"Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write." H.G. Wells

Syllabus F.Y. B.Sc Statistics Course I

SEMESTER I

COURSE USST101 DESCRIPTIVE STATISTICS-1

UNIT – I

Types of Data and Tabulation : Concepts of statistical population and sample .

Different types of scales nominal, ordinal , interval and ratio.

Types of Data from a population : Qualitative and quantitative data; Time series data; discrete and continuous data.

Primary data : Concept of a questionnaire and a schedule and distinction between them. Verification for consistency.

Construction of tables with one, two or three factors of classification. Independence and Association for 2 X 2 table using Yule's coefficient of association and coefficient of colligation.

Requirements of good statistical table.

UNIT – II

Diagrammatic representation using bar diagrams and pie chart.

Univariate frequency distribution of discrete and continuous variables. Cumulative frequency distribution.

Graphical representation of frequency distribution by Histogram, frequency polygon, Stem and leaf diagram and Cumulative frequency polygon.

Bivariate frequency distribution. Marginal and Conditional frequency distributions.

UNIT – III

Measures of central tendency: Concept of central tendency or location of data.

Measures of central tendency or location.: Arithmetic mean (simple and weighted), Combined mean, Geometric mean and Harmonic mean.

Median, Quartiles, Deciles, Percentiles, Mode.

Uses of Mean, Medan and Mode their Merits and demerits.

Requirements of good average.

Statistics is a science which deals with collection, tabulation , presentation, analysis, and interpretation of the data.

Data values are the values of the characteristics under study.

QUALITATIVE DATA & QUANTITATIVE DATA

The characteristics of the unit which can be measured numerically, is called 'Quantitative characteristics' or 'variable', and the values of such characteristics are called quantitative data. For ex. ; height, weight, marks scored, percentage of marks.

Continuous variable and Discrete variable.

A continuous variable is one for which there is a possible value between any two possible values. If the set of all possible values of variable is an interval then variable is continuous. Eg. Consider plant height as a variable, then within the range 35c.m. to 36c.m. there are infinite possible values such as 35.05c.m., 35.588c.m., 35.975c.m. ….

But when we consider number of leaves as a variable then this variable takes only certain values. Such variable is known as discrete variable. If the set of all possible values of the random variable is either finite or countable infinite then variable is discrete. Generally discrete variable take integer values, but it not necessarily so. Eg if we consider the ratio of number of wings to number of legs of insects then variable have values 0, 2/6, 4/6.

The characteristics of the unit which can not be measured numerically but can be classified into different groups is called 'Qualitative characteristics' or 'attribute', and the data representing such classification are called qualitative data. For ex sex, religion, class obtained.

Another factor that determines the amount of information that can be provided by a variable is its "type of measurement scale." Specifically variables are classified as (a) Nominal, (b) Ordinal, (c) Interval or (d) Ratio.

Nominal scale allow for only qualitative classification. That is, they can be measured only in terms of whether the individual items belong to some distinctively different categories, but we cannot quantify or even rank order those categories. For example, all we can say is that 2 individuals are different in terms of variable A (e.g., they are of different race), but we cannot say which one "has more" of the quality represented by the variable. Typical examples of nominal variables are gender, race, color, city, etc.

Ordinal scale allow us to rank order the items we measure in terms of which has less and which has more of the quality represented by the variable, but still they do not allow us to say "how much more." A typical example of an ordinal variable is the socioeconomic status of families.

For example, we know that upper-middle is higher than middle but we cannot say that it is, for example, 18% higher. Also this very distinction between nominal, ordinal, and interval scales itself represents a good example of an ordinal variable. For example, we can say that nominal measurement provides less information than ordinal measurement, but we cannot say "how much less" or how this difference compares to the difference between ordinal and interval scales.

Interval variables allow us not only to rank order the items that are measured, but also to quantify and compare the sizes of differences between them. For example, temperature, as measured in degrees Fahrenheit or Celsius, constitutes an interval scale. We can say that a temperature of 40 degrees is higher than a temperature of 30 degrees, and that an increase from 20 to 40 degrees is twice as much as an increase from 30 to 40 degrees.

Ratio variables are very similar to interval variables; in addition to all the properties of interval variables, they feature an identifiable absolute zero point, thus they allow for statements such as x is two times more than y. Typical examples of ratio scales are measures of time or space. For example, as the Kelvin temperature scale is a ratio scale, not only can we say that a temperature of 200 degrees is higher than one of 100 degrees, we can correctly state that it is twice as high. Interval scales do not have the ratio property. Most statistical data analysis procedures do not distinguish between the interval and ratio properties of the measurement scales.

Sources of data

1. **Primary Source** A source is primary for one who creates source by collecting the data by direct investigation . Such data are called primary data .

2 **Secondary sources** Collected data ,which is made available as a reference to others, if anyone else makes use of it ,then source of data becomes secondary source for that user.

Collection of primary data : Primary data are collected by survey . Surveys are of two types 1. Census survey or population survey 2. Sample survey

If we collect the data on each unit of the population then survey is known as population survey. But population survey is time consuming, expensive. Some times it is not advisable as it results in destruction of the population .hence we go for sample survey. From a given population we select a set of objects and the observations are made on these selected objects or units. This is known as sample survey.

SAMPLING TECHNIQUES

Depending on the nature of the population we adopt different sampling procedures namely 1) Simple random sampling 2) Stratified random sampling.

Most commonly used technique is simple random sampling.(SRS)

In SRS each unit from the population has an equal and independent chance of being included in the sample.(independent means selection of any unit from the population must in no way influence the selection of any other unit.)

To draw a random sample we use 1) Lottery method 2) Method of random numbers.

Suppose population contains N objects, then we assign numbers 1 to N for these units. Let the number of sample units be selected be n.

In lottery method we write these N numbers on N identical slips after well shuffling we select n slips , then we consider the units corresponding to the numbers on the selected slips . This constitutes a random sample. If we select the slips one after another then we get a sample without replacement, but if we replace the slip after noting the number on it before selecting the next slip then we get a sample with replacement.

Lottery method is time consuming hence we use the alternative method the method of random numbers. From random number table we take n random numbers serially, divide each number by N get remainder r, select the population unit corresponding to this r. if we want sample without replacement then see that there is no repetition of random number.

Now random numbers are available on calculator and computer.

In with replacement sampling a unit from the population can be selected more than once but in without replacement sampling the unit can be selected at most once.

When the population is heterogeneous to reduce sampling error we divide the population in to homogeneous subgroups known as strara, then from each statum we select a random sample, such sample is known as stratified random sample and the sampling procedure as stratified random sampling.

A measure based on population units is known as parameter and measure for sample units is called a statistic. By evaluating statistic we can talk about parameter.

Collection of Primary Data

Primary data are always collected from the source. It is collected either by the investigator himself or through his agents. There are different methods of collecting primary data. Each method has its relative merits and demerits. The investigator has to choose a particular method to collect the information. The choice to a large extent depends on the preliminaries to data collection some of the commonly used methods are discussed below.

1. Direct Personal observation:

This is a very general method of collecting primary data. Here the investigator directly contacts the informants, solicits their cooperation and enumerates the data. The information are collected by direct personal interviews.

The novelty of this method is its simplicity. It is neither difficult for the enumerator nor the informants. Because both are present at the spot of data collection. This method provides most accurate information as the investigator collects them personally. But as the investigator alone is involved in the process, his personal bias may influence the accuracy of the data. So it is necessary

that the investigator should be honest, unbiased and experienced. In such cases the data collected may be fairly accurate. However, the method is quite costly and time-consuming. So the method should be used when the scope of enquiry is small.

2 .Indirect Oral Interviews :

This is an indirect method of collecting primary data. Here information are not collected directly from the source but by interviewing persons closely related with the problem. This method is applied to apprehend culprits in case of theft, murder etc. The information relating to one's personal life or which the informant hesitates to reveal are better collected by this method. Here the investigator prepares 'a small list of questions relating to the enquiry. The answers (information) are collected by interviewing persons well connected with the incident. The investigator should cross-examine the informants to get correct information.

This method is time saving and involves relatively less cost.

The accuracy of the information largely depends upon the integrity of the investigator. It is desirable that the investigator should be experienced and capable enough to inspire and create confidence in the informant to collect accurate data.

3 Mailed Questionnaire method:

This is a very commonly used method of collecting primary data. Here information are collected through a set of questionnaire. A questionnaire is a document prepared by the investigator containing a set of questions. These questions relate to the problem of enquiry directly or indirectly. Here first the questionnaires are mailed to the informants with a formal request to answer the question and send them back. For better response the investigator should bear the

postal charges. The questionnaire should carry a polite note explaining the aims and objective of the enquiry, definition of various terms and concepts used there. Besides this the investigator should ensure the secrecy of the information as well as the name of the informants, if required.

Success of this method greatly depends upon the way in which the questionnaire is drafted. So the investigator must be very careful while framing the questions. The questions should be

i) Short and clear

ii) Few in number

iii) Simple and intelligible

iv) Corroboratory in nature or there should be provision for cross check

v) Impersonal, non-aggressive type

vi) Simple alternative, multiple-choice or open-end type

(a) In the simple alternative question type, the respondent has to choose between alternatives such as 'Yes or No', 'right or wrong' etc.

(b) In the multiple choice type, the respondent has to answer from any of the given alternatives.

(c) In the Open-end or free answer questions the respondents are given complete freedom in answering the questions.

The questionnaire method is very economical in terms of time, energy and money. The method is widely used when the scope of enquiry is large. Data collected by this method are not affected by the personal bias of the investigator. However the accuracy of the information depends on the cooperation and honesty of the informants. This method can be used only if the informants are cooperative, conscious and educated. This limits the scope of the method.

4. Schedule Method:

In case the informants are largely uneducated and non-responsive data cannot be collected by the mailed questionnaire method. In such cases, schedule method is used to collect data. Here the questionnaires are sent through the enumerators to collect information. Enumerators are persons appointed by the investigator for the purpose. They directly meet the informants with the questionnaire. They explain the scope and objective of the enquiry to the informants and solicit

their cooperation. The enumerators ask the questions to the informants and record their answers in the questionnaire and compile them. The success of this method depends on the sincerity and efficiency of the enumerators. So the enumerator should be sweet-tempered, good-natured, trained and well-behaved.

Schedule method is widely used in extensive studies. It gives fairly correct result as the enumerators directly collect the information. The accuracy of the information depends upon the honesty of the enumerators. They should be unbiased. This method is relatively more costly and time-consuming than the mailed questionnaire method.

5. From Local Agents:

Sometimes primary data are collected from local agents or correspondents. These agents are appointed by the sponsoring authorities. They are well conversant with the local conditions like language, communication, food habits, traditions etc. Being on the spot and well acquainted with the nature of the enquiry they are capable of furnishing reliable information.

The accuracy of the data collected by this method depends on the honesty and sincerity of the agents as they actually collect the information from the spot. Information from a wide area at less cost and time can be collected by this method. The method is generally used by government agencies, newspapers, periodicals etc. to collect data.

Information are like raw materials or inputs in an enquiry. The result of the enquiry basically depends on the type of information used. Primary data can be collected by employing any of the above methods. The investigator should make a rational choice of the methods to be used for collecting data.

Secondary data is the data that have been already collected by and readily available from other sources. Such data are cheaper and more quickly obtainable than the primary data and also may be available when primary data can not be obtained at all.

Advantages of Secondary data

1. It is economical. It saves efforts and expenses.
2. It is time saving.
3. It helps to make primary data collection more specific since with the help of secondary data, we are able to make out what are the gaps and deficiencies and what additional

information needs to be collected.
4. It helps to improve the understanding of the problem.
5. It provides a basis for comparison for the data that is collected by the researcher.

Disadvantages of Secondary Data

1. Secondary data is something that seldom fits in the framework of the marketing research factors. Reasons for its non-fitting are:-
    a. Unit of secondary data collection-Suppose you want information on disposable income, but the data is available on gross income. The information may not be same as we require.
    b. Class Boundaries may be different when units are same.
    c. Thus the data collected earlier is of no use to you.
2. Accuracy of secondary data is not known.
3. Data may be outdated.

Evaluation of Secondary Data

Because of the above mentioned disadvantages of secondary data, we will lead to evaluation of secondary data. Evaluation means the following four requirements must be satisfied:-

1. **Availability-** It has to be seen that the kind of data you want is available or not. If it is not available then you have to go for primary data.
2. **Relevance-** It should be meeting the requirements of the problem. For this we have two criterion:-
    a. Units of measurement should be the same.
    b. Concepts used must be same and currency of data should not be outdated.
3. **Accuracy-** In order to find how accurate the data is, the following points must be considered: -
    a. Specification and methodology used;
    b. Margin of error should be examined;
    c. The dependability of the source must be seen.
4. **Sufficiency-** Adequate data should be available.

TABULATION: The process of placing classified data into tabular form is known as tabulation.

A table is a logical and symmetric arrangement of statistical data in rows and columns.

The main objectives of tabulation are:

- It simplifies complex data and the data presented are easily understood.
- It facilitates comparison of related facts.
- It facilitates computation of various statistical.

- It presents facts in minimum possible space.
- Moreover, the needed information can be easily located.
- Tabulated data are good for references and to present the information

Components of Table

An ideal table should consist of the following main parts:

- Table number For easy reference assign number for the table
- Title of the table : Each table must have title . It must be brief and self explanatory.
- Captions or column headings
- Stubs or row heading
- Body of the table It is most important part of the table It contains numerical information including row and column totals.
- Footnotes normally written at the bottom of the table.
- Sources of data: It mentions the source of the information used in the table.

Table may be simple, double or complex depending upon the type of classification.

When data is collected on attributes then data can be tabulated .The table should be precise, easy to understand and self explanatory.

When the data is classified with respect to one characteristic then the table is known as one way table

For ex. Data is collected on number of students in the degree college for the year 2005-06 , then we can tabulate the information according to classes.

No. of students in the college for the year 2005-06

F. Y. B. Sc.	S. Y. B. Sc.	T. Y. B. Sc.	TOTAL

When the data is classified with respect to two characteristics then the table is known as two way table

For ex. Data is collected on number of students in the degree college for the year 2005-06 , then we can tabulate the information according to classes and sex.

No. of students in the college for the year 2005-06

CLASS → SEX |	F. Y. B. Sc.	S. Y. B. Sc.	T. Y. B. Sc.	Total
Male				
Female				

Total				

In the above data if we add one more characteristic namely category (open, reserved) , then we can tabulate the information according to classes , sex and category then we have to prepare three way table..

No. of students in the college for the year 2005-06

CLASS → SEX \ category		F. Y. B. Sc.	S. Y. B. Sc.	T. Y. B. Sc.	Total
Male	Open				
	Reserved				
Total					
Female	Open				
	Reserved				
Total					
Total	Open				
	Reserved				
Total					

Characteristic of good table:

A good statistical table should be precise and easy to understand. It should be self explanatory. Table should contain

- Title /Heading for the table
- Stubs/Captions: Row heading and column heading arranged according to alphabetical or chronological order according to order/importance.
- Body of the table is most important part. It contains numerical information
- It should contain meaningful row/column totals.
- It should contain head note/footnote if necessary.

Advantages of Tabulation

- It facilitates comparison relationship between different sets of data can be easily studied and compared.

- Simplification of complex data.
- It facilitates further statistical treatment.
- It provides unique identity to data collected.
- They are simple to conduct.
- It is only after tabulation that some vital omissions are detected.
- It easier to present the information in the form of graphs and diagrams.
- The needed information can be easily located.

Disadvantages of Tabulation

- It provides incorrect information while comparing with other data.
- It is time consuming. It takes long time to tabulate the information.
- Only people with practical experience and knowledge can construct good table.
- Difficulty in accumulation of sufficient data.
- Lack of description.

Theory of attributes:

An attribute means a quality on characteristic which are not related to quantitative measurements. Examples of attributes are health, honesty, blindness etc. They cannot be measured directly. The observer may find the presence or absence of these attributes. Statistics of attributes based on descriptive character

Association of attribute is studied by the presence or absence of a particular attribute. If only one attribute is studied, the population is divided into two classes according to its presence or absence and such classification is termed as division by dichotomy. If a class is divided into more than two scale-classes, such classification is called manifold classification. Positive class which denotes the presence of attribute is generally denoted by Roman letters generally A,B,….etc and the negative class denoting the absence of the attribute and it is denoted by the Greek letters a, b….etc For example, A represents the attribute ‘ Literacy’ and B represents ‘ Criminal’ . a and b represents the ‘ Illiteracy’ and ‘ Not Criminal’ respectively.

When with respect to a given characteristic we can classify the unit in to two classes such as possessing given attribute and not possessing given attribute then classification is known as dichotomous classification. Ex: Sex- male and female. Result – Pass and fail

We use capital letters A,B,C, to denote class corresponding to the attribute and Greek small letters α, β, γ .. to denote complimentary classes

Classes and Class frequencies:

Different attributes, their sub-groups and combinations are called different classes and the numbers of observations assigned to them are called their class frequencies. If two attributes are studied the number of classes will be 9. (i.e.) (A), (α), (B), (β), (AB),(A β), (α β), (α B) and N. The number of observations or units belonging to class is known as its frequency are denoted within bracket. Thus (A) stands for the frequency of A and (AB) stands for the number objects possessing the attribute both A and B. The contingency table of order (2X2) for two attributes A and B can be displayed as given below

	A	α	Total
B	(AB)	(α B)	(B)
β	(A β)	(α β)	(β)
Total	(A)	(α)	N

When we consider three attributes A, B, C, then we can display various combination by following table

		A	α	Total
B	C	(ABC)	(α BC)	(BC)
	γ	(AB γ)	(α B γ)	(Bγ)
	Total	(AB)	(α B)	(B)
β	C	(A β C)	(α β C)	(β C)
	γ	(A β γ)	(α β γ)	(β γ)
	Total	(A β)	(α β)	(β)
Total	C	(A C)	(α C)	(C)
	γ	(A γ)	(α γ)	(γ)
	Total	(A)	(α)	N

If the complete series of frequencies arrived at by noting n attributes is being tabulated, frequencies of the same order should be kept together. Those of the same order are best arranged by taking separately the set o r "aggregate" of frequencies, derivable from each positive class by substituting negatives for one or more of the positive attributea Thus the frequencies for the case of three attributes may conveniently be tabulated in the order—

Order 0. *N*

Order 1. *(A), (a) : (B),* (β): *(C),(γ)*

Order 2. *(AB), (Aβ), (αB),* (αβ): *(AC), (Aγ), (αC), (αγ): (BC), (Bγ), etc*

Order 3. *(ABC), (aBC), (AβC), (ABy), (aβC), (aBy), (Aβγ),*

Relationship between the class frequencies:

The frequency of a lower order class can always be expressed in terms of the higher order class frequencies.

i.e., $N = (A) + (\alpha) = (B) + (\beta)$

$(A) = (AB) + (A\beta)$

$(\alpha) = (\alpha B) + (\beta\alpha)$

$(B) = (AB) + (\alpha B)$

$(\beta) = (A\beta) + (\alpha\beta)$

If the number of attributes is n, then there will be 3^n classes and we have 2^n cell frequencies.

i.e., $N = (A) + (\alpha) = (B) + (\beta) = (C) + (\gamma)$

$(A) = (AB) + (A\beta) = (AC) + (A\gamma)$, $(\alpha) = (\alpha B) + (\beta\alpha) = (\alpha C) + (\alpha\gamma)$,

$(B) = (AB) + (\alpha B) = (BC) + (B\gamma)$, $(\beta) = (A\beta) + (\alpha\beta) = (\beta C) + (\beta\gamma)$

$(C) = (AC) + (\alpha C) = (BC) + (\beta C)$, $(\gamma) = (A\gamma) + (\alpha\gamma) = (B\gamma) + (\beta\gamma)$

$(AB) = (ABC) + (AB\gamma)$, $(A\beta) = (A\beta C) + (A\beta\gamma)$,

$(BC) = (ABC) + (\alpha BC)$, $(\beta C) = (A\beta C) + (\alpha\beta C)$

$(AC) = (ABC) + (A\beta C)$, $(A\gamma) = (AB\gamma) + (A\beta\gamma)$

$(\alpha B) = (\alpha BC) + (\alpha B\gamma)$, $(\alpha C) = (\alpha BC) + (\alpha\beta C)$

$(\alpha\beta) = (\alpha\beta C) + (\beta\alpha\gamma)$ $(\alpha\gamma) = (\alpha B\gamma) + (\alpha\beta\gamma)$

$(\beta\gamma) = (A\beta\gamma) + (\alpha\beta\gamma)$, $(B\gamma) = (AB\gamma) + (A\beta\gamma)$

Consistency of the data:

In order to find out whether the given data are consistent or not we have to apply a very simple test. The test is to find out whether any or more of the ultimate class-frequencies is negative or not. If none of the class frequencies is negative we can safely calculate that the given data are consistent (i.e the frequencies do not conflict in any way each other). On the other hand, if any of the ultimate class frequencies comes to be negative the given data are inconsistent.

When we have one attributes

1. $(A) \geq 0$
2. $(\alpha) \geq 0$ i.e, $(A) \leq N$

   $0 \leq (A) \leq N$

When we have two attributes

1. $(AB) \geq 0$
2. $(\alpha B) \geq 0$ i.e, $(AB) \leq (B)$

3 $(A\beta) \geq 0$ i.e. $(AB) \leq (A)$

4 $(\alpha\beta) \geq 0$ i.e. $(AB) \geq (A)+(B)-N$

$\text{Max}\{0, (A)+(B)-N\} \leq (AB) \leq \text{Min}\{(A), (B)\}$

When we have Three attributes

1 $(ABC) \geq 0$

2 $(A\beta C) \geq 0$ i.e. $(ABC) \leq (AC)$

3 $(AB\gamma) \geq 0$ i.e. $(ABC) \leq (AB)$

4 $(A\beta\gamma) \geq 0$ i.e. $(ABC) \geq (AB)+(AC)-(A)$

5 $(\alpha BC) \geq 0$ i.e, $(ABC) \leq (BC)$

6 $(\alpha B\gamma) \geq 0$ i.e, $(ABC) \geq (BC)+(AB)-(B)$

7 $(\alpha\beta C) \geq 0$ i.e. $(ABC) \leq (AC)+(BC)-(C)$

8 $(\alpha\beta\gamma) \geq 0$ i.e. $(ABC) \leq (AB)+(BC)+(AC)-(A)-(B)-(C)+N$

$\text{Max}\{0, (AB)+(AC)-(A), (BC)+(AB)-(B), (AB)+(BC)-(C)\} \leq (ABC)$

$\leq \text{Min}\{(AB), (BC), (AC), (AB)+(BC)+(AC)-(A)-(B)-(C)+N\}$

Example 1:

Given N = 2500, (A) = 420, (AB) = 85 and (B) = 670.Find the missing values.

Solution: We know

1 $N = (A) + (\alpha) = (B) + (\beta)$

2 $(A) = (AB) + (A\beta)$

3 $(\alpha) = (\alpha B) + (\alpha\beta)$

4 $(B) = (AB) + (\alpha B)$

5 $(\beta) = (A\beta) + (\alpha\beta)$

From (2) $420 = 85 + (A\beta)$ i.e. $(A\beta) = 420 - 85$, $(A\beta) = 335$

From (4) $670 = 85 + (\alpha B)$; $(\alpha B) = 670 - 85$; $(\alpha B) = 585$

From (1) $2500 = 420 + (\alpha)$; $(\alpha) = 2500 - 420$; $(\alpha) = 2080$

From (1) $(\beta) = 2500 - 670 = 1830$

From (3) $2080 = 585 + (\alpha\beta)$; $(\alpha\beta) = 1495$

Example 2:

Test the consistency of the following data with the symbols having their usual meaning.

i) N = 1000 (A) = 600 (B) = 500 (AB) = 50

ii) (A)=48; (AB)=50

Solution: i) Since $(\alpha\beta) = -50$, the given data is inconsistent

ii) Here (AB) >(A) but for consistency $(AB) \leq (A)$, hence data is inconsistent.

Independence of Attributes:

If the attributes are said to be independent the presence or absence of one attribute does not affect the presence or absence of the other. For example, the attributes skin colour and intelligence of persons are independent.

If two attributes A and B are independent then the actual frequency is equal to the expected frequency

(AB) = (A).(B)/ N

Similarly (α β) = (α).(β)/ N

Association of attributes:

Two attributes A and B are said to be associated if they are not independent but they are related with each other in some way or other.

The attributes A and B are said to be positively associated if (AB) > (A).(B)/ N

If (AB) < (A).(B)/N ,. Then they are said to be negatively associated.

Example 3:

Show that whether A and B are independent, positively associated or negatively associated.

(AB) = 128, (α B) = 384, (A β) = 24 and (α β) = 72

Solution:

(A) = (AB) + (A β) = 128 + 24 = 152 , (B) = (AB) + (α B)= 128 +384 = 512

(α) = (α B) + (α β)= 384 + 72 = 456, (N) = (A) + (α)= 152 + 456=608

(A) * (B) /N=152´ 512/608= 128

(AB) = 128

(AB) =(A) * (B)/N Hence A and B are independent

Yule's co-efficient of association:

The above example gives a rough idea about association but not the degree of association. For this Prof. G. Undy Yule has suggested a formula to measure the degree of association. It is a relative measure of association between two attributes A and B.

If (AB), (α B), (A β) and (α β) are the four distinct combination of A, B, α and β then Yules' co-efficient of association is

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

Note:

I. If Q = +1 there is perfect positive association

If Q = -1 there is perfect negative association

If Q = 0 there is no association (i.e) A and B are independent

Example 4:

Investigate the association between darkness of eye colour in father and son from the following data.

Fathers' with dark eyes and sons' with dark eyes = 50

Fathers' with dark eyes an sons' with no dark eyes = 79

Fathers' with no dark eyes and sons with dark eyes = 89

Neither son nor father having dark eyes = 782

Solution:

Let A denote the dark eye colour of father and B denote dark eye colour of son.

	A	α	Total
B	50	89	139
β	79	782	861
Total	129	871	1000

Yule's co-efficient of association is

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)} = \frac{50*782 - 89*79}{50*782 + 89*79} = 0.69$$

There is a positive association between the eye colour of fathers' and sons' .

Yule's co-efficient of Colligation:

If (AB), (α B), (A β) and (α β) are the four distinct combination of A, B, α and β then Yule's co-efficient of Colligation is

$$Y = \frac{\sqrt{(AB)*(\alpha\beta)} - \sqrt{(A\beta)*(\alpha B)}}{\sqrt{(AB)*(\alpha\beta)} + \sqrt{(A\beta)*(\alpha B)}}$$

Relation between Q and Y is $Q = \frac{2Y}{1+Y^2}$