

CLASSIFICATION AND TABULATION OF DATA



Odisha Social Science
Research & Consultancy

LEARNING OBJECTIVES

- **The trainees will be able to do meaningful classification of large mass of data and interpret the same.**
- **They will be able to construct frequency distribution table and interpret the same.**
- **They will be able to describe different parts of tables and types of table**

CONTENT



- ☐ **Concept of Variable**
- ☐ **Ordered array**
- ☐ **What is data Classification ?**
- ☐ **Objectives of Classification**
- ☐ **Frequency distributions**
 - ☐ Variables and attributes
- ☐ **Tabulation of data**
 - ☐ Parts of a table
 - ☐ Type of tables

Concept of Variable

- **Variable**

A characteristic which takes on different values in different persons, place or things.

Example: Diastolic/Systolic blood pressure, heart rate, the heights of adult males, the weights of preschool children and the ages of patients seen in a dental clinic

- **Quantitative Variable:-** One that can be measured and expressed numerically. The measurements convey information regarding amount.

Example: Diastolic/Systolic blood pressure, heart rate, the heights of adult males, the weights of preschool children and the ages of patients seen in a dental clinic

- **Qualitative Variable:-** The characteristics that can't be measured quantitatively but can be categorized. The measurement convey information regarding the attribute. The measurement in real sense can't be achieved but persons, places or things belonging to different categories can be counted.

Example: sex of a patient, colour and odour of stool and urine samples etc.

Random Variable

- Values obtained arise as a result of chance event/factor, so that can't be exactly predicted in advance.

Example: heights of a group of randomly selected adult.

- **Discrete Random Variable:-** Characterized by gaps or interrupts in the values that it can assume.

It assumes values with definite jumps.

It can't take all possible values within a range.

It is observed through counting only

Example: No. of daily admission to a general hospital, the no. of decayed, missing or filled teeth per child in an elementary school.

■ Continuous Random Variable:-

- It can take all possible values positive, negative, integral and fractional values within a specified relevant interval.
- Doesn't possess the gaps or interruptions within a specified relevant interval of values assumed by the variable.
- Derived through measurement

Example: height, weight and skull circumference

Because of limitations of available measuring instruments, however observations on variables that are inherently continuous are recorded as if they are discrete.

The ordered array

- A first step in organizing data is preparation of an ordered array.
- It is a listing of values of a data series from the smallest to the largest values.
- It enables one to quickly determine the smallest and largest value in the data set and other facts about the arrayed data that might be needed in a hurried manner.
- Look at the unordered and ordered data in the file DataExample.xls

- **DATA CLASSIFICATION:** *The grouping of related facts/data into different classes according to certain common characteristic.*

- **Basis of data Classification:**

- **Broadly 4 broad basis**

1. **Geographical** *i.e. area wise*

- Total Population of Orissa by districts
- No. of death due to malaria by districts.
- Infant deaths in Orissa by districts

2. Chronological or Temporal

- *i.e. on the basis of time*

Table: 2 Death by lightening

Year	Number
1990	10
1991	5
1992	12
1993	6
1994	9
1995	3
1996	3
1997	5
1998	12
1999	12
2000	8
2001	7
2002	8
Total	100

3. Qualitative *i.e. on the basis of some attributes*

Example: People by place of residence, sex and literacy

Place of residence							
Rural				Urban			
Male		Female		Male		Female	
Literate	Illiterate	Literate	Illiterate	Literate	Illiterate	Literate	Illiterate

4. Quantitative: On the basis of quantitative class intervals

For example students of a college may be classified according to weight as follows

Table 3 :Weight of students of a college

Wt. In (LBS)	No. of students
90-100	50
100-110	200
110-120	260
120-130	360
130-140	90
140-150	40
Total	1000

Classification of Age of 600 person in the Social Survey

Class Interval	Frequency	Relative frequency
15 -24	56	09.3
25-34	153	25.5
35-44	149	24.8
45-54	75	12.5
55 - 64	61	10.2
65 - 74	70	11.7
75 - 84	28	4.7
85 - 94	8	1.3
Total	600	100.0

In a survey of 35 families in a village,
the number of children per family was recorded data were obtained.

1	0	2	3	4	5	6
7	2	3	4	0	2	5
8	4	5	9	6	3	2
7	6	5	3	3	7	8
9	7	9	4	5	4	3

OBJECTIVES OF CLASSIFICATION

- Helps in *condensing the mass of data* such that similarities and dissimilarities can be readily distinguished.

No. of children	No of families (Frequency)	Cum. Fre. Less than	Cum. Fre. Greater than
0-2	7	7	35
3-5	16	23	28
6 and above	12	35	12
Total	35		

■ Facilitate *comparison*

No. of children	No of families (Frequency)	Cum. Fre. Less than	Cum. Fre. Greater than
0-2	7	7 (20%)	35 (100)
3-5	16	23 (65.7%)	28 (80%)
6 and above	12	35 (100%)	12 (34%)
Total	35		

- Most *significant features* of the data can be pin pointed at a glance
- Enables *statistical treatment* of the collected data
 - Averages can be computed
 - Variations can be revealed
 - Association can be studied
 - Model for prediction / forecasting can be built
 - Hypothesis can be formulated and tested etc.

Principles of Classification:

There is no hard and fast rules for deciding the class interval, however it depends upon:

- Knowledge of the data
- Lowest and highest value of the set of observations
- Utility of the class intervals for meaningful comparison and interpretation

- **The classes should be collectively exhaustive and non-overlapping i.e. *mutually exclusive*.**
- **The number of classes should not be too large other wise the purpose of class i.e. *summarization of data will not be served*.**
- **The number of classes should not be too small either, for this also may obscure the true nature of the distribution.**
- **The class should preferable of equal width. Other wise the class frequency would not be comparable, and the computation of statistical measures will be laborious.**

- More specifically Struges formula can be used to decide the no. of class interval;
 - $K=1+3.322(\log_{10} n)$
Where k = no. of classes, n=no. of observation
- The width of the class interval may be determined by dividing the range by k

$$w = \frac{R}{k}$$

where R= difference between the highest and the lowest observation.

w = width of the class interval

- When the nature of data make them appropriate class interval width of 5 units, 10 units and width that are multiple of 10 tend to make the summarization more comprehensible.

Classification will be called **exclusive (Continuous)**, when the class intervals are so fixed that the upper limit of one class is the lower limit of the next class and the upper limit is not included in the class.

An example

Income (Rs.)	No. of families
1000 – 1100 = (1000 but under 1100)	15
1100 – 1200 = (1100 but under 1200)	25
1200 – 1300 = (1200 but under 1300)	10
Total	50

- Classification will be **inclusive (discontinuous)** when the upper and lower limit of one class is include in that class itself

Income (Rs.)	No. of persons
1000 – 1099 = (1000 but $\leq$ 1099)	50
1100 – 1199 = (1100 but $\leq$ 1199)	100
1200 – 1299 = (1200 but $\leq$ 1299)	200
Total	300

- Discontinuous class interval can be made continuous by applying the Correction factor.

$$\text{CF} = \frac{\text{Lower limit of 2}^{\text{nd}} \text{ Class} - \text{Upper limit of the 1}^{\text{st}} \text{ Class}}{2}$$

The correction factor is subtracted from the lower limit and added to the upper limit to make the class interval continuous.

Frequency distributions

- **Quantitative Variables:**
 - Discrete variable
 - Continuous variable
- **Qualitative variable (attributes)**
- **The manner in which the total number of observations are distributed over different classes is called a frequency distribution.**

Frequency distribution of an attribute

- **In 1993, 1674 inhabitants of Calcutta, Bombay and Madras were surveyed. Each was asked, among, other questions, whether he/she knew about the HIV / AIDS. The results is tabulated.**

Table 4 : Results of survey on Awarenesson HIV / AIDs

State of Knowledge	Number of people
Aware	620
Unaware	1054
Total	1674

Table 5 : Proportion of people Aware of HIV / AIDS

State of Knowledge	Relative frequency
Aware	0.370
Unaware	0.630
Total	1.000

Frequency distribution of a discrete variable

- Data grouped in to classes and the number of cases which fall in each class are recorded

Example: In a survey of 35 families in a village, the number of children per family was recorded data were obtained.

1	0	2	3	4	5	6
7	2	3	4	0	2	5
8	4	5	9	6	3	2
7	6	5	3	3	7	8
9	7	9	4	5	4	3

Steps for frequency distribution

- Find the largest & smallest value; those are 9 and 0 respectively.
- Form a table with 10 classes for the 10 values 0,1,2.....9
- Look at the given values of the variable one by one and for each value put a tally mark in the table against the appropriate class.
- To facilitate counting, the tally marks are arranged in the blocks of five every fifth stroke being drawn across the proceeding four. This is done below.

Table 6: Frequency Table

No. of children	Tallies	Frequency	Cumulative Frequency	Cumulative Frequency
			Less than type	More than type
0		2	2	35
1		1	3	33
2		4	7	32
3	 	6	13	28
4	 	5	18	22
5	 	5	23	17
6		3	26	12
7		4	30	9
8		2	32	5
9		3	35	3

TABULATION OF DATA

- Compress the data into rows and columns and relation can be understood.
- Tabulation simplifies complex data, facilitate comparison, gives identify to the data and reveals pattern

Different parts of a table

- **Table number**
- **Title of the table**
- **Caption: Column Heading**
- **Stub : Row heading**
- **Body : Contains data**
- **Head notes: Some thing that is not explained in the title, caption, stubs can be explained in the head notes on the top of the table below the title.**
- **Foot notes: Source of data, some exception in the data can be given in the foot notes.**

Table can be classified into 3 ways

Type of table	Characteristic Feature
1. Simple table	only one characteristic is shown
2. Complex table	
a. Two way table	shows two characteristics and is formed when either the stub or the caption is divided in to two co-ordinate parts
b. Higher order table	When three or more characteristic are represented in the same table, such a table is called higher order table
3. General and special purpose table	published by Govt. such as in the statistical Abstract of India, or census reports are general purpose table

Complex table

Average Number of OPD patients in a PHC in a tribal area in different age group according of sex

Age in yrs	OPD Patients		
	Male	Female	Total
Below 25	25	5	30
25-35	30	4	34
35-45	25	5	30
45-55	22	3	25
Above 55	15	1	16
Total	117	18	135

Number of patients in OPDs of Public sector hospital by Religion, Age, Rank and Sex

[illegible]

■ Exercise

Age of 169 subjects who participated in a study of Sparteine and Mephynytoin Oxidation is given. Calculate the frequency, cumulative frequency both less than and more than type, relative frequency and cumulative relative frequency. From these find out what is the proportion of patient aged 60 and more. (Data in the file DataExample.xls)

- Next Class:
 - Understanding data through presentations

The background is a solid blue color with a subtle grid pattern of thin, lighter blue lines. The grid consists of both horizontal and vertical lines, creating a series of small squares across the entire image.

THANK YOU