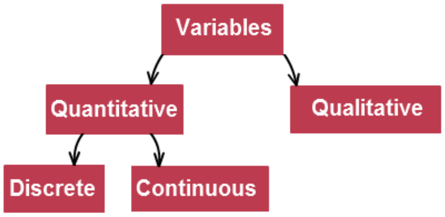
Chapter 1

Introduction

In todays society, decisions are made on the basis of data. Most scientific or industrial studies and experiments produce data, and the analysis of these data and drawing useful conclusions from them become one of the central issues. The field of statistics is concerned with the scientific study of collecting, organizing, analyzing, and drawing conclusions from data. Statistical methods help us to transform data to knowledge. Statistical concepts enable us to solve problems in a diversity of contexts, add substance to decisions, and reduce guesswork. The discipline of statistics stemmed from the need to place knowledge management on a systematic evidence base. Earlier works on statistics dealt only with the collection, organization, and presentation of data in the form of tables and charts. In order to place statistical knowledge on a systematic evidence base, we require a study of the laws of probability. In mathematical statistics we create a probabilistic model and view the data as a set of random outcomes from that model. Advances in probability theory enable us to draw valid conclusions and to make reasonable decisions on the basis of data.

1.1 Variable



Variable is a characteristic that vary from one element to another element.For example, Income, score, taste,temperature etc. There are two types of variables , qualitative and quantitative variables.

1.1.1 Quantitative variable

Quantitative variables are those which represent quantity of something. These variables assign numeric values in them. Quantitative variables are the variables that describe:how much or how many. For example:Marks obtained, height and weight. These variables are quantitative since they assign only numerical data.

Discrete variable

A variable whose values are countable is called discrete variable. Discrete variable assume integer value over a certain interval. For example, marks of students, number of sales man in a company etc.

Continuous variable

A variable that can assume any numerical value over a certain interval.For example, height and weight are of students of management department are continuous variables

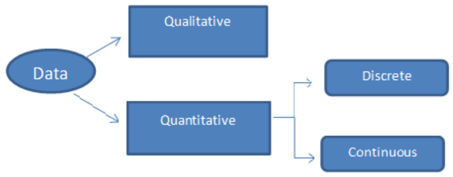
1.1.2 Qualitative Variables

A variable that cannot assume numerical values, but can be classified into two or more non-numeric categories is called qualitative or categorical variable. For example color, gender of a person etc.

Example

Data from a pharmaceutical company medical study contain values of many variables for each of the people who were the subjects of the study.Which of the following variables are categorical and which are quantitative?
(a) Gender(female or male)
(b) Age(years)
(c) Race(Asian,black,white,or other)
(d) Smoker(yes or no)
(e) Systolic blood pressure(millimeters of mercury)
(f) Level of calcium in the blood(micrograms per milliliter)

1.2 Data and its types



A data set is a collection of information on one or more variable.For example, marks of 40 students in mathematics, opinion of 100 voters.

1.2.1 Cross-Sectional data

Cross-sectional data are data collected on different elements or variables at the same point in time or for the same period of time.

1.2.2 Time series data

Time series data are data collected on the same element or the same variable at different points in time or for different periods of time.

1.2.3 Quantitative and Qualitative data

Quantitative data are observations measured on a numerical scale. Non numerical data that can only be classified into one of the groups of categories are said to be qualitative or categorical data.

Example

Data on response to a particular therapy could be classified as no improvement, partial improvement, or complete improvement. These are qualitative data. The number of minority-owned businesses in Florida is quantitative data. The marital status of each person in a statistics class as married or not married is qualitative or categorical data. The number of car accidents in different U.S. cities is quantitative data. The blood group of each person in a community as O, A, B, AB is qualitative data

1.2.4 Primary data and its Sources

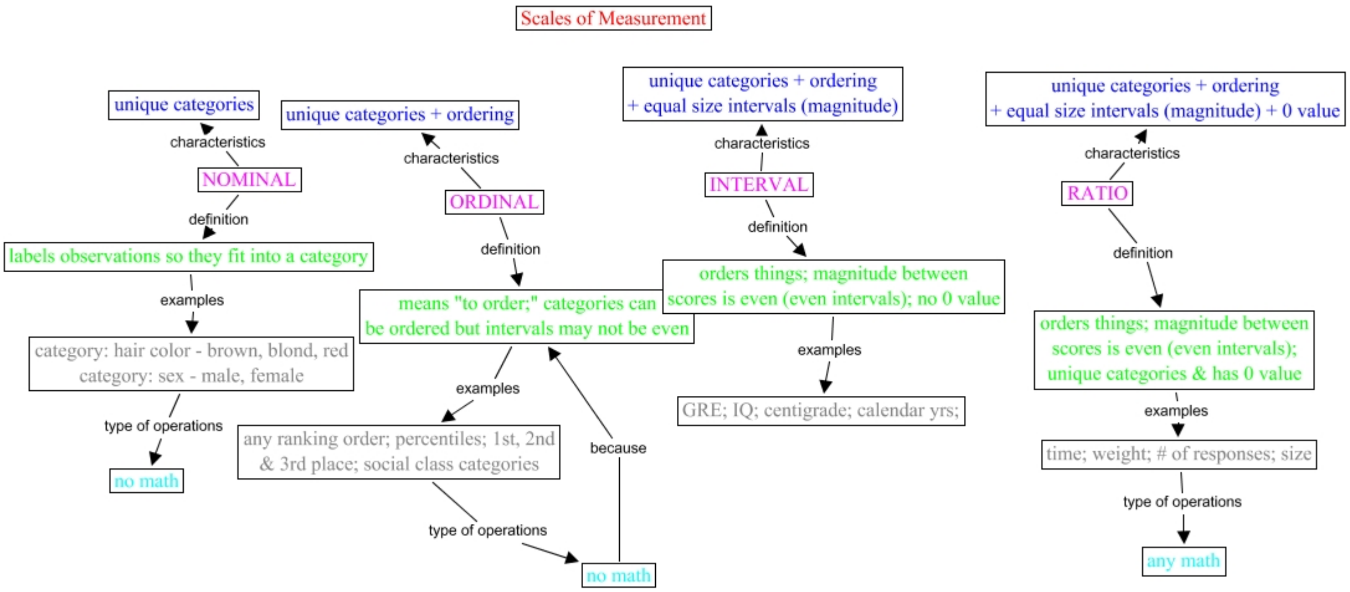
Data observed or collected directly from first-hand experience.

1.2.5 Secondary data and its Sources

Published data and the data collected in the past or other parties is called secondary data

1.3 Measurement Scales

The scaling techniques are used in measuring in different variables and assessing the validity and reliability of various measures.


Scales of Measurement
unique categories
characteristics
NOMINAL
definition
labels observations so they fit into a category
examples
category: hair color - brown, blond, red
category: sex - male, female
type of operations
no math
unique categories + ordering
characteristics
ORDINAL
definition
means "to order;" categories can be ordered but intervals may not be even
examples
any ranking order; percentiles; 1st, 2nd & 3rd place; social class categories
type of operations
no math
because
unique categories + ordering + equal size intervals (magnitude)
characteristics
INTERVAL
definition
orders things; magnitude between scores is even (even intervals); no 0 value
examples
GRE; IQ; centigrade; calendar yrs;
unique categories + ordering + equal size intervals (magnitude) + 0 value
characteristics
RATIO
definition
orders things; magnitude between scores is even (even intervals); unique categories & has 0 value
examples
time; weight; # of responses; size
type of operations
any math

1.3.1 Nominal scale

Categorical variables are measures on nominal scale where the measurements involve the naming or categorization of possible values of a variables. The measurements are produced as qualitative in the sense that the categories are different from each other merely. The numbers are assigned to the categories in any order, hence they are simply label and do not represent real quantity. Sex, blood type, religion,nationality, ethnicity, language etc are examples of nominal scaling.

The categorical equality and equivalence are operations which apply to objects of the nominal scale. The mode is allowed as the measure of central tendency for the nominal type. On the other hand, the median, i.e. the middle-ranked item, makes no sense for the nominal type of data since ranking is meaningless for the nominal type.

1.3.2 Ordinal scale

If the data obtained is classified into different categories with respect to value of a variable with proper ordering, it is referred to as ordinal scaling. This scale is used in ranking respondents according to some characteristic such as students in order of grade points, completely agree', 'mostly agree', 'mostly disagree', 'completely disagree' when measuring opinion.

The mode and median is allowed as the measure of central tendency. Thus means and standard deviations have no validity, but they can be used to get ideas for how to improve operationalization of variables used in questionnaires.

1.3.3 Interval Scale

Interval scale is an ordered scale in which difference between the measurements is meaningful quantity. In other words, this scale not only separates individuals by rank order, it is also measures the distance between rank positions in equal units. For example, the distance between position 2 and 4 is stated to be the same as between position 5 and 7, but not how much because a zero position has not been established.

The mode, median, and arithmetic mean are allowed to measure central tendency of interval variables, while measures of statistical dispersion include range and standard deviation.

1.3.4 Ratio scale

If the difference between the measurement is meaning full quantity and is equal at all points on a scale with respect to zero point, it is referred to as ratio scale. With this scale one can compare interval and rank objects according to the magnitude. Few example of ratio scale measurement are salary, weight, blood pressure etc.

The geometric mean and the harmonic mean are allowed to measure the central tendency, in addition to the mode, median, and arithmetic mean. The range and the coefficient of variation are allowed to measure statistical dispersion. All statistical measures are allowed because all necessary mathematical operations are defined for the ratio scale.

Example

The Wall Street Journal subscriber survey (October 13, 2003) asked 46 questions about subscriber characteristics and interests. State whether each of the following questions provided

qualitative or quantitative data and indicate the measurement scale appropriate for each.
a. What is your age?
b. Are you male or female? c. When did you first start reading the WSJ? High school, college, early career, midcareer, late career, or retirement?
d. How long have you been in your present job or position?
e. What type of vehicle are you considering for your next purchase? Nine response categories include sedan, sports car, SUV, minivan, and so on.

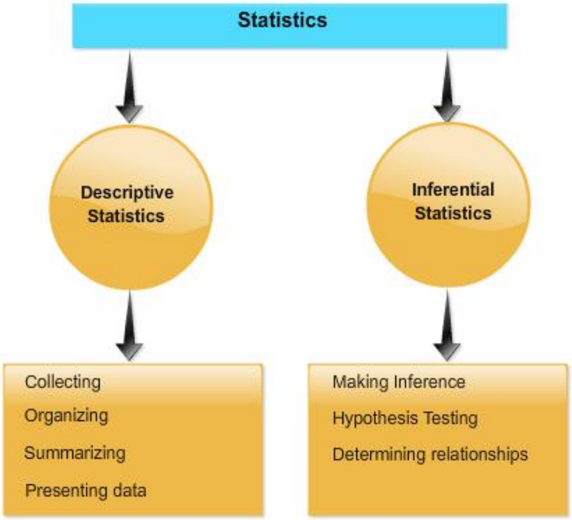
Example

Bill scored 1200 on the Scholastic Aptitude Test and entered college as a physics major. As a freshman, he changed to business because he thought it was more interesting. Because he made the deans list last semester, his parents gave him $30 to buy a new Casio calculator. For this situation, identify at least one piece of information in the
a. nominal scale of measurement.
b. ordinal scale of measurement.
c. interval scale of measurement.
d. ratio scale of measurement

Example

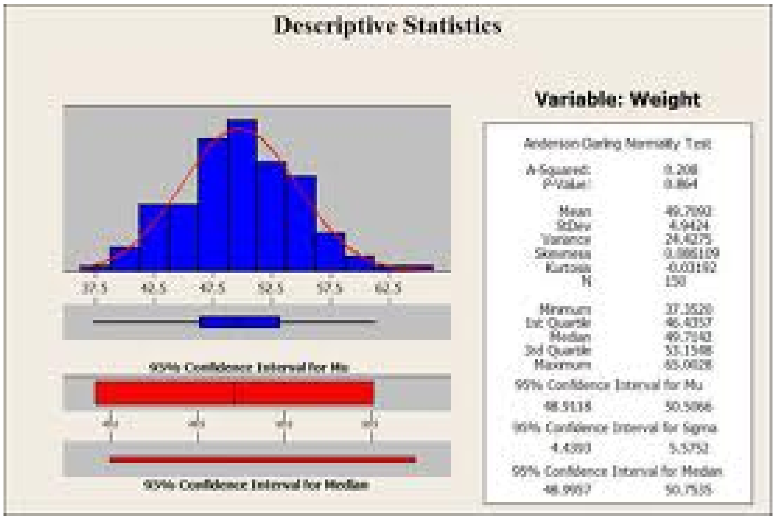
In studying the performance of the companys stock investments over the past year, the research manager of a mutual fund company finds that only 43% of the stocks returned more than the rate that had been expected at the beginning of the year.
a. Could this information be viewed as representing the nominal scale of measurement? If so, explain your reasoning. If not, why not?
b. Could this information be viewed as representing the ratio scale of measurement? If so, explain your reasoning. If not, why not?

1.4 Statistics



Statistics is a science and art that consists of methods and techniques used to collect, present, summarize,analyze and interpret the data.

1.4.1 Descriptive Statistics


Descriptive Statistics
37.5
42.5
47.5
52.5
57.5
62.5
95% Confidence Interval for Mu
95% Confidence Interval for Median

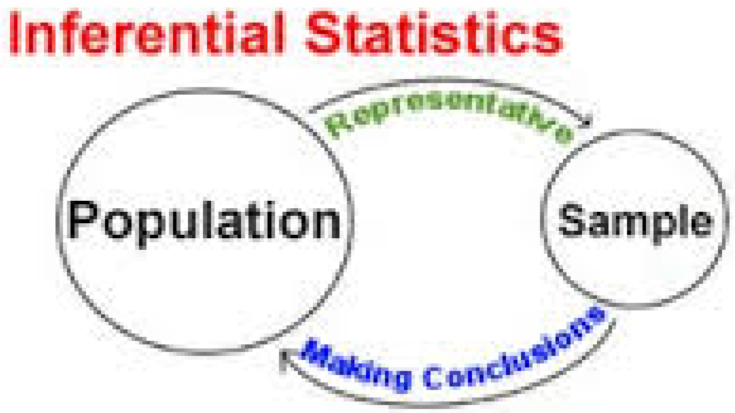
```
Variable: Weight

    Anderson-Darling Normality Test
    A-Squared:          0.208
      P-Value:          0.864

         Mean           49.7092
        StDev           4.9424
     Variance           24.4275
     Skewness           0.086109
     Kurtosis           -0.03192
            N           150

      Minimum           37.3520
 1st Quartile           46.4357
       Median           49.7142
 3rd Quartile           53.1548
      Maximum           65.0028
  95% Confidence Interval for Mu
      48.9118           50.5066
  95% Confidence Interval for Sigma
       4.4393           5.5752
  95% Confidence Interval for Median
      48.9957           50.7535
```

In descriptive statistics, we deal with collection of data, its presentation in various forms, such as tables, graphs and diagrams and findings averages and other measures which would describe the data.

For example.

Industrial statistics, population statistics, trade statistics etc Such as businessman make to use descriptive statistics in presenting their annual reports, final accounts, bank statements.

1.4.2 Inferential Statistics



In inferential statistics, we study the techniques used for analysis of data, making the estimates and drawing conclusions from limited information taken on sample basis and testing the reliability of the estimates

For example.

Suppose, we want to have an idea about the percentage of illiterates in our country. We take a sample from the population and find the proportion of illiterates in the sample. This sample proportion with the help of probability enables us to make some inferences about the population proportion. This study belongs to inferential statistics

Example

What is the difference between descriptive statistics and inferential statistics? Which branch is involved when a state senator surveys some of her constituents in order to obtain guidance on how she should vote on a piece of legislation?

Example

In 2002, the Cinergy Corporation sold 35,615 million cubic feet of gas to residential customers, an increase of 1.1% over the previous year. Does this information represent descriptive statistics or inferential statistics? Why?

Example

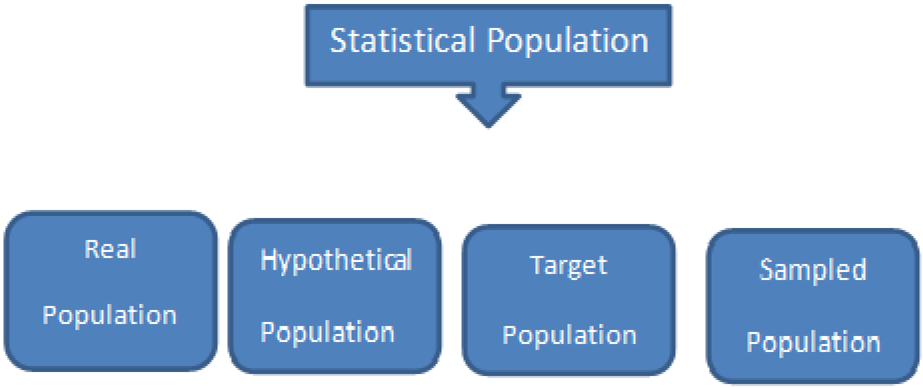
An article in Runners World magazine described a study that compared the cardiovascular responses of 20 adult subjects for exercises on a treadmill, on a mini trampoline, and jogging in place on a carpeted surface. Researchers found average heart rates were significantly less on the mini trampoline than for the treadmill and stationary jogging. Does this information represent descriptive statistics or inferential statistics? Why? SOURCE: Kate Delhagen, Health Watch, Runner's World

Example

A research firm observes that men are twice as likely as women to watch the Super Bowl on television. Does this information represent descriptive statistics or inferential statistics? Why?

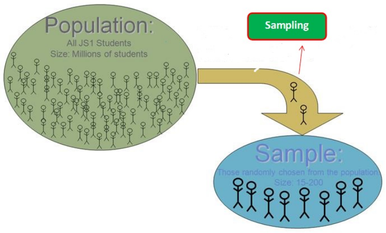
.

1.5 Statistical population



A population is a collection of objects or units about which we want to know something or draw an inference. The population may be finite or infinite . Assume a population consists of electric bulbs produced by a plant. Suppose, we want to estimate the average life of electric bulbs produced by a plant. The population consists of number of bulbs produced by the plant.

1.6 Sample


Population:
All JS1 Students
Size: Millions of students
Sampling
Sample:
Those randomly chosen from the population
Size: 15-200

A sub set of the population, which represents the entire population, is called a sample. Sample is a representative part of the population. The sample is denoted by s and its size by n

Example

We wish to estimate the percentage of defective parts produced in a factory during a given week (five days) by examining 20 parts produced per day. The parts will be examined each day at randomly chosen times. In this case all parts produced during the week is the population and the (100) selected parts for five days constitutes a sample.

Example

Consider we want to find the percentage of ticket less travelers in the Train. Then all persons in all the Train will constitute the population and the persons checked by a particular checker(s) will form a sample.

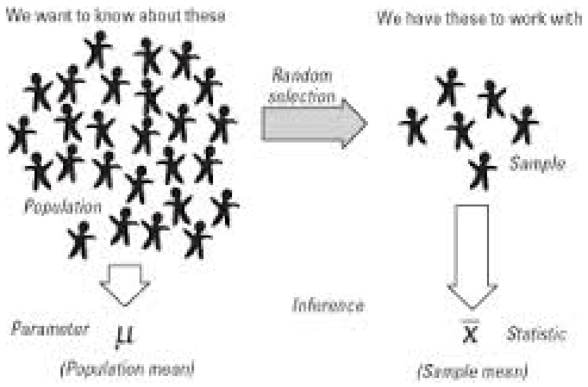
Example

Roger Amster teaches an English course in which 40 students are enrolled. After yesterday?s class, Roger questioned the 5 students who always sit in the back of the classroom. Three of the 5 said ?yes? when asked if they would like A Tale of Two Citiesas the next class reading assignment.

a. Identify the population and the sample in this situation.

b. Is this likely to be a representative sample? If not, why not?

1.7 Parameter and statistic



A numerical value that describes the characteristic of a population is known as parameter. Parameter is a fixed quantity. A numerical value that describes the characteristic of a sample is known as Statistic. Statistic is not fixed but vary from sample to sample

1.8 Types of Statistical polulation

1.8.1 Finite population

If the number of objects or units in the population is countable , it is said to be a finite population. For example, the number of houses in Attock is a finite population.

1.8.2 Real population

Real population consists of units that physically exist.For example units are Students, items produced by company.

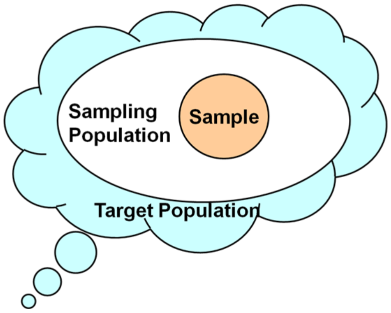
1.8.3 Hypothetical population

Hypothetical population consists of units that are not physically exit, bur are outcomes of random experiment. For example, population of all possible outcomes of lab experiment.

1.8.4 Infinite population

If the number of objects or units in the population is infinite, it is said to be an infinite population. For example, the number of stars in the sky forms an infinite population.In general, the population is denoted by Ω and its size is denoted by N . In the case of infinite population,$N \to \infty$

Target population

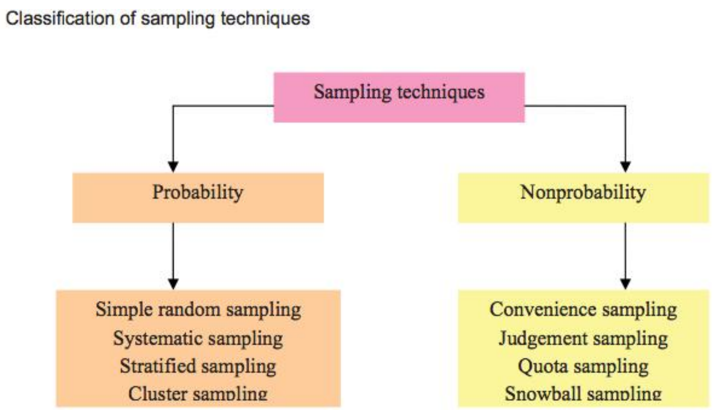


A finite or infinite population about which we require information is called target population. For example, old students of management department of CIIT Attock

Sampled or Study population

This is the basic finite set of individuals we intend to study. For example, all 18 year old students of management department of CIIT Attock whose permanent address is in Attock city.

1.9 Sampling



The process of selecting a sample from the population is known as sampling. There are two types sampling, probability / random sampling and non-probability / non random sampling.

Sampling with replacement

Sampling is called with replacement when a unit selected at random from the population is returned to the population and then a second element is selected at random. Whenever a unit is selected, the population contains all the same units. A unit may be selected more than once. There is no change at all in the size of the population at any stage. We can assume that a sample of any size can be selected from the given population of any size.
Example:
Consider a population: 2,5,7
Draw all possible sample of size 2
The list of 9 possible sample is:
(2,2) (2,5) (2,7), (5,2) (5,5) (5,7) (7,2) (7,5) (7,7)

Sampling without replacement

Sampling is called without replacement when a unit is selected at random from the population and it is not returned to the main lot. First unit is selected out of a population of size N and the second unit is selected out of the remaining population of N-1 units and so on. Thus the size of the population goes on decreasing as the sample size n increases. The sample size n cannot exceed the population size N. The unit once selected for a sample cannot be repeated in the same sample. Thus all the units of the sample are distinct from one another.
Example: Example:
Suppose population size is N = 6 with the observable values 10, 4, 17, 6, 8, 15.

Suppose we want a sample of size 2 without replacement.

How many possibilities are there to choose 2 out of 6 members?

A list of all 15 possible sample is (10 ,4), (10 ,17), (10 ,6), (10 ,8), (10 ,15),(4 ,17), (4 ,6), (4 ,8), (4 ,15)(17 ,6), (17 ,8), (17,15) , (6 ,8), (6,15), (8,15)

1.10 Reason for sampling

When studying characteristics of a population, there are many practical reasons why we prefer to select portions or samples of a population to observe and measure. Some of the reasons for sampling are:
1. The time to contact the whole population may be prohibitive. A candidate for a national office may wish to determine her chances for election. A sample poll using the regular staff and field interviews of a professional polling firm would take only 1 or 2 days. By using the same staff and interviewers and working 7 days a week, it would take nearly 200 years to contact all the voting population! Even if a large staff of interviewers could be assembled, the benefit of contacting all of the voters would probably not be worth the time.
2. The cost of studying all the items in a population may be prohibitive. Public opinion polls and consumer testing organizations, such as Gallup Polls and Roper ASW, usually contact fewer than 2,000 of the nearly 60 million families in the United States. One consumer research organization charges about $40,000 to mail samples and tabulate responses in order to test a product (such as breakfast cereal, cat food, or perfume). The same product test using all 60 million families would cost about $1 billion. A table of random numbers is an efficient way to select members of the sample
3. The physical impossibility of checking all items in the population. The populations of fish, birds, snakes, mosquitoes, and the like are large and are constantly moving, being born, and dying. Instead of even attempting to count all the ducks in Canada or all the fish in Lake Erie, we make estimates using various techniques-such as counting all the ducks on a pond picked at random, making creel checks, or setting nets at predetermined places in the lake.
4. The destructive nature of some tests. If the wine tasters at the Sutter Home Winery in California drank all the wine to evaluate the vintage, they would consume the entire crop, and none would be available for sale. In the area of indus- trial production, steel plates, wires, and similar products must have a certain minimum tensile strength. To ensure that the product meets the minimum standard, the Quality Assurance Department selects a sample from the current production. Each piece is stretched until it breaks, and the breaking point (usually measured in pounds per square inch) recorded. Obviously, if all the wire or all the plates were tested for tensile strength, none would be available for sale or use. For the same reason, only a sample of photographic film is selected and tested by Kodak to determine the quality of all the film produced, and only a few seeds are tested for

germination by Burpee prior to the planting season.
5. The sample results are adequate. Even if funds are available, it is doubtful the additional accuracy of a 100 percent sample-that is, studying the entire population-is essential in most problems. For example, the federal government uses a sample of grocery stores scattered throughout the United States to determine the monthly index of food prices. The prices of bread, beans, milk, and other major food items are included in the index. It is unlikely that the inclusion of all grocery stores in the United States would significantly affect the index, since the prices of milk, bread, and other major foods usually do not vary by more than a few cents from one chain store to another. When selecting a sample, researchers or analysts must be very careful that the sample is a fair representation of the population. In other words, the sample must be unbiased. In Chapter 1, an example of abusing statistics was the intentional selection of dentists to report that ”2 out of 3 dentists surveyed indicated they would recommend Brand X toothpaste to their patients.” Clearly, people can select a sample that supports their own biases. The ethical side of statistics always requires unbiased sampling and objective reporting of results. Next, several sampling methods show how to select a fair and unbiased sample from a population.

1.11 Probability or Random sampling and its types

Probability sampling is any method of selection of a sample based on the theory of probability. In this sampling scheme, each unit of population has known chance being included in the sample.

1.11.1 Simple random sampling

A sample selected in such a way that every element of the population has an equal chance of being chosen is called a simple random sample. Before applying this method, complete list of all units in population must be prepared.

This sampling is suitable for sampling if (1) population is homogeneous. (2) Assumption of independence of sample members is required for statistical test.

1.11.2 Stratified random sampling

Sampling in which the population is divided into homogeneous groups (called strata) according to some characteristic. Then A random sample is selected from each strata.

Chapter 2

Presentation of Data

Objectives

- It condense the raw data into a form suitable for statistical analysis.
- It remove the complexities and highlights the feature of data.
- It facilitates comparison and drawing conclusion from data.
- it helps in statistical analysis by separating elements of data set into homogeneous groups and hence bring out the points of similarity and dissimilarity.

2.1 Textual method

Data can be presented using paragraphs or sentences. It involves enumerating important characteristics, emphasizing significant figures and identifying the important features of data.

- **Rearrangement from lowest to highest**.
- **Stem and leaf plot**.

A stem-and-leaf display conveys information about the following aspects of the quantitative data:

- Identification of a typical or representative value
- Extent of spread about the typical value
- Presence of any gaps in the data

- Extent of symmetry in the distribution of values
- Number and location of peaks
- Presence of any outlying values

Although the stem-and-leaf display may appear to offer the same information as a histogram, it has two primary advantages. First, The stem-and-leaf display is easier to construct by hand. Second, Within a class interval, the stem-and-leaf display provides more information than the histogram because the stem-and-leaf shows the actual data.

STEPS for Stem and Leaf

1. Select one or more leading digits for the stem values. The trailing digits become the leaves.
2. .List possible stem values in a vertical column.
3. Record the leaf for every observation beside the corresponding stem value.
4. Order the leaves from smallest to largest on each line.
5. Indicate the units for stems and leaves someplace in the display

EXAMPLE
You are asked to present the performance of your section in the Statistics test. The following are the test scores of your class:
34 42 20 50 17 9 34 43 50 18 35 43 50 23 23 35 37 38 38 39 39 38 38 39 24 29 25 26 28 27 44 44 49 48 46 45 45 46 45 46
Rearrangement of data from lowest to highest
9 17 18 20 23 23 24 25 25 26 27 28 29 34 34 3 35 37 38 38 38 38 39 39 39 42 43 43 44 44 45 45 45 46 46 46 48 49 50 50 50
Stem-and-leaf plot.

Steam	Leaves
0	9
1	7, 8
2	0, 3, 3, 4, 5, 6, 7, 8, 9
3	4, 4, 5, 5, 7, 8, 8, 8, 8, 9, 9, 9
4	2, 3, 3, 4, 4, 5, 5, 5, 6, 6, 6, 8, 9
5	0, 0, 0,

2.2 Tabular method

Frequency Distribution
A frequency distribution is a table that distributes a data set into a suitable number of categories (classes). Rather than retaining the entire set of data in a display, a frequency table essentially provides only a count of those observations that are associated with each class. Once the data are summarized in the form of a frequency distribution, a graphical representation can be given through bar graphs, pie charts, and histograms. Data presented in the form of a frequency table are called grouped data.

Merits

- The data are expressed as more compact form.
- one can quickly note the pattern of distribution of observations falling in various classes.
- It permits the uses of more complex statistical techniques which reveal certain other hidden characteristics of data.

Demerits

In the process of grouping, individual observation lose their identity
STEPS for Frequency Distribution

1. Decide on the number of non overlapping groups or classes. H.A Sturges proposed a rule,
   K=1+3.22 log N, k denotes the number of classes
2. Determine the range of variation in data. R=Maximum value- Minimum value
3. Determine the class width or interval. Class Interval=R/K
4. Determine where to locate the class limit
5. Distribute the data into appropriate classes.
6. Make frequency column

EXAMPLE
The following data represents the marks of 30 students in the subject" Introduction to Statistics" out of 100 marks. 61, 65, 35, 62, 80, 70, 67, 90, 67, 80, 70, 54, 75, 62, 71, 80, 65, 71,

76, 67, 54, 54, 60, 70, 70, 90, 60, 60,76,82
Make the frequency distribution of marks and discuss it.

2.3 Graphical method

Bar chart
A graph of bars whose heights represent the frequencies (or relative frequencies) of respective categories is called a bar graph.
Steps
Measure the categories on x-axis and numerical value on y-axis
Plot the data and draw bars.
Note:
Width of each bar should be same.
Height of each bar should be twice of its width
Gap between bars should be half of its width.
Pie Chart
EXAMPLE
The data represent the percentages of price increases of some consumer goods and services for the period December 2000 to December 2102 in a certain city. Construct a bar chart for given data'. Increases of Some Consumer Goods and Services is,
Medical Care 83.3%, Electricity 22.1%, Residential Rent 43.5%, Food 41.1%, Consumer Price Index 35.8%

Histogram
A histogram is a graph in which the bars are drawn adjacent to each other without any gaps. A histogram compresses a data set into a compact picture that shows the location of the mean and modes of the data and the variation in the data, especially the range. It identifies patterns in the data. This is a good aggregate graph of one variable. In order to obtain the variability in the data, it is always a good practice to start with a histogram of the data.
Steps

1. Construct frequency distribution
2. Measure class boundaries on horizontal axis with uniform interval.
3. Measure frequency (relative frequency, percentage frequency) on vertical axis.
4. Plot the data and draw bars.

EXAMPLE

Frequency Polygon
The frequency polygon is a graph that displays the data by using lines that connect points plotted for the frequencies at the midpoints of the classes. The frequencies are represented by the heights of the points Steps:
(1)- Draw histogram
(2)- Identify mid point of each bar and join these points with straight lines

Frequency Curve
Solution

Exercise:
1.1. Construct a stem-and-leaf display for the following data.
70 72 75 64 58 83 80 82 76 75 68 65 57 78 85 72
1.2. Construct a stem-and-leaf display for the following data.
11.3 9.6 10.4 7.5 8.3 10.5 10.0 9.3 8.1 7.7 7.5 8.4 6.3 8.8
1.3. A psychologist developed a new test of adult intelligence. The test was administered to 20 individuals, and the following data were obtained.
114 99 131 124 117 102 106 127 119 115 98 104 144 151 132 106 125 122 118 118
Construct a stem-and-leaf display for the data.
1.4. The 2004 Naples, Florida, minimarathon (13.1 miles) had 1228 registrants (Naples Daily News, January 17, 2004). Competition was held in six age groups. The following data show the ages for a sample of 40 individuals who participated in the marathon.
49 33 40 37 56 44 46 57 55 32 50 52 43 64 40 46 24 30 37 43 31 43 50 36 61 27 44 35 31 43 52 43 66 31 50 72 26 59 21 47
a. Show a stretched stem-and-leaf display.
b. What age group had the largest number of runners?
c. What age occurred most frequently?
d. A Naples Daily News feature article emphasized the number of runners who were 20-something. What percentage of the runners were in the 20-something age group? What do you suppose was the focus of the article?
1.5. Consider the following data.
14 21 23 21 16 19 22 25 16 16 24 24 25 19 16 19 18 19 21 12 16 17 18 23 25 20 23 16 20 19 24 26 15 22 24 20 22 24 22 20
a. Develop a frequency distribution
b. Develop a relative frequency distribution and a percentage frequency distribution.
c. Construct cumulative frequency distribution, cumulative relative frequency distribution, cumulative percentage frequency distribution
d. Draw Histogram, frequency polygon and frequency curve.
1.6. A doctors office staff studied the waiting times for patients who arrive at the office with a request for emergency service. The following data with waiting times in minutes were collected over a one-month period.
2 5 10 12 4 4 5 17 11 8 9 8 12 21 6 8 7 13 18 3
Use classes of 0–4, 5-9, and so on in the following:

a. Show the frequency distribution.
b. Show the relative frequency distribution.
c. Show the cumulative frequency distribution.
d. Show the cumulative relative frequency distribution.
e. What proportion of patients needing emergency service wait 9 minutes or less?
f. Draw Histogram, frequency polygon and frequency curve.

1.7 A shortage of candidates has required school districts to pay higher salaries and offer extras to attract and retain school district superintendents. The following data show the annual base salary ($1000s) for superintendents in 20 districts in the greater Rochester, New York, area (The Rochester Democrat and Chronicle, February 10, 2008).
187 184 174 185 175 172 202 197 165 208 215 164 162 172 182 156 172 175 170 183
Use classes of 150-159, 160-169, and so on in the following.
a. Show the frequency distribution.
b. Show the percent frequency distribution.
c. Show the cumulative percent frequency distribution.
d. Develop a histogram for the annual base salary.
e. Do the data appear to be skewed? Explain.
f. What percentage of the superintendents make more than $200,000?
g. Draw Histogram, frequency polygon and frequency curve.

1.8 NRF/BIG research provided results of a consumer holiday spending survey (USA Today, December 20, 2005). The following data provide the dollar amount of holiday spending for a sample of 25 consumers.
1200 850 740 590 340 450 890 260 610 350 1780 180 850 2050 770 800 1090 510 520 220 1450 280 1120 200 350
a. What is the lowest holiday spending? The highest?
b. Use a class width of $250 to prepare a frequency distribution and a percent frequency distribution for the data.
c. Prepare a histogram and comment on the shape of the distribution.
d. What observations can you make about holiday spending?
e. Draw Histogram, frequency polygon and frequency curve.

1.9 Sorting through unsolicited e-mail and spam affects the productivity of office workers. An Insight Express survey monitored office workers to determine the unproductive time per day devoted to unsolicited e-mail and spam (USA Today, November 13, 2003). The following data show a sample of time in minutes devoted to this task.
2 4 8 4 8 1 2 32 12 1 5 7 5 5 3 4 24 19 4 14
Summarize the data by constructing the following:
a. A frequency distribution .
b. A relative frequency distribution
c. A cumulative frequency distribution
d. A cumulative relative frequency distribution
e. What percentage of office workers spend 5 minutes or less on unsolicited e-mail and spam? What percentage of office workers spend more than 10 minutes a day on this task?
f. Histogram, frequency polygon and frequency curve.

1.10 The Nielsen Home Technology Report provided information about home technology and its usage. The following data are the hours of personal computer usage during one week

for a sample of 50 persons.
4.1 1.5 10.4 5.9 3.4 5.7 1.6 6.1 3.0 3.7 3.1 4.8 2.0 14.8 5.4 4.2 3.9 4.1 11.1 3.5 4.1 4.1 8.8 5.6 4.3 3.3 7.1 10.3 6.2 7.6 10.8 2.8 9.5 12.9 12.1 0.7 4.0 9.2 4.4 5.7 7.2 6.1 5.7 5.9 4.7 3.9 3.7 3.1 6.1 3.1
Summarize the data by constructing the following:
a. A frequency distribution (use a class width of three hours)
b. A relative frequency distribution
c. Comment on what the data indicate about personal computer usage at home.
d. Draw Histogram, frequency polygon and frequency curve.
1.11 According to the recent U.S. Federal Highway Administration Highway Statistics, the per- percentages of freeways and expressways in various road mileagerelated highway pavement conditions are as follows:
Poor 10%, Mediocre 32%, Fair 22%, Good 21%, and Very good 15%.
(a) Construct a bar graph.
(b) Construct a pie chart.
1.12 The following data give the letter grades of 20 students enrolled in a statistics course.
A B F A C C D A B F C D B A B A F B C A
(a) Construct a bar graph. (b) Construct a pie chart.
1.13 In a fiscal year, a city collected \$32.3 million in revenues. City spending for that year is expected to be nearly the same, with no tax increase projected.
Expenditure: Reserves 0.7%, capital outlay 29.7%, operating expenses 28.9%, debt service 3.2%, transfers 5.1%, personal services 32.4%.
Revenues: Property taxes 10.2%, utility and franchise taxes 11.3%, licenses and permits 1%, inter governmental revenue 10.1%, charges for services 28.2%, fines and forfeits 0.5%, interest and miscellaneous 2.7%, transfers and cash carryovers 36%.
(a) Construct bar graphs for expenditure and revenues and interpret.
(b) Construct pie charts for expenditure and revenues and interpret

1.14 The data shown are the number of grams per serving of 30 selected brands of cakes. Construct a frequency distribution using 5 class
. 32 47 51 41 46 30 46 38 34 34 52 48 48 38 43 41 21 24 25 29 33 45 51 32 32 27 23 23 34 35

1.15. Weights of the NBAs Top 50 Players Listed are the weights of the NBAs top 50 players. Construct a grouped frequency distribution and a cumulative frequency distribution . Analyze the results in terms of peaks, extreme values, etc.
240 210 220 260 250 195 230 270 325 225 165 295 205 230 250 210 220 210 230 202 250 265 230 210 240 245 225 180 175 215 215 235 245 250 215 210 195 240 240 225 260 210 190 260 230 190 210 230 185 260
Source: www.msn.foxsports.com
1.16 The number of stories in each of the worlds 30 tallest buildings follows. Construct a grouped frequency distribution and a cumulative frequency distribution with 7 classes
88 88 110 88 80 69 102 78 70 55 79 85 80 100 60 90 77 55 75 55 54 60 75 64 105 56 71 70 65 72
Source: New York Times Almanac.
1.17 The average quantitative GRE scores for the top 30 graduate schools of engineering are listed. Construct a grouped frequency distribution and a cumulative frequency distribution . 767 770 761 760 771 768 776 771 756 770 763 760 747 766 754 771 771 778 766 762 780 750 746 764 769 759 757 753 758 746
Source: U.S. News & World Report, Best Graduate Schools.
1.18. The number of passengers (in thousands) for the leading U.S. passenger airlines in 2004 is indicated below. Use the data to construct a grouped frequency distribution and a cumulative frequency distribution with a reasonable number of classes, and comment on the shape of the distribution.
91,570 86,755 81,066 70,786 55,373 42,400 40,551 21,119 16,280 14,869 13,659 13,417 13,170 12,632 11,731 10,420 10,024 9,122 7,041 6,954 6,406 6,362 5,930 5,585 5,427
Source: The World Almanac and Book of Facts.
1.19 The ages of the signers of the Declaration of Independence are shown. (Age is approximate since only the birth year appeared in the source, and one has been omitted since his birth year is unknown.) Construct a grouped frequency distribution and a cumulative frequency distribution
41 54 47 40 39 35 50 37 49 42 70 32 44 52 39 50 40 30 34 69 39 45 33 42 44 63 60 27 42 34 50 42 52 38 36 45 35 43 48 46 31 27 55 63 46 33 60 62 35 46 45 34 53 50 50
Source: The Universal Almanac.
1.20. The number of total vetoes exercised by the past 20 Presidents is listed below. Use the data to construct a grouped frequency distribution and a cumulative frequency distribution .

44 39 37 21 31 170 44 635 30 78 42 6 250 43 10 82 50 181 66 37
1.20 The data are the salaries (in hundred thousands of dollars) of a sample of 30 colleges and university coaches in the United States. Construct a frequency distribution for the data
164 225 225 140 188 210 238 146 201 544 550 188 415 261 164 478 684 330 307 435 857 183 381 275 578 450 385 297 390 515

1.21 The data show the NFL team payrolls (in millions of dollars) for a specific year. Construct a frequency distribution for the payroll
99 105 106 102 102 93 109 106 77 91 103 118 97 100 107 103 94 109 100 98 84 92 98 110 94 104 98 123 102 99 100 107
Source: NFL.
1.22 The state gas tax in cents per gallon for 25 states is given below. Construct a grouped frequency distribution and a cumulative frequency distribution .
7.5 16 23.5 17 22 21.5 19 20 27.1 20 22 20.7 17 28 20 23 18.5 25.3 24 31 14.5 25.9 18 30 31.5
1.23 The age at inauguration for each U.S. President is shown. Construct a stem and leaf plot and analyze the data.
57 54 52 55 51 56 47 61 68 56 55 54 61 51 57 51 46 54 51 52 57 49 54 42 60 69 58 64 49 51 62 64 57 48 51 56 43 46 61 65 47 55 55 54
Source: New York Times Almanac.
1.24 A listing of calories per one ounce of selected salad dressings (not fat-free) is given below. Construct a stem and leaf plot for the data.
100 130 130 130 110 110 120 130 140 100 140 170 160 130 160 120 150 100 145 145 145 115 120 100 120 160 140 120 180 100 160 120 140 150 190 150 180 160
1.25 The following data are based on a survey from American Travel Survey on why people travel. Construct a pie graph for the data and analyze the results.

Purpose	Number
Personal business	146
Visit friends or relatives	330
Work-related	225
Leisure	299

Source: USA TODAY.

1.26 The popular vehicle car colors are shown. Construct a pie graph for the data.

White	19
Silver	18
Black	16
Red	13
Blue	12
Gray	12
Other	10

Source: Dupont Automotive Color Popularity Report.
1.27 In a recent survey, 3 in 10 people indicated that they are likely to leave their jobs when the economy improves. Of those surveyed, 34% indicated that they would make a career change, 29% want a new job in the same industry, 21% are going to start a business, and

16% are going to retire. Make a pie chart and a Pareto chart for the data. Which chart do you think better represents the data? Source: National Survey Institute

1.28 The Brunswick Research Organization surveyed 50 randomly selected individuals and asked them the primary way they received the daily news. Their choices were via newspaper (N), television (T), radio (R), or Internet (I). Construct a categorical frequency distribution , Pie chart and bar chart for the data.

N N T T T I R R I T I N R R I N N I T N I R T T T T N R R I R R I N T R T I I T T I N T T I R N R T

1.29 A sporting goods store kept a record of sales of five items for one randomly selected hour during a recent sale. Construct a frequency distribution and draw pie chart for the data (B = baseballs, G =golf balls, T = tennis balls, S = soccer balls, F = footballs).

F B B B G T F G G F S G T F T T T S T F S S G S B

1.30 7. The percentage (rounded to the nearest whole percent) of persons from each state completing 4 years or more of college is listed below.

23 25 24 34 22 24 27 37 33 24 26 23 38 24 24 17 28 23 30 25 30 22 33 24 28 36 24 19 25 31 34 31 27 24 29 28 21 25 26 15 26 22 27 21 25 28 24 21 25 26

Source: New York Times Almanac

. Construct frequency distribution, cumulative, relative, percentage frequency distribution. frequency polygon and frequency curve.

Chapter 3

Measure of Central Tendency

We can easily present things as we wish them, to be..." (Aesop)
If at first you do not succeed, you are just about average. (Bill Cosby)

3.1 Objectives

1. Summarize the data
2. Comparison between two or more data sets
3. Provide base for calculating other statistical measures

3.2 Measures of Central Tendency

1. Arithematic Mean
2. Geometric Mean
3. Harmonic Mean
4. Median
5. Mode

3.3 Criteria or Properties of Good Measures of Central Tendency

1. Easily calculated and understandable.

2. Based on all values
3. Does not effected by extreme values
4. Rigidly define
5. It can be used for further statistical analysis
6. Sampling Stability

3.4 Arithematic Mean

Arithematic mean is a value which is obtained by dividing the sum of all values by their number.
Let $x_1, x_2, ..., x_N$ be the value of population of size N. The population mean denoted by μ defined as

$$\mu = \frac{x_1 + x_2 + ... + x_N}{N}$$

Let $x_1, x_2, ..., x_n$ be the value of sample of size n. The population mean denoted by $\bar{x}$ defined as

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}$$

Example
The following data represents the marks of 6 students selected at random from class of 40 students in the subject“ Statistics and probability” out of 100 marks.
54, 60, 70, 90, 60, 60,
Find the arithematic mean.
Solution

$$\bar{x} = \frac{54 + 60 + 70 + 90 + 60 + 60}{6} = \frac{394}{6} = 65.67$$

3.5 Geometric Mean

Geometric mean is a value which is obtained by N^{th} under-root of product of all N values
Or
Geometric mean is the antilog of arithematic of log of the values.
Let $x_1, x_2, ..., x_N$ be the value of population of size N. The Geometric mean denoted by $G.M$ defined as

$$G.M\left(x_1.x_2...x_N\right)^{\frac{1}{N}}$$

or

$$G.M = Antilog\left(\frac{logx_1 + logx_2 + ... + logx_N}{N}\right)$$

Example
The following data represents the marks of 6 students selected at random from class of 40 students in the subject“ Statistics and probability” out of 100 marks.
54, 60, 70, 90, 60, 60,
Find the Geometric mean.
Solution

$$G.M = (54.60.70.90.60.60)^{\frac{1}{6}} = ...$$

Or

$$G.M = \frac{log54 + log60 + log70 + log90 + log60 + log60}{6} = ...$$

Example
The return on investment earned by Atkins Construction Company for four successive years was: 30 percent, 20 percent, -40 percent, and 200 percent. What is the geometric mean rate of return on investment?

3.6 Harmonic Mean

Harmonic mean is the reciprocal of arithematic mean of reciprocal of values. Let $x_1, x_2, ..., x_N$ be the value of population of size N. The Harmonic mean denoted by $H.M$ defined as

$$H.M = \frac{N}{\frac{1}{x_1} + \frac{1}{x_2} + ... + \frac{1}{x_N}}$$

Example
The following data represents the marks of 6 students selected at random from class of 40 students in the subject“ Statistics and probability” out of 100 marks.
54, 60, 70, 90, 60, 60, Find the Harmonic mean.
Solution

$$H.M = \frac{6}{\frac{1}{54} + \frac{1}{60} + \frac{1}{70} + \frac{1}{90} + \frac{1}{60} + \frac{1}{60}} = ...$$

3.7 Median

Median is a value which divide an array data into two equal parts. Median is a position measure and equal number of values lie above and below the median value. Let $x_1, x_2, ..., x_N$ be the value of population of size N. Median is denoted by $\hat{X}$

$$\hat{X} = \left(\frac{N+1}{2}\right) th\ value, if N\ is\ odd$$

$$\hat{X} = \frac{1}{2}\left[\left(\frac{N}{2}\right) th\ value + \left(\frac{N}{2}+1\right) th\ value\right]\ , if\ N\ is\ even$$

Example

The following data represents the marks of 6 students selected at random from class of 40 students in the subject“ Statistics and probability” out of 100 marks. 54, 60, 70, 90, 65, 60, Find the Median Marks.

Solution

Array data: 54, 60 , 60, 65, 70, 90

$$\hat{X} = \frac{1}{2}\left[\left(\frac{6}{2}\right) th\ value + \left(\frac{6}{2}+1\right) th\ value\right]$$

$$\hat{X} = \frac{1}{2}\left[3rd\ value + 4th\ value\right] = \frac{1}{2}\left[60+65\right] = 62.5$$

3.8 Mode

Mode is the value(s) which occur most frequent number of time in data set It is denoted by$\hat{X}$. If any value does not repeat more then once. Then we say, Mode does not exit. If more then one values repeat equal number of times. Then these values will be mode

Example

The following data represents the marks of 6 students selected at random from class of 40 students in the subject“ Statistics and probability” out of 100 marks. 54, 60, 70, 90, 65, 60, Find the Mode.

Solution

Mode=$\hat{X} = 60$

Example

Erika operates a website devoted to providing information and support for persons who are interested in organic gardening. According to the hit counter that records daily visitors to her site, the numbers of visits during the past 8 days have been as follows:
70, 37, 55, 63, 59, 68, 56, 54.
Determine the mean, Geometric mean, Harmonic mean, Mode and Median

Find the arithematic mean, Geometric mean, Harmonic mean, Mode and Median using data given in question number 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9, 1.10, 1.11, 1.12

3.9 Properties of Mean

1. Weighted mean
2. Incorrect values heve been used for calculation of mean
3. Mathematical properties of arithematic mean
4. For any set of information. $A.M \geq G.M \geq H.M$
5. If a data set take the values $a, ar, ar^2, ..., ar^{n-1}$ then

$$(G.M)^2 = A.M * H.M$$

Exercise

Exercise 3.1
A social scientist for a childrens advocacy organization has randomly selected 10 Saturday-morning television cartoon shows and carried out a content analysis in which he counts the number of incidents of verbal or physical violence in each. For the 10 cartoons examined, the counts were as follows:
27, 12, 16, 22, 15, 30, 14, 30, 11, 21.
Find the mean, Geometric mean, Harmonic mean, Mode and Median
Exercise 3.2
Wageweb conducts surveys of salary data and presents summaries on its Web site. Salaries reported for benefits managers ranged from \$50,935 to \$79,577 (http://www.Wageweb.com, April 12, 2000). Assume the following data are a sample of the annual salaries for 00 benefits managers. Data are in thousands of dollars.
57.7 64.4 62.1 59.1 71.1 62.1 64.4 61.2 66.8 61.8
Find the arithematic mean, Geometric mean, Harmonic mean, Mode and Median
Exercise 3.3
There are 10 salespeople employed by Moody Insurance Agency in Venice, Florida. The numbers of new life insurance policies sold last month by the respective salespeople were: 15,23,4,19,18,10,10,8,28,19. Find the arithematic mean, Geometric mean, Harmonic mean
Exercise 3.4
The accounting department at a mail-order company counted the following numbers of incoming calls per day to the company's toll-free number during the first 7 days in May 2003: 14,24,19,31,36,26,17. Find the arithematic mean, Geometric mean, Harmonic mean
Exercise 3.5
The percent increase in sales for the last 4 years at Combs Cosmetics were: 4.91, 5.75, 8.12, and 21.60. (a) Find the geometric mean percent increase.
(b) Find the arithmetic mean percent increase.
(c) Is the arithmetic mean equal to or greater than the geometric mean?
Exercise 3.6

The cost of consumer purchases such as single-family housing, gasoline, Internet services, tax preparation, and hospitalization were provided in The Wall-Street Journal (January 2, 2007). Sample data typical of the cost of tax-return preparation by services such as H&R Block are shown below. 120 230 110 115 160 130 150 105 195 155 105 360 120 120 140 100 115 180 235 255 Compute the mean, median, and mode.

Exercise 3.7

In automobile mileage and gasoline-consumption testing, 13 automobiles were road tested for 300 miles in both city and highway driving conditions. The following data were recorded for miles-per-gallon performance.

City: 16.2 16.7 15.9 14.4 13.2 15.3 16.8 16.0 16.1 15.3 15.2 15.3 16.2

Highway: 19.4 20.6 18.3 18.6 19.2 17.4 17.2 18.6 19.0 21.1 19.4 18.5 18.7

Use the mean, median, and mode to make a statement about the difference in performance for city and highway driving.