hypothesis can be based on just one of the rank sums.

In general, suppose that n_1 observations are available from first population and n_2 from the second, and R_1 denote the sum of the ranks of the observations from the first population. The Mann-Whitney test statistic is then define as

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1 \tag{7}$$

In testing the null hypothesis that the central locations of the two population distributions are the same, it assume that apart from any possible differences in central location, the two population distributions are identical. It can be shown, then, that if the null hypothesis is true, the random variable U has mean

$$E(U) = \mu_u = \frac{n_1 n_2}{2} \tag{8}$$

and variance

$$Var(U) = \sigma_u^2 = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12} \tag{9}$$

Furthermore, under the null hypothesis, the distribution of U approaches the normal quite rapidly as the number sample observations increases. Hence, for moderately large sample size, the distribution of the random variable

$$Z = \frac{U - \mu_u}{\sigma_u} \tag{10}$$

is well approximated by the standard normal. This allows tests to be carried out in a straightforward, as described following;

3.3 Multiple Regression

Many business and economics problems are interested in the relationship between one variable Y, and several, say k, other variables X_1, X_2,..., X_k. For example, the amount of gas pumped at each of several service stations might be related to the location of the station as well as the number of pumps. If X_1 represent the number of pumps at each of the gas stations in a sample and let X_2 represent daily traffic count past each of these stations, then it can use these k = 2 variables together to explain or predict the variations in the amount of gas pumped at each station Y [1].

It obtains from the i th gas station the set of numbers (Y_i, X_{i1},Xi_2). If there are n gas stations in the sample, it has a sample of size n that consists of the n sets of values for Y, X_1, and X_2, as follow [3];

Y, X_1, X_2

Y_1, X_{11}, X_{12}

Y_2, X_{21}, X_{22}

Y_n, X_{n1}, X_{n2}

The object in this chapter is to study the relationship between the variables X_1, X_2,..., X_k and the variable Y.

One way to study the relationship between two or more X variable and Y is multiple regression. Any time a regression equation uses two or more X variables in explaining or predicting variation in the dependent Y variable, the regression equation is called a multiple regression equation, because there are multiple independent or X variable. Multiple regression are used any time and dependent variable can be made more accurate by using more than one associated variable. In the gas station example, by using the number of pumps operated by the station $X_1$and the traffic count past a station X_2 it should be able to predict more accurately what volume of gasoline the station will pump.

In this chapter present the multiple regression model and the procedures used to develop and evaluate multiple regression equations. These procedure, however, are rather tedious to implement in practical problems.

3.3.1 Curvilinear regression

The sample regression equations have consideration and represented straight lines or planes, in order to worlds in linear. At times, however, theory, experience, or a

scatter diagram may suggest a curvilinear relationship between the dependent variable Y and the independent variable X [3].

One way to study the relationship between the dependent variable Y and the independent variable X_1 when the relationship between the two is curvilinear is to use a polynomial regression model. Polynomial regression models can have one or more independent variables, and the independent variable cab be raised to various power in another powers. However, it will be restricted to a polynomial regression model with one independent variable. Furthermore, the independent variable will be raised to the first power in one term of the regression model and to the second power in another term of the regression model. This particular polynomial regression model that has the independent variable X_1 raised to the second power is a *quadratic regression model,* and its regression equation can be expressed as follows;

Equation for Quadratic Regression Model

$$Y_i = A + B_1 X_{i1}, + B_2 X_i^2{}_1 + e_i \qquad (11)$$

To estimate the quadratic regression model's parameters, A, B_1 and B_2 simply let [6],

$$X_2 = X_1{}^2 \qquad (12)$$

And use the ordinary least squares method of multiple linear regression to find the values for the estimators a, b_1, and b_2 in the following equation:

$$Y = a + b_1 X_1 + b_2 X_2 \qquad (13)$$

Data analysis allows to establish new variables from variables that already exist. That is X_2 can be established by the computer in this case by $X_2 = X_1{}^2$.