

UNIVERSITY OF RWANDA



College of Science and Technology

School of ICT

Department of Computer Science

**Applied probability and statistics for
Level 2, computer science**

Academic year: 2015-2016

By
Dr Gahamanyi Marcel

Content

Quick Review of Descriptive Statistics

Representation of data--Frequency distribution-Histograms-Frequency polygon-Cumulative Frequency Polygon-Measure of Central tendency (Mean, Mode and Median)-Grouped and Ungrouped data-Measure of dispersion (Range, Standard deviation)-Peaked ness (Skew ness, Kurtosis)

Elementary Probability theory and Distributions

Definition of Probability-Conditional probability (Baye's theorem)

Random variables-Probability distribution-Some discrete and continuous probability -distribution (Binomial, Poisson and normal distributions)-Mathematical expectation and Variance

Hypothesis Testing

Significance of hypothesis testing-Type I and II errors-Tests involving distributions-One and two tail tests-Tests for large and small samples-Goodness of fit-Chi square-distribution-Student t distribution

Regression Analysis and Correlation

Curve fitting by least squares method-Pearson's coefficient correlation-Confidence-intervals for the regression coefficients-Auto and cross correlations

Quality control

Detecting process change-Control chart-X chart-Runs Analysis-P-chart-C-chart-Tolerant limits-Acceptance sampling for defectives.

Part I: Introduction to statistics

Course Objective:

Statistics is essentially a decision making tool. This introductory course imparts knowledge on fundamental statistical concepts and how to convert *data* into *information* which enable managers to make informed decisions.

Course Outline:

Chapter 1. Introduction to basic concepts -Statistics-What and Why, Statistics Defined, Descriptive Statistics, Inferential Statistics; statistical model building, Role of Probability, random sampling, , Population, Sample, Parameter, Statistic, Variable, Statistical Data, Quantitative Data, Qualitative Data, Discrete and Continuous data, Level of Measurement-Nominal, Ordinal, Interval and Ratio, Data Sources ,scope of statistics, statistical applications.

Chapter 2. Graphical and Tabular presentations of data.

2.1 Univariate descriptive statistics- Presenting frequencies: Frequency distribution tables, Formation of a discrete frequency distribution, Formation of a continuous frequency distribution-class limits, class intervals, class frequency, and class mid-point. Cumulative and relative frequency distribution, graphical forms, discrete-bar charts and Ogive, continuous- Histogram and Ogive.

2.2 Bivariate descriptive statistics- Presenting frequencies: Bivariate frequency distribution, Cross tabulations, scatter diagrams.

Chapter 3. Univariate descriptive Measures:

3.1 Measures Of location (central Value)- Mode ,Median, Arithmetic mean, Weighted Arithmetic mean, Geometric mean, Quadratic mean, Harmonic mean,

3.2 Measures of Dispersion:The Range, Inter Quartile Range, Quartile Deviation, Coefficient of Quartile Deviation, The Variance, The Standard Deviation, Co-efficient of Variation, Percentiles, Chebycheff's rule, Empirical rule.

Chapter 4. Bivariate descriptive statistics: Measures of linear relationship - covariance, coefficient of correlation, Least squares method, Regression line.

Chapter 1: Introduction to Basic concepts of Statistics

1.1 What is statistics?

The term *statistics*, derived from the word *state*, was used to refer to a collection of facts of interest to the State. The idea of collecting data spread from the descriptive science of states and became increasingly identified with numbers as we see it in the following.

Statistics is a way to get information from data. It is concerned with the collection of data, its subsequent description, and its analysis, which often leads to the drawing of conclusions.

Statistics is a body of methods and theory that is applied to quantitative data (numerical evidence) when making decisions in the face of uncertainty. It enables us to recognize and evaluate the errors involved in quantifying our experience, especially when generalizing from what is known of some small group (a sample) to some wider group (the population).

It (Statistics) is the science of collecting, organizing, analyzing and interpreting data in order to make decisions. Quantitative Methods (Statistics) are a set of mathematical techniques used by social scientists (and managers) to organize and manipulate data for the purpose of answering questions, testing theories and making managerial decisions.

Generally, there are the two bodies of methods that together constitute the subject called statistics: *descriptive statistics* and *inferential statistics*.

1.2 Descriptive Statistics

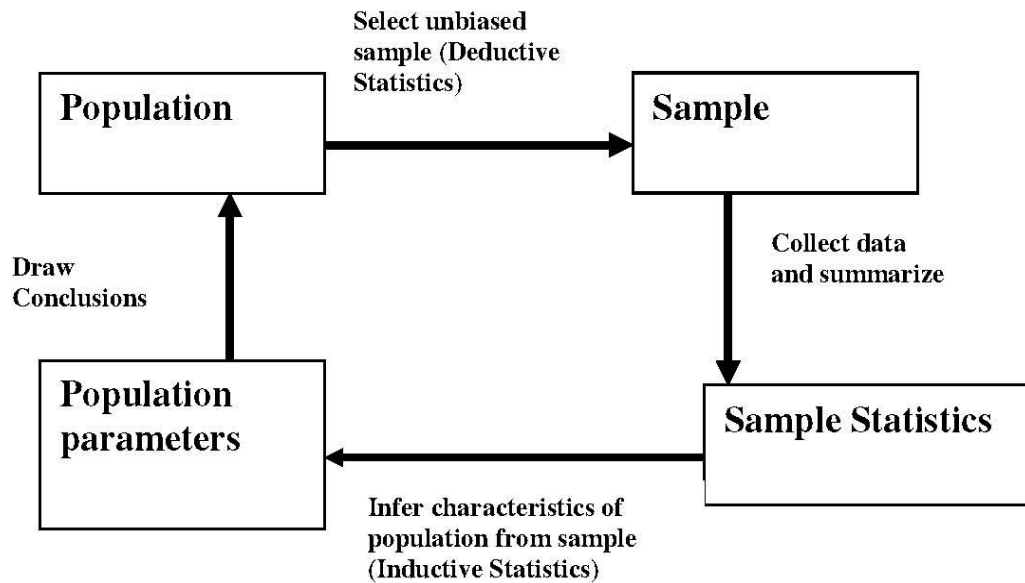
Descriptive statistics deals with methods for organizing, summarizing, and presenting data in a convenient and usable form. One form of descriptive statistics uses graphical techniques, which allow statistics practitioners to present data in ways that make it easy for the reader to extract useful information. Another form of descriptive statistics uses numerical techniques such as measures of central location (mean, mode, median) and measures of variability (range, variance, standard deviation, etc) to summarize data.

1.3 Inferential Statistics

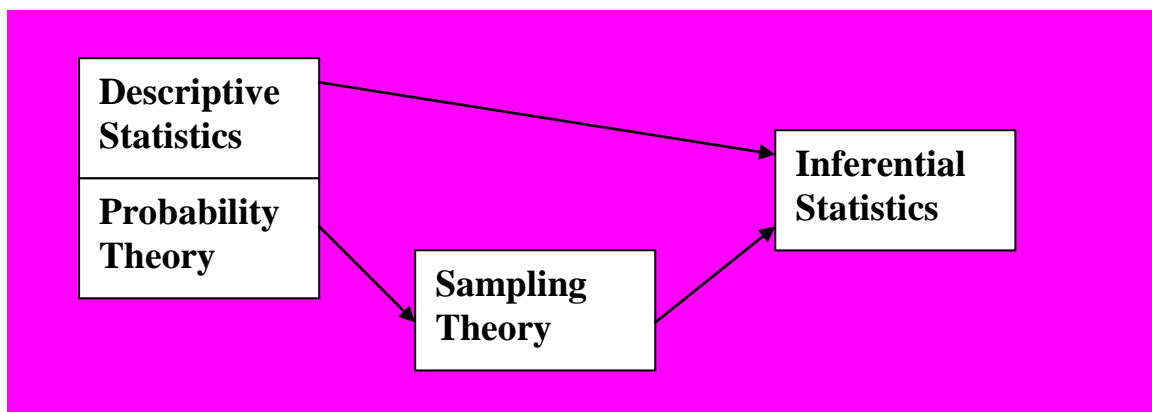
Inferential statistics consists of a body of methods for drawing conclusions or inferences about characteristics of a population based on information contained in a sample taken from the population.

A major contribution of statistics is that data from a sample can be used to make estimates and test hypotheses about the characteristics of a population because populations are almost very large; investigating each member of the population would be impractical and expensive. However, such conclusions and estimates are not going to be correct. For this reason, we build into the statistical inference a measure of reliability. There are two such measures, the *confidence level* and the *significance level*.

This process of making an estimate, prediction, or decision about a population based on sample data is referred to as statistical inference (see figure below).



In inferential statistics, we have to build a statistical model (margin of errors, confidential interval, level of significance) to provide a statement of quality, then probability and sampling theory are gathered, then descriptive statistics which inferences will be based on are computed. See the figure below.



1.4 Some Important Definitions

- A **Population** (Universe): the population refers to the collection of all items or elements of interest in a particular study. In the language of statistics, population does not necessarily refer to a group people. It may, for example, refer to the salaries of all workers in a community.

For example, an election is to be held soon to determine who will be the mayor of a certain city. Based on the results of a survey of 400 eligible voters, a newspaper has reported the proportion of eligible voters who favor Ms. John, one of the candidates. The population is the collection of all residents of the city who are eligible to vote in the election.

- A **Parameter** is a summary descriptive measure computed to describe a characteristic of the Population, such as the average salary of workers in a community. In the example 1.1, the population parameter is the proportion of all eligible voters who favor Ms. John.

- A **Sample**: this refers to a subset or a portion of the entire population selected for analysis. In the example above, the 400 eligible voters surveyed by the pollster constitute the sample.
- A **Statistic** is a summary descriptive measure computed to describe a characteristic of the sample. In the example above, the sample statistic is the proportion of these 400 eligible voters who favor Ms. John.
- A **variable** is a characteristic of interest for the elements of a population or a sample. For example, the mark on a statistics exam is a characteristic of statistics exams that certainly is of interest and the marks will vary from student to student, thus the name *variable*.

The most important characteristic of a variable is that it can change; in other words, it can take more than one value, either across entities (**cross section data**) or within the same entity over time (**time series data**).

Furthermore, these different values are capable of being observed and/or measured. For e.g. 'age' and 'qualifications' are variables, which are attributes of individuals. They can vary from one person to another and if you are considering only one individual, over time.

- **Data** are the observed values of a variable. They are the facts and figures that are collected, summarized, analyzed, and interpreted. The data collected in a particular study are referred to as the data set.
- **Values** of a variable are the possible observation of the variable.

The total number of data values in a **data set** is the number of elements multiplied by the number of variables. The *data set* is all the data collected in a particular study.

- **The elements are** the entities on which data are collected e.g. each student in the class is an element.
- **Cross sectional data** are data collected at a certain point in time, for example, test score in a statistics course.
- **Time series data** is collected over successive points in time, for example, the amount of crude oil imported monthly in Rwanda.

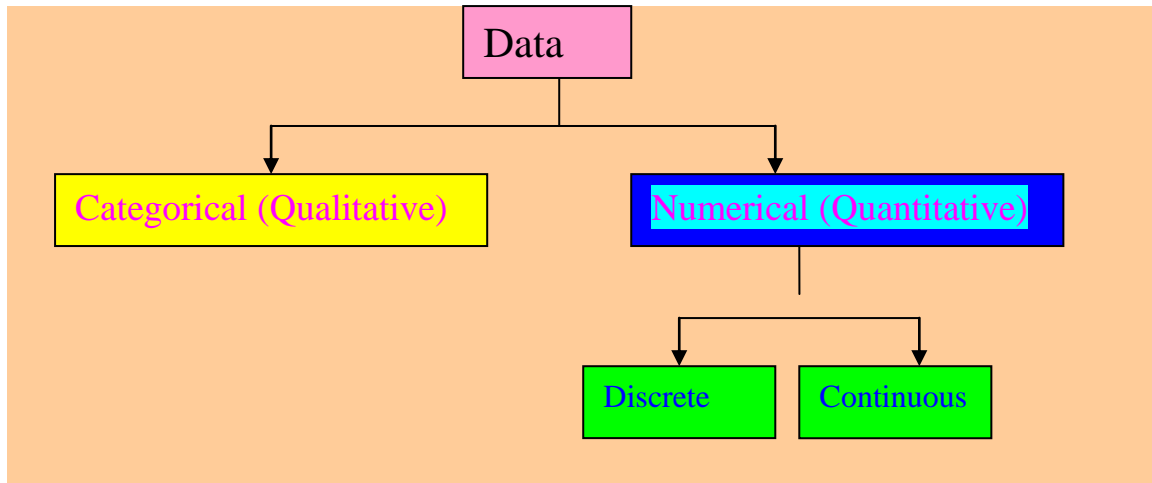
1.5 Types of data and level of measurement of variables

1.5.1 Types of data

When most people think of data, they think of sets of numbers. However, there are two main types of data: *numerical data (quantitative)* and *categorical data (qualitative)*.

The statistical analysis that is appropriate depends on whether the data for the variable are qualitative or quantitative.

Figure: types of statistical data



❖ Qualitative Data

Qualitative data, often referred to as categorical data, are labels or names used to identify an attribute of each element. Qualitative data are categorical observations and qualitative variable is a non-numerical attribute of an individual or ‘object’. For example Gender, color, sex, etc.

Qualitative data use either the nominal or ordinal scale of measurement and they may be either numeric or nonnumeric. The statistical analyses for qualitative data are rather limited.

❖ Quantitative Data

Quantitative data require numeric values that indicate either how many or how much. They should be *discrete* or *continuous*. Quantitative data are said to be **discrete** when they take values that are isolated, there is separation between the possible values for the data, for example the number of children per family.

Quantitative data are said to be **continuous** when there is no separation between the possible values for the data. A variable can take any other intermediate values between two given values, for example the test scores. The figure above shows the types of statistical data.

1.5.2 Types of measurement scales

Data collection requires one of the following scales of measurement: **nominal**, **ordinal**, **interval**, or **ratio**. The scale of measurement determines the amount of information contained in the data and indicates the most appropriate data summarization and statistical analysis. The figure below shows the hierarchy of level of measurement for a variable:

Figure: The hierarchy of measurement scales.

RATIO		Absolute zero
INTERVAL		Distance is meaningful
ORDINAL		Attributes can be ordered
NOMINAL	Attributes are only named	

- ❖ When the data for a variable consist of labels or names used to identify an attribute of the element, the scale of measurement is considered a **nominal scale**. For analysis, a nonnumeric label or a numeric code may be used but the scale of measurement remains nominal even though the data appear as numeric values.

Example: Students of a university are classified by the school in which they are enrolled using a nonnumeric label such as Business, Humanities, Education, and so on.

Alternatively, a numeric code could be used for the school variable (e.g. 1 denotes Business, 2 denotes Humanities, 3 denotes Education, and so on).

Others examples of nominal scale variables: sex, marital status (single-1, married-2, Divorced-3, widowed-4), religion, race, company's stock, etc.

- ❖ The scale of measurement of a variable is called an **ordinal scale** if the data exhibit the properties of nominal data and the order or rank of the data is meaningful. A nonnumeric label or a numeric code may be used when recording values.

Examples: Students of a university are classified by their class standing using a nonnumeric label such as *freshman*, *sophomore*, *Junior*, or *Senior*.

Alternatively, a numeric code could be used for the class standing variable (e.g. 1 denotes freshman, 2 denotes sophomore, and so on).

Again, for example, at the completion of most college and university courses, students are asked to evaluate the course. The variables are the ratings of various aspects of the course, including the professor. Suppose that in a particular college the values poor, fair, good, very good, and excellent. We can record students' evaluation as poor-1, fair-2, good-3, very good-4, and excellent-5 or we can use other codes by respecting the order.

- ❖ The scale of measurement for a variable becomes an **interval scale** if the data have the properties of ordinal data and the interval between values is expressed in terms of a fixed unit of measure. Interval data are always numeric.

Examples:

1. Three students with economics exam scores of 18, 13, and 16 can be ranked or ordered in terms of best performance to poorest performance. In addition, the differences between the scores are meaningful.

For instance, student 1 scored $18 - 13 = 5$ points more than student 2.

2. Melissa has an SAT score of 1205, while Kevin has an SAT score of 1090. Melissa scored 115 points more than Kevin.

- ❖ The scale of measurement of a variable is a **ratio scale** if the data have all the properties of interval data and the ratio of two values is meaningful. Variables such as distance, height, weight, and time use the ratio scale.

This scale must contain a zero value that indicates that nothing exists for the variable at the zero point. For example, consider the cost of an automobile. A zero value for the cost would indicate that the automobile has no cost and is free.

In addition, if we compare the cost of Rwf 30,000 for one automobile to the cost of Rwf 15,000 for a second automobile, the ratio property shows that the first automobile is $\text{Rwf } 30,000 / \text{Rwf } 15,000 = 2$ times, or twice, the cost of the second automobile.

Other example: Melissa's college record shows 36 credit hours earned, while Kevin's record shows 72 credit hours earned. Kevin has twice as many credit hours earned as Melissa.

In short words:

Interval or ratio data: values are real numbers; all calculations are valid; and data may be treated as ordinal or nominal.

Ordinal data: values must present the ranked order of the data; calculations based on an ordering process are valid; data may be treated as nominal but not as interval or ratio.

Nominal: values are arbitrary numbers that represent categories; only calculations based on the frequencies of occurrence are valid; data may not be treated as ordinal or interval.

Exercises.

1. A psychologist has interviewed 250 school children throughout New York State and found that 80% of them spend at least 25 hours a week watching television.

a) Identify the population

b) What is the sample?

c) Identify the population parameter and the sample statistic of interest here.

d) Comment on the following inference, which is based on the results of the psychologist's interviews: 80 percent of American school children spend at least 25 hours a week watching television.

2. What should be a scale measurement for variable "temperature?" Explain your answer.

3. Complete the following table by yes or not in blank spaces for the possible operations at different measurement scales:

Table -Operations at the four levels of measurement				
Levels of Measurements	Put data in Categories	Arrange data in order	Subtract data values	Determine if one Data value is a multiple of another
NOMINAL				
ORDINAL				
INTERVAL				
RATIO				

4. For each of the following examples of data, determine whether the data type is quantitative, qualitative, or ranked.

a) The weekly level of the prime interest rate during the past year

b) The make of car driven by each of a sample of executives

c) The number of contacts made by each of a company's salespersons during a week

d) The rating (excellent, good, fair, or poor) given to a particular television program by each of a sample of viewers

e) The number of shares traded on the New York Stock Exchange each week throughout 1987.

Solutions

1.a) The population is made of all school children of New York State.

b) The sample is made of 250 school children interviewed.

c) The population parameter is the proportion of all school children that spend at least 25 hours a week watching television. The sample statistic is the proportion of these 250 who spend at least 25 hours a week watching television and is equal to 80% of them.

d) The inference, which is based on the results of the psychologist's interviews: 80 percent of American school children spend at least 25 hours a week watching television is not true because the data collected were based on only one State rather than for a sample which represents all States of USA.

2. The scale measurement for Fahrenheit temperature is interval because one should compute the difference between two temperature values or you can add them. The variable temperature should not be ratio scale of measurement because the ratio of two temperature values is meaningless.

3.

Table -Operations at the four levels of measurement				
Levels of measurements	Put data in categories	Arrange data in order	Subtract data values	Determine if one data value is a multiple of another
NOMINAL	YES	No	No	No
ORDINAL	YES	YES	No	No
INTERVAL	YES	YES	YES	No
RATIO	YES	YES	YES	YES

4.a) Quantitative, if the interest rate level is expressed as a percentage. If the level is simply observed as being high, moderate, or low, then the data type is qualitative.

b) Qualitative

c) Quantitative

d) Ranked, because the categories can be ordered

e) Quantitative

1.6 Data sources and data acquisition considerations

Data can be obtained from existing sources or from surveys and experimental studies designed to collect new data. One can ask why we need data. The answer should be that we need data to provide input to survey, to provide input to study, to measure performance of ongoing service or production process, to evaluate conformance to standards, to assist in formulating alternative courses of action, or to satisfy curiosity.

1.6.1 Existing Sources

In some cases, data needed for a particular application might already exist. Within a firm, detailed information is often kept on customers, suppliers, and employees for example.

Substantial amounts of business and economic data are available from organizations that specialize in collecting and maintaining data.

Government agencies are another important source of data.

Data are also available from a variety of industry associations and special-interest organizations.

Currently, the *Internet* has become an important source of data. Most government agencies, like the Bureau of the Census (www.census.gov), make their data available through a web site. More and more

companies are creating web sites and providing public access to them. A number of companies now specialize in making information available over the Internet.

1.6.2 Statistical Studies

Sometimes the data needed for a particular application are not available through existing sources. In such cases, the data can often be obtained by conducting a statistical study.

Statistical studies can be classified as either *experimental* or *observational*.

In *experimental studies* the variables of interest are first identified. Then one or more factors are controlled so that data can be obtained about how the factors influence the variables. For example, a pharmaceutical firm might be interested in conducting an experiment to learn about how a new drug affects blood pressure.

The blood pressure is the variable of interest. To obtain the effect of new drug, researchers select a sample of individuals.

The dosage level of new drug is controlled, as different groups of individuals are given different dosage levels. Statistical analysis of the experimental data can determine how the new drug affects blood pressure.

In *observational (non-experimental)*, statistical studies make no attempt to control or influence the variables of interest. A survey is perhaps the most common type of observational study.

1.6.3 Data acquisition considerations

❖ Time Requirement

- Searching for information can be time consuming.
- Information might no longer be useful by the time it is available.

❖ Cost of Acquisition

Organizations often charge for information even when it is not their primary business activity.

❖ Data Errors

Using any data that happens to be available or that were acquired with little care can lead to poor and misleading information.

1.7 Functions and Scope of Statistics

a) Functions

The Following are the important functions of the science of statistics:

- _ It presents facts in a definite form (definiteness)
- _ It simplifies mass of figures (condensation)
- _ It facilitates comparison
- _ It helps in formulating and testing hypothesis
- _ It helps in prediction
- _ It helps in the formulation of suitable policies

b) Scope of Statistics

Statistics pervades all subject matter - its use has permeated almost every facet of our lives. It is a tool of all sciences indispensable to search and intelligent judgment and has become a recognized discipline in its own right. [Statistics (singular) is a subject of study; statistics (plural) are numerical facts.] Statistics

and the State: State collects statistics on several problems, which help in framing policies. State is the biggest collector and user of statistical data. [The word Statistics comes from Italian word 'STATISTA' meaning statesman, and/or the German word 'STATISTIK' which means political state.] Statistics and Business: Valuable tool for decision making in all areas of business: Production, sale, purchase, finance, personnel, Accounting, market & product research, and Quality control. Statistics and Economics: Economic policies would leap in the dark in the absence of statistical information. Statistics and Research: Indispensable in research work.

1.7 Statistical applications

As mentioned above, statistics can be applied in all subject of matter. In what follows, we present solely the statistical applications in Business and Economics.

❖ Accounting

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. In that case, the audit staff can select a subset of accounts called a sample to assess the claim.

❖ Finance

Financial analysts use a variety of statistical information, including price-earnings ratios and dividend yields, to guide their investment recommendations. For example, data suppliers can process data, and then sell statistical summaries of data to manufacturers.

❖ Marketing

Electronic point-of-sale scanners at retail checkout counters are being used to collect data for a variety of marketing research applications. For example, the company can assess the impact of quality, price, publicity, promotion, or marketable place on its sales.

❖ Production

A variety of statistical quality control charts are used to monitor the output of a production process.

❖ Economics

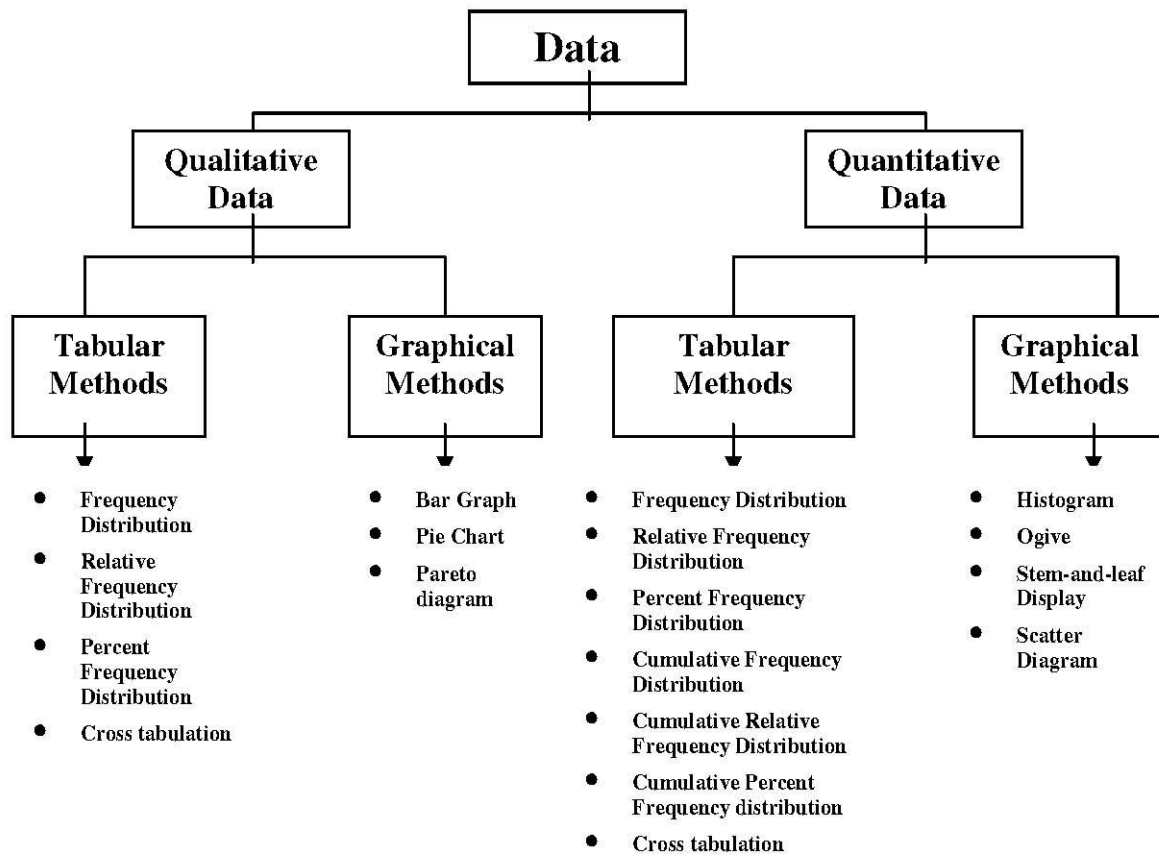
Economists use statistical information in making forecasts about the future of the economy or some aspect of it. For instance, in forecasting inflation rates, economists use statistical information on such indicators as Producer Price Index (PPI), the unemployment rate, and the manufacturing capacity utilization.

Chapter 2: Tabular and Graphical Presentations of data

2.1 Introduction

This chapter introduces tabular and graphical descriptive methods used to summarize and describe sets of data. These methods depend on which type is for data under consideration (quantitative, qualitative) and how many variables are we interested in (univariate variable, i.e single variable or bivariate, i.e relationship between two variables).

The figure below shows possibilities of tabular and graphical presentations of data depending on the data type.



Classification is the first step in Tabulation – items having common characteristics

must be brought together before the data can be displayed in tabular form. Classification is the grouping of related facts into classes. This will be developed when will come to continuous data.

Objectives of classification

- (i) To condense the mass data in such a manner that similarities and dissimilarities can be readily apprehended.
- (ii) To facilitate comparison
- (iii) To pinpoint the most significant features of the data at a glance.
- (iv) To give prominence to the important information gathered.
- (v) To enable statistical treatment of the material collected.

General Rules for constructing diagrams

- (i) Title – every diagram must be given a suitable title
- (ii) Proportion between width and height of the diagram should be maintained
- (iii) Selection of Scale – the scale should also specify the size of the unit and what it represents. E.g. million tons; number of persons (000).
- (iv) Footnotes – may be given
- (v) Index – illustrating different types of lines or different shades, colours should be given
- (vi) Simplicity, neat and clean.

2.2 Summarizing Qualitative data and quantitative-discrete data (one variable)

- **Frequency distribution**

A frequency distribution is a tabular summary of data showing the number (frequency) of items in each of several no overlapping classes.

Example: a sample of 50 NUR students has been selected to know their marital statuses. The reported answers are shown in the table below. Construct a frequency table to summarize the information given.

Marital statuses for NUR 50 students

Divorced	Divorced	Divorced	Divorced	Divorced
Married	Married	Married	Married	Married
Separated	Separated	Separated	Separated	Separated
Single	Single	Single	Single	Single
Single	Single	Married	Single	Separated
Single	Separated	Married	Married	Divorced
Divorced	Single	Single	Married	Divorced
Separated	Single	Single	Single	Married
Separated	Single	Single	Single	Single
Single	Single	Separated	Married	Single

To develop a frequency table for these data, we count the number of times each category of marital status appears in the table above.

Frequency distribution of students' marital status

Marital Status	Frequency
Divorced	8
Married	11
Separated	10
Single	21
Total	50

- **Relative frequency and percent frequency distribution**

A frequency distribution shows the number (frequency) of items in each of several no overlapping classes. However, we often interested on proportion, or percentage, of items in each class. A *relative frequency* of the class equals the fraction or proportion of items belonging to class. For a data set with n observations, the relative frequency of each class can be determined as follows:

$$\text{Relative frequency of a class} = \frac{\text{frequency of the class}}{n}$$

The *percent frequency* of the class is the relative frequency multiplied by 100.

Relative and percent frequency distributions of students' marital statuses

Marital Status	Frequency	Relative frequency	Percent frequency
Divorced	8	0.16	16
Married	11	0.22	22
Separated	10	0.2	20
Single	21	0.42	42
Total	50	1	100

- **Bar graph**

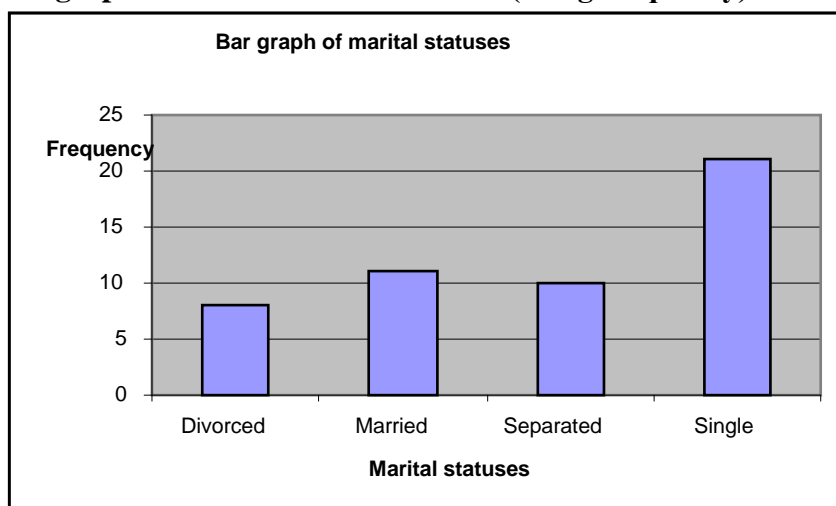
A *bar graph*, or a bar chart, is a graphical device for depicting qualitative data summarized in a frequency, relative frequency, or a percent frequency distribution. On the horizontal axis we specify the labels that are used for the classes (categories) of data. A frequency, relative frequency, or a percent frequency scale can be used for the vertical axis.

Types of Bar Diagrams

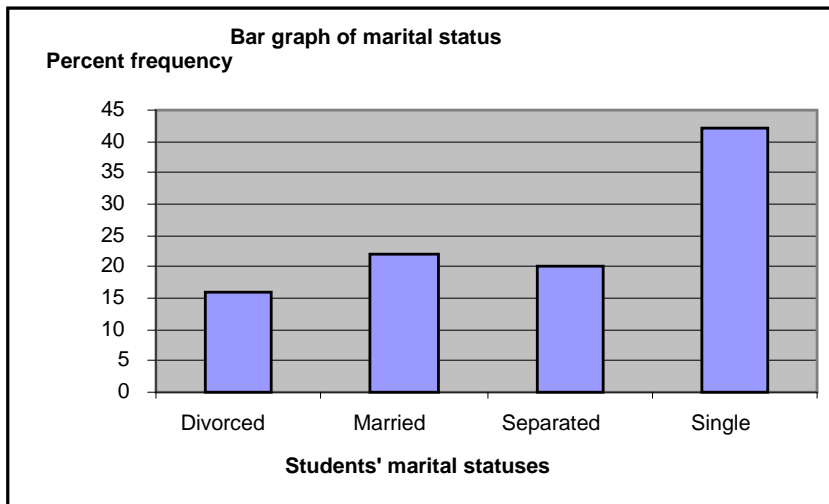
a) Simple Bar diagram:

- (i) The width of the bars should be uniform throughout the diagram.
- (ii) The gap between one bar and another should be uniform through out
- (iii) Bars may be either horizontal or vertical.
- (iv) It is desirable to write the respective figures at the end of each bar.

Bar graph of students' marital status (using frequency)



Bar graph of students' marital status (using percent Frequency)

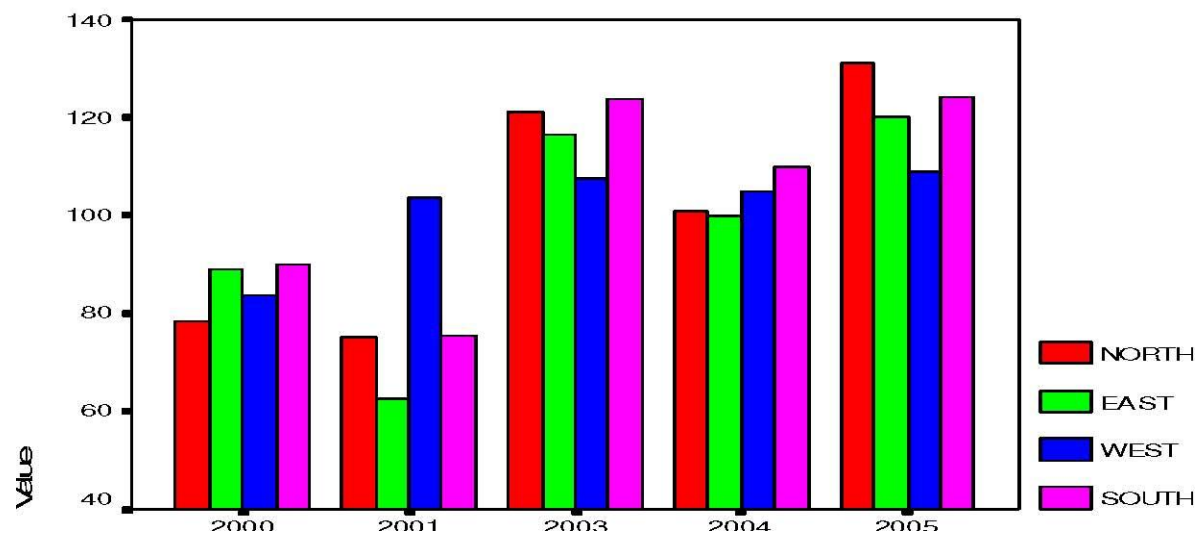


b) Multiple bars (Clustered bar Diagram):

In a multiple bar diagram two or more set of interrelated data are represented. Illustration: The regional coffee production in Rwanda is given below: Represent it in a Multiple Bar diagram

Year	North	East	West	South
2000	78.40	88.90	83.7	89.90
2001	75.10	62.50	103.60	75.50
2003	121.20	116.50	107.70	123.90
2004	101.00	100.00	105.00	110.00
2005	131.00	120.00	109.00	124.00

Multiple bar for coffee production in Rwanda

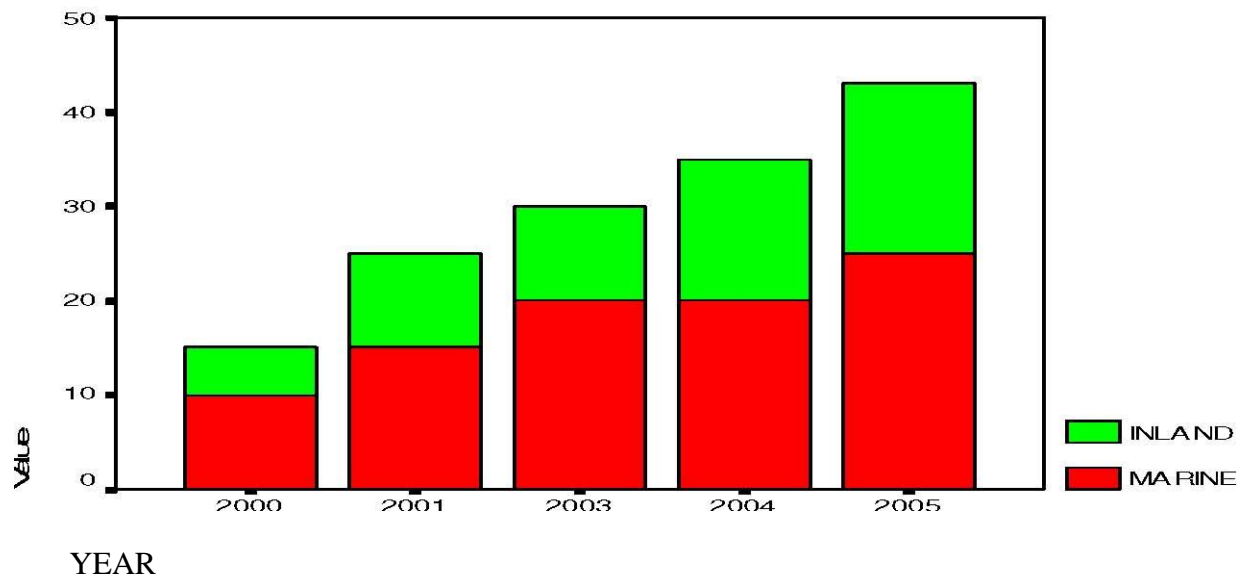


C) Sub-divided bar diagram (component bar chart, Stack bar)

In a subdivided bar diagram each bar representing the magnitude of a given phenomenon is further

subdivided in its various components. Each component occupies a part of the bar proportional to its share in total. Example: Growth of Fish production in Africa (in billion tons).

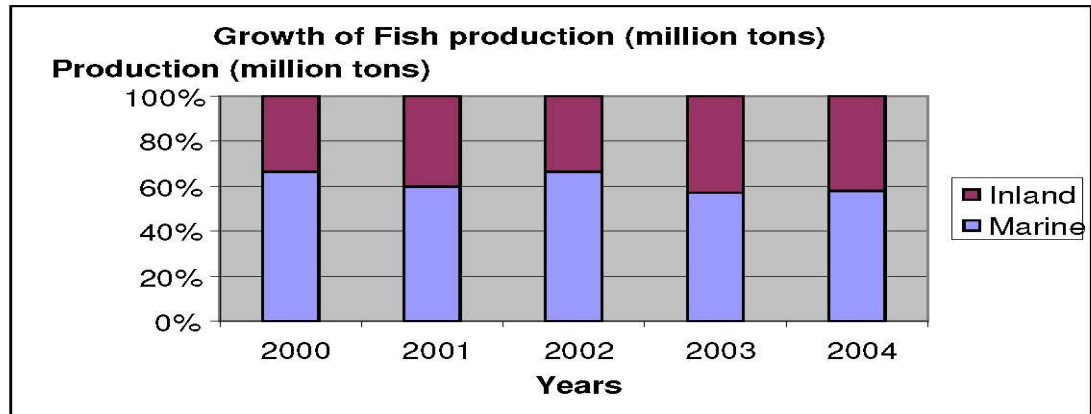
Year	Marine	Inland
2000	10.00	5.00
2001	15.00	10.00
2003	20.00	10.00
2004	20.00	15.00
2005	25.00	18.00



d) Percentage Bars:

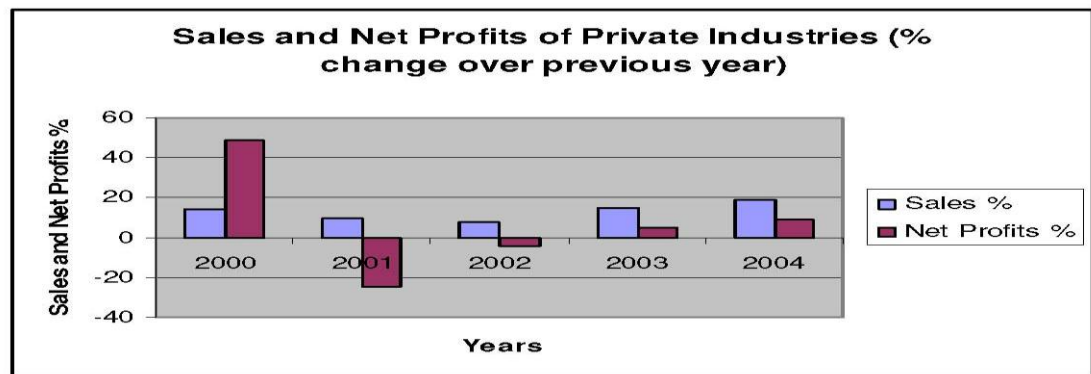
The length of the bars is kept equal to 100 and segments are cut in these bars to represent the components (%) of an aggregate. Represent the above fish production by subdivided bar diagram drawn on the % basis.

Year	Marine	Inland
2000	66,67	33,33
2001	60,00	40,00
2002	66,67	33,33
2003	57,14	42,86
2004	58,14	41,86



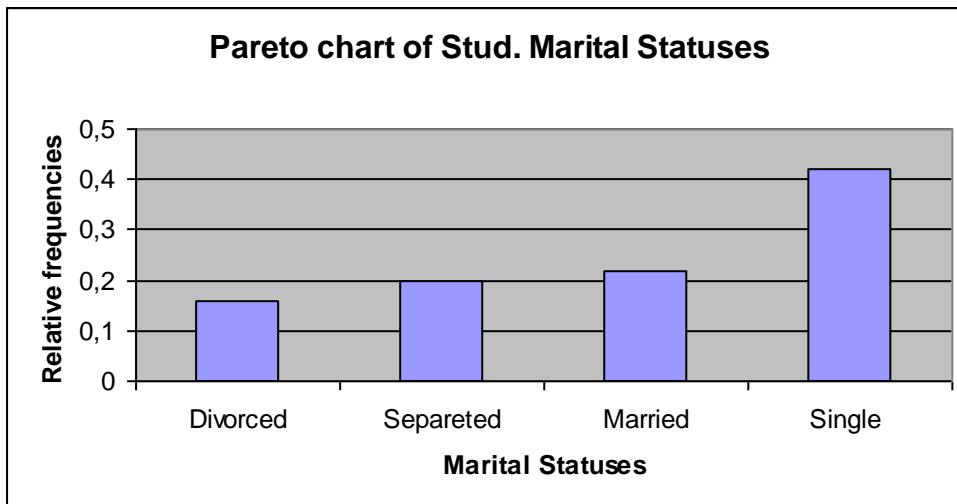
d) Deviation Bars:

Such bars can have both positive and negative values. Positive values are shown above the base line and negative values below it.



- **Pareto diagram**

A *pareto diagram* has the same rules as a simple bar chart but for a pareto chart we solely use relative frequencies of items and by respecting the order (ascending or descending) with respect to relative frequencies.

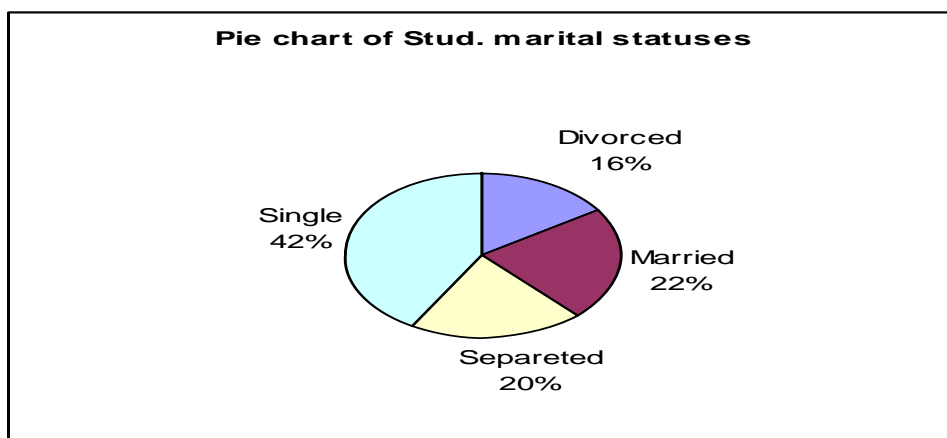


- **Pie Diagram or Pie chart**

A pie chart is a useful method for displaying the percentage of observations that fall into each category of qualitative data, while a bar chart can be used to display the frequency of observations that fall into each category.

To construct a pie chart, we first draw a circle to represent all of data. Then we use the relative frequencies to subdivide the circle into sectors, or parts, that correspond to the relative frequency for each class (because a circle contains 360 degrees, we take a relative frequency for each class times 360 to obtain the angle of the sector from the centre of the circle).

Marital Status	Relative frequency	Degrees
Divorced	0,16	57,6
Married	0,22	79,2
Separated	0,2	72
Single	0,42	151,2
Total	1	360



- **A line chart for time series**

If the categories consist of points in time (time series data) and the objective is to focus on the trend in frequencies over time, a line chart is useful.

Rules for constructing line graphs of time series:

- (i) Take the 'Time' on the X axis and variable on the Y-axis.
- (ii) Begin Y-axis with zero and select a suitable scale
- (iii) Corresponding to the time factor plot the value of the variable and join the various points by straight lines
- (iv) If on one graph more than one variable is shown, they should be distinguished by the use of thick, thin, dotted lines etc.
- (v) False base line: One of the fundamental rules while constructing graphs is that the scale on the Y-axis should begin from zero. Some times when the values to be plotted are relatively high, the Y scale is broken. The X-axis can also be broken in a similar manner.

Example1: According to the *New York Times* (27 September 1987, p. 1F), the June levels of unemployment in the United States for five years were as follows:

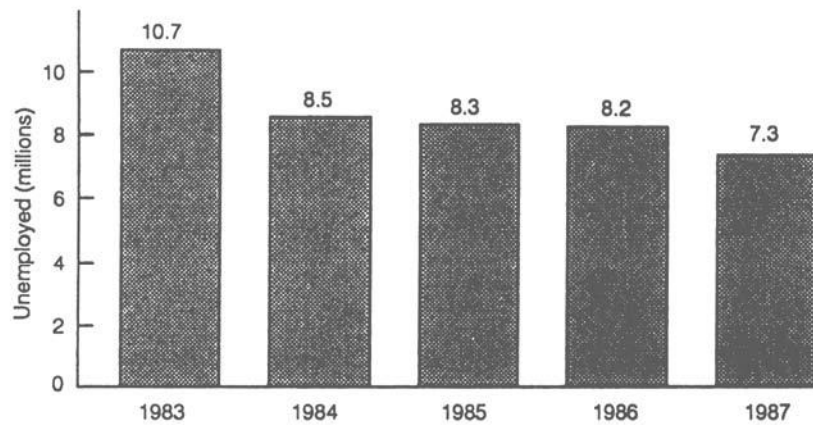
Year	Unemployed (millions)
1983	10.7
1984	8.5
1985	8.3
1986	8.2
1987	7.3

- a) Use a bar chart to depict these data.
- b) Use a line chart to depict these data.

Solution

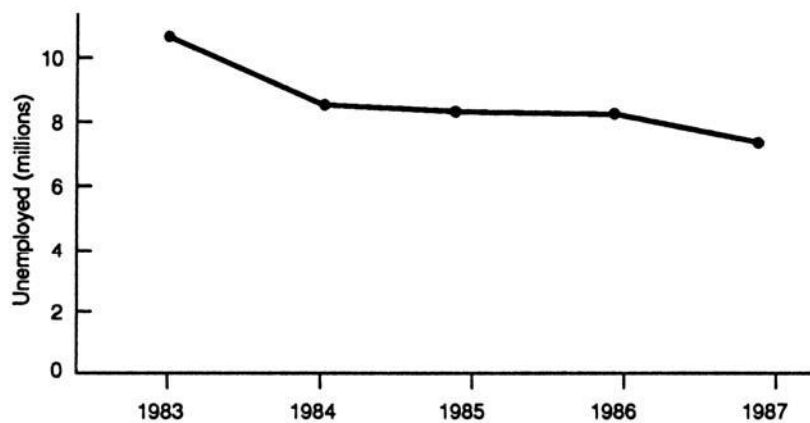
- a) The five years, or categories, are represented by intervals of equal width on the horizontal axis. The height of the vertical bar erected above any year is proportional to the frequency (number of unemployed) corresponding to that year.

A bar chart of unemployment in the United States



- b) A line chart is obtained by plotting the frequency of a category above the point on the horizontal axis representing that category and then joining the points with straight lines.

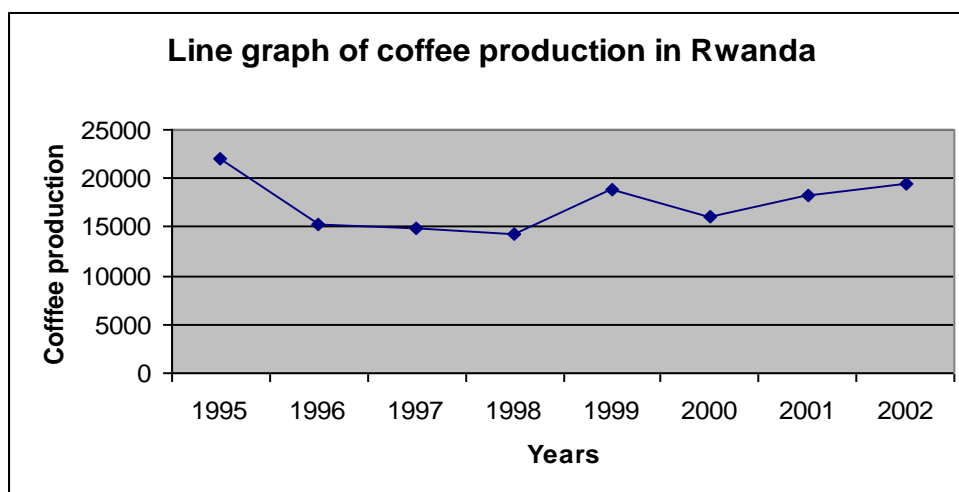
A line chart of unemployment in the United States



Example2: Coffee production in Rwanda between 1995 and 2005. When only one variable is to be represented, on the X-axis measure time and on the Y-axis the value of the variable and plot the various points and join them by straight lines. The fluctuation of this line shows the variations in the variable, and the distance of the plotting from the base line of the graph indicates the magnitude.

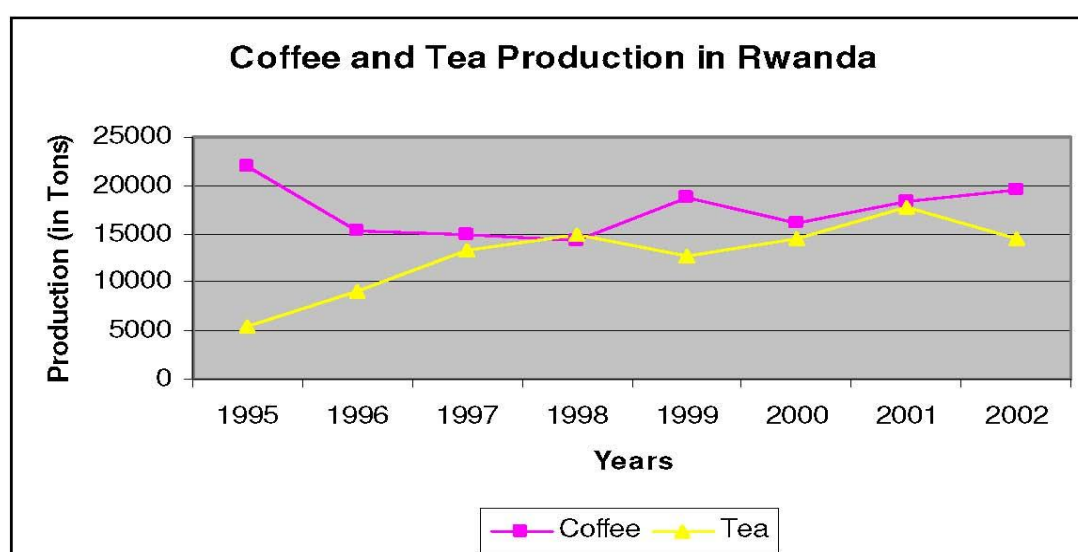
Year	Coffee Production in tones
1995	21952
1996	15285
1997	14830
1998	14268
1999	18817

2000	16098
2001	18268
2002	19467



If the unit of measurement is the same, we can represent two or more variables on the same graph. This facilitates comparison. When two or more variables are shown on the same graph it is desirable to use thick, thin, dotted lines etc. to distinguish between the various variables.

Year	Coffee Production	Tea Production
1995	21952	5414
1996	15285	9058
1997	14830	13239
1998	14268	14875
1999	18817	12669
2000	16098	14481
2001	18268	17686
2002	19467	14543



Exercises

1. A small survey was carried out into the mode of travel to work. The information is presented in the following table:

	1	2	3	4	5	6	7	8	9	10
Person										
Mode of travel	Car	Car	Bus	Car	Walk	Cycle	Car	Bus	Train	Car

	11	12	13	14	15	16	17	18	19	20
Person										
Mode of travel	Cycle	Bus	Train	Car	Bus	Walk	Car	Train	Bus	Car

	21	22	23	24	25	26	27	28	29	30
Person										
Mode of travel	Cycle	Car	Car	Bus	Train	Car	Walk	Cycle	Bus	Car

- a) Classify these data into a frequency distribution.
 - b) Construct a bar chart, pareto diagram, and a pie chart to represent the above data.
2. The *New York Times* article reported, “6 million Americans who say they want work are not even seeking jobs.” A breakdown of these 6 million Americans by race follows:

Race	Frequency
White	4,320,000
Black	1,500,000
Other	180, 000

Use a bar chart, pareto diagram and a pie chart to depict these data.

3. The number of men and women who have received an M.B.A. degree from a particular university in each of five years is shown below.

Year	Men	Women
1988	74	12
1989	85	20
1990	90	32
1991	112	48
1992	128	67

- a) Use a component bar chart to depict these data.
- b) Use a line chart to depict these data and add comments.

2.3 Graphical Techniques for Quantitative (continuous) Data

This section introduced the basic methods of descriptive statistics used for organizing a set of numerical data in tabular form and presenting it graphically. Summarizing data in this way requires that you first group the data into classes. Judgment is required concerning the number and the size of the classes to be used. The important point to bear in mind when making this judgment is that the presentation of the grouped data should enable the user to quickly grasp the general shape of the distribution of the data.

- **Frequency distribution.**

The three steps are necessary to define the classes for a frequency distribution with continuous data:

Step 1. Determine the number of classes

As a general guideline, we recommend using between 5 and 20 classes depending on how large is the number of data items.

The number can be either fixed arbitrarily keeping in view the nature of problem under study or it can be decided with the help of STURGE'S rule. To him, number of classes can be determined by the formula: $K = 1 + (3.322 * \log N)$

Thus if 10 observations are being studied the number of classes shall be:

$$k = 1 + (3.322 * 1) = 4.322 \text{ or } 4$$

And if 100 observations:

$$k = 1 + (3.322 * 2) = 7.6 \text{ or } 8$$

Step 2. Determine the class width of each class

As a general guideline, we recommend that class width be the same for each class.

A simple formula to obtain the estimate of appropriate class interval 'i' is:

$i = L - S / k$ where L = largest item, S = Smallest item; k = the number of classes.

For example, if the salary of 100 employees in a commercial undertaking varies between Rwf 500 and 5500 and one wants to form 10 classes, then the class interval would be: $i = L - S / k = 5500 - 500 / 10 = 500$

STURGE suggested the following formula for determining the magnitude of class interval. $i = L - S / 1 + 3.322 * \log N$

Step 3. Determine the class limits

The class limits are the lowest and the highest values that can be included in the Class. The starting class would be the lower value – the lower value + class width

N.B:

- Class Intervals:** The difference between upper and lower limit of a class is known as class interval of that class e.g. $40 - 20 = 20$
- Class frequency:** The number of observations corresponding to a particular class is known as the frequency of that class.

c) **Class mid-point or class mark:**

Upper limit of the class + Lower limit of the Class/2

- d) When making classes you should use the **Exclusive method** of classification or **Inclusive method** of classification. Better to use the first.

Exclusive Method: When the class intervals are so fixed that the upper limit of one class is the lower limit of the next class, it is known as the exclusive method of classification – it ensures continuity of data. It is always presumed that upper limit is exclusive i.e. the item of that value is not included in that class.

Inclusive method: Under the inclusive method of classification, the upper limit of one class is included in that class itself.

Example:

The weights in pounds of a group of workers are as follows:

173...165...171...175...190...183...177...160...151...169

162...179...140...171...175...168...158...186...182...162

154...180...164...166...157

Construct a frequency and relative frequencies distributions for these data

Solution

The hardest, and most important, step in constructing a frequency distribution is choosing the number and width of the classes.

As we have small size of observations, let us take the number of classes as 5.

Then $i = L - S / k = 190 - 140 / 5 = 10$

A frequency table of the weights in pounds of a group of 25 workers

Class Limits	Frequency
140 up to 150	1
150 up to 160	4
160 up to 170	8
170 up to 180	7
180 up to 190	5
Total	25

The relative frequencies, obtained by dividing each frequency by 25, are shown below:

Class Limits	Frequency	Relative Frequency	Cumulative Relative Frequency
140 up to 150	1	.04	.04
150 up to 160	4	.16	.20
160 up to 170	8	.32	.52
170 up to 180	7	.28	.80

180 up to 190	5	.20	1.00
---------------	---	-----	------

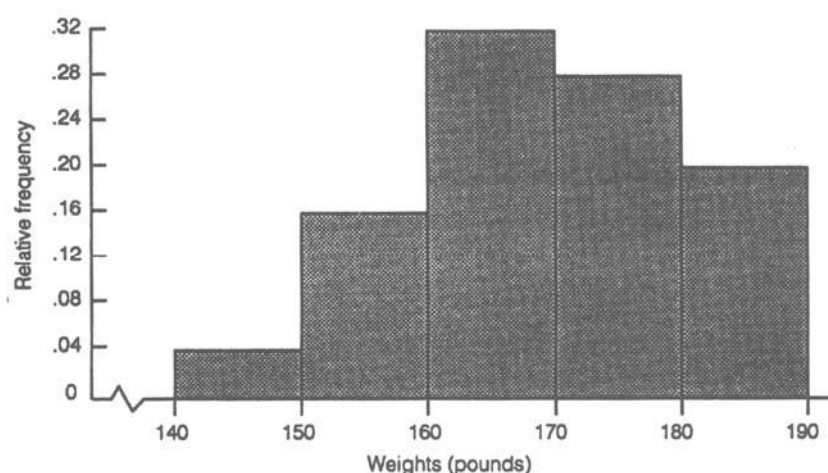
- **Histogram**

A common graphical presentation of quantitative-continuous data is a *histogram*.

A **histogram** is a graphical display of tabulated frequencies or relative frequencies, shown as bars. It shows what proportion of cases fall into each of several categories: it is a form of data binning. The categories are usually specified as non-overlapping intervals of some variable. The categories (bars) must be adjacent. The intervals (or bands, or bins) are generally of the same size.

A histogram is a set of vertical bars whose areas are proportional to the frequencies represented. While constructing the histogram the variable is always taken on the X-axis and the frequencies depending on it on the Y-axis. A distance on the scale that is proportional to its class interval represents each class. In this manner we get a series of rectangle each having a class interval distance as its width and the frequency distance as its height. The area of the histogram represents the total frequency as distributed through out the classes. Bar diagram is one-dimensional i.e. only the length of the bar is material and not the width. Histogram is two-dimensional that is in a histogram both lengths as well as the width is important.

Histogram of the weights in pounds of a group of 25 workers



- **Frequency polygon**

Frequency polygons are a graphical device for understanding the shapes of distributions. They serve the same purpose as histograms, but are especially helpful in comparing sets of data. Frequency polygons are also a good choice for displaying cumulative frequency distributions.

A frequency polygon is a graph of frequency distribution. We may draw a histogram of the given data and then join by straight lines the mid points of the upper horizontal sides of each rectangle with the adjacent ones.

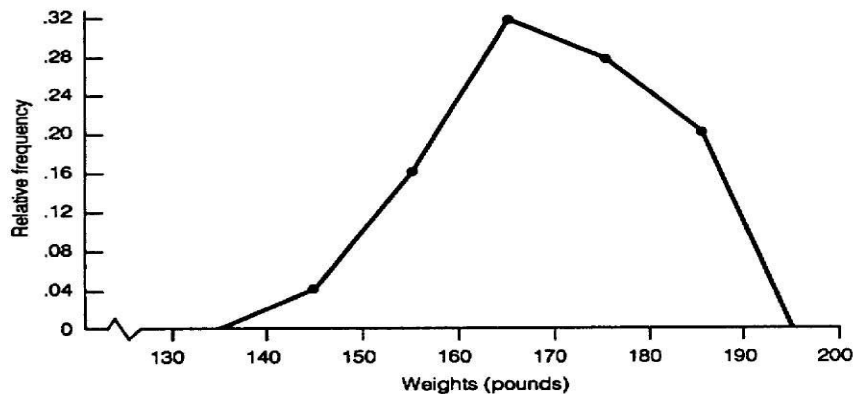
To create a frequency polygon, start just as for histograms, by choosing a class interval. Then draw an X-axis representing the values of the scores in your data. Mark the middle of each class interval with a tick mark, and label it with the middle value represented by the class. Draw the Y-axis to indicate the frequency of each class. Place a point in the middle of each class interval at the height corresponding to its frequency.

Finally, connect the points. You should include one class interval below the lowest value in your data and

one above the highest value. The graph will then touch the X-axis on both sides.

Plotting the relative frequency of each class above the midpoint of that class and then joining the points with straight lines construct the relative frequency polygon. The polygon is closed by considering one additional class (with zero frequency) at each end of the distribution and extending a straight line to the midpoint of each of these classes.

A relative frequencies polygon of the weights in pounds of a group of 25 workers



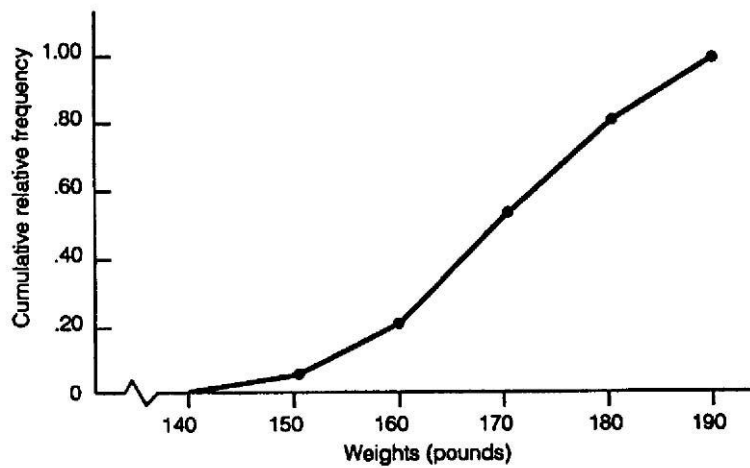
You can easily discern the shape of the distribution from Figure above. Most of the weights are between 155 and 185. It is clear that the distribution is not symmetric in as much as good weights (to the right) trail off more gradually than poor weights (to the left). The distribution is skewed.

- **Cumulative Frequency curve or OGIVE**

The curve obtained by plotting cumulative frequencies is called a Cumulative Frequency curve or an OGIVE.

To construct the ogive, the cumulative relative frequency of each class is plotted above the upper limit of that class, and straight lines then join the points representing the cumulative frequencies. The ogive is closed at the lower end by extending a straight line to the lower limit of the first class.

An Ogive of the weights in pounds of a group of 25 workers



Data in Ordered Array:

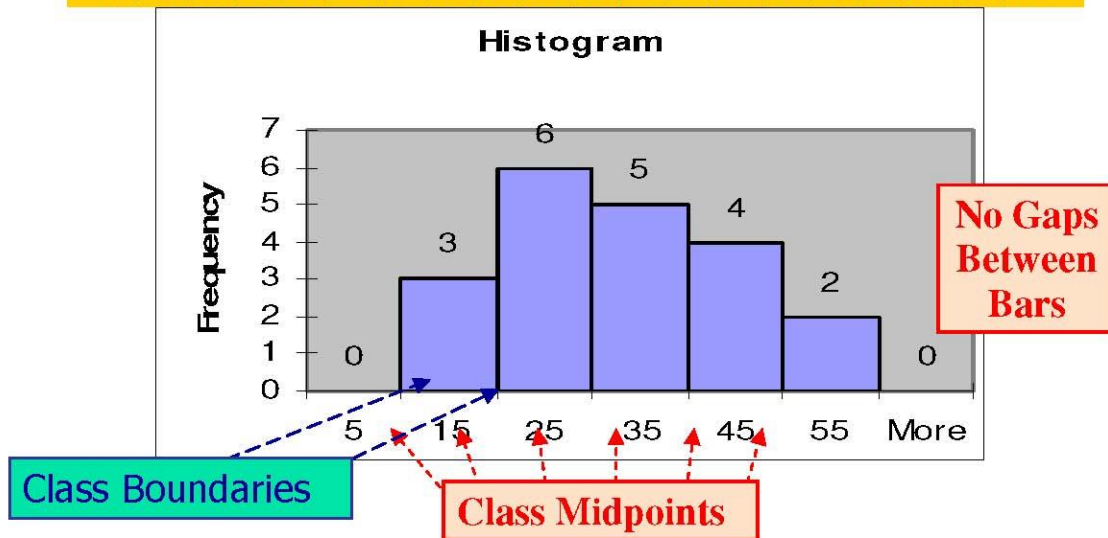
12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

Class	Frequency	Relative Frequency	Percentage
10 but under 20	3	.15	15
20 but under 30	6	.30	30
30 but under 40	5	.25	25
40 but under 50	4	.20	20
50 but under 60	2	.10	10
Total	20	1	100

The Histogram

Data in Ordered Array:

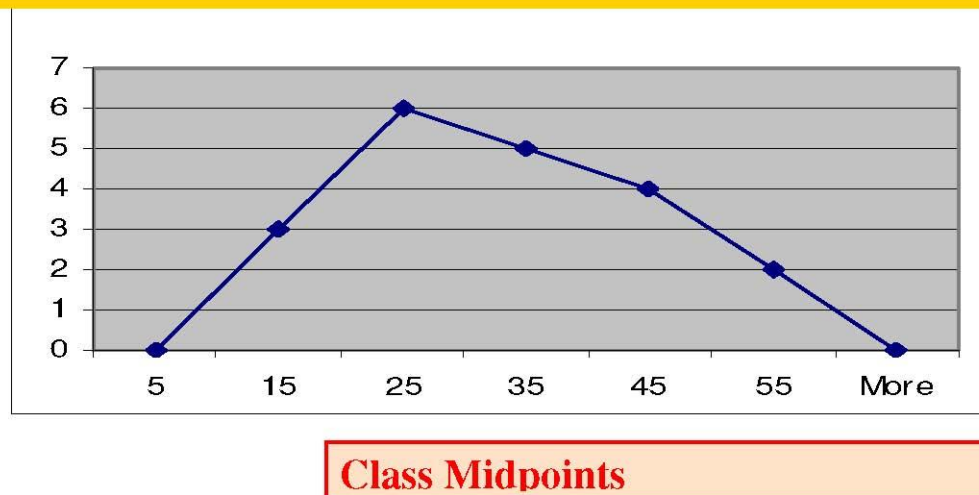
12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58



The Frequency Polygon

Data in Ordered Array:

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58



Cumulative Frequency

Data in Ordered Array:

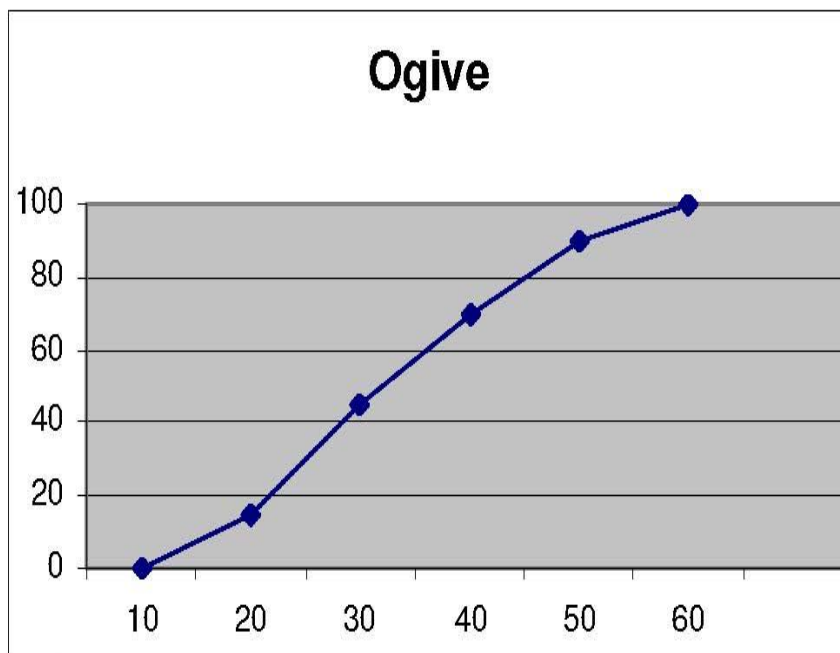
12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

Lower Limit	Cumulative Frequency	Cumulative % Frequency
10	0	0
20	3	15
30	9	45
40	14	70
50	18	90
60	20	100

The Ogive (Cumulative % Polygon)

Data in Ordered Array :

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58



Class Boundaries (Not Midpoints)

EXERCISES

1.a) Prepare a frequency table for the following data (marks in Stat Examination out of 100) with width of each class interval as 10. Use exclusive method of classification.

57 44 80 75 00 18 45 14 04 64 72 51 69 34 22 83 70 20 57 28 96 56 50 47

10 34 61 66 80 46 22 10 84 50 47 73 42 33 48 65 10 34 66 53 75 90 58 46 38 69.

b) Represent the above data using a Histogram.

2. The salaries (in hundreds of dollars) of a sample of 40 government employees are as follows:

208 160 175 334 228 211 179 354 265 215 191 239 298 226 220

260 173 263 226 165 252 422 284 232 225 348 290 180 300 200

245 204 256 281 230 275 158 224 315 217

a) Construct a frequency distribution for these data.

b) Construct a relative frequency histogram for the data.

c) Construct a relative frequency polygon for the data.

d) Construct an ogive for the data.

2.4 Graphical Techniques for bivariate data

There are many situations we wish to depict the relationship between two variables; in such cases bivariate methods are required.

- **A cross-tabulation or bivariate (or Two-Way) Frequency Distribution**

A cross-tabulation is a tabular summary of data of two variables.

If the data corresponding to one variable say 'X', is grouped into 'm' classes and the data corresponding to other variable say 'Y' is grouped into 'n' classes then the bivariate table will consist of $m \times n$ cells.

Illustration: The data given below relate to the height and weight of 20 persons. You are required to form a two-way frequency table with class interval 62" to 64", 64" to 66" and so on; and 115 to 125 lb, 125 to 135 lb etc.

Sl No.	Weight	Height	Sl No.	Weight	Height
1	170	70	11	163	70
2	135	65	12	139	67
3	136	65	13	122	63
4	137	64	14	134	68
5	148	69	15	140	67
6	121	63	16	132	69
7	117	65	17	120	65
8	128	70	18	148	68
9	143	71	19	129	67
10	129	62	20	152	67

A two-way table related to the height and weight of 20 persons

Weight (Y) Height (X)	115 - 125	125 - 135	135 - 145	145 - 155	155 - 165	165 - 175	Total
62 – 64	(2)	(1)					3
64 – 66	(2)		(3)				5
66 – 68		(1)	(2)	(1)			4
68 – 70		(2)		(2)			4
70 – 72		(1)	(1)		(1)	(1)	4
Total	4	5	6	3	1	1	20

- **Scatter Diagrams**

Statistics practitioners frequently need to know the relationship between two *quantitative* variables. Economists, for example, are interested in the relationship between inflation rates and unemployment rates. Business owners are interested in many variables, including the relationship between their advertising expenditures and sales levels. The graphical technique used to depict the relationship between the variables X and Y is the **scatter diagram**, which is a plot of all pairs of values (x, y) for the variables X and Y.

Example:

An educational economist wants to establish the relationship between an individual's income and education. She takes a random sample of 10 individuals and asks for their income (in \$1,000s) and education (in years). The results are shown below. Construct a scatter diagram for these data, and describe the relationship between the number of years of education and income level.

x (education)	y (income)
11	25
12	33
11	22
15	41
8	18
10	28
11	32
12	24
17	53
11	26

Solution

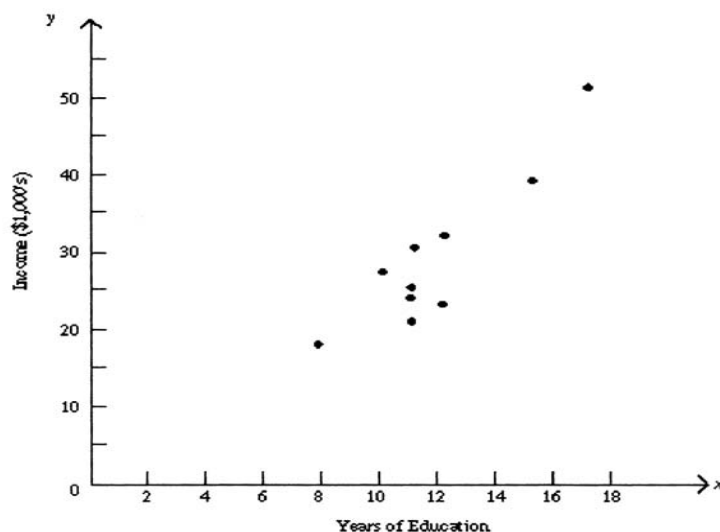
If we feel that the value of one variable (such as income) depends to some degree on the value of the other variable (such as years of education), the first variable (income) is called the **dependent variable** and is plotted on the vertical axis. The ten pairs of values for education (x) and income (y) are plotted in Figure 2.1, forming a scatter diagram.

The scatter diagram allows us to observe two characteristics about the relationship between education (x) and income (y):

1. Because these two variables move together—that is, their values tend to increase together and decrease together—there is a **positive relationship** between the two variables.
2. The relationship between income and years of education appears to be **linear**, since we can imagine drawing a straight line (as opposed to a curved line) through the scatter diagram that approximates the positive relationship between the two variables.

The pattern of a scatter diagram provides us with information about the relationship between two variables. Figure 2.1 depicts a positive linear relationship. If two variables move in opposite directions, and the scatter diagram consists of points that appear to cluster around a straight line, then the variables have a **negative linear relationship** (see Figure 2.2). It is possible to have **nonlinear relationships** (see Figures 2.3 and 2.4), as well as situations in which the two variables are **unrelated** (see Figure 2.5). In topic 4, we will compute numerical measures of the **strength** of the linear relationship between two variables.

Figure 2.1
Scatter Diagram for Example 2.4



Some figures of possible relationship:

Figure 2.2

Negative Linear Relationship

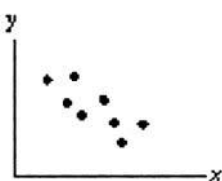


Figure 2.3

Nonlinear Relationship



Figure 2.4

Nonlinear Relationship

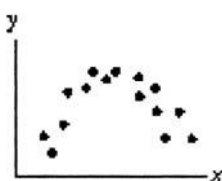
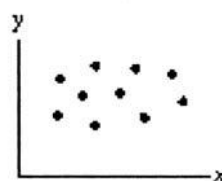


Figure 2.5

No Relationship



EXERCISES

1. The manager of a large furniture store wanted to determine the effectiveness of her advertising. The furniture store regularly runs several ads per month in the local newspaper. The manager wanted to know if the number of ads influenced the number of customers. During the past eight months, she kept track of both figures, which are shown below. Construct a scatter diagram for these data, and describe the relationship between the number of ads and the number of customers.

Month	Number of Ads, x	Number of customers, y
1	5	528
2	12	876
3	8	653
4	6	571
5	4	556
6	15	1,058
7	10	963
8	7	719

2. Present the following information in a tabular form. In 1975 out of a total of 1750 workers of a factory, 1200 were members of a trade union. The number of women employees was 200 of which 175 did not belong to a trade union. In 1980, the number of union workers increased to 1580 of which 1290 were men. On the other hand the number of non – union workers fell down to 208 of which 180 were men. In 1985 there were 1800 employees who belonged to a trade union and 50 who did not belong to a trade union. Of all the employees in 1985, 300 were women of whom only 8 did not belong to a trade union.

Chapter 3: Univariate Descriptive Measures

a. 3.1 Measures of Central Location

This section discussed three commonly used numerical measures of the central, or average, value of a data set: the mean, the median, and the mode. You are expected to know how to compute each of these measures for a given data set. Moreover, you are expected to know the advantages and disadvantages of each of these measures, as well as the type of data for which each is an appropriate measure.

Question: How do I determine which measure of central location should be used—the mean, the median, or the mode?

Answer: If the data are qualitative, the only appropriate measure of central location is the mode. If the data are ranked, the most appropriate measure of central location is the median.

For quantitative data, however, it is possible to compute all three measures. Which measure you should

use depends on your objective. The mean is most popular because it is easy to compute and to interpret. (In particular, the mean is generally the best measure of central location for purposes of statistical inference. It has the disadvantage, however, of being unduly influenced by a few very small or very large measurements.

To avoid this influence, you might choose to use the median. This could well be the case if the data consisted, for example, of salaries or of house prices. The mode, representing the value occurring most frequently (or the midpoint of the class with the largest frequency) should be used when the objective is to indicate the value (such as shirt size or house price) that is most popular with consumers.

3.1.1 Measures of central location: Case of discrete series

1. Mean

- **Arithmetic mean**

The **arithmetic mean** (or simply the **mean**) of a list of numbers is the sum of the entire list divided by the number of items in the list. If the list is a statistical population, then the mean of that population is called a **population mean**. If the list is a statistical sample, we call the resulting statistic a **sample mean**.

If we denote a set of data by $X = (x_1, x_2, \dots, x_n)$, then the sample mean is typically denoted with a horizontal bar over the variable (\bar{x} , enunciated "x bar"). The [Greek letter \$\mu\$](#) is used to denote the arithmetic mean of an entire population.

The **sample mean** is calculated as follows:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} (x_1 + \dots + x_n).$$

For example, the arithmetic mean of six values (let's say marks of six students during an exam out of 60): 34, 27, 45, 55, 22, 34 is:

$$\frac{34 + 27 + 45 + 55 + 22 + 34}{6} = \frac{217}{6} \approx 36.167.$$

The **population mean** is calculated as:

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i = \frac{1}{N} (X_1 + X_2 + \dots + X_N)$$

Mathematical properties of Arithmetic Mean:

- The sum of the deviations of the items from the arithmetic mean is always zero.
i.e. $\sum (x_i - \bar{x}) = 0$
- The sum of the squared deviations of the items from arithmetic mean is the minimum, that is, less than the sum of the squared deviations of the items from any other values.

- iii) Since $\bar{x} = (\sum x_i)/n$; $n\bar{x} = \sum x_i$
 iv) If we have the arithmetic mean of the number of items of two or more than two related groups, we can compute combined average of these groups by applying the following formula:

$$\bar{x}_{1,2} = n_1 \bar{x}_1 + n_2 \bar{x}_2 / n_1 + n_2$$

Where $\bar{x}_{1,2}$ = Combined mean of the two groups.

\bar{x}_1 = Arithmetic mean of first group

\bar{x}_2 = Arithmetic mean of second group

n_1 = Number of items in the first group

n_2 = Number of items in the second group

N.B: When the values are presented in a frequency table,

- **Sample mean** $\bar{x} = (\sum f_i x_i) / n$

Where f_i = the frequency; x_i = the values; n = the number of observations, i.e. $\sum f_i$.

- **Population mean** $\mu = \frac{1}{N} \sum f_i X_i$

Exercise: Calculate the Arithmetic mean of the following data.

Marks X	20	30	40	50	60	70
No of Students; f	8	12	20	10	6	4

2. Median

The **median** is another measure of central location for a variable. The median is the value in the middle when the data are arranged in ascending order (smallest value to highest value). With an odd number of observations, the median is the middle value. An even number of observation has no single middle value. In this case, we follow the convention and define the median as the average of the values for the middle two observations. The sample and population median are computed in the same way.

- **Calculation of Median: Individual Series.**

Arrange the data in ascending or descending order of magnitude.

Median = $(n+1) / 2^{\text{th}}$ item.

Example: Find the mean and median of the following sample of measurements:

8, 12, 6, 6, 10, 8, 4, 6

Solution

$$\text{The mean value is} = \frac{8+12+6+6+10+8+4+6}{8} = \frac{60}{8} = 7.5$$

To find the median, we must first arrange the measurements in ascending order:

4, 6, 6, 6, 8, 8, 10, 12

Since the number of measurements is even, the median is the midpoint between the two middle values, 6 and 8. Thus, the median is 7.

- **Calculation of Median: frequency Series.**

- Arrange the data in ascending or descending order
- Find out the cumulative frequencies
- Apply the formula: Median = Size of $n+1 / 2$
- Now look at the cumulative frequency column and find that total Which is either equal to $n+1/2$ or next higher to that and determine the value of the variable corresponding to it? That gives the value of the median.

Exercise: Calculate the Median from following data.

Income	800	1000	1500	1800	2000	2500
No of Persons	16	24	26	30	20	6

3. Mode:

This is the third measure of location. The **mode** is defined as the value that occurs with greatest frequency. Both sample and population mode are computed in the same way.

Example: for the example above of the following sample of measurements:

8, 12, 6, 6, 10, 8, 4, 6

The mode is 6, because that is the value that occurs most frequently.

Consider the following sample of measurements, which is obtained from the sample in Example above by adding one extreme value, 21:

8, 12, 6, 6, 10, 8, 4, 6, 21

Which measure of central location is most affected by the addition of the single value?

Solution

The mean value is now 9; and the mode is still 6. We arrange the new sample of measurements in ascending order: 4, 6, 6, 6, 8, 8, 10, 12, and 21. The median is now equal to 8, the middle value. Thus, the mean is the measure that is most affected by the addition of the extreme value.

N.B: The mode is not affected by extreme values, there may not be a mode, there may be several modes, and it is used for either numerical or categorical data.

Exercise: Consider the following sample of measurements:

27, 32, 30, 28, 30, 32, 35, 28, 32, 29

Compute each of the following: a) the mean; b) the median; c) the mode

3.1.2 Measures of central location: Case of continuous series

1. Calculation of Mean

When data are grouped into classes, we lose the individual values and we use the mid points of the class. The formula for each mean becomes the following:

- **Population arithmetic mean** $\mu = \frac{1}{N} \sum_{i=1}^c f_i m_i$
- **Sample arithmetic mean** $= \frac{1}{n} \sum_{i=1}^c f_i m_i$

Where n = sample size, N = size of the population, c = number of classes in the frequency distribution, m_i = midpoint of i class.

- **Geometric mean** $= \sqrt[n]{m_1^{f_1} * m_2^{f_2} * ... * m_n^{f_n}}$

Example: Calculate the arithmetic mean, geometric mean for the following frequency distribution

Marks	5 – 10	10– 15	15-20	20-25	25-30	30-35	35-40	40-45	45-50
No of Students	7	15	24	31	42	30	26	15	10

Solution

Marks	f_i	m_i	$f_i m_i$	$m_i^{f_i}$	$m_1^{f_1} * ... * m_9^{f_9}$
5–10	7	7.5	52.5	1334838.867	1334838.867
10–15	15	12.5	187.5	2.84217E+16	3.79384E+22
15-20	24	17.5	420	6.80633E+29	2.58221E+52
20-25	31	22.5	697.5	8.27291E+41	2.13624E+94
25-30	42	27.5	1155	2.83122E+60	6.0482E+154
30-35	30	32.5	975	2.27248E+45	1.3744E+200
35-40	26	37.5	975	8.41033E+40	1.1559E+241
40-45	15	42.5	637.5	2.66584E+24	3.0816E+265
45-50	10	47.5	475	5.84704E+16	1.8018E+282
Total	200		5575		1.8018E+282

$$\text{Sample arithmetic mean} = \frac{1}{n} \sum_{i=1}^c f_i m_i = \frac{1}{200} * 5575 = 27.88$$

$$\text{Geometric mean} = \sqrt[n]{m_1^{f_1} * m_2^{f_2} * ... * m_n^{f_n}} = \sqrt[200]{1.8018E + 282} = 25.78$$

2. Calculation of Median:

After ascertaining the class in which median lies (use $N+1 / 2$ and see cumulative frequency); the formula

$$\text{is: Median} = L + \left(\frac{(n/2 + 1) - cf}{f} \right) * i$$

Where: L = Lower limit of the median class (the class in which the middle item of the distribution lies).
 $c.f$ = cumulative frequency of the class preceding the median class or sum of the frequencies of all classes lower than the median class. f = simple frequency of the median class. And i = the class interval of the median class.

Calculation of the median for the above example frequency distribution is ;

Marks	5–10	10–15	15-20	20-25	25-30	30-35	35-40	40-45	45-50
No of Students	7	15	24	31	42	30	26	15	10
Cumulative frequency	7	22	46	77	119	149	175	190	200

Median:

Size of $N+1/2$ th item = $201/2 = 100.5^{\text{th}}$ item. Median lies in the class 25 – 30

$L = 25$; $N+1/2 = 100.5$, $cf = 77$; $f = 42$; $i = 5$

$$\text{Median} = 25 + \left(\frac{100.5 - 77}{42} \right) * 5 = 27.80$$

3. Calculation of Mode: Continuous Series.

Steps:

- By inspection ascertain the modal class.
- Determine the value of mode by applying the following formula:

$$\text{Mode} = L + \left(\frac{D_1}{D_1 + D_2} \right) * i$$

Where: L = Lower limit of the modal class D_1 = the difference between the frequency of the modal class and the frequency of the pre-modal class. D_2 = the difference between the frequency of the modal class and the frequency of the post-modal class. And i = the class interval of the modal class.

For the example above, Modal class is 25 – 30, $D_1 = 42-31=11$, $D_2=42-30=12$, $i=5$

$$\text{The mode} = 25 + \left(\frac{11}{11+12} \right) * 5 = 27.39$$

Exercise: calculate the mode, median, arithmetic mean, geometric mean for the following distributions related with marks obtained by students during the statistics exam.

1.

Marks	010	1020	2030	3040	4050	5060	6070	7080	8090	90100
Students	3	5	7	10	12	15	12	6	2	8

2.

Marks	0-10	10-20	20-30	30-40	40-50	50-60
No of Students	10	20	50	60	40	20

3.2 Measures of Variability or dispersion

Dispersion or spread is the degree of the scatter or variation of the variables about the central value. Measures of variation are used:

- (i) to determine the reliability of an average. If widely dispersed, the central values are less representative.
- (ii) To serve as a basis for the control of variability.
- (iii) To compare two or more series with regard to their variability.
- (iv) To facilitate the use of other statistical measure

- **Range**

This is the simplest measure of variability.

Range = Largest observation- Smallest observation.

Because the range is calculated from only two observations, it tells us nothing about the other observations. Consider the following two sets of data:

Set1: 4 4 4 4 4 50

Set2: 4 8 15 24 39 50

The range of both sets is 46. The two sets are completely different and yet ranges are the same!

- **Inter-quartile range**

This measure of variability is simply the difference between the third quartile ,Q3,and the first quartile,Q1.

IQR=Q3-Q1

- **Quartile deviation:** is the average of IQR. That is $QD = \frac{Q3 - Q1}{2}$
- **Coefficient of quartile deviation :** $CQD = \frac{Q3 - Q1}{Q3 + Q1}$

Example: Find out the value of QD and its coefficient from the following data:

Roll No	1	2	3	4	5	6	7
Marks	20	28	40	12	30	15	50

Solution

Arrange data: 12 15 20 28 30 40 50

$$Q1 = (N+1) * \frac{25}{100} = (7+1) * 0.25 = 2^{\text{nd}} \text{ item} = 15, Q2 = (N+1) * \frac{50}{100} = (7+1) * 0.5 = 4^{\text{th}} \text{ item} = 28$$

$$Q3 = (N+1) * \frac{75}{100} = (7+1) * 0.75 = 6^{\text{th}} \text{ item} = 40, \text{IQR} = Q3 - Q1 = 40 - 15 = 25$$

$$QD = \frac{Q3 - Q1}{2} = \frac{40 - 15}{2} = 12.5, CQD = \frac{Q3 - Q1}{Q3 + Q1} = \frac{40 - 15}{40 + 15} = .455$$

Summary:

Interquartile Range

- Measure of Variation
- Also Known as Midspread
 - Spread in the middle 50%
- Difference between the First and Third Quartiles

Data in Ordered Array: 11 12 13 16 16 17 17 18 21

Interquartile Range = $Q_3 - Q_1 = 17.5 - 12.5 = 5$

- Not Affected by Extreme Values

© 2003 Prentice-Hall, Inc. Chap 3-16

Quartiles

- Split Ordered Data into 4 Quarters

- Position of i-th Quartile $(Q_i) = \frac{i(n+1)}{4}$

Data in Ordered Array: 11 12 13 16 16 17 18 21 22

Position of $Q_1 = \frac{1(9+1)}{4} = 2.5$ $Q_1 = \frac{(12+13)}{2} = 12.5$

- Q_1 and Q_3 are Measures of Noncentral Location
- Q_2 = Median, a Measure of Central Tendency

© 2003 Prentice-Hall, Inc. Chap 3-13

Exercise: Calculate the Range and the Inter-quartile range (Petrol consumption of cars)

40.6	34.6	38.8	39.7	38.3	39.2	38.3	36.4	35.3	37.7
38.5	37	36	29.8	32.6	35.3	34.7	30.2	35.9	


• Variance and Standard deviation

The variance and the closely related standard deviation are measures of how spread out a distribution is. In other words, they are measures of variability.

Numerical Measures of Variability		
Samples and Populations - Notation		
	Sample	Population
Variance	s^2	σ^2
Standard Deviation	s	σ

The **variance** is computed as *the average squared deviation* of each number from its mean.
 The **standard deviation formula** is very simple: it is *the square root* of the variance. It is the most

commonly used measure of spread. A low standard deviation indicates that the data points tend to be very close to the same value (the mean), while high standard deviation indicates that the data are spread out over a large range of values.




Variance

- Important Measure of Variation
- Shows Variation about the Mean
 - Sample Variance:

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$$
 - Population Variance:

$$\sigma^2 = \frac{\sum_{i=1}^N (X_i - \mu)^2}{N}$$

© 2003 Prentice-Hall, Inc.
Chap 3-17



Standard Deviation

- Most Important Measure of Variation
- Shows Variation about the Mean
- Has the Same Units as the Original Data
 - Sample Standard Deviation:

$$S = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}}$$
 - Population Standard Deviation:

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (X_i - \mu)^2}{N}}$$

© 2003 Prentice-Hall, Inc.
Chap 3-18

Example: Compute the variance, and standard deviation for the following sample of data:
6, 10, 2, 4, 12, 8

Solution

The sample mean is :

$$\begin{aligned}\bar{x} &= \frac{\sum_{i=1}^n x_i}{n} = \frac{\sum_{i=1}^6 x_i}{6} \\ &= \frac{6 + 10 + 2 + 4 + 12 + 8}{6} = \frac{42}{6} = 7\end{aligned}$$

The sample variance is

$$\begin{aligned}s^2 &= \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} = \frac{\sum_{i=1}^6 (x_i - 7)^2}{5} \\ &= \frac{(6 - 7)^2 + (10 - 7)^2 + (2 - 7)^2 + (4 - 7)^2 + (12 - 7)^2 + (8 - 7)^2}{5} \\ &= \frac{(-1)^2 + (3)^2 + (-5)^2 + (-3)^2 + (5)^2 + (1)^2}{5} \\ &= \frac{70}{5} \\ &= 14\end{aligned}$$

In an example such as this one, where we have already computed in order to find $\sum_{i=1}^n x_i$, it is easier to

use the shortcut formula for computing the sample variance. This saves us from having to compute numerous squared deviations. To use the shortcut formula, we need to compute

$$\begin{aligned}\sum_{i=1}^n x_i^2 &= \sum_{i=1}^6 x_i^2 \\ &= 6^2 + 10^2 + 2^2 + 4^2 + 12^2 + 8^2 = 364\end{aligned}$$

Using the shortcut formula for s^2 , we obtain

$$\begin{aligned}s^2 &= \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i \right)^2}{n} \right] \\ &= \frac{1}{5} \left[364 - \frac{(42)^2}{6} \right] \\ &= 14\end{aligned}$$

Thus, the sample standard deviation is

$$s = \sqrt{s^2} = \sqrt{14} = 3.74$$

Treating the data in example above as a population, calculations of the mean, variance, and standard deviation of the population of data change.

Solution

The population mean is

$$\begin{aligned}\mu &= \frac{\sum_{i=1}^N x_i}{N} = \frac{\sum_{i=1}^6 x_i}{6} \\ &= \frac{6 + 10 + 2 + 4 + 12 + 8}{6} = \frac{42}{6} = 7\end{aligned}$$

The population variance is

$$\begin{aligned}\sigma^2 &= \frac{\sum_{i=1}^N (x_i - \mu)^2}{N} = \frac{\sum_{i=1}^6 (x_i - 7)^2}{6} \\ &= \frac{(6-7)^2 + (10-7)^2 + (2-7)^2 + (4-7)^2 + (12-7)^2 + (8-7)^2}{6} \\ &= \frac{70}{6} \\ &= 11.67\end{aligned}$$

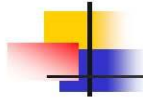
Thus, the population standard deviation is

$$\sigma = \sqrt{\sigma^2} = \sqrt{11.67} = 3.42$$

N.B 1. When data are presented into a frequency distribution, the formula for computing the standard

deviation becomes: $S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n f_i (x_i - \bar{x})^2}$

2. In case of grouped data into classes, the standard deviation is:



Standard Deviation

(continued)

■ Approximating the Standard Deviation

- Used when the raw data are not available and the only source of data is a frequency distribution

- $$S = \sqrt{\frac{\sum_{j=1}^c (m_j - \bar{X})^2 f_j}{n - 1}}$$

n = sample size

c = number of classes in the frequency distribution

m_j = midpoint of the j th class

f_j = frequencies of the j th class

© 2003 Prentice-Hall, Inc.

Chap 3-19

Using the shortcut method:

$$\bar{x} = \frac{\sum_{i=1}^k f_i m_i}{n}$$
$$s^2 = \frac{1}{n-1} \left[\sum_{i=1}^k f_i m_i^2 - \frac{\left(\sum_{i=1}^k f_i m_i \right)^2}{n} \right]$$

Example

Approximate the mean and standard deviation of the sample data that have the following frequency distribution:

Class	Frequency
15 up to 25	30
25 up to 35	78
35 up to 45	96
45 up to 55	62
55 up to 65	34

Solution

putation

A convenient method of computing the summations needed in the approximation formulas is to use a table such as the following one:

Class i	Class Limits	Frequency f_i	m_i	Midpoint $f_i m_i$	$f_i m_i^2$
1	15 up to 25	30	20	600	12,000
2	25 up to 35	78	30	2,340	70,200
3	35 up to 45	96	40	3,840	153,600
4	45 up to 55	62	50	3,100	155,000
5	55 up to 65	34	60	2,040	122,400
	Total	$n = 300$		11,920	513,200

$$\bar{x} \cong \frac{\sum_{i=1}^5 f_i m_i}{n} = \frac{11,920}{300} = 39.7$$

$$s^2 \cong \frac{1}{n-1} \left[\sum_{i=1}^5 f_i m_i^2 - \frac{\left(\sum_{i=1}^5 f_i m_i \right)^2}{n} \right]$$

$$= \frac{1}{299} \left[513,200 - \frac{(11,920)^2}{300} \right]$$

$$= 132.37$$

$$s = \sqrt{132.37} = 11.51$$

❖ Interpreting Standard Deviation: Chebyshev's theorem and Empirical Rule

An important attribute of the standard deviation as a measure of spread is that if the mean and standard deviation of a normal distribution are known, it is possible to compute the percentile rank associated with any given score.

The standard deviation of a set of measurements, taken by itself, is difficult to interpret. This section described two ways in which we can use the standard deviation to make a statement regarding the proportion of measurements that fall within various intervals of values centered at the mean value. The information depends on the shape of histogram. If the histogram is bell shaped, we can use the Empirical Rule. Otherwise we use Chebyshev's theorem.

The Empirical Rule:

The Empirical Rule makes more precise statements, but it can be applied only to samples of measurements with a mound-shaped distribution.

For such a sample of measurements, the Empirical Rule states that:

1. **Approximately 68%** of the observations fall within 1 standard deviation of the mean,
2. **Approximately 95%** of the observations fall within 2 standard deviations of the mean,
3. And virtually all the observations (**approximately 99.7%**) fall within 3 standard deviations of the mean.

Chebyshev's theorem:

Chebyshev's theorem which applies to any set of measurements (all shapes of histograms), states that the proportion of observations that lie within k standard deviations of the mean is at least

$$1 - 1/k^2, \text{ where } k \geq 1.$$

When $k=2$, the Chebyshev's theorem states that at least three-quarters (75%) of all observations lie within two standard deviations of the mean. With $k=3$, Chebyshev's theorem states that at least eight-ninths (88.9%) of all observations lie within three standard deviations of the mean.

Example1:

A professor has announced that the grades on a statistics exam have a mean value of 72 and a standard deviation of 6. Not knowing anything about the shape of the distribution of grades, what can we say about the proportion of grades that are between:

- a) 66 and 78?
- b) 60 and 84?
- c) 54 and 90?

Solution

- a) Expressing the interval (66, 78) as

$$(66, 78) = (72 - 6, 72 + 6) = (\mu - s, \mu + s)$$

We observe that we are dealing with the interval of values that fall within $k = 1$ standard deviations of the mean. With $k = 1$, Chebyshev's theorem states that at least $1 - 1/1^2 = 0$ of the grades fall within the interval (66, 78). This statement provides no useful information. It may be that no grades fall within this interval, or that many do.

- b) We can express the interval (60, 84) as

$$(60, 84) = (72 - 12, 72 + 12) = (\mu - 2s, \mu + 2s)$$

With $k = 2$, Chebyshev's theorem states that at least $1 - 1/2^2 = 3/4$ of the grades fall within the interval (60, 84). This implies that at most $1/4$ of the grades do not fall between 60 and 84.

- c) We can express the interval (54, 90) as

$$(54, 90) = (72 - 18, 72 + 18) = (\mu - 3s, \mu + 3s)$$

With $k = 3$, Chebyshev's theorem states that at least $1 - 1/3^2 = 8/9$ of the grades fall within the interval (54, 90). This implies that at most 1/9 of the grades do not fall between 54 and 90.

Example 2:

What would your answer to example above be if the professor also announced that the grades have a mound-shaped distribution?

Solution

- Since the grades have a mound-shaped distribution, we can use the Empirical Rule, which states that approximately 68% of the grades fall within 1 standard deviation of the mean. In other words, 68% of the grades fall between 66 and 78. Notice that this implies that approximately 32% of the grades do not fall between 66 and 78. Furthermore, since a mound-shaped distribution is symmetrical, approximately 16% of the grades are lower than 66 and 16% are higher than 78.
- Approximately 95% of the grades fall between 60 and 84.
- Virtually all the grades fall between 54 and 90.

- **Coefficient of variation (C.V)**

The coefficient of variation represents the ratio of the standard deviation to the mean, and it is a useful statistic for comparing the degree of variation from one data series to another, even if the means are drastically different from each other.

$$c_v = \frac{\sigma}{\mu}$$

This is only defined for *non-zero* mean, and is most useful for variables that are always *positive*. The CV is often presented as the given ratio multiplied by 100

The standard deviations of two variables, while both measure dispersion in their respective variables, cannot be compared to each other in a meaningful way to determine which variable has greater dispersion because they may vary greatly in their units and the means about which they occur. The standard deviation and mean of a variable are expressed in the same units, so taking the ratio of these two allows the units to cancel. This ratio can then be compared to other such ratios in a meaningful way: between two variables (that meet the assumptions outlined below), the variable with the smaller CV is less dispersed than the variable with the larger CV.

Let us assume a given distribution, which has a mean of 52.645 and a standard deviation of 9.368,
CV=(9.368/52.645)*100=17.796%

Summary on CV for a sample data

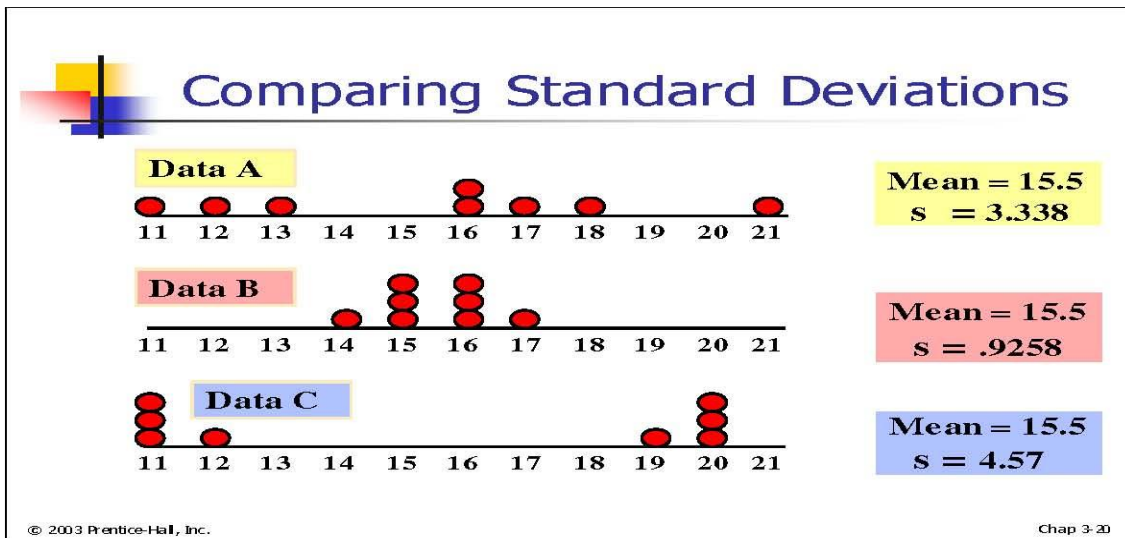
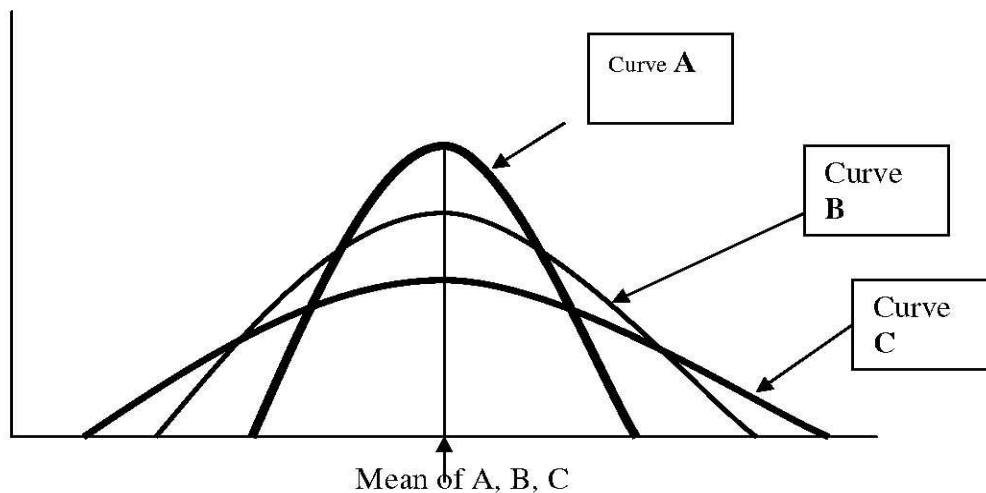


Figure: Three curves with the same mean but with different variations.



Coefficient of Variation

- Measure of Relative Variation
- Always in Percentage (%)
- Shows Variation Relative to the Mean
- Used to Compare Two or More Sets of Data Measured in Different Units
- $CV = \left(\frac{S}{\bar{X}} \right) 100\%$
- Sensitive to Outliers

© 2003 Prentice-Hall, Inc. Chap 3-21

Exercises:

1 Consider the following sample of data:

17, 25, 18, 14, 28, 21

i) Compute each of the following for this sample: a) the mean; b) the range; c) the variance; d) the

standard deviation

ii) Treating the data in Exercise 1 as a population, calculate each of the following for this population:

a) The mean; b) the variance; c) the standard deviation

2. The mean and standard deviation of the weights of the sample of 25 workers are 168.8 pounds and 11.2 pounds, respectively.

i) Knowing nothing else about the distribution of the 25 weights, what can you say about the proportion of weights that fall between 146.4 and 191.2 pounds?

ii) If the distribution of these weights were mound-shaped, what proportion of the weights would be:

a) Between 157.6 and 180.0 pounds?

b) Below 146.4 pounds?

c) If we estimate the standard deviation of the sample of 25 weights using the range approximation of s . Why should your answer not be closer to the true value of $s = 11.2$ pounds?

3. Approximate the mean and standard deviation of the sample data that have the following frequency distribution:

Class	Frequency
0 up to 5	7
5 up to 10	55
10 up to 15	82
15 up to 20	44
20 up to 25	12

Chapter 4. Bivariate descriptive Measures

4.1 Measures of linear relationship

In Chapter 2 (Section 2.4), we learned how a scatter diagram can be used to assess whether or not there is a relationship between two variables, and to determine if the relationship is linear, nonlinear, positive, or negative. This section introduced numerical measures of the linear relationship between two variables X and Y :

- **The covariance**

If we had access to the entire population of values for X and Y , we could compute the population covariance:

$$\text{Population covariance} = \text{COV}(X, Y) = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{N}$$

where μ_x is the population mean of the first variable, X ; μ_y is the population mean of the second variable,

Y ; and N is the size of the population.

Usually, we will be working with only a sample of observations for X and Y . The sample covariance is defined in a similar manner, where n is the number of pairs of observations in the sample.

$$\text{Sample covariance} = \text{cov}(X, Y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

For convenience, we label the population covariance $\text{COV}(X, Y)$ and the sample covariance $\text{cov}(X, Y)$.

Recall, from the discussion of scatter diagrams (Section 2.4), that there is a **positive relationship** between two variables if their values tend to increase together and decrease together. Moreover, the relationship is **linear** if a straight line can be drawn through the scatter diagram that approximates the positive relationship between the two variables.

In general, if two variables have a positive linear relationship, the covariance between the two variables will be a positive number (Figure 4.1(a)).

On the other hand, if two variables have a negative linear relationship, the covariance between the two variables will be a negative number (Figure 4.1(b)). Finally, if two variables are unrelated the covariance will be close to zero (Figure 4.1(c)).

It is difficult to ascertain the **strength** of the relationship between X and Y from their covariance. A better measure for this purpose is the coefficient of correlation, obtained by dividing the covariance by the standard deviations of X and Y .

- **The coefficient of correlation**

The **population** coefficient of **correlation** is labeled ρ (Greek letter rho) and is defined as

$$\rho = \frac{\text{COV}(X, Y)}{\sigma_x \sigma_y}$$

Where σ_x and σ_y are the standard deviations of X and Y , respectively.

As mentioned previously, you will usually be working with only a sample of observations for X and Y . The **sample** coefficient of **correlation**, r , is defined as

$$r = \frac{\text{cov}(X, Y)}{s_x s_y}$$

where s_x and s_y are the sample standard deviations of X and Y , respectively.

The coefficient of correlation will always lie between -1 and $+1$. Its sign will be the same as the sign of the covariance and is interpreted in the same way. The closer the correlation is to $+1$, the stronger is the **positive** relationship between X and Y . Figure 4.1(d) depicts two variables whose correlation coefficient is $+1$. On the other hand, the closer the correlation is to -1 , the stronger is the **negative** relationship between X and Y . Figure 4.1(e) depicts two variables that are perfectly negatively correlated. Finally, a correlation close to zero indicates that **no linear relationship** exists, as in Figure 4.1(f).

Example:

In Chapter 2 (section 2.4), we used a scatter diagram to determine that there is a positive linear relationship between years of education and income level (in \$1,000s), based on the sample of data shown below. Using the same data, measure how these two variables are related by computing their covariance and coefficient of correlation.

x (years of education)	y (income in \$1,000s)
11	25
12	33
11	22
15	41
8	18
10	28
11	32
12	24
17	53
11	26

Solution:

We begin by calculating the sample means and standard deviations:

$$\bar{x} = 11.8$$

$$s_x = 2.53$$

$$\bar{y} = 30.2$$

$$s_y = 10.28$$

We then compute the deviations from the mean for both x and y , and compute their products. The following table summarizes these calculations.

x	y	$(x - \bar{x})$	$(y - \bar{y})$	$(x - \bar{x})(y - \bar{y})$
11	25	-0.8	-5.2	4.16
12	33	0.2	2.8	0.56
11	22	-0.8	-8.2	6.56
15	41	3.2	10.8	34.56
8	18	-3.8	-12.2	46.36
10	28	-1.8	-2.2	3.96
11	32	-0.8	1.8	-1.44
12	24	0.2	-6.2	-1.24
17	53	5.2	22.8	118.56
11	26	-0.8	-4.2	3.36
				Total = 215.4

Thus,

$$\text{cov}(X, Y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{215.4}{9} = 23.9333$$

The coefficient of correlation is

$$r = \frac{\text{cov}(X, Y)}{s_x s_y} = \frac{23.9333}{(2.53)(10.28)} = 0.92$$

There is a reasonably strong linear relationship between X and Y .

• Least Squares Method and Regression Line

The method of least squares assumes that the best-fit curve of a given type is the curve that has the minimal sum of the deviations squared (*least square error*) from a given set of data.

Suppose that the data points are (x_1, y_1) , (x_2, y_2) , ..., (x_n, y_n) where x is the independent variable and y is the dependent variable.

The fitting curve $f(x)$ has the deviation (error) d from each data point, i.e., $d_1 = y_1 - f(x_1)$, $d_2 = y_2 - f(x_2)$, ..., $d_n = y_n - f(x_n)$. According to the method of least squares, the best fitting curve has the property that:

$$\Pi = d_1^2 + d_2^2 + \dots + d_n^2 = \sum_{i=1}^n d_i^2 = \sum_{i=1}^n [y_i - f(x_i)]^2 = \text{a minimum}$$

The least squares method is used to produce the line that provides the best possible fit to the points in a scatter diagram. The coefficients b_0 and b_1 in the least squares (regression) line

$$\hat{y}_i = b_0 + b_1 x_i$$

are best calculated using the shortcut formulas:

$$b_1 = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} \quad \text{and} \quad b_0 = \bar{y} - b_1 \bar{x}$$

If $\text{cov}(X, Y)$ is already known, an alternative way of calculating b_1 is

$$b_1 = \frac{\text{cov}(X, Y)}{s_x^2}$$

Example:

- Find the least squares (regression) line for the data in Example 4.12 using the shortcut formulas for the coefficients.
- Use the relationship between b_1 and $\text{cov}(X, Y)$ to check the value you calculated for b_1 in part (a).
- Interpret the coefficients

Solution

- Notice that, in the example above, we've labeled years of education x and income y , because we believe that income is affected by years of education. Our first step is to calculate the sums $\sum x_i$, $\sum x_i^2$, $\sum y_i$, and $\sum x_i y_i$.

We find : $\sum x_i = 118$; $\sum x_i^2 = 1,450$, $\sum y_i = 302$, $\sum x_i y_i = 3,779$

Therefore,

The least squares line is therefore

$$b_1 = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} = \frac{3,779 - \frac{(118)(302)}{10}}{1,450 - \frac{(118)^2}{10}} = 3.74$$

$$b_0 = \bar{y} - b_1 \bar{x} = 30.2 - (3.74)(11.8) = -13.93$$

$$\hat{y} = b_0 + b_1 x = -13.93 + 3.74x$$

b) The relationship between b_1 and $\text{cov}(X, Y)$, We calculate the sample regression coefficients using the following formulas:

$$b_1 = \frac{\text{cov}(X, Y)}{s_x^2} \quad \text{and} \quad b_0 = \bar{y} - b_1 \bar{x}$$

Note that we've labeled education x and income y because income is affected by education. Our first step is to calculate the sums, , , $\sum x_i$, $\sum x_i^2$, $\sum y_i$, $\sum y_i^2$, and $\sum x_i y_i$. The sum $\sum y_i^2$ is not required in the least squares method but is usually needed for other techniques involved with regression. We find $\sum y_i^2$

$$\sum x_i = 118 ; \sum x_i^2 = 1,450 ; \sum y_i = 302 ; \sum y_i^2 = 10,072 ; \sum x_i y_i = 3,779$$

Next, we compute the covariance and the variance of x :

$$\begin{aligned} \text{cov}(X, Y) &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1} \\ &= \frac{1}{n-1} \left[\sum x_i y_i - \frac{\sum x_i \sum y_i}{n} \right] \\ &= \frac{1}{10-1} \left[3,779 - \frac{(118)(302)}{10} \right] \\ &= 23.93 \end{aligned}$$

$$s_x^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{1}{n-1} \left[\sum x_i^2 - \frac{(\sum x_i)^2}{n} \right] = \frac{1}{10-1} \left[1,450 - \frac{(118)^2}{10} \right] = 6.40$$

Therefore,

$$b_1 = \frac{\text{cov}(X, Y)}{s_x^2} = \frac{23.93}{6.40} = 3.74$$

This is the same value for b_1 that was calculated in part (a).

$$\text{The sample means are : } \bar{x} = \frac{\sum x_i}{n} = \frac{118}{10} = 11.8 ; \bar{y} = \frac{\sum y_i}{n} = \frac{302}{10} = 30.2$$

We can now compute the y-intercept

$$b_0 = \bar{y} - b_1\bar{x} = 30.2 - (3.74)(11.8) = -13.93$$

Thus, the least squares regression line is

$$\hat{y} = -13.93 + 3.74x$$

c) Interpret the coefficients

The sample slope $b_1 = 3.74$ tells us that on average for each additional year of education, an individual's income rises by \$3.74 thousand. The y-intercept is $b = -13.93$. It should be obvious that this value has no meaning. Recall that whenever the range of the observed values of x does not include zero it is usually pointless to try to interpret the meaning of the y-intercept.

N.B:

1. The techniques we developed above require that both variables are quantitative.
2. One should ask Question: How do I know which the dependent variable is and which is the independent variable?

Answer: The dependent variable is the one that we want to forecast or analyze. The independent variable is hypothesized to affect the dependent variable. In example above, we wish to analyze income, and we choose as the variable that most affects income the individual's education. Hence, we label income y and education x .

3. Having the regression line, we should estimate the dependent values by substituting numerical values for the independent variable on the right side of equation.

- **Coefficient of Determination**

This tells us the proportion of the variation of dependent variable explained by the variation in independent variable.

The coefficient of determination is defined as

$$R^2 = \frac{[cov(X,Y)]^2}{s_x^2 s_y^2}$$

Calculate the coefficient of determination for Example above:

Solution

$$R^2 = \frac{[cov(X,Y)]^2}{s_x^2 s_y^2} = \frac{[23.93]^2}{(6.40)(105.73)} = .8463$$

The coefficient of determination tells us that 84.63% of the variation in y is explained by the variation in x . The remaining 15.37% is unexplained.

Exercises

1. In Exercise (section 2.4), you were asked to construct a scatter diagram and use it to assess the relationship between the number of ads placed by a store and the number of customers, based on the sample of data shown below.

Month	Number of Ads, x	Number of customers, y
-------	-----------------------	--------------------------

1	5	528
2	12	876
3	8	653
4	6	571
5	4	556
6	15	1,058
7	10	963
8	7	719

- Using the same data, measure how these two variables are related by computing their covariance and coefficient of correlation and interpret the strength of the relation.
- Find the equation of the regression line and interpret its coefficients
- Calculate the coefficient of determination and interpret it.
- Estimate the number of customers if the number of ads is 14.

2. Fifteen observations were taken to estimate a simple regression model. The following summations were produced:

$$\sum x = 50 \quad \sum x^2 = 250 \quad \sum y = 100 \quad \sum y^2 = 1,100 \quad \sum xy = 500$$

- Find the least squares regression line.
 - Estimate the value of y if x=8
3. Solve the question (1) by using logarithmic transformation of number of customers' values.

4. The owner of the concession stands at a football stadium would like to be capable of predicting the number of cups of coffee sold during each game. He believes that the most important variable is the temperature at game time. To investigate the relationship, he recorded the number of cups of coffee sold and the temperature during nine randomly selected games. These data are shown below.

Game	Temperature, x (°F)	Number of Cups of Coffee Sold (100's), y
1	53	50
2	50	47

3	75	43
4	48	58
5	45	57
6	63	44
7	40	64
8	55	48
9	30	71

- Calculate covariance and coefficient of correlation and interpret the strength of the relation.
- Find the least squares regression line.
- Calculate the coefficient of determination and interpret it.
- Estimate the number of cups of temperature if the temperature is fixed at 60.