

Unit-7

DATA PROCESSING AND ANALYSIS

Coding, entering and classifying the data:

Once the field data begin to flow in, the attention is turned to data processing and analysis. Data processing implies editing, coding, classifying and tabulation of the collected data so that they are amenable to analysis.

EDITING:

The first step in data processing is to edit the raw data. Editing of data is the process of examining the collected raw data to detect errors and omissions and to correct these when possible. It involves a careful scrutiny of the completed questionnaires one edits to assure that the data are:

- Accurate
- Concise with other information / facts gathered
- Uniformly entered
- As complete as possible
- Arrange to facilitate coding and tabulation.

The editing can be done at two levels; on the field where the data is collected or in the office.

CODING:

Coding refers to the process of assigning numerical or other symbols to answer so that responses can be put into a limited number of categories or classes. By this several answers are reduced to a few categories.

CLASSIFICATION:

Classification is the process of arranging data in groups or classes on the basis of common characteristics. Data having common characteristics are placed together and in this way the entire data gets divided into a number of groups or classes. Classification can be in the form of either attribute (literacy, sex, honesty etc) or class intervals (like income, population, age, weight etc).

TABULATION:

Tabulation is the process of summarizing raw data and displaying in the compact form (in the form of statistical tables) for further analysis. It is an orderly arrangement of data in columns and rows

Descriptive Vs Inferential statistics:

Descriptive Statistics: It is a type of statistics used to describe or summarize information about population or a sample.

Inferential statistics: Statistics used to make inferences or judgment about a population on the basis of sample information.

POPULATION PARAMETERS V/S SAMPLE STATISTICS:

Population parameters it is used to designate variables in a population. It is a measure of population characteristics.

μ - to represent population mean

σ -to represent standard deviation of the population

N-to represent population size

P-to represent population proportion

Sample Statistics: It is used to designate variables in a sample. It measures characteristics of a sample

$\bar{X}$ – to represent sample mean

s- to represent sample standard deviation

n- to represent sample size

p-to represent sample proportion

MEASURES OF CENTRAL TENDENCY:

The Mean: it is the arithmetic average; it involves all observations and is affected by extreme cases.

The Medium: It is the mid point / middle value. It does not involve all observations.

The Mode: The value that occurs most often is referred to as the modal value.

MEASURES OF DISPERSION:

Range: The difference between the smallest and the largest value of a frequency distribution is known as the range of the distribution.

Deviation scores: A method of calculating how far way any observation is from the mean is to calculate the individual deviation

$$d = x - \bar{x}$$

Average deviation: taking the average of individual deviations.

$$AD = \sum (x - \bar{x}) / n$$

Since $\sum (x - \bar{x})$ is always zero AD is also zero.

Mean Absolute Deviation: $MAD = \sum |x - \bar{x}| / n$

Mean Square Deviation: $MSD = \sum (x - \bar{x})^2 / n$

Standard Deviation $= \sqrt{MSD} = \sqrt{\sum (x - \bar{x})^2 / n}$

Data analysis:

Tabulation: It refers to the orderly arrangement of data in a table or other summary format. At the very beginning, we have to develop a master sheet, and transform all information from each questionnaire on this master sheet. Counting the number of responses to a question and putting them in a frequency distribution is tabulation.

Ex: Do you have TV? a. Yes b. No

Possession of TV:

RESPONSES	FREQUENCY
YES	27
NO	93
TOTAL	100

Percentage: For simplicity frequency distribution has to be followed by percentages.

Cross – tabulation: a technique of organizing data by group category or classes.

Ex: Question1: What is your Occupation a. Business b. Civil Servant

Question2: Do you have a TV a.) Yes b.) No.

Television possession by occupation:

Occupation	Response		Total
	Yes	No	
Business	8	20	28
Civil Servant	19	73	92
Total	27	93	120

Percentage – Cross Tabulation: Changing the figures in the above table into percentages.

Data Transformation: It is a process of changing data's original form to a format that is more suitable to perform a data analysis that will achieve the research objective.

To a question to students of a college, whether the college life is beautiful?

- a. Strongly agree b. Agree c. Neither agree nor disagree d. Disagree
e. Strongly disagree

On a 5-point scale, it can be converted as:

Strongly Agree as + 2 or 5

Agree as +1 or 4

Neither agree or disagree as 0 or 3

Disagree as -1 or 2

Strongly disagree as -2 or 1

Using index numbers:

Index numbers are data summary values calculated based on figures for some base period to facilitate comparisons over time.

The index number shows percentage changes from a base number (if the data are time related, a base year is chosen). The index numbers are computed by dividing each year's activity by the base year activity and multiplying by hundred.

Ex: The following is a hypothetical data of a given region:

Year	Land (in hectares)	Index	Population size (in millions)	Index
1990	20000	100	3	100
1991	21000	105	3.1	103
1992	21500	107.5	3.2	107
1993	22000	110.0	3.3	110
1994	23000	115	3.3	110
1995	22000	110	3.4	113
1996	22500	112.5	3.5	117
1997	21500	107.5	3.6	117
1998	23000	115	3.7	120

Note: We took the beginning year value as base year.

Calculating rank order: Respondents often indicate a rank ordering of preference. To summarize these data for all respondents researchers perform a data transformation by multiplying the frequencies by the rank (score) to develop a new scale.

Ex: Individual ranking of places of work of selected second year students of a particular university gave the following results:

Respondents	Mekalle	Bahirdar	Nazareth	Awassa	Jimma
1	5	4	1	3	2
2	1	2	4	5	3
3	5	1	2	3	4
4	3	1	5	2	4
5	5	4	2	1	3

Preference Rank

Place of Work	1 st	2 nd	3 rd	4 th	5 th
Mekalle	1	-	1	-	3
Bahirdar	2	1	-	2	-
Nazareth	1	2	-	1	1
Awassa	1	1	2	-	1
Jimma	-	1	2	2	-

In this case we have to multiply number of respondents by the rank score and then summarizing up the scores. The lowest total score show the first preference ranking.

Accordingly:

Mekalle: $(1 \times 1) + (1 \times 3) + (3 \times 5) = 19$ Ranked 5th

Bahirdar: $(2 \times 1) + (1 \times 2) + (2 \times 4) = 12$ Ranked 1st

Nazareth: $(1 \times 1) + (2 \times 2) + (1 \times 4) + (1 \times 5) = 14$ Ranked 2nd

Awassa: $(1 \times 1) + (1 \times 2) + (2 \times 3) + (1 \times 5) = 14$ Ranked 2nd

Jimma: $(1 \times 2) + (2 \times 3) + (2 \times 4) = 16$ Ranked 4th

STATISTICAL ANALYSIS:

This requires choosing the appropriate statistical techniques. The choice of the method of statistical analysis depends on:

1. The type of questions to be asked: i.e., is it to measure central tendency, relationship between the variables, or category differences.
2. The number of variables:
 - ❖ **Univariate data analysis:** When a researcher generalizes from a sample about one variable
 - ❖ **Bivariate data analysis:** When the desire is to explain the relationship between two variables at a time
 - ❖ **Multivariate data analysis:** Is the simultaneous investigation of more than two variables.
3. **Scale measurement of data:** There are four scale measurements of data as follows:
 - a) **Nominal Data:** (scale) Data that fall into different categories. Where one cannot array the category in any order of magnitude, no mathematical operations can be conducted on this data Ex: about sex, religion, the chest number's worn by the athletes or the numbers on the t-shirts of the football players meant for their identification.
 - b) **Ordinal Data:** (scale) Data that permits a ranking by order of magnitude but it is not possible to determine how much one item is compared with another. Not much of mathematical operations can be conducted on data in this type as well ex: one might rank from 1 to 5 for five towns in which they might prefer to work, the position of a athlete at the end of a race.
 - c) **Interval Data:** (scale) / ratio data – provides detailed information but not much of mathematical operations as addition, subtraction, multiplication, division can be worked out Ex: time of finishing a 100 meters race by the athletes in a competition etc
 - d) **Ratio Data:** (scale) Provides detailed information and all the mathematical operations can be conducted on this type of data. Ex: income, age, weight, height, price, output etc.

PARAMETRIC Vs NON-PARAMETRIC:

Note:

Parametric analysis based on interval data and nonparametric analysis is based on nominal or ordinal data

Parametric statistics: Statistical procedures that use interval scales or ratio scale and assume population or sampling distribution is normal.

Non-parametric statistics: Statistical procedures that use nominal or ordinal data and make no assumptions about the distribution of the population or sampling distribution.

Examples of selecting the appropriate statistical methods:

Scale of measurement	Problems	Statistical questions to be answered	Possibilities of statistical significance
Interval or ratio scale	Compare actual and hypothetical values of average	Is the sample mean significantly different from the hypothesized population mean.	Z-test if sample size is large ($n > 30$) t-test if sample size is small ($n < 30$)
Ordinal Scale	Compare actual and expected values	Does the distribution differ from the expected	The Chi – square test (χ^2)
Nominal Scale	Identify sex of key executives	Is the number of female executives equal to the number of male executives	The Chi – square test (χ^2)
Nominal scale	For comparing more than two variables	Is there significant difference between more than two variables in terms of mean?	One way and two ANOVA (analysis of variance test)

Hypothesis testing procedure:

1. **Setting of the hypothesis :**

- a. **The Null Hypothesis:** a statement about a status-quo that asserts any change from what has been thought to be true will be entirely due to random error.
- b. **The alternative hypothesis:** is a statement indicating the opposite of null hypothesis.

The purpose of hypothesis testing is to determine which one of these two hypotheses is correct.

2. **Level of significance and critical values:** This determines the chances of committing a type I error i.e., the chances that the null hypothesis is true and is rejected, generally represented with α . (type II error is the chances that the null hypothesis is false and is accepted) Generally 5% is taken when nothing is mentioned for α . And the relevant critical values based on the tail of the test and the test statistic adopted is represented diagrammatically.
3. **Test statistic:** appropriate statistical tool in the form of Z-test, t-test, χ^2 test or ANOVA is conducted.
4. **Computation:** based on the relevant test statistic the values obtained after editing is introduced into the relevant formula.
5. **Decision:** In this step based on the results obtained after computation acceptance or rejection of the null hypothesis is done.