Debre Markos University

College of social sciences

Department of Geography and Environmental Studies

Teaching note for the course Quantitative methods in Geography and Environmental studies[[1]](#footnote-1)

Contents

[1 Introduction 1](#_Toc455715073)

[1.1 Concept of quantitative Geography 1](#_Toc455715074)

[1.2 Applications of quantitative Geography 2](#_Toc455715075)

[1.3 Types of statistics-Descriptive and inferential 3](#_Toc455715076)

[1.4 Statistical data 3](#_Toc455715077)

[1.5 Scale of measurement 4](#_Toc455715078)

[2 Descriptive Statistics and Graphical display of quantitative data 4](#_Toc455715079)

[2.1 Descriptive Statistics 4](#_Toc455715080)

[2.1.1 Measures of Central Tendency 5](#_Toc455715081)

[2.1.2 Measures of dispersion (variability) 12](#_Toc455715082)

[2.1.3 Measures of Shape of Distribution 20](#_Toc455715083)

[2.2 Graphical display of quantitative data 22](#_Toc455715084)

[3 Concept of Inferential statistics and probability 24](#_Toc455715085)

[3.1 Inferential statistics and hypothesis testing 24](#_Toc455715086)

[3.2 Concept of probability 27](#_Toc455715087)

[4 Parametric and Non parametric tests 29](#_Toc455715088)

[4.1 Parametric tests 30](#_Toc455715089)

[4.1.1 One sample T-test 30](#_Toc455715090)

[4.1.2 Independent t-test 31](#_Toc455715091)

[4.1.3 Paired sample T-test (dependent T-test) 33](#_Toc455715092)

[4.1.4 One way ANOVA 35](#_Toc455715093)

[4.2 Non parametric test 41](#_Toc455715094)

[4.2.1 Chi square test for independence 42](#_Toc455715095)

[4.2.2 The mann-whitney U-test 43](#_Toc455715096)

[4.2.3 The wilcoxon signed-rank test 46](#_Toc455715097)

[4.2.4 Kruskal-Wallis Test 47](#_Toc455715098)

[5 Correlation and Regression Analysis 49](#_Toc455715099)

[5.1 Correlation 49](#_Toc455715100)

[5.1.1 Types of Correlation 51](#_Toc455715101)

[5.2 Coefficient of determination(CD) 55](#_Toc455715102)

[5.3 Regression Analysis 55](#_Toc455715103)

[5.3.1 Types of Regression Models 55](#_Toc455715104)

**Chapter one**

# 1 Introduction

## Concept of quantitative Geography

* Quantitative geography consists of one or more of the following activities:
  + The analysis of spatial numerical data;
  + The development of spatial theory;
  + The construction and testing of mathematical models of spatial processes.

The goal of all these activities is to add to our understanding of spatial processes. This can be done directly, as in the case of spatial choice modeling, where mathematical models are derived based on theories of how individuals make choices from a set of spatial alternatives. Or it can be done indirectly, as in the analysis of spatial point patterns, from which spatial process might be inferred.

Since investigation made by Quantitative physical geographers more likely involve predictable processes, tend to adopt the naturalist view point more frequently than their human geography counterparts. In human geography, where the subject matter is typically clouded by human idiosyncrasies, measurement problems and uncertainty, the search is not generally for hard evidence that global laws of human behavior exist. Rather the quantitative analysis in human geography is to accumulate sufficient evidence which makes the adoption of particular line of thought. One of the strength of quantitative approach is that it enables the measurements of determinants that can be measured (and in many cases these provide very useful and very practical information for real world decision making).

However these measurements might be subject to some uncertainty. This recognition of the role of uncertainty is often more important in the application of quantitative techniques to human geography than to physical geography and makes the former in some ways more challenging and at the same time more receptive to innovative ideas about how to handle this uncertainty.

What distinguishes quantitative geography from other subjects such as econometrics, quantitative sociology, physics, engineering or operations research, is its predominant focus on spatial data. Spatial data are those which combine attribute information with location information. The increasing recognition that *‘spatial is special’* reflects the maturing of quantitative geography from being predominantly a user of other disciplines' techniques to being an exporter of ideas about the analysis of spatial data.

Sometimes the concept of statistical analysis and mathematical model blurred. A Model might for example be developed from mathematical principles and then be standardized by statistical methods. Typically areas such as the analysis of point patterns, spatial regression, and various descriptive measures of spatial data such as autocorrelation are thought of as ‘statistical’. Whereas topics such as spatial interaction modeling, and location-allocation modeling are thought of as ‘mathematical.

## Applications of quantitative Geography

A major goal of geographical research ,whether it be quantitative or qualitative, empirical or theoretical, humanistic or positivists is to generate knowledge about the processes influencing the spatial patterns ,both human and physical, that we observe on the earth’s surface. The advantages of quantitative analysis in this framework are fourfold.

***Firs,*** Quantitative methods allow the reduction of large data sets to a smaller amount of more meaningful information. This is important in analyzing the increasingly large spatial data sets obtained from varieties of sources such as satellite imagery, census counts, local government, market research firms and various land survey.

***Secondly,*** an increasing role for quantitative analysis is in explanatory data analysis which consists of a set of techniques to explore data (and also model out puts) in order to suggest hypothesis or to examine the presence of outliers.

***Thirdly*** quantitative analysis allows us to examine the role of randomness in generating observed spatial patterns of data and to test hypotheses about such patterns. In spatial analysis we typically, although not always, deal with a sample of observations from a larger population and we wish to make some inferences about the population from the sample. Statistical analysis will allow such an inference to be made. For instance, suppose we want to investigate the possible linkage between the location of a nuclear power station and nearby incidences of childhood leukemia. We could use statistical techniques to inform us of the probability that such a spatial cluster of the disease could have arisen by chance. The statistical test would not provide us with a definite answer –we would just have a better basis on which to judge the reliability of our conclusion.

***Fourthly,*** the mathematical modeling of spatial processes is useful in a number of ways. The calibration of spatial models provides information on the determinants of those processes through the estimates of the models’ parameters. They also provide a framework in which predictions can be made of the spatial impact of various actions such as the building of a new shopping development on traffic patterns or the building of a seawall on coastal erosion Finally, models can be used normatively to generate expected values under different scenarios against which reality can be compared.

## Types of statistics-Descriptive and inferential

**Descriptive statistics** – Methods of organizing, summarizing, and presenting data in an informative way. **Inferential statistics** – The methods used to determine something about a population on the basis of a sample information.

**Population –**The entire set of individuals or objects of interest or the measurements obtained from all individuals or objects of interest. **Sample –** A portion, or part, of the population of interest

## Statistical data

Statistical data are usually obtained by counting or measuring items. ***Most statistical data are either quantitative or qualitatitve.***  **Qualitative -** data are measurements that each fall into one of several categories (types of land use, sex, ethnic groups and other attributes of the population). **Quantitative** - data are observations that are measured on a numerical scale (temperature, rainfall amount, slope in degree or percent, population density, family size etc.).

**1. Qualitative data**

Qualitative data are generally described by words or letters. They are not as widely used as quantitative data because many numerical techniques do not apply to the qualitative data. For example, it does not make sense to find an average temperature or rain fall.

Qualitative data can be separated into two subgroups:

* **Dichotomic** (if it takes the form of a word with two options (gender - male or female, poor or non poor, yes or no).
* **polynomic** (if it takes the form of a word with more than two options (education - primary school, secondary school and university, income level- low, medium, high,Agro ecological zones-Highland, midland, lowland etc. ).

**2. Quantitative data**

Quantitative data are always numbers and are the **result of counting or measuring** attributes of a population. Quantitative data can be separated into two subgroups:

* **Discrete** (If it is the result of *counting* (e.g. number of farmers, family members, number cropping seasons etc.)
* **continuous** (if it is the result of *measuring* (e.g. farm land size, yield, weight, height, temperature, rainfall etc)

## Scale of measurement

**1. Nominal** – Consist of categories in each of which the number of respective observations is recorded. The categories are in no logical order and have no particular relationship. The categories are said to be ***mutually exclusive*** since an individual, object, or measurement can be included in only one of them. Ethnicity, religion, nationality, residence region or place or agro ecology are some of examples.

**2. Ordinal-**Consists of distinct categories in which order is implied. Values in one category are larger or smaller than values in other categories (e.g. rating-excelent, good, fair, poor)

**3. Interval** – is a set of numerical measurements in which the distance between numbers is of a known, constant size e.g. temperature

**4. Ratio** – consists of numerical measurements where the distance between numbers is of a known, constant size, in addition, there is a ***nonarbitrary zero point.*** Height, weight, width of tree, individuals are some of examples.

**Chapter two**

# Descriptive Statistics and Graphical display of quantitative data

## Descriptive Statistics

* **Measures of Central Tendency:-** A single number to serve as a representative value around which all the numbers in the set tend to cluster. Sometimes it is referred to as a “middle” number of the data.
* **Measures of Variation or Variability:-** measures of *dispersion* (or *variability* or *spread*) indicate the extent to which the observed values are “spread out” around that center — how “far apart” observed values typically are from each other and therefore from some average value.
* **Measures of Shape of distribution:-**They describe the relative position of specific measurements in the data.

### Measures of Central Tendency

Measures of central tendency are numbers that tend to cluster around the “middle” of a set of values. Common measures of central tendencyare Mean (e.g. arithmetic mean, harmonic mean, and geometric mean), Median and Mode

**A. Mean (Arithmetic, harmonic and geometric mean)**

**1. Arithmetic mean**

The arithmetic mean is the "standard" average, often simply called the "mean".

* ***Arithmetic mean for raw Data***

Suppose, we have 'n' observations (or measures) x1, x2, x3... xn then the Arithmetic mean is

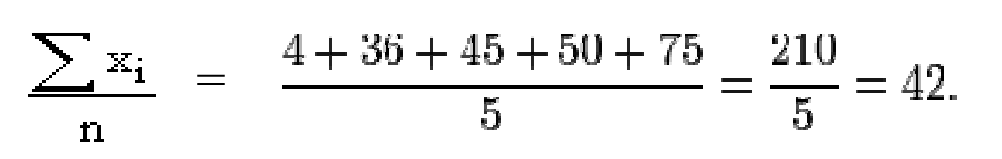
http://www.pinkmonkey.com/studyguides/subjects/stats/chap4/Image299.gif

We shall use the symbol (pronounced as x bar) to denote the Arithmetic mean. Therefore, the Arithmetic mean of the set x1 + x2 + x3 + .......+ xn is given by,

http://www.pinkmonkey.com/studyguides/subjects/stats/chap4/Image302.gif

This method is known as the ''Direct Method".

**Example1.** The arithmetic mean of five values: 4, 36, 45, 50, and 75 is



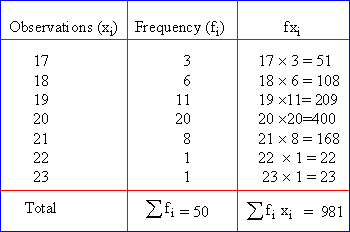
* **Arithmetic mean for ungrouped frequency data (Discrete Series)**

If the observations x1 + x2 + x3 + .......+ xn are repeated f1 + f2 + f3 + ......+ fn times, then we have:

Arithmetic mean=

http://www.pinkmonkey.com/studyguides/subjects/stats/chap4/Image313.gif

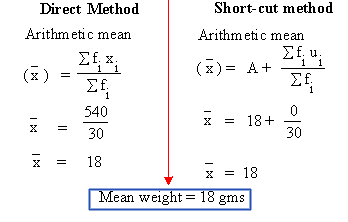
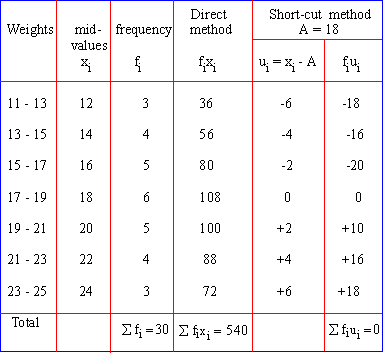
**Example1:**The arithmetic mean of the following 50 observations.

http://www.pinkmonkey.com/studyguides/subjects/stats/chap4/Image318.gif

* ***Arithmetic Mean for intervals or grouped frequency distribution (continuous series)***

The procedure of finding the arithmetic mean in this series is the same as we have used in the discrete series. The only difference is that in this series, we are given class-intervals, whose mid-values (class-marks) are to be calculated first.

***Example:*** Find the arithmetic mean for the following data.



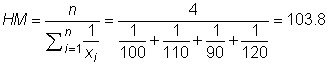
**2. Harmonic Mean**

The ***harmonic mean*** is a better "average" when the numbers are defined in relation to some unit.  The common example is averaging speed. **Example1:** suppose that you have four 10 km segments to your automobile trip.  You drive your car:

* + - * 100 km/hr for the first 10 km
      * 110 km/hr for the second 10 km
      * 90 km/hr for the third 10 km
      * 120 km/hr for the fourth 10 km.

What is your average speed?

**Solution:**



**3. Geometric Mean**

In economic evaluation work, the ***geometric mean***is often useful.

The formula to calculate Geometric mean is



***Example:*** The geometric mean of five values: 4, 36, 45, 50, and 75 is:

(4 \times 36 \times 45 \times 50 \times 75)^{^1/_5} = \sqrt[5]{24\;300\;000} = 30.

**B. Median**

It is the value of the middle item and divides the series in to two equal parts. A number such that at most half of the measurements are below it and at most half of the measurements are above it. It is the value of the size of the central item of the arranged data (data arranged in the ascending or the descending order).

* **Median for raw Data (In Individual Series)**

Set the individual series either in the ascending (increasing) or in the descending (decreasing) order, of the size of its items or observations. If the total number of observations be 'n' then and n is odd

The median = http://www.pinkmonkey.com/studyguides/subjects/stats/chap4/Image380.gif size of   observation

**Example:**Find the median for the following sets of numbers; 96, 180, 98, 75, 270, 80, 102, 100, 94, 75 and 200.

**Solution:**

* Arrange the data in the ascending order as 75, 75, 80, 94, 96, 98, 100, 102,180, 200, 270.

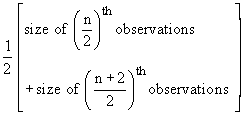
Now the total number of items 'n'= 11 (odd)

Therefore, the median = size of item

=size of itemhttp://www.pinkmonkey.com/studyguides/subjects/stats/chap4/Image380.gif

=size of 6th itemhttp://www.pinkmonkey.com/studyguides/subjects/stats/chap4/Image382.gif  
    =    98

If 'n' is even, the median=



**Example:** The population (in thousands) of 36 metropolitan cities is as follows:

2468, 591, 437, 20, 213, 143, 1490, 407, 284, 176, 263, 19, 181, 777, 387, 302, 213, 204, 153, 733, 391, 176 178, 122, 532, 360, 65, 260, 193, 92, 672, 258, 239, 160, 147, 151. Then Calculate the median.

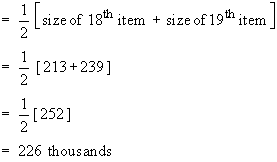
**Solution:**

First Arrang the terms in the ascending order as:

20, 65, 92, 131, 142, 143, 147, 151, 153, 160, 169, 176, 178, 181, 193, 204, (213, 39), 258, 263, 260, 384, 302, 360, 387, 391, 407, 437, 522, 591, 672, 733, 777, 1490, 2488.

Since total number of items n = 36 (Even).

The median=http://www.pinkmonkey.com/studyguides/subjects/stats/chap4/Image383.gif



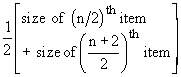
* **Median for ungrouped frequency data (In Discrete Series)**

***Steps:***

1. Arrange the data in ascending or descending order of magnitude.
2. Find the cumulative frequencies.
3. Apply the formula :
4. If 'n' =  http://www.pinkmonkey.com/studyguides/subjects/stats/chap4/Image385.gif (odd) then,

Median = size of item http://www.pinkmonkey.com/studyguides/subjects/stats/chap4/Image386.gif

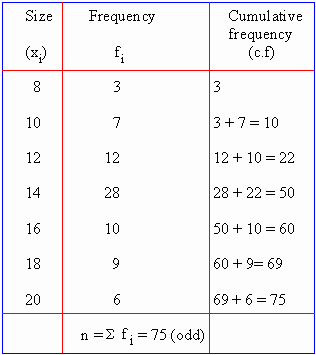
B. If 'n' =    http://www.pinkmonkey.com/studyguides/subjects/stats/chap4/Image385.gif (even) then

Median=

**Example:** Locate the median for the following distribution.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Temperature(Celsius) | 8 | 10 | 12 | 14 | 16 | 18 | 20 |
| Frequency | 3 | 7 | 12 | 28 | 10 | 9 | 6 |

**Solution**



Therefore, the median = http://www.pinkmonkey.com/studyguides/subjects/stats/chap4/Image389.gif

**http://www.pinkmonkey.com/studyguides/subjects/stats/chap4/Image390.gif**

= size of 38th item

In the order of the cumulative frequency, the 38th term is present in the 50th cumulative frequency, whose size is 14. Therefore, the median = 14

* **Median for grouped interval data (Continuous Series)**

***Steps:***

1. Determine the particular class in which the value of the median lies using the formulahttp://www.pinkmonkey.com/studyguides/subjects/stats/chap4/Image391.gif. Use as the rank of the median and

2. After ascertaining the class in which median lies, the following formula is used to determine the exact value of the median.

Median = http://www.pinkmonkey.com/studyguides/subjects/stats/chap4/Image393.gif

***Where,***

http://www.pinkmonkey.com/studyguides/subjects/stats/chap4/Image394.gif= lower limit of the median class, the class in which the middle item of the distribution lies.

http://www.pinkmonkey.com/studyguides/subjects/stats/chap4/Image395.gif = upper limit of the median class

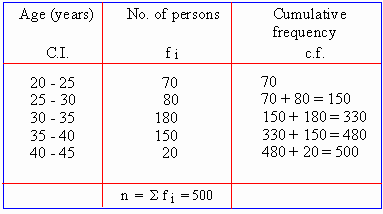
c.f = cumulative frequency of the class preceding the median class

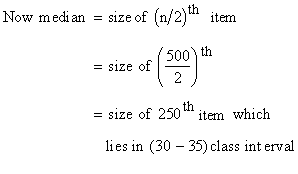
f = sample frequency of the median class

**Example** Calculate the median for the age of 500 people given below;

**Solution:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Age(years) | 20-25 | 25-30 | 30-35 | 35-40 | 40-45 |
| No.of person | 70 | 80 | 180 | 150 | 20 |

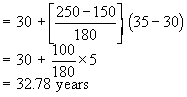




Median=http://www.pinkmonkey.com/studyguides/subjects/stats/chap4/Image393.gif

***Where;***

http://www.pinkmonkey.com/studyguides/subjects/stats/chap4/Image394.gif = 30,  http://www.pinkmonkey.com/studyguides/subjects/stats/chap4/Image398.gif = 35,  http://www.pinkmonkey.com/studyguides/subjects/stats/chap4/Image399.gif= 250, c.f. = 150 and f = 180

Therefore, Median=

1. **Mode**:

It is the size of items which possesses the maximum frequency (most frequently occurring value). In short the number which occurs most often is the mode.

* ***Mode for Ungrouped Data (individual series)***

**Example:** Locate mode for the data; 7, 12, 8, 5, 9, 6, 10, 9, 4, 9 and 9**.** Since 9 occur most frequently compared to others it is the mode. Set of data may have **uni-modal, *Bi-modal or Multi-modal series.***

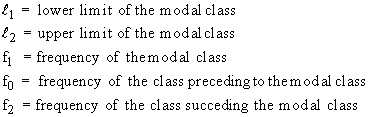
* ***Mode for Grouped frequency (continuous series) Data***

***Steps:***

1. Determine the modal class which has the maximum frequency.

2. By interpolation the value of the mode can be calculated as

Mode =



**Example:**Calculate the modal wages for the data given below;

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Daily wages in$ | 20-25 | 25-30 | 30-35 | 35-40 | 40-45 | 45-50 |
| No. Of workers | 1 | 3 | 8 | 12 | 7 | 5 |

***Solution:***

* Here the maximum frequency is 12, corresponding to the class interval (35 - 40) which is the modal class.

Therefore =http://www.pinkmonkey.com/studyguides/subjects/stats/chap4/Image410.gif

***Then;***

Mode =







* Modal wages is $37.22

### Measures of dispersion (variability)

The measures of central tendencies (i.e. means) indicate the general magnitude of the data and locate only the center of a distribution of measures. They do not establish the degree of variability or the spread out or scatter of the individual items and their deviation from the means. Therefore, Measures of central tendency very often present an incomplete picture of data. In order to evaluate more completely any group of scores it is necessary to measure the spread or dispersion of the data being studied. From this discussion we now focus our attention on the scatter or variability which is known as dispersion. Thus the ’dispersion’ is also known as the "average of the second degree." Methods of Computing Dispersion include range, Mean deviation, Standard Deviation, variance, Quartile deviation and coefficient of variation (CV).

**1. Range**

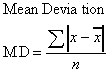
In any statistical series, the difference between the largest and the smallest values is called as the range. The range is the simplest measure of finding variation. ***Range = Maximum - Minimum***

**2. Mean deviation**

Mean deviation is the average amount of scatter of the items in a distribution from either the mean or the median or the mode, ignoring the signs of these deviations. It is the mean of the distances between each value and the mean. It gives us an idea of how values spread out from the center.

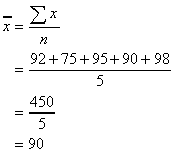
* **Mean deviation for Raw Data (individual series).**

In this case the mean deviation is given by the formula;



**Example:** Find the mean deviation for the data set 92, 75, 95, 90, and 98.

***First find the mean.***



Now subtract the mean from each score, take the absolute value of each difference, total the absolute values, then divide by the number of values.

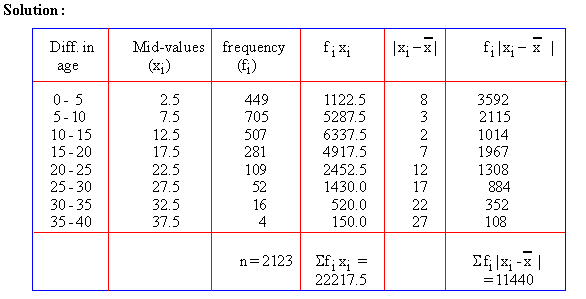
|  |  |  |
| --- | --- | --- |
| **x** | http://infinity.cos.edu/faculty/woodbury/stats/tutorial/IMG00113.GIF | http://infinity.cos.edu/faculty/woodbury/stats/tutorial/IMG00114.GIF |
| 92 | 2 | 2 |
| 75 | -15 | 15 |
| 95 | 5 | 5 |
| 90 | 0 | 0 |
| 98 | 8 | 8 |
|  | http://infinity.cos.edu/faculty/woodbury/stats/tutorial/IMG00115.GIFhttp://infinity.cos.edu/faculty/woodbury/stats/tutorial/IMG00117.GIF | http://infinity.cos.edu/faculty/woodbury/stats/tutorial/IMG00116.GIF30 |

So, we can say that on the average, these student's test scores deviated by 6 points from the mean. We follow the same procedure for discrete series data except we multiply the value with the corresponding frequency.

* ***Mean deviation for grouped data (*Continuous series)**

**Example:** calculate the mean deviation from the following data. Difference in ages between boys and girls of a class.

|  |  |
| --- | --- |
| **Diff. in years:** | **No of students:** |
| 0 – 5 | 449 |
| 5 – 10 | 705 |
| 10 – 15 | 507 |
| 15 – 20 | 281 |
| 20 – 25 | 109 |
| 25 – 30 | 52 |
| 30 – 35 | 16 |
| 35 – 40 | 4 |



http://www.pinkmonkey.com/studyguides/subjects/stats/chap5/Image189.gif

MDhttp://www.pinkmonkey.com/studyguides/subjects/stats/chap5/Image190.gif

**3. Standard Deviation and Variance**

Standard deviation is the most commonly used measure of the spread or dispersion of data around the mean. The standard deviation is defined as the square root of the variance (V).  The **variance** (abbreviated SD2 or s2) is another measure of variability.  Operationally, there are several ways of calculating standard deviation. The following is one of those.

Standard deviation and variance for ungrouped data

* **Standard deviation and variance for raw data**

**Example:** Find the standard deviation and variance for the following data

|  |  |  |
| --- | --- | --- |
| Slope in (%) | Xi – x | (Xi – x)2 |
| 5 | –1 | 1 |
| 4 | –2 | 4 |
| 8 | +2 | 4 |
| 2 | –4 | 16 |
| 8 | +2 | 4 |
| 7 | +1 | 1 |
| 2 | –4 | 16 |
| 9 | +3 | 9 |
| 7 | +1 | 1 |
| 8 | +2 | 4 |
| 60 | Σ(xi – x)2 = | 60 |
|  | (Variance)SD2 = | 60 / 9 = 6.667 |
|  | (Standard deviation)SD = |  |

Variance and standard deviation for discrete series and contentious series data is calculated using the following formula except we calculate mid value for contentious series data first.

* **Standard deviation and variance for discrete series data**

***Example:*** Compute standard deviation for the slope of 30 separate plot of lands

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Slope(%) (xi) | Number of plot of lands (f) | fixi |  |  |  |
| 5 | 3 | 15 | -2.5 | 6.25 | 18.75 |
| 6 | 4 | 24 | -1.5 | 2.25 | 9 |
| 7 | 2 | 14 | -0.5 | 0.25 | 0.5 |
| 4 | 5 | 20 | -3.5 | 12.25 | 61.25 |
| 8 | 4 | 32 | 0.5 | 0.25 | 1 |
| 9 | 6 | 54 | 1.5 | 2.25 | 13.5 |
| 11 | 6 | 66 | 3.5 | 12.25 | 73.5 |
|  | =30 | 225 |  |  | =177.5 |

**Solution**

* **Standard deviation and variance for continuous data**

**Example:** Find standard deviation for the following data set

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Class | Mid value(xi) | Frequency(fi) | fixi |  |  |  |
| 5-10 | 7.5 | 5 | 37.5 | -16.3448 | 267.1534 | 1335.7669 |
| 11-16 | 13.5 | 6 | 81 | -10.3448 | 107.0155 | 642.09275 |
| 17-22 | 19.5 | 4 | 78 | -4.34483 | 18.87753 | 75.510107 |
| 23-28 | 25.5 | 3 | 76.5 | 1.655172 | 2.739596 | 8.2187872 |
| 29-34 | 31.5 | 4 | 126 | 7.655172 | 58.60166 | 234.40666 |
| 35-40 | 37.5 | 2 | 75 | 13.65517 | 186.4637 | 372.92747 |
| 41-46 | 43.5 | 5 | 217.5 | 19.65517 | 386.3258 | 1931.629 |
|  |  | **29** | **691.5** |  |  | ***4600.5517*** |

1. ***Quartile deviation***

Before discussing quartile deviation it is good to understand about quartile and inter quartile range because it is helpful to calculate quartile deviation. If we concentrate on two extreme values (as in the case of range), we don’t get any idea about the scatter of the data within the range (i.e. the two extreme values). If we discard these two values the limited range thus available might be more informative. For this reason the concept of Inter-quartile range is developed. It is the range which includes middle 50% of the distribution. Here 1/4 (one quarter of the lower end) and 1/4 (one quarter of the upper end) of the observations are excluded. Now the lower quartile (Q1) is the 25th percentile and the upper quartile (Q3) is the 75th percentile. It is interesting to note that the 50th percentile is the middle quartile (Q2) which is in fact what you have studied under the title’ Median ".

***Thus symbolically***

Inter quartile range = Q3 - Q1

Hence Quartile deviation is half of the deference between Q3 and Q1. It is also known as semi inter quartile range. The formula of Quartile Deviation is

In getting the quartile deviation from ungrouped data, the following steps are used

1. Arrange the scores from highest to lowest.
2. Assign serial numbers to each score. The first serial number is assigned to the lowest score, while the last serial number is assigned to the highest score.
3. Determine the first quartile (*Q1*). To be able to locate *Q*1, divide *N* by 4. Use the obtained value in locating the serial number of the score that falls under *Q1*­.
4. Determine the third quartile *(Q3),* by dividing 3*N* by 4. Locate the serial number corresponding to the obtained answer.
5. Subtract *Q1* from *Q3* and divide the difference by 2.

* **Quartile deviation for raw data(individual series)**

Example:find quartile deviation for the following daily temperature record for ten days in degree Celsius. 9,11,12,13,14,15,15.16,18 and 20.

**Solution:**

2nd +o.75 (3rd -2nd)

11+ 0.75(12-11)

**=11.75**

8th +0.25(9th -8th )

16+0.25(18-16)

16+0.5

**=16.5**

* **Quartile deviation for discrete data**

**Example:** The following is the height of 26 trees. Examine variation of height using quartile deviation.

|  |  |  |
| --- | --- | --- |
| Height of trees in meter | Number of trees | Cf |
| 3 | 5 | 5 |
| 4 | 7 | 12 |
| 5 | 4 | 16 |
| 6 | 3 | 19 |
| 7 | 2 | 21 |
| 8 | 1 | 22 |
| 9 | 4 | 26 |

**Solution:**

6th +0.75(7th -6th)

4+0.75(4-4)

20th +0.25 (21th -20th)

7+0.25(7-7)

**Getting the Quartile Deviation from Grouped Data**

In getting the quartile deviation from grouped data, the following steps are used

1. Cumulate the frequencies from the bottom to the top of the grouped frequency distribution.
2. For the first quartile, use the formula Q1 =L +

***Where:***

L = lower limit of the Q1 class

N/4 = locator of the Q1 class

N = total number of scores

CF = cumulative frequency preceding the Q1 class

i = class size/interval

1. For the third quartile, use the formula Q3 = L +

***Where***

L = lower limit of the Q3 class

3N/4 = locator of the Q3 class

N = total number of scores

CF = cumulative frequency preceding the Q3 class

i = class size/interval

**Example: Computation of the Quartile Deviation for livestock holding of households**

|  |  |  |
| --- | --- | --- |
| ***Classes*** | ***Frequency (f)*** | ***Cumulative Frequency (CF)*** |
| 11-15  16-20  21-25  26-30  31-35  36-40  41-45  46-50 | 4  4  6  8  10  9  7  5  N = 53 | 4  8  14  22  32  41  48  53 |

*For the first quartile*

= = 13. 25

*CF* = 8 *f* = 6 *L* = 21

Q1 = L +

= 21 + (5)

= 21 + (5)

= 21 + = **25.375**

*For the third quartile*

= = **39.75**

*CF* = 32 *f* = 9 *L* = 36

Q3 = L +

= 36 + (5)

= 36 + (5)

= 36 + 4.3

= 40.3

After obtaining the first and third quartiles, we can now compute *QD.* Thus *QD =* ().

*QD* =()

= ()

= 7.462

***5. Coefficient of Variation***

Coefficient of variation (**CV**) measures the **spread** of a set of data as a proportion of its mean. It is the **ratio** of the sample **standard deviation** to the sample **mean**. It is sometimes expressed as a **percentage**. A standard application of the **Coefficient of Variation** (CV) is to characterize the **variability** of **geographic variables** over space or time.the formula to calculate CV is;

http://www.fao.org/docrep/W7295E/w7295e08.gif

**Coefficient of Variation** (CV) is particularly applied to characterize the **inter-annual** **variability** of **climate variables** (e.g., temperature or precipitation) or **biophysical variables** (leaf area index (LAI), biomass, etc). **Example**: The mean annual rainfall (mm) data for certain semiarid area of six years are 389, 540, 576, 510, 456, and 430. Calculate the coefficient of variation of the six years mean annual rainfalls of the area.

***Solution;***

Mean **=** 483.5 mm, Standard Deviation = 70.7 mm

The CV shows that a high variability of rainfall of the area, in some years the area gets 14.6% higher and in others the same percentage lower rain from the six years average.

http://www.fao.org/docrep/W7295E/w7295e08.gif



### Measures of Shape of Distribution

A Fundamental task in many statistical analyses is to characterize the **location** and **variability** of a data set (Measures of **central tendency** vs. measures of **dispersion).** But these statistical means and measures of variation are not enough to draw sufficient inferences about the data. Another aspect of the data is to know its ***symmetry and flatness.***  Symmetry is well studied by the knowledge of the ***"skewness"*** while flatness is understood by knowledge of the ***Kurtosis***

Therefore, A further characterization of the data includes **skewness** and **kurtosis**. The **histogram** is an effective **graphical** technique for showing both the **skewness** and **kurtosis** of a data set.

**1. Skewness**

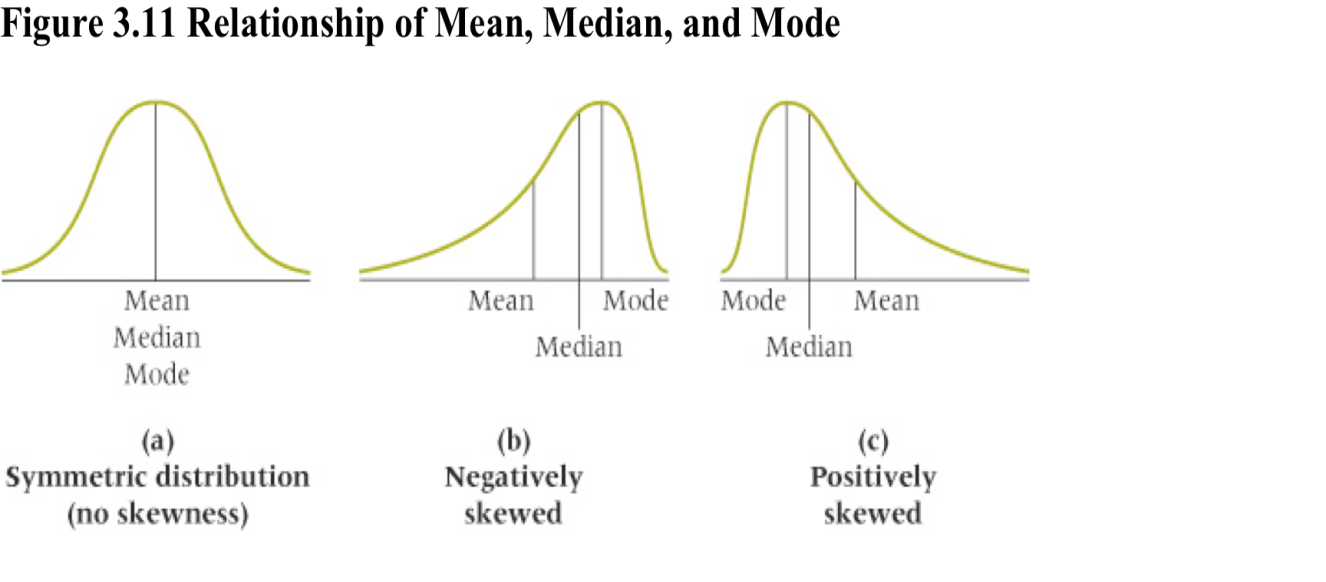
**Skewness** measures the degree of asymmetry exhibited by the data

If **skewness** equals zero, the histogram is **symmetric** about the mean.

**Positive skewness:** -There are more observations below the mean than above it-When the mean is greater than the median.

**Negative skewness:-**There are a small number of low observations and a large number of high ones-When the median is greater than the mean.

***Skewness graphically***



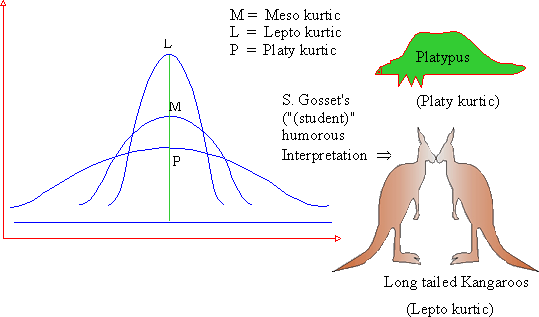
***2. Kurtosis***

**Kurtosis** measures how peaked the histogram is.

The **kurtosis** of a **normal distribution** is 0. **Kurtosis** characterizes the relative **peakedness** or **flatness** of a distribution compared to the normal distribution

* **Platykurtic**– When the **kurtosis < 0**, the frequencies throughout the curve are closer to be equal (i.e., the curve is more **flat** and **wide**). Thus, **negative kurtosis** indicates a relatively **flat** distribution
* **Mesokurtic-**when kurtosis=0
* **Leptokurtic**– When the **kurtosis > 0**, there are high frequencies in only a small part of the curve (i.e, the curve is more **peaked**). Thus, **positive kurtosis** indicates a relatively **peaked** distribution.

***Kurtosis graphically***

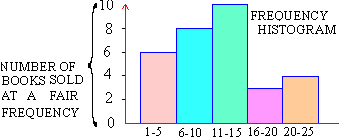


## Graphical display of quantitative data

Quantitative data can be displayed using ***histogram, bar graph, pie chart or line graph***

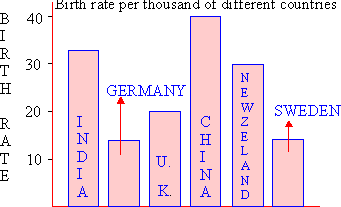
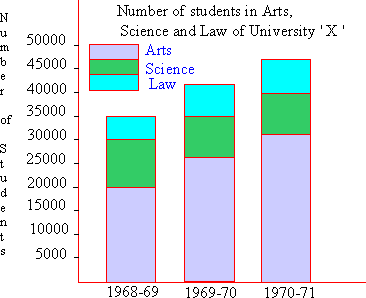
***1. Histogram***

It is defined as a pictorial representation of a grouped frequency distribution by means of adjacent rectangles, whose areas are proportional to the frequencies. To construct a Histogram, the class intervals are plotted along the x-axis and corresponding frequencies are plotted along the y - axis. A histogram is similar to a bar chart. However histograms are used for ***continuous (as opposed to discrete or qualitative)*** data.



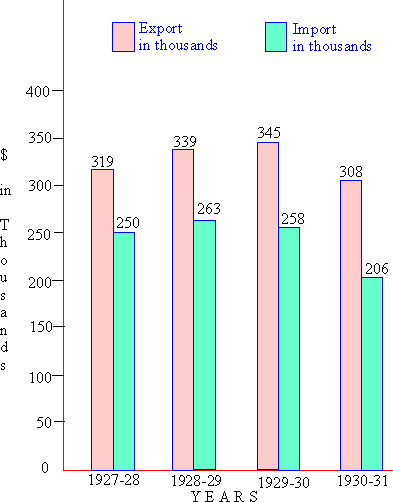
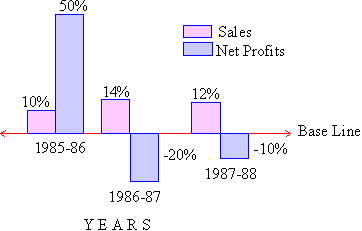
**2. Bar graph/chart**

The Bar Chart (or Bar Graph) is one of the most common ways of displaying categorical/qualitative data. The different types of bar graphs are **simple, Sub – divided, Multiple and Deviation Bar graph**

**Sub- divided bar graph**

Simple bar Graph

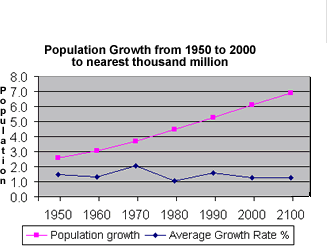
 

Deviation bar graph

Multiple bar graph

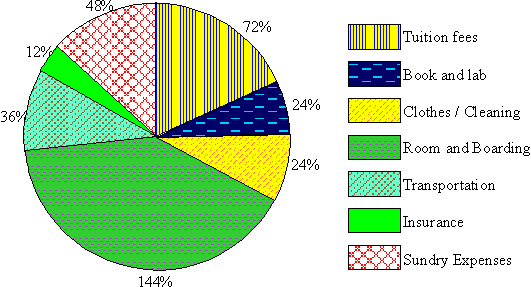
**3. Line graph**

A line graph is used to display continuing data over time, e.g. rainfall or population over a year. The lines in the graph indicate whether the data value rises or falls during the time period.



**4. Pie chart**

A Pie-Chart/Diagram is a graphical device - a circular shape broken into sub-divisions. The sub-divisions are called "*sectors*“. A pie diagram is useful when we want to show relative positions (proportions) of the figures which make the total. It is also useful when the components are many in number.



**Chapter three**

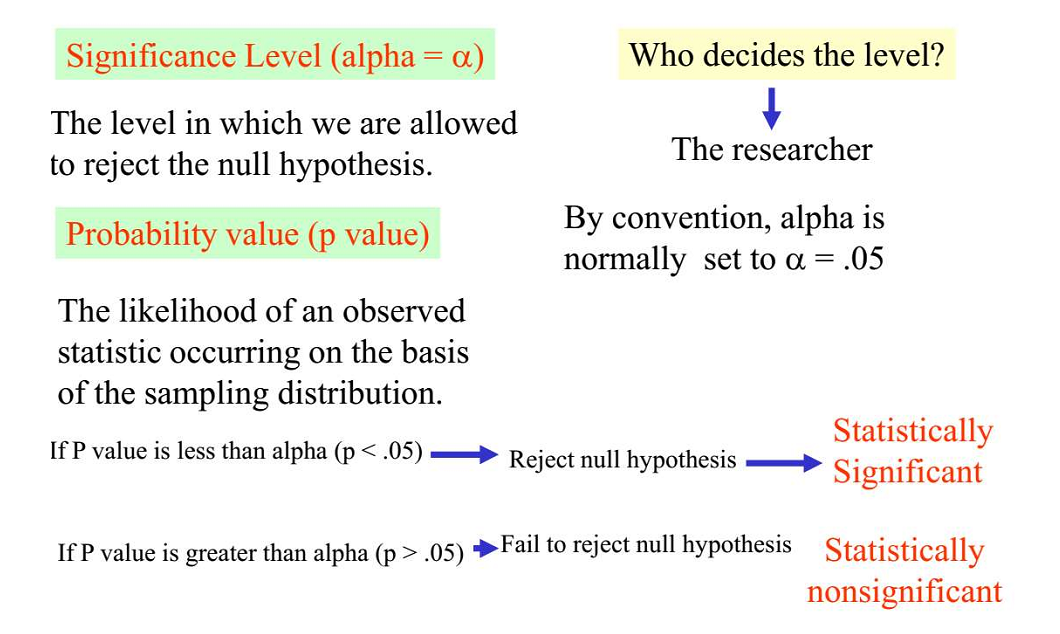
# Concept of Inferential statistics and probability

## Inferential statistics and hypothesis testing

It is a mathematical tool that permit the researcher to generalize to a population of individuals based upon information obtained from a limited number of research participants (samples) using hypothesis testing procedures.

Hypothesis is a statement supposed to be true till it is proved false. It may be based on previous experience or may be derived theoretically. The approach here is to set up an assumption that there is no contraction between the **believed result** and the **sample result**and that the difference therefore can be ascribed solely to chance. Such a hypothesis is called a **null hypothesis ( Ho)**. If the null hypothesis is rejected, that is taken as evidence in favor of the research hypothesis which is called as the **alternative hypothesis (denoted by H**a**)**.

**Null hypothesis**. The null hypothesis, denoted by H0, is usually the hypothesis that sample observations result purely from chance.  **Alternative hypothesis**. The alternative hypothesis, denoted by H1 or Ha, is the hypothesis that sample observations are influenced by some non-random cause. Null hypothesis is the statement that the difference between two sample means is due to random, chance, sampling error. It indicates that there is no true difference or relationship between parameters in the populations.

**Testing for statistical significance** 

Statistical analyses determine whether to accept or reject the null hypothesis

**Alpha level**

An established probability level which serves as the criterion to determine whether to accept or reject the null hypothesis. It represents the confidence that your results reflect true relationships.

***Common significance levels***

* + *p* < .01 (I will correctly reject the null hypothesis 99 of 100 times)
  + *P <* .05 (I will correctly reject the null hypothesis 95 of 100 times)
  + *p <* .10 (I will correctly reject the null hypothesis 90 of 100 times)

**Steps in Hypothesis Testing**

Here are the steps to performing hypothesis testing

1. Formulate the null hypothesis
2. Determining the appropriate statistical test
3. Selecting the significance level
4. Calculate the statistical test
5. State the rule of rejecting the null hypothesis
6. Decision: accept or reject the null hypothesis.
7. Statement of the result(conclusion)

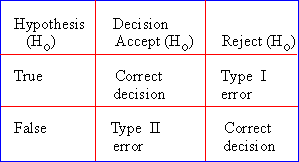
**Decision Rules**

The analysis plan includes decision rules for rejecting the null hypothesis. In practice, statisticians describe these decision rules in two ways - with reference to a ***P-value or with reference to a region of acceptance.*** P-value- If the P-value is less than the significance level, we reject the null hypothesis. Region of acceptance- The region of acceptance is a range of values. If the test statistic falls within the region of acceptance, the null hypothesis is not rejected.

The region of acceptance is defined so that the chance of making a Type I error is equal to the significance level. Rejecting null hypothesis does not mean that either proofing or disproving the hypothesis, but it is to check whether the sample information support the hypothesis or not.

**Errors in Testing of Hypothesis**

Two types of errors can result from a hypothesis test. **Type I error**. A Type I error occurs when the researcher rejects a null hypothesis when it is true. The probability of committing a Type I error is called the **significance level** or **alpha(**α). **Type II error**. A Type II error occurs when the researcher fails to reject a null hypothesis that is false. The probability of committing a Type II error is called **Beta.**



**One-Tailed and Two-Tailed Significance Tests**

When do you use a one-tailed or two-tailed test of significance? The answer is that it depends on your hypothesis. When your research hypothesis states the direction of the difference or relationship, then you use a one-tailed probability. For example, a one-tailed test would be used to test these null hypotheses: Females will not score significantly higher than males on an IQ test.

In each case, the null hypothesis (indirectly) predicts the direction of the difference. **A two-tailed** test would be used to test these null hypotheses (***no direction***): There will be no significant difference in IQ scores between males and females. There will be no significant difference in the amount of product purchased between blue collar and white collar workers.

## Concept of probability

**Probabilities in Geography**

The analyses of many problems (daily or geographic) are often based on **probabilities**, such as: What are the “**chances**” of having rain over the weekend? What is the “**likelihood”** that the 100-year floods will occur within the next ten years How “**likely**” is it that a pixel on a satellite image is correctly classified or misclassified?

**Probability & Probability Distribution**

We **summarize** a sample statistically and want to make some **inferences** about the population (e.g., what **proportion** of the **population** has values within a given range). The concept of **probability** is the key to making **statistical inferences** by sampling a population. What we are doing is trying to ascertain the probability of **an event having a given outcome**. This requires us to be able to specify the **distribution** of a variable before we can **make inferences**

**Probability-Related Concepts**

**An event** – Any phenomenon you can observe that can have **more than one outcome** (e.g., flipping a coin). **An outcome** – Any **unique condition** that can be the result of an event (e.g., flipping a coin: **heads** or **tails**), a.k.a simple event or sample points. **Sample space** – The set of **all** possible outcomes associated with an event. e.g., flip a coin – **heads** (H) and **tails** (T). e.g., flip a coin twice – HH, HT, TH, TT. Associated with each possible outcome in a **sample space** is a **probability**. **Probability** is a measure of the **likelihood** of each possible outcome. **Probability** measures the degree of **uncertainty**. Each of the probabilities is greater than or equal to zero, and less than or equal to one. The **sum of probabilities** over the **sample space** is equal to one

**Example I** – when we flip a coin, **two** possible outcomes: “**heads**” and “**tails**”. Each outcome is **equally** likely.” **Heads**” and “**tails**” have the same probability (0.5). The **sum of probabilities** over the **sample space** is one

**How to Assign Probabilities to Experimental Outcomes?**

There are numerous ways to assign probabilities to the elements of sample spaces. Classical method assigns probabilities based on the assumption of equally likely outcomes. Relative frequency method assigns probabilities based on experimentation or historical data*.* Subjective method assigns probabilities based on the assignor’s judgment or belief.

1. **Classical Method:** This approach assumes that each outcome is equally likely. If an experiment has *n* possible outcomes, this method would assign a probability of 1/*n* to each outcome. It is an appropriate way to assign probabilities to the outcomes in special kinds of experiments. **Example I**: during rolling a die Sample Space may be *S* = {1, 2, 3, 4, 5, 6}. Then the Probabilities each sample point has a **1/6** chance of occurring. **Example** **II** – Flip four coins. Let “0” represent “**heads**” and “1” represents “**tails**”. For each toss, the probability of “heads” or “tails” is ½. Assuming that outcomes of the four tosses are **independent** from one another there are **sixteen** possible outcomes.

**Probability of each outcome**:

½ \* ½ \* ½ \* ½ = 1/16 = 0.0625

|  |  |  |  |
| --- | --- | --- | --- |
| 0000 | 0100 | 1000 | 1100 |
| 0001 | 0101 | 1001 | 1101 |
| 0010 | 0110 | 1010 | 1110 |
|  |  |  |  |

1. **Relative Frequency Method:** The second way is to assign them on the basis of relative frequencies. Example, given a weather pattern, a meteorologist may note that in 65 out of the last 100 times that such a pattern prevailed there was measurable precipitation the next day. If there were such a weather pattern today, what would the probability of having rain tomorrow be? The possible outcomes – rain or no rain tomorrow – are assigned probabilities of 0.65 and 0.35, respectively
2. **Subjective Method:** When extreme weather conditions occur it might be inappropriate to assign probabilities based solely on historical data. We can use any data available as well as our experience and intuition, but ultimately a probability value should express our degree of belief that the experimental outcome will occur. The best probability estimates often are obtained by combining the estimates from the classical or relative frequency approach with the subjective estimates.

**Chapter Four**

# Parametric and Non parametric tests

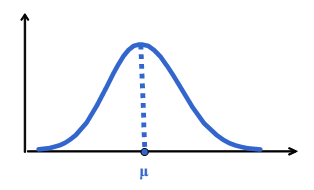
***Comparing parametric & Non parametric***

**Parametric Test Procedures**

1. Involve Population Parameters Example: Population Mean



2. Have Stringent Assumptions Example: Normal Distribution



3. Require Interval Scale or Ratio Scale Whole Numbers or Fractions Example: Height in Inches (72, 60.5, 54.7)

4. Examples: z-Test, t-Test, ANOVA

**Nonparametric Test Procedures**

1. Do Not Involve Population Parameters

2. No Stringent Distribution Assumptions  
“Distribution-free”

3. Data Measured on Any Scale Ratio or Interval Ordinal Example: Good-Better-Best Nominal Example: Male-Female

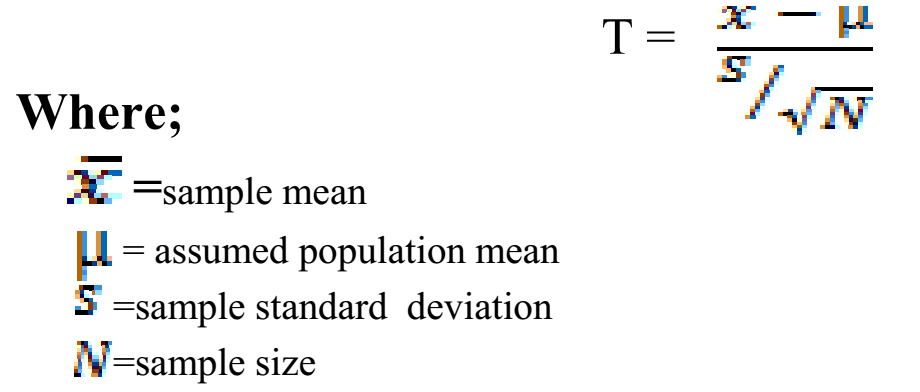
4. Example: Wilcoxon rank sum test, Kruskal wallis H-test, Chi-square test and

## Parametric tests

### One sample T-test

We use one sample T-test to test against a hypothetical mean and when the population variance is unknown. We use the following formula to test against the hypothetical mean.

.



***Example: A*** researcher collected a temperature value for 25 days and found that the average daily temperature to be 150c and sample standard deviation of 9. Does this average value of temperature significantly differed from the population value 120c at 0.05 significance value?

***Solution***

For one sample t-statistic the degrees of freedom (df) is equal to the sample size minus 1 or df=N-1 where N is the sample size.

***Steps of testing this hypothesis***;

1. State null and alternative hypothesis

HO:µ=12

H1:µ≠12

2. Set alpha value and degree of freedom=0.05 Df=25-1=24

3. Calculate the value of the proper statistics



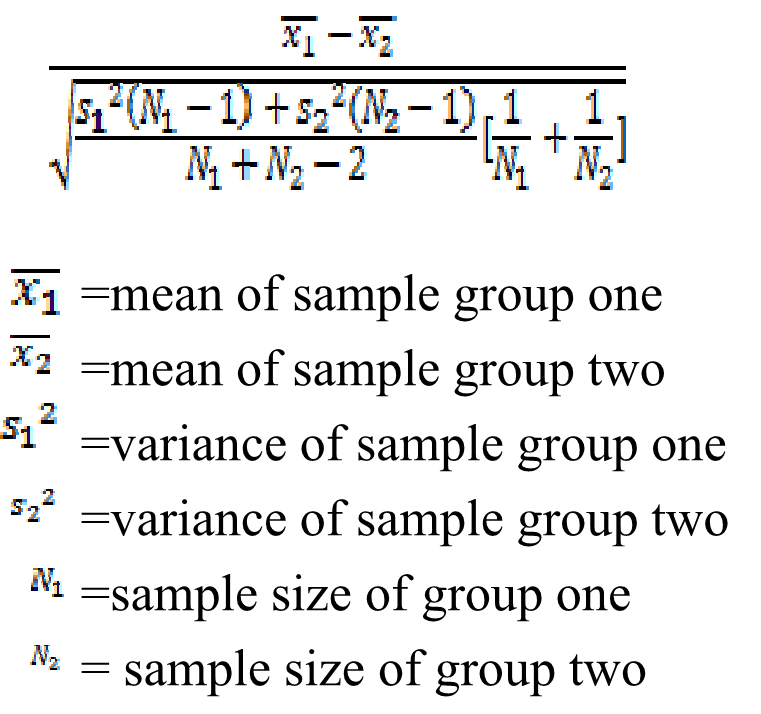
4. State the rule of rejecting the null hypothesis: reject Ho if t≥2.064 or if t≤-2.064 otherwise accept it.

5. Decision ***Ho accepted.*** Since the cal calculated value of t(1.667) is not greater than or equal to 2.064 or less than or equal to -2.064 we cannot reject the null hypothesis. That is the critical value at 24 df (i.e N-1=25-1) and 0.05 significance level in two tailed is equal to 2.064.

***6. Statement of the result:*** The average daily temperature is not different from the usual/population average temperature.

### Independent t-test

It is used to compare differences between separate groups. Any differences between groups can be explored with the independent t-test, as long as the tested members of each group are reasonably representative of the population. The following is the formula to calculate independent t-test.

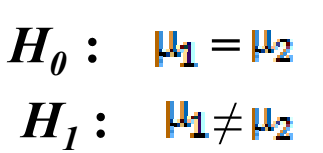


**Example: A** researcher hypothesized that there is no difference between male and female headed households in terms of the amount of Teff yield they get from their respective farm land. To test his hypothesis he collected the amount of Teff yield in kuintal from sample of 13 female headed households and 13 male headed households as follow. Then do you think that the researcher will reject his hypothesis at 0.05 ***significance level***

|  |  |  |
| --- | --- | --- |
| No | males headed (in quintal)(group1) | Female headed (in quintal)(Group2) |
| 1 | 7 | 7 |
| 2 | 8 | 8 |
| 3 | 9 | 4 |
| 4 | 3 | 10 |
| 5 | 5 | 14 |
| 6 | 10 | 9 |
| 7 | 15 | 9 |
| 8 | 12 | 10 |
| 9 | 6 | 4 |
| 10 | 8 | 3 |
| 11 | 10 | 2 |
| 12 | 4 | 6 |
| 13 | 5 | 7 |

***Solution***

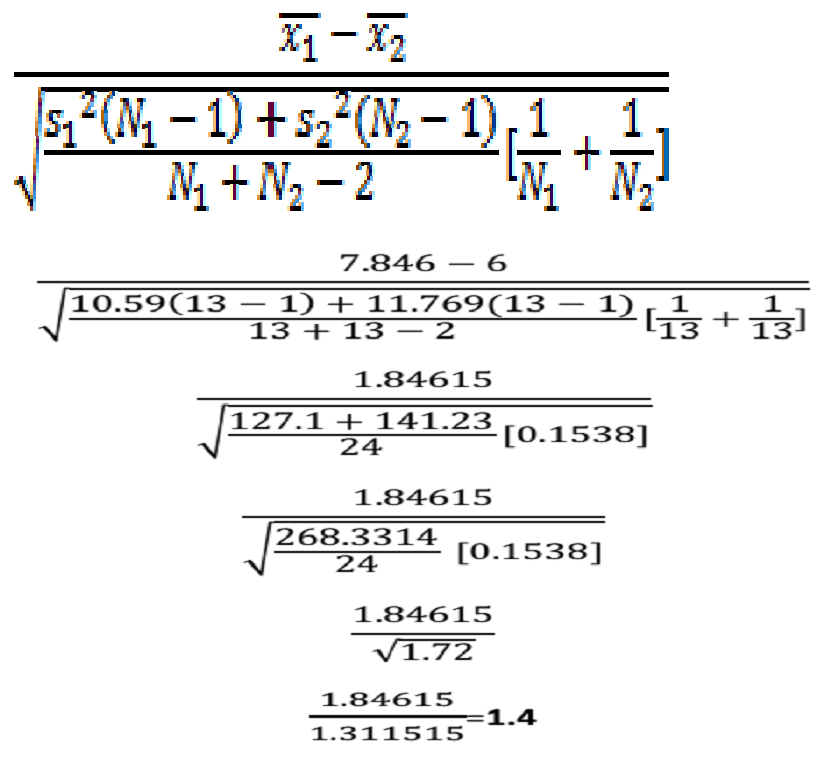
***1. Formulate the null hypothesis***

******

2. Determining the appropriate statistical test we use “independent t-test “because it is comparing the mean of the two groups”

3. Selecting the significance level i.e. 0.05

4. Calculate the statistical test



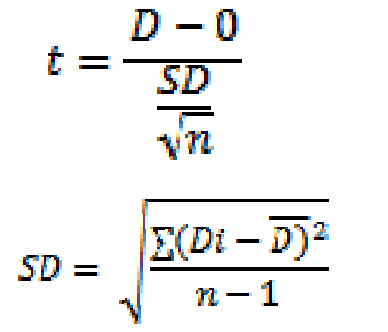
5. State the rule of decision: reject Ho if t≥2.064 or if t≤-2.064 otherwise accept it. This is because the critical value at 24 df(i.e.N1+N2-2=13+13-2=24) and two tailed 0.05 significance level is equal to 2.064.

**6. Decision accept Ho .** Since the calculated value of t(1.4) is not greater than or equal to 2.064 or less than or equal to -2.064 we cannot reject the null hypothesis.

**7. Staement of the result:** the average yield of Teff produced by female and male headed households does not differed. *The difference is not statistically significant.*

### Paired sample T-test (dependent T-test)

In many research designs, it is helpful to measure the same people more than once. A common example is testing for performance improvements (or decrements) over time or after treatment or intervention. However, in any circumstance where multiple measurements are made on the same person (or “experimental unit”), it may be useful to observe if there are mean differences between these measurements. The paired t-test will show whether the differences observed in the 2 measures will be found reliably in repeated samples. We use the following formula to calculate paired t-test (dependant t-test).



**Example:** A researcher hypothesized that manure application does not have an effect on yield of wheat. To ensure his hypothesis he took a sample of 16 equal size separate plot of lands. By the year 2015 he harvested wheat from each land without manure application and recorded the amount. By the next year he added 10 kg of manure to each land, harvested maize and recorded for each as follow. Dou you think that his hypothesis retained from rejection at 5% significance level?

|  |  |  |
| --- | --- | --- |
| NO | Yield (quintal) before manure added | Yield (quintal) after manure added |
| 1 | 8 | 7 |
| 2 | 7 | 9 |
| 3 | 9 | 3 |
| 4 | 10 | 4 |
| 5 | 11 | 6 |
| 6 | 14 | 8 |
| 7 | 13 | 9 |
| 8 | 4 | 3 |
| 9 | 2 | 4 |
| 10 | 5 | 6 |
| 11 | 8 | 8 |
| 12 | 9 | 9 |
| 13 | 6 | 3 |
| 14 | 7 | 4 |
| 15 | 8 | 5 |
| 16 | 4 | 7 |

***Solution;***

***Steps***

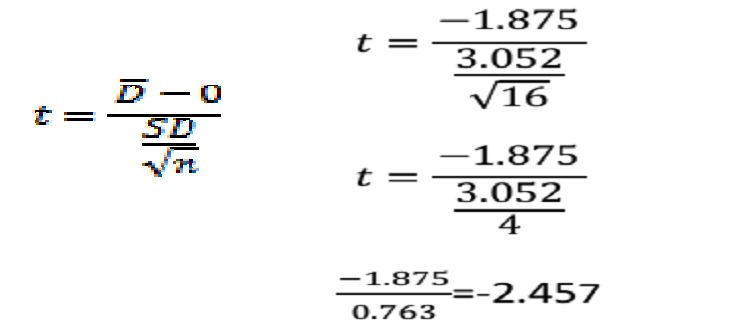
1. Formulate the null hypothesis



2.Determining the appropriate statistical test we use here “dependant t-test“ because it is comparing the mean difference of a single sample group before and after treatment in this case adding manure is the treatment.

3. Selecting the significance level i.e. 0.05

4. Calculate the statistical test



5. State the rule of decision. reject Ho if t≥1.753 or if t≤ -1.753 otherwise accept it. This is because the critical value for 15 df (i.e. N-1=16-1=15) and 0.05 significance level in two tailed is 1.753).

**6. Decision: Ho rejected.** Since the calculated value of t(-2.457) is less than -1.753 reject the null hypothesis.

**7. Statement of the result:** adding manure makes the yield to decline.

### One way ANOVA

A One-Way Analysis of Variance (***One way ANOVA)*** is a way to test the equality of three or more means at one time by using variances.

**Hypotheses**

The null hypothesis will be that all population means are equal; the alternative hypothesis is that at least one means is different.



***The grand mean*** of a set of samples is the total of all the data values divided by the total sample size(N). i.e



***The total variation*** is comprised the sum of the squares of the differences of each mean with the grand mean. Denoted by SS(T). There is the between group variation and the within group variation. The whole idea behind the analysis of variance is to compare the ratio of between group variance to within group variance.

**Between Group Variation**

The variation due to the interaction between the samples is denoted SS(B) for Sum of Squares Between groups. There are k samples involved with one data value for each sample (the sample mean), so there are k-1 degrees of freedom. The variance due to the interaction between the samples is denoted MS(B) for Mean Square Between groups. This is the between group variation divided by its degrees of freedom. The degree of freedom is equal to the sum of the individual degrees of freedom for each sample. Since each sample has degrees of freedom equal to one less than their sample sizes, and there are k samples, the total degrees of freedom is k less than the total sample size: df = N - k.

**Within Group Variation**

The variation due to differences within individual samples, denoted SS(W) for Sum of Squares Within groups. Each sample is considered independently, no interaction between samples is involved. The variance due to the differences within individual samples is denoted MS(W) for Mean Square Within groups. This is the within group variation divided by its degrees of freedom.

**F test statistic**

The F test statistic is found by dividing the between group variance by the within group variance. The degrees of freedom for the numerator are the degrees of freedom for the between group (k-1) and the degrees of freedom for the denominator are the degrees of freedom for the within group (N-k).

**Summary Table**

All of this sounds like a lot to remember, and it is. However, there is a table which makes things really nice.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | SS | df | MS | F |
| Between | SS(B) | k-1 | SS(B) -----------  k-1 | MS(B) --------------  MS(W) |
| Within | SS(W) | N-k | SS(W) -----------  N-k |
| Total | SS(W) + SS(B) | N-1 |  |

**Decision Rule**

The decision will be to reject the null hypothesis if the test statistic from the table is greater than the F critical value with k-1 numerator and N-k denominator degrees of freedom. If the decision is to reject the null, then at least one of the means is different. However, the ANOVA does not tell you where the difference lies. For this, you need another test, either the ***Scheffe' or Tukey test*** test.

***Example:*** A researcher hypothesized that there is no significant difference between farmer groups (rich, medium and poor) with regard to the total grain yield amount they produce . To test the hypothesis she collected yield amount in quintal from 15,18 and 24 rich, medium and poor farm groups respectively as shown in the following table. Then do you think that the researcher will reject her hypothesis at 5% significance level?

|  |  |  |  |
| --- | --- | --- | --- |
| No | Rich Farm group's Grain quintal | Medium's Farm group Grain quintal | poor's Farm group Grain quintal |
| 1 | 25 | 14 | 5 |
| 2 | 24 | 13 | 6 |
| 3 | 22 | 12 | 9 |
| 4 | 21 | 12 | 8 |
| 5 | 20 | 13 | 9 |
| 6 | 30 | 11 | 7 |
| 7 | 35 | 11 | 7 |
| 8 | 19 | 13 | 6 |
| 9 | 18 | 10 | 6 |
| 10 | 26 | 16 | 8 |
| 11 | 25 | 17 | 8 |
| 12 | 28 | 13 | 7 |
| 13 | 20 | 14 | 8 |
| 14 | 22 | 14 | 7 |
| 15 | 25 | 13 | 9 |
| 16 | N=15 | 13 | 7 |
| 17 | **Mean=24** | 12 | 7 |
| 18 |  | 11 | 6 |
| 19 |  | N=18 | 6 |
| 20 |  | **Mena =12.88888889** | 8 |
| 21 |  |  | 8 |
| 22 |  |  | 7 |
| 23 |  |  | 8 |
| 24 |  |  | 7 |
|  |  |  | n=24 |
|  |  |  | **Mena =7.25** |

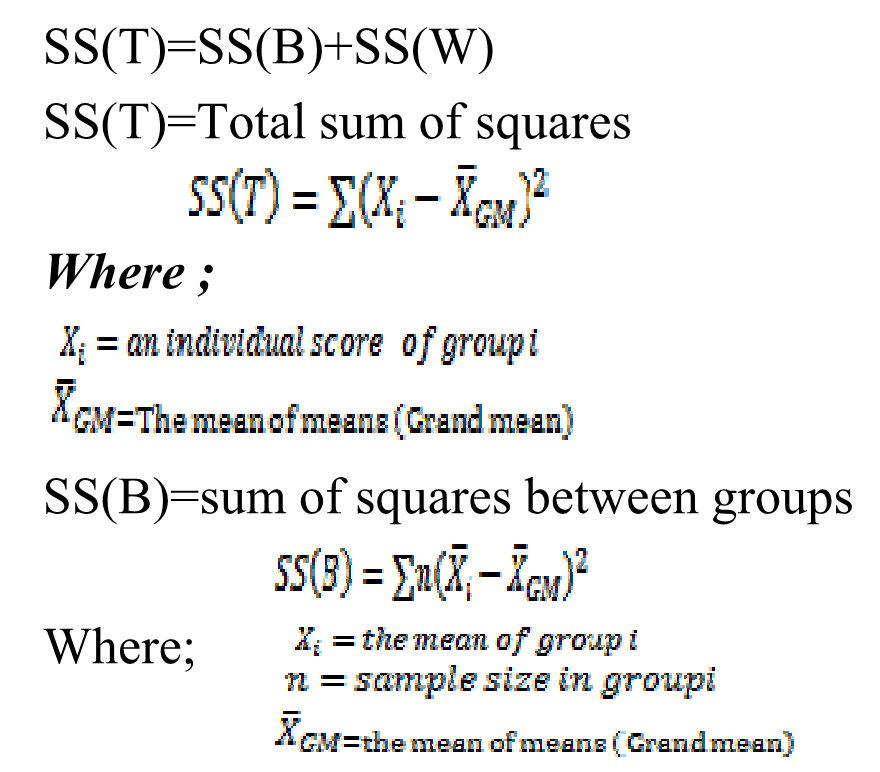
**Solution:-**

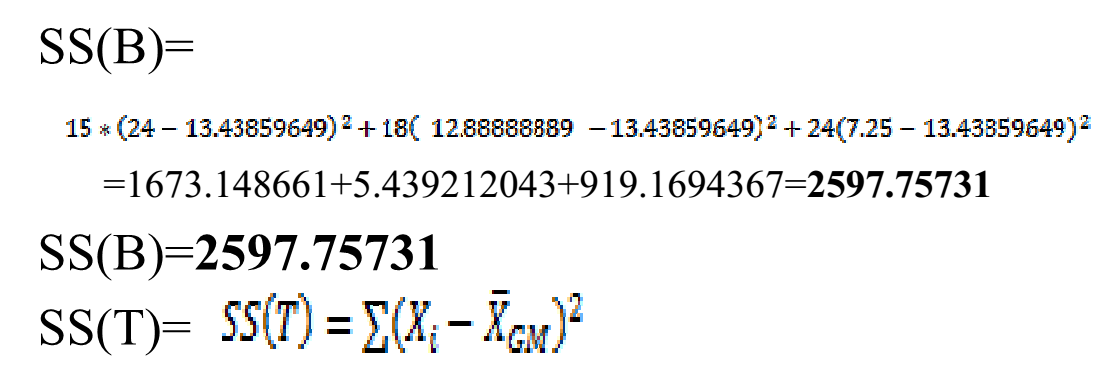
**Steps**

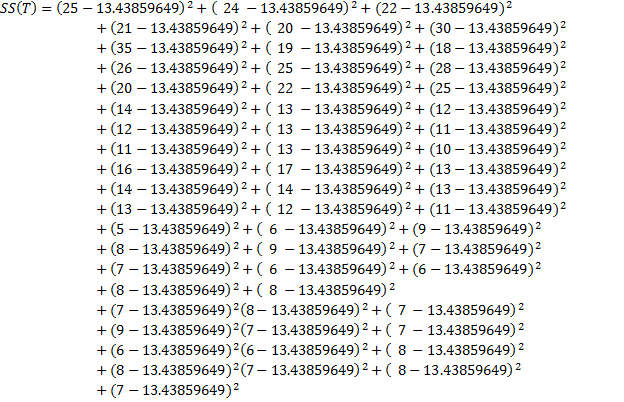
1. Formulate the hypothesis



1. Determining the appropriate statistical test we use here “one way ANOVA “because it is comparing the mean difference of three farmer groups.
2. The significance level i.e. 0.05
3. Calculate the statistical test







SS(T)=111.5432+133.6661+73.29763+57.17482+43.05202+274.2801+464.8941+30.92921+20.8064+157.7889+133.6661+212.0345+43.05202+73.29763+133.6661+0.315174+0.192367+2.06956+2.06956+0.192367+5.946753+5.946753+0.192367+11.82395+6.560788+12.68359+0.192367+0.315174+0.315174+0.192367+0.192367+2.06956+5.946753+71.20991+55.33272+19.70114+29.57833+19.70114+41.45552+41.45552+55.33272+55.33272+29.57833+29.57833+41.45552+29.57833+41.45552+19.70114+41.45552+41.45552+55.33272+55.33272+29.57833+29.57833+41.45552+29.57833+41.45552=**2966.035**

SS(W)=SS(T)-SS(B)

SS(W)=**2966.035- 2597.75731= 368.27769**

* Then represented as follow on the table of

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | SS | df | MS | F |
| Between | 2597.75731 | 3-1=2 | 2597.75731 -----------  2    **=**1298.878655 | 1298.878655 --------------  6.812  =**190.675** |
| Within | 368.27769 | 57-3=54 | 368.27769 -----------  54  =6.812 |
| Total | 2966.035 | 57-1=56 |  |

5. State the rule of decision: reject Ho if F≥3.17 otherwise accept it.

**6. Decision:** Since the calculated value of F (**190.675**) much greater than the tabulated value (**3.17)** the null hypothesis should be rejected at 0.05 significance value and (2, 54) degree of freedom (df).

**7. Staement of the result:** there is a significant difference between farmer groups with regard to the yield they produce. To check which group exceeds the other we have to calculate LSD or Tuket or bonferroni and see the result on SPSS or other statistical packages.

## Non parametric test

Parametric tests involve specific probability distributions (e.g., the normal distribution) and the tests involve estimation of the key parameters of that distribution (e.g., the mean or difference in means) from the sample data. But nonparametric tests are sometimes called distribution-free tests because they are based on fewer assumptions (e.g., they do not assume that the outcome is approximately normally distributed). The cost of fewer assumptions is that nonparametric tests are generally less powerful than their parametric counterparts (i.e., when the alternative is true, they may be less likely to reject H0). In nonparametric tests, the hypotheses are not about population parameters (e.g., μ=50 or μ1=μ2). Instead, the null hypothesis is more general. For example, when comparing two independent groups in terms of a continuous outcome, the null hypothesis in a parametric test is H0: μ1 =μ2. In a nonparametric test the null hypothesis is that the two populations are equal, often this is interpreted as the two populations are equal in terms of their central tendency.

Low power is a major issue when the sample size is small – which unfortunately is often when we wish to employ these tests. The most practical approach to assessing normality involves investigating the distributional form of the outcome in the sample using a histogram and to augment that with data from other studies, if available, that may indicate the likely distribution of the outcome in the population.

### Chi square test for independence

A chi-square statistic is computed to measure the amount of discrepancy between the ideal sample (expected frequencies) and the actual sample data (the observed frequencies). A large discrepancy results in a large value for chi-square and indicates that the data do not fit the null hypothesis and the hypothesis should be rejected. For chi-square, the data are frequencies rather than numerical scores.

The two types of chi-square tests: - **chi-square test for independence and chi-square test for goodness-of-fit.**

**1. The Chi-Square Test for Independence**

It can be used and interpreted in two different ways:

1. Testing hypotheses about the relationship between two variables in a population, or

2. Testing hypotheses about differences between proportions for two or more populations.

**2. Chi-square test for goodness-of-fit**

The **chi-square test for goodness-of-fit** uses frequency data from a sample to test hypotheses ***about the shape or proportions of a population.***  The data, called **observed frequencies**, simply count how many individuals from the sample are in each category. The null hypothesis specifies the proportion of the population that should be in each category. Both chi-square tests use the same statistic.

* The calculation of the chi-square statistic requires two steps:

1. Calculating **expected frequencies**

* For the goodness of fit test, the expected frequency for each category is obtained by Expected frequency = fe = pn (p is the proportion from the null hypothesis and n is the size of the sample). For the test for independence, the expected frequency for each cell in the matrix is obtained by



The calculation of chi-square is the same for all chi-square tests:



The fact that chi-square tests do not require scores from an interval or ratio scale makes these tests a valuable alternative to the t tests, ANOVA, or correlation, because they can be used with data measured on a nominal or an ordinal scale.

***Example:*** A researcher hypothesized that there is no association between sex and education. In other word gender difference does not make difference in education level. To test his hypothesis he collected the following data from 180 respondents. Then do you think that he will reject his hypothesis at 0.05 significance level?

Sex distribution by education level

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | Education | | Total |
| Illiterate | Literate |
| Gender | Female | 36 | 24 | 60 |
| Male | 24 | 96 | 120 |
| Total | | 60 | 120 | 180 |

1. Formulate the hypothesis

HO: There is no significant difference between expected and observed frequency of male and females with regard to their education level.

2. The appropriate statistical test. Here we use “chi-square test“because it is non parametric and efficient to test such association when the data is in nominal measurement scale form.

3. Selecting the significance level i.e. 0.05

4. Calculate the statistical test

|  |  |  |  |
| --- | --- | --- | --- |
| **Observed frequency(O)** | **Expected frequency(E)** | **(Oi-Ei)2** |  |
| 36 | 60\*60/180=20 | 256 | 12.8 |
| 24 | 60\*120/180=40 | 256 | 6.4 |
| 24 | 120\*60/180=40 | 256 | 6.4 |
| 96 | 120\*120/180=80 | 256 | 3.2 |
|  |  |  | =28.8 |

5.State the rule of decision; the tabulate value of chi-square test at 1 degree of freedom(df)(r-1)(c-) i.e (2-1)(2-1) and at 0.05 significance level is 3.84. Accept the null hypothesis if the calculated value is less than the tabulated value otherwise reject it.

**6. Decision reject Ho.** Because the calculated value of chi square test (28.8) is much greater than the tabulated value (3.84)

**7. Staement of the result:** there is an association between sex and literacy rate. That is more male are literate and while more females are illiterate.

### The mann-whitney U-test

Also known as Wilcoxon-test, Wilcoxon rank sum test, U-test, Mann-Whitney-U-test. It tests Two Independent Population. It Corresponds to t-Test for 2 Independent Means.

The mann-whitney U-test is used to calculate the probability of finding any given difference in rank sums under the null hypothesis that the samples were drawn from populations with the same distribution. As with other tests, the significance attached to the calculated value of the test statistics depends on the significance level chosen, and on whether the test is directional or non directional.

Let the number of scores in the smaller group (if groups are unequal in size) be N1 and the number in the larger group N2.obviuosly,if the groups are equal in size ,N1-N2. We now rank the whole combined set of N1 and N2 scores from lowest (rank1) to the highest. If there is more than one occurrence of the sample scores (that is,’tied ranks), each occurrence is given the mean of the ranks which would have been allocated if there had not been a tie.

For example, in the series 1,2,3,3,3,5,the three occurrences of the scores 3 would have occupied ranks 3,4 and 5 if there had been no tie, and they are therefore each given the mean of these ranks(that is,4). Since three ranks have been used, however, the next number (the 5 in our series) will have the rank of six. We now obtain the sum of ranks (R1) for the smaller sample. If samples are of equal size, either may be used. The value of R1, together with those of N1 and N2, is substituted in the following expression of N1 and N2, is substituted in the following expression for U1:



We can now do the same for the larger group



We now take the smaller of U1 and U2 and call it U. If our calculated value is smaller than or equal to the critical value we can reject the null hypothesis.

***Example;*** suppose that we have a group of 17 people to investigate the satisfaction of households with their access to climate change information. The subjects are allocated randomly to two groups, one of 9(from lowland) and the other of 8 people (from high land). The subjects are asked to grade their satisfaction with their access to climate change information on a scale of being informed from 0(totally uninformed) to 10(highly informed). The investigator wishes to know whether there is any significant difference between the two sets of ratings at the 5 percent level in a non-directional test. We first rank the combined scores, giving an average rank to tied scores, as rank. The rank sum for the smaller group (R1) is 47. We can now use this value to find U1.

Lowland and highland household’s satisfaction rating with their access to climate change information.

|  |  |  |  |
| --- | --- | --- | --- |
| **lowland**  **households(N2=9)** | **Rank** | **Highland**  **households (N1=8)** | **Rank** |
| 7 | 9 | 7 | 9 |
| 8 | 13.5 | 4 | 1 |
| 6 | 5.5 | 5 | 3 |
| 9 | 16 | 6 | 5.5 |
| 10 | 17 | 8 | 13.5 |
| 7 | 9 | 5 | 3 |
| 7 | 9 | 5 | 3 |
| 8 | 13.5 | 7 | 9 |
| 8 | 13.5 |  | R1=47 |





* Using the simple formula for U2,we have

U2=N1N2-U1=(8\*9)-61=72-61=11

Since U2 is the smaller of the two values, we have U=U2=11. We find that the critical value of U for the 5 percent level in a non directional test is 15. Since our calculated value is less than the critical value this(i.e. 11<15), we can reject the null hypothesis, and claim that there is a significant difference between the two sets of ratings.

Above values of about 20 for N1 and N2, the test statistics U conforms to an approximately normal distribution. For large samples, then we calculate a z value as follows



If the calculated value of z is greater than or equal to the critical value for the required significance level we may reject the null hypothesis. Let us suppose that we have calculated U for two samples, one of 30 and one of 35 scores, and have obtained a value of 618.we calculate z as follow:



* If the test is non directional and at the 5 percent level, the critical value of z is 1.96.since the calculated value is smaller than this, we cannot reject the null hypothesis.

### The wilcoxon signed-rank test

The non-parametric equivalent of the t-test for dependant samples is the wilcoxon signed-ranks test. This test assumes that we can rank differences between paired observations. The data for the test will consist of a numbers of pairs of scores, each derived from a single subject or from a pair of matched subjects. In order to calculate the test statistics, we first find the difference between each pair of scores, subtracting consistently (second from first, or vice versa) and recording the signs.

We then rank the absolute values of the difference, ignoring the sign. If two scores in a pair are the same (that is, if the difference is zero), that pair is ignored altogether. If two values of the difference are tied, they are given the mean of the ranks they would have had if they had been different in value (compare the similar procedure in the mann-whitney test). Each rank is now given the sign of the difference it corresponds to. The sum of the positive ranks is found, and also that of the negative ranks.

The smaller of these two sums is the test statistics “w”. If the value of “w” is smaller than or equal to the critical value the null hypothesis will be rejected otherwise it will be accepted. It should be noted that in taking an appropriate value for the number of pairs of scores N, pairs which are tied, and so have been discarded, are not counted.

**Example:** A researcher recorded 10 farmers’ satisfaction, rated in likert scale form (ranging 0 to20) with their maize yield before and after they get an extension services during two consecutive seasons as shown in the following table. Use wilcoxon signed-rank test and ensure whether there is significant yield difference between the two seasons. Test at 5 % significance level.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Farmer** | **Satisfaction**  **Before ex. service** | **Satisfaction**  **After ex. service** | **Difference** | **Rank** |
| A | 8 | 10 | -2 | -4.5 |
| B | 7 | 6 | +1 | +2 |
| C | 4 | 4 | 0 | - |
| D | 2 | 5 | -3 | -7.5 |
| E | 4 | 7 | -3 | -7.5 |
| F | 10 | 11 | -1 | -2 |
| G | 17 | 15 | +2 | +4.5 |
| H | 3 | 6 | -3 | -7.5 |
| I | 2 | 3 | -1 | -2 |
| J | 11 | 14 | -3 | -7.5 |

***Solution:***

We first calculate the differences and rank them, giving a mean rank in the case of ties. We give each rank the sign of the difference it corresponds to. The pair with zero difference is dropped from the analysis, and N, the number of pairs, is decreased accordingly to 9. The sum of the positive ranks is 6.5, while that of the negative ranks is 38.5.

We take the smaller of these, 6.5, as our value for “w”. The critical value of “w” for N=9 and a 5 percent significance level in a non directional test is 5. Since our value of “w” is larger than the critical value, we accept the null hypothesis and conclude that no significant difference has been demonstrated in satisfaction of farmers.

Where the number of pairs is greater than about 20, the distribution of “w” is almost normal, and a z- score can be calculated from the computed “w” value using the following expression



Let us consider the case where a value of “w”=209 has been calculated for a set of data with N=35. We then have



We can ignore the sign, which is normally negative. We know that values of 1.96 and 1.64 are required at the 5 percent level for a non-directional or a directional test, respectively. Therefore, if our test was non directional we must conclude that no significant difference has been demonstrated. If, on the other hand, we had made a directional prediction, we could claim significance at the 5 percent level.

### Kruskal-Wallis Test

The Kruskal-Wallis test is a nonparametric (distribution free) test, and is used when the assumptions of ANOVA are not met.  They both- Kruskal-Wallis test and one way ANOVA assess for significant differences on a continuous dependent variable by a grouping independent variable (with three or more groups).  In the ANOVA, we assume that distribution of each group is normally distributed and there is approximately equal variance on the scores for each group.

The **Kruskal-Wallis test** is used to determine whether three or more independent samples were selected from populations having the same distribution. Given three or more independent samples, the test statistic *H* for the Kruskal-Wallis test is



Where *k* represents the number of samples, *ni* is the size of the *i*th sample, *N* is the sum of the sample sizes, and *Ri* is the sum of the ranks of the *i*th sample.

**Assumptions**

1. We assume that the samples drawn from the population are random.
2. We also assume that the cases of each group are independent.
3. The measurement scale for should be at least ordinal.

***Null hypothesis*** assumes that the samples are from identical populations. ***Alternative hypothesis*** assumes that the samples come from different populations.

Reject the null hypothesis when H is greater than the critical number (always use a right tail test.). As the one-way ANOVA is an extension of the two independent groups t-test, the Kruskal-Wallis test is an extension of the Mann-W hitney U test. If k = 3 and n (i)'s are less than or equal to 5, use Table N to determine the critical value for a specified level of significance.

Decision rule: Reject ***Ho*** if H < H (alpha), otherwise do not reject ***Ho.*** The conclusion will relate the decision with the actual hypothesis being tested. If k, the number of groups is greater than 3 or any of the n(i)‘s are greater than 5, the critical value is found by entering the chi-square distribution (Table F) with k-1 degrees of freedom. The sampling distribution is a chi-square distribution with *k* - 1 degrees of freedom (Where *k* = the number of samples.)

***Procedure***

1. Arrange the data of samples in a single series in ascending order.
2. Assign rank to them in ascending order. In the case of a repeated value, or a tie, assign ranks to them by averaging their rank position.
3. Then sum up the different ranks, e.g. R1 R2 R3….Rn, for each of the different groups.
4. Calculate the statistics.

***Example1:*** Twelve hectars of land under teff have been randomly selcted from three soil types-soil type A,B and C and the yield in each soil type recorded, are given below.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Soil typeA** | **Soil type B** | **Soil type C** |
| Teff yield | 19 | 18 | 14 |
| 20 | 16 | 23 |
| 17 | 22 | 13 |
| 24 | 15 |  |
| 21 |  |  |
|  | N=5 | N=4 | N=3 |

* Based on above information

a. Formulate hypothesis

b. Test using kruskal-wallis test(H-test)

***Ho:*** There is no significant difference in teff yield among the three soil types

***H1***: There is significant difference in Teff yield.

Rejection level is: α=0.05. In order to compute the H-statistics we have to rank the yields in a single sequence ascending order, i.e. giving rank 1 to the lowest value.

Ranks of teff yield in three soil types

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Soil type A** | **Rank** | **Soil Type B** | **Rank** | **Soil C** | **Rank** |
|  | 19 | 7 | 18 | 6 | 14 | 2 |
|  | 20 | 8 | 16 | 4 | 23 | 11 |
|  | 17 | 5 | 22 | 10 | 13 | 1 |
|  | 24 | 12 | 15 | 3 |  |  |
|  | 21 | 9 |  |  |  |  |
| Sum of Ranks =R |  | 41 |  | 23 |  | 14 |

* The H- statistics is therefore







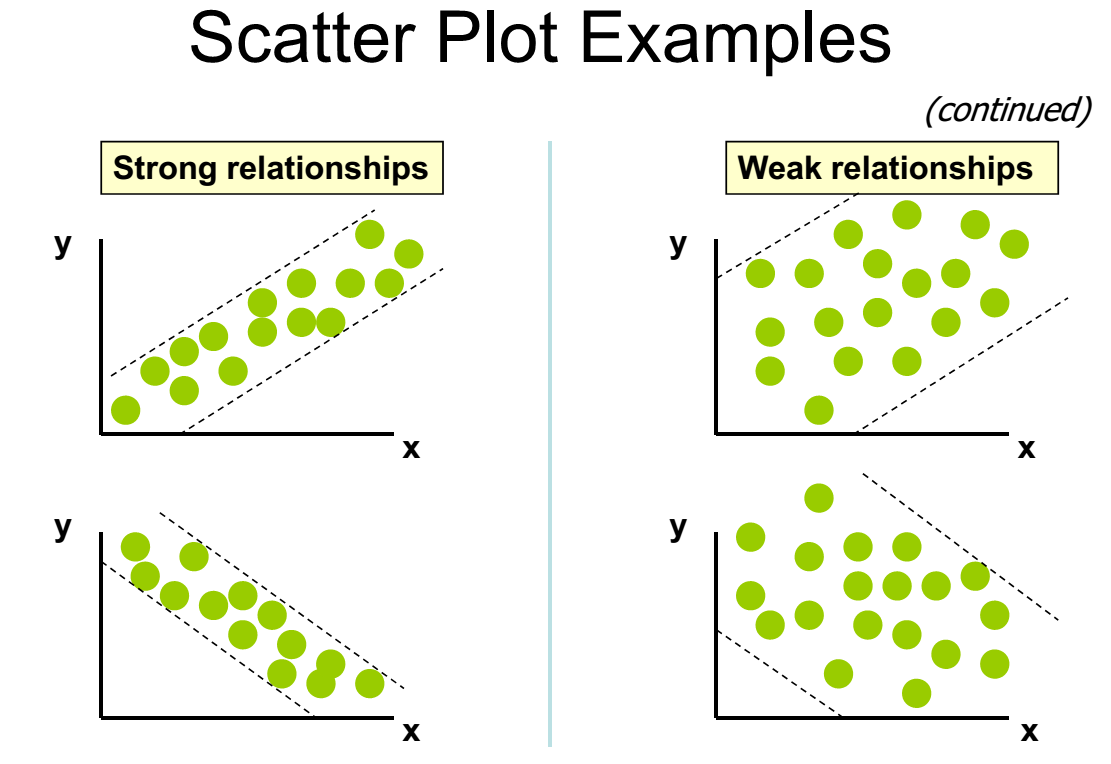
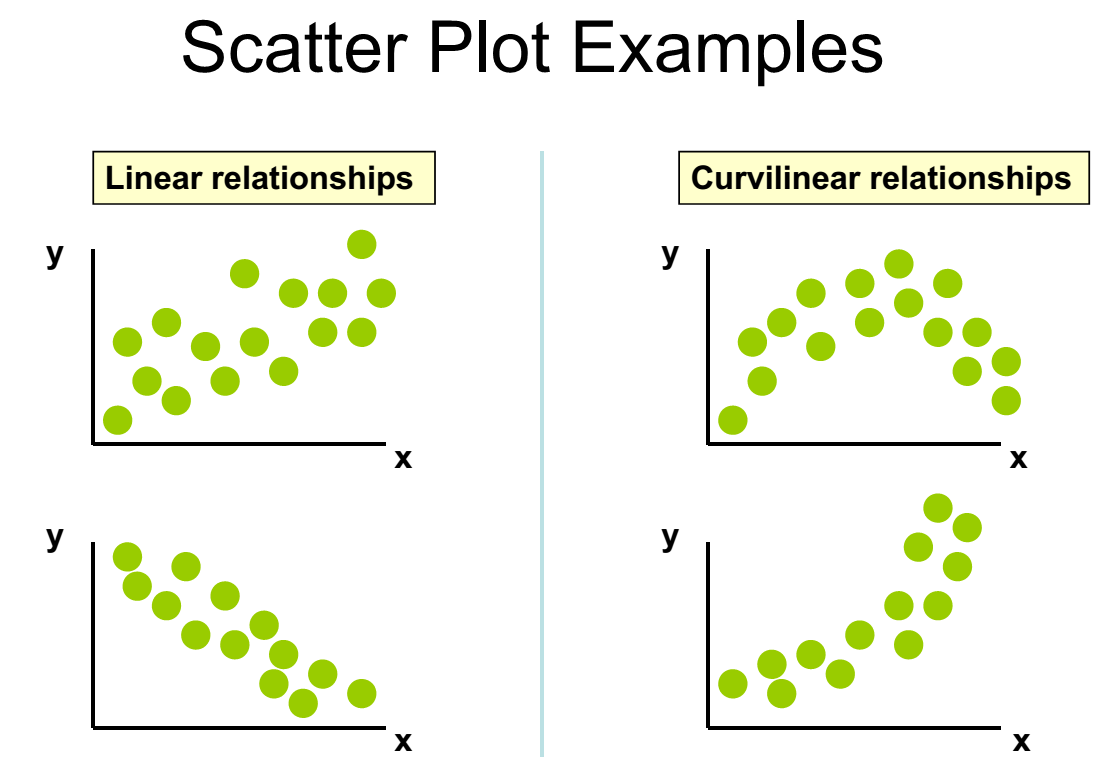
Since we have no tied ranks, correction is not required. Referencing to H-table we note that for three samples of sizes, n1=5, n2=4 and n3=3, then the critical value at 0.05 significance level is 5.656. This value is greater than the computed H value, and the null hypothesis cannot be rejected, implying teff yield does not appear to be affected by soil type in as far as the data in or sample shows.

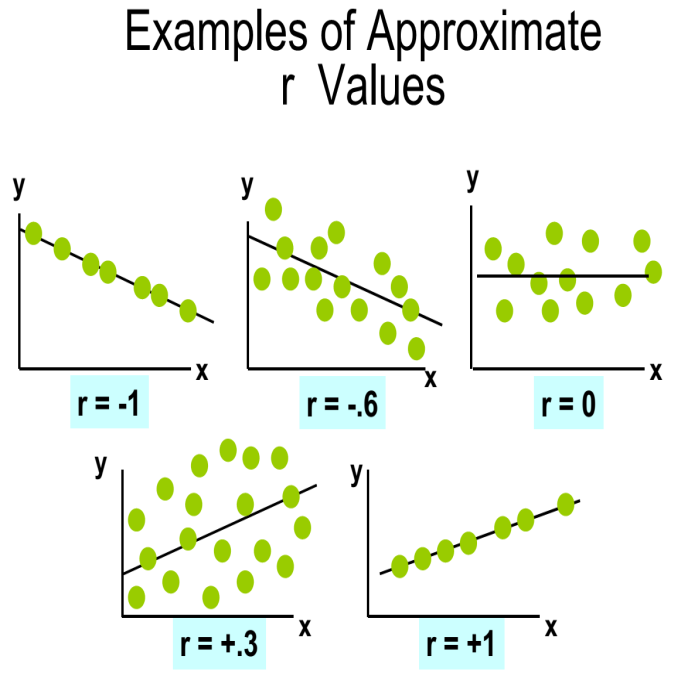
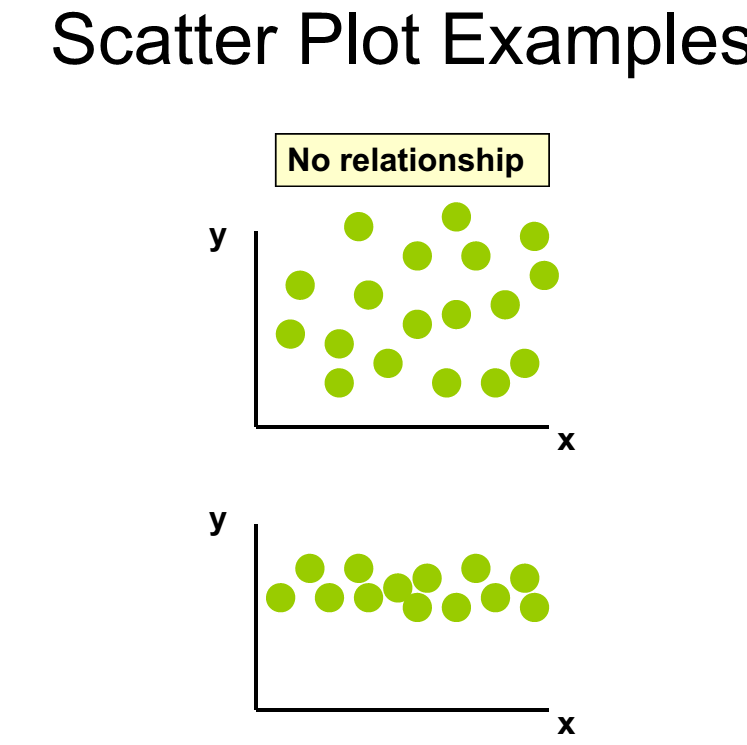
***Chapter Five***

# Correlation and Regression Analysis

## Correlation

Correlation analysis is used to measure strength of the association (linear relationship) between two variables. Only concerned with strength of the relationship. No causal effect is implied. A scatter plot (or scatter diagram) is used to show the relationship between two variables





Correlation coefficient r measures the strength of relationship between one dependant variable and one or more independent variables.

***Properties of “r”***

* Unit free
* Range between -1 and 1
* The closer to -1, the stronger the negative linear relationship
* The closer to 1, the stronger the positive linear relationship
* The closer to 0, the weaker the linear relationship
* r has the same sign as the slope of the regression (best fit) line
* r does not change if the independent (x) and dependent (y) variables are interchanged
* r does not change if the scale on either variable is changed.

The following table reveals the effect (or degree) of coefficient or correlation.

|  |  |  |
| --- | --- | --- |
| Degrees | Positive | Negative |
| Perfect correlation | + 1 | -1 |
| High degree | + 0.75 to + 1 | - 0.75 to -1 |
| Moderate degree | + 0.25 to + 0.75 | - 0.25 to - 0.75 |
| Low degree | 0 to 0.25 | 0 to - 0.25 |
| Absence of correlation | Zero | 0 |

### Types of Correlation

1. Pearson product-moment correlation coefficient
2. Spearman Rank- Correlation coefficient

#### Pearson product-moment correlation coefficient

Pearson product-moment correlation coefficient is a measure of the correlation (linear dependence) between two variables X and Y. Pearson correlation coefficient is calculated using the following formula



**Or the algebraic equivalent**

***Where:***

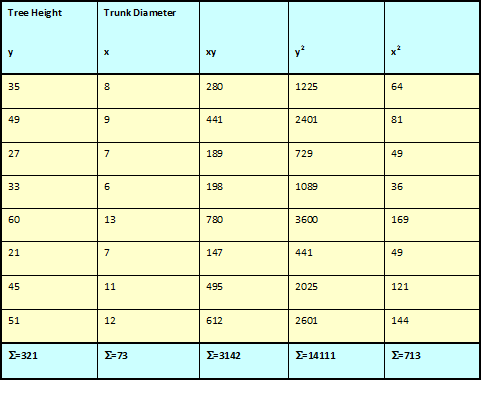
r = Sample correlation coefficient

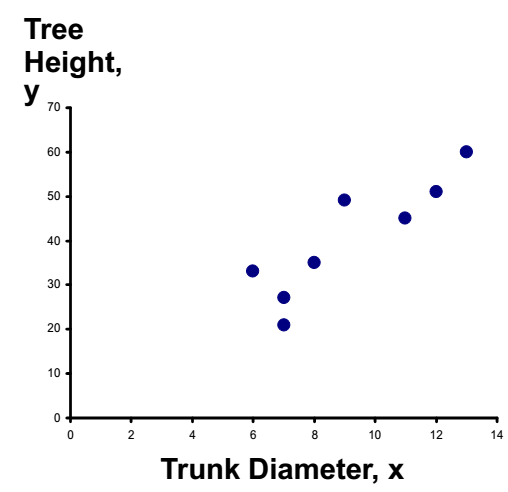
n = Sample size

x = Value of the independent variable

y = Value of the dependent variable

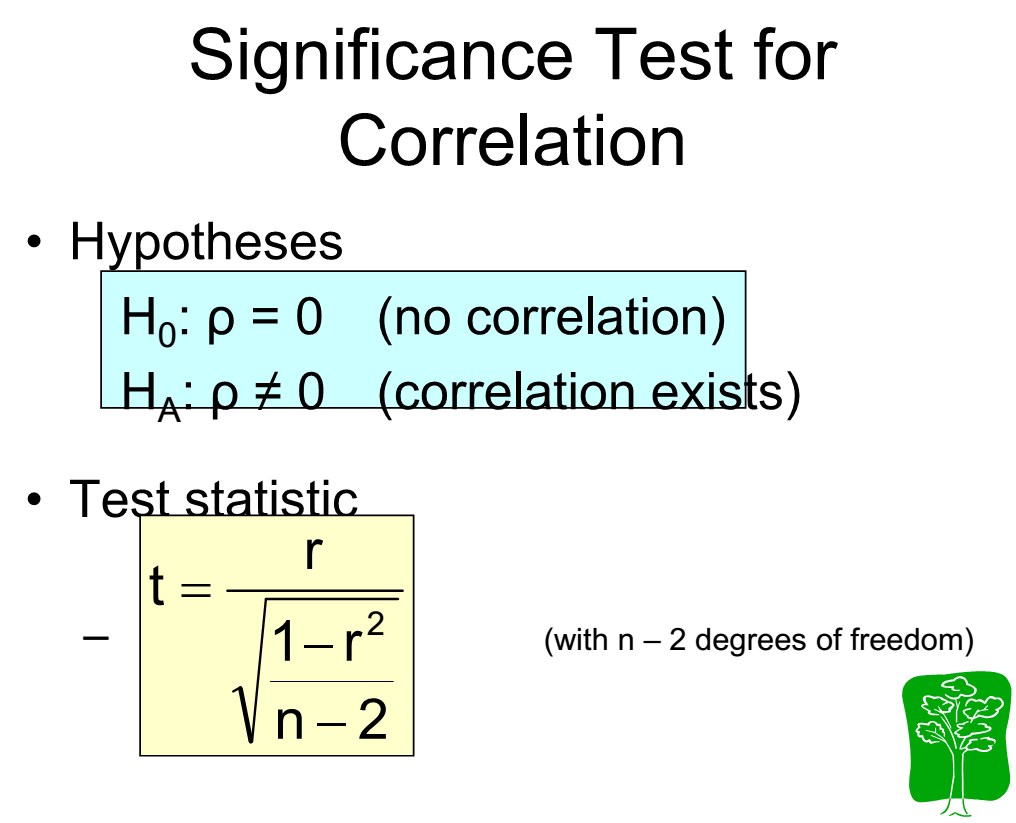
**Example Calculate “r” for the following data**

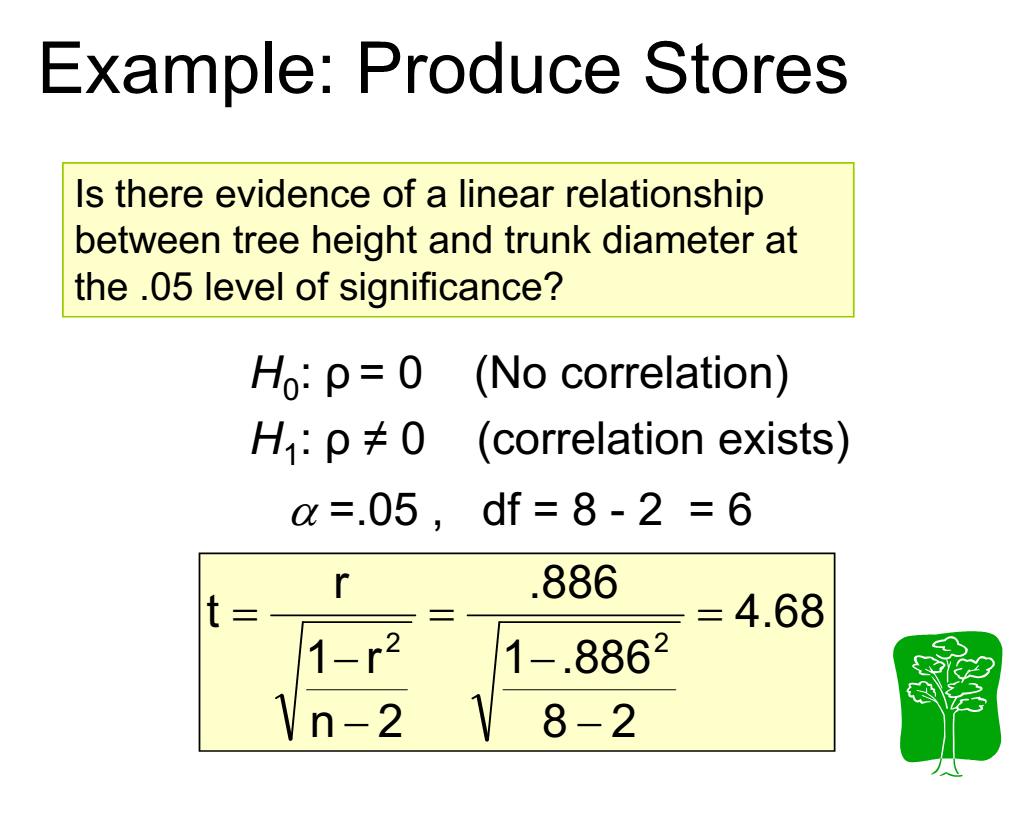




**r = 0.886** → relatively strong positive

linear association between x and y





#### 

#### Spearman Rank- Correlation coefficient

* This method is based on the ranks of the items rather than on their actual values.
* The advantage of this method over the others in that it can be used even when the actual values of items are unknown.

***The formula is***

***R=*** http://www.pinkmonkey.com/studyguides/subjects/stats/chap6/Image114.gif

**Where;**

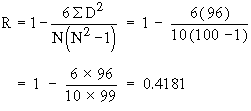
R = Rank correlation coefficient

D = Difference between the ranks of two items

N = the number of observations.

**Example:** calculate correlation coefficient to check whether the farmers’ rank in Adoption of extension service (R1) have correlation with their wealth rank (R2).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Farmers** | **Adoption of extension service(R1)** | **Wealth rank (R2)** | **R1 - R2 D** | **(R1 - R2 )2 D2** |
| 1 | 1 | 3 | -2 | 4 |
| 2 | 3 | 1 | 2 | 4 |
| 3 | 7 | 4 | 3 | 9 |
| 4 | 5 | 5 | 0 | 0 |
| 5 | 4 | 6 | -2 | 4 |
| 6 | 6 | 9 | -3 | 9 |
| 7 | 2 | 7 | -5 | 25 |
| 8 | 10 | 8 | 2 | 4 |
| 9 | 9 | 10 | -1 | 1 |
| 10 | 8 | 2 | 6 | 36 |
|  |  |  | S D = 0 | S D2 = 96 |



* Therefore the association between the two ranks is moderate

## Coefficient of determination(CD)

R = the magnitude of the relationship between the dependent variable and the best linear combination of the predictor variables. R2 = the proportion of variation in Y accounted for by the set of independent variables (X’s). The percentage of variance in one variable that is accounted for by the variance in the other variable. CD= square of correlation coefficient

Example: r= yield. Rainfall=0.7

49% of the variance in yield can be explained by the variance in rainfall amount

## Regression Analysis

***Regression analysis:*** is a set of statistical procedures used to explain or predict the values of a quantitative dependent variable based on the values of one or more independent variables.

### Types of Regression Models



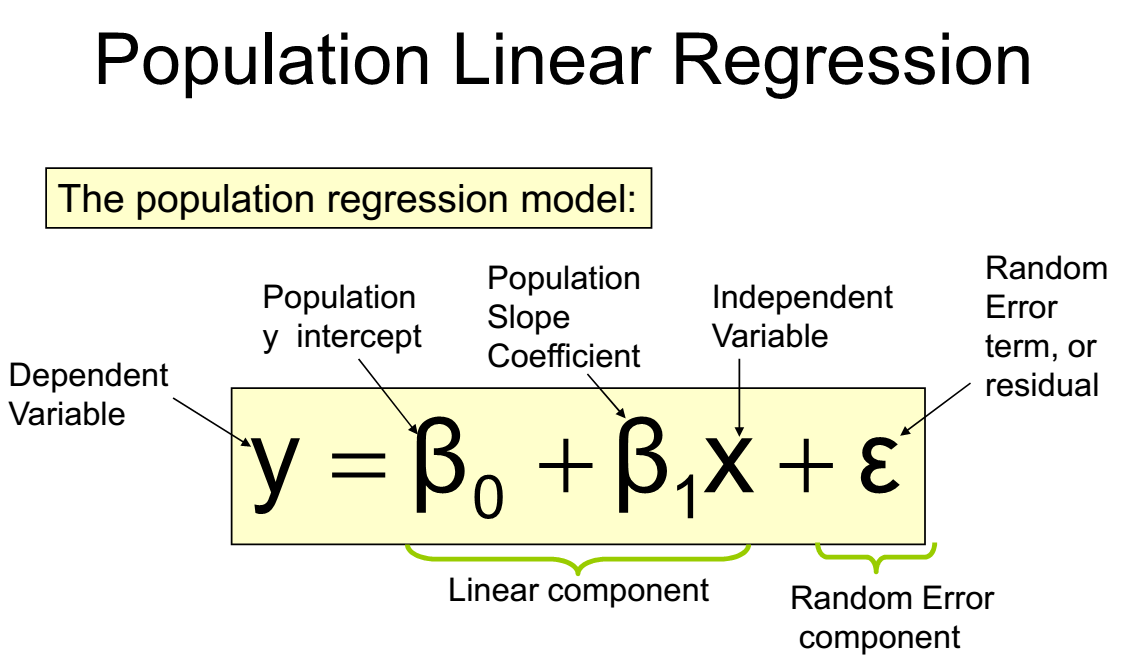
#### Simple regression

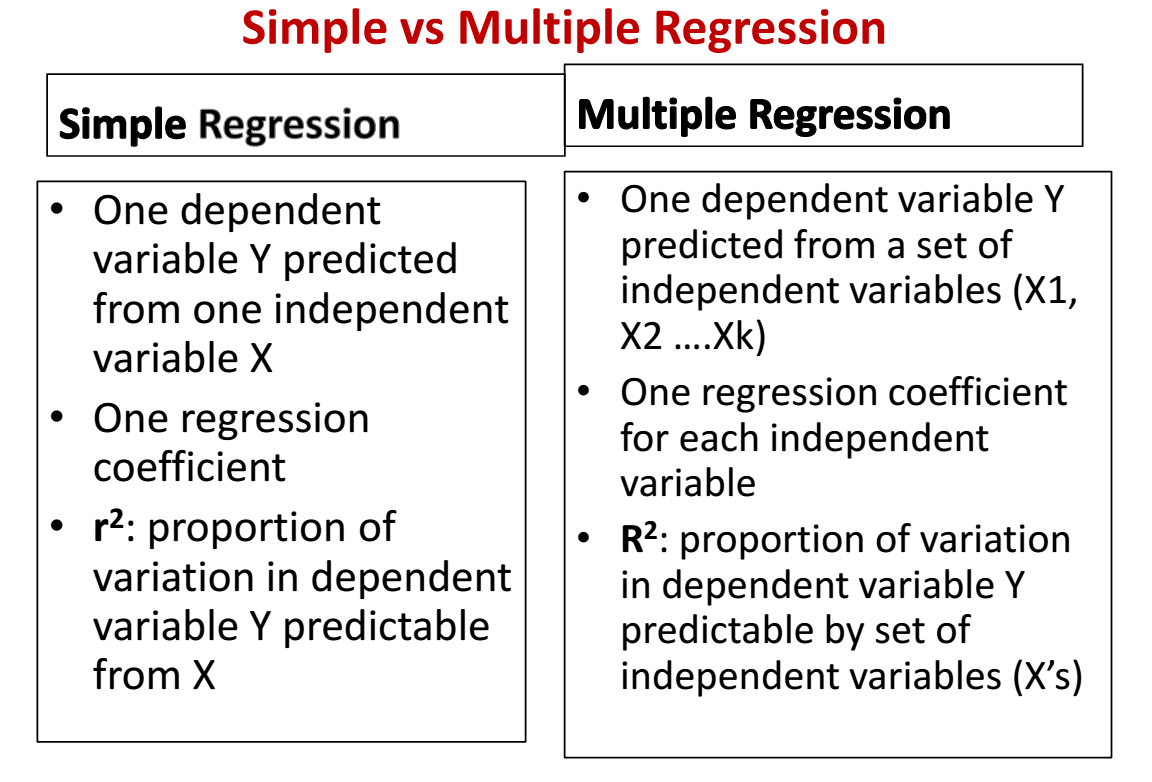
In ***simple regression***, there is one quantitative dependent variable and one independent variable. The model

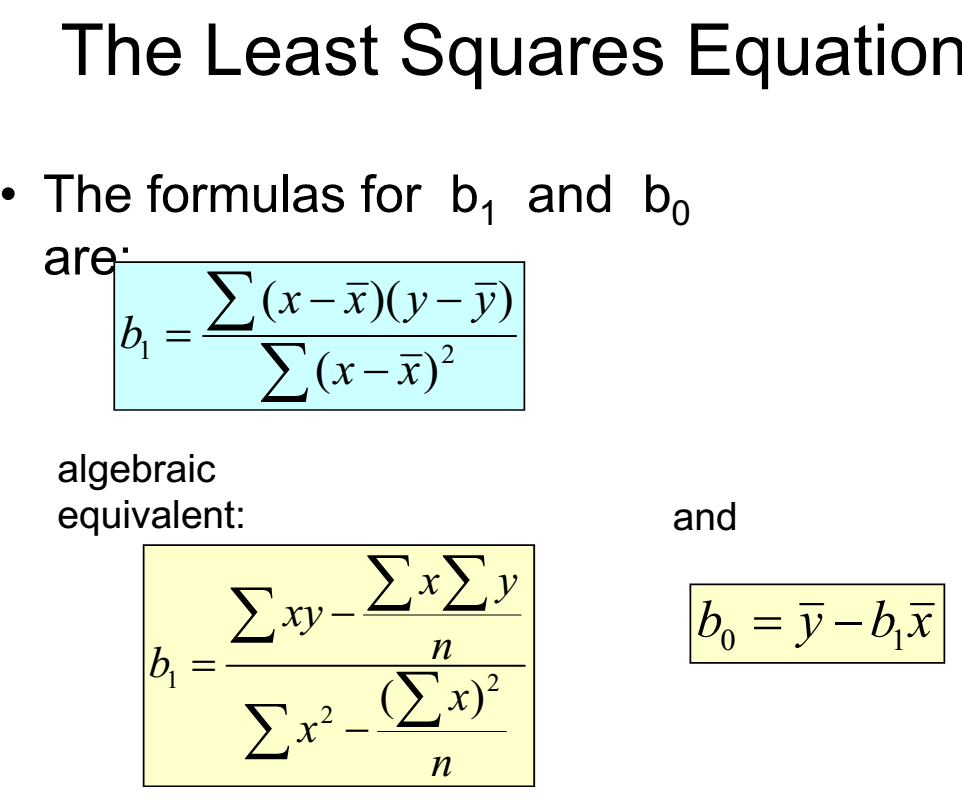


In ***multiple regression***, there is one quantitative dependent variable and two or more independent variables.









**Interpretation of the Slope and the Intercept**

* b0 is the estimated average value of y when the value of x is zero
* b1 is the estimated change in the average value of y as a result of a one-unit change in x

**Finding the Least Squares Equation**

The coefficients b0 and b1 will usually be found using computer software, such as Excel or SPSS. Other regression measures will also be computed as part of computer-based regression analysis.

**Example for simple Linear Regression**

The weekly advertising expenditure (x) and weekly sales (y) are presented in the following table.



* + From previous table we have:



* The least squares estimates of the regression coefficients are:





* The estimated regression function is:



* + This means that if the weekly advertising expenditure is increased by $1 we would expect the weekly sales to increase by $10.8.

#### Multiple regression

**4.2.1.2 Multiple linear regressions**

**What is Multiple Linear Regressions**

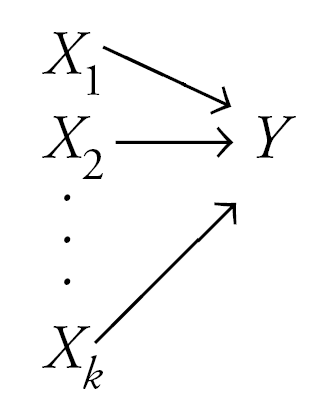
Multiple Regressions is a statistical method for estimating the relationship between a dependent variable and two or more independent (or predictor) variables.

**Why is this important?**

Behavior is rarely a function of just one variable, but is instead influenced by many variables. So the idea is that we should be able to obtain a more accurate predicted score if using multiple variables to predict our outcome.

With multiple linear regressions, we form a 'linear combination' of multiple variables to best predict an outcome, and then we assess the contribution that each predictor variable makes to the equation. For example my research question might be “How much does an independent variable contribute to explaining dependent variable after the effect of another independent variable is taken into account?”

**In short multiple regressions** simultaneously consider the influence of multiple explanatory variables on a response variable Y. The intent is to look at the independent effect of each variable.



**Multiple linear regression model equation**

**Y = A + B1X1 + B2X2 … + BnXn**

Y= dependent variable or the variable to be predicted.

Xs = the independent or predictor variables

A = “raw score equations” include a constant

B = B weights; or partial regression coefficients.

The Bs show the relative contribution of their independent variable on the dependent variable when controlling for the effects of the other predictors

***Multiple linear regressions (***MLR) ***out put***

The following notions are essential for the understanding of MLR output: R2, adjusted R2, constant, b coefficient, beta, F-test, t-test. For MLR “R2” (the coefficient of multiple determination) is used rather than “r” (Pearson’s correlation coefficient) to assess the strength of this more complex relationship (as compared to a bivariate correlation).

**Adjusted R square and b coefficient**

The adjusted R2 adjusts for the inflation in R2 caused by the number of variables in the equation. As the sample size increases above 20 cases per variable, adjustment is less needed (and vice versa). b coefficient measures the amount of increase or decrease in the dependent variable for a one-unit difference in the independent variable, controlling for the other independent variable(s) in the equation.

Ideally, the independent variables are uncorrelated. Consequently, controlling for one of them will not affect the relationship between the other independent variable and the dependent variable.

**Inter-correlation or collinearlity**

If the two independent variables are uncorrelated, we can uniquely partition the amount of variance in Y due to X1 and X2 and bias is avoided. Small inter-correlations between the independent variables will not greatly biased the b coefficients. However, large inter-correlations will biased the b coefficients and for this reason other mathematical procedures are needed.

Computation of the estimates by hand is tedious. They are ordinarily obtained using a regression computer program. Standard errors also are usually part of output from a regression program. Regression coefficients are interpreted as “the change in the mean dependent variable for each unit change in the corresponding independent variable, all other variables held constant”.

**Calculating Coefficients and building multiple linear regression equation**

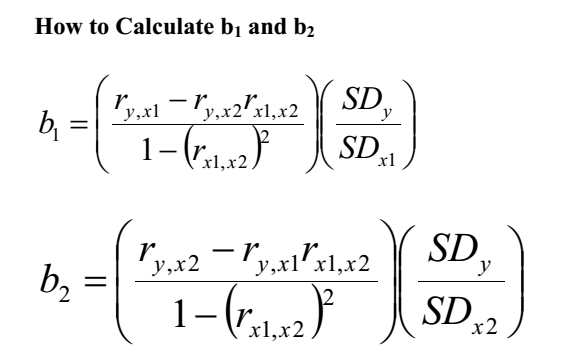
**The formula for two independent variables is**



***Where ;***

*  =A predicted value of Y (which is your dependent variable)
* A= the y intercept
* B1=the change in y for each 1 increment change in x1(in our case …
* B2=the change in y for each 1 increment change in x2(in

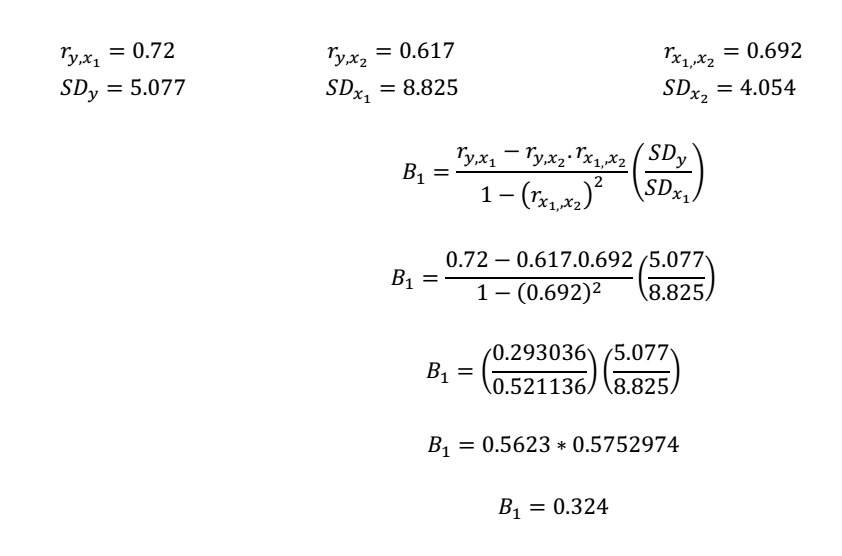
X=an X score (X is your Independent Variable) for which you are trying to predict a value of Y

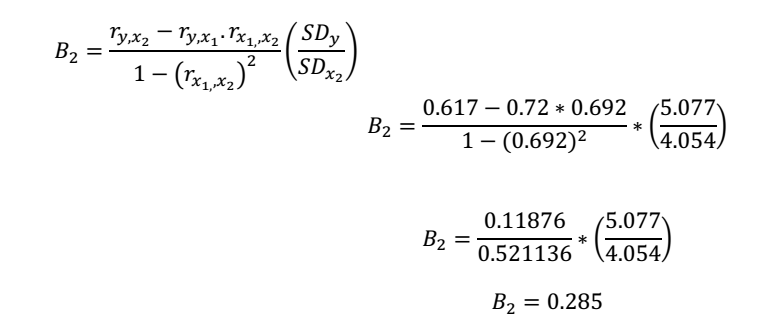


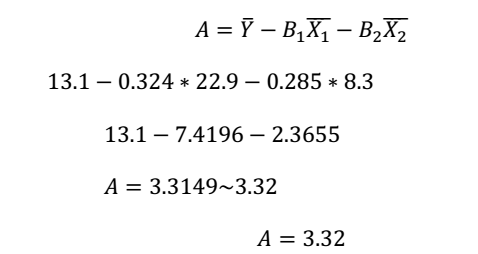
***Example:*** Construct the multiple linear regression equation for the following data(yield as dependant variable and amount of manure and degree of slope as independent variables).

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No** | **yield(kuintal) (y)** | **amount of manure (kg)(x1)** | **slope(%)(x2)** |
| 1 | 5 | 10 | 3 |
| 2 | 6 | 17 | 4 |
| 3 | 8 | 12 | 5 |
| 4 | 6 | 13 | 7 |
| 5 | 7 | 14 | 8 |
| 6 | 8 | 15 | 4 |
| 7 | 11 | 23 | 3 |
| 8 | 12 | 26 | 6 |
| 9 | 14 | 13 | 10 |
| 10 | 15 | 25 | 13 |
| 11 | 17 | 27 | 9 |
| 12 | 10 | 29 | 8 |
| 13 | 17 | 15 | 6 |
| 14 | 18 | 19 | 5 |
| 15 | 19 | 34 | 7 |
| 16 | 20 | 30 | 11 |
| 17 | 15 | 32 | 12 |
| 18 | 17 | 33 | 15 |
| 19 | 18 | 35 | 16 |
| 20 | 19 | 36 | 14 |

***Solution***

******

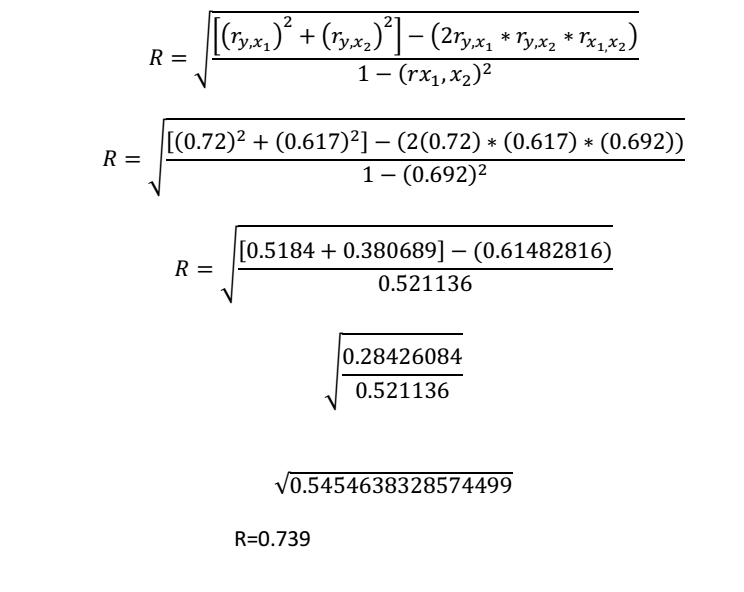
******



* Then the equation for the association of yield with manure and slope(steepness of land) is the following



* *The correlation coefficient is computed as follow ;*



**References**

Allan G.bluman((2007).elementary statistics :step by step(6th edition),Mcgraw-Hill College; 6th  
edition. Allan G. Bluman(2009). Elementary Statistics a Step by Step Approach 7th edition

McGraw-Hill, a business unit of The McGraw-Hill Companies, Inc., 1221 Avenue of the Americas, New York, NY 10020. Copyright

Aadne Aasland(2008). A User Manual for SPSS Analysis: Norwegian institute of urban and regionalresearch

Eshetu Wencheko(2000).Introduction to statistics ,department of statistics ,Addis Ababa university

Jack Levin(1983).Elementary statistics in social research(3rd edition), Pearson; 3 edition (February 25,2010)

J.S CHANDAN.(1998) statistics for business and economics, Vikas Publishing House PVT LTD. (New Delhi )

K.Briggs(1989).practical geography: presentation and analysis, Hodder Arnold H&S (March 1, 1989),**ISBN-13:** 978-0340415672

Mc Graw –hill companies (1998). Elementary statistics: step by stepm approach

M.S.Gill(2003).Quantitative methods in geographic research

M .Z.A .khan(1998).text book of practical Geography Concept Publishing Company, 1998, ISBN 8170226821, 9788170226826

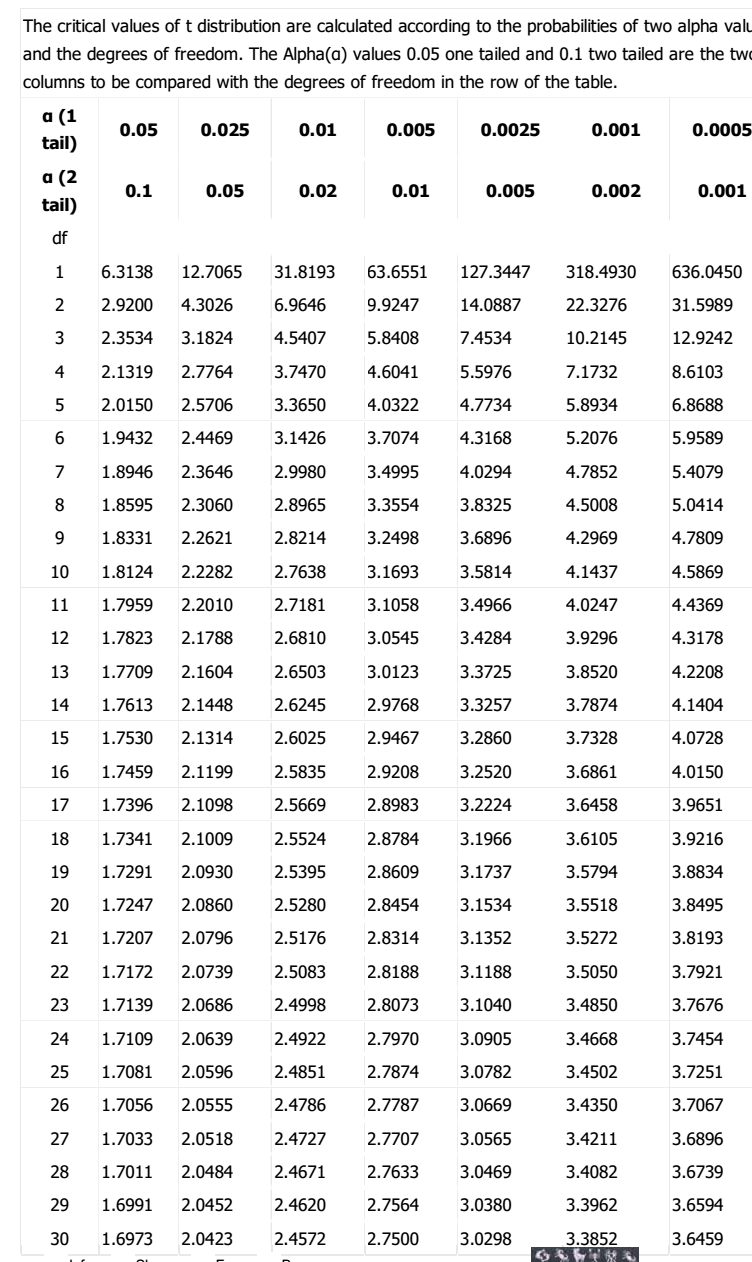
Nabendu Palsahadeb Sarkar(2005).statistis: concepts and applications

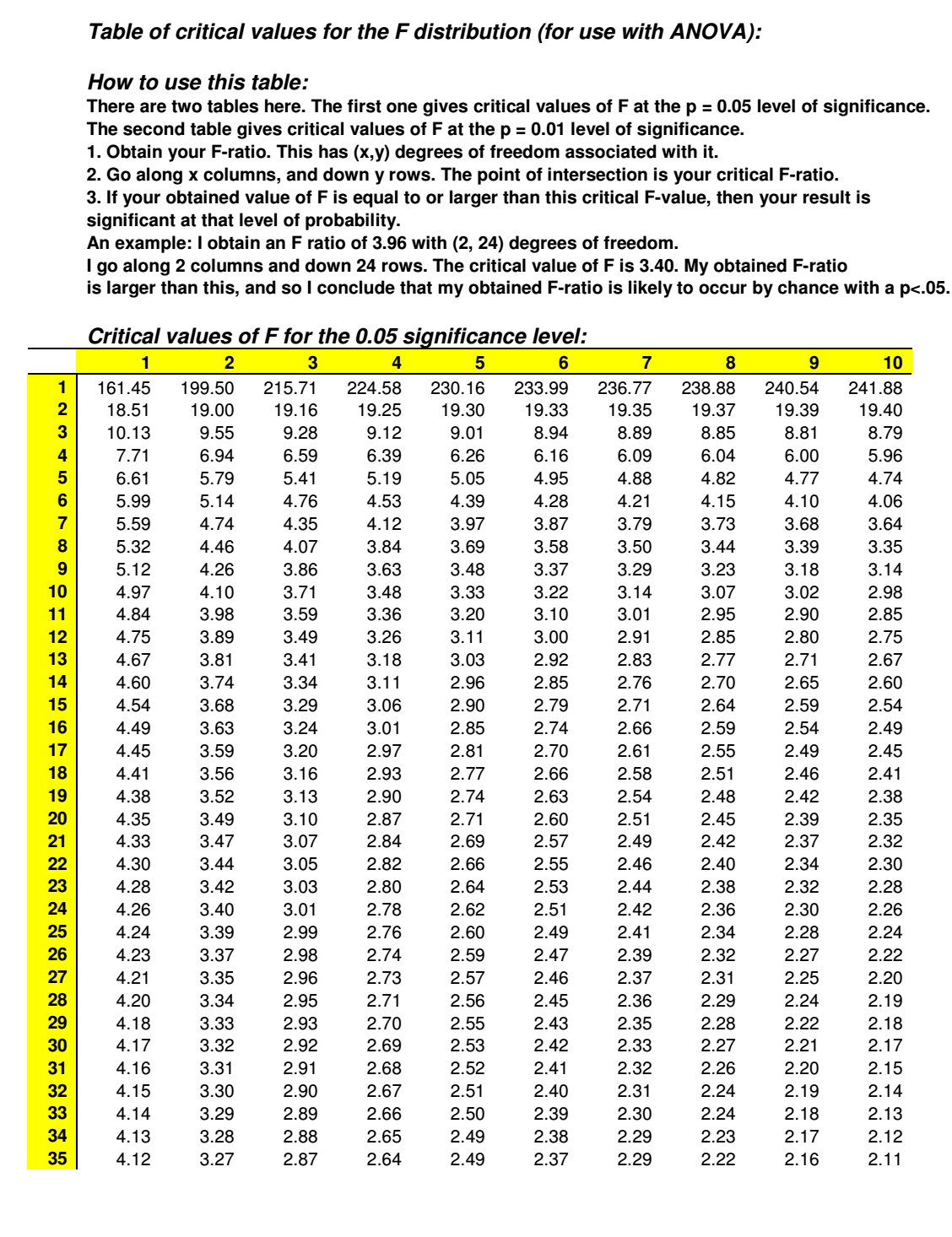
Sheldon M. Ross (2010) introduction to statistics(3RD edition) Academic Press ISBN 0080922104, 9780080922102

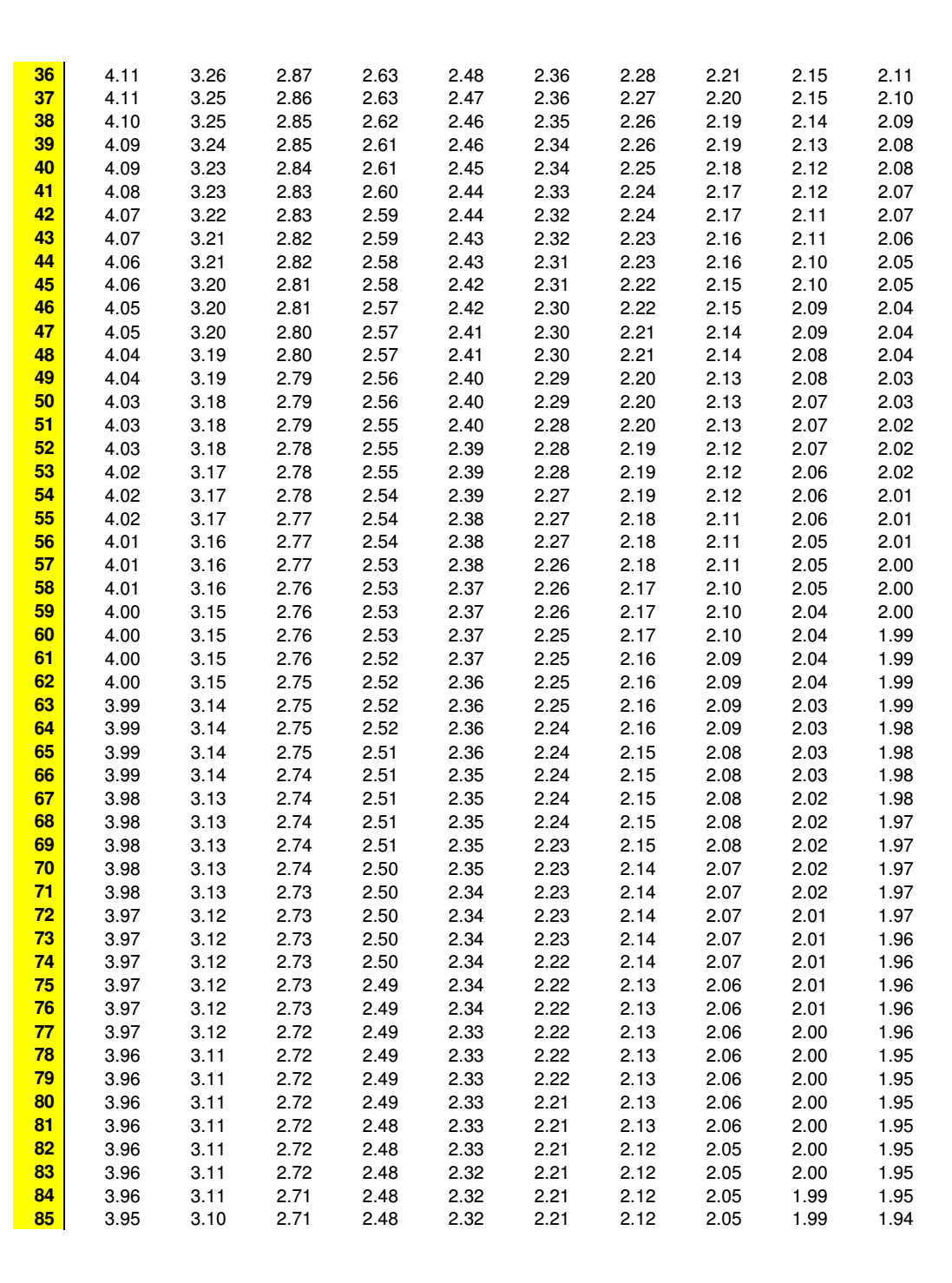
SPSS Inc.(2003).*SPSS 12.0 Brief Guide*

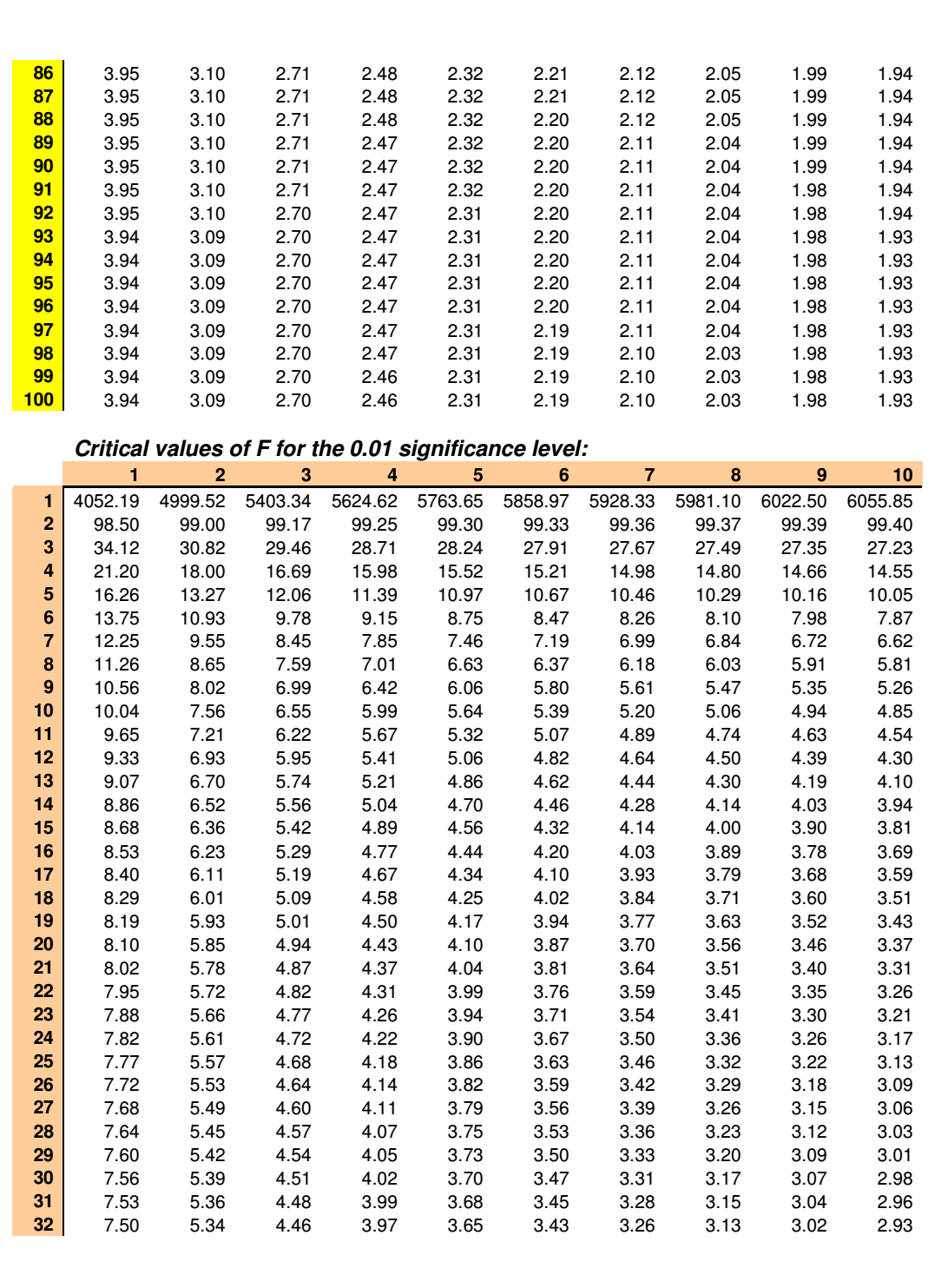
Vijay Gupta(1999). SPSS for Beginners,Published by VJBooks Inc.

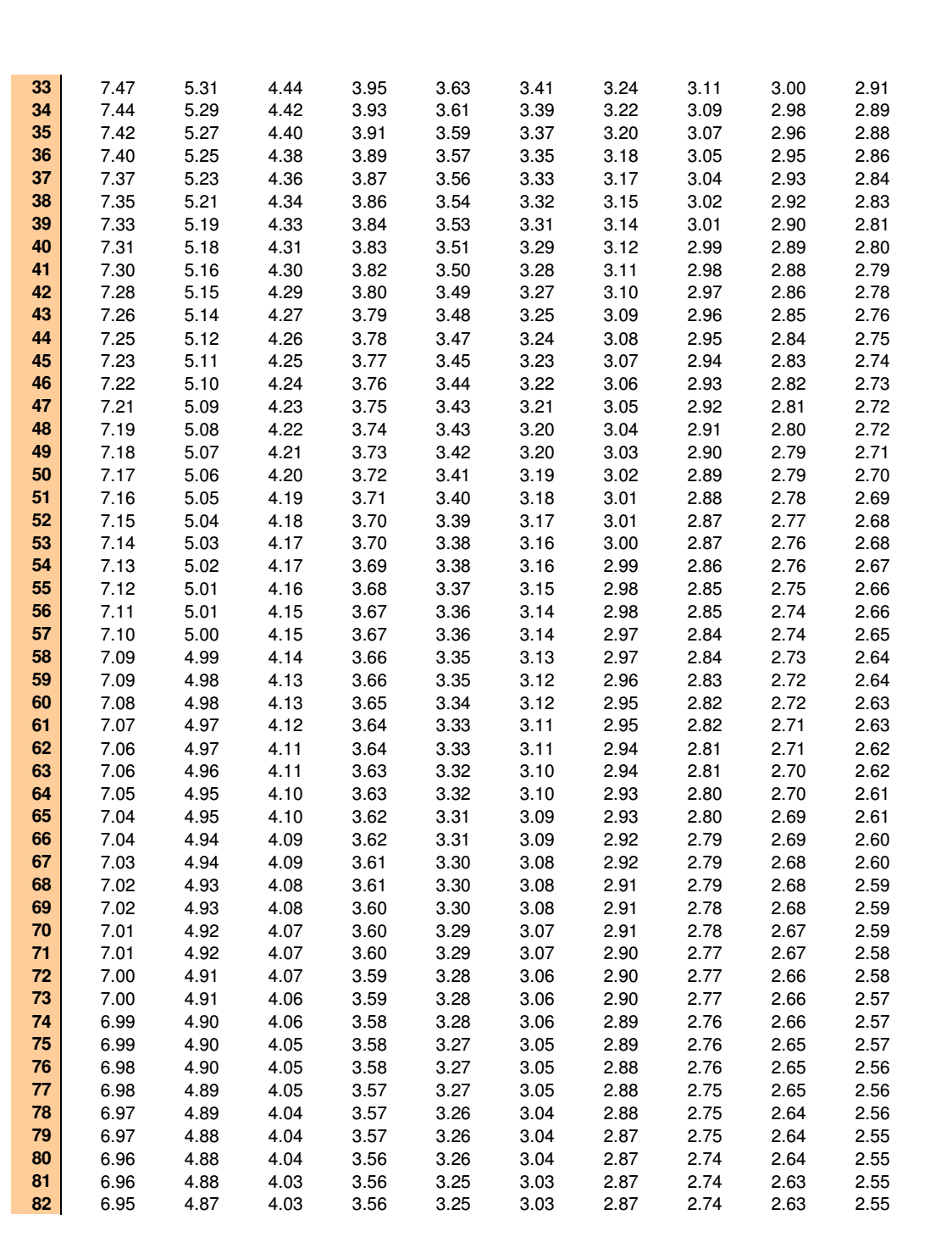
Vilay K.Roha tgi and A.K.MD.ehsanes saleh(2001) introduction to probability and statistics(2nd edition )

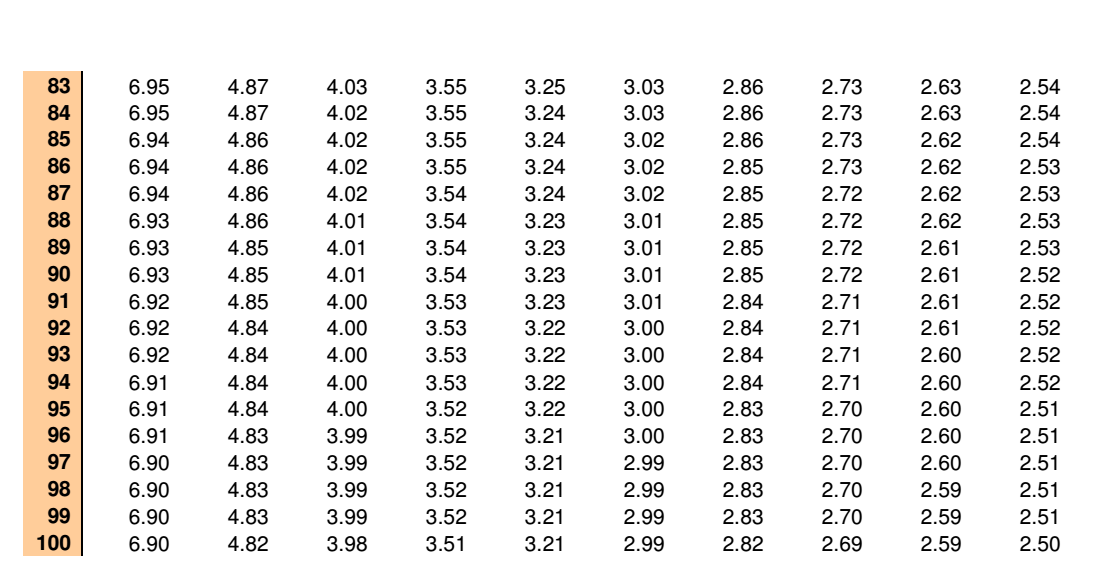


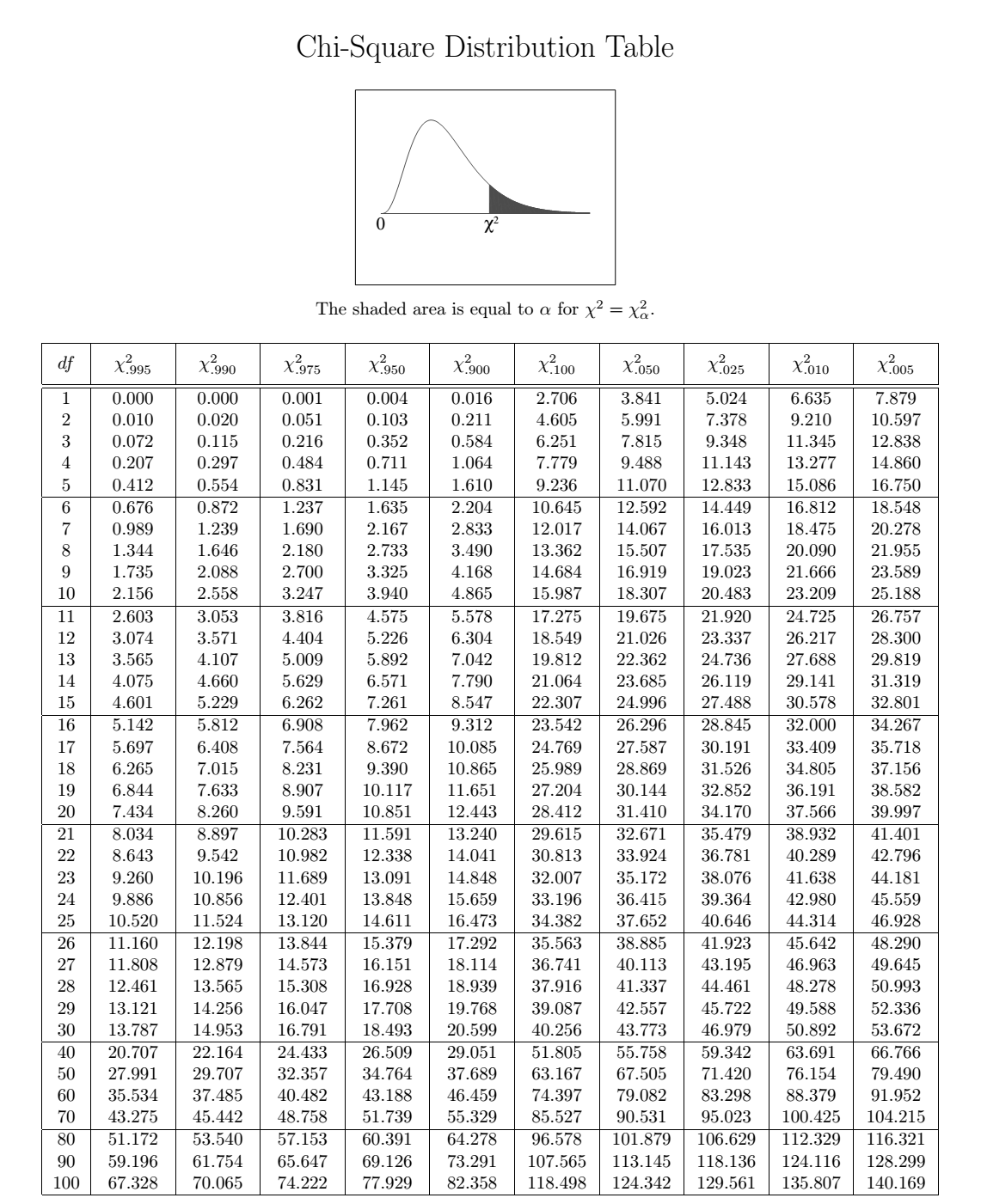


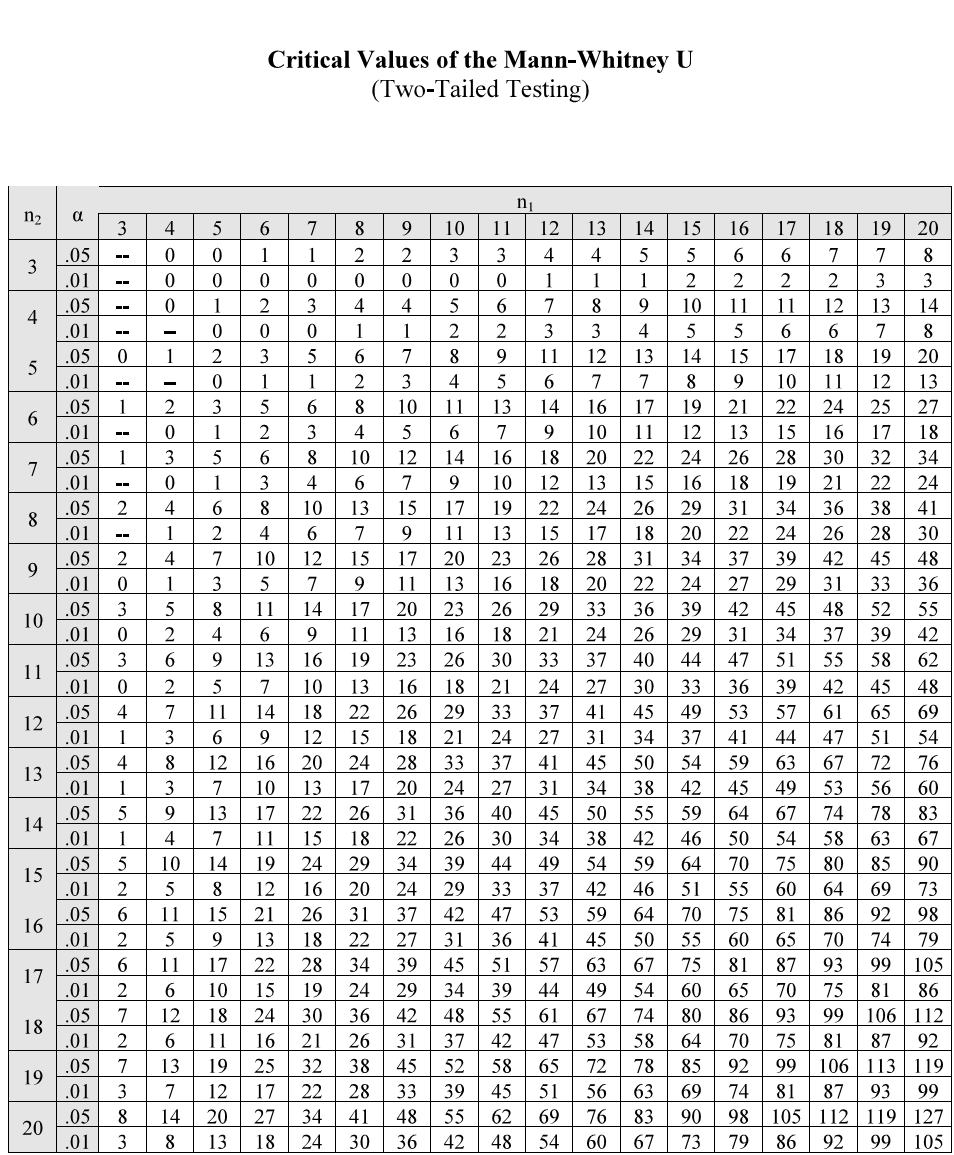


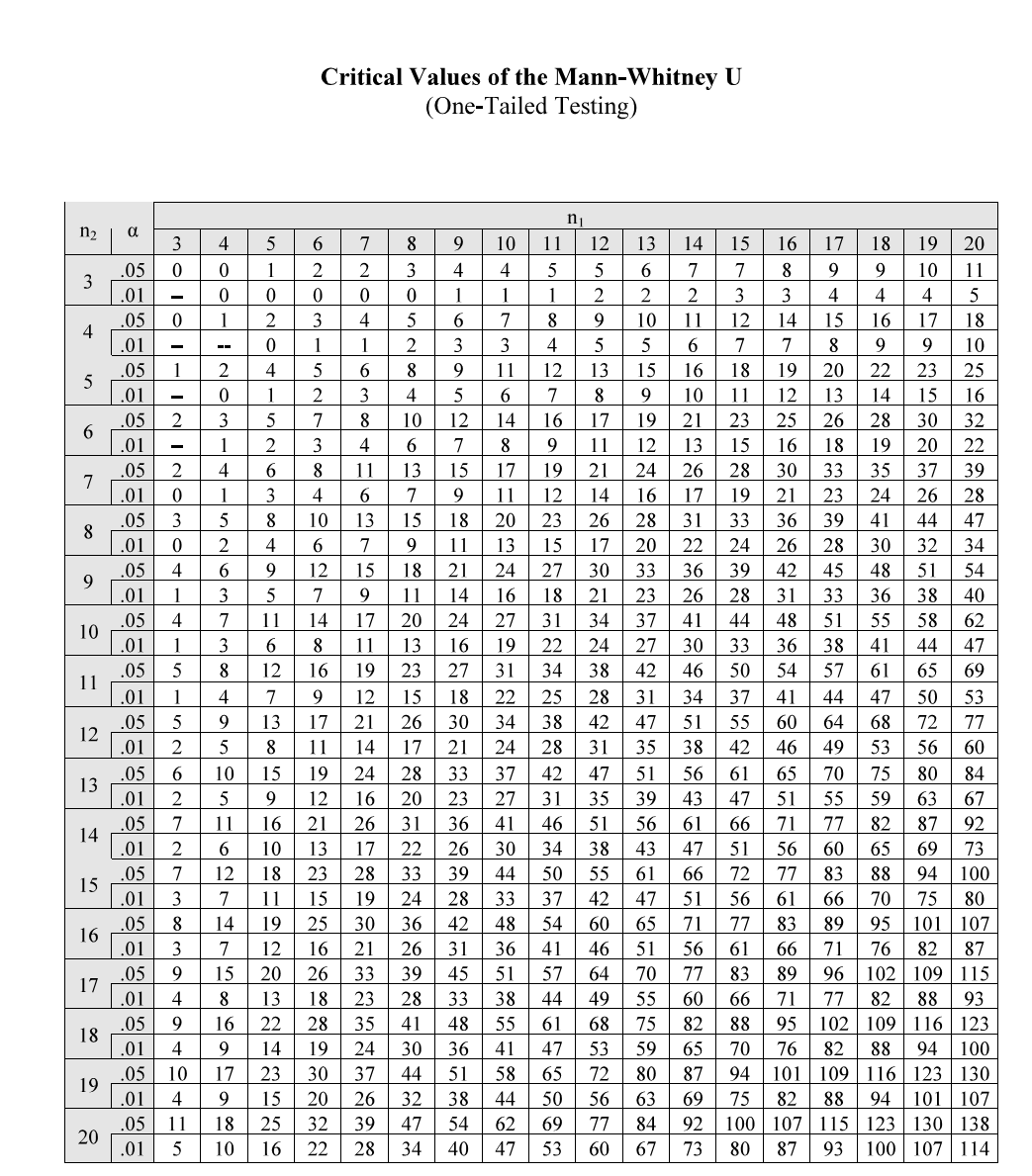


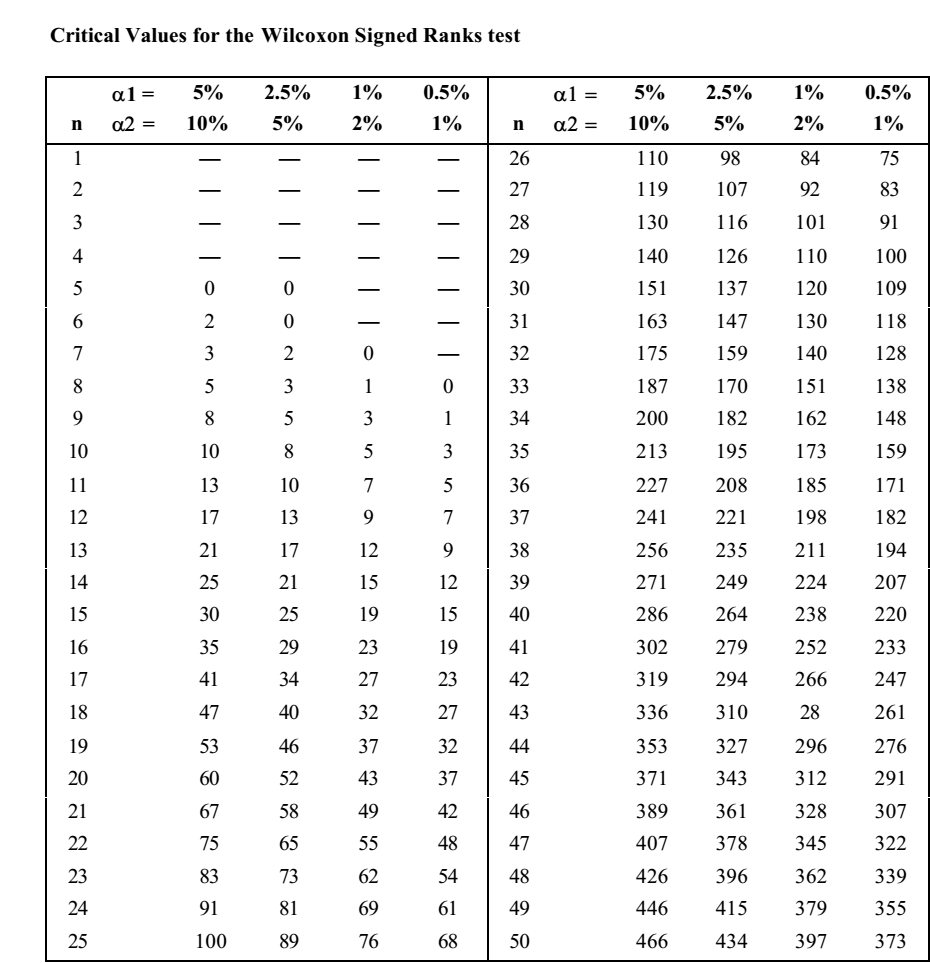


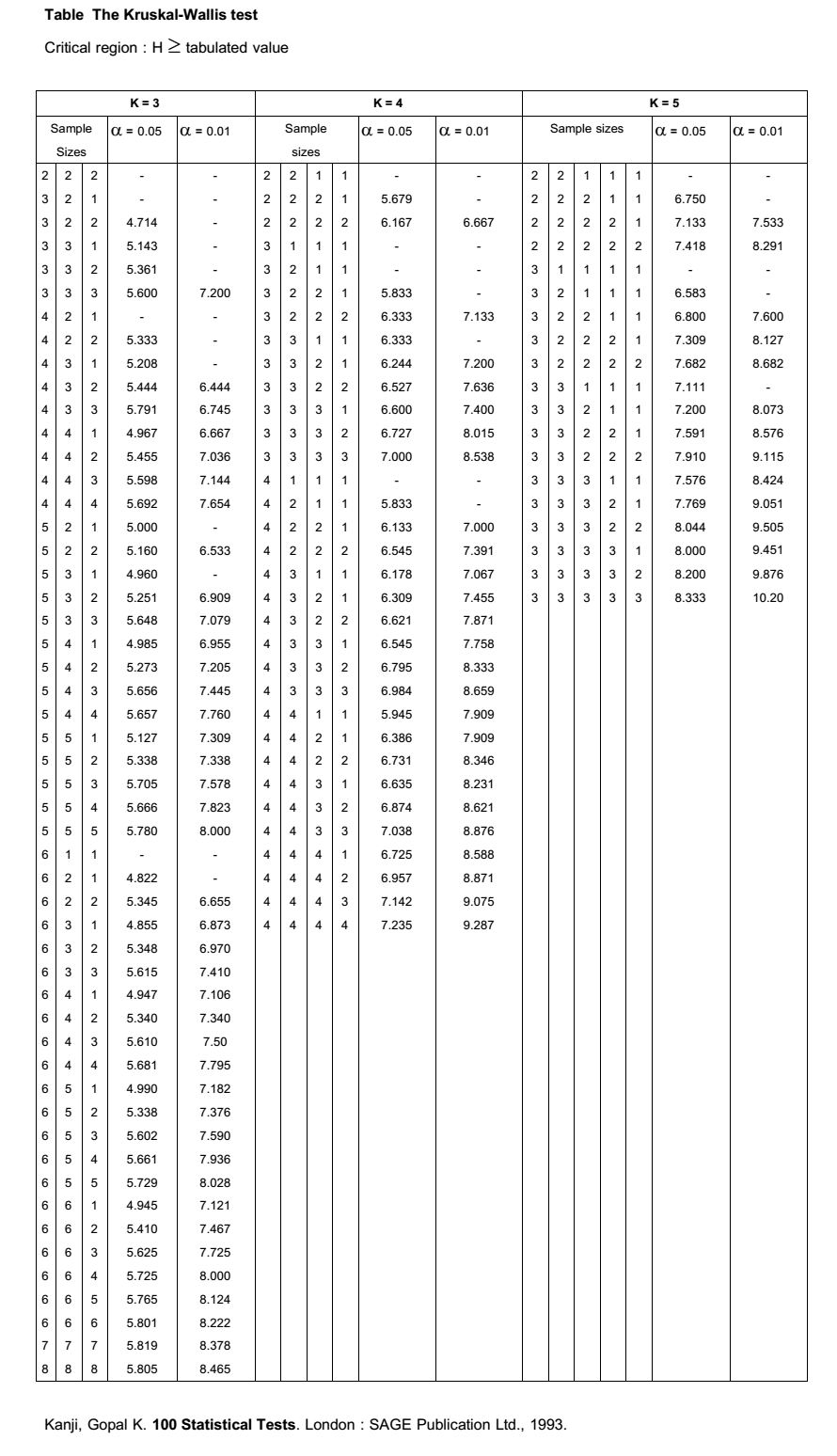












1. [↑](#footnote-ref-1)