

CHAPTER 1

1. Introduction to Statistics

Definition and classifications of statistics

Definition of Statistics

We can define statistics in two ways.

1. **Plural sense** (lay man definition).
It is an aggregate or collection of numerical facts.
2. **Singular sense** (formal definition)

Statistics is defined as the art or science of collecting, organizing, presenting, analyzing and interpreting numerical data for the purpose of assisting in making a more effective decision. So, managers and decision makers a better understanding of the business and economic environment and thus enables them to make more informed and better decisions.

Definitions of some terms

- a. **Statistical Population:** It is the collection of all possible observations of a specified characteristic of interest (possessing certain common property) and being under study. An example is all of the students in Debre Markos College those who take statistics course in 2013 E.C.
- b. **Sample:** It is a subset of the population, selected using some sampling technique in such a way that they represent the population.
- c. **Sampling:** The process or method of sample selection from the population.
- d. **Sample size:** The number of elements or observation to be included in the sample.
- e. **Census:** Complete enumeration or observation of the elements of the population. Or it is the collection of data from every element in a population
- f. **Parameter:** Characteristic or measure obtained from a population.
- g. **Statistic:** Characteristic or measure obtained from a sample.
- h. **Variable:** It is an item of interest that can take on many different numerical values.
- i. **Data :** Are the facts and figures collected, analyzed, and summarized for presentation and interpretation.
- j. **Data set:** All the data collected in a particular study.
- k. **Elements:** Are the entities on which data are collected.

- l. **Variable:** is a characteristic of interest for the elements.
- m. **Observation:** Measurements collected on each variable for a particular element

Example

Data set for 5 selected workers from 67 workers in a certain Company

Working IN	Sex	Age	Education status	Income
2	Male	34	Degree	8900
33	Female	24	Diploma	6700
41	Male	28	Degree	7400
52	Male	42	Master	1200
62	Female	52	PHD	1400

Questions : So, identify the above possible statistical terms from the table?

1.2. Classifications of Statistics: Descriptive versus Inferential

Depending on how data can be used statistics is sometimes divided in to two main areas or branches.

1. **Descriptive Statistics:** It is concerned only with the collection, organisation, summarizing, analysis and presentation of an array of numerical qualitative or quantitative data without drawing conclusions. It includes the mean, median, mode, standard deviation, range, percentile, kurtosis, correlation coefficient, proportions.....etc.
2. **Inferential Statistics:** This is built on the descriptive statistics by going a step further to make interpretation with a view to population upon which a decision would be based. Valid and reliable decisions, generalisations, predictions and conclusions could be drawn using this statistics tools such as stochastic process, queuing theory, game theory, quality control, chi-square, t-test, f-test etc. Or it is a method used to generalize from a sample to a population. For example, the average income of all families (the population) in Ethiopia can be estimated from figures obtained from a few hundred (the sample) families. Statistical techniques based on probability theory or Model are required.

1.3. Types of Variables or Data and Scales of Measurement

1.3.1 .Types of Variables or Data

Distinguishing between the different types of variables is a basic and integral part of applied statistics. The methods to analyze these data are very different and therefore it

is important to make the distinction. The two types of variables are Qualitative and Quantitative.

1. **Qualitative Variables** are nonnumeric variables and can't be measured. Examples include gender, religious affiliation, and state of birth.

2. **Quantitative Variables** are numerical variables and can be measured. Examples include balance in checking account, number of children in family. Note that quantitative variables are either discrete (which can assume only certain values, and there are usually "gaps" between the values, such as the number of bedrooms in your house) or continuous (which can assume any value within a specific range, such as income, wage, height...)

1.3.2. Measurement Scales

Measurement is the assignment of numbers to objects or events in a systematic fashion. Four levels of measurement scales are commonly distinguished: nominal, ordinal, interval, and ratio and each possessed different properties of measurement systems.

Nominal Scales

Nominal scales are measurement systems that possess none of the three properties stated above.

- Level of measurement which classifies data into mutually exclusive, all inclusive categories in which no order or ranking can be imposed on the data.
- No arithmetic and relational operation can be applied.

Examples:

- Political party preference (Republican, Democrat, or Other,)
- Sex (Male or Female.)
- Marital status(married, single, widow, divorce)
- Country code
- Regional differentiation of Ethiopia.

Ordinal Scales

Ordinal Scales are measurement systems that possess the property of order, but not the property of distance. The property of fixed zero is not important if the property of distance is not satisfied.

- Level of measurement which classifies data into categories that can be ranked. Differences between the ranks do not exist.
- Arithmetic operations are not applicable but relational operations are applicable.
- Ordering is the sole property of ordinal scale.

Examples:

- Letter grades (A, B, C, D, F).
- Rating scales (Excellent, Very good, Good, Fair, poor).
- Military status.

Interval Scales

Interval scales are measurement systems that possess the properties of Order and distance, but not the property of fixed zero.

- Level of measurement which classifies data that can be ranked and differences are meaningful. However, there is no meaningful zero, so ratios are meaningless.
- All arithmetic operations except division are applicable.
- Relational operations are also possible.

Examples:

- IQ
- Temperature in °F.

Ratio Scales

Ratio scales are measurement systems that possess all three properties: order, distance, and fixed zero. The added power of a fixed zero allows ratios of numbers to be meaningfully interpreted;

i.e. the ratio of Bekele's height to Martha's height is 1.32, whereas this is not possible with interval scales.

- Level of measurement which classifies data that can be ranked, differences are meaningful, and there is a true zero. True ratios exist between the different units of measure.
- All arithmetic and relational operations are applicable.

Examples:

- Weight
- Height
- Income(Salary)
- Number of students
- Age

The following present a list of different attributes and rules for assigning numbers to objects. Try to classify the different measurement systems into one of the four types of scales. **(Exercise)**

1. Your checking account number as a name for your account.
2. Your checking account balance as a measure of the amount of money you have in that account.
3. The order in which you were eliminated in a spelling bee as a measure of your spelling ability.
4. Your score on the first statistics test as a measure of your knowledge of statistics.
5. Your score on an individual intelligence test as a measure of your intelligence.
6. The distance around your forehead measured with a tape measure as a measure of your intelligence.
7. A response to the statement "Abortion is a woman's right" where "Strongly Disagree" = 1, "Disagree" = 2, "No Opinion" = 3, "Agree" = 4, and "Strongly Agree" = 5, as a measure of attitude toward abortion.
8. Times for swimmers to complete a 50-meter race
9. Months of the year Meskerm, Tikimit...

10. Socioeconomic status of a family when classified as low, middle and upper classes.
11. Blood type of individuals, A, B, AB and O.
12. Pollen counts provided as numbers between 1 and 10 where 1 implies there is almost no pollen and 10 that it is rampant, but for which the values do not represent an actual counts of grains of pollen.
13. Regions numbers of Ethiopia (1, 2, 3 etc.)
14. The number of students in a college;
15. the net wages of a group of workers;
16. the height of the men in the same town;

Applications in Business and Economics

In today's global business and economic environment, anyone can access vast amounts of statistical information. The most successful managers and decision makers understand the information and know how to use it effectively. In this section, we provide examples that illustrate some of the uses of statistics in business and economics.

How Accountants Use Statistics

In an age of Big Data, statistical analysis is becoming an increasingly powerful tool for accountants. Taking a course in introductory statistics will help every accountant improve their efficiency and help their clients make better decisions.

Public accounting firms use statistical sampling procedures when conducting audits for their clients. For instance, suppose an accounting firm wants to determine whether the amount of accounts receivable shown on a client's balance sheet fairly represents the actual amount of accounts receivable. Usually the large number of individual accounts receivable makes reviewing and validating every account too time-consuming and expensive. As common practice in such situations, the audit staff selects a subset of the accounts called a sample.

After reviewing the accuracy of the sampled accounts, the auditors draw a conclusion as to whether the accounts receivable amount shown on the client's balance sheet is acceptable.

Finance: Financial analysts use a variety of statistical information to guide their investment recommendations.

In the case of stocks, the analysts review a variety of financial data including price/earnings ratios and dividend yields. By comparing the information for an individual stock with information about the stock market averages, a financial analyst can begin to draw a conclusion as to whether an individual stock is **over- or underpriced**. For example, *in 2012 years* reported that the average dividend yield for the 30 stocks *in certain* Company-A / Industrial Average was 2.45%. Company-B showed a dividend yield of 3.05%. In this case, the statistical information on dividend yield indicates a higher dividend yield for Company-B than the average for Company-A stocks. Therefore, a financial analyst might conclude that Company-B was underpriced. This and other information about Company-B would help the analyst make a buy, sell, or hold recommendation for the stock.

Auditors: Accountants who perform audits benefit greatly from understanding and using **statistical analysis**. For example, when conducting a **reliability assessment**, one of the accountant's first tasks is to gather evidence. Auditors know that the easiest way to do this is by looking at a portion of the whole, rather than gathering every bit of data available. Statistically **representative samples** are preferred in this area as they help auditors work more efficiently and objectively.

Benefit Accountants: Accounting standards are front and center when managers determine retirement and other benefits. Accountants set premium adjustments to account for future risk and account for artificial fluctuations in short-term interest rates using statistical models and methods. Recently, accountants and others used historical statistical data to develop policy recommendations to help control defined benefit plans and promote retirement security.

Controllers: Jacks-of-all-trades, controllers typically work for a single company, overseeing all of its finances including cost analyses, budget reports and forecasting, as well as giving financial analysis and advice to the head of the organization. Having a thorough understanding of the statistical principles used in creating analyses and forecasts, controllers ensure that their organization operates profitably and efficiently.

Forecasters: Accountants use statistics to forecast consumption, earnings, cash flow and book value. Simply put as accounting for the future, forecasting involves an amount of guesswork about the future – and when people guess, they frequently make errors. Having a thorough

understanding of the distribution and metrics for evaluating that error, accountants are better able to more efficiently make predictions about the future.

Mortgage Underwriter :With foreclosures at near record levels, anticipating and predicting the risks associated with any given loan has never been more important. Mortgage underwriters assess that risk, and, therefore, need to have a thorough understanding of statistics in order to set a premium price that is reasonable for the borrower and profitable for the lender.

Risk Managers: Hand in glove with forecasting is risk management. Accountants are frequently required to specify a premium that reflects the risk, or range of error, with any given forecast. Known as the discount rate, accountants often use statistical principles, such as correlation and distribution, to anticipate this risk and account for it when setting a valuation. More recently, accountants are using more sophisticated statistical techniques, such as co-variance and beta models, to limit valuation error.

Research area applications of statistic

Our world and our lives are filled with counting and measuring. It is applied in every sphere of human endeavour – social as well as physical sciences – like, Accounting and finance, Management, Economics, Education, Planning, Politics, Biology, Information Technology, etc. It is almost impossible to find a single department of human activity where statistics is not applicable. It is often described as the art of *decision under uncertainty*.

Here are some of the many ways that statistics play a role in our studies, our work, and our daily lives.

Business

- ✓ A manager maintaining inventory needs to know how many products are being sold at which time of year so she can place orders before she runs out of materials.
- ✓ What's the range of estimated sales next year that has a 95% chance of being correct?
- ✓ What's the reliability or predictability of sales for our 14 sales professionals? Are they similar or are there differences in reliability that go beyond just random fluctuations?
- ✓ Supervisors have to monitor the quality of their production lines and service levels to spot problem areas, inefficient processes, and people who need to improve their knowledge, skills, and performance.

- ✓ Sales managers need to know which customers
 - Buy the most?
 - And which complain the most?
 - Are increasing their purchase levels the fastest?
 - Which salesperson has reduced productivity the most in the last quarter?
 - Is the reduction in sales by this salesperson due to chance alone or could there be something else going on?
- ✓ Marketing managers ask about the millions of dollars spent on several advertising campaigns; are any of them obviously better than the others?
- ✓ Obviously worse?
- ✓ Are the differences just temporary fluctuations or are they something we should take seriously?
- ✓ A health-club manager asks if a particular member's absence is just part of her normal routine or whether they should phone her to see if she would like to suspend her membership payments temporarily as part of the customer service efforts?

Finance

- ✓ Which of several brokerages has a reliable record of higher-than-average return on investment?
- ✓ Is the share price for this firm rising predictably enough for a day-trader to invest in it?
- ✓ What has our return on investment been for these two brands of computer equipment over the last four years?
- ✓ Should we pay the new premium being proposed by our insurance company for liability insurance?
- ✓ What is the premium that our actuaries are calculating for a fire-insurance policy on this type of building?
- ✓ Should we invest in bonds or in stock this year? What are the average rates of return? How predictable are these numbers?
- ✓ Which country has the greatest chance of maximizing our profit on investment over the next decade? Which industry? Which company?

Management

- ✓ One of our division managers claims that his below-average profit figures are just the luck of the draw; how often would he get such a big drop in profit by chance alone if we look at our historical records?
- ✓ Another manager claims that her above-average productivity is associated with the weekly walkabout that is part of her management-by-walking-around philosophy; can we test that idea to see if there is any rational basis for believing her analysis?

Uses of statistics:

The main function of statistics is to enlarge our knowledge of complex phenomena. The following are some general uses of statistics in different fields:

1. It presents facts in a definite and precise form.
2. Data reduction.
3. Measuring the magnitude of variations in data.
4. Furnishes a technique of comparison
5. Estimating unknown population characteristics.
6. Testing and formulating of hypothesis.
7. Studying the relationship between two or more variable.
8. Forecasting future events....etc.

CHAPTER TWO

2. Methods of Data Collection and presentation

2.1 Methods of Data Collection

Data:- is the raw material of statistics. It can be obtained either by measurement or counting. When we determine that the appropriate approach to seeking an answer to a question will require the use of statistics, we begin to search for suitable data to serve as the raw material for our investigation.

2.1.1 Sources of data

The statistical data may be classified under two categories depending up on the sources.

1. **Primary data:** - Data collected by the investigator himself for the purpose of a specific inquiry or study. Such data are original in character & are mostly generated by surveys conducted by individuals or research institutions.

It is more reliable & accurate since the investigator can extract the correct information by removing doubts, if any, in the minds of the respondents regarding certain questions (Questionnaire).

Primary data can be collected by:-

Telephone survey:- Surveys conducted through telephone interviews.

Mailed Questionnaire Method: In the mailed questionnaire method a questionnaire in the form of a set of questions is sent by mail to the informants. They are expected to answer the questions and also the additional needed information, whenever needed and mail them back to the investigator.

This method should be used in the following

- i. When the area under investigation is wide.
- ii. When the informants are educated.
- iii. When the informants are expected to leave for faraway places.

Indirect Personal Interview: In some cases the informants cannot be contacted directly. In this situation an indirect personal inquiry is conducted to get the desired information. The indirect personal investigation is done through some agencies that have some knowledge of the phenomenon under enquiry. This method is useful in the following cases:

- a. When the area of investigation is large.
- b. When the information cannot be obtained directly from the informants.

2. **Secondary data:** - When an investigator uses data, which have already been collected by others, such data are called secondary data. Such data are primary data for the agency that collected them, and become secondary for someone else who uses these data for his own purposes. Example of secondary data: books, reports, magazines, etc.

The sources of secondary data can be classified into two:

- a. Published sources
- b. Un-published sources

Published Sources

Some of the published sources providing secondary data are:

- a. Government Publications:
- b. International Publications:
- c. Semi-official Publications:
- d. Reports of Committees and Commissions:
- e. Private Publications.

Unpublished Sources

In some cases data are collected but these are not put in published form. For example

Research scholars in the institutes and universities, trade associations and labour bureaus do collect data but they never put it in the published form.

When our source is secondary data check that:

- ✓ The type and objective of the situations.
- ✓ The purpose for which the data are collected and compatible with the present problem.
- ✓ The nature and classification of data is appropriate to our problem.
- ✓ There are no biases and misreporting in the published data.

Note: Data which are primary for one may be secondary for the other.

2.2 Methods of Data Presentation

Having collected and edited the data, the next important step is to organize it. That is to present it in a readily comprehensible condensed form that aids in order to draw inferences from it. It is also necessary that the like be separated from the unlike ones.

The presentation of data is broadly classified in to the following two categories:

- **Tabular presentation**
- **Diagrammatic and Graphic presentation.**

The process of arranging data in to classes or categories according to similarities technically is called **classification**.

Classification is a preliminary and it prepares the ground for proper presentation of data.

Definitions:

- **Raw data:** recorded information in its original collected form, whether it be counts or measurements, is referred to as raw data.
- **Frequency:** is the number of values in a specific class of the distribution.
- **Frequency distribution:** is the organization of raw data in table form using classes and frequencies.

There are three basic types of frequency distributions

- Categorical frequency distribution
- Ungrouped frequency distribution
- Grouped frequency distribution

There are specific procedures for constructing each type.

1) Categorical frequency Distribution:

Used for data that can be place in specific categories such as nominal, or ordinal. E.g. Marital status, satisfaction level of customers....

Example: A social worker collected the following data on marital status for 20 accountants. (M=married, S=single, W=widowed, D=divorced)

M	S	D	W	D
S	S	M	M	M
W	D	S	M	M
W	D	D	S	S
S	W	W	D	D

Solution:

Since the data are categorical, discrete classes can be used. There are four types of marital status M, S, D, and W. These types will be used as class for the distribution. We follow procedure to construct the frequency distribution.

Step 1: Make a table as shown.

Class	Tally	Frequency	Percent
(1)	(2)	(3)	(4)
M			
S			
D			
W			

Step 2: Tally the data and place the result in column (2).

Step 3: Count the tally and place the result in column (3).

Step 4: Find the percentages of values in each class by using;

$$\% = \frac{f}{n} * 100 \quad \text{Where } f = \text{frequency of the class, } n = \text{total number of value.}$$

Percentages are not normally a part of frequency distribution but they can be added since they are used in certain types diagrammatic such as pie charts.

Step 5: Find the total for column (3) and (4).

Combing the entire steps one can construct the following frequency distribution.

Class	Tally	Frequency	Percent
(1)	(2)	(3)	(4)
M	<del> </del>	5	20
S	<del> </del> //	7	28
D	<del> </del> //	7	28
W	<del> </del>	6	24

Ungrouped frequency Distribution

-Is a table of all the potential raw score values that could possible occur in the data along with the number of times each actually occurred.

-Is often constructed for small set or data on **discrete variable**.

Constructing ungrouped frequency distribution:

- First find the smallest and largest raw score in the collected data.
- Arrange the data in order of magnitude and count the frequency.
- To facilitate counting one may include a column of tallies.

Example:

The following data represent the number of accountants employed for 20 Organizations.

80	76	90	85	80
70	60	62	70	85
65	60	63	74	75

76 70 70 80 85

Construct a frequency distribution, which is ungrouped.

Solution:

Step 1: Find the range, $\text{Range} = \text{Max} - \text{Min} = 90 - 60 = 30$.

Step 2: Make a table as shown

Step 3: Tally the data.

Step 4: Compute the frequency.

Acc.N ⁰ -	Tally	Frequency
60	//	2
62	/	1
63	/	1
65	/	1
70	////	4
74	/	1
75	//	2
76	/	1
80	///	3
85	///	3
90	/	1

Each individual value is presented **separately**, that is why it is named **ungrouped** frequency distribution.

1) Grouped frequency Distribution:

-When the range of the data is **large**, the data must be grouped in to classes that are more than one unit in width.

Definitions:

- **Grouped Frequency Distribution:** a frequency distribution when several numbers are grouped in one class.

- **Class limits:** Separates one class in a grouped frequency distribution from another. The limits could actually appear in the data and have gaps between the upper limits of one class and lower limit of the next.
- **Units of measurement (U):** the distance between two possible consecutive measures. It is usually taken as 1, 0.1, 0.01, 0.001, -----.
- **Class boundaries:** Separates one class in a grouped frequency distribution from another. The boundaries have one more decimal places than the row data and therefore do not appear in the data. There is no gap between the upper boundary of one class and lower boundary of the next class. The lower class boundary is found by subtracting $U/2$ from the corresponding lower class limit and the upper class boundary is found by adding $U/2$ to the corresponding upper class limit.
- **Class width:** the difference between the upper and lower class boundaries of any class. It is also the difference between the lower limits of any two consecutive classes or the difference between any two consecutive class marks.
- **Class mark (Mid points):** it is the average of the lower and upper class limits or the average of upper and lower class boundary.
- **Cumulative frequency:** is the number of observations less than/more than or equal to a specific value.
- **Cumulative frequency above:** it is the total frequency of all values greater than or equal to the lower class boundary of a given class.
- **Cumulative frequency below:** it is the total frequency of all values less than or equal to the upper class boundary of a given class.
- **Cumulative Frequency Distribution (CFD):** it is the tabular arrangement of class interval together with their corresponding cumulative frequencies. It can be more than or less than type, depending on the type of cumulative frequency used.
- **Relative frequency (rf):** it is the frequency divided by the total frequency.
- **Relative cumulative frequency (rcf):** it is the cumulative frequency divided by the total frequency.

Guidelines for classes

1. There should be between 5 and 20 classes.
2. The classes must be mutually exclusive. This means that no data value can fall into two different classes
3. The classes must be all inclusive or exhaustive. This means that all data values must be included.
4. The classes must be continuous. There are no gaps in a frequency distribution.
5. The classes must be equal in width. The exception here is the first or last class. It is possible to have a "below ..." or "... and above" class. This is often used with ages.

Steps for constructing Grouped frequency Distribution

1. Find the largest and smallest values
2. Compute the Range(R) = Maximum - Minimum
3. Select the number of classes desired, usually between 5 and 20 or use Sturges rule
 $k = 1 + 3.32 \log n$ where k is number of classes desired and n is total number of observation.
4. Find the class width by dividing the range by the number of classes and rounding up, not off.
 $w = \frac{R}{k}$.
5. Pick a suitable starting point less than or equal to the minimum value. The starting point is called the lower limit of the first class. Continue to add the class width to this lower limit to get the rest of the lower limits.
6. To find the upper limit of the first class, subtract U from the lower limit of the second class. Then continue to add the class width to this upper limit to find the rest of the upper limits.
7. Find the boundaries by subtracting $U/2$ units from the lower limits and adding $U/2$ units from the upper limits. The boundaries are also half-way between the upper limit of one class and the lower limit of the next class. !may not be necessary to find the boundaries.
8. Tally the data.
9. Find the frequencies.
10. Find the cumulative frequencies. Depending on what you're trying to accomplish, it may not be necessary to find the cumulative frequencies.
11. If necessary, find the relative frequencies and/or relative cumulative frequencies

Example: The following data is the salary of accountants in dollar (\$) per day in certain organization. Construct a frequency distribution for the following data.

11 29 6 33 14 31 22 27 19 20
18 17 22 38 23 21 26 34 39 27

Solutions:

Step 1: Find the highest and the lowest value $H=39$, $L=6$

Step 2: Find the range; $R=H-L=39-6=33$

Step 3: Select the number of classes desired using Sturges formula;

$$k = 1 + 3.32 \log n = 1 + 3.32 \log (20) = 5.32 = 6 (\text{rounding up})$$

Step 4: Find the class width; $w=R/k=33/6=5.5=6$ (rounding up)

Step 5: Select the starting point, let it be the minimum observation.

- 6, 12, 18, 24, 30, 36 are the lower class limits.

Step 6: Find the upper class limit; e.g. the first upper class $=12-U=12-1=11$

- 11, 17, 23, 29, 35, 41 are the upper class limits.

So combining step 5 and step 6, one can construct the following classes.

Class limits

6 – 11

12 – 17

18 – 23

24 – 29

30 – 35

36 – 41

Step 7: Find the class boundaries;

E.g. for class 1 Lower class boundary= $6-U/2=5.5$

Upper class boundary = $11+U/2=11.5$

- Then continue adding w on both boundaries to obtain the rest boundaries. By doing so one can obtain the following classes.

Class boundary

5.5 – 11.5

11.5 – 17.5

17.5 – 23.5

23.5 – 29.5

29.5 – 35.5

35.5 – 41.5

Step 8: tally the data.

Step 9: Write the numeric values for the tallies in the frequency column.

Step 10: Find cumulative frequency.

Step 11: Find relative frequency or/and relative cumulative frequency.

The complete frequency distribution follows:

Class limit	Class boundary	Class Mark	Tally	Freq.	Cf (less than type)	Cf (more than type)	rf.	rcf (less than type)
6 – 11	5.5 – 11.5	8.5	//	2	2	20	0.10	0.10
12 – 17	11.5 – 17.5	14.5	//	2	4	18	0.10	0.20
18 – 23	17.5 – 23.5	20.5	///	7	11	16	0.35	0.55
24 – 29	23.5 – 29.5	26.5	////	4	15	9	0.20	0.75

30 – 35	29.5 – 35.5	32.5	///	3	18	5	0.15	0.90
36 – 41	35.5 – 41.5	38.5	//	2	20	2	0.10	1.00

Contingency tables (Two Way Table)

Relationships between two categorical variables can be shown through a two-way table (also known as a contingency table, cross tabulation table or a cross classifying table).

This led to many influential reports on the topic, including the study of Saving habits of 4019 students in Debre Markos University in 2013 years. Here is a two-way table summarizing some of the results. Is a saving habit of family affects saving habit of students? The information result is given below.

Parents save	Students Save	Students donot save	Total
Parents save	400	1380	1780
Parents don not save	416	1823	2239
total	816	3203	4019

- ✓ The row variable is the parents saving habits.
- ✓ The column variable is the student saving habits.
- ✓ The cells display the counts for each of the categories of row and column variables.

A two-way table with **r** rows and **c** columns is often called an **r by c** table (written $r \times c$).

The totals along each of the rows and columns give the **marginal distributions**.

Scatter plots

We now consider two dimensional data. The values of the first variable $x_1; x_2; \dots; x_n$ are assumed **known** and in an experiment and are often set by the experimenter. This variable is called the **explanatory**, predictor, regressors or input variables and in a two dimensional scatter plot of the data display its values on the horizontal axis. The values

$y_1; y_2; \dots; y_n$, taken from observations with input $x_1; x_2; \dots; x_n$ are called the **response** or target variable and its values are displayed on the vertical axis. In describing a scatterplot, take into consideration. In describing a scatterplot, take into consideration

- The **form**, for example,
 - ✓ linear
 - ✓ curved relationships
 - ✓ clusters

- The **direction**,
- ✓ a positive or negative association
- ✓ and the strength of the aspects of the scatterplot.

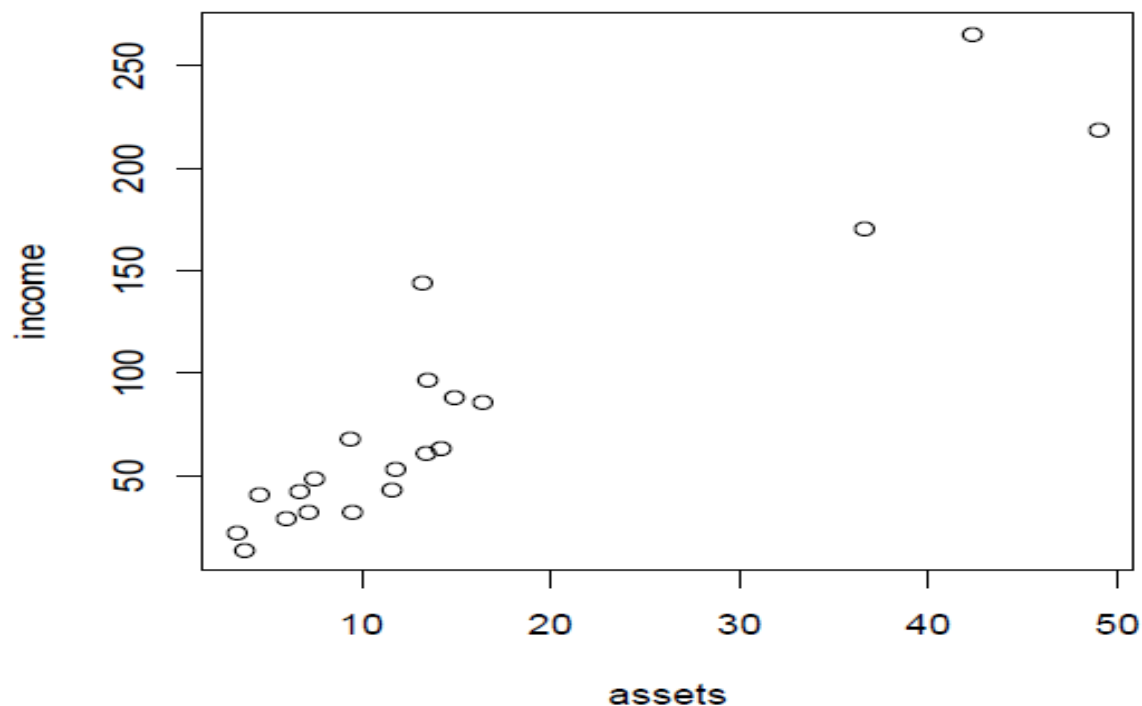
Example: This historical data show the 20 largest banks in 2013. Values given in billions of dollars.

Bank	1	2	3	4	5	6	7	8	9	10
Assets	49.0	42.3	36.6	16.4	14.9	14.2	13.5	13.4	13.2	11.8
Income	218.8	265.6	170.9	85.9	88.1	63.6	96.9	60.9	144.2	53.6
Bank	11	12	13	14	15	16	17	18	19	20
Assets	11.6	9.5	9.4	7.5	7.2	6.7	6.0	4.6	3.8	3.4
Income	42.9	32.4	68.3	48.6	32.2	42.7	28.9	40.7	13.8	22.2

Questions?

1. Identify the predictor and response variables?
2. Describe the scatter plots
3. What would you expect to be a feature on this scatter plot of a failing bank?
4. Is the increase of assets increases the income of Bank?

Income vs. Assets (in billions of dollars)



Scatter plot of Income Versus Assets

Diagrammatic and Graphic presentation of data.

These are techniques for presenting data in **visual displays** using **geometric** and **pictures**.

Importance:

- They have greater attraction.
- They facilitate comparison.
- They are easily understandable.

-Diagrams are appropriate for presenting **discrete data**.

-The three most commonly used diagrammatic presentation for discrete as well as qualitative data are:

- Pie charts
- Pictogram
- Bar charts

Pie chart

- A pie chart is a circle that is divided into sections or wedges according to the percentage of frequencies in each category of the distribution.
- The angle of the sector is obtained using:

$$\text{Angle of sector} = \frac{\text{Value of the part}}{\text{the whole quantity}} * 100$$

Note that pie-charts are usually used for depicting **nominal level data**.

Example: A survey showed that a car owner spends birr 2,950 per year on operating expenses. Below is the breakdown of the various expenditure items. Draw an appropriate chart to portray the data.

Expenditure item	Amount (in birr)
Fuel	603
Interest on car loan	279
Repairs	930
Insurance and license	646
Depreciation	492
Total	2,950

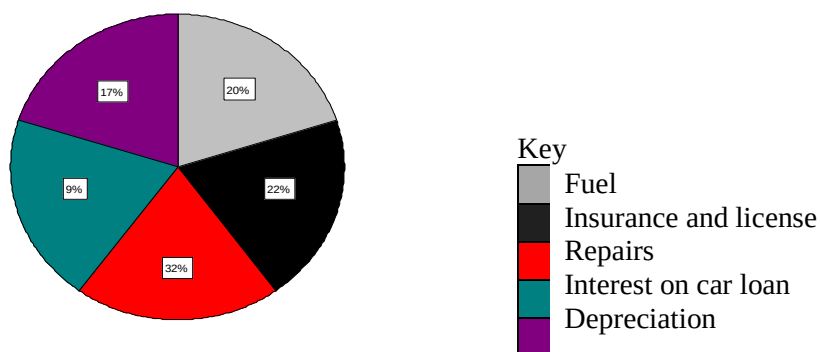
How to draw a pie-chart

- First find the percentages of each class
- Next calculate the degree measures for each class

- Finally, using a protractor, put each sector /degree measure/ in a circle and give a key for explanation.

Expenditure item	Amount (in birr)	Percentage (approx)	Degree (approx)
Fuel	603	20	74
Interest on car loan	279	9	34
Repairs	930	32	113
Insurance and license	646	22	79
depreciation	492	17	60
Total	2,950	100	360

Now we can draw the pie-chart for the data.



Exercise: Draw a suitable diagram to represent the following population in a town.

Men	Women	Girls	Boys
2500	2000	4000	1500

Pictogram

-In this diagram, we represent data by means of some picture symbols. We decide about a suitable picture to represent a definite number of units in which the variable is measured.

Example: Draw a pictorial diagram to present the following data (number of students in a certain school for four years.)

Year	2010	2011	2012	2013
No. of students	2000	3000	5000	7000

Let a single picture () represents one thousand students.

2010	     
2011	    
2012	  
2013	 

Key:  = 1000 students

Exercise: Draw a pictogram to represent the following population of a Kebele.

Year	2010	2011	2012	2013
Population	2000	3000	5000	7000

Bar Charts:

- A set of bars (thick lines or narrow rectangles) representing some magnitude over time space.
- They are useful for comparing aggregate over time space.
- Bars can be drawn **either vertically or horizontally**.
- There are different types of bar charts.

The most common are:

- Simple bar chart
- Component or sub divided bar chart.
- Multiple bar charts.
- Deviation or two way bar chart
- Broken bar chart

Simple Bar Chart

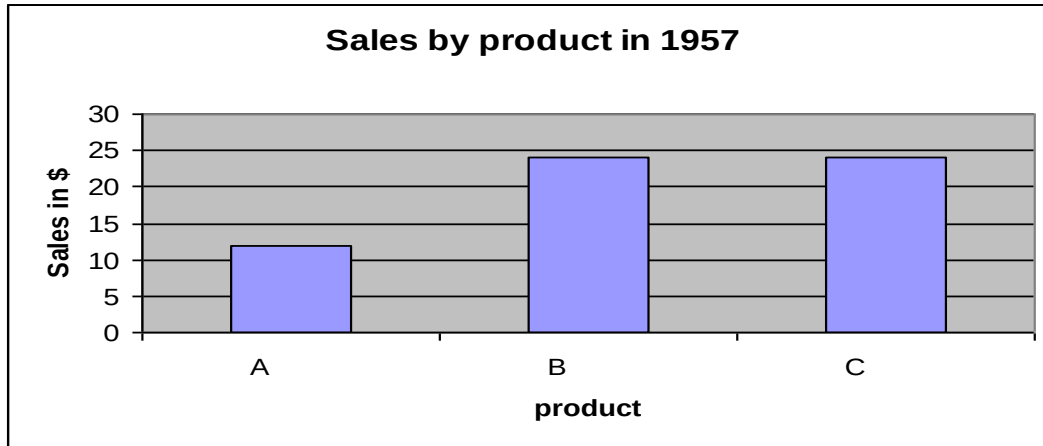
-Are used to display data on **one variable**.

-They are thick lines (narrow rectangles) having the same breadth. The magnitude of a quantity is represented by the height /length of the bar.

Example: The following data represent sale by product, 1957- 1959 of a given company for three products A, B, C.

Product	Sales(\$) In 1957	Sales(\$) In 1958	Sales(\$) In 1959
A	12	14	18
B	24	21	18
C	24	35	54

Solutions:



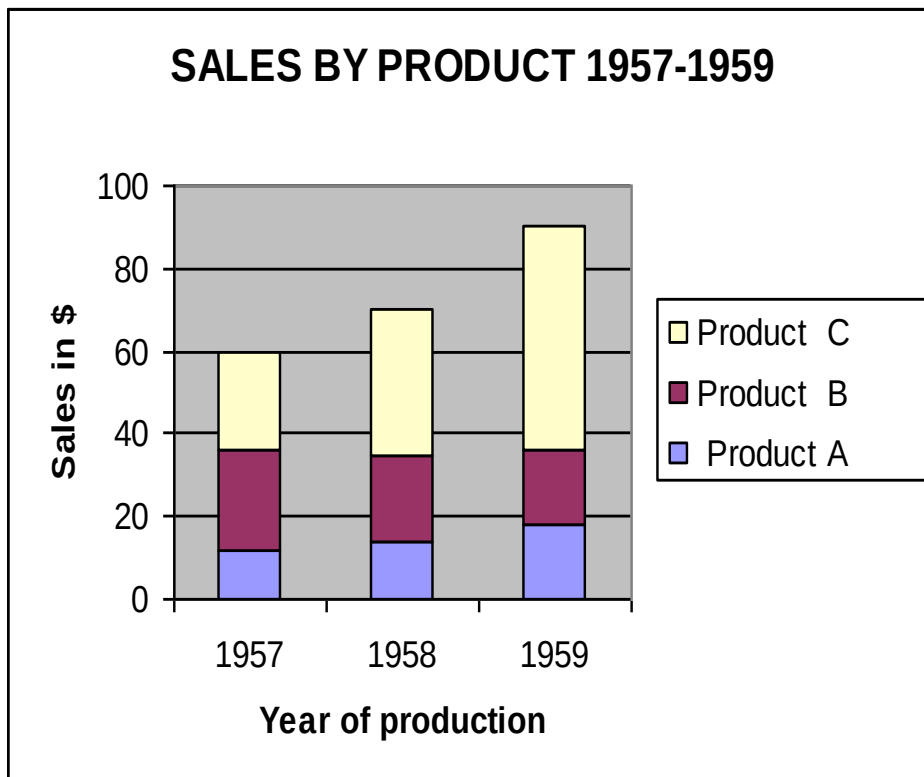
Component Bar chart

-When there is a desire to show how a total (or aggregate) is divided in to its component parts, we use component bar chart.

-The bars represent total value of a variable with each total broken in to its component parts and different colours or designs are used for identifications

Example: Draw a component bar chart to represent the sales by product from 1957 to 1959.

Solutions:



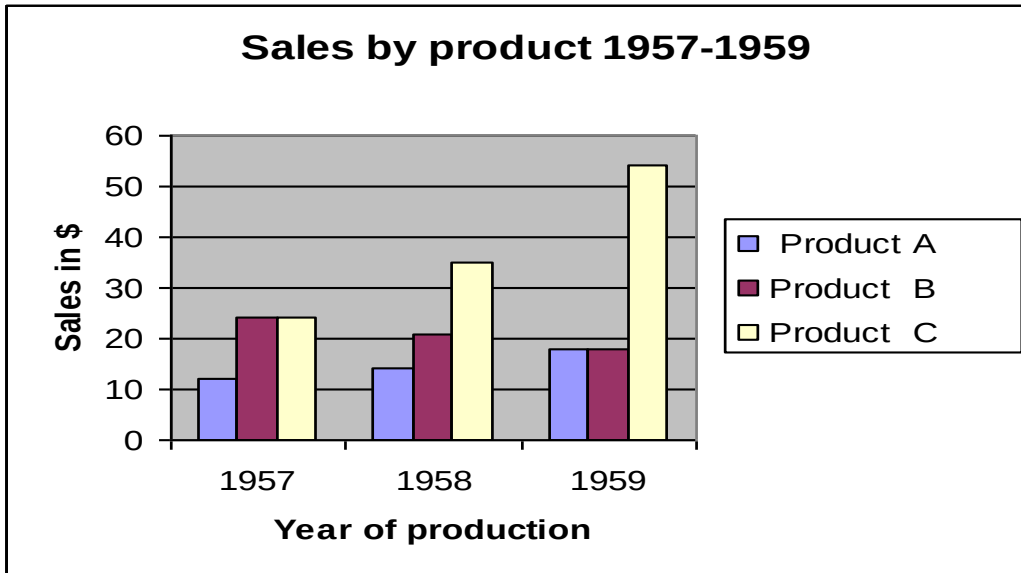
Multiple Bar charts

- These are used to display data on more than one variable.
- They are used for comparing different variables at the same time.

Example:

Draw a component bar chart to represent the sales by product from 1957 to 1959.

Solutions:



Graphical Presentation of data

A graphical display helps to visualize the data and obtain a preliminary understanding of the data distribution. Stem-and-Leaf plots, bar graphs, histograms, and box plots are popular graphical methods.

- The Stem-and-Leaf plots, histogram, frequency polygon and cumulative frequency graph or ogive is most commonly applied graphical representation for **continuous data**.

Stem-and-Leaf plots

This is a simple graphical representation of a data set. It shows all data information and their distribution. Hence is suitable for relatively **smaller data** sets. This is prepared in following steps:

1. Separate each data point into a stem component and a leaf component.
2. Write all stem digits left of a vertical line.
3. Write all leaf digits right of the vertical bar.
4. Re-arrange each leaf digit

Example: No. of accountants employed in 13 Banks in 2012 who were graduated in 2010-2011 years from Debre Markos University are given below:

6 24 7 9 15 10 31 6 23 7 13 11 4.

Draw a stem-and-leaf diagram for this data.

Solution:

Step 1: Separate the data into stem (10 digit) and leaf(unit digit) components.

Step 2 & 3: Write stem digits to the left and leaf digits to the right

Step 4: Order the leaf digits.

The decimal point is 1 digit(s) to the right of the

0 | 4667791

1 | 01352

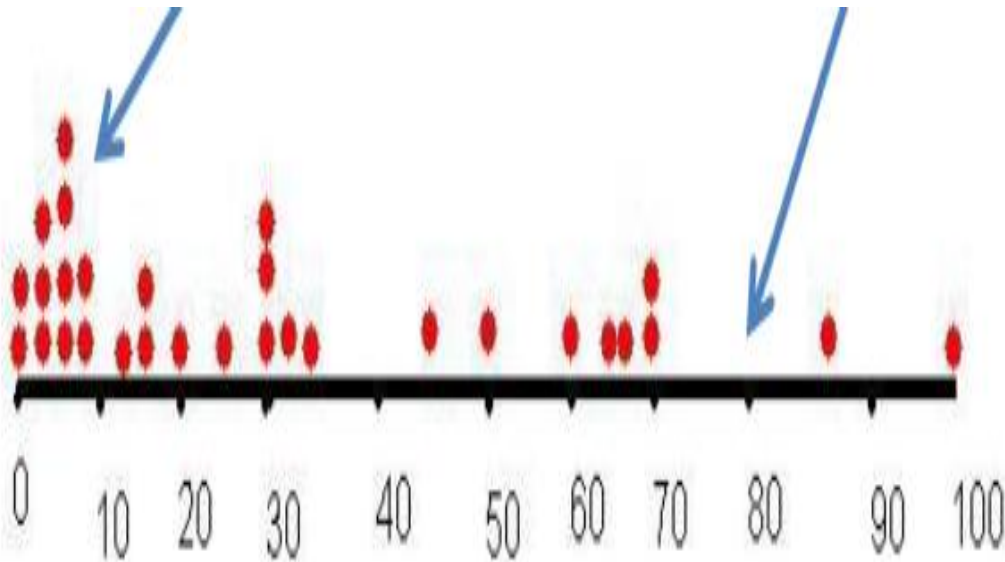
2 | 343

3 | 1

Dot Plots

A data display in which each data item is shown as a dot above a number line

- In a dot plot a cluster shows where a group of data points fall.
- A gap is an interval where there are no data items.



Steps to Create a Dot Plot

1. Order numbers from least to greatest.
2. Draw a number line, labeling the number line with the minimum and the maximum and then all the numbers that fall between them.
3. Put a dot above each number on the number line for each data entry in your set.

Don't forget a title and labels!

Example: For the purpose of employing the new graduated accountants in Commercial Bank of Ethiopia training program was prepared, the new graduated accountants are given a test in which they are given a set of tasks and the time it takes them to complete the tasks is measured. The following is a list of the time (in seconds) for a group of new trainees. 61, 61, 64, 67, 70, 71, 71, 71, 72, 73, 74, 74, 75, 77, 79, 80, 81, 81, 83.

Display the data in a dot plot.

Solution



Other types of Graphs

Procedures for constructing statistical graphs:

- Draw and label the X and Y axes.
- Choose a suitable scale for the frequencies or cumulative frequencies and label it on the Y axes.
- Represent the class boundaries for the histogram or ogive or the mid points for the frequency polygon on the X axes.
- Plot the points.
- Draw the bars or lines to connect the points.

Histogram

A graph which displays the data by using vertical bars with various height to represent frequencies. **Class boundaries** are placed along the horizontal axes. Class marks and class limits are sometimes used as quantity on the X-axes.

The feature of this histogram is that the rectangles are adjacent to each other without any gap amongst them.

Exercise: Draw a histogram presenting the following previous data in tabulation part.

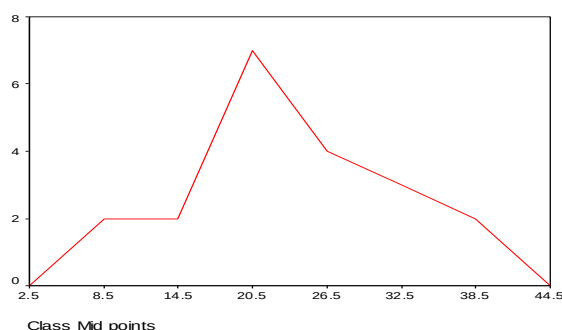
Class Boundaries	Class Mark	Frequency	Cumulative Frequency (less than type)	Cumulative Frequency (more than type)
5.5 – 11.5	8.5	2	2	20
11.5 – 17.5	14.5	2	4	18
17.5 – 23.5	20.5	7	11	16
23.5 – 29.5	26.5	4	15	9
29.5 – 35.5	32.5	3	18	5
35.5 – 41.5	38.5	2	20	2

Frequency Polygon:

- The frequency is placed along the vertical axis and classes mid points are placed along the horizontal axis. It is customer to the next higher and lower class interval with corresponding frequency of zero, this is to make it a complete polygon.

Example: Draw a frequency polygon for the above data (example *).

Solutions:



Ogive (cumulative frequency polygon)

- A graph showing the cumulative frequency (less than or more than type) plotted against upper or lower class boundaries respectively. That is class boundaries are plotted along the horizontal axis and the corresponding cumulative frequencies are plotted along the vertical axis. The points are joined by a free hand curve.

Example: Draw an ogive curve(less than type and more than type) for the above data.

- Less than type cumulative frequency polygon
- More than type cumulative frequency polygon

