**UNIT ONE**

1. **INTRODUCTION TO STATISTICS**
   1. **Definition and classification of Statistics**

The word statistics is defined in different ways depending on its use in the plural and singular sense.

**In the plural sense**: - statistics is defined as the collection of numerical facts or figures (or the raw data themselves).

Example: 1. Vital statistics (numerical data on marriage, births, deaths, etc).

2. The average mark of statistics course for students is 70% would be considered as a statistics whereas Abebe has got 90% in statistics course is not statistics.

**Remark:** statistics are aggregate of facts. Single and isolated figures are not statistics as they cannot be compared and are unrelated.

**In its singular sense**:- the word Statistics is the subject that deals with the methods of collecting, organizing, presenting, analyzing and interpreting statistical data.

**Classification of Statistics**

Statistics is broadly divided into two categories based on how the collected data are used.

**Descriptive Statistics**:- deals with describing the data collected without going further conclusion.

**Example 1.1:** Suppose that the mark of 6 students in Statistics course for computer science students is given as 40, 45, 50, 60, 70 and 80. The average mark of the 6 students is 57.5 and it is considered as descriptive statistics.

**Inferential Statistics**:- It deals with making inferences and/or conclusions about a population based on data obtained from a sample of observations. It consists of performing hypothesis testing, determining relationships among variables and making predictions.

**Example 1.2:** In the above example, if we say that the average mark in Statistics course for Mathematics students is 57.5, then we talk about inferential statistics (draw conclusion based on the sample observation).

**1.2 Stages of Statistical Investigation**

The area of statistics points out the following five stages. These are collection, organization, presentation, analysis and interpretation of data.

**Collection of data**: This is the process of obtaining measurements or counts or obtaining raw data. Data can be collected in a variety of ways; one of the most common methods is through the use of sample or census survey.

**Organization of data**: - Data collected from published sources are generally in organized form. However if an investigator has collected data through a survey, it is necessary to edit these data in order to correct any apparent inconsistencies, ambiguities, and recording errors.

This phase also includes correcting the data for errors, grouping data into classes and tabulating.

**Presentation of data**:- After the data have been collected and organized they can be presented in the form of tables, charts, diagrams and graphs. This presentation in an orderly manner facilitates the understanding as well as analysis of data.

**Analysis of data**: - the basic purpose of data analysis is to dig out useful information for decision making. This analysis may simply be a critical observation of data to draw some meaningful conclusions about it or it may involve highly complex and sophisticated mathematical techniques.

**Interpretation of data**: - Interpretation means drawing conclusions from the data collected and analyzed. Correct interpretation will lead to a valid conclusion of the study & thus can aid in decision making.

**1.3 Definition of some statistical terms**

**Population**: - It is the totality of objects under study. The population represents the target of an investigation, and the objective of the investigation is to draw conclusions about the population hence we sometimes call it target population. The word population doesn’t necessarily refer to people.

**Examples**:- All clients of Telephone Company, Population of families, etc.

The population could be finite or infinite (an imaginary collection of units).

**Sample: -** is part or subset of population under study.

**Sampling frame: -** is the list of all possible units of the population that the sample can be drawn from it.

Eg. List of all students of AAU, List of all residential houses in A.A city, etc

**Survey**: - is an investigation of a certain population to assess its characteristics. It may be census or sample.

**Census survey**: a complete enumeration of the population under study.

**Sample survey**: the process of collecting data covering a representative part or portion of a population.

**Parameter**: - is a statistical measure of a population, or summary value calculated from a population. Examples: Average, Range, proportion, variance, etc

**Statistic**: - is a descriptive measure of a sample, or it is a summary value calculated from a sample.

**Sampling**: - The process or method of sample selection from the population.

**Sample size**: - The number of elements or observation to be included in the sample.

**An element**: - is a member of sample or population. It is specific subject or object (for example a person, firm, item, etc.) about which the information is collected.

**Variable**: - It is an item of interest that can take numerical or non-numerical values for different elements. It may be qualitative or quantitative. Example: age, weight, sex, marital status, etc.

**Observation (measurement)**:- is the value of a variable for an element.

**Qualitative variables**:- are variables that assume non-numerical values. They can be categorized and they are usually called attributes. Example: - Sex, marital status, ID number, etc.

**Quantitative variables**: - are variables which assume numerical values. eg. Age, weight, etc.

**1.4 Applications, uses and limitations of Statistics**

Statistics can be applied in any field of study which seeks quantitative evidence. For instance, Engineering, Economics, Natural Science, etc

Engineering: Statistics have wide application in engineering.

* To compare the breaking strength of two types of materials
* To determine the probability of reliability of a product.
* To control the quality of products in a given production process.
* To compare the improvement of yield due to certain additives such as fertilizer, herbicides, e t c.

**Function/Uses of Statistics**

The following are some uses of statistics:

* It **condenses and summarizes a mass of data**: the original set of data (raw data) is normally voluminous and disorganized unless it is summarized and expressed in few presentable, understandable & precise figures.
* Statistics **facilitates comparison of data**: measures obtained from different set of data can be compared to draw conclusion about those sets. Statistical values such as averages, percentages, ratios, rates, coefficients, etc, are the tools that can be used for the purpose of comparing sets of data.
* Statistics **helps to predict future trends**: statistics is very useful for analyzing the past and present data and forecasting future events.
* Statistics **helps to formulate & review policies**
* **Formulating and testing hypothesis:** Statistical methods are extremely useful in formulating and testing hypothesis and to develop new theories.

**Limitations of Statistics**

Some of these limitations are:

* **It does not deal with individual values**: as discussed earlier, statistics deals with aggregate of facts. For example, wage earned by an individual worker at any one time, taken by itself is not a statistics.
* **It does not deal with qualitative characteristics directly**: statistics is not applicable to qualitative characteristics such as beauty, honesty, poverty, standard of living and so on since these cannot be expressed in quantitative terms.
* **Statistical conclusions are not universally true**: since statistics is not an exact science, as is the case with natural sciences, the statistical conclusions are true only under certain assumptions.
* **It can be misused**: statistics cannot be used to full advantage in the absence of proper understanding of the subject matter.

**1.5 Levels of Measurement**

Proper knowledge about the nature and type of data to be dealt with is essential in order to specify and apply the proper statistical method for their analysis and inferences.

**Scale Types**

Measurement is the assignment of values to objects or events in a systematic fashion. Four levels of measurement scales are commonly distinguished: nominal, ordinal, interval, and ratio. The first two are qualitative while the last two are quantitative.

**Nominal scale**: The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. Qualities with no ranking or ordering, no numerical or quantitative value;these types of data are consists of names, labels and categories.

**Example 1.3**: Eye color: brown, black, etc, sex: male, female.

* In this scale, one is different from the other
* Arithmetic operations (+, -, \*, ÷) are not applicable, comparison (<, >, ≠, etc) is impossible

**Ordinal scale**: - defined as nominal data that can be ordered or ranked.

* Can be arranged in some order, but the differences between the data values are meaningless.
* Data consisting of an ordering of ranking of measurements are said to be on an ordinal scale of measurements. That is, the values of an ordinal scale provide enough information to order objects.
* One is different from and greater /better/ less than the other
* Arithmetic operations (+, -, \*, ÷) are impossible, comparison (<, >, ≠, etc) is possible.

**Example 1.4 -**Letter grading (A, B, C, D, F), -Rating scales (excellent, very good, good, fair, poor), military status (general, colonel, lieutenant, etc).

**Interval Level**: data are defined as ordinal data and the differences between data values are meaningful. However, there is **no true zero**, or **starting point**, and the ratio of data values are meaningless. Note: Celsius & Fahrenheit temperature readings have no meaningful zero and ratios are meaningless.

In this measurement scale:-

* One is different, better/greater and by a certain amount of difference than another.
* Possible to add and subtract. For example; 800c – 500c = 300c, 700c – 400c = 300c.
* Multiplication and division are not possible. For example; 600c = 3(200c). But this does not imply that an object which is 600c is three times as hot as an object which is 200c.

Most common examples are: temperature, IQ.

**Ratio scale**: Similar to interval, except there is a true zero (absolute absence), or starting point, and the ratios of data values have meaning.

* Arithmetic operations (+, -, \*, ÷) are applicable. For ratio variables, both differences and ratios are meaningful.
* One is different/larger /taller/ better/ less by a certain amount of difference and so much times than the other.
* This measurement scale provides better information than interval scale of measurement.

**Example 1.5**: weight, age, number of students.

**CHAPTER TWO**

1. **METHODS OF DATA COLLECTION AND PRESENTATION**
   1. **Method of Data Collection**

There are three major methods of data collection.

1. Observation or measurement
2. Interviews and questionnaires
3. The use of documentary sources
4. **Observation or measurement** ( direct personal observation)

In this case data can be obtained through direct observation or measurement. This requires training and monitoring of the measure to ensure the use of standard procedure.

* Provides accurate information but it is expensive and inconvenient.

**Example**: physical examination, clinical measurements, laboratory tests etc.

1. **Interviews and questionnaires**
2. **Direct personal Interview**

In direct personal interview method, the investigator presents himself/herself personally before the informant and questions him/her personally.

This method is best suited to situations where the problems are not completely understood and where questions cannot be formulated before hand and one question leads to another. The disadvantage of is that it is time consuming and is not suited for large group of informants.

1. **Enumeration**

In this method, well trained enumerators ask the selected group of respondent a set of questions relevant to the study. A form that contains the set of questions to be asked and recorded the response of information afterward by an enumerator is known as schedule.

E.g. census in Ethiopia

**Respondents (Interviewees)** are individuals who are to answer the question on the questionnaire.

**Interviewers**: the person to record the responses given to the questions by the respondent.

1. **Face To Face Interviews** (questionnaires in charge of enumerators)

The interviewer knows exactly who is responding to the questionnaire. This may not be the case with either telephone or mail administered questionnaire since anyone in the household can answer.

* The interviewer can help the respondent if he/she has difficulty in understanding the questions. The difficulty could be due to language, concentration or limited intellectual capacity.
* There is more flexibility in presenting the items; they can range from closed to open.
* There is the ability to use the method of skip patterns.
* Skip patterns means skipping a questions or a group of questions which are not applicable.

Disadvantages

* Untrained interviewer may distort the meaning of the questions.
* Attribute of the interviewer may affect the responses given due to:

1. Bias of the interviewer and b) his/her social or ethnic characteristics.

* It costs much in terms of time and money. Training of interviewers, salary for interviewers. In many instances interviewers go house to house in order to locate the respondents.
* Employing a bilingual interviewer can also increase cost.

1. **Telephone Interviews**

Advantages

It is less expensive in time and money compared with face to face interviews. The interviewer is able to help the respondent if he/she doesn’t understand the question (as seen with face to face interview)

* Broad representative samples can be obtained for those who have telephone lines.
* May assure the uniformity of interviewer.

Disadvantage

* Under representation of those groups which do not have telephones.
* Problem with unlisted telephone number in the directory.
* Respondent may be substituted by another.
* Depending on the time of day the phone calls are made, different types of persons are reached which will create bias in the sample.
* Problem with questions with multiple options for answers and complicated questions.
* Repeated calls may be necessary.

1. **Self administered questionnaire returned by mail** (mailed questionnaire)

Here the questionnaire is mailed to respondents to be filled. Sometimes it is known as self enumeration.

**Advantages**

* These are the cheapest. There is no need for trained interviewer. There is no interviewer bias.
* Mailed questionnaire can be coordinated from one central location.

Disadvantage

* Low response rate
* Uncompleted questionnaires due to omission or invalid responses.
* No assurance that the questionnaire was answered by the right person
* Needs intense follow up to get a high response rate.

1. **The use of documentary sources**

Extracting information from existing sources (e.g. Hospital records) is much less expensive than the other two methods. It can be an important source of data.

**Limitation**: It is difficult to get information needed, when records are compiled in unstandardized manner.

* 1. **Source and Types of Data**

Data may be obtained from two sources, primary and secondary.

1. **Primary sources**: sources that can supply first hand information for immediate use.

**Primary data**: data originally collected for the purpose at hand.

Example: observe signs, measure characteristics, record symptoms and interview respondent, etc.

1. **Secondary sources**: the source in which data are obtained from records of individual that have been collected by persons other than the investigator for other purpose.

Example: Hospital records, vital statistics and registers, etc.

**Secondary data**: the data obtained from secondary sources.

* 1. **Method of Data Presentation**

The data collected in survey or other empirical inquiry are called **raw** data. These unorganized data are not in a way to be assimilated. It is therefore, necessary to reduce and present the data with their relevant features.

**Tables**

Tables include the systematic arrangement of statistical data in column and rows. Important features are:-

1. Tables should be simple and self explanatory.
2. Each row or column should be labeled concisely and clearly giving units of measurement for all quantitative data.
3. The title should describe the content of the table and the scale should be understood without reference to the text. A good title will answer the question: what? When? And where?
4. Percentage should add up to 1000%.
5. Any necessary explanatory footnotes should be included at the bottom of the table.
   * 1. **Frequency distribution and tables**

**Frequency**: - is the number of counts assigned to individuals having a particular characteristic.

**Frequency distribution**: the set of frequencies of all possibilities is called frequency distribution of the variable.

Based on the type of data, we can have two type of frequency distribution, tables.

1. **Qualitative frequency tables** (categorical frequency distribution)

Table 1 Data on smoking status by gender of a sample of health workers, Jimma Hospital 1986 E.C.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Observation | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| Gender | M | F | M | M | F | F | F | M | M | M | F | F | F | F | M | F | M | F | M | M |
| Smoking status | Y | N | N | Y | N | N | Y | N | N | N | N | N | N | Y | Y | Y | N | N | Y | Y |

Information on each of the characteristics (gender, and smoking status) is displayed for each health worker.

1. Characteristics Tally Frequency

Gender

Male //// //// 10

Female //// //// 10

Smoking status

No //// //// // 12

Yes //// /// 8

Summarize what is presented in (a)

1. Characteristics Frequency (%)
2. **Gender**
3. Male 10 50
4. Female 10 50
5. **Smoking status**
6. No 12 60
7. Yes 8 40

Provides both frequency and percentage

The presentation in (b) is one way table. The classification by gender is one way classification. The classification by smoking status is another one way table. When a single variable is used for classification, the table formed is considered as one way table.

**Table c**

No Yes Total

Female 7 3 10

35% 15% 50%

70% 30%

58.33% 37.5%

Male 5 5 10

25% 25% 50%

50% 50%

41.67% 62.5%

12 8 20

60 40 100

What is presented in (c) is two way table with four cells sometimes known as **Contingency Table**. When two variables are used for classification then the table is called two way or contingency table.

Note: the sum of all cell percentage must be equal to 100%.

35+15+25+25=100%

70+30=100% the sum of row 1 pct

58.33+41.67=100% the sum of coll. Pct

When more than one variable are used for classification, the table formed is called High order table.

Example: the following is an example for high order table. Three classification variables are used.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Smoking status | |
| Health center | Gender | Yes | No |
| 1 | Male | 10 | 32 |
|  | Female | 23 | 98 |
| 2 | Male | 33 | 65 |
|  | Female | 12 | 21 |
| 3 | Male | 11 | 32 |
|  | Female | 21 | 21 |
|  |  |  |  |

**b) Quantitative Frequency Table**

1. **Ungrouped Frequency Distribution**

When measurements are taken on the entities of a population, the resulting values usually comes to the researcher as a mass of data.

The first step in organizing these data is the preparation of an ordered array.

**An ordered Array**: is a listing of the values of a collection (either population or sample) in order of magnitude, from the smallest value to the largest value.

Table 2: Age in year of 20 women who attended health education at Jimma Health center in 1986

|  |
| --- |
| 30 25 23 41 39 27 41 24 32 29 29 35 31 36 33 36 42 35 37 41 |

II. Table 3: frequency distribution showing the array age of 20 women

|  |
| --- |
| 23 24 25 27 29 29 30 31 32 33 35 35 36 36 37 39 41 41 41 42 |

Slightly better than Table2

Ungrouped frequency distribution of the age of 20 women

|  |
| --- |
| Age(xj) 23 24 25 27 29 30 31 32 33 35 36 37 39 41 42 total |
| Tally / / / / // / / / / // // / / /// / |
| Frequency(f) 1 1 1 1 2 1 1 1 1 2 2 1 1 3 1 20 |

Each individual value is presented separately, that is why it is named ungrouped frequency distribution. Before the day of computers the objective of grouping data was to facilitate the calculation of various descriptive measures and summarization. The main purpose in grouping is now summarization.

Class Intervals: are non overlapping intervals such that each value in the set of observations can be placed in one, and only one, of the intervals.

**Procedures for grouping into class intervals**

1. Decide on the number of classes you want. Too few intervals are undesirable because of the loss of information. On the other hand, if too many intervals are used, the objective of summarization is not being met.

A commonly followed rule of thumb states 6≤N.C ≤15.

The formula

**K=1+3.322(Log10n)** is a formula by “Sturges”.

**K**=number of class intervals.

**n**=number of values in the data.

But this should not be regarded as final answer. E.g. if we have a sample of 275 observations, that we want to group,

K= 1+3.322(log10275) = 1+3.322(2.4393) = 9

In practice other considerations might cause us to use 8 or fewer or perhaps 10 or more class intervals.

1. Determine width(W) of the class interval

W= (b-a)/k

b=the largest observation in the data set.

a=the smallest value in the data set.

k=the number of class interval.

1. Approximate W to the nearest integer.

It is preferable for “w” to be odd since it has advantage of having a mid point which is a number.

i.e. 10-19 width=10 (10+19)/2= 14.5

5-9 width=5 (5+9)/2= 7

Example: consider the age data given previously.

n=20

k=1+3.322(log20) =1+3.322(1.3010) = 5.196

k=5 w= (42-23)/54

The grouped frequency table using Sturges formula

|  |
| --- |
| Class 23-26 27-30 31-34 35-38 39-42 |
| Frequency (fi) 3 4 3 5 5 |

Note: this is only example

The data are grouped in to a set of non-overlapping intervals.

**Class limits (CL):** these are extreme values for each class. They are called lower and upper class limits and are used for discrete values.

For our example:

Lower class limits (LCL): are 23, 27, 31, 35, and 39

Upper class limits (UCL): are 26, 30, 34, 38, and 42

Note: I. usually class intervals are ordered from smallest to largest.

II. The lower limit of the first class interval should be equal to or smaller than the smallest measurement in the data.

1. The upper limit of the last class interval should be equal to or greater than the largest measurement.

**Class Boundaries (CB**)

With continuous data, values such as 26.5 will not fit any of the class given above. It is therefore necessary to set exact limit or true limits which are known as class boundaries.

Exact limits refer to values of continuous measurement.

1. **Lower class boundary (LCB):** given a class limit, the LCB is obtained by subtracting half the unit of measurement from the LCL of the class.

The unit of measurement is the gap between the UCL of the class and the LCL of the next higher class.

Thus LCBi=LCLi - (LCLi+1-UCLi)/2

1. **Upper class Boundary (UCB)**: UCB is the average of the upper class limit and the next lower class limit.

i.e. UCB is obtained by adding half the unit of measurement to the UCL of the class.

Thus

UCBi= UCLi+ (LCLi+1-UCLi)/2

= (UCLi+LCLi+1)/2

Note: UCBi=LCBi+1

Proof:

Consider:

LCLi+1=LCLi+1-(LCLi+2-UCLi+1)/2

But UCLi+1-UCLi=LCLi+2-LCLi+1=w

⇒UCLi+1=LCLi+2 +UCLi-LCLi+1

Substituting this in to the formula, we have

LCBi+1= (LCLi+1+UCLi)/2

Examples:

Convert the following class limit into class boundaries

1. 5-9 b) 44.5-49.4 c) 78.25-80.24

10-14 49.5-54.4 80.25-82.24

15-19 54.5-59.4 82.25-84.24

1. LCB1=LCL1- (LCL2-UCL1 )/2=5- (10-9)/2=4.5

The UCB1= (UCL1+LCL1+1)/2= (9+10)/2=9.5

1. 4.5-9.5 b) 44.45-49.45 c) 78.245-80.245

9.5-14.5 49.45-54.45 80.245-82.245

14.5-19.5 54.45-59.45 82.245-84.245

Class Marks (mi): are the mid points of the classes.

Mi= (LCLi + UCLi)/2 or (LCBi + UCBi)/2

e.g. (5+9)/2=7 or (4.5+9.5)/2=7

Note: mi+1=mi + w

m2=7+5=12 also (10+14)/2=12

**Advantage of grouping**

* Provides information about the range of the data.
* Gives an impression about the values that are frequent and infrequent.
* It provides data that can be easily used for graphical representation.

**Disadvantage of grouping**

* Information may be lost, since individual values displayed.
* Something that can be determined from the original data cannot be determined from grouped data.

**Modified frequency distribution**

1. **The cumulative frequency distribution**: is used when one is interested to know how often the measurements fall below or above a certain level.

**Less than cumulative frequency (LCF)**: the LCF of a value of a variable is the number of individual with value less than or equal to that value.

**More than cumulative frequency (MCF)**: the MCF of a value of a variable is the number of cases with value greater or equal to that value.

Example:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Class limit | Frequency | Less than | LCF | More than | MCF |
| 23-26 | 3 | 22.5(<23) | 0 | 22.5(>22) | 20 |
| 27-30 | 4 | 26.5(<27) | 3 | 26.5(>26) | 17 |
| 31-34 | 3 | 30.5(<31) | 7 | 30.5(>30) | 13 |
| 35-38 | 5 | 34.5(<35) | 10 | 34.5(>34) | 10 |
| 39-42 | 5 | 38.5(<39) | 15 | 38.5(>38) | 5 |
|  |  | 42.5(<43) | 20 | 42.5(>42) | 0 |

Relative frequency distribution: the proportion of individuals expressed as percentage of the total.

Rel.freq. =

⇒R.fi=fi/n x 100

Example

|  |  |  |  |
| --- | --- | --- | --- |
| Class limit | Freq. | Relative freq. (%) | Cumulative R.freq. |
| 23-26 | 3 | 3/20x100=15 | 15 |
| 27-30 | 4 | 4/20x100=20 | 35 |
| 31-34 | 3 | 3/20x100=15 | 50 |
| 35-38 | 5 | 5/20x100=25 | 75 |
| 39-42 | 5 | 5/20x100=25 | 100 |
| Total | 50 | 100 |  |

Note:- about 75% of the women are in the age group 23-28 years.

* + 1. **DIAGRAMATIC PRESENTATION OF DATA**

The essential advantages of these presentations lie in the fact that they facilitate comparisons.

**Bar chart**

There are three types of bar charts

1. Simple
2. Component
3. Multiple
4. **Simple Bar chart**: the bars may be vertical or horizontal, with their height or width representing the size of the data. It helps to make simple comparison b/n data.

* The bars do not overlap.
* The space b/n the bars must be equal and narrow.
* It shows changes in the totals of different categories.

**Example** Construct a simple bar diagram for the following table showing annual cases of HIV reported in Ethiopia as of July 31, 1993.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Year of report | 1986 | 1987 | 1988 | 1989 | 1990 | 1991 | 1992 | 1993 |
| Cases | 2 | 17 | 87 | 190 | 448 | 885 | 3256 | 2814 |



1. **Component Bar chart**

For each category in the bar are subdivided in to components to allow comparison between parts. It is used to present more than one variable. These are two types

1. Actual
2. Percentage

**Actual component bar chart**: height or length of individual components is represented by its actual figure. Different parts of a bar are shaded or colored differently to provide contrast. It shows not only changes in the total of different categories but also the change in component figures within each category.

**Example**

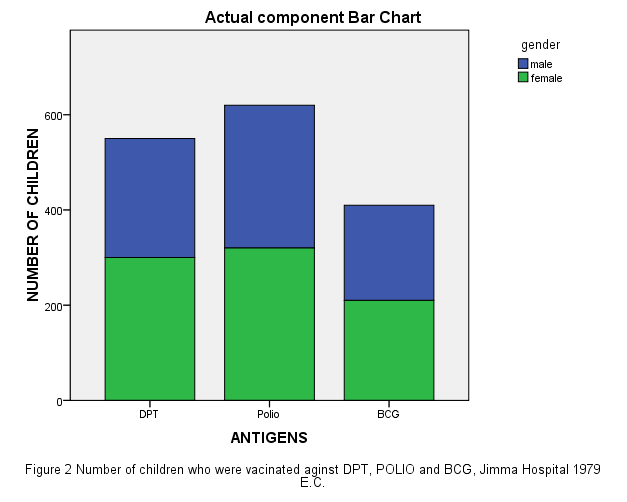
Construct actual component bar chart for the number of children who were vaccinated with DPT, POLIO and BCG antigens in Jimma Hospital in 1979 E.c.

Antigen Male Female

DPT 250 300

Polio 300 320

BCG 200 210



**Percentage component Bar chart**

Similar to actual component bar, except the components are expressed as percentages of the total.

* All bars are equal in height.
* Mostly used to compare relative variation b/n data.

Example: draw a percentage component bar chart for the vaccination data, previously described

**Soln**

Male Female

DPT= 250/550 X 100=45.5% 300/550 X 100=54.5%

POLIO=300/620 X100=48.4% 320/620 X100=51.6%

BCG=200/410 X100=48.8% 210/410 X100=51.2%

1. **Multiple Bar chart**

These are used when two or more inter-related data are to be compared. Height of bars shows actual values of each component. It is used to present more than one variable.

**Example**: draw a multiple bar chart for the vaccination data.



**PIE CHART**

Pie charts are used to show the partitioning of a total into its components parts using circles. The circles should be divided into sectors proportional in size to the frequencies of the categories they represent.

**Steps in drawing a pie chart**

1. Convert freq. distribution into percentage frequency distribution.
2. Draw a circle of any of radius and note that the circle is represented by an angle of 3600.
3. Convert percentage into degree measures. Since the whole circle (3600) represents 100% of the observation, 3.60 will represent 1%.

**Example**

Draw the pie chart for the following table. First construct a table providing the central angles.

|  |  |  |  |
| --- | --- | --- | --- |
| Wards | Frequency | Percentage | Central angle |
| Medical A | 55 | 27.5 | 99 |
| Medical B | 30 | 15 | 54 |
| Surgical A | 40 | 20 | 72 |
| Surgical B | 25 | 12.5 | 45 |
| pediatrics | 50 | 25 | 90 |
| Total | 200 | 100 | 360 |



**Histogram**

A histogram presents grouped frequency distribution of a continuous type. The real limits of the class make up the horizontal axis, while the vertical axis has as its scale the frequency of occurrence. Comparison can be made using the height or areas of the bar.

Method of construction histogram

1. Obtain a frequency distribution with class boundaries and class midpoints.
2. Construct bars on the horizontal axis with center at the class midpoint and width equal to the class width.
3. The height of each bar should correspond to the respective class frequency.

**Example**: consider the following grouped age data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S.N | Class limit | Class boundaries | Mid point | Frequency |
|  | 15-19 | 14.5-19.5 | 17 | 2 |
|  | 20-24 | 29.5-24.5 | 22 | 8 |
|  | 25-29 | 24.5-29.5 | 27 | 6 |
|  | 30-34 | 39.5-34.5 | 32 | 12 |
|  | 35-39 | 34.5-39.5 | 37 | 7 |
|  | 40-44 | 39.5-44.5 | 42 | 6 |
|  | 45-49 | 44.5-49.5 | 47 | 4 |
|  | 50-54 | 49.5-54.5 | 52 | 3 |
|  | 55-59 | 54.5-59.5 | 57 | 1 |
|  | 60-64 | 59.5-64.5 | 62 | 1 |

**Note**: each cell contains a certain proportion of the total area, depending on the frequency.

For example, the fourth cell contains 12/50 of the area. ⇒The relative frequency of occurrence of values between 29.5-34.5.

**Histogram**



**Frequency polygon:-** is a multi-sided figure where the frequency is plotted against the class midpoint. The steps are:

1. Construct a histogram
2. Mark the midpoint on the top of each bar
3. Join these marks with straight lines
4. Extend these lines on both ends so that it reaches the horizontal axis at the class mid points. This allows the total area to be enclosed.

Frequency distributions of age

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Class limit | 15-19 | 20-24 | 25-29 | 30-34 | 35-39 | 40-44 | 45-49 | 50-54 | 55-59 | 60-64 |
| Mid point | 17 | 22 | 27 | 32 | 37 | 42 | 47 | 52 | 57 | 62 |
| Frequency | 2 | 8 | 6 | 12 | 7 | 6 | 4 | 3 | 1 | 1 |



**Note**: the total area under the frequency polygon is equal to the area under the histogram.

**Ogives or cumulative frequency curve**

Points are plotted in association with the exact values on the horizontal axis and the cumulative frequency values on the vertical axis. Then connect the points with straight lines.

-the curves obtained are called the “less than” and “more than” curves.

**Note**: Cumulative frequencies are plotted at the class boundaries. Frequency polygons are plotted at class marks. The sum of the frequencies of two or more classes is cumulative frequency.

Consider the age data

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Class limit | Frequency | Less than | LCF | More than | MCF |
| <23 | 0 | 22.5(<23) | 0 | 22.5(>22) | 20 |
| 23-26 | 3 | 26.5(<27) | 3 | 26.5(>26) | 17 |
| 27-30 | 4 | 30.5(<31) | 7 | 30.5(>30) | 13 |
| 31-34 | 3 | 34.5(<35) | 10 | 34.5(>34) | 10 |
| 35-38 | 5 | 38.5(<39) | 15 | 38.5(>38) | 5 |
| 39-42 | 5 | 42.5(<43) | 20 | 42.5(>42) | 0 |
|  |  |  |  |  |  |

