INFERENTIAL STATISTICS AND DATA ANALYSIS

CHAPTER 5

Statistics is the science of collecting, analyzing, and interpreting data to extract meaningful insights and inform decision-making. At the heart of statistical analysis lies sampling and data analysis, essential components of statistical research. **Sampling** involves selecting a subset of individuals, objects, or observations from a larger population to estimate population characteristics. Effective sampling methods ensure that the sample accurately represents the population, enabling reliable conclusions. **Data analysis** involves processing, transforming, and examining data to uncover patterns, trends, and relationships. Statistical techniques, such as descriptive statistics, inferential statistics, and graphical methods, are used to extract insights from data.

In this chapter we will learn about;

- Introduction to statistics, probabilistic models
- Data and distribution of data, tabulation, Graph, stem and leaf diagram, charts
- Misleading graphs, shapes of distribution, bivariate analysis
- Measuring relationships via regression analysis and correlation
- Correlation and scatter plot, Variability and measure of dispersion
- Data tendencies via measure of location and spread of data
- Box and whisker plot
- Sampling techniques, estimation, hypothesis and hypothesis testing
- Margin of error and confidence interval
- Exercises about inferential statistics and data analysis

Introduction to Statistics

The word “Statistics” is derived from the Latin word **Status** or the Italian word **Statista** or the German word **Statistik** or the French word **Statistique** meaning “a political state” or “the state – man’s art”. It is the discipline that includes procedures or techniques used to collect process, analyzes numerical data to inferences and to reach decision in the face of uncertainty. It is the science of collection, presentation, analysis and interpretation of numerical data.

According to this definition, there are four stages.

- Collection of data
- Presentation of data
- Analysis of data
- Interpretation

Use/Importance of Statistical Information

- To inform general public
- To explain things that have happened
- To justify a claim
- To provide general comparison
- To estimate the unknown quantities

Limitations of Statistical Information

- Deals with aggregates and not with individuals.
- Deals with numerically specified characteristics.

Types of Statistics

- **Descriptive/Deductive Statistics:** Branch which deals with concepts and methods concerned with summarization and description of the important aspects of numerical data. In it no conclusion is drawn about the population.
- **Inferential/Inductive Statistics:** Branch which deals with drawing inferences about the population on the basis of sample information. Its two most types are; Estimation and Testing of Hypothesis.

Introduction to Probabilistic Models

Probabilistic models are mathematical frameworks used to represent and analyze uncertain relationships between variables, providing a powerful tool for understanding complex phenomena in various fields, including engineering, economics, biology, and social sciences. These models quantify uncertainty using probability theory, enabling predictions, decision-making, and risk assessment under uncertainty. By capturing stochastic relationships, probabilistic models facilitate:

1. Uncertainty quantification
2. Predictive modeling
3. Decision-making under uncertainty
4. Risk analysis

Applications

1. Machine learning (deep learning, natural language processing)
2. Signal processing and communication
3. Finance (risk management, portfolio optimization)
4. Reliability engineering and maintenance
5. Biostatistics and epidemiology

Benefits:

Handling uncertainty and ambiguity**,** Incorporating prior knowledge**,** Flexibility and adaptability**,** Interpretable results**,** Scalability.

Deterministic Models/Relations: If there is an extract relationship between dependent and independent variables, and there is no chance of error, the regression line developed for such variables is called deterministic models. The following equation is an example of deterministic model; Y = a + bX

Probabilistic Models/Relations: If there is no extract relationship between dependent and independent variables, and there is a chance of error, the regression line developed for such variables is called Probabilistic models. The following equation is an example of Probabilistic model; Y = a + bX + error

Statistical Modeling: Statistical modeling is an elaborate method of generating sample data and making real world predictions using numerous statistical models and explicit assumptions. It helps data scientists visualize the relationship between random variables and strategically interpret dataset.

Types of Probabilistic Models

There are several statistical models, each designed to solve a specific research issue or data format. Here are a few common types of statistical models and their applications.

- **Linear Regression Models:** These models are used to represent the connection between a continuous result variable and one or more predictor variables. For example, depending on a person's height, age, and gender, a linear regression model may be used to estimate their weight.
- **Logistic Regression Models:** Logistic regression models are used to represent the connection between a binary outcome variable (for example, yes/no) and one or more predictor variables. For example, depending on age, blood pressure and cholesterol levels, a regression logistic model may be used to predict if a patient would have a heart attack.
- **Time Series Models:** Time series models are used to model data that change over time, such as stock prices, weather trends, or monthly sales numbers. These type of models may be applied to data to find trends, sessional patterns and other forms of temporal correlations.
- **Multilevel Models:** These models are used to model data having a hierarchical structure, such as pupil in school are patients in hospitals. Multilevel models can be used to investigate how individual – level and group – level factors impact outcomes, as well as to account for the fact that people in the same group maybe more similar to each other than those in different groups.
- **Structural Equation Models:** These types of models are used to represent complicated interactions between several variables. Structural equation model can be used to evaluate ideas regarding casual links between variables and to quantify their strength and direction.
- **Clustering Models:** Clustering models are used to bring together comparable observations based on their similarities in terms of features. Clustering algorithm can be used to uncover patterns in data that would be difficult to detect using other approaches.

Distributions

Distribution is a fundamental concept in statistics that refers to the way that data or observations are dispersed or spread out. It describes the shape, central tendency, and variability of a dataset, providing valuable insights into the underlying patterns and relationships. Understanding distribution is crucial in statistics, as it enables researchers and analysts to summarize and describe data, make informed decisions, and draw meaningful conclusions.

- A listing of all classes of the data and their frequencies is called a **Frequency distribution**. It is a tabular arrangement for classifying data into different groups and the number of observations falling in each group corresponds to the respective group. On the basis of type of variables, it has two types;
  - Discrete frequency distribution
  - Continuous frequency distribution
- The data presented in the form of frequency distribution is called **Grouped Data**.
- A listing of all classes and their relative frequencies is called a **Relative Frequency distribution**. Most distributions show frequencies as well as relative frequencies.

1. Discrete Frequency Table by using a Tally Column: 20 coins are tossed 5 times and the number of heads recorded at each toss are given below;
3,4,2,3,3,5,2,2,2,1,1,2,1,4,2,2,3,3,4,2.
Make frequency distribution of number of heads observed.

Solution: Let X = number of heads. The frequency distribution is given below;

X	Tally Marks	frequency (f)
1	|||	3
2	𝍸 |||	8
3	𝍸	5
4	|||	3
5	|	1

2. Continuous Frequency Table by listing Actual Values: For data given below;
51,55,32,41,22,30,35,53,30,60,59,15,7,18,40,49,40,25,14,18,19,2,43,22,39,26,34, 19,10,17, 47,38,13,30,34,54,10,21,51,52.
Make frequency distribution with a class interval of size 10.

Solution:

Class/Groups	Observations	frequency *(f)*
0 – 9	2,7	2
10 – 19	10,10,13,14,15,17,18,18,19,19	10
20 – 29	21,22,22,25,26	5
30 – 39	30,30,30,32,34,34,35,38,39	9
40 – 49	40,40,41,43,47,49	6
50 – 59	51,51,52,53,54,55,59	7
60 – 69	60	1

3. Continuous Frequency Table: Bradley worked a summer job to earn money for college. His weekly hours over a 12 week period were 25, 32, 36, 32, 18, 28, 30, 36, 12, 16, 35, 36. Find Distribution table.

Solution:

The Distribution would be as follows:

Hours	Frequencey	Relative Frequency
10-19	3	$\frac{3}{12} = 0.25 = 25\%$
20-29	3	$\frac{3}{12} = 0.167 = 17\%$
30-39	7	$\frac{7}{12} = 0.583 = 58\%$
Total	12	100%

Remember:

- **Class Limits:** The minimum and the maximum values defined for a class or group are called Class Limits. The minimum value is called the **lower class limit** and maximum value is called the **upper class limit** of the class.
- **Class Boundaries:** The real class limits of a class are called **class boundaries**. A class boundary is obtained by adding two successive class limits and dividing the sum by 2. The value so obtained is taken as **upper class boundary** for the previous class and **lower class boundary** for the next class.
- **Mid – Point/ Class Mark:** For a given class the average of that class obtained by dividing the sum of upper and lower class by 2, is called the mid – point of class mark of that class.
- **Interval/ Class Width:** Difference between the class boundaries.
- **Cumulative Frequency:** The total of frequency up to an upper class limit or boundary is called the cumulative frequency.

Classes	Frequency *(f)*	Class Boundaries	Mid Point	Cumulative Frequency
10 – 14	5	9.5 – 14.5	12	5
15 – 19	12	14.5 – 19.5	17	5 + 12 = 17
20 – 24	30	19.5 – 24.5	22	17 + 30 = 47
25 – 29	25	24.5 – 29.5	27	47 + 25 = 72
30 – 34	6	29.5 – 34.5	32	72 + 6 = 78

Some Facts about Data

- **Observation:** It is a fact or figure; we collect about a given variable. It can be expressed as a number or as a quality.
- **Data:** The collection of raw fact and figures is called data.
- **Data Set:** The collection of observations on one or more variables.

- **Cross Section Data:** Data collected on different elements at the same point in time or for the same period of time are called cross section data.
- **Time Series Data:** Data collected on the same element for the same variable at different points in time or for different periods of time are called time series data.
- **Discrete Data:** A data which is generated by a discrete variable is called discrete data.
- **Continuous Data:** A data which is generated by a continuous variable is called continuous data.
- **Datum:** A single numerical fact is datum.
- There are two types of data: Primary data and Secondary data.
- **Primary Data:** The data that have been initially collected and have not undergone any statistical treatment are called primary data.
- **Source of Primary Data:**
  Direct personal investigation
  Indirect investigation or interviews
  Collection through questionnaires
  Collection through local sources
  Through internet
  Experimental research
- **Secondary Data:** The data which has undergone any statistical treatment at least once is called primary data.
- **Source of Secondary Data:**
- **Official:** Using the publications of statistical divisions, ministry of finance, the federal and provincial bureau of statistics, ministries of food, agriculture and industry etc.
- **Semi Official:** State Bank of Pakistan, Railway Board, Central Cotton Committee, Board of Economic Inquiry, District Councils, Municipalities etc.
- Publications of Trade Association, Chamber of Commerce etc.
- Technical and Trade Journals and newspapers.
- Research organizations such as universities and other institutions.

Presentation of Data

The device of gathering data often results in a massive volume of statistical data which are in the form of individual measurement of counts. These are as follows;

- **Classification:** The process of dividing a set of observations or objects into classes or groups. It is the sorting of data into homogeneous classes or groups according to their being alike or not.
- **Tabulation:** A systematic presentation of data classified under suitable heads and subheads and placed in columns and rows. It is an orderly arrangement of data in columns and rows.

- **Graphical Display:** The visual display of statistical data in the form of point lines, areas and other geometrical forms and symbols is in the most general term called graphical display. Such graphical representation divided into **graphs** and **diagrams**.

Data Handling

Data handling is the process of securing the research data is gathered, archived or disposed of in a protected and safe way during and after the completion of the analysis process. Data handling means collecting the set of data and presenting in a different form.

Data Handling Steps

The steps involved in the data handling process are as follows:

- Problem Identification
- Data Collection
- Data Presentation
- Graphical Representation
- Data Analysis
- Conclusion

  From the analysis of the data, we can derive the solution to our problem statement. The data can be usually represented in any one of the following ways. They are: Bar Graph, Line Graphs, Histograms, Stem and Leaf Plot, Dot Plots, Frequency Distribution, Cumulative Tables and Graphs

Tabulation

An orderly arrangement of data in columns and rows

Graphs of Data

As the old saying goes, a picture is worth a thousand words. Data summaries can come in pictures or graphs. Here are some of the typical types of graphs to display distributions. They can give us a quick overview of the big picture and the characteristics of the data.

Histogram / Frequency Histogram

A **histogram** is a bar graph where the data is represented in equal intervals. A Frequency Histogram is a graph that displays the classes on the horizontal axis and the frequencies on the vertical axis. It consists of vertical bars, whose height is equal to the frequency of the class(interval). The bars are drawn next to each other (without gaps), since they encompass the range of the data in numerical order. The left side of each bar start sat the lower limit of the class interval. The right side goes up to the lower limit of the next interval. A Histogram is only for quantitative data, not qualitative.

Relative Frequency Histogram

A Relative Frequency Histogram is the same as a frequency histogram, except it uses relative frequencies for the vertical axis and the bar heights.

Cumulative Frequency Polygon or Ogive

It is graphical representation of frequency distribution taking upper class boundary along with x – axis and cumulative frequency along y – axis.

Frequency Polygon

It is graphical representation of frequency distribution taking mid point along x – axis and frequency along y – axis.

Relative Frequency

It is a frequency which is derived by dividing frequency of any particular class by total frequency.

Historiogram

A Historiogram is a graph showing changes in the values of one period of time to the next is known as graph of a time series or histogram. Graph of frequency distribution is called **Histogram** and the graph of time series is called **Historiogram**.

Relative Frequency versus Grouped Frequency

One advantage of using a relative frequency distribution instead of a grouped frequency distribution is that there is a direct correspondence between the percent values of the relative frequency distribution and probabilities. For instance, in the relative frequency distribution in Table, the percent of the data that lie between 35 and 40 seconds is 14.9%. Thus, if a subscriber is chosen at random, the probability that the subscriber will require at least 35 seconds but less than 40 seconds to download the music file is 0.149.

4. An Internet service provider (ISP) has installed new computers. To estimate the new download times its subscribers will experience, the ISP surveyed 1000 of its subscribers to determine the time required for each subscriber to download a particular file from the Internet sitemusic.net. Summarized the survey by frequency distribution and histogram.

Solution: The results of that survey are summarized in Table.

Download Time (in seconds)	Number of Subscribers
0-5	6
5-10	17
10-15	43
15-20	92
20-25	151
25-30	192
30-35	190
35-40	149
40-45	90
45-50	45
50-55	15
55-60	10

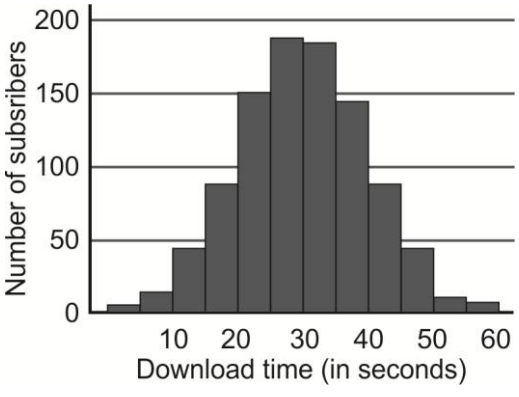


Table is called a **grouped frequency distribution**. It shows how often (frequently) certain events occurred.

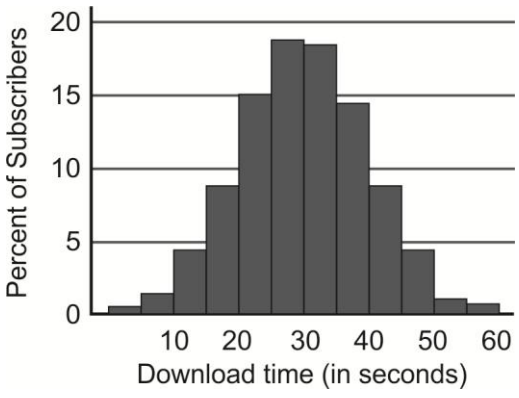
Each interval, 0–5, 5–10, and so on, is called a **class**. This distribution has 12 classes. For the 10–15 class, 10 is the **lower class boundary** and 15 is the **upper class boundary**. Any data value that lies on a common boundary is assigned to the higher class. The graph of a frequency distribution is called a **histogram**. A histogram provides a pictorial view of how the data are distributed. In Figure, the height of each bar of the histogram indicates how many subscribers experienced the download times shown by the class at the base of the bar.

5. An Internet service provider (ISP) has installed new computers. To estimate the new download times its subscribers will experience, the ISP surveyed 1000 of its subscribers to determine the time required for each subscriber to download a particular file from the Internet sitemusic.net. Summarized the survey by relative frequency distribution and relative frequency histogram.

Solution:

The results of that survey are summarized in Table.

Download Time (in seconds)	Number of Subscribers
0-5	0.6
5-10	1.7
10-15	4.3
15-20	9.2
20-25	15.1
25-30	19.2
30-35	19.0
35-40	14.9
40-45	9.0
45-50	4.5
50-55	1.5
55-60	1.0



Examine the distribution in Table. It shows the percent of subscribers that are in each class, as opposed to the frequency distribution in Table, which shows the number of customers in each class. The type of frequency distribution that lists the percent of data in each class is called a **relative frequency distribution**. The **relative frequency histogram** in Figure was drawn by using the data in the relative frequency distribution. It shows the percent of subscribers along its vertical axis.

6. Use the relative frequency distribution in Table to determine

a. The percent of subscribers who required at least 25 seconds to download the file.

b. The probability that a subscriber chosen at random will require at least 5 but less than 20 seconds to download the file.

Download Time (in seconds)	Percentage of Subscribers	
0-5	0.6	Sum is 15.2%
5-10	1.7	
10-15	4.3	
15-20	9.2	
20-25	15.1	Sum is 69.1%
25-30	19.2	
30-35	19.0	
35-40	14.9	
40-45	9.0	
45-50	4.5	
50-55	1.5	
55-60	1.0	

Solution:

a. The percent of data in all the classes with a lower boundary of 25 seconds or more is the sum of the percents for all of the classes highlighted in red in the distribution below. Thus the percent of subscribers who required at least 25 seconds to download the file is 69.1%. See Table.

b. The percent of data in all the classes with a lower boundary of at least 5 seconds and an upper boundary of 20 seconds or less is the sum of the percents in all of the classes highlighted in blue in the distribution above. Thus the percent of subscribers who required at least 5 but less than 20 seconds to download the file is 15.2%. The probability that a subscriber chosen at random will require at least5 but less than 20 seconds to download the file is 0.152. See Table.

7. Use the relative frequency distribution in Table to determine

a. the percent of subscribers who required less than 25 seconds to download the file.

b. the probability that a subscriber chosen at random will require at least 10 seconds but less than 30 seconds to download the file.

Download Time (in seconds)	Percentage of Subscribers
0-5	0.6
5-10	1.7
10-15	4.3
15-20	9.2
20-25	15.1
25-30	19.2
30-35	19.0
35-40	14.9
40-45	9.0
45-50	4.5
50-55	1.5
55-60	1.0

Solution:

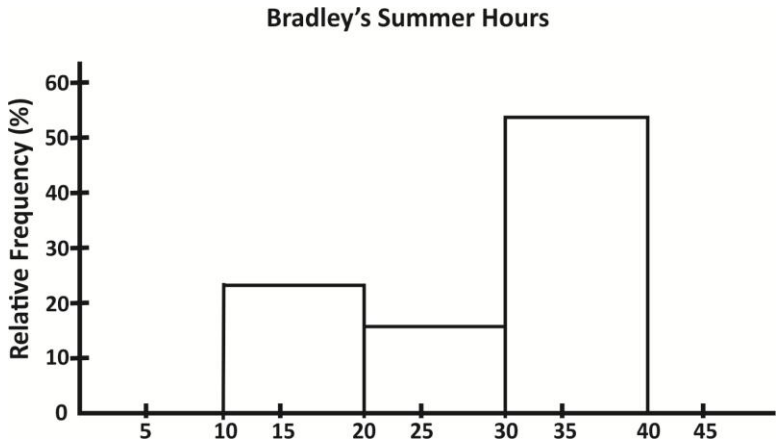
a. The percent of data in all classes with an upper bound of 25 seconds or less is the sum of the percents for the first five classes in Table. Thus the percent of subscribers who required less than 25 seconds to download the file is 30.9%.

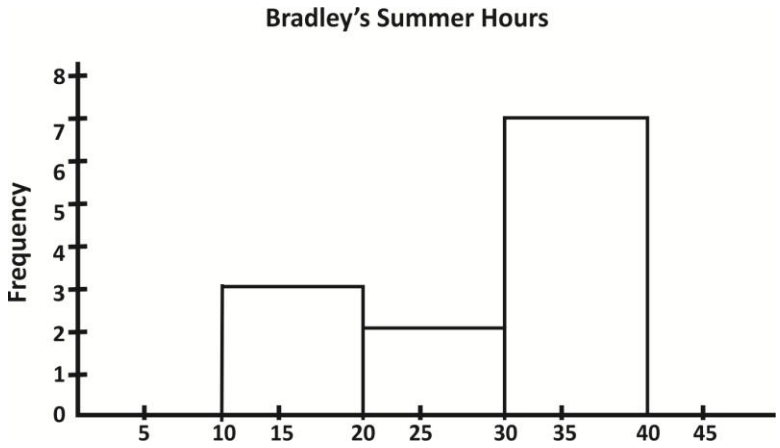
b. The percent of data in all the classes with a lower bound of at least 10 seconds and an upper bound of 30 seconds or less is the sum of the percents in the third through sixth classes in Table. Thus the percent of subscribers who required from10 to 30 seconds to download the file is 47.8%. The probability that a subscriber chosen at random will require from 10 to30 seconds to download the file is 0.478.

8. For Bradley's weekly hours at a summer job: 25, 32, 36, 32, 18, 28, 30, 36, 12, 16, 35, 36. Find the frequency and relative frequency histograms hours he worked in a week.

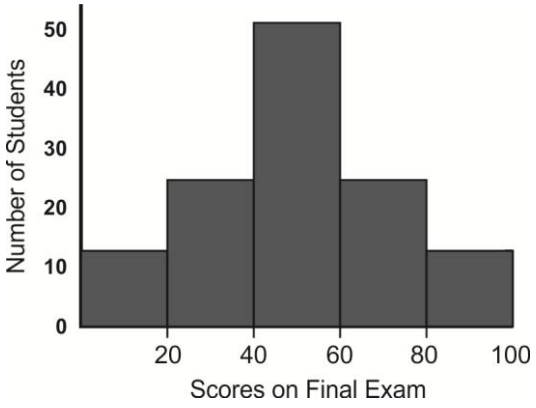
Solution:

The frequency and relative frequency histograms for Bradley's summer job data are shown below.



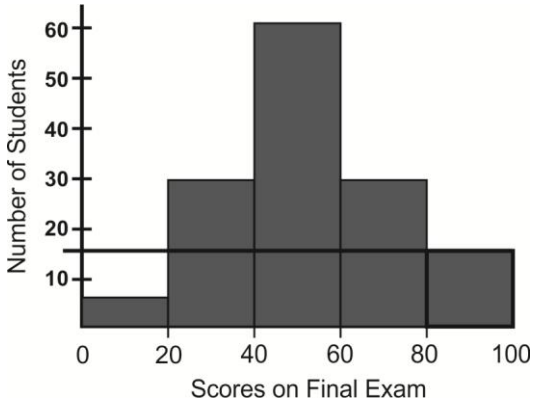


9. In the following graph, how many students got the highest score?



Solution:

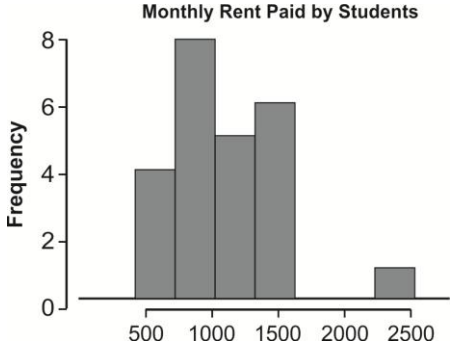
The highest score ranges from 80 to 100. Based on the histogram chart, there were about 15 students. Therefore, the answer is 15.



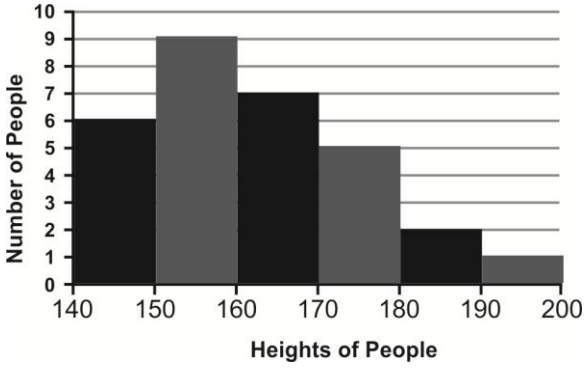
10. Draw a histogram for the distribution for the following data

1500	1350	350	1200	850	900
1500	1150	1500	900	1400	1100
1250	600	610	960	890	1325
900	800	2550	495	1200	690

Solution:



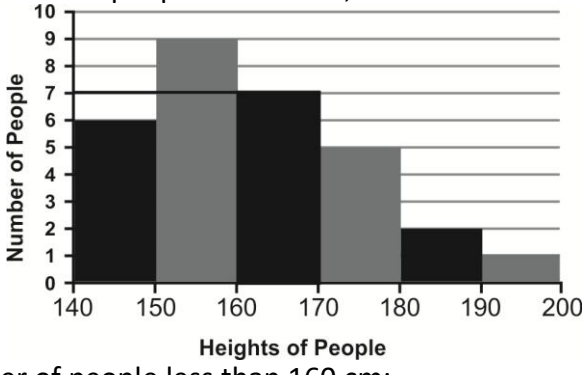
11. The histogram below shows the height (in cm) distribution of 30 people. How many people have heights between 160 and 170 cm? How many people have heights less than 160 cm?



Solution:

Solution for heights between 160 and 170 cm:

We need to look at the third bar because it ranges from 160 cm to 170 cm. The bar indicates that there are 7 people. Therefore, the answer is 7.



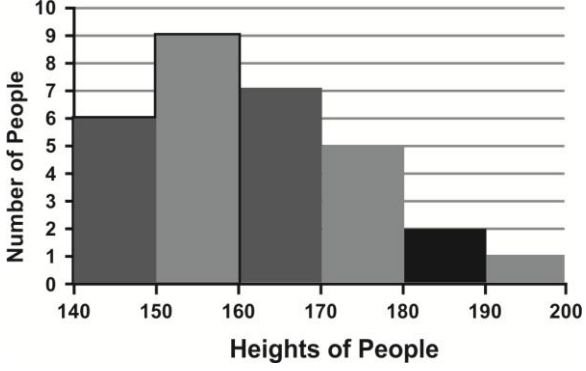
Solution of number of people less than 160 cm:

For the people less than 160 cm height, we have to look at the bars from two categories – 140 cm to 150 cm and 150 cm to 160 cm. Therefore,

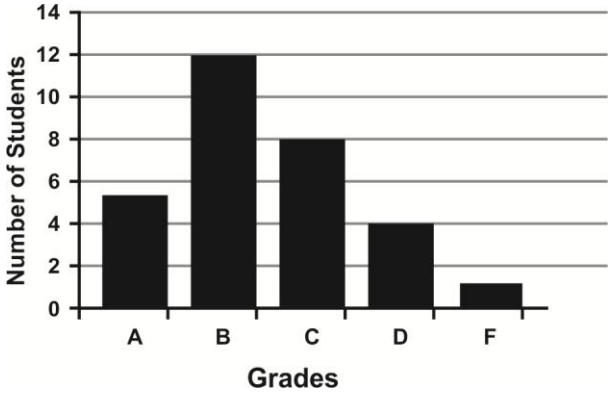
140-150: 6 people

150-160: 9 people

Total people less than 160 cm is 6 + 9 = 15 people

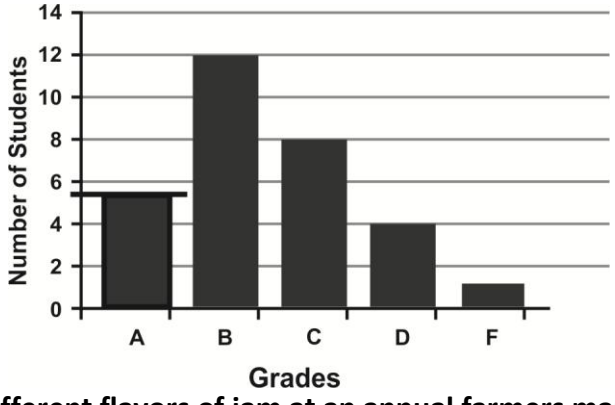


12. How many students got a grade of 'A' based on the following chart.

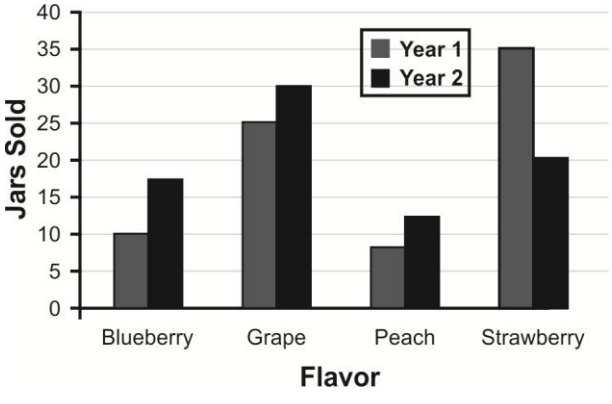


Solution:

When we observe the bar A, the bar lines up with 5. Therefore, the answer is 5 students.



13. Logan sells four different flavors of jam at an annual farmers market. The graph below shows the number of jars of each type of jam they sold at the market during the first two years. Which flavor of jam had the greatest increase in number of jars sold from Year 1 to Year 2?



Solution:

For this problem, we have to calculate the difference between year 1 and year 2 of all the jars.

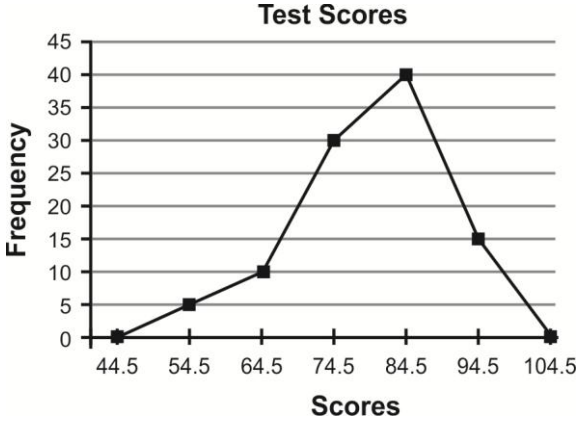
Jars	Year 1	Year 2	Difference(year2 - Year1)
Bluberry	10	18	8
Grape	25	30	5
Peach	8	12	4
Strawberry	35	20	-15

Although strawberry has the highest difference, it is not the answer because the number of jars decreased. Blueberry jars increased from 10 to 18 with an increase of 8 jars. Therefore, blueberry is the answer.

14. Construct a frequency polygon from the following frequency table.

Frequency Distribution of Calculus Final Test Scores			
Lower Bound	Upper Bound	Frequency	Cumulative Frequency
49.5	59.5	5	5
59.5	69.5	10	15
69.5	79.5	30	45
79.5	89.5	40	85
89.5	99.5	15	100

Solution:



Time Series Graphs/ Line Graphs

Line Graphs show trends over a period of time. Time is located on the horizontal axis and amounts will be located on the vertical axis. Information will be organized into ordered pairs. To draw the graph, you plot these points and then connect them with straight lines.

In line graph, the x-axis (horizontal axis) consists of data values and the y-axis (vertical axis) consists of frequency points. The frequency points are connected using line segments. The graph of time series is called **Historiogram**.

Difference between Line Chart and a Time-Series Graph

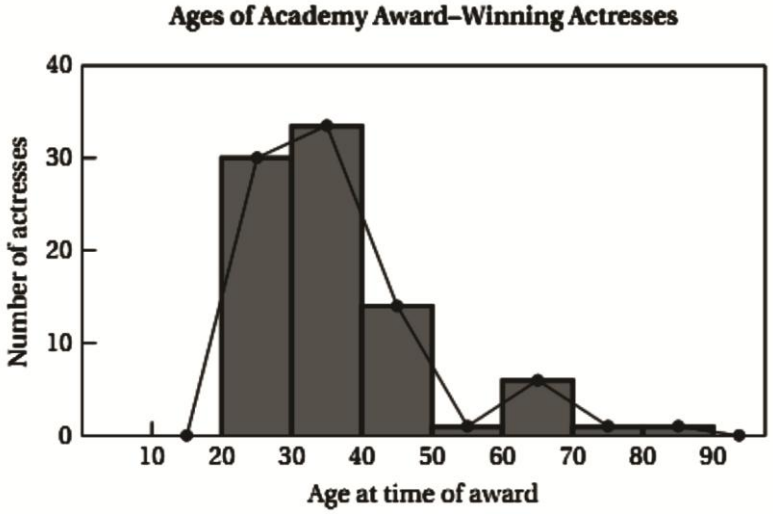
A **line chart** shows the data value for each category as a dot, and the dots are connected with lines. For each dot, the horizontal position is the center of the bin it represents and the vertical position is the data value for the bin.

While a **time-series graph** is a line chart or histogram in which the horizontal axis represents time.

15. Oscar-Winning Actresses: Table shows the ages (at the time when they won the award) of all Academy Award–winning actresses through 2013. Make a histogram and a line chart to display these data. Discuss the results.

Age	Number of actresses
20-29	30
30-39	34
40-49	13
50-59	1
60-69	6
70-79	1
80-89	1

Solution:

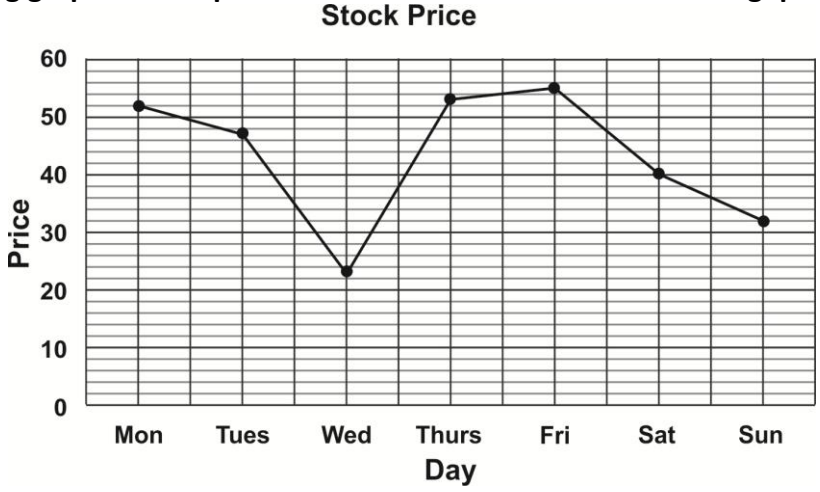


16. Time-Series graph: Figure shows a time-series graph of homicide rates in the United States. Briefly summarize what it shows.

Solution:

The graph shows how the homicide rate per 100,000 people has changed since 1960. We see that the homicide rate rose dramatically—more than doubling—from a minimum around 1962 to a first peak around 1974. It then remained high, with some variations, through about 1993. After 1993, it fell dramatically to the year 2000, then stayed nearly constant until a slight drop from 2008 through 2012. The decrease in the homicide rate during the 1990s has been attributed to tougher enforcement of drug laws and a crackdown on gangs.

17. Veronica is a stock trader. She followed the value of a stock and recorded the following graph for this past week. Use it to answer the following questions.



a) What was the value of the stock on Tuesday?

b) What day of the week was the stock price equal to $40?

c) Between which two consecutive days was the biggest change in stock prices?

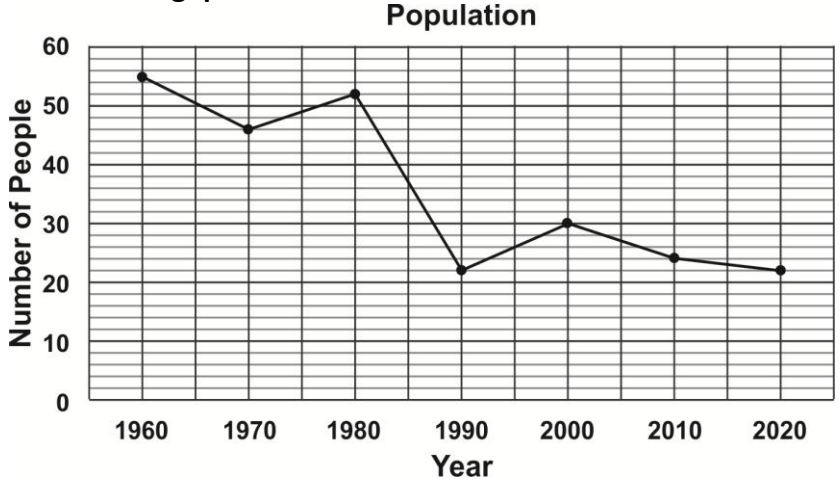
Solution:

a) When reading the graph the horizontal axis is the day and the vertical axis tells us the value of the stock that day. For Tuesday the value of the stock was $47.

b) Start by reading the vertical axis and find where the stock price is $40 and then find the corresponding day on the horizontal axis which is Saturday.

c) We need to look at the graph and find two consecutive points who have the biggest deference between stock prices. Looking at the graph this occurs between Wednesday and Thursday.

18. Edge Hill is the smallest incorporated city in the state of Georgia. The line graph below shows its population every ten years starting in 1960. Use this graph to answer the following questions.



a) What was the population in 2000?

b) What year was the population equal to 46 people?

c) At approximately what rate did population decrease from the year 1980 to 1990?

Solution:

a) When reading this graph the horizontal axis represents the year and the vertical axis represents the population size. To answer this problem look for the year 2000 on the horizontal axis and then read the population size from the vertical axis which is 30 people.

b) For this problem look for the point on the line graph where the population is 46 people. This happens for the year 1970.

c) To answer this question we need to compute a rate. Start by identifying two points from the graph at the years 1980 and 1990: (1980,53) and (1990, 22). Now compute the rate. To do this we will use the slope formula.

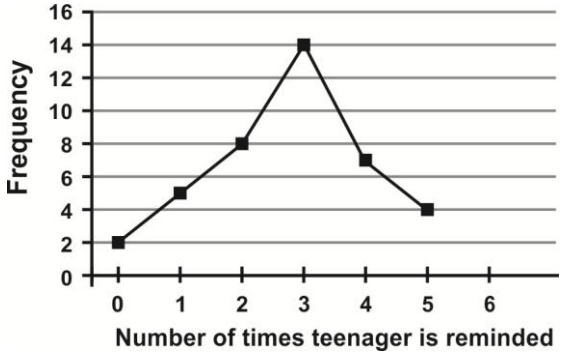
$$Rate = \frac{y_2 - y_1}{x_2 - x_1} = \frac{22 - 53}{1990 - 1980} = \frac{-31}{10} = -3.1$$

What this tells us is that on average the population decreased by approximately 3 peoples each year between 1980 and 1990.

19. In a survey, 40 mothers were asked how many times per week a teenager must be reminded to do his or her chores. The results are shown in Table and in Figure.

Numbers of times teenager is remindedd	Frequency
0	2
1	5
2	8
3	14
4	7
5	4

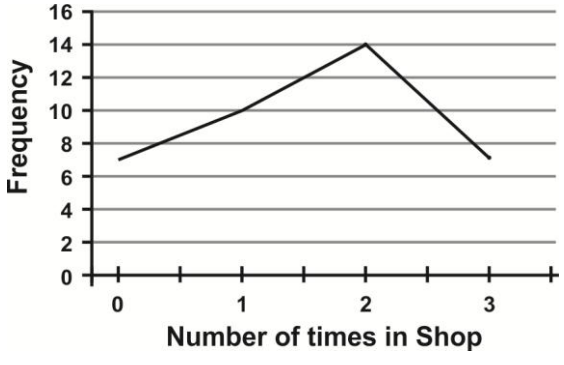
Solution:



20. In a survey, 40 people were asked how many times per year they had their car in the shop for repairs. The results are shown in Table. Construct a line graph.

Numbers of times in shop	Frequency
0	7
1	10
2	14
3	9

Solution:

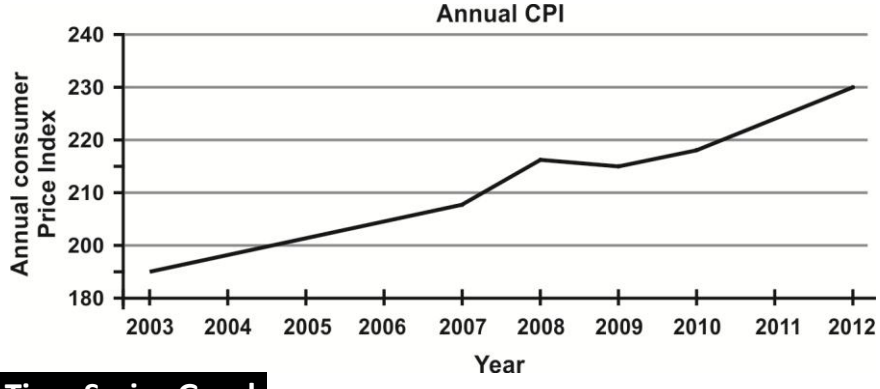


21. The following data shows the Annual Consumer Price Index, each month, for ten years. Construct a time series graph for the Annual Consumer Price Index data only.

Year	Jan	Feb	Mar	Apr	May	Jun	Jul
2003	181.7	183.1	184.2	183.8	183.5	183.7	183.9
2004	185.2	186.2	187.4	188.0	189.1	189.7	189.4
2005	190.7	191.8	193.3	194.6	194.4	194.5	195.4
2006	198.3	198.7	199.8	201.5	202.5	202.9	203.5
2007	202.416	203.499	205.352	206.686	207.949	208.352	208.299
2008	211.080	211.693	213.528	214.823	216.632	218.815	219.964
2009	211.143	212.193	212.709	213.240	213.856	215.693	215.351
2010	216.687	216.741	217.631	218.009	218.178	217.965	218.011
2011	220.223	221.309	223.467	224.906	225.964	225.722	225.922
2012	226.665	227.663	229.392	230.085	229.815	229.478	229.104

Year	Aug	Sep	Oct	Nov	Dec	Annual
2003	184.6	185.2	185.0	184.5	184.3	184.0
2004	189.5	189.9	190.9	191.0	190.3	188.9
2005	196.4	198.8	199.2	197.6	196.8	195.3
2006	203.9	202.9	201.8	201.5	201.8	201.6
2007	207.917	208.490	208.936	210.177	210.036	207.342
2008	219.086	218.783	216.573	212.425	210.228	215.303
2009	215.834	215.969	216.177	216.330	215.949	214.537
2010	218.312	218.439	218.711	218.803	219.179	218.056
2011	226.545	226.889	226.421	226.230	225.672	224.939
2012	230.379	231.407	231.317	230.221	229.601	229.594

Solution:



Uses of a Time Series Graph

Time series graphs are important tools in various applications of statistics. When recording values of the same variable over an extended period of time, sometimes it is difficult to discern any trend or pattern. However, once the same data points are displayed graphically, some features jump out. Time series graphs make trends easy to spot.

Stem-and-Leaf Diagram

A clear disadvantage of using a frequency table is that the identity of individual observations is lost in grouping process. To overcome this draw back, **John Turkey** introduced a technique known as *the Stem and Leaf Display*. This technique offers a quick and novel way for simultaneously sorting and simplifying data sets, where each number in data set is divided into two parts, a stem and a leaf. A **stem** is the leading digit(s) of each number and used in sorting, while a **leaf** is the rest of the number or the trailing digit(s) and shown in display.

Importance

The stem and leaf display is a useful step for listing the data in an array, leaves are associated with the stem to know the numbers. The stem and leaf table provide a useful description of the data set and can easily be converted to a frequency table. It is a common practice to arrange the trailing digits in each case from smallest to highest.

Steps in the Construction of a Stem-and-Leaf Diagram

1. Determine the stems and list them in a column from smallest to largest or largest to smallest.
2. List the remaining digit of each stem as a leaf to the right of the stem.
3. Include a legend that explains the meaning of the stems and the leaves. Include a title for the diagram.

22. Construct a stem-and-leaf display from the data and list the data in an array.
48,31,54,37,18,64,61,43,40,71,51,12,52,65,53,42,39,62,74,48,29,67,30,49,68,35, 57,26,27,58

Solution:

A scan of the data indicates that the observation range is 12 to 74. We use the first observation is 48, which has a stem of 4 and a leaf of 8, the second a stem of 3 and a leaf of 1, etc. Then our required stem and leaf display is as follows;

Stem (Leading Digit)	Leaf (Trailing Digit)
1	8 2
2	9 6 7
3	1 7 9 0 5
4	8 3 0 2 8 9
5	4 1 2 3 7 8
6	4 1 5 2 7 8
7	1 4

23. Construct stem-and-leaf diagram for the following history test scores:
65, 72, 96, 86, 43, 61, 75, 86, 49, 68, 98, 74, 84, 78, 85, 75, 86, 73

Solution:

In the stem-and-leaf diagram on the following page, we have organized the history test scores by placing all of the scores that are in the 40s in the top row, the scores

that are in the 50s in the second row, the scores that are in the 60s in the third row, and so on. The tens digits of the scores have been placed to the left of the vertical line. In this diagram they are referred to as stems. The ones digits of the test scores have been placed in the proper row to the right of the vertical line. In this diagram they are the leaves. It is now easy to make observations about the distribution of the scores. Only two of the scores are in the 90s. Six of the scores are in the 70s, and none of the scores are in the 50s. The lowest score is 43 and the highest is 98.

Stem	Leaves
4	3 9
5	
6	1 5 8
7	2 3 4 5 5 8
8	4 5 6 6 6
9	6 8

Legend: $8|6$
Represents 86

Remark: The choice of how many leading digits to use as the stem will depend on the particular data set.

24. Construct stem-and-leaf diagram for the following data set, in which a travel agent has recorded the amounts spent by customers for a cruise.

$3600	$4700	$7200	$2100	$5700	$4400	$9400
$6200	$5900	$2100	$4100	$5200	$7300	$6200
$3800	$4900	$5400	$5400	$3100	$3100	$4500
$4500	$2900	$3700	$3700	$4800	$4800	$2400

Solution:

One method of choosing the stems is to let each thousands digit be a stem and each hundreds digit be a leaf. If the stems and leaves are assigned in this manner, then the notation 2|1 with a stem of 2 and a leaf of 1, represents a cost of $2100, and 5|4 represents a cost of $5400. A stem-and-leaf diagram can now be constructed by writing all of the stems in a column from smallest to largest to the left of a vertical line, and writing the corresponding leaves to the right of the line.

Stem	Leaves
2	1 1 4 9
3	1 1 6 7 7 8
4	1 4 5 55 7 8 8 9
5	2 4 4 7 9
6	2 2
7	2 3
8	
9	4

Legend: $7|3$
Represents $\$7300$

Remark: Sometimes two sets of data can be compared by using a back-to-back stem-and- leaf diagram, in which common stems are listed in the middle column of the diagram. Leaves from one data set are displayed to the right of the stems, and leaves from the other data set are displayed to the left.

25. Construct stem-and-leaf diagram for the following data set below shows the test scores for two classes that took the same test.
Scores in the range of 70s, 80s, 90s, 40s, 50s, and 60s

Solution:

The back-to-back stem-and-leaf diagram below shows the test scores for two classes that took the same test. It is easy to see that the 8 A.M. class did better on the test because it had more scores in the 80s and90s and fewer scores in the 40s, 50s, and 60s. The number of scores in the 70s was the same for both classes.

Biology Test Scores		
8 A.M. Class		**10 A.M. Class**
2	4	5 8
7	5	6 7 9 9
5 8	6	2 3 4 8
1 2 3 3 3 7 8	7	1 3 3 5 5 6 8
4 4 5 5 6 8 8 9	8	2 3 6 6 6
2 4 5 5 8	9	4 5
Legend: 3|7		Legend: 8|2
Represents 73		Represents 82

26. Draw Stem and Leaf graph for Susan Dean's spring pre-calculus class, scores for the first exam were as follows (smallest to largest):
33; 42; 49; 49; 53; 55; 55; 61; 63; 67; 68; 68; 69; 69; 72; 73; 74; 78; 80; 83; 88; 88; 88; 90; 92; 94; 94; 94; 94; 96; 100

Solution:

Stem and Leaf Graph	
Stem	**Leaves**
3	3
4	2 9 9
5	3 5 5
6	1 3 7 8 8 9 9
7	2 3 4 8
8	0 3 8 8 8
9	0 2 4 4 4 4 6
10	0

The stem plot shows that most scores fell in the 60s, 70s, 80s, and 90s. Eight out of the 31 scores or approximately 26% were in the 90s or 100, a fairly high number of As.

27. For the Park City basketball team, scores for the last 30 games were as follows (smallest to largest):
32; 32; 33; 34; 38; 40; 42; 42; 43; 44; 46; 47; 47; 48; 48; 48; 49; 50; 50; 51; 52; 52; 52; 53; 54; 56; 57; 57; 60; 61
Construct a stem plot for the data.

Solution:

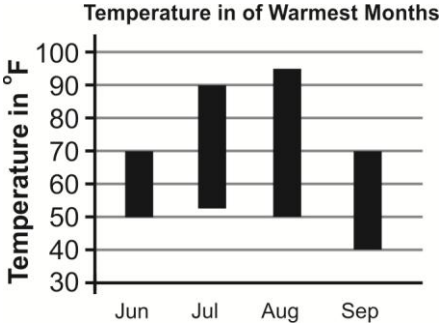
Stem	Leaves
3	2 2 3 4 8
4	0 2 2 3 4 6 7 7 8 8 8 9
5	0 0 1 2 2 2 3 4 6 7 7
6	0 1

The stem plot is a quick way to graph data and gives an exact picture of the data. You want to look for an overall pattern and any outliers. An **outlier** is an observation of data that does not fit the rest of the data. It is sometimes called an **extreme value**. When you graph an outlier, it will appear not to fit the pattern of the graph. Some outliers are due to mistakes (for example, writing down 50 instead of 500) while others may indicate that something unusual is happening.

Range Bar

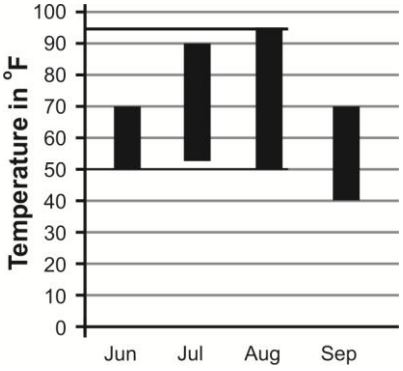
A range bar graph represents a range of data for each independent variable.

28. In the following chart, which month is the warmest month?

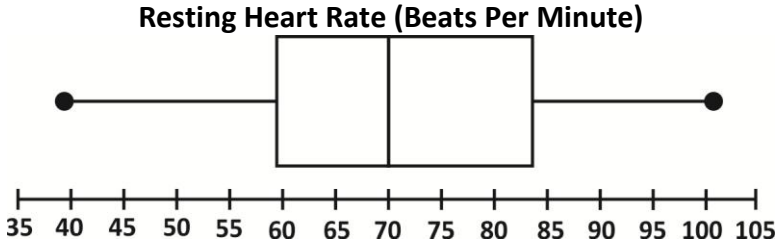


Solution:

In June, temperatures range from 50° F to 70° F
In July, temperatures range from 52° F to 90° F
In August, temperatures range from 50° F to 95° F, 95 being the highest.
In September, temperatures range from 40° F to 70° F
Therefore, the warmest month is **August**
The highest temperature is 95° F.

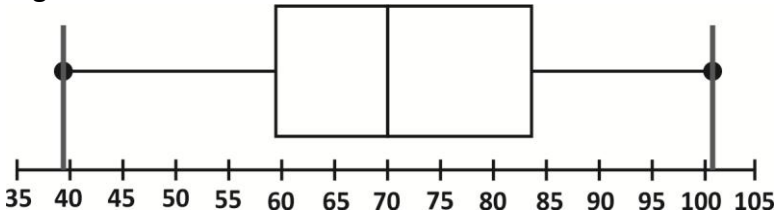


29. The box plot below summarizes the resting heart rates, in beats per minute, of the members at a gym. Which of the following could be the range of resting heart rates, in beats per minute?



Solution:

The chart shows resting heart rates (beats/minute), and we are asked to calculate range of resting heart rates.



We know that

Range = Maximum – Minimum

Range = 101 – 139 = 62

Therefore, the answer is 62 beats per minute.

Chart

A chart is the diagrammatic representation of a spatial series where the data is split into different categories.

Different types of Charts

Histograms are for quantitative data. There is a similar graph for qualitative data (categories), called a **Bar Graph**. In a bar graph, the width of the bars is arbitrary and the bars are not connected. A Bar Graph is a graph that consists of bars for each category with the length/height of the bars specifying the frequency for each category. One axis will indicate the categories and the other axis will indicate the frequency. Bar graphs are often used for comparing different characteristics of items. Bar graphs can have horizontal or vertical bars. Can show frequency or relative frequency. Few types are as follows;

Bar Graph/ Bar Chart/ Simple Bar Chart: A simple bar chart is used when the data consists of a single component and also do not involve much variation.

Multiple Bar Charts: When we represent two or more sets of inter related data and also we want to compare different phenomena diagrammatically we use multiple bar chart.

Component Bar Charts: Component bar charts are used when the data consists of more than one homogeneous component, they represents the commutations of the various components of data.

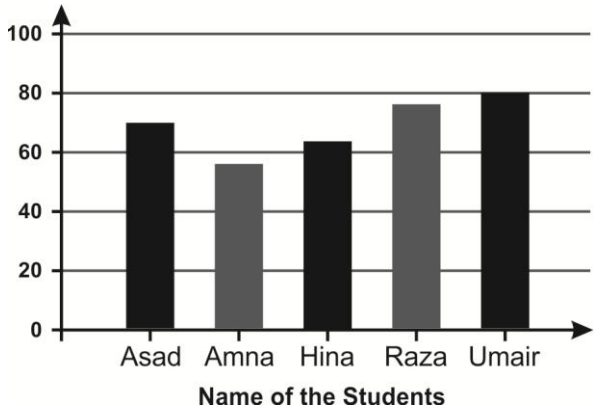
Percentage Component Bar Charts: Percentage component bar charts are used if the comparison of relative values are concerned. These charts are drawn on percentage basis.

Pareto Chart: A **Pareto chart** is a specific type of bar graph where the classes are reordered so that the bars are in size order.

30. Draw the simple bar chart for the following

Name	Asad	Amna	Hina	Raza	Umair
Marks	70	57	65	75	82

Solution:



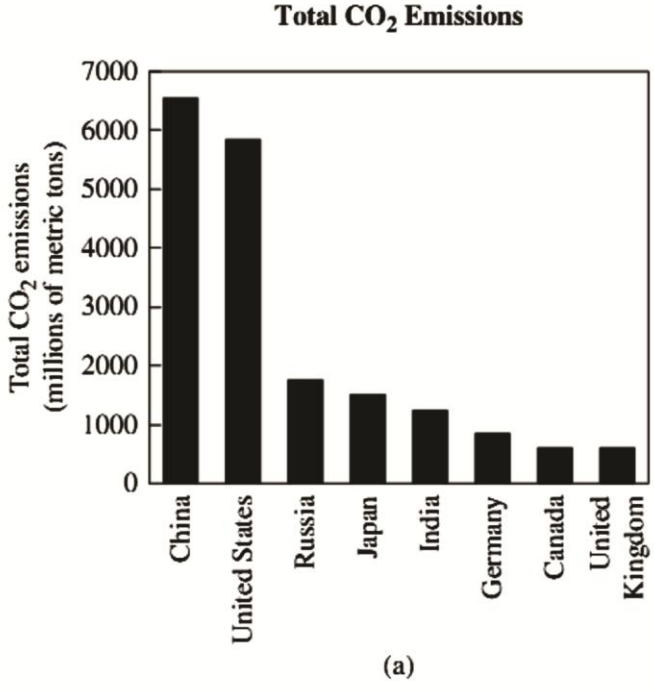
31. Carbon Dioxide emissions: Carbon dioxide is released into the atmosphere primarily by the combustion of fossil fuels (oil, coal, natural gas). Table lists the eight countries that emit the most carbon dioxide each year. Make bar graphs for the total emissions and the emissions per person. Put the bars in descending order of size.

Country	Total carbon dioxide Emissions (millions of metric tons of carbon)	Pre-Person Carbon Dioxide Emissions (metric tons of carbons)
China	6534	4.91
United States	5833	19.18
Russia	1729	12.29
Japan	1495	9.54
India	1214	1.31
Germany	829	10.06
Canada	574	17.27
United kingdom	572	9.38

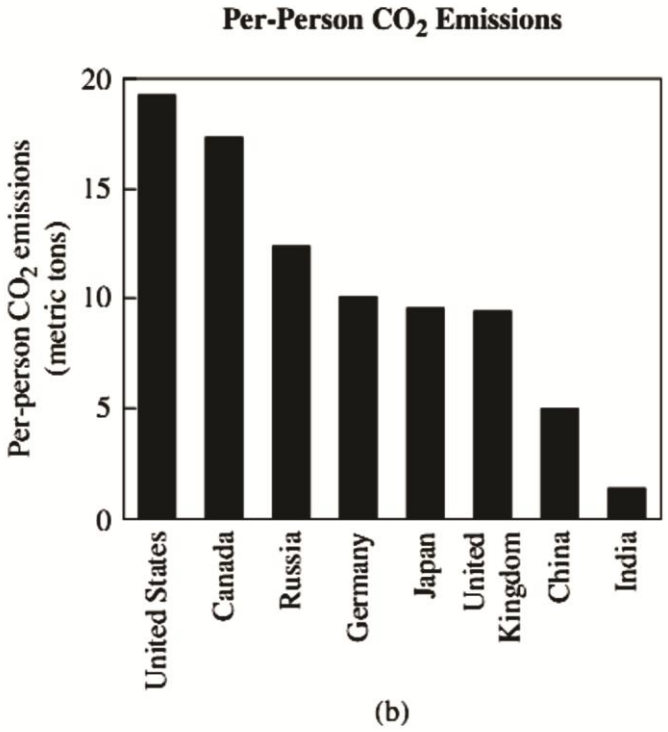
Solution:

The categories are the countries, and the frequencies are the data values. The total emissions are given in units of "millions of metric tons," and the highest value in these units is 6534; therefore, a range of 0 to 7000 makes a good choice for the vertical scale. The per-person emissions are given in metric tons, and the highest value is 19.18 for the United States; therefore, a range of 0 to 20 works well. Figure shows the two bar graphs, with bars placed in order of descending height.

The values for total carbon dioxide emissions go from 145 to 1802 (millions of tons), so a range of 0 to 2000 makes a good choice for the vertical scale. Each bar's height corresponds to its data value, and we label the category (country) under the bar. Figure (a) shows the bar graph for total emissions, with bars in order of decreasing height and (b) per-person carbon dioxide emissions by country.



(a)

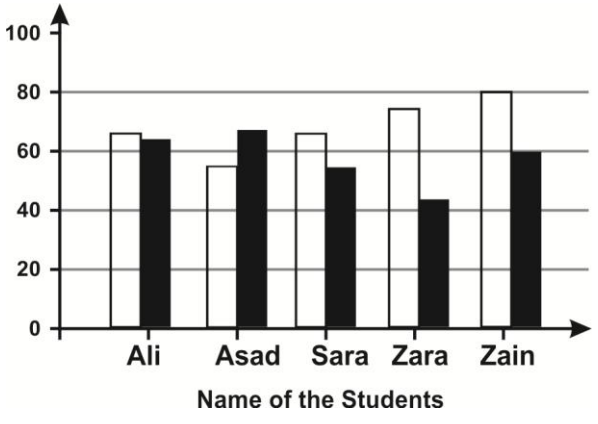


(b)

32. Draw the multiple bar chart for the following

Name	Ali	Asad	Sara	Zara	Zain
MBF	70	57	65	75	82
I to B	65	70	57	45	60

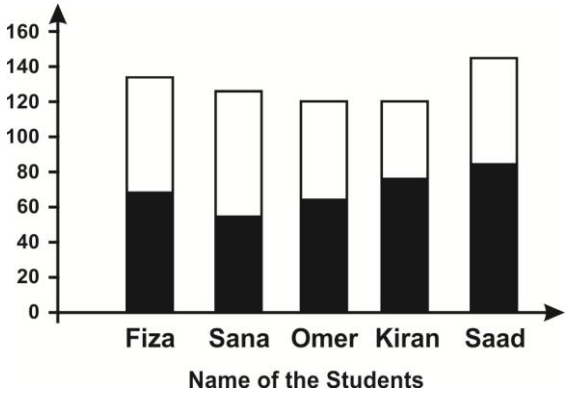
Solution:



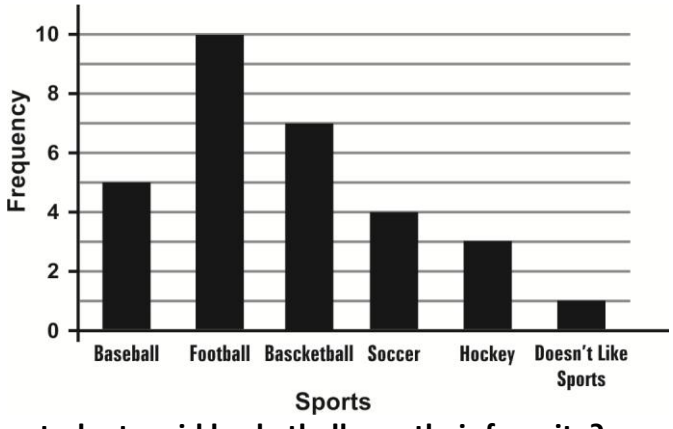
33. Draw the component bar chart for the following

Name	Fiza	Sana	Omer	Kiran	Saad
MBF	130	120	115	115	140
I to B	65	45	57	70	80

Solution:



34. A group of students surveyed their class about what sport was their favorite. The results are given below.



a) How many students said basketball was their favorite?

b) How many more students liked Football than Soccer?

Solution:

a) Find the bar above that is labelled "Basketball" on the horizontal axis. The number of students whose favorite is basketball is equivalent to the height of the bar. Read the height of the bar from the vertical axis. The number of students who said basketball was their favorite sport was 7.

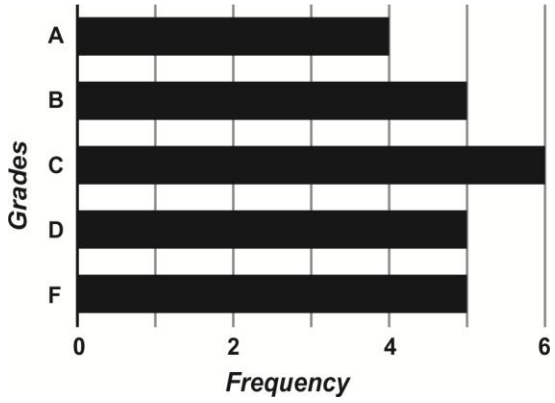
b) Begin by observing the number of students who liked Football and Soccer from the bar graph. The number of students who liked Football is 10 and the number of students who liked Soccer is 4. There are 10 − 4 = 6 more students that liked Football than Soccer.

35. The bar graph below shows scores on a Math test.

a) How many B's were there?

b) How many test grades are there?

c) How many students scored a C or higher?



Solution:

a) This is a horizontal bar graph. The grade categories are on the vertical axis. Begin by looking for "B". Then look at the horizontal axis for the length of the bar for B's. This gives 5 B's.

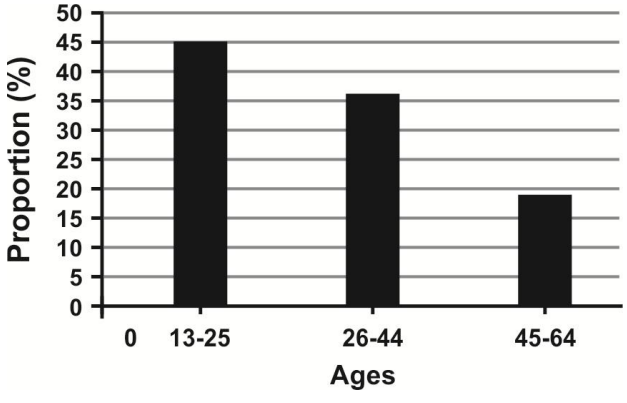
b) Add all the lengths of each bar for each letter grade together to find the number of test grades. This gives a total of 4 + 5 + 6 + 2 + 3 = 20

c) Begin by finding the number of test grades for A's, B's, and C's, which are4, 5, and 6 respectively. Number of grades of C or better are 4 + 5 + 6 = 15.

36. By the end of 2011, Facebook had over 146 million users in the United States. Table shows three age groups, the number of users in each age group, and the proportion (%) of users in each age group. Construct a bar graph using this data.

Age groups	Number of facebook users	Proportion (%) of facebook users
13-25	65082280	45%
26-44	53300200	36%
45-64	27885100	19%

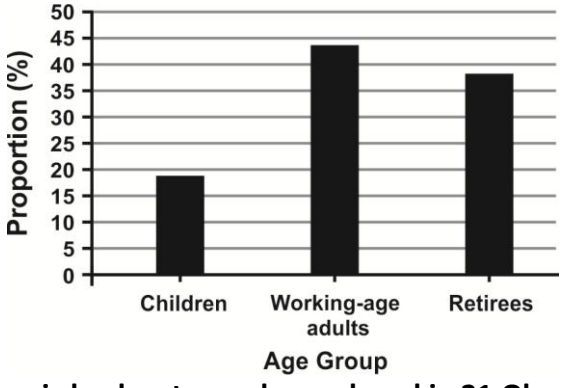
Solution:



37. The population in Park City is made up of children, working-age adults, and retirees. Table shows the three age groups, the number of people in the town from each age group, and the proportion (%) of people in each age group. Construct a bar graph showing the proportions.

Age groups	Number of peoples	Proportion of population
Children	67059	19%
Working-age-adults	152198	43%
Retirees	131662	38%

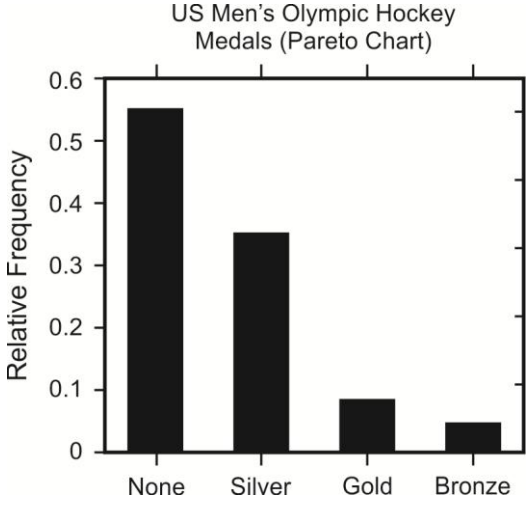
Solution:



38. The US Men's Olympic hockey teams have played in 21 Olympic games, winning 11 medals (2 gold, 8 silver, 1 bronze). Make a frequency bar graph and a relative frequency Pareto chart.

Solution:

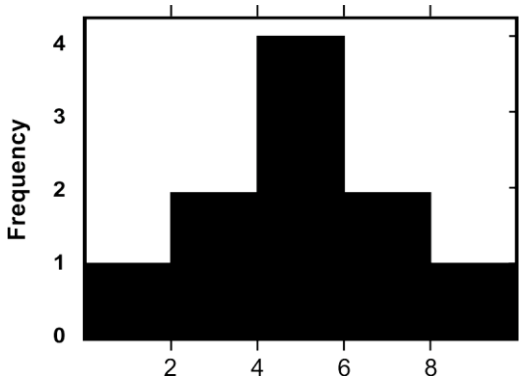
Below are a frequency bar graph (ordered from worst to best finishes) and a relative frequency Pareto chart (ordered from highest to lowest).



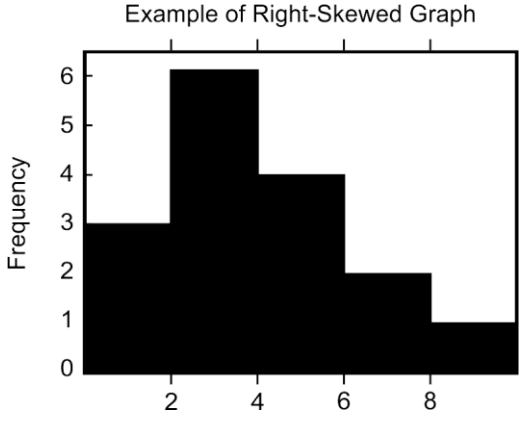
Remember: Graphs can help us get an overall view of the data set. When looking at a graph, pay attention to the following:

- **Center:** where is the middle of the graph, and the highest point.
- **Spread:** how are the parts of the graph spread out from each other?
- **Shape:** what shape does the graph have? Bell shape, straight across, repeatedly up and down, random?
- **Symmetry:** graph can be split in half with two mirror image parts, almost equal amount on both sides. Graph that extends more out to left is Left-skewed. Graph that extends more out to right is Right-skewed.
- **Outliers:** are data values (small parts of graph) that are far from other data (parts).

39. Give an example of symmetric graph.
Solution:

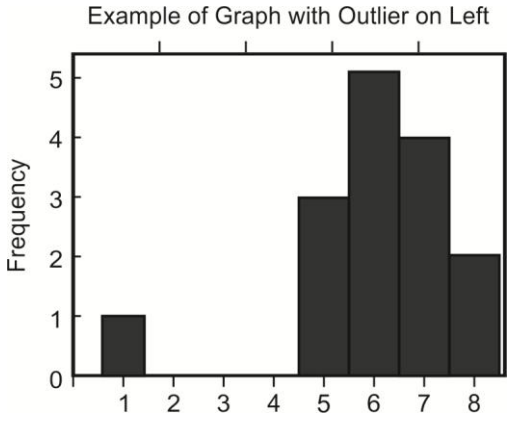


40. Give an example of right skewed graph.
Solution:



41. Give an example of graph for outlier.

Solution



Circle Graph /Pie Chart

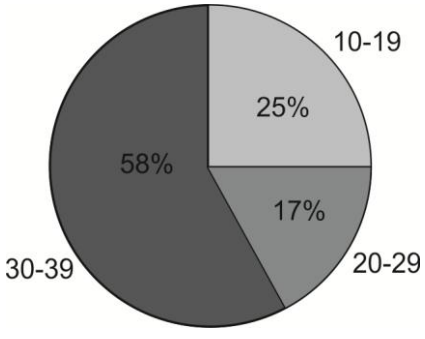
A Circle Graph (or Pie Chart) is a circle cut into sections with varying sizes shaped like slices of a pizza. The sizes of the sections are based on the relative frequencies of the categories. The percent or frequency for each category can be specified on the sections of the pie chart.

It is a graphical device for presenting qualitative data summaries based on subdivision of a circle into sectors that corresponds to the relative frequency for each class. It is a disk (circle) divided into pie-shaped pieces proportional to the relative frequencies. A pie chart should be labeled well, with class and the relative frequency for each slice. If a slice is very small, then the labels can go outside with an arrow pointing to the corresponding slice. The preferred way to sketch a pie chart is to start slices at 12 o'clock and rotate clockwise.

42. For Bradley's weekly hours at a summer job:

25, 32, 36, 32, 18, 28, 30, 36, 12, 16, 35, 36. Construct a Pie Chart.

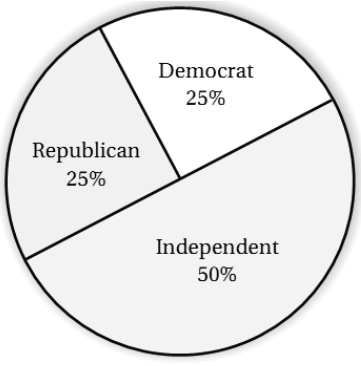
Solution



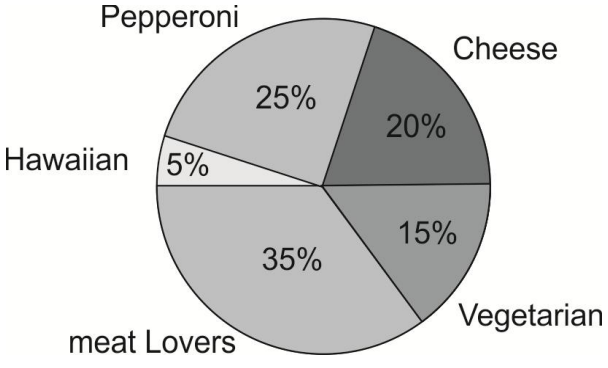
43. Among the registered voters in Rochester County, 25% are Democrats, 25% are Republicans, and 50% are Independents. Make a pie chart showing the breakdown of party affiliations in Rochester County.

Solution:

Because Democrats and Republicans each represent 25% of the voters, the wedges for Republicans and Democrats each occupy 25%, or one-fourth, of the pie. Independents represent half of the voters, so their wedge occupies the remaining half of the pie. Figure shows the result. As always, note the importance of clear labeling.



44. A pizza chef at Mario's Pizza makes a list of all the types of pizza he made on a particularly busy Tuesday. He then used this list to create a pie chart for the pizza types.



a) What was the least ordered type of pizza?

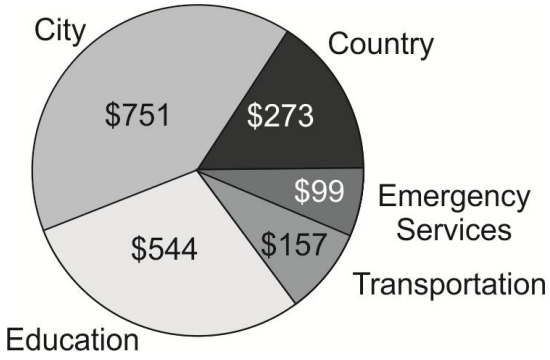
b) What was the most ordered type of pizza?

Solution

a) Start by identifying the smallest piece on the pie chart. Looking at this graph shows that this category is that of Hawaiian at 5%.

b) Identifying the largest piece on the pie chart. Looking at this graph shows that this category is that of Meat Lovers at 35%.

45. Using the Circle Graph answer the following questions.



a) How much in taxes did Tyrone pay in total?
b) How much of Tyrone's Taxes went to the City?
c) Which of the categories on the pie chart received the smallest amount of money?

Solution

a) Sum of the money amount for each category represented on the chart. The total amount is $99 + $157 + $273 + $544 + $751 = $1824.
b) Look on the chart for the piece labeled City and read the value there which is $751.
c) Look at the chart and look for the category with the lowest amount of money. In this case the category is Emergency Services at $99.

Misleading Graphs

Misleading graphs are visual representations of data that intentionally or unintentionally deceive or mislead the viewer, often by manipulating the presentation, scale, or data selection to support a particular agenda or conclusion.

Examples of Misleading Graphs in Statistics

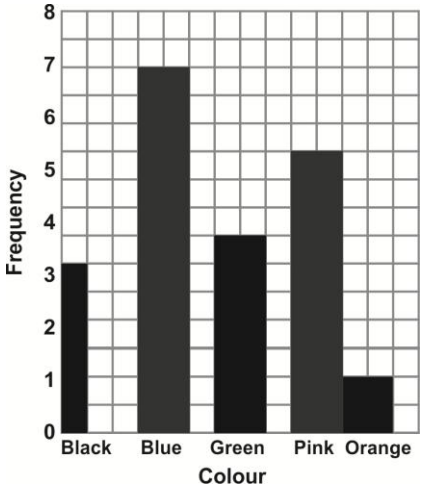
Truncated Axis: Omitting parts of the axis to exaggerate changes.
Example: A graph showing a 10% increase in sales, but the y-axis starts at 90% instead of 0%.
Logarithmic Scale: Using a logarithmic scale to make small changes appear large.
Example: A graph showing stock prices, using a logarithmic scale to exaggerate fluctuations.
Biased Labeling: Using emotive or misleading labels.
Example: A graph labeled "Huge Increase in Crime Rate" when the actual increase is small.
Cherry-Picked Data: Selectively presenting data to support a claim.
Example: Showing only data that supports a trend, while ignoring contradictory data.
Omitting Data: Leaving out important data to change the narrative.

Example: Omitting data points that contradict a trend.
Dual Scales: Using two different scales on the same graph.
Example: A graph with two different scales, making it difficult to compare data.
Cumulative Graphs: Using cumulative data to make a trend appear more impressive.
Example: Showing cumulative sales over time, rather than monthly sales.
Average vs. Median: Using averages instead of medians to misrepresent data.
Example: Using average income instead of median income to hide income inequality.
Misleading Colors: Using colors to influence interpretation.
Example: Using red to indicate a "bad" trend and green to indicate a "good" trend.
3D Graphs: Using 3D graphs to make data appear more impressive.
Example: A 3D pie chart that exaggerates the size of a particular slice.

46. Using an example show that bar chart is a misleading graph.

Solution:

Colour	Frequency
Black	3
Blue	7
Green	4
Pink	6
Orange	1

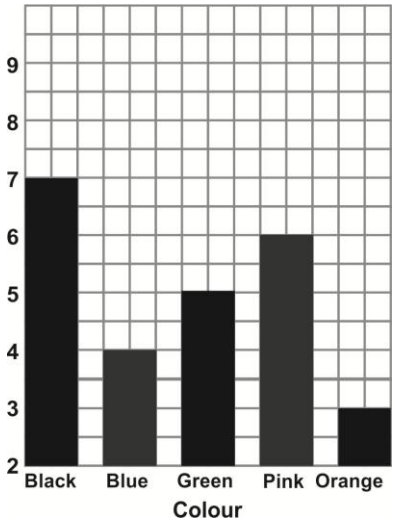


- Bars have not equal width.
- Spacing between bars not equal.
- Bars drawn at the wrong height.

47. Using an example show that bar chart is a misleading graph.

Solution:

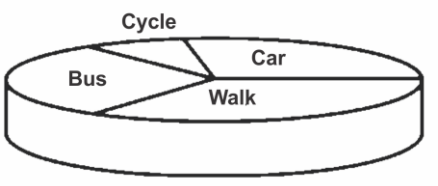
Colour	Frequency
Black	7
Blue	4
Green	5
Pink	6
Orange	3



- Frequency axis not starting at zero.
- Frequency axis not labeled.
- One of the bars not labeled.
- Spacing between numbers on frequency axis (or missed).

48. Using an example show that Pie chart is a misleading graph.

Solution:

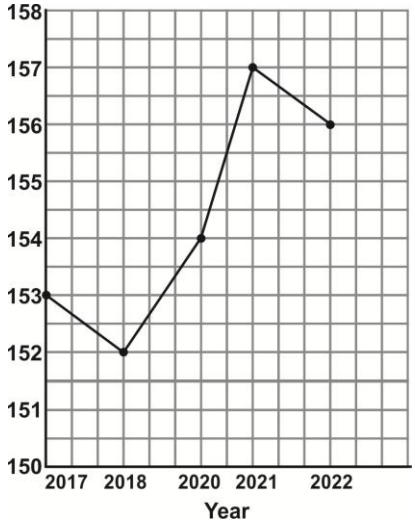


- 3D pie charts can be misleading.
- Sector not labeled.
- Points plotted in the wrong place.
- Wrong angle drawn.
- Wrong angle calculated.

49. Using an example show that line graph is a misleading graph.

Solution:

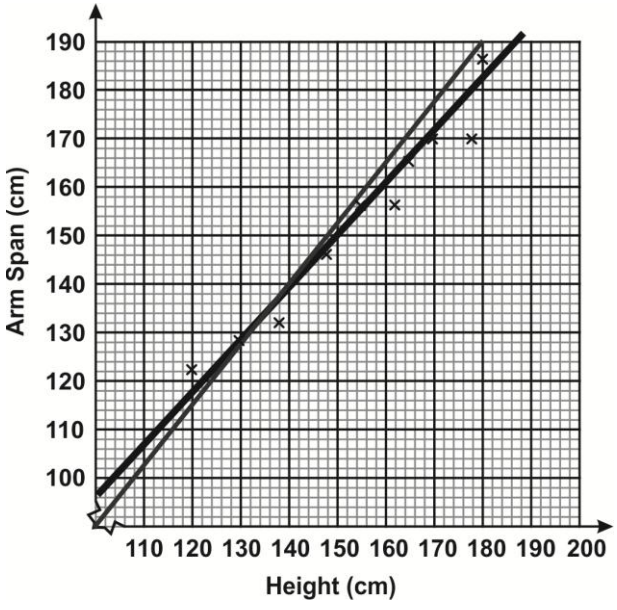
Year	Share Price
2017	154
2018	152
2019	154
2020	157
2021	156



- Vertical axis not starting at zero.
- Vertical axis not labeled.
- Points plotted in the wrong place.
- Number missing on horizontal axis.
- Spacing on the axes.
- Curves rather than lines (not shown).

50. Using an example show that scatter graph is a misleading graph.

Solution:



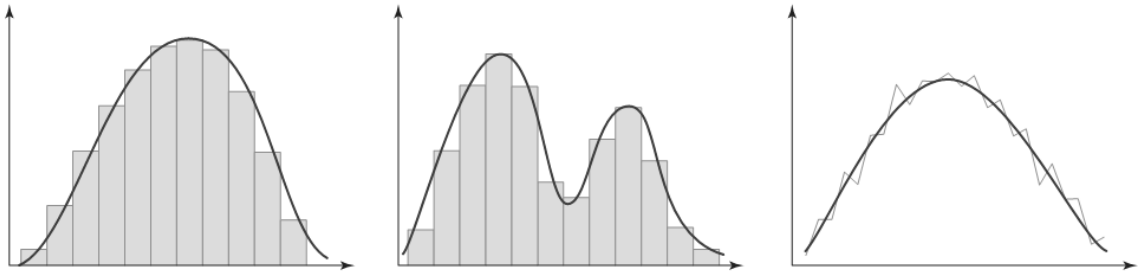
- Points plotted in the wrong place.
- Problems with axes – scale, label, suitable etc.
- Line of best fit not drawn in a suitable location.

Shapes of Distributions

We next turn our attention to describing the overall shape of a distribution. We can see the complete shape of a distribution on a graph. Our goal is to describe the general shape in words, which we do by focusing on three characteristics visible on the graph of a distribution: its number of modes, its symmetry or skewness, and its variation.

Number of Modes

Because we are interested in the general shapes of distributions, it's easier to work with smooth curves that fit the data, rather than with the actual data. Figure shows three examples of this idea, two in which the distributions are shown as histograms and one in which the distribution is shown as a line chart. In each case, the smooth curves make good approximations to the original distributions.



(a) (b) (c)

The first way we might characterize the shape of each distribution is by its number of modes, or peaks. The distributions in Figure (a) and Figure (c) have one mode, so we say that they are unimodal, or single-peaked. The distribution in Figure (b) has two modes, even though the second peak is lower than the first; it is a bimodal distribution. Other distributions may have no modes (they are called uniform distributions) or more than two modes.

51. Number of Modes: How many peaks would you expect for each of the following distributions? Why? Make a rough sketch for each distribution, with clearly labeled axes.

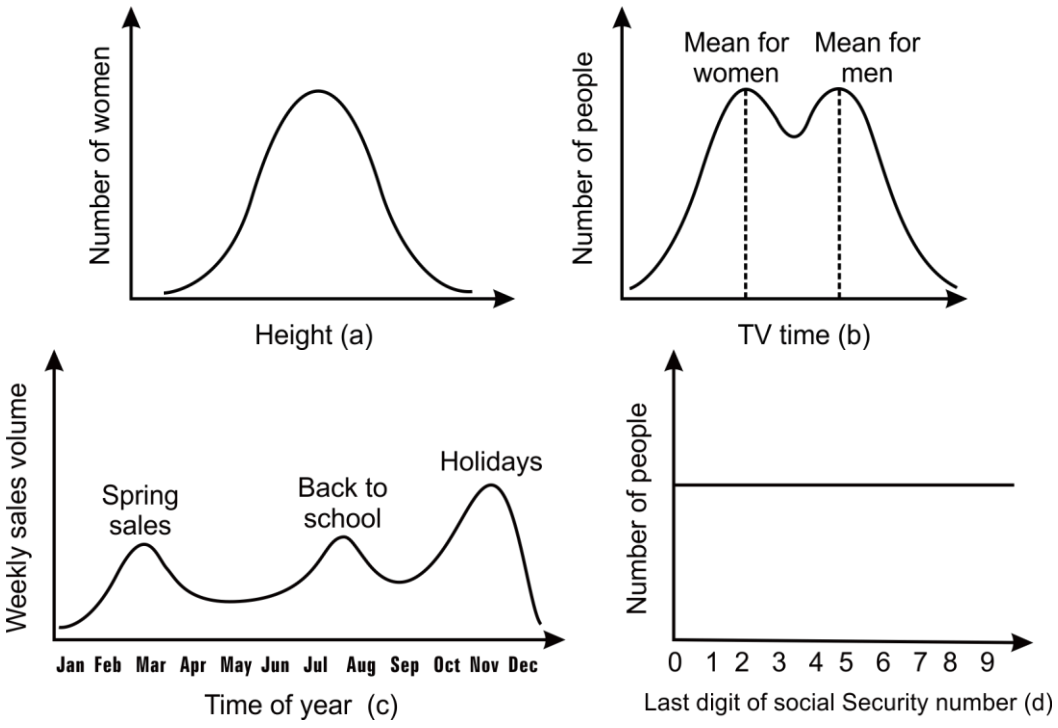
a. Heights of 1,000 randomly selected adult women

b. Hours spent watching football on TV in January for 1,000 randomly selected adult Americans

c. Weekly sales throughout the year at a retail clothing store for children

d. The number of people with particular last digits (0 through 9) in their Social Security numbers

Solution: Figure shows sketches of the distributions.

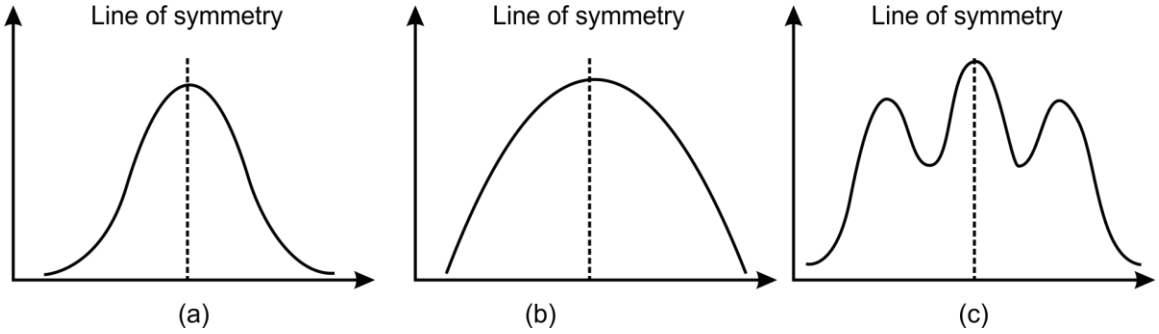


a. The distribution of heights of women is single-peaked (unimodal) because many women are at or near the mean height, with fewer and fewer women at heights much greater or less than the mean.
b. The distribution of times spent watching football on TV for 1000 randomly selected adult Americans is likely to be bimodal (two modes). One mode represents the mean watching time of men, who tend to watch more football than women, and the other represents the mean watching time of women.
c. The distribution of weekly sales throughout the year at a retail clothing store for children is likely to have several modes. For example, it will probably have a mode in spring for sales of summer clothing, a mode in late summer for back-to-school sales, and another mode in winter for holiday sales.
d. The last digits of Social Security numbers are essentially random, so the number of people with each different last digit (0 through 9) should be about the same. That is, about 10% of all Social Security numbers end in 0, 10% end in 1, and so on. It is therefore a uniform distribution with no mode.

Symmetry or Skewness

A second way to describe the shape of a distribution is in terms of its symmetry or skewness.

The distributions in Figure are all symmetric.



Symmetric Distribution: A single-peaked distribution is symmetric if its left half is a mirror image of its right half.

Skewed Distribution: A distribution that is not symmetric must have values that tend to be more spread out on one side than the other. In this case, we say that the distribution is skewed.

Left-skewed Distribution (or negatively skewed): A single-peaked distribution is left-skewed if its values are more spread out on the left side of the mode.

Right-skewed Distribution (or positively skewed): A single-peaked distribution is right-skewed if its values are more spread out on the right side of the mode.

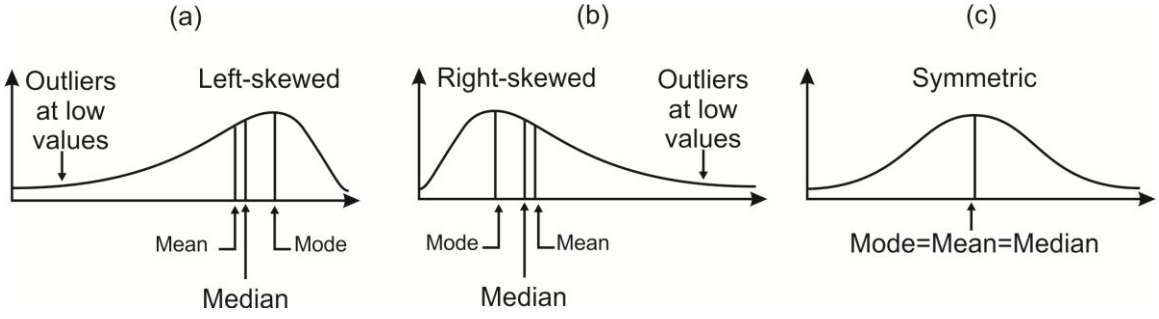


Figure also shows how skewness affects the relative positions of the mean, median, and mode. By definition, the mode is the peak in a single-peaked distribution. A left-skewed distribution pulls both the mean and median to the left of the mode; that is, to values less than the mode. In addition, outliers at the low end of the data set make the mean less than the median. Similarly, a right-skewed distribution pulls the mean and median to the right—to values greater than the mode—and the large outliers make the mean greater than the median. When the distribution is symmetric and single-peaked, both the mean and the median are equal to the mode.

52. Skewness: For each of the following situations, state whether you expect the distribution to be symmetric, left-skewed, or right-skewed. Explain.

- **a. Heights of a sample of 100 women**
- **b. Number of books read during the school year by fifth graders**
- **c. Speeds of cars on a road where a visible patrol car is using radar to detect speeders**

Solution:

a. The distribution of heights of women is symmetric, because roughly equal numbers of women are shorter and taller than the mean and extremes of height are rare on either side of the mean.

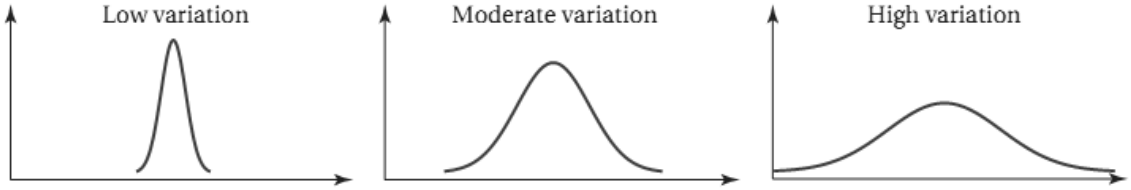
b. The distribution of the number of books read is right-skewed. Most fifth-grade children read a moderate number of books during the school year, but a few voracious readers will read far more than most other students. These students will therefore be outliers with high values for the number of books read, creating a tail on the right side of the distribution.

c. Drivers usually slow down when they are aware of a patrol car looking for speeders. Few if any drivers will be exceeding the speed limit, but some drivers slow to well below the speed limit. The distribution of speeds is therefore left-skewed, with a mode near the speed limit but a few cars going well below the speed limit.

Variation

A third way to describe a distribution is by its variation, which is a measure of how much the data values are spread out. It describes how widely data values are spread out about the center of a distribution.

A distribution in which most data values are clustered together has a low variation. As shown in Figure (a), such a distribution has a fairly sharp peak. The variation is higher when the data are distributed more widely around the center, which makes the peak broader. Figure (b) shows a distribution with moderate variation, and Figure (c) shows a distribution with high variation.



53. Variation in Marathon times: How would you expect the symmetry and variation to differ between times in the Olympic marathon and times in the New York marathon? Explain.

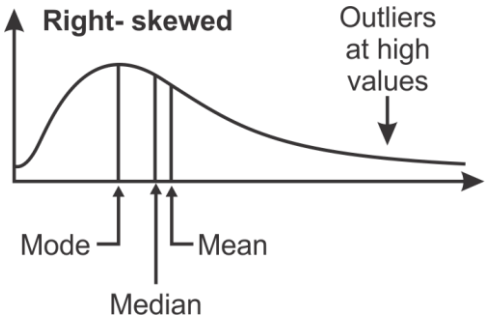
Solution:

The Olympic marathon invites only elite runners, whose times are likely to be clustered not far above world record times. The New York marathon allows runners of all abilities, whose times are spread over a very wide range (from near the world record to many hours). Therefore, the variation among the times should be greater in the New York marathon than in the Olympic marathon.

54. Changes in Household Income: Question offered two seemingly different claims about changes in household income since 1978. One claim stated that the average had risen less than 10%, while the other stated that it had risen more than 27%. Explain how both statements can be true and what this fact tells you about the shape of the distribution.

Solution:

The claims can both be true because the first one refers to the median while the second refers to the mean. The fact that the mean is significantly higher than the median tells us that the distribution of household incomes is right-skewed (see Figure).



This makes sense because most households are middle-class, so the mode of the household income distribution is a middle-class income. But a small number of very high-income households pull the mean to a considerably higher value than either the mode or the median, stretching the distribution to the right (high-income) side. The fact that some incomes are far higher than others also tells us that the distribution has a relatively large variation, at least if we measure variation in terms of the total range of values.

Bivariate Analysis

Bivariate analysis is a statistical technique used to examine the relationship between two continuous or categorical variables. One variable is dependent while other is independent.

Purpose

- Identify relationships between variables
- Determine correlation or association strength
- Visualize data distribution

Types of Bivariate Analysis

- **Scatter Plots:** Visualize relationship between two continuous variables
- **Correlation Coefficient:** Measure strength and direction of linear relationship
- **Contingency Tables:** Examine relationship between two categorical variables
- **T-Tests:** Compare means of two groups
- **Regression Analysis:** Model relationship between independent and dependent variables

Applications

- **Market Research:** Analyze customer behavior
- **Medical Research:** Investigate disease relationships
- **Social Sciences:** Examine relationships between demographic variables
- **Business:** Identify relationships between sales and marketing strategies

Regression line or Least-Squares Line:

The regression line or least-squares line is the line that minimizes the sum of the squares of the vertical distances between the data points and the line. **Or** it is the line for which the sum of the squares of the vertical distances between the data points and the line is a minimum.

Linear Regression:

The process of finding the regression line is called **linear regression**.

Regression:

The dependence of variable upon one or more other variables is called regression.

Simple Regression:

The dependence of one variable upon single independent variable is called simple regression.

Multiple Regressions:

The dependence of one variable upon two or more independent variables is called simple regression.

Regressor:

The variable that forms the basis of estimation or prediction is called regressor. It is also called as the predictor variable or independent variable or controlled variable or explanatory variable or non – random variable.

Regressand:

The variable whose resulting value depends upon the selected value of the independent variable is called the regressand. It is also called the response variable or the predicted variable or dependent variable or explained variable or random variable.

Regression Analysis:

Regression analysis refers to the methods of describing the functional dependent on the basis of the other independent variable.

The Principle of Least Square:

This principle says that the sum of the squares of the residual of the observed values from their corresponding estimated values should be the least or minimum.

Mathematically; $Least = S = \sum \left(Y_i - \hat{Y} \right)^2$

Estimated Regression Line

The estimated regression line of Y on X; $Y = a_{YX} + a_{YX}X$	The estimated regression line of X on Y; $X = a_{XY} + a_{XY}Y$
$b_{YX} = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sum(X-\overline{X})^2}$	$b_{XY} = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sum(Y-\overline{Y})^2}$
$b_{YX} = \frac{\sum(XY-n\overline{X}\,\overline{Y})}{\sum(X-\overline{X})^2}$	$b_{XY} = \frac{\sum(XY-n\overline{X}\,\overline{Y})}{\sum(Y-\overline{Y})^2}$
$b_{YX} = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}}$	$b_{XY} = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sum Y^2 - \frac{(\sum Y)^2}{n}}$
$b_{YX} = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2}$	$b_{XY} = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum Y^2 - (\sum Y)^2}$
$b_{YX} = \frac{S_{XY}}{S_X^2} = r\frac{S_Y}{S_X}$	$b_{XY} = \frac{S_{XY}}{S_Y^2} = r\frac{S_X}{S_Y}$
$a_{YX} = \frac{\sum Y - b_{YX}\sum X}{n}$	$a_{XY} = \frac{\sum X - b_{XY}\sum Y}{n}$
$a_{YX} = \overline{Y} - b_{YX}\,\overline{X}$	$a_{XY} = \overline{X} - b_{XY}\,\overline{Y}$

55. Fit a regression line Y on X from percentage of marks scored by 12 students in statistics X and economics Y.

x	30	34	26	49	60	62	65	51	44
y	27	18	34	28	26	30	32	30	28

Solution:

The estimated regression line is $\hat{y} = a + bx$

x	y	x^2	xy
30	27	900	810
34	18	1156	612
26	34	676	884
49	28	2401	1372
60	26	3600	1560
62	30	3844	1860
65	32	4225	2080
51	30	2601	1530
44	28	1936	1232
$\sum x = 421$	$\sum y = 253$	$\sum x^2 = 21339$	$\sum xy = 11940$

The least square estimate for n = 9 are

$$b = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} = 0.064 \;\; ; \;\; a = \frac{\sum y - b\sum x}{n} = 25.12$$

The best fitted line is $\hat{y} = 25.12 + 0.064x$

56. Fit a regression line X on Y from percentage of marks scored by 12 students in statistics X and economics Y.

x	30	34	26	49	60	62	65	51	44
y	27	18	34	28	26	30	32	30	28

Solution:

The estimated regression line is $\hat{y} = a + by$

x	y	y^2	xy
30	27	729	810
34	18	324	612
26	34	1156	884
49	28	784	1372
60	26	676	1560
62	30	900	1860
65	32	1024	2080
51	30	900	1530
44	28	784	1232
$\sum x = 421$	$\sum y = 253$	$\sum y^2 = 21339$	$\sum xy = 11940$

The least square estimate for n = 9 are

$$b = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum y^2 - (\sum y)^2} = 0.64 \;\; ; \;\; a = \frac{\sum x - b\sum y}{n} = 28.79$$

The best fitted line is $\hat{y} = 28.79 + 0.64y$

Method of Least Square:

A procedure in which regression equation is obtained by minimizing the sum of squares of residuals (errors). The parameter values obtained are called least square estimates.

Normal Equations of Least Square Regression Line

- $\sum Y = na + b\sum X$ this is the normal equation for 'a'
- $\sum XY = a\sum X + b\sum X^2$ this is the normal equation for 'b'

Residual:

Residual is the difference between actual value Y_i from predicted value $\hat{Y}_i$. This error term is denoted by e_i.

i.e. Residual $= e_i = Y_i - \hat{Y}_i = Y_i - (a + bX_i) = Y_i - a - bX_i$

57. Fit a straight line by method of least squares to the following data and estimate Y for X = 30 where X = Supply and Y = Demand.

Score(x)	0	5	10	15	20	25
GPA(y)	12	15	17	22	24	30

Solution:

The estimated equation of straight line is $\hat{y} = a + bx$

x	y	x^2	xy
0	12	0	0
5	15	25	75
10	17	100	170
15	22	225	330
20	24	400	480
25	30	625	750
$\sum x = 75$	$\sum y = 120$	$\sum x^2 = 1375$	$\sum xy = 1805$

The normal equations are

$\sum Y = na + b\sum X \Rightarrow 6a + 75b = 120$ ..... (i)

$\sum XY = a\sum X + b\sum X^2 \Rightarrow 75a + 1375b = 1805$ ..... (ii)

By $25(i) - 2(ii)$ we have $a = 11.25, b = 0.7$

The required fitted straight line is $\hat{y} = 11.25 + 0.7x$

To estimate the value of Y put x = 30; $\hat{y} = 11.25 + 0.7(30) = 32.25$

58. Find a straight line by the method of least squares and show that sum of errors is always zero.

Score(x)	0	1	2	3	4	5	6
GPA(y)	12	10	14	11	13	15	16

Solution:

The estimated equation of straight line is $\hat{y} = a + bx$

x	y	x^2	xy
0	12	0	0
1	10	1	10
2	14	4	28
3	11	9	33
4	13	16	25
5	15	25	75
6	16	36	96
$\sum x = 21$	$\sum y = 91$	$\sum x^2 = 91$	$\sum xy = 294$

The normal equations are

$\sum Y = na + b\sum X \Rightarrow 7a + 21b = 91$ ..... (i)

$\sum XY = a\sum X + b\sum X^2 \Rightarrow 21a + 91b = 294$ ..... (ii)

By $(ii) - 3(i)$ we have $a = 10.75, b = 0.75$

The required fitted straight line is $\hat{y} = 10.75 + 0.75x$

x	y	$\hat{y}$	$y - \hat{y}$
0	12	$10.75 + 0.75(0) = 10.75$	1.25
1	10	$10.75 + 0.75(1) = 10.75$	-1.5
2	14	$10.75 + 0.75(2) = 10.75$	1.75
3	11	$10.75 + 0.75(3) = 10.75$	-2
4	13	$10.75 + 0.75(4) = 10.75$	-0.75
5	15	$10.75 + 0.75(5) = 10.75$	0.50
6	16	$10.75 + 0.75(6) = 10.75$	0.75
$\sum x = 21$	91	91	0

Hence, Sum of errors is always zero. i.e. $\sum\left(Y - \hat{Y}\right) = 0$

59. The following sample of 8 grade point averages and marks in matriculation was observed for students from a college.

Score(x)	480	490	510	510
GPA(y)	2.7	2.9	3.3	2.9
Score(x)	530	550	610	640
GPA(y)	3.1	3.0	3.2	3.7

Find the least square line. Estimate the mean GPA of student scoring 600 marks.

Solution:

The estimated regression line is $\hat{y} = a + bx$

x	y	x^2	xy
480	2.7	230400	1296
490	2.9	240100	1421
510	3.3	260100	1683
510	2.9	260100	1479
530	3.1	280900	1643
550	3.0	302500	1650
610	3.2	372100	1952
640	3.7	409600	2368
$\sum x = 4320$	$\sum y = 24.8$	$\sum x^2 = 2355800$	$\sum xy = 13492$

The least square estimate for n = 8 are

$$b=\frac{n\sum xy-(\sum x)(\sum y)}{n\sum x^2-(\sum x)^2}=0.00435 \quad ; \quad a=\frac{\sum y-b\sum x}{n}=0.751$$

The best fitted line is $\hat{y}=0.751+0.00435x$

For x = 600 we have $\hat{y}=0.751+0.00435(600)=3.361$

60. Given the following data

x	0	1	2	3	4
y	1.0	1.8	3.3	4.5	6.3

Determine the least square line taking x as independent variable. Find the estimated values for the given value of x and show that; $\sum y=\sum\hat{y};\sum e=0$ **.**

Also calculate the sum of squares of the residual. And verify that

$\sum e^2=\sum y^2-a\sum y-b\sum xy$

Solution:

The estimated regression line is $\hat{y}=a+bx$

x	**y**	**x^2**	**xy**
0	1.0	0	0
1	1.8	1	1.8
2	3.3	4	6.6
3	4.5	9	13.5
4	6.3	16	25.2
$\sum x=10$	$\sum y=16.9$	$\sum x^2=30$	$\sum xy=47.1$

The least square estimate for n = 5 are

$$b=\frac{n\sum xy-(\sum x)(\sum y)}{n\sum x^2-(\sum x)^2}=1.33 \; ; \quad a=\frac{\sum y-b\sum x}{n}=0.72$$

The best fitted line is $\hat{y}=0.72+1.33x$

The estimated values $\hat{y}$ for the given values of x and the residuals $e=y-\hat{y}$ are obtained as shown in the following table.

x	y	$\hat{y}$	e	e^2	y^2
0	1.0	$0.72+1.33(0)=0.72$	0.28	0.0784	1.00
1	1.8	$0.72+1.33(1)=2.05$	-0.25	0.0625	3.24
2	3.3	$0.72+1.33(2)=3.38$	-0.08	0.0064	10.89
3	4.5	$0.72+1.33(3)=4.71$	-0.21	0.0441	20.25
4	6.3	$0.72+1.33(4)=6.04$	0.26	0.0676	39.69
Sum	16.9	16.90	0	0.2590	75.07

It is verified that

$\sum y = 16.90 = \sum \hat{y}$ and $\sum e = \sum (y - \hat{y}) = 0$

The sum of squares of residual is $\sum e^2 = 0.2590$

Clearly it is verified that $\sum e^2 = \sum y^2 - a\sum y - b\sum xy$

As $\sum e^2 = 75.07 - 0.72(16.9) - 1.33(47.1) = 0.2590$

61. Sodium thiosulfate is used by photographers to develop some types of film. The amount of this chemical that will dissolve in water depends on the temperature of the water. The table below gives the numbers of grams of sodium thiosulfate that will dissolve in 100 milliliters of water at various temperatures.

Temperature, x, in degrees Celsius	20	35	50	60	75	90	100
Sodium thiosulfate Dissolved, y, in grams	50	80	120	145	175	205	230

a. Find the linear regression equation for these data.

b. How many grams of sodium thiosulfate does the model predict will dissolve in100 milliliters of water when the temperature of the water is 70°C? Round to the nearest tenth of a gram

Solution:

a. the regression equation is $y = 2.2517731x + 5.2482270$

b. Evaluate the regression equation when x = 70

$y = 2.2517731x + 5.2482270$

$= 2.2517731(70) + 5.2482270 \quad = 162.872344$

Approximately 162.9 grams of sodium thiosulfate will dissolve when the temperature of the water is 70°C.

62. The heights and weights of women swimmers on a college swim team are given in the table below.

Height, x, in inches	68	64	65	67	62	67	65
Weight, y, in pounds	132	108	108	125	102	130	105

a. Find the linear regression equation for these data.

b. Use your regression equation to estimate the weight of a woman swimmer who is 63 inches tall. Round to the nearest pound

Solution:

a. The regression equation is approximately $y = 5.6333x - 252.8667$.

b. When x = 63, we have $y = 5.6333(63) - 252.8667 \approx 102$.

The estimated weight of a woman swimmer who is 63 inches tall is approximately 102 pounds.

63. The table below lists data from an experiment speed and stride length for several adult men.

Stride length (meters)	2.5	3.0	3.3	3.5	3.8	4.0	4.2	4.5
Speed (meters per second)	3.4	4.9	5.5	6.6	7.0	7.7	8.3	8.7

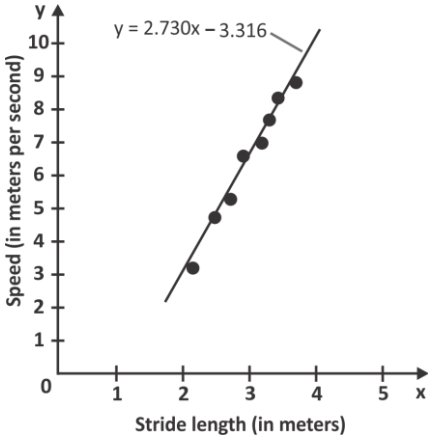
a. Find the equation of the regression line for these data.

b. Use your regression equation to predict the average speeds of adult men with stride lengths of 2.8 meters and 4.8 meters. Round your results to the nearest tenth of a meter per second

Solution:

a. the regression equation is approximately $y = 2.730x - 3.316$

The graph is shown as; you can see that the line fits the data well.



b. Evaluate the regression equation when x = 2.8.

$$y = 2.730(2.8) - 3.316$$
$$= 4.328$$

Rounded to the nearest tenth, the predicted average speed for an adult man with a stride length of 2.8 meters is 4.3 meters per second. Similarly, substituting 4.8 for gives $y = 2.730(4.8) - 3.316 = 9.788$, so 9.8meters per second is the predicted average speed for an adult man with a stride length of 4.8 meters.

64. The table below lists data from an experiment comparing speed and stride length for several camels.

Stride length (meters)	2.5	3.0	3.2	3.4	3.5	3.8	4.0	4.2
Speed (meters per second)	2.3	3.9	4.4	5.0	5.5	6.2	7.1	7.6

a. Find the equation of the regression line for these data.

b. Use your regression equation to predict the average speeds of camels with stride lengths of 2.7 meters and 4.5 meters. Round your results to the nearest tenth of a meter per second

Solution:

a. The equation of the regression line is approximately $y = 3.130x - 5.55$.

b. When x = 2.7, we have $y = 3.130(2.7) - 5.55 \approx 2.9$

and when x = 4.5, we have $y = 3.130(4.5) - 5.55 \approx 8.5$

The predicted average speed of a camel with a stride length of2.7 meters is about 2.9 meters per second, and the predicted average speed for a camel with a stride length of 4.5 meters is approximately 8.5 meters per second.

Correlation: The interdependence between two or more variables is called correlation. It measures the strength or closeness of relationships between two variables.

Linear Correlation Coefficient: The linear correlation coefficient r is a measurement of the interdependence between the variables. It measures the numerical strength or closeness of linear relationships between two variables. It is a measure of how well the regression line fits the given data. If r is positive, then the closer r is to 1, the stronger the linear relationship between the domain and range values and the better the fit of the regression line to the data. If r is negative, then the closer r is to -1, the stronger the linear relationship between the domain and range values and the better the fit of the regression line to the data.

Types of Correlation

Positive or direct correlation: If r is positive, the relationship between the domain and range values has a **positive or direct correlation**. In this case, if the domain value increases, the range value also tends to increase and vice versa. In this case both variables move in the same direction. The value of correlation coefficient for positive correlation is between 0 and 1. i.e. $0 < r < 1$.

Examples

- Relationship between lung cancer and smoking habits.
- Increase in temperature in summer increase the sale of room coolers.
- Here are the two variables moves in the same direction

x	12	15	20	27	30
y	8	10	12	19	25

Negative or inverse correlation: If r is negative, the linear relationship between the domain and range values has a **negative or inverse correlation**. In this case, if the domain value increases, the range value tends to decrease. In this case both variables move in the opposite direction. The value of correlation coefficient for negative correlation is between -1 and 0. i.e. $-1 < r < 0$

Examples

- The volume of gas will decrease as the pressure increases.
- Increase in supply of a commodity decreases its price.
- Here are the two variables moves in the opposite direction

x	15	18	25	30	33
y	40	35	30	25	20

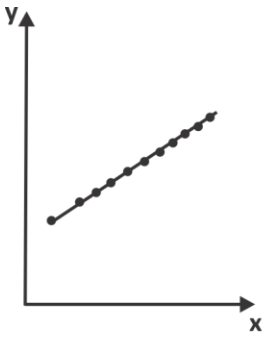
Zero or null correlation: The absence of any relation between the variables is called zero correlation. In this case variables are independent to each other .i.e. $r = 0$.

Examples

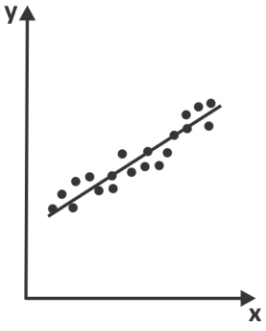
- Amount of rainfall and the head sizes.
- Here are the two variables with no effect to each other.

x	1	2	3	4	5
y	7	7	7	7	7

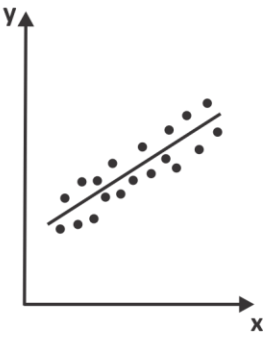
Some Graph of Correlation Types



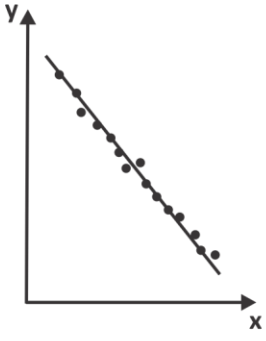
a. Perfect positive correlation, r = 1



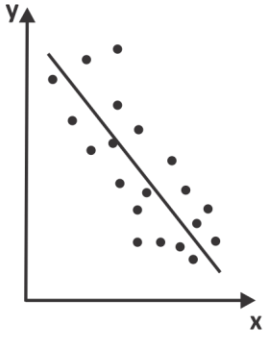
b. Strong positive correlation, r ≈ 0.8



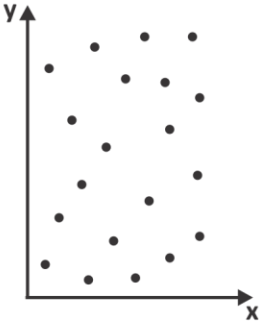
c. Positive correlation r ≈ 0.6



d. Perfect negative r ≈ - 0.9



b. Strong positive r ≈ - 0.5



c. Little or no linear

Correlation Coefficient Formulae

i. $$r = \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}}$$

ii. $$r = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{\left[n\sum X^2 - (\sum X)^2\right]\left[n\sum Y^2 - (\sum Y)^2\right]}}$$

iii. $$r = \frac{n\sum XY - nXY}{\sqrt{\left[\sum X^2 - nX^2\right]\left[\sum Y^2 - nY^2\right]}}$$

iv. $$r = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sqrt{\sum (X - \overline{X})^2 \sum (Y - \overline{Y})^2}}$$

v. $r = \dfrac{\sum XY - nXY}{\sqrt{\left[\sum X^2 - nX^2\right]\left[\sum Y^2 - nY^2\right]}}$

vi. $r = \dfrac{\sum XY - \dfrac{(\sum X)(\sum Y)}{n}}{\sqrt{\left[\sum X^2 - \dfrac{(\sum X)^2}{n}\right]\left[\sum Y^2 - \dfrac{(\sum Y)^2}{n}\right]}}$

Properties of Correlation Coefficient

- The Correlation Coefficient always lies between -1 and $+1$. i.e. $-1 \le r \le +1$
- The Correlation Coefficient is symmetric with respect to the variables X and Y. i.e. $r_{XY} = r_{YX}$

i. The Correlation Coefficient is the geometric mean of the two regression coefficients. i.e. $r_{XY} = \pm\sqrt{b_{XY}.b_{YX}}$

ii. The Correlation Coefficient is a pure number and it has no unit.

iii. For two independent random variables Correlation Coefficient is zero.

iv. The Correlation Coefficient is independent of the origin and unit of measurement. i.e. $r_{XY} = r_{UV}$

65. Calculate the correlation coefficient between percentage of marks scored by 12 students in statistics X and economics Y.

x	30	34	26	49	60	62	65	51	44
y	27	18	34	28	26	30	32	30	28

Solution:

The estimated regression line is $\hat{y} = a + bx$

x	y	x^2	y^2	xy
30	27	900	729	810
34	18	1156	324	612
26	34	676	1156	884
49	28	2401	784	1372
60	26	3600	676	1560
62	30	3844	900	1860
65	32	4225	1024	2080
51	30	2601	900	1530
44	28	1936	784	1232
$\sum x = 421$	$\sum y = 253$	$\sum x^2 = 21339$	$\sum y^2 = 7277$	$\sum xy = 11940$

Correlation coefficient between X and Y is

$$r = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{\left[n\sum X^2 - (\sum X)^2\right]\left[n\sum Y^2 - (\sum Y)^2\right]}} = 0.202$$

66. Find the linear correlation coefficient for the data on stride length versus speed of an adult man.

Stride length (meters)	2.5	3.0	3.3	3.5	3.8	4.0	4.2	4.5
Speed (meters per second)	3.4	4.9	5.5	6.6	7.0	7.7	8.3	8.7

Then find the linear correlation coefficient for the speed data for dogs.

Stride length (meters)	1.5	1.7	2.0	2.4	2.7	3.0	3.2	3.5
Speed (meters per second)	3.7	4.4	4.8	7.1	7.7	9.1	8.8	9.9

Which regression line is a better fit for the corresponding data?

Solution: After entering the data for adult men and finding the linear regression equation, we have the linear correlation coefficient as approximately $r = 0.9937$. Now clear the data for adult men and enter the data for dogs. The regression equation as approximately $y = 3.212x - 1.092.$ and the correlation coefficient as approximately $r = 0.9864.$ Both correlation coefficients are positive, but because the value of for the adult men data is closer to 1 than the value of for the dog data, the regression line for the adult men fits better than the one for the dogs.

67. Find the linear correlation coefficient for stride length versus speed of a camel as given.

Stride length (meters)	2.5	3.0	3.2	3.4	3.5	3.8	4.0	4.2
Speed (meters per second)	2.3	3.9	4.4	5.0	5.5	6.2	7.1	7.6

Round your result to the nearest hundredth.

Solution:

The table below lists data from an experiment comparing speed and stride length for several camels.

Stride length (meters)	2.5	3.0	3.2	3.4	3.5	3.8	4.0	4.2
Speed (meters per second)	2.3	3.9	4.4	5.0	5.5	6.2	7.1	7.6

Here r = 0.998497842, so the linear correlation coefficient is approximately 1.00.

68. Find the correlation coefficient for the data giving the number of persons employed and cloth manufactured in a textile mill.

Person Employed (x)	137	209	113	189	176	200	219
Cloth manufactured (y)	23	47	22	40	39	51	49

Solution:

x	y	x^2	y^2	xy
137	23	18769	529	3151
209	47	43681	2209	9823
113	22	12769	484	2486
189	40	35721	1600	7560
176	39	30976	1521	6864
200	51	40000	2601	10200
219	49	47961	2401	10731
$\sum x = 1243$	$\sum y = 271$	$\sum x^2 = 229877$	$\sum y^2 = 11345$	$\sum xy = 50815$

Correlation Coefficient Formula for n = 7

$$r_{xy} = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]\left[n\sum y^2 - (\sum y)^2\right]}}$$

$$r_{xy} = \frac{7(50815) - (1243)(271)}{\sqrt{\left[7(229877) - (1243)^2\right]\left[7(11345) - (271)^2\right]}}$$

$$r_{xy} = \frac{355705 - 336853}{\sqrt{[64090][5974]}} \Rightarrow r_{xy} = \frac{18852}{19567.16} \Rightarrow r_{xy} = 0.96$$

69. Find the correlation coefficient by using the deviation from their mean for the data giving the height and weight of 8 men.

Height (x)	78	89	97	69	59	79	68	61
Weight (y)	125	137	156	112	107	136	123	106

Solution:

$$\bar{x} = \frac{\sum x}{n} = \frac{600}{8} = 75 \text{ and } \bar{y} = \frac{\sum y}{n} = \frac{1000}{8} = 1255$$

x	y	$x - \bar{x}$	$y - \bar{y}$
78	125	3	0
89	137	11	12
97	156	22	31
69	112	6	-13
59	107	16	-18
79	136	4	11
68	123	-7	-2
61	106	14	-21
Sum 600	1000	0	0

$(x - \bar{x})^2$	$(y - \bar{y})^2$	$(x - \bar{x})(y - \bar{y})$
9	0	0
196	144	168
484	961	682
36	169	78
256	324	288
16	121	44
49	4	14
196	441	294
Sum 1242	2164	1568

$$r = \frac{\sum\left(x-\bar{x}\right)\left(y-\bar{y}\right)}{\sqrt{\sum\left(x-\bar{x}\right)^2 \sum\left(y-\bar{y}\right)^2}} = 0.956$$

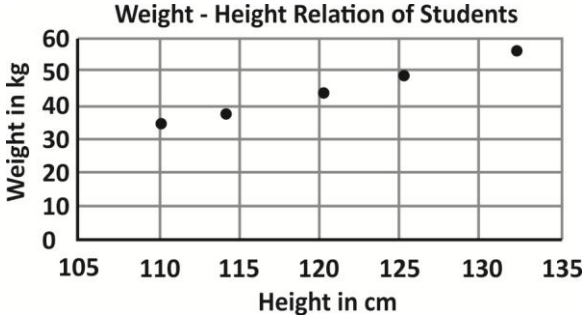
Scatter Plot

A scatter plot is a graph which represents a set of points on the xy - axes. It is a chart type that is normally used to observe and visually display the relationship between variables. Also used for identification of co relational relationships and identification of data patterns

70. A sample of 5 students has the following body weights and heights. The data is represented in a scatter plot graph. Based on the graph, if a student has a weight of 44 kg, how tall is the student?

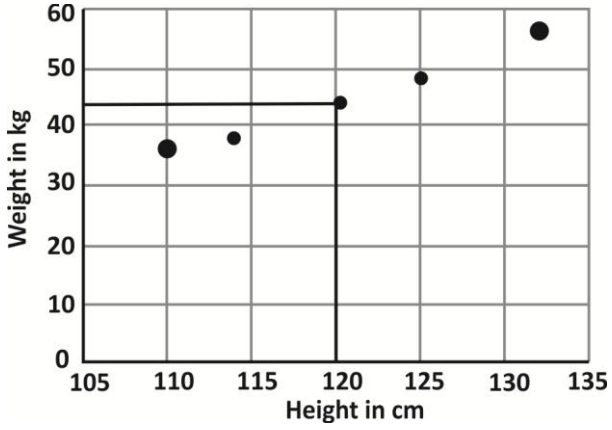
Height in cm	110	114	120	125	132
Weight in kg	35	38	44	49	56

Solution:

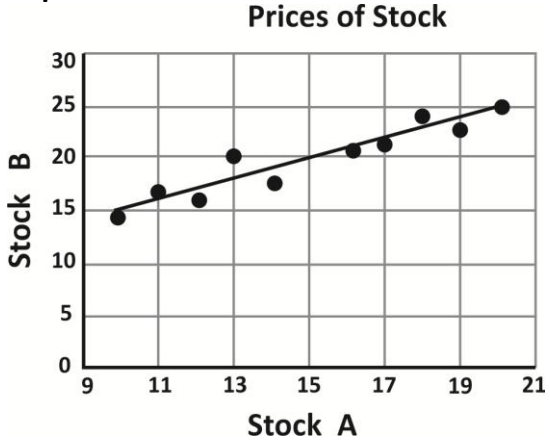


As previously stated, the points are xy points of the given data. The point (110, 35) is the first point, and the (132, 56) is the last point.

For the height of a student who weighs 44 kg (y-coordinate), we need to find the x-coordinate. The point lines up with **120 cm** of height.

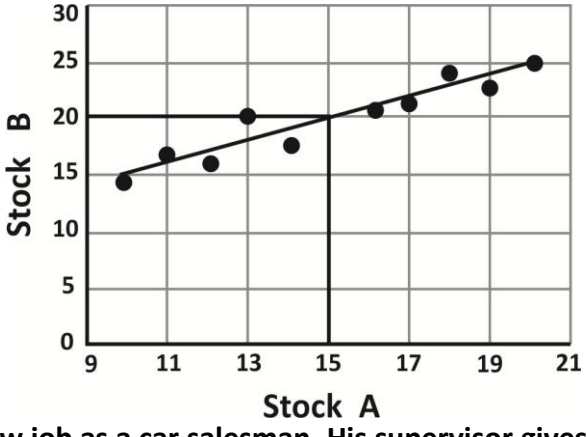


71. A research on two stock companies reveals that the closing prices of stocks were positively correlated to each other. The following chart shows the stock prices of the companies and a line of best fit. Based on the chart, if the price of stock A is $15, what would the price of stock B be?



Solution:

If the price of stock A is $15, the price of stock B is about **$20**.

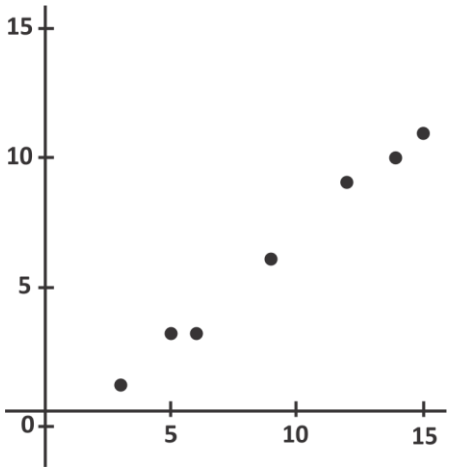


72. Ben started a new job as a car salesman. His supervisor gives him the advice that the more test drives per day he gets his customers to take the more sale she will make per day. He records the following data over the past week.

X(Number of Test Drives Per Day)	Y(Number of Sales Per Day)
3	1
5	3
6	3
9	6
12	9
14	10
15	11

Solution:

When we plot this set of data as a scatter plot we get the following graph.



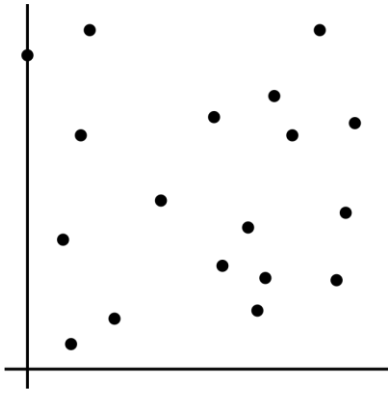
This clearly shows there is a relationship between the two variables. As x increases we see that y also increases. This shows there is what's called a positive linear correlation between the two variables.

Correlation and Scatter Plots

Specifically, linear correlation, is where there appears to be a linear relationship between the two variables. There are three types of correlation:

- Positive Correlation. As x increases y increases.
- Negative Correlation. As x increases y decreases.
- Zero Correlation. There is no relationship.

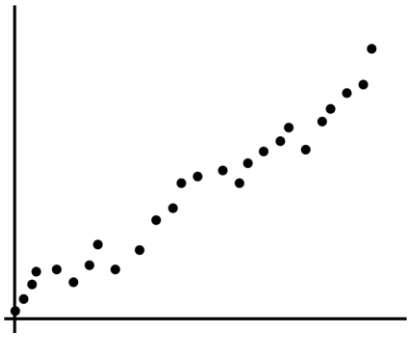
73. Determine if the following scatter plot exhibit positive linear correlation, negative linear correlation, or no linear correlation.



Solution:

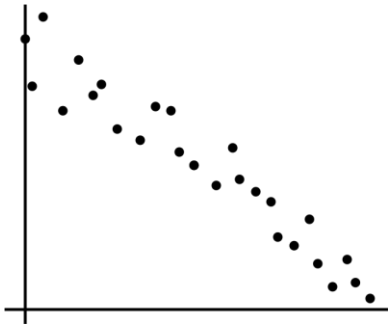
The graph shows no pattern. In this case there is no correlation.

74. Determine if the following scatter plot exhibit positive linear correlation, negative linear correlation, or no linear correlation.



Solution: The graph shows the pattern that as x increases, y increases. This is positive linear correlation.

75. Determine if the following scatter plot exhibit positive linear correlation, negative linear correlation, or no linear correlation.



Solution:

The graph shows the pattern that as x increases, y decreases. This is negative linear correlation.

76. The table below shows the maximum exercise heart rate for specific individuals of various ages who exercise regularly.

Age, x, in years	20	25	30	32	43	55	28	42	50	55	62
Heart rate, y, in maximum beats per minute	160	150	148	145	140	130	155	140	132	125	125

a. Sketch a scatter diagram of the data.

b. Find a linear function that models the data.

c. Use this function to predict the maximum exercise heart rate recommended for a 28-year-old person.

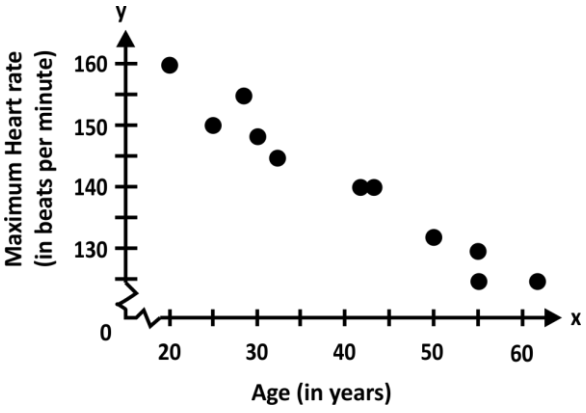
Solution:

The table below shows the maximum exercise heart rate for specific individuals of various ages who exercise regularly.

Age, x, in years	20	25	30	32	43	55	28	42	50	55	62
Heart rate, y, in maximum beats per minute	160	150	148	145	140	130	155	140	132	125	125

The graph as follows, called a **scatter diagr**am, is a graph of the ordered pairs of the table. These ordered pairs suggest that the maximum exercise heart rate for an individual decreases as the person's age increases.

Although these points do not lie on one line, it is possible to find a line that approximately fits the data. One way to do this is to select two data points and then find the equation of the line that passes through the two points. To do this, we first find the slope of the line between the two points and then use the point–slope formula to find the equation of the line. Suppose we choose $(20,160)$ as P_1 and $(62,125)$ as P_2.Then the slope of the line between P_1 and P_2 is



$$m=\frac{y_2-y_1}{x_2-x_1}=\frac{125-160}{62-20}=-\frac{35}{42}=-\frac{5}{6}$$

Now use the point–slope formula.

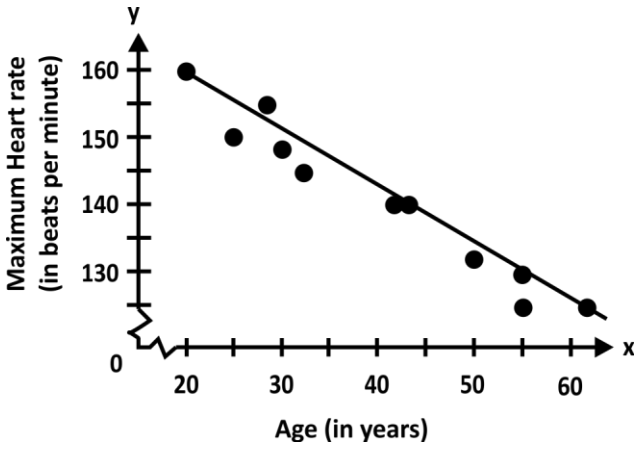
$$y-y_1=m(x-x_1)$$

$$y-160=-\frac{5}{6}(x-20)$$

$$y-160=-\frac{5}{6}x+\frac{50}{3}$$

$$y=-\frac{5}{6}x+\frac{530}{3}$$

The graph of $y=-\frac{5}{6}x+\frac{530}{3}$ is shown as follows.



This line approximates the data and can be used to estimate maximum exercise heart rates for different ages. For example, an exercise physiologist could determine the recommended maximum exercise heart rate for a 28-year-old individual by replacing in the equation by 28and determining the value of y.

$$y = -\frac{5}{6}x + \frac{530}{3} \Rightarrow y = -\frac{5}{6}(28) + \frac{530}{3} \approx 153.3$$

The maximum exercise heart rate recommended for a 28-year-old person is approximately 153 beats per minute.

77. The average prices per pound of white bread in U.S. cities, as recorded in August of various years, are listed in the table below.

Year	1997	1998	1999	2000	2001	2002	2003	2004	23005
Price per pound	0.872	0.869	0.884	0.923	0.991	1.012	0.996	0.996	1.060

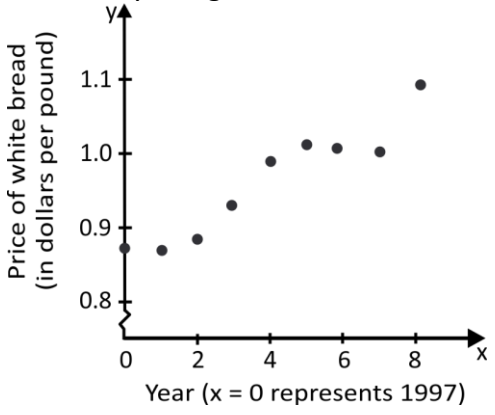
a. Sketch a scatter diagram of the data.

b. Find a linear function that models the data.

c. Use this function to predict the price per pound of white bread in August of 2011.

Solution:

a. We can simplify the data by using 0 for 1997, 1 for 1998, etc.



Looking at the scatter diagram, it appears that a line through the points (3, 0.923) and (6, 0.996) will fit the data points reasonably well.

The slope of the line through these points is

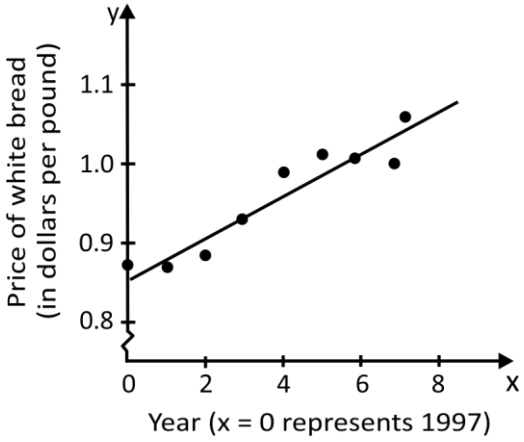
$$m=\frac{y_2-y_1}{x_2-x_1}=\frac{0.996-0.923}{6-3}=\frac{0.073}{3}$$

and the equation of the line is $y-y_1=m(x-x_1)$

$$y-0.923=\frac{0.073}{3}(x-3)$$

$$y-0.923=\frac{0.073}{3}x-0.073$$

$$y=\frac{0.073}{3}x+0.85$$



or approximately $y=0.0243x+0.85$. Thus the price per pound of white bread is approximately $y=0.0243x+0.85$, where is the number of years after August 1997.

The year 2011 corresponds to x = 14, so we evaluate the function when x = 14.

$$f(14)=0.0243(14)+0.85=1.1902$$

The estimated price of white bread in August 2011 is about $1.19 per pound.

78. The populations of a city for various years are given in the table below.

Year	1992	1994	1996	1998	2000	2002	2004	2006
Population (thousands)	20.28	26.31	32.16	37.38	40.11	46.62	49.87	52.91

a. Sketch a scatter diagram of the data.

b. Write an equation for a linear function that models the data.

c. Use the function from part b to predict the city's population in the year 2025.

Solution:

a. The regression equation is approximately $y=5.6333x-252.8667$

b. When x = 63, we have $y=5.6333(63)-252.8667\approx 102$.

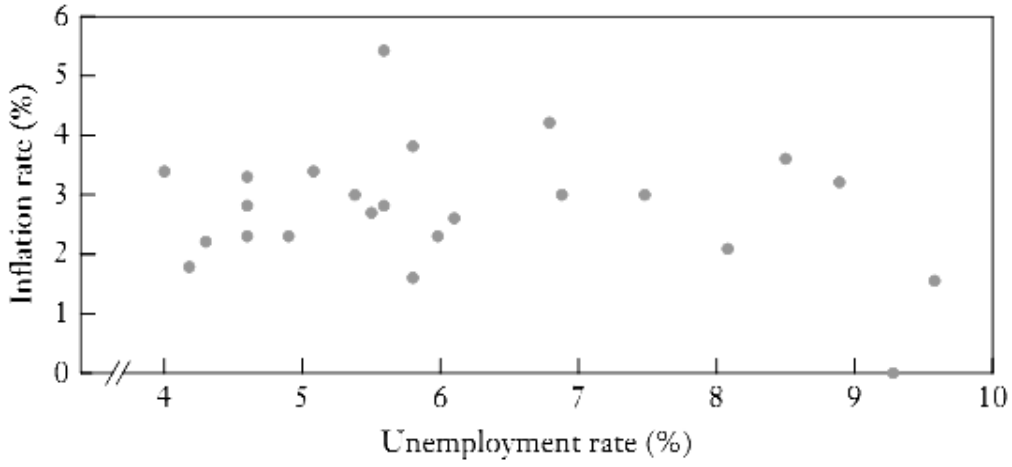
The estimated weight of a woman swimmer who is 63 inches tall is approximately 102 pounds.

79. Inflation and Unemployment: Prior to the 1990s, most economists assumed that the unemployment rate and the inflation rate were negatively correlated. That is, when unemployment goes down, inflation goes up, and vice versa. Table shows unemployment and inflation data for the period 1990–2012. Make a **scatterplot** for these data. Based on your diagram, does it appear that the data support the historical claim of a link between the unemployment and inflation rates?

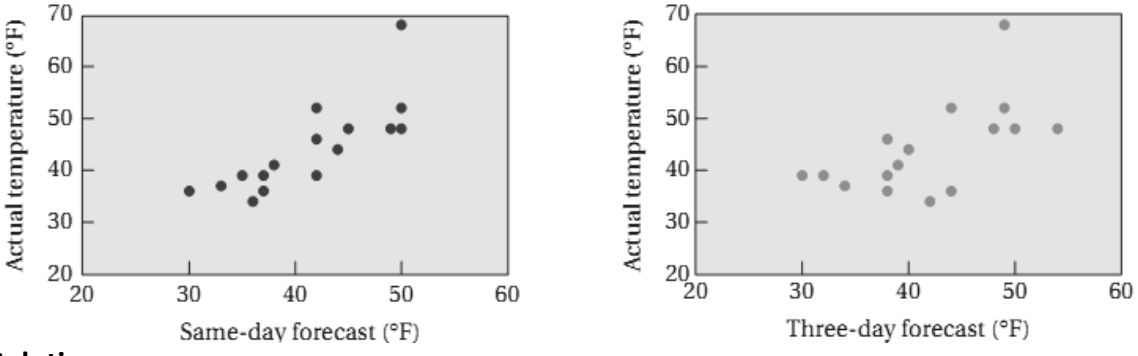
Year	Unemployment Rate (%)	Inflation Rate (%)	Year	Unemployment Rate (%)	Inflation Rate (%)
1990	5.6	5.4	2002	5.8	1.6
1991	6.8	4.2	2003	6.0	2.3
1992	7.5	3.0	2004	5.5	2.7
1993	6.9	3.0	2005	5.1	3.4
1994	6.1	2.6	2006	4.6	3.3
1995	5.6	2.8	2007	4.6	2.8
1996	5.4	3.0	2008	5.8	3.8
1997	4.9	2.3	2009	9.3	0*
1998	4.6	2.3	2010	9.6	1.6
1999	4.3	2.2	2011	8.9	3.2
2000	4.0	3.4	2012	8.1	2.1
2001	4.2	1.8			

Solution:

We make the scatterplot by plotting the variable unemployment rate on the horizontal axis and the variable inflation rate on the vertical axis. To make the graph easy to read, we use values ranging from 3.5% to 10% for the unemployment rate and from 0 to 6% for the inflation rate. Figure 5.40 shows the result. To the eye, there does not appear to be any obvious correlation between the two variables. (A calculation confirms that there is nearly zero correlation.) These data do not support the historical claim of a negative correlation between the unemployment and inflation rates.



80. Accuracy of Weather Forecasts: The scatter plots in Figure show two weeks of data comparing the actual high temperature for the day with the same-day forecast (left diagram) and the three-day forecast (right diagram). Discuss the types of correlation on each diagram.

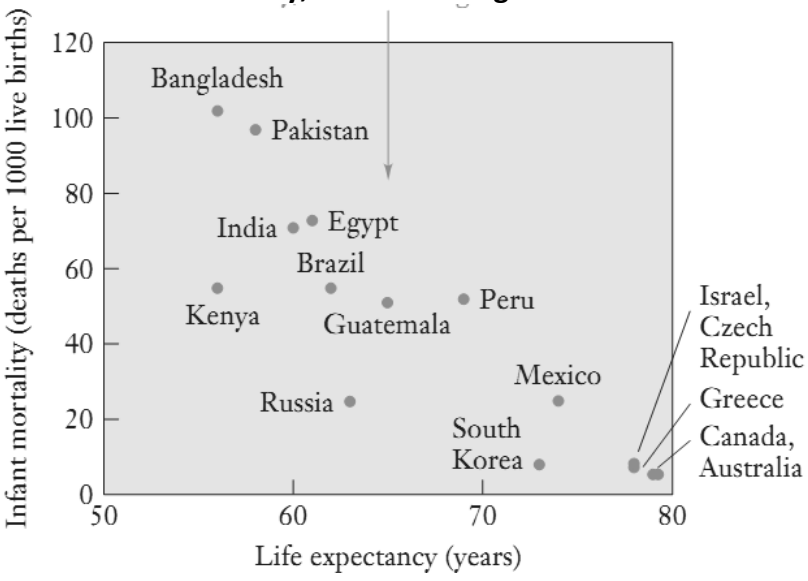


Solution:

Both scatterplots show a general trend in which higher predicted temperatures mean higher actual temperatures. That is, both show positive correlations. However, the points in the left diagram lie more nearly on a straight line, indicating a stronger correlation than in the right diagram. This makes sense, because we expect weather forecasts to be more accurate on the same day than three days in advance.

81. Explanation for a correlation: Consider the correlation between infant mortality and life expectancy in Figure. Which of the three possible explanations for a correlation applies? Explain.

Higher life expectancy generally goes with lover Infant mortality, so this is a negative correlation.



Solution: The negative correlation between infant mortality and life expectancy is probably an example of a common underlying cause. Both variables depend on an underlying variable that we might call *quality of health care*. In countries where health care is better in general, infant mortality is lower and life expectancy is higher.

Variability and Measure of dispersion

Variability is an inherent characteristic of data, referring to the spread or dispersion of individual data points from the central tendency. In other words, it measures how much the data deviates from the average value. Understanding variability is crucial in statistics, as it provides valuable insights into the consistency and reliability of the data. Measures of dispersion, such as range, variance, and standard deviation, are used to quantify variability, enabling researchers and analysts to compare and contrast datasets, identify patterns and trends, and make informed decisions. Variability is an inherent characteristic of data, referring to the spread or dispersion of individual data points from the central tendency. In other words, it measures how much the data deviates from the average value. Understanding variability is crucial in statistics, as it provides valuable insights into the consistency and reliability of the data. Measures of dispersion, such as range, variance, and standard deviation, are used to quantify variability, enabling researchers and analysts to compare and contrast datasets, identify patterns and trends, and make informed decisions.

Dispersion: The degree to which numerical data tend to spread about an average value is called the dispersion or variation of data. **Or** Dispersion measures the spread or variability of data points within a dataset.

Types of Dispersion:

There are two main types of dispersion

- **Absolute Dispersion:** It measures the variation among the values in the same unit of measurement in which the original data are given such as rupees, kg, inches etc. commonly used absolute measures are range, quartiles, mean, standard deviation and variance.
- **Relative Dispersion:** If we compare the dispersion of two dissimilar distributions we need relative term is called relative dispersion. **Or** The ratio between the measure of dispersion and corresponding measure of location is called relative dispersion. These measures are free of unit in which the original data is measure. Commonly used relative measures are coefficient of range, coefficient of quartiles, coefficient of mean and coefficient of variation.

Measure of Central Tendencies

Measures of central tendency, also known as measures of location, are statistical tools used to describe the central or typical value within a dataset. These measures aim to provide a single value that best represents the entire distribution of data, giving an idea of the data's "center" or "average" value. The three most common measures of central tendency are the mean, median, and mode, each providing a unique perspective on the data's central value. By calculating and interpreting these measures, researchers and analysts can gain a better understanding of the data's overall characteristics, identify patterns and trends, and make informed decisions.

Central Tendency and Spread of Data

Central tendency is a statistical measure that identifies the middle or typical value of a dataset or distribution. It aims to provide a single value that best represents the entire dataset. It include mean, median, mode and mid ranges.

Some of the specific measures of center are shown below.

Types of Averages

- **Mathematical Averages**
  - Mean / Arithmetic Mean
  - Geometric Mean
  - Harmonic Mean
- **Positional Averages**
  - Median
  - Mode

Arithmetic Mean / Mean

The Mean is sum of all the values of the observations, divided by the number of observations. The mean is also more commonly just called **average**.

The **Sample mean for ungrouped data** is represented by the symbol $\bar{X}$, which is called 'X-bar': $\bar{X} = \frac{\sum x}{n}$, where n is the number of values in the sample.

The **Mean of Frequency Table/ Sample mean for grouped data** is represented by the symbol $\bar{X}$, which is called 'X-bar': $\bar{X} = \frac{\sum fx}{\sum f}$, where f is the frequency in the sample.

The **Population mean** is represented by the symbol μ, which is called 'mu': $\mu = \frac{\sum x}{N}$, Where N is the number of values in the population.

82. For Bradley's weekly hours at a summer job: 25, 32, 36, 32, 18, 28, 30, 36, 12, 16, 35, 36. Find the mean (average) hours he worked in a week.

Solution:

$$\mu = \frac{\sum X}{N} = \frac{336}{12} = 28$$

Which we will report, according to the round off rule, as 28.0 hours worked in an average week.

83. (Sample): Calculate mean for ungrouped data; 2,3,5,7,4,1.

Solution:

$$\bar{X} = \frac{\sum x}{n} = \frac{22}{6} = 3.6$$

84. (Population): Calculate mean for ungrouped data; 1,4,8,4,3,5,1,2,6,3.

Solution:

$$\mu = \frac{\sum x}{N} = \frac{37}{10} = 3.7$$

85. Calculate mean for grouped data;

Classes	Frequency f	Mid Point (X)	fX
20 –24	5	22	110
25 –29	8	27	216
30 –34	13	32	416
35 –39	22	37	814
40 –44	15	42	630
45 –49	10	47	470
50 –54	8	52	416
	$\sum f = 81$		$\sum fX = 3072$

Solution:

$$\bar{X} = \frac{\sum fX}{\sum f} = \frac{3072}{81} = 37.9$$

86. Calculate mean for grouped data;

Classes	Frequency f	Mid Point (X)	fX
1 – 3	2	2	4
4 – 6	5	5	25
7 – 9	7	8	56
10 – 12	5	11	55
13 – 15	6	14	84
16 – 18	5	17	85
	$\sum f = 30$		$\sum fX = 309$

Solution:

$$\bar{X} = \frac{\sum fX}{\sum f} = \frac{309}{30} = 10.3$$

87. Six friends in a biology class received test grades of 92, 84, 65, 76, 88, and 90. Find the mean of these test scores.

Solution:

$$\bar{X} = \frac{\sum x}{n} = \frac{495}{6} = 82.5$$

88. Estimate the mean of the following frequency distribution.

Class	1-8	9-16	17-24	25-32	33-40
Frequency	3	5	7	2	1

Solution:

The midpoints of the classes (X) are: 4.5; 12.5; 20.5; 28.5; 36.5.

Then the formula would be

$$\frac{\sum fx}{\sum f} = \frac{3(4.5)+5(12.5)+7(20.5)+2(28.5)+1(36.5)}{3+5+7+2+1} = \frac{313}{18} = 17.388$$

89. Compute the overall semester GPA for a college student who earned the following grades.

Cource	Grade	Credits
Physics	B = 3.0	4
English	C = 2.0	3
Math	C = 2.0	3
Study Skills	B = 3.0	1
History	A = 4.0	3

Solution:

Here we are looking for the average grade points, so the data values are grade (quality) points. The credit amounts are the weights.

$$GPA = \frac{\sum fx}{\sum f} = \frac{4(3.0)+3(2.0)+3(3.0)+1(3.0)+3(4.0)}{4+3+3+1+3} = \frac{39}{14} = 2.79$$

90. Find the mean of the data in Table.

Observed Event Number of Cable Television Connections, x	Frequency Number of Households, f, with x Cable Television Connections
0	5
1	12
2	14
3	3
4	2
5	3
6	0
7	1
	40 total

This row indicates that there are 14 households with two cable television connections.

Solution:

The numbers in the right-hand column of Table are the frequencies *f* of the numbers in the first column. The sum of all the frequencies is 40.

$$Mean = \frac{\sum fx}{\sum f} = \frac{4(3.0)+3(2.0)+3(3.0)+1(3.0)+3(4.0)}{4+3+3+1+3} = \frac{39}{14} = 2.79$$

$$Mean = \frac{(0.5)+(1.12)+(2.14)+(3.3)+(4.2)+(5.3)+(6.0)+(7.1)}{40}$$

$$Mean = \frac{79}{40} = 1.975$$

The mean number of cable connections per household for the homes in the sub-division is 1.975.

91. Table shows the number of movies (original and sequels or prequels) in each of five popular science fiction series. Find the average number of films in these series using mean?

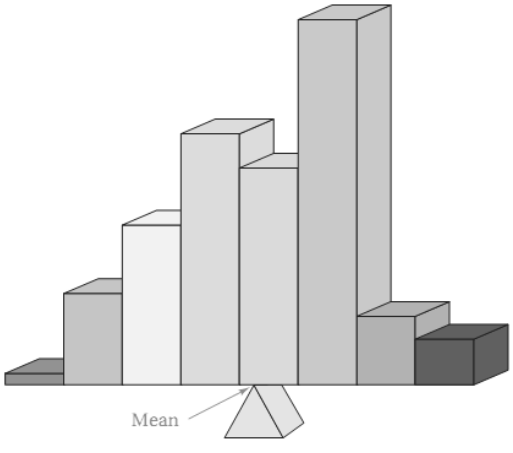
Title	Number of Movies in Series (as of 2013)
Alien	4
Planet of the Apes	7
Star Trek	12
Star Wars	6
Terminator	4

Solution:

We find the mean by dividing the total number of movies by five (because there are five series):

$$Mean = \frac{4+7+12+6+4}{5} = \frac{33}{5} = 6.6$$

In other words, these five series have a mean of 6.6 movies. More generally, we find the mean of any data set by summing all the data values and then dividing by the number of data values. The mean is what most people think of as the average. In essence, it represents the balance point for a data distribution, as shown in Figure.



Outlier: An outlier in a data set is a data value that is much higher or much lower than almost all other values. An outlier can change the mean of a data set but does not affect the median or mode.

92. Calculate effect of outlier on mean for ungrouped data;
245670, 176200, 360280, 272440, 450394, 310160, 393610, 3874480

Solution:

Mean without Outlier

$$= \frac{245670+176200+360280+272440+450394,310160+393610}{7}$$

$$\text{Mean without Outlier} = \frac{2208754}{7} = 315536.29$$

Mean with Outlier

$$= \frac{245670+176200+360280+272440+450394,310160+393610+3874480}{7}$$

$$\text{Mean with Outlier} = \frac{6083234}{7} = 760404.25$$

93. Calculate effect of outlier on mean for ungrouped data;
0.8161, 0.8194, 0.8165, 0.8176, 0.7901, 0.8143, 0.8126

Solution:

$$\text{Mean without Outlier} = \frac{0.8161+0.8194+0.8165+0.8176+0.8143+0.8126}{7}$$

$$\text{Mean without Outlier} = 0.812$$

Mean with Outlier

$$= \frac{0.8161+0.8194+0.8165+0.8176+0.7901+0.8183+0.8126}{7}$$

$$\text{Mean with Outlier} = 0.812$$

Geometric Mean: When we have the data in grades, ratio and percentage form, then we apply geometric mean. Geometric Mean of a variable X is the n^{th} positive root of the product of the $x_1, x_2, x_3, ..., x_n$ observations. In symbol we write; $G.M = \left(x_1 x_2 x_3 ... x_n\right)^{1/n}$

For ungrouped data

Geometric Mean $= anti\log\left(\frac{\sum \log x}{n}\right)$, where n is the number of values in the sample.

For grouped data

Geometric Mean $= anti\log\left(\frac{\sum f \log x}{\sum f}\right)$

94. Calculate geometric mean for ungrouped data; 8,40, 175, 1209, 2000.
Solution:

X	logX
8	0.91
40	1.61
175	2.24
1209	3.08
2000	3.30

Geometric Mean $= anti\log\left(\frac{\sum \log x}{n}\right) = 169.04$

95. Calculate geometric mean for ungrouped data; 3,13,11,15,5,4,2.
Solution:

X	logX
3	0.477
13	1.113
11	1.041
15	1.176
5	0.698
4	0.602
2	0.301

Geometric Mean $= anti\log\left(\frac{\sum \log x}{n}\right) = 5.923$

96. Calculate geometric mean for grouped data;

Classes	Frequency (f)
15 – 19	5
20 – 24	3
25 – 29	8
30 – 34	2
35 – 39	4

Solution:

Classes	Frequency *(f)*	Mid Point *(X)*	Log X	F Log X
15 –19	5	17	1.23	6.15
20 –24	3	22	1.34	4.02
25 –29	8	27	1.43	11.44
30 –34	2	32	1.50	3
35 –39	4	37	1.56	6.24

Geometric Mean $= anti\log\left(\frac{\sum f \log x}{\sum f}\right) = anti\log\left(\frac{30.85}{22}\right) = 25.24$

97. Calculate geometric mean for grouped data;
Solution:

Classes	Frequency *(f)*	Mid Point *(X)*	Log X	*f* Log X
60 – 80	5	70	1.84	9.2
80 - 100	14	90	1.95	27.3
100 - 120	17	110	2.04	34.68
120 - 140	10	130	2.11	21.1
140 - 160	1	150	2.17	2.17
160 - 180	0	170	2.23	0
180 - 200	2	190	2.27	4.54

Geometric Mean $= anti\log\left(\frac{\sum f \log x}{\sum f}\right) = anti\log\left(\frac{166.31}{49}\right) = 2477.88$

Harmonic Mean

Harmonic mean is defined as the reciprocal of mean and the reciprocal of the values. This type of average is also used as grades, ratio and percentage form of data.

- **For ungrouped data:**

  Harmonic Mean $= \frac{n}{\sum\left(\frac{1}{x}\right)}$, where n is sample size.

- **For grouped data:**

  Harmonic Mean $= \frac{f}{\sum\left(\frac{f}{x}\right)}$,

98. Calculate Harmonic mean for ungrouped data; 13.2,14.2,14.8,15.2,16.1.
Solution:

X	1/X
13.2	0.07
14.2	0.07
14.8	0.06
15.2	0.06
16.1	0.06

Harmonic Mean $= \frac{n}{\sum\left(\frac{1}{x}\right)} = \frac{5}{0.32} = 15.62$

99. Calculate Harmonic mean for grouped data;

Solution:

Classes	Frequency *(f)*	Mid Point (X)	(*f*/X)
15 – 19	5	17	0.29
20 – 24	3	22	0.13
25 – 29	2	27	0.07
30 – 34	4	32	0.125
35 – 39	2	37	0.05
	$\sum f = 16$		

Harmonic Mean $= \dfrac{\sum f}{\sum\left(\dfrac{f}{x}\right)} = \dfrac{16}{0.66} = 24.24$

Median

It is the middle most value of the observations arranged in ascending and descending order of magnitude. To find the median, first arrange the data in increasing order. If there are an odd number of observations, the median is the middle value in order. If there is an even number of observations, the median is the average of two middle values in order

Example

- **Odd data:** 1,2,**3**,4,5 here median is 3
- **Even data:** 1,2,**3**,**4**,5,6 here median is $\dfrac{3+4}{2} = 3.5$

Formulae

- **For Ungroup Data:**

$$M = \left(\frac{n+1}{2}\right)^{th}$$ for Odd data

$$M = \frac{1}{2}\left[\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2}+1\right)^{th}\right] = \frac{1}{2}\left[\left(\frac{n}{2}\right)^{th} + \left(\frac{n+2}{2}\right)^{th}\right]$$ for Even data

- **For Group Data:**

$$M = L + \frac{H}{f}\left(\frac{\sum f}{2} - C\right) = L + \frac{H}{f}\left(\frac{n}{2} - C\right)$$ we may write $\sum f = n$

L= Lower class boundary of median class

H= Class interval

F= Frequency of median class

C= Cumulative Preceding frequency

100.Find the median of Bradley's summer weekly hours 25, 32, 36, 32, 18, 28, 30, 36, 12, 16, 35, 36.

Solution:

First we must put the values in order from lowest to highest: 12, 16, 18, 25, 28, 30, 32, 32, 35, 36, 36, 36. There are two values in the middle, 30 and 32, with five values below and above. The average of the two middle values is 31, which is the median.

101.Find the median of 45,32,21,65,36,53,48,76,27.

Solution:

$$M = \left(\frac{n+1}{2}\right)^{th} = \left(\frac{9+1}{2}\right)^{th} = \left(\frac{10}{2}\right)^{th} = 5^{th}\ term = 45$$

102.Find the median of 45,32,21,65,36,53,48,27.

Solution:

$$M = \frac{1}{2}\left[\left(\frac{8}{2}\right)^{th} + \left(\frac{8}{2}+1\right)^{th}\right] = \frac{1}{2}\left[4^{th} + 5^{th}\right] = \frac{36+45}{2} = 40.5$$

103.Calculate Median for grouped data;

Solution:

Classes	Frequency *(f)*	Class Boundaries	Cumulative Frequency
10 – 14	5	9.5 – 14.5	5
15 – 19	12	14.5 – 19.5	5 + 12 = **17**
20 – 24	**30**	**19.5 – 24.5**	**17 + 30 = 47**
25 – 29	25	24.5 – 29.5	47 + 25 = 72
30 – 34	6	29.5 – 34.5	72 + 6 = 78
$\sum f = 78$			

$$M = L + \frac{H}{f}\left(\frac{\sum f}{2} - C\right) = 19.5 + \frac{5}{30}\left(\frac{78}{2} - 17\right) = 385.5$$

104. Find the median for the data in each of the following lists.

a. 4, 8, 1, 14, 9, 21, 12

b. 46, 23, 92, 89, 77, 108

Solution:

a. Ranking the numbers from smallest to largest gives 1, 4, 8, 9, 12, 14, 21. The middle number is 9. Thus 9 is the median.

b. Ranking the numbers from smallest to largest gives 23, 46, 77, 89, 92, 108. The two middle numbers are 77 and 89. The mean of 77 and 89 is 83. Thus 83 is the median of the data.

105. The median of the ranked list 3, 4, 7, 11, 17, 29, 37 is 11. If the maximum value 37 is increased to 55, what effect will this have on the median?

Solution:

The median will remain the same because 11 will still be the middle number in the ranked list.

106.Table shows the number of movies (original and sequels or prequels) in each of five popular science fiction series. Find the average number of films in these series using median?

Title	Number of Movies in Series (as of 2013)
Alien	4
Planet of the Apes	7
Star Trek	12
Star Wars	6
Terminator	4

Solution:

We could describe the average number of films in these five series by computing the median, or middle value, of the data set. To find a median, we arrange the data values in ascending (or descending) order, repeating data values that appear more than once. If the number of values is odd, there is exactly one value in the middle of the list, and this value is the median. If the number of values is even, there are two values in the middle of the list, and the median is the number that lies halfway between them. Putting the data in Table in ascending order gives the list **4, 4, 6, 7, 12**. The median number of movies is **6** because 6 is the middle number in the list.

Mode

The French word Mode mean 'Fashion' has been adopted to convey the idea of most frequent. The Mode is the value that has the most number of observations (frequency), but must occur more than once. Most occurring value of the data set is called mode.

Remember for Ungrouped Data

- A distribution having a single mode called **unimodel** mode. e.g. 2,1,7,2,6,4,2 has a unimodel mode that is 2.
- A distribution having two modes called **bimodel** mode. e.g. 2,5,3,2,4,3,2,3 has a bimodel mode that is 2 and 3.
- A distribution having more than two modes called **multimodel** mode. e.g. 4,1,2,3,6,4,5,2,3,4,2,1,3 has a multimodel mode that is 2,3 and 4.

107.Find the mode of Bradley's summer weekly hours 25, 32, 36, 32, 18, 28, 30, 36, 12, 16, 35, 36.

Solution:

Having the values in order makes this easier: 12, 16, 18, 25, 28, 30, 32, 32, 35, 36, 36, 36. The values that occurs the most often is 36, which is the only mode here.

Remember for Grouped Data

$$\textbf{Mode} = L + \frac{(f_m - f_1)}{(f_m - f_1) + (f_m - f_2)} \times H$$

L= Lower class boundary of maximum frequency

H= Class interval

f_m= Maximum frequency

f_1= Preceding frequency of Maximum frequency

f_2= Following frequency of Maximum frequency

108.Calculate Mode for grouped data;

Solution:

Classes	Frequency *(F)*	Class Boundaries
15 – 19	5	14.5 – 19.5
20 – 24	**10**	**19.5 – 24.5**
25 – 29	2	24.5 – 29.5
30 – 34	4	29.5 – 34.5
35 – 39	3	34.5 – 39.5

$$\textbf{Mode} = L + \frac{(f_m - f_1)}{(f_m - f_1) + (f_m - f_2)} \times H = 19.5 + \frac{(10-5)}{(10-5)+(10-2)} \times 5 = 9.42$$

109. Find the mode of the data in each of the following lists.

a. 18, 15, 21, 16, 15, 14, 15, 21 **b. 2, 5, 8, 9, 11, 4, 7, 23**

Solution:

a. In the list 18, 15, 21, 16, 15, 14, 15, 21, the number 15 occurs more often than the other numbers. Thus 15 is the mode.

b. Each number in the list 2, 5, 8, 9, 11, 4, 7, 23 occurs only once. Because no number occurs more often than the others, there is no mode.

110. Table shows the number of movies (original and sequels or prequels) in each of five popular science fiction series. Find the average number of films in these series using mode?

Title	Number of Movies in Series (as of 2013)
Alien	4
Planet of the Apes	7
Star Trek	12
Star Wars	6
Terminator	4

Solution:

In the case of the movies, the mode is 4 because this value occurs twice in the data set, while the other values occur once.

111. Eight grocery stores sell the PR energy bar for the following prices: $1.09, $1.29, $1.29, $1.35, $1.39, $1.49, $1.59, $1.79. Find the mean, median, and mode for these prices.

Solution:

The mean price is $1.41:

$$Mean = \frac{\begin{pmatrix} \$1.09 + \$1.29 + S1.29 + \$1.35 \\ +\$1.39 + \$1.49 + \$1.59 + \$1.79 \end{pmatrix}}{8} = \$1.41$$

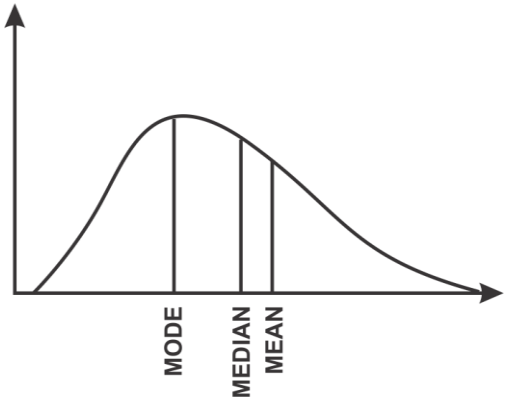
To find the median, we first sort the data in ascending order:
$1.09, $1.29, $1.29, $1.35, $1.39, $1.49, $1.59, $1.79.
Because there are eight prices (an even number), there are two values in the middle of the list: $1.35 and $1.39. Therefore, the median lies halfway between these two values, which we calculate by adding them and dividing by 2:

$$\text{Meadian } = \frac{\$1.35 + \$1.39}{2} = \$1.37$$

The mode is $1.29, because this is the only price that occurs more than once.

Empirical Relation between Mean, Median and Mode

In a single – peaked frequency distribution, the values of the mean, median and mode coincide if the frequency distribution is absolutely symmetrical. But if these values differ, the frequency distribution is said to be skewed or asymmetrical.



Experience tells us that in a unimodal curve of moderate skewness, the median is usually sandwiched between the mean and the mode and between them the following approximate relation hold good.

Mean – Mode = 3 (Mean – Median)

Or Mode = 3Median – 2Mean

This empirical relation does not hold in case of a J – Shaped or an extremely skewed distribution.

Comparative Properties of the Mean, the Median, and the Mode

The mean of a set of data:

- Is the most sensitive of the averages. A change in any of the numbers changes the mean.
- Can be different from each of the numbers in the set.
- Can be changed drastically by changing an extreme value.

The median of a set of data:

- Is usually not changed by changing an extreme value.
- Is generally easy to compute.

The mode of a set of data:

- may not exist, and when it does exist it may not be unique.
- Is one of the numbers in the set, provided a mode exists.
- Is generally not changed by changing an extreme value.
- Is generally easy to compute.

Note:

In the following example, we compare the mean, the median, and the mode of the salaries of five employees of a small company.

Salaries: $370,000 $60,000 $32,000 $16,000 $16,000

The sum of the five salaries is $494,000. Hence the mean is
The median is the middle number, $32,000. Because the $16,000 salary occurs the most, the mode is $16,000. The data contain one extreme value that is much larger than the other values. This extreme value makes the mean considerably larger than the median. Most of the employees of this company would probably agree that the median of $32,000 better represents the average of the salaries than does either the mean or the mode. $\frac{\$494,000}{5} = \$98,800$

112.A track coach wants to determine an appropriate heart rate for her athletes during their workouts. She chooses five of her best runners and asks them to wear heart rate monitors during a workout. In the middle of the workout, she reads the following heart rates for the five athletes: 130, 135, 140, 145, 325. Which is a better measure of the average in this case: the mean or the median? Why?

Solution:

Four of the five values are fairly close together and seem reasonable for mid-workout heart rates. The high value of 325 is an outlier. This outlier seems likely to be a mistake (perhaps caused by a faulty heart monitor), because anyone with such a high heart rate should be in cardiac arrest. If the coach uses the mean as the average, she will be including this outlier—which means she will be including any mistake made when it was recorded. If she uses the median as the average, she'll have a more reasonable value, because the median won't be affected by the outlier.

"average" confusion

The different meanings of "average" can lead to confusion. Sometimes this confusion arises because we are not told whether the "average" is the mean or the median, and other times because we are not given enough information about how the average was computed. The following two examples illustrate such situations.

113.A newspaper surveys wages for assembly workers in regional high-tech companies and reports an average of $22 per hour. The workers at one large firm immediately request a pay raise, claiming that they work as hard as employees at other companies but their average wage is only $19. The management rejects their request, telling them that they are overpaid because their average wage, in fact, is $23. Can both sides be right? Explain.

Solution:

Both sides can be right if they are using different definitions of average. In this case, the workers may be using the median while management uses the mean. For example, imagine that there are only five workers at the firm and their wages are $19, $19, $19, $19, and $39. The median of these five wages is $19 (as the workers claimed), but the mean is $23 (as management claimed).

114.All 100 first-year students at a small college take three courses in the Core Studies Program. The first two courses are taught in large lectures, with all 100 students in a single class. The third course is taught in ten classes of 10 students each. Students and administrators get into an argument about whether classes are too large. The students claim that the mean size of their Core Studies classes is 70. The administrators claim that the mean class size is only 25 students. Can both sides be right? Explain.

Solution:

Both sides are right, but they are using different means. The students have calculated the mean size of the classes in which each student is personally enrolled. Each student is taking two classes with enrollments of 100 each and one class with an enrollment of 10, so the mean size of each student's classes is

$$\frac{total\ enrollment\ in\ student's\ classes}{number\ of\ classes\ student\ is\ taking} = \frac{100+100+10}{3} = 70$$

The administrators have calculated the mean enrollment in all classes. There are two classes with 100 students each and ten classes with 10 students each, making a total enrollment of 300 students in 12 classes. The mean enrollment per class is

$$\frac{total\ enrollment}{number\ of\ classes} = \frac{300}{12} = 25$$

The two claims about the mean are both correct, but very different, because the students and administrators used different means. The students calculated the mean class size per student, while the administrators calculated the mean number of students per class.

Range

Measures of Variation (or measures of spread) are descriptive measures that indicate how much variation is in the data or how spreads out the data values are from each other.

- The Minimum (Min) is the lowest value in the data set.
- The Maximum (Max) is the highest value in the data set.

The **Range** is the difference between Minimum and Maximum values. It is the difference between largest and smallest values of a data set.

Range $= X_{\max} - X_{\min}$

Coefficient of Range $= \dfrac{X_{\max} - X_{\min}}{X_{\max} + X_{\min}}$

115.Find the min, max, and range of Bradley's summer weekly hours.

Solution:

Having the values in order makes this easier as well: 12, 16, 18, 25, 28, 30, 32, 32, 35, 36, 36, 36. Here min = 12, max = 36, and Range = 36 -12 = 24 hours. We could also say the range is from 12 to 36 hours.

116.Calculate range for 2,12,15,4,7,10,13.

Solution:

$$\text{Range} = X_{\max} - X_{\min} = 15 - 2 = 13$$

$$\text{Coefficient of Range} = \frac{X_{\max} - X_{\min}}{X_{\max} + X_{\min}} = \frac{15-2}{15+2} = \frac{13}{17} = 0.76$$

117.Calculate Range for grouped data;

Solution:

Classes	Frequency *(f)*	Class Boundaries
15 – 19	5	**14.5** – 19.5
20 – 24	3	19.5 – 24.5
25 – 29	2	24.5 – 29.5
30 – 34	4	29.5 – 34.5
35 – 39	2	34.5 – **39.5**

$$\text{Range} = X_{\max} - X_{\min} = 39.5 - 14.5 = 25$$

$$\text{Coefficient of Range} = \frac{39.5-14.5}{39.5+14.5} = \frac{15-2}{15+2} = \frac{25}{54} = 0.46$$

118.Consider the following two sets of quiz scores for nine students. Which set has the greater range? Would you also say that the scores in this set are more varied?

Quiz 1: 1 10 10 10 10 10 10 10 10

Quiz 2: 2 3 4 5 6 7 8 9 10

Solution:

The range for Quiz 1 is 10 - 1 = 9 points, which is greater than the range for Quiz 2 of 10 - 2 = 8 points. However, aside from a single low score (an outlier), Quiz 1 has no variation at all because every other student got a 10. In contrast, no two students got the same score on Quiz 2, and the scores are spread throughout the list of possible scores. The scores on Quiz 2 are more varied even though Quiz 1 has the greater range.

Percentiles/ pth Percentiles

Percentiles are the values of the variate that divide a set of data into one hundred equal parts after arranging the observations in ascending order of magnitude. **Or** A value x is called the p^{th} percentile of a data set provided p% of the data values are less than x.

For ungrouped data

Position of $P_1 = \left(\frac{n+1}{100}\right)^{th}$ value and Position of $P_i = \left[\frac{i(n+1)}{100}\right]^{th}$ value

$P_i = \text{Positional value} + \text{Decimal Part}(\text{difference})$

For grouped data

$P_1 = l + \frac{h}{f}\left(\frac{n}{100} - C\right)$ and $P_i = l + \frac{h}{f}\left(\frac{in}{100} - C\right)$

Where

$l =$ Lower class boundary of the model class

$h =$ Size of class interval of the model class

$f =$ Frequency of the model class

$n =$ Sum of the frequencies

$C =$ Cumulative frequency of the preceding class of the model class

119.Obtain D65 and D94 from the following data;

Class	Frequency	Class	Frequency
110 – 119	2	160 – 169	18
120 – 129	4	170 – 179	13
130 – 139	17	180 – 189	6
140 – 149	28	190 – 199	5
150 – 159	25	200 – 209	2

Solution:

Class	Frequency	C.B	C.F
110 – 119	2	109.5 – 119.5	2
120 – 129	4	119.5 – 129.5	6
130 – 139	17	129.5 – 139.5	23
140 – 149	28	139.5 – 149.5	51
150 – 159	25	149.5 – 159.5	76
160 – 169	18	159.5 – 169.5	94
170 – 179	13	169.5 – 179.5	107
180 – 189	6	179.5 – 189.5	113
190 – 199	5	189.5 – 199.5	118
200 – 209	2	199.5 – 209.5	120
Sum	120		

65th Percentile (P_{65})

$l = 159.5,\ f = 18, h = 10, C = 76$

$$P_{65} = l + \frac{h}{f}\left(\frac{65n}{100} - C\right) = 160.61$$

94th Percentile (P_{94})

$l = 179.5,\ f = 6, h = 10, C = 107$

$$P_{94} = l + \frac{h}{f}\left(\frac{94n}{100} - C\right) = 189.17$$

120. According to the U.S. Department of Labor, the median annual salary in 2003 for a physical therapist was $57,720. If the 85th percentile for the annual salary of a physical therapist was $71,500, find the percent of physical therapists whose annual salaries were

a. more than $57,720. **b. less than $71,500.**

c. between $57,720 and $71,500.

Solution:

a. By definition, the median is the 50th percentile. Therefore, 50% of the physical therapists earned more than $57,720 per year.

b. Because $71,500 is the 85th percentile, 85% of all physical therapists made less than $71,500.

c. From parts a and b, $85\% - 50\% = 35\%$ of the physical therapists earned between $57,720 and $71,500.

121. According to the U.S. Department of Labor, the median annual salary in 2003 for a police dispatcher was $28,288. If the 30th percentile for the annual salary of a police dispatcher was $25,640, find the percent of police dispatchers whose annual salaries were

a. less than $28,288. **b. more than $25,640.**

c. between $25,640 and $28,288.

Solution:

a. By definition, the median is the 50th percentile. Therefore, 50% of the police dispatchers earned less than $28,288 per year.

b. Because $25,640 is the 30th percentile, $100\% - 30\% = 70\%$ of all police dispatchers made more than $25,640.

c. From parts a and b, $50\% - 30\% = 20\%$ of the police dispatchers earned between $25,640 and $28,288.

Percentile for a Given Data Value

Given a set of data and a data value x,

$$\text{Percentile of score x} = \frac{\text{number of data values less than x}}{\text{total number of data values}}.100$$

122.On a reading examination given to 900 students, Elaine's score of 602 was higher than the scores of 576 of the students who took the examination. What is the percentile for Elaine's score?

Solution:

$$\text{Percentile} = \frac{\text{number of data values less than } 602}{\text{total number of data values}}.100 = \frac{576}{9000}.100 = 64$$

Elaine's score of 602 places her at the 64th percentile.

123. On an examination given to 8600 students, Hal's score of 405 was higher than the scores of 3952 of the students who took the examination. What is the percentile for Hal's score?

Solution:

$$\text{Percentile} = \frac{\text{number of data values less than } 405}{\text{total number of data values}}.100 = \frac{3952}{8600}.100 = 46$$

Hal's score of 405 places him at the 46th percentile.

Deciles

These are the values, which divide the set of observations into ten equal parts after arranging the observations in ascending order of magnitude.

Formulae for deciles

For ungrouped data

$$\text{Position of } D_1 = \left(\frac{n+1}{10}\right)^{th} \text{ value}$$

$$\text{Position of } D_2 = \left[\frac{2(n+1)}{10}\right]^{th} \text{ value}$$

$$\text{Position of } D_3 = \left[\frac{3(n+1)}{10}\right]^{th} \text{ value}$$

$$\text{Position of } D_5 = \left[\frac{5(n+1)}{10}\right]^{th} \text{ value}$$

$$\text{Position of } D_9 = \left[\frac{9(n+1)}{10}\right]^{th} \text{ value}$$

$$D_i = \text{Positional value} + \text{Decimal Part}(\text{difference})$$

124.Find 4th and 7thdeciles from the following data;

17,22,27,29,38,40,42,45,50,54,56,57,60

Solution:

4thdecile (D_4)

$$\text{Position of } D_4 = \left[\frac{4(n+1)}{10}\right]^{th} \text{ value} = \left[\frac{4(13+1)}{10}\right]^{th} \text{ value} = (5.6)^{th} \text{ value}$$

$$D_4 = \text{Positional value} + \text{Decimal Part}(\text{difference})$$

$$D_4 = 5^{th} + 0.6(6^{th} \text{ value} - 5^{th} \text{ value})$$

$$D_4 = 38 + 0.6(40 - 38) = 39.2$$

7th decile (D_7)

$$\text{Position of } D_7 = \left[\frac{7(n+1)}{10}\right]^{th} \text{ value} = \left[\frac{7(13+1)}{10}\right]^{th} \text{ value} = (9.8)^{th} \text{ value}$$

$$D_7 = \text{Positional value} + \text{Decimal Part}(\text{difference})$$

$$D_7 = 9^{th} + 0.8(10^{th} \text{ value} - 9^{th} \text{ value})$$

$$D_4 = 50 + 0.8(54 - 50) = 53.2$$

For grouped data

$$D_1 = l + \frac{h}{f}\left(\frac{n}{10} - C\right) \text{ and } D_i = l + \frac{h}{f}\left(\frac{in}{10} - C\right)$$

Where

$l =$ Lower class boundary of the model class

$h =$ Size of class interval of the model class

$f =$ Frequency of the model class

$n =$ Sum of the frequencies

$C =$ Cumulative frequency of the preceding class of the model class

125.Obtain D_3 and D_7 from the following data;

126.

Class	Frequency	Class	Frequency
110 – 119	2	160 – 169	18
120 – 129	4	170 – 179	13
130 – 139	17	180 – 189	6
140 – 149	28	190 – 199	5
150 – 159	25	200 – 209	2

Solution:

Class	Frequency	C.B	C.F
110 – 119	2	109.5 – 119.5	2
120 – 129	4	119.5 – 129.5	6
130 – 139	17	129.5 – 139.5	23
140 – 149	28	139.5 – 149.5	51
150 – 159	25	149.5 – 159.5	76
160 – 169	18	159.5 – 169.5	94
170 – 179	13	169.5 – 179.5	107
180 – 189	6	179.5 – 189.5	113
190 – 199	5	189.5 – 199.5	118
200 – 209	2	199.5 – 209.5	120
Sum	120		

3rd Decile

$$l = 139.5,\ f = 28, h = 10, C = 23$$

$$D_3 = l + \frac{h}{f}\left(\frac{3n}{10} - C\right) = 144.14$$

7rd Decile

$$l = 159.5,\ f = 18, h = 10, C = 76$$

$$D_7 = l + \frac{h}{f}\left(\frac{7n}{10} - C\right) = 163.94$$

Quartiles

Quartiles are the values of the variate that divide a set of data into four equal parts after arranging the observations in ascending order of magnitude.

Quartile Deviation/Semi Inter – Quartile Range

It is also called semi inter – quartile range. The SIQR is the measure of dispersion defined by the difference between third quartile and the first quartile and half of the range is called quartile deviation. The QD is also an absolute measure of dispersion. Its relative measure called coefficient of quartile deviation.

First Quartile

The **lower quartile (or first quartile)** divides the lowest fourth of a data set from the upper three-fourths. It is the median of the data values in the lower half of a data set. (Exclude the middle value in the data set if the number of data points is odd.)

Second Quartile

The **middle quartile (or second quartile)** is the overall median.

Third Quartile

The upper quartile (or third quartile) divides the lowest three-fourths of a data set from the upper fourth. It is the median of the data values in the upper half of a data set. (Exclude the middle value in the data set if the number of data points is odd.)

Formulae

$$Q.D = \frac{Q_3 - Q_1}{2} \quad \text{Where } Q_3 = \left(\frac{3(n+1)}{4}\right)^{th} \quad \text{and } Q_1 = \left(\frac{n+1}{4}\right)^{th}$$

Coefficient of Q. $D = \frac{Q_3 - Q_1}{Q_3 + Q_1}$

127. Find QD of 45,32,21,65,36,53,48,76,27.

Solution:

Arrange: 21,27,32,36,45,48,53,65,76

$$Q_1 = \left(\frac{n+1}{4}\right)^{th} = \left(\frac{9+1}{4}\right)^{th} = \left(\frac{10}{4}\right)^{th} = (2.5)^{th} = 27.5$$

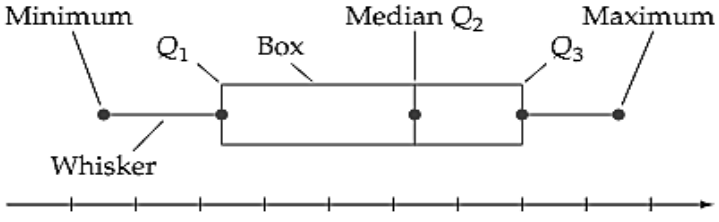
$$Q_3 = \left(\frac{3(n+1)}{4}\right)^{th} = \left(\frac{3(9+1)}{4}\right)^{th} \left(\frac{3(10)}{4}\right)^{th} = \left(\frac{30}{4}\right)^{th} = (7.5)^{th} = 53.5$$

$$Q.D = \frac{Q_3 - Q_1}{2} = \frac{53.5 - 27.5}{2} = \frac{26}{2} = 13$$

Coefficient of Q $D = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{53.5 - 27.5}{53.5 + 27.5} = \frac{26}{81} = 0.32$

Box-and-Whisker Plots

A box-and-whisker plot (sometimes called a **box plot**) is often used to provide a visual summary of a set of data. A box-and-whisker plot shows the median, the first and third quartiles, and the minimum and maximum values of a data set. Seethe figure below.



Construction of a Box-and-Whisker Plot

1. Draw a horizontal scale that extends from the minimum data value to the maximum data value.
2. Above the scale, draw a rectangle (box) with its left side at $Q_1$and its right side at Q_3
3. Draw a vertical line segment across the rectangle at the median, Q_2
4. Draw a horizontal line segment, called a whisker, that extends fromQ_1 to the minimum and another whisker that extends from Q_3 to the maximum.

Importance of a Box-and-Whisker Plot

Box plots have become popular because they are easy to construct and they illustrate several important features of a data set in a simple diagram. That is we can easily estimate

- The quartiles of the data.
- The range of the data.
- The position of the middle half of the data as shown by the length of the box.

128.The following table lists the calories per 100 milliliters of 25 popular beers. Find the quartiles of the data and Construct a box-and-whisker plot for the data set.

Calories, per 100 Milliliters, of Selected Beer

43	37	42	40	53	62	36	32	50	49
26	53	73	48	45	39	45	48	40	56
41	36	58	42	39					

NOTE: It is important to start a box plot with a scaled number line. Otherwise the box plot may not be useful.

Solution:

Step 1: Rank the data, as shown in the following table.

1) 26	**2)** 32	**3)** 36	**4)** 36	**5)** 37	**6)** 39	**7)** 39	**8)** 40	**9)** 40
10) 41	**11)** 42	**12)** 42	**13)** 43	**14)** 45	**15)** 45	**16)** 48	**17)** 48	**18)** 49
19) 50	**20)** 53	**21)** 53	**22)** 56	**23)** 58	**24)** 62	**25)** 73		

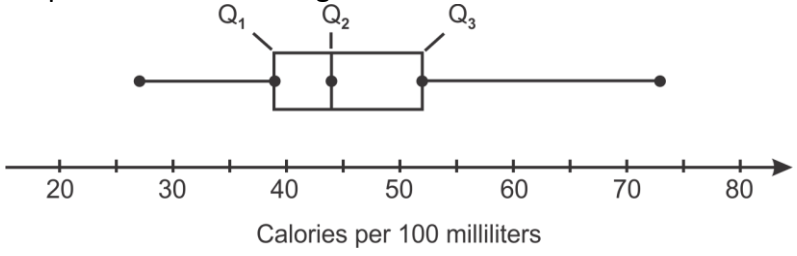
Step 2: The median of these 25 data values has a rank of 13. Thus the median is 43. The second quartile Q_2 is the median of the data, so $Q_2 = 43$

Step 3: There are 12 data values less than the median and 12 data values greater than the median. The first quartile is the median of the data values less than the median. Thus Q_1 is the mean of the data values with ranks of 6 and 7.

$$Q_1 = \frac{39+39}{2} = 39$$

The third quartile is the median of the data values greater than the median. Thus Q_3 is the mean of the data values with ranks of 19 and 20. $Q_3 = \frac{50+53}{2} = 51.5$

For the data set, we determined that $Q_1 = 39, Q_2 = 43$ and $Q_3 = 51.5$. The minimum data value for the data set is 26 and the maximum data value is 73. Thus the box-and-whisker plot is as shown in Figure.



129.Graph a box-and-whisker plot for the data values shown.

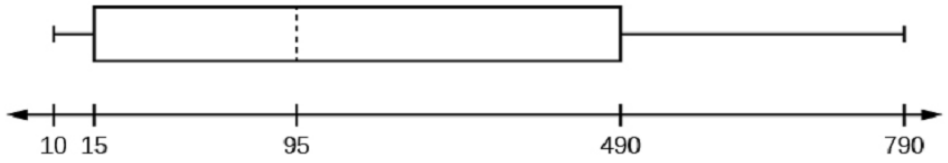
10; 10; 10; 15; 35; 75; 90; 95; 100; 175; 420; 490; 515; 515; 790

The five numbers used to create a box-and-whisker plot are:

Min: 10, Q_1: 15, Med: 95, Q_3: 490, Max: 790

Solution:

The following graph shows the box-and-whisker plot.



130.Construct a box-and-whisker plot for the following data.

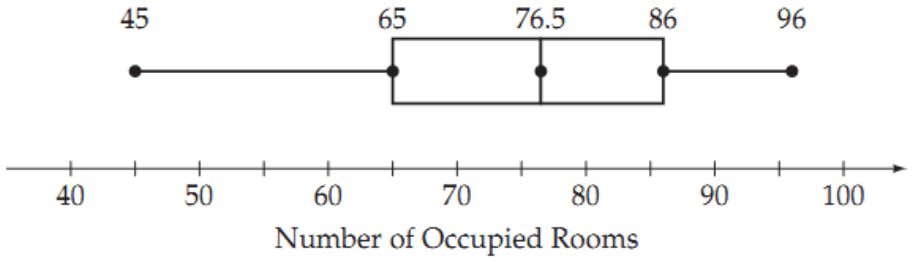
The Numbers of Occupied Rooms in a Resort during an 18-Day Period

86	77	58	45	94	96	83	76	75
65	68	72	78	85	87	92	55	61

Solution:

For the data set, we determined that

$Q_1 = 65, Q_2 = 76.5$ and $Q_3 = 86$.

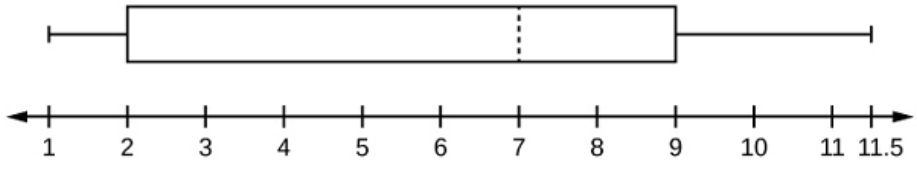


131. Construct box plot for the dataset. 1; 1; 2; 2; 4; 6; 6.8; 7.2; 8; 8.3; 9; 10; 10; 11.5

Solution: Consider the dataset.

1; 1; 2; 2; 4; 6; 6.8; 7.2; 8; 8.3; 9; 10; 10; 11.5

The first quartile is two, the median is seven, and the third quartile is nine. The smallest value is one, and the largest value is 11.5. The following image shows the constructed box plot.



The two whiskers extend from the first quartile to the smallest value and from the third quartile to the largest value. The median is shown with a dashed line.

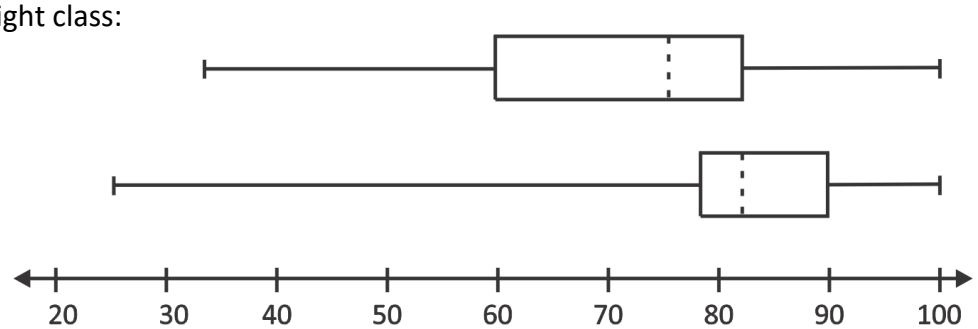
132.Test scores for a college statistics class held during the day are:
99; 56; 78; 55.5; 32; 90; 80; 81; 56; 59; 45; 77; 84.5; 84; 70; 72; 68; 32; 79; 90
Test scores for a college statistics class held during the evening are:
98; 78; 68; 83; 81; 89; 88; 76; 65; 45; 98; 90; 80; 84.5; 85; 79; 78; 98; 90; 79; 81; 25.5

a. Find the smallest and largest values, the median, and the first and third quartile for the day class.
b. Find the smallest and largest values, the median, and the first and third quartile for the night class.
c. For each data set, what percentage of the data is between the smallest value and the first quartile? the first quartile and the median? the median and the third quartile? The third quartile and the largest value? What percentage of the data is between the first quartile and the largest value?
d. Create a box plot for each set of data. Use one number line for both box plots.
e. Which box plot has the widest spread for the middle 50% of the data (the data between the first and third quartiles)? What does this mean for that set of data in comparison to the other set of data?

Solution:

a. Min = 32, Q_1 = 56, M = 74.5, Q_3 = 82.5, Max = 99
b. Min = 25.5, Q_1 = 78, M = 81, Q_3 = 89, Max = 98
c. Day class: There are six data values ranging from 32 to 56: 30%. There are six data values ranging from 56to 74.5: 30%. There are five data values ranging from 74.5 to 82.5: 25%. There are five data values ranging from 82.5 to 99: 25%. There are 16 data values between the first quartile, 56, and the largest value, 99: 75%.

Night class:



d. The first data set has the wider spread for the middle 50% of the data. The IQR for the first data set is greater than the IQR for the second set. This means that there is more variability in the middle 50% of the first dataset.

Variance and Standard Deviation

Variance and standard deviation are two fundamental concepts in statistics that measure the spread or dispersion of a dataset from its central tendency. While the mean provides a snapshot of the data's central value, variance and standard deviation offer a complementary perspective by quantifying the degree of variation or uncertainty within the data. Variance represents the average of the squared differences between individual data points and the mean, while standard deviation is the square root of variance, providing a more interpretable measure of dispersion. Together, variance and standard deviation provide essential insights into the consistency and reliability of the data, enabling researchers and analysts to make informed decisions and draw meaningful conclusions.

Variance: Variance is defined as the mean of the squared deviation of $x_i;(i=1,2,3,\ldots,n)$ observations from their arithmetic mean.

Standard Deviation: Standard Deviation is defined as positive square root of the mean of the squared deviation of $x_i;(i=1,2,3,\ldots,n)$ observations from their arithmetic mean

Standard Deviation is the measure of how far, on average, the data is from the mean. Another related measure, is the **Variance** which is standard deviation squared. The standard deviation and variance for a **SAMPLE** are calculated by the following symbols and formulas:

- $\textbf{Variance} = s^2 = \dfrac{\sum\left(X-\overline{X}\right)^2}{n}$ for ungroup data
- $\textbf{Variance} = s^2 = \dfrac{\sum f\left(X-\overline{X}\right)^2}{\sum f}$ for group data
- $\textbf{Standard Deviation} = s = \sqrt{\dfrac{\sum\left(X-\overline{X}\right)^2}{n}}$ for ungroup data
- $\textbf{Standard Deviation} = s = \sqrt{\dfrac{\sum f\left(X-\overline{X}\right)^2}{\sum f}}$ for group data

The standard deviation and variance for a **POPULATION** are calculated by the following symbols and formulas:

- $\textbf{Variance} = \sigma^2 = \dfrac{\sum\left(X-\mu\right)^2}{n}$
- $\textbf{Standard Deviation} = \sigma = \sqrt{\dfrac{\sum\left(X-\mu\right)^2}{n}}$

Computational formulae of Variance and Standard Deviation:

- **Variance** $= s^2 = \dfrac{\sum X^2}{n} - \left(\dfrac{\sum X}{n}\right)^2$ for ungroup data
- **Variance** $= s^2 = \dfrac{\sum fX^2}{\sum f} - \left(\dfrac{\sum fX}{\sum f}\right)^2$ for group data
- **Standard Deviation** $= s = \sqrt{\dfrac{\sum X^2}{n} - \left(\dfrac{\sum X}{n}\right)^2}$ for ungroup data
- **Standard Deviation** $= s = \sqrt{\dfrac{\sum fX^2}{\sum f} - \left(\dfrac{\sum fX}{\sum f}\right)^2}$ for group data

133.The following numbers were obtained by sampling a population.

2, 4, 7, 12, 15

Find the standard deviation of the sample.

Solution:

$$\overline{X} = \frac{\sum X}{n} = \frac{2+4+7+12+15}{5} = \frac{40}{5} = 8$$

X	$X - \overline{X}$	$\left(X - \overline{X}\right)^2$
2	2 - 8 = -6	$(-6)^2 = 36$
4	4 - 8 = -4	$(-4)^2 = 16$
7	7 - 8 = -1	$(-1)^2 = 1$
12	12 - 8 = 4	$(4)^2 = 16$
15	15 - 8 = 7	$(7)^2 = 49$
		118

$$\textbf{Variance} = s^2 = \frac{\sum\left(X - \overline{X}\right)^2}{n} = \frac{118}{5} = 23.6$$

$$\textbf{Standard Deviation} = s = \sqrt{\frac{\sum\left(X - \overline{X}\right)^2}{n}} = \sqrt{23.6} = 4.858$$

134.The marks of six students in Mathematics are as follows. Determine variance and standard deviation.

Student No.	1	2	3	4	5	6
Marks	60	70	30	90	80	42

Solution:

Let X = marks of a student.

$$\overline{X} = \frac{\sum X}{n} = \frac{372}{6} = 62$$

X	X^2	$X - \overline{X}$	$\left(X - \overline{X}\right)^2$
60	3600	– 2	4
70	4900	8	64
30	900	– 32	1024
90	8100	28	784
80	6400	18	324
42	1764	– 20	400
$\sum X = 372$	$\sum X^2 = 25664$	$\sum\left(X - \overline{X}\right) = 0$	$\sum\left(X - \overline{X}\right)^2 = 2600$

$$\textbf{Variance} = s^2 = \frac{\sum\left(X - \overline{X}\right)^2}{n} \approx 433.3333$$

$$\textbf{computational Variance} = s^2 = \frac{\sum X^2}{n} - \left(\frac{\sum X}{n}\right)^2 \approx 433.3333$$

$$\textbf{Standard Deviation} = s = \sqrt{\frac{\sum\left(X - \overline{X}\right)^2}{n}} \approx 20.81666$$

$$\textbf{computational Standard Deviation} = s = \sqrt{\frac{\sum X^2}{n} - \left(\frac{\sum X}{n}\right)^2} \approx 20.81666$$

135.For the following data showing weights of toffee boxes in gm. Determine variance and standard deviation.

X	4.5	14.5	24.5	34.5	44.5	54.5	64.5
f	2	10	5	9	6	7	1

Solution:

Let X = marks of a student.

$$\overline{X} = \frac{\sum X}{\sum f} = \frac{241.5}{40} = 6.0376$$

X	f	$X-\overline{X}$	$\left(X-\overline{X}\right)^2$	$f\left(X-\overline{X}\right)^2$	fX	fX^2
4.5	2	– 28	784	1568	9	40.5
14.5	10	–18	324	3240	145	2102.5
24.5	5	– 8	64	320	122.5	3001.5
34.5	9	2	4	36	310.5	10712.25
44.5	6	12	144	864	267	11881.5
54.5	7	22	484	3388	381.5	20791.75
64.5	1	32	1024	1024	64.5	4160.25

$$\textbf{Variance} = s^2 = \frac{\sum f\left(X-\overline{X}\right)^2}{\sum f} = \frac{10600}{40} = 265$$

$$\textbf{computational Variance} = s^2 = \frac{\sum fX^2}{\sum f} - \left(\frac{\sum fX}{\sum f}\right)^2 = \frac{52690}{40} - \left(\frac{1300}{40}\right)^2 = 261$$

$$\textbf{Standard Deviation} = s = \sqrt{261} = 16.155$$

$$\textbf{computational Standard Deviation} = s = \sqrt{261} = 16.155$$

136.A consumer group has tested a sample of eight size D batteries from each of three companies. The results of the tests are shown in the following table. According to these tests, which company produces batteries for which the values representing hours of constant use have the smallest standard deviation?

Company	Hours of Constant Use per batter
Ever So Brignt	6.2, 6.4, 7.1, 5.9, 8.3, 5.3, 7.5, 9.3
Dependable	6.8, 6.2, 7.2, 5.9, 7.0, 7.4, 7.3, 8.2
Beacon	6.1, 6.6, 7.3, 5.7, 7.1, 7.6, 7.1, 8.5

Solution:

The mean for each sample of batteries is 7 hours.

The batteries from Ever So Bright have a standard deviation of

$$\textbf{Standard Deviation} = s = \sqrt{\frac{\sum\left(X-\overline{X}\right)^2}{n}} = \sqrt{\frac{12.34}{7}} \approx 1.328 \text{ Hours}$$

The batteries from Dependable have a standard deviation of

$$\textbf{Standard Deviation} = s = \sqrt{\frac{\sum\left(X-\overline{X}\right)^2}{n}} = \sqrt{\frac{3.62}{7}} \approx 0.719 \text{ Hours}$$

The batteries from Beacon have a standard deviation of

$$\textbf{Standard Deviation} = s = \sqrt{\frac{\sum\left(X-\overline{X}\right)^2}{n}} = \sqrt{\frac{5.38}{7}} \approx 0.877 \text{ Hours}$$

The batteries from Dependable have the smallest standard deviation. According to these results, the Dependable company produces the most consistent batteries with regard to life expectancy under constant use.

137. Can the variance of a data set be smaller than the standard deviation of the data set?

Solution:

Yes. The variance is smaller than the standard deviation whenever the standard deviation is less than 1.

Sampling Techniques

Sampling techniques are methods used to select a representative subset of data from a larger population. It is used in almost every field of life. Some examples are as follows;

- A cook taste a bit of cooked food to find whether it has been properly cooked or not.
- A food inspector takes a sample of food or items like milk, flour etc. to find whether they are pure or not.
- Cement, steel and bricks are examined before using them in different places.

Below are some important terminologies we will be using in sampling.

- **Data** is the collection of all observations for a particular variable or variables, from one more people or things.
- A **Population** is the collection of all individuals or items under consideration in a study **or** the totality of individuals is called population **e.g.** Total number of absent students, number of color TV sets, Monthly salaries of all employees, number of computers sold out. Total number of objects in a population is called **population size**.
- A **Sample** is the part of a population from which information is actually collected. **e.g.** wheat yield per acre for 5 pieces of land. Total number of objects in a sample is called **sample size**.
- **Sampling** is the process of selecting a small portion of the population which represent all the characteristics of the population.
- **Sample Survey** is the collection of information from a representative part of the population. It is carried out by an experimental design.
- A **Census** is information (data) obtained from the entire population.
- A **Parameter** is a numerical measurement describing some characteristic of a population. Such as mean, median or standard deviation calculated from the population.

Examples: The average starting salary of elementary school teachers in Georgia is $33,673.The average for the whole United States is $35,763.

- A **Statistic** is a numerical measurement describing some characteristic of a sample. Example: A survey of ten job postings for elementary school teachers in the Atlanta area, had an average starting salary of $38,541.
- **Sampling Error** is the difference between the sample static and the population parameter is called sampling error. i.e. $\in = t - \theta$, where t is sample static and θ is corresponding population parameter.
- **Standard Error** is the standard deviation of sampling distribution of any static.
- **Bias** is the difference between the expected value of the sample static and the true value of the population parameter. i.e. $B = \in(t) - \theta$, where t is sample static used to estimate the population static θ .

Some useful Formulae

- **Formulae of Parameters;**

Population Mean $= \mu = \dfrac{\sum x}{N}$

Population Variance $= \sigma^2 = \dfrac{\sum x^2}{N} - \left(\dfrac{\sum x}{N}\right)^2$

Population Proportion $= \pi = \dfrac{x}{N}$

- **Formulae of Static;**

Sample Mean $= x = \dfrac{\sum x}{n}$

Sample Variance $= s^2 = \dfrac{\sum x^2}{n} - \left(\dfrac{\sum x}{n}\right)^2$

Sample Proportion $= p = \dfrac{x}{n}$

Sampling with replacement: Sampling is said to be with replacement when from a finite population, a sampling unit is drawn, observed and then returned to the population before another unit is drawn. The population in this case remains the same and a sampling unit might be selected more than once.

Formula to find number of samples, when sampling is done with replacement is $\mathbf{m = N^n}$ where m is possible number of samples, N is population size and n is sample size.

Results for Sampling Distribution of Means (with replacement)

- **Formulae of Parameters;**

Mean of Means or Mean of Sampling Distribution of Means $= \mu_{\bar{x}} = \sum \bar{x}\, f\left(\bar{x}\right)$

Also we may use in need: $\mu_{\bar{x}} = \mu = \frac{\sum x}{N}$

Standard Error of Means $= \sigma_{\bar{x}} = \sqrt{\sum \bar{x}^2 f(\bar{x}) - (\mu_{\bar{x}})^2}$

Standard Error of Sampling Distribution of Means $= \sigma_{\bar{x}} = \sqrt{\sum \bar{x}^2 f(\bar{x}) - (\mu_{\bar{x}})^2}$

Also we may use in need: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{\frac{\sum x^2}{N} - \left(\frac{\sum x}{N}\right)^2}}{\sqrt{n}}$

Variance of Means $= \sigma_{\bar{x}}^2 = \sum \bar{x}^2 f(\bar{x}) - (\mu_{\bar{x}})^2$

Variance of Sampling Distribution of Means $= \sigma_{\bar{x}}^2 = \sum \bar{x}^2 f(\bar{x}) - (\mu_{\bar{x}})^2$

Also we may use in need: $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} = \frac{\frac{\sum x^2}{N} - \left(\frac{\sum x}{N}\right)^2}{n}$

138. A population consists of 3,7,11,15,19. Take all possible sample of size 2 with replacement. Form the sampling distribution of sample mean $\bar{x}$. Find its means and variance. Compare it with population mean and variance.

Solution:

$x = 3,7,11,15,19$ N = 5 n = 2 with replacement

Population Mean $= \mu = \frac{\sum x}{N} = \frac{55}{5} = 11$ ..... (i)

x	3	7	11	15	19
x^2	9	49	121	225	361

Population Variance $= \sigma^2 = \frac{\sum x^2}{N} - \left(\frac{\sum x}{N}\right)^2 = 32$

$\frac{\sigma^2}{n} = \frac{32}{2} = 16$ ..... (ii)

Possible samples of size 2 WR =m = $N^n = 5^2 = 25$

Samples	$\bar{x}$	Samples	$\bar{x}$
3,3	3	7,3	5
3,7	5	7,7	7
3,11	7	7,11	9
3,15	9	7,15	11
3,19	11	7,19	13

Samples	$\bar{x}$	Samples	$\bar{x}$
11,3	7	19,3	11
11,7	9	19,7	13
11,11	11	19,11	15
11,15	13	19,15	17
11,19	15	19,19	19

Sampling distribution of sample mean $\bar{x}$ is

$\bar{x}$	f	$f(\bar{x})$	$\bar{x}f(\bar{x})$	$\bar{x}^2 f(\bar{x})$
3	1	1/25	3/25	9/25
5	2	2/25	10/25	50/25
7	3	3/25	21/25	147/25
9	4	4/25	36/25	324/25
11	5	5/25	55/25	605/25
13	4	4/25	52/25	676/25
15	3	3/25	45/25	675/25
17	2	2/25	34/25	578/25
19	1	1/25	19/25	361/25
total	25	1	275/25	3425/25

$$\text{Mean of Means} = \mu_{\bar{x}} = \sum \bar{x} f(\bar{x}) = \frac{275}{25} = 11 \quad \ldots\ldots\ldots\ldots (\text{iii})$$

From (i) and (iii) $\quad \boldsymbol{\mu_{\bar{x}} = \mu}$

$$\text{Variance of Means} = \sigma^2_{\bar{x}} = \sum \bar{x}^2 f(\bar{x}) - (\mu_{\bar{x}})^2 = \frac{3425}{25} - (11)^2$$

$$\sigma^2_{\bar{x}} = 16 \quad \ldots\ldots\ldots\ldots (\text{iv})$$

From (ii) and (iv) $\quad \boldsymbol{\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}}$

139.A population consists of 4,6,8. Take all possible sample of size 3 with replacement. Form the sampling distribution of sample mean $\bar{x}$. Calculate mean and standard error of mean. Verify the results with population mean and standard deviation.

Solution:

$x = 4,6,8$ N = 3 n = 3 with replacement

$$\text{Population Mean} = \mu = \frac{\sum x}{N} = \frac{18}{3} = 6 \quad \ldots\ldots\ldots\ldots (\text{i})$$

x	4	6	8	**18**
x^2	16	36	64	**116**

Population Variance $= \sigma = \sqrt{\frac{\sum x^2}{N} - \left(\frac{\sum x}{N}\right)^2} = 1.63$

$$\frac{\sigma}{\sqrt{n}} = \frac{1.63}{\sqrt{3}} = 0.94 \ldots\ldots\ldots\ldots (ii)$$

Possible samples of size 3WR =m = $N^n = 3^3 = 27$

Samples	$\bar{x}$	Samples	$\bar{x}$
4,4,4	4	6,4,4	4.67
4,4,6	4.67	6,4,6	5.33
4,4,8	5.33	6,4,8	6
4,6,4	4.67	6,6,4	5.33
4,6,6	5.33	6,6,6	6
4,6,8	6	6,6,8	6.67
4,8,4	5.33	6,8,4	6
4,8,6	6	6,8,6	6.67
4,8,8	4.67	6,8,8	7.33

Samples	$\bar{x}$
8,4,4	5.33
8,4,6	6
8,4,8	6.67
8,6,4	6
8,6,6	6.67
8,6,8	7.33
8,8,4	6.67
8,8,6	7.33
8,8,8	8

Sampling distribution of sample mean $\bar{x}$ is

$\bar{x}$	f	$f(\bar{x})$	$\bar{x} f(\bar{x})$	$\bar{x}^2 f(\bar{x})$
4	1	1/27	4/27	16/27
4.67	3	3/27	14.01/27	65.4267/27
5.33	6	6/27	31.98/27	170.4534/27
6	7	7/27	42/27	252/27
6.67	6	6/27	40.02/27	266.9334/27
7.33	3	3/27	21.99/27	161.1867/27
8	1	1/27	8/27	64/27
total	27	1	162/27	996.0002/27

$$\text{Mean of Means} = \mu_{\bar{x}} = \sum \bar{x} f\left(\bar{x}\right) = \frac{162}{27} = 6 \quad \ldots\ldots\ldots\ldots \left(\text{iii}\right)$$

From $\left(\text{i}\right)$ and $\left(\text{iii}\right)$ $\boldsymbol{\mu_{\bar{x}} = \mu}$

$$\text{Standard Error of Means} = \sigma_{\bar{x}} = \sqrt{\sum \bar{x}^2 f\left(\bar{x}\right) - \left(\mu_{\bar{x}}\right)^2} = \sqrt{\frac{996.0002}{27} - \left(6\right)^2}$$

$$\sigma_{\bar{x}} = 0.94 \quad \ldots\ldots\ldots\ldots \left(\text{iv}\right)$$

From $\left(\text{ii}\right)$ and $\left(\text{iv}\right)$ $\boldsymbol{\sigma_{\bar{x}} = \frac{\sigma^2}{n}}$

Sampling without replacement:

If the sample is taken without replacement from a finite population, the selected element is not returned to the population before drawing the next element. In without replacement sampling an element can be selected only once. Formula to find number of samples, when sampling is done without replacement is **$m = {}^{N}C_{n}$** where m is possible number of samples, N is population size and n is sample size.

Results for Sampling Distribution of Means (with replacement)

- **Formulae of Parameters;**

$$\text{Mean of Means or Mean of Sampling Distribution of Means} = \mu_{\bar{x}} = \sum \bar{x} f\left(\bar{x}\right)$$

Also we may use in need: $\mu_{\bar{x}} = \mu = \frac{\sum x}{N}$

$$\text{Standard Error of Means} = \sigma_{\bar{x}} = \sqrt{\sum \bar{x}^2 f\left(\bar{x}\right) - \left(\mu_{\bar{x}}\right)^2}$$

$$\text{Standard Error of Sampling Distribution of Means} = \sigma_{\bar{x}} = \sqrt{\sum \bar{x}^2 f\left(\bar{x}\right) - \left(\mu_{\bar{x}}\right)^2}$$

Also we may use in need: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = \frac{\sqrt{\frac{\sum x^2}{N} - \left(\frac{\sum x}{N}\right)^2}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$

$$\text{Variance of Means} = \sigma_{\bar{x}}^2 = \sum \bar{x}^2 f\left(\bar{x}\right) - \left(\mu_{\bar{x}}\right)^2$$

$$\text{Variance of Sampling Distribution of Means} = \sigma_{\bar{x}}^2 = \sum \bar{x}^2 f\left(\bar{x}\right) - \left(\mu_{\bar{x}}\right)^2$$

Also we may use in need: $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) = \frac{\frac{\sum x^2}{N} - \left(\frac{\sum x}{N}\right)^2}{n}\left(\frac{N-n}{N-1}\right)$

140.A population consists of 6 members 2,4,6,8,10 and 12. Take all possible sample of size 2 without replacement. Form the sampling distribution of means. Find its means and variance. Compare it with population mean and variance.

Solution:

$x = 2,4,6,8,10,12$ N = 6 n = 2 WOR

Population Mean $= \mu = \frac{\sum x}{N} = \frac{42}{6} = 7$ ..... (i)

x	2	4	6	8	10	12
x^2	4	16	36	64	100	144

Population Variance $= \sigma^2 = \frac{\sum x^2}{N} - \left(\frac{\sum x}{N}\right)^2 = 11.67$

$\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) = \frac{11.67}{2}\left(\frac{6-2}{6-1}\right) = 4.67$ ..... (ii)

Possible samples of size 2 WOR =m = ${}^{N}C_n = {}^{6}C_2 = 15$

Samples	$\bar{x}$	Samples	$\bar{x}$
2,4	3	4,6	5
2,6	4	4,8	6
2,8	5	4,10	7
2,10	6	4,12	8
2,12	7	6,8	7

Samples	$\bar{x}$
6,10	8
8	9
8,10	9
8,12	10
10,12	11

Sampling distribution of sample mean $\bar{x}$ is

$\bar{x}$	f	$f(\bar{x})$	$\bar{x} f(\bar{x})$	$\bar{x}^2 f(\bar{x})$
3	1	1/15	3/15	9/15
4	1	1/15	4/15	16/15
5	2	2/15	10/15	50/15
6	2	2/15	12/15	72/15
7	3	3/15	21/15	147/15
8	2	2/15	16/15	128/15
9	2	2/15	18/15	162/15
10	1	1/15	10/15	100/15
11	1	1/15	11/15	121/15
total	15	1	105/15	805/15

$$\text{Mean of Means} = \mu_{\bar{x}} = \sum \bar{x} f(\bar{x}) = \frac{105}{15} = 7 \quad \ldots\ldots\ldots\ldots (\text{iii})$$

From (i) and (iii) $\quad \boldsymbol{\mu_{\bar{x}} = \mu}$

$$\text{Variance of Means} = \sigma^2_{\bar{x}} = \sum \bar{x}^2 f(\bar{x}) - (\mu_{\bar{x}})^2 = \frac{805}{15} - (7)^2$$

$$\sigma^2_{\bar{x}} = 46.7 \quad \ldots\ldots\ldots\ldots (\text{iv})$$

From (ii) and (iv) $\quad \boldsymbol{\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)}$

141.A population consists of 6 members 3,6,9,12,15 and 18. Take all possible sample of size 3 without replacement. Form the sampling distribution of means. Find its means and standard deviation. Also find the standard error.

Solution:

$x =$ 3,6,9,12,15,18 N = 6 n = 3 WOR

Population Mean $= \mu = \frac{\sum x}{N} = \frac{63}{6} = 10.5$

x	3	6	9	12	15	18
x^2	9	36	81	144	225	324

Population S.D $= \sigma = \sqrt{\frac{\sum x^2}{N} - \left(\frac{\sum x}{N}\right)^2} = 5.12$

Possible samples of size 3 WOR =m = ${}^{N}C_3 = {}^{6}C_3 = 20$

Samples	$\bar{x}$	Samples	$\bar{x}$
3,6,9	6	6,9,12	9
3,6,12	7	6,9,15	10
3,6,15	8	6,9,18	11
3,6,18	9	6,12,15	11
3,9,12	8	6,12,18	12
3,9,15	9	6,15,18	13
3,9,18	10	9,12,15	12
3,12,15	10	9,12,18	13
3,12,18	11	9,15,18	14
3,15,18	12	12, 15,18	15

Sampling distribution of sample mean $\bar{x}$ is

$\bar{x}$	f	$f(\bar{x})$	$\bar{x}f(\bar{x})$	$\bar{x}^2 f(\bar{x})$
6	1	1/20	6/20	36/20
7	1	1/20	7/20	49/20
8	2	2/20	16/20	128/20
9	3	3/20	27/20	243/20
10	3	3/20	30/20	300/20
11	3	3/20	33/20	363/20
12	3	3/20	36/20	432/20
13	2	2/20	26/20	338/20
14	1	1/20	14/20	196/20
15	1	1/20	15/20	225/20
total	20	1	210/20	2310/20

$$\text{Mean of Means} = \mu_{\bar{x}} = \sum \bar{x} f(\bar{x}) = \frac{210}{20} = 10.5$$

$$\text{Standard Error of Means} = \sigma_{\bar{x}} = \sqrt{\sum \bar{x}^2 f(\bar{x}) - (\mu_{\bar{x}})^2} \Rightarrow \sigma_{\bar{x}} = \sqrt{\frac{2310}{20} - (10.5)^2} = 2.29$$

Estimation

Estimation is a procedure by which we obtain the value of unknown population parameters by using the sample information. It is the process of approximating a population parameter (e.g., mean, proportion, variance) using sample data. It involves making an educated guess about a population characteristic based on a representative subset of data.

Types of Estimation

Estimation is divided into two types;

- **Point Estimation:** The process of finding a single value from the sample.
- **Interval Estimation:** The process of finding a range of values within which the population parameter is expected to lie with a certain degree of confidence.

Point Estimate: A single numerical value calculated from the sample.

Point Estimator: The rule or formula that is used to estimate a population parameter. Its important characteristics are unbiasedness, consistency, efficiency and sufficiency.

Estimate

An estimate is defined as numerical values of the unknown population parameter obtained by apply an estimator. Estimate is divided into two types;

- **Point Estimate:** It is a single numerical value from the sample.
- **Interval Estimate:** It is a range of values within which the population parameter is expected to lie with a certain degree of confidence.

Degree of Freedom

The term degree of freedom is defined as the number of independent or freely chosen variables.

Central Limit Theorem

The distribution of the means of a large number of samples of size taken from a population is approximately a normal distribution.

Confidence Intervals

A **confidence interval** is a type of estimate but, instead of being just one number, it is an interval of numbers. It provides a range of reasonable values in which we expect the population parameter to fall. There is no guarantee that a given confidence interval does capture the parameter, but there is a predictable probability of success. Confidence intervals are based on the Central Limit Theorem and the hypothesis testing equations.

Let $(1-\alpha)$ be a specified high probability and L and U be the functions of sample observations $X_1, X_2, X_3, \ldots, X_n$ such that: $P(L < \theta < U) = 1-\alpha; 0 < \alpha < 1$

Then the interval (L, U) is called a $100(1-\alpha)\%$ confidence interval for the parameter θ.

Level of Confidence

The probability of accepting a true null hypothesis is called level of hypothesis. It is denoted by $(1-\alpha)$.

Level of Significance

The probability of rejecting a null hypothesis when it is actually true is called level of significance. It is denoted by α .

Test Statistic

A test statistics is a function or formula of sample observations that provides a basis for testing a null hypothesis. The most commonly used test statistics are Z and T – Tests.

Test Significance

Test of significance is a procedure which enables us, on the basis of sampling distribution, whether to accept or reject a hypothesis.

T-Test/ Small Sample Test

A t-test is a statistical hypothesis test used to determine if there's a significant difference between the means of two groups. It's commonly used for:

1. Comparing two independent samples (e.g., control vs. treatment).
2. Comparing a sample mean to a known population mean.
3. Testing the significance of regression coefficients.

Types of T-Tests

1. One – sample T-test.
2. Independent samples T-test (two-sample t-test).
3. Paired samples T-test.

Z–Test/Z – Score Test

A Z-test is a statistical hypothesis test used to determine if a sample mean is significantly different from a known population mean. It's commonly used when:

1. Sample size is large (n > 30).
2. Population standard deviation is known.
3. Data is normally distributed.

Key differences between T-Test and Z-Test

1. Sample size: T-test is used for smaller samples (n < 30), while Z-test is used for larger samples.
2. Population standard deviation: T-test estimates standard deviation from sample data, while Z-test requires known population standard deviation.
3. Distribution: T-test assumes normal distribution, while Z-test assumes normal distribution and large sample size.

Student T-Test vs. T-Test

Student's T-test and T-test are often used interchangeably. However, technically:
1. Student's T-test refers specifically to the test developed by William Sealy Gosset (under the pseudonym "Student") for small samples.
2. T-test is a broader term encompassing various types of T-tests.
In practice, the terms are used synonymously, and the distinction is often ignored.

142.A random sample selected from a normal population with mean μ and variance σ^2 gave the values 25,31,23,33,28,36,22,26. Give the point estimator for μ and σ^2 and find their point estimates.

Solution:

x	x^2
25	625
31	961
23	529
33	1089
28	784
39	1296
22	484
26	676
224	6444

Point estimator of population mean $\mu = \bar{x} = \frac{\sum x}{n} = \frac{224}{8} = 28$

Point estimator of population variance $\sigma^2 = \hat{s}^2 = \frac{\sum x^2 - n\bar{x}^2}{n-1} = \frac{6444 - 8(28)^2}{8-1} = 24.57$

Z –values for Commonly used Confidence Levels

Confidence Level	Area	Z value
90%	0.0500 and 0.9500	1.64 or 1.65
95%	0.0250 and 0.9750	1.96
96%	0.0200 and 0.9800	2.05
97%	0.0150 and 0.9850	2.17
98%	0.0100 and 0.9900	2.33
99%	0.0050 and 0.9950	2.57 or 2.58

143. Confidence Interval for Mean (Z – test): A normal population has a variance of 100. A random sample of size 16 selected from the population has mean of 52.50. Construct the 90% confidence interval estimate of the population mean μ. Interpret the result.

Solution:

$$n = 16, \sigma^2 = 100, \sigma = 10, \bar{x} = 52.50, 90\%CI = ?$$

$$1-\alpha = 90\% \Rightarrow \alpha = 1-0.90 \Rightarrow \alpha = 0.10 \Rightarrow \frac{\alpha}{2} = 0.05 \Rightarrow Z_{\frac{\alpha}{2}} = Z_{0.05} = 1.645$$

$$90\%CI = \bar{x} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = 52.50 \pm (1.645)\frac{10}{\sqrt{16}} = 52.50 \pm 4.11$$

$$90\%CI = 52.50 - 4.11 = 48.39$$

$$90\%CI = 52.50 + 4.11 = 56.64$$

Hence 90% confidence interval for population mean μ obtained from the observed sample is $(48.39, 56.64)$.

144. Confidence Interval for Mean (Z – test): A particular component in a transistor circuit has a lifetime which is known to follow a skew distribution. A random sample of 250 components from a week's production given an average lifetime of 840 hours, and the variance of lifetime is 483(Hours^2). Find approximately 95% confidence limits to the true mean lifetime in the whole population of the product.

Solution:

$$n = 250, \hat{s}^2 = 483, \hat{s} = \sqrt{483} = 21.98, \bar{x} = 840, 95\%CI = ?$$

$$1-\alpha = 95\% \Rightarrow \alpha = 1-0.95 \Rightarrow \alpha = 0.05 \Rightarrow \frac{\alpha}{2} = 0.025 \Rightarrow Z_{\frac{\alpha}{2}} = Z_{0.025} = 1.96$$

$$95\%CI = \bar{x} \pm Z_{\frac{\alpha}{2}} \frac{\hat{s}}{\sqrt{n}} = 840 \pm (1.96)\frac{21.98}{\sqrt{250}} = 840 \pm 2.72$$

$$95\%CI = 840 - 2.72 = 837.28 \Rightarrow 95\%CI = 840 + 2.72 = 842.72$$

Hence 95% confidence interval for population mean μ obtained from the observed sample is $(837.28, 842.72)$.

145. Confidence Interval for Mean (Z – test): A random sample of size n = 200 selected without replacement from a finite population of size N = 1000 with $\sigma = 1.28$ showed that $\bar{x} = 68.60$. Construct a 97% confidence interval for the mean of the population.

Solution:

$$n = 200, N = 1000, \sigma = 1.28, \bar{x} = 68.60, 97\%CI = ?$$

$$1-\alpha = 97\% \Rightarrow \alpha = 1-0.97 \Rightarrow \alpha = 0.03 \Rightarrow \frac{\alpha}{2} = 0.015 \Rightarrow Z_{\frac{\alpha}{2}} = Z_{0.015} = 2.17$$

$$97\%CI = \bar{x} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = 68.60 \pm (2.17) \frac{1.28}{\sqrt{220}} \sqrt{\frac{1000-200}{1000-1}} = 68.60 \pm 0.17$$

$$97\%CI = 68.60 - 0.17 = 68.43$$

$$97\%CI = 68.60 + 0.17 = 68.77$$

Hence 97% confidence interval for population mean μ obtained from the observed sample is $(68.43, 68.77)$.

146. Confidence Interval for Mean (t – test): Ten packets of a particular brand of biscuits are chosen at random and their mass measured in grams. The results are; $n = 10, x = 3978.70, \sum x^2 = 1583098.30$ assuming that the sample is taken from a normal population with mean mass μ calculate the 98% confidence interval for μ .

Solution:

$$n = 10, x = 3978.70, \sum x^2 = 1583098.30, v = n-1 = 10-1 = 9, 98\%CI = ?$$

$$\bar{x} = \frac{\sum x}{n} = \frac{3978.70}{10} = 397.87, \hat{s} = \sqrt{\frac{\sum x^2 - n\bar{x}^2}{n-1}} = 3.21$$

$$1-\alpha = 98\% \Rightarrow \alpha = 1-0.98 \Rightarrow \alpha = 0.02 \Rightarrow \frac{\alpha}{2} = 0.01 \Rightarrow t_{\frac{\alpha}{2}(v)} = t_{0.01(9)} = 2.821$$

$$98\%CI = \bar{x} \pm t_{\frac{\alpha}{2}(v)} \frac{\hat{s}}{\sqrt{n}} = 397.87 \pm (2.821) \frac{3.21}{\sqrt{10}} = 397.87 \pm 2.86$$

$$98\%CI = 397.87 - 2.86 = 395.01$$

$$98\%CI = 397.87 + 2.86 = 400.73$$

Hence 98% confidence interval for population mean μ obtained from the observed sample is $(395.01, 400.73)$.

147.Confidence Interval for Mean (t – test): A random sample of eight observations of a normal variable gave $\sum x = 261.20, \sum \left(x - \bar{x}\right)^2 = 3.22.$ **Calculate a 95 % confidence interval for the population mean** μ **.**

Solution:

$n = 8, \hat{s} = ?, \bar{x} = ?, v = n - 1 = 8 - 1 = 7, 95\%CI = ?$

$$\bar{x} = \frac{\sum x}{n} = \frac{261.20}{8} = 32.65, \hat{s} = \sqrt{\frac{\sum \left(x - \bar{x}\right)^2}{n-1}} = 0.68$$

$$1 - \alpha = 95\% \Rightarrow \alpha = 1 - 0.95 \Rightarrow \alpha = 0.05 \Rightarrow \frac{\alpha}{2} = 0.025 \Rightarrow t_{\frac{\alpha}{2}(v)} = t_{0.025(7)} = 2.365$$

$$95\%CI = \bar{x} \pm t_{\frac{\alpha}{2}(v)} \frac{\hat{s}}{\sqrt{n}} = 32.65 \pm (2.365)\frac{0.68}{\sqrt{8}} = 32.65 \pm 0.57$$

$95\%CI = 32.65 - 0.57 = 32.08$

$95\%CI = 32.65 + 0.57 = 33.22$

Hence 95% confidence interval for population mean μ obtained from the observed sample is $(32.08, 33.22)$.

Hypothesis and Hypothesis testing

Hypothesis and hypothesis testing are fundamental concepts in statistical inference, enabling researchers to systematically test and validate theories, assumptions, and predictions about a population or phenomenon. A hypothesis is a clear, concise, and testable statement that proposes a relationship, difference, or association between variables. Hypothesis testing involves the systematic evaluation of this statement using sample data, statistical methods, and probability theory. By testing hypotheses, researchers can determine whether observed patterns or relationships are due to chance or whether they reflect real effects, ultimately informing decision-making, guiding future research, and advancing knowledge in various fields.

Hypothesis

Any statement which may or may not be true is called hypothesis.

Hypothesis Testing

It is a procedure which enables us to decide on the basis of information obtained from sample data whether to accept or reject any specified statement or hypothesis or assumption about the value of population parameter.

Statistical Hypothesis

A statistical hypothesis is a statement about one or more parameter of a population. This statement may or may not be true. Its validity is tested on the basis of sample obtained from the population.

Null Hypothesis

Any hypothesis which is tested for possible rejection under the assumption that it is true is called null hypothesis. It is generally denoted by **H_0**.

Alternative Hypothesis

Any hypothesis which is different from the null hypothesis. It is accepted when null hypothesis is rejected. It is generally denoted by **H_A** or **H_1**.

Types of Null Hypothesis and Alternative Hypothesis

Null Hypothesis (H_0)	Alternative Hypothesis (H_1)
$\theta = \theta_0$	$\theta \neq \theta_0$
$\theta \leq \theta_0$	$\theta > \theta_0$
$\theta \geq \theta_0$	$\theta < \theta_0$

Simple Hypothesis

A hypothesis in which all parameter of the distribution are specified is called simple hypothesis. For example, if the average age of ICS students is 16 year. i.e. $H_0 = \mu = 16$ is a simple hypothesis.

Composite Hypothesis

A hypothesis in which all parameter of the distribution are not specified is called composite hypothesis. For example, $H_1 = \mu < 16$ or $H_1 = \mu > 16$ years are composite hypothesis.

Power of a Test

If we reject a false null hypothesis, it is called power of a test. It is denoted by $(1-\beta)$.

True Situation	Decision	
	Accept H_0	Reject H_0
H_0 is true	Correct decision Or level of confidence $= 1-\alpha$	Wrong decision $= \alpha$
H_0 is false	Wrong decision $= \beta$	Correct decision $= 1-\beta$

Critical Values

The values of test statistic which separates the rejection and acceptance region are called as critical values.

Critical Region/ Rejection Region

Critical region is the part of sampling distribution of a statistics which leads to the rejection of the null hypothesis.

Acceptance Region

Acceptance region is the part of sampling distribution of a statistics which leads to the acceptance of the null hypothesis.

Some useful Formulae

- Z – test for Hypothesis Testing: $z = \dfrac{\bar{x} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$
- T – test for Hypothesis Testing: $t = \dfrac{\bar{x} - \mu}{\dfrac{\hat{s}}{\sqrt{n}}}$

Choice of Test for Testing of Hypothesis

Variance	Sample Size	Test	Use Variance in Formula
σ^2 known	n < 30 or n > 30	Z – test	σ^2
σ^2 unknown	n > 30	Z – test	$\hat{s}^2$
σ^2 unknown	n < 30	T – test	$\hat{s}^2$

General Procedure of testing of Hypothesis (or Null Hypothesis)

The procedure for testing a hypothesis about population parameter involves the following steps;

- State your problem and formulate appropriate null hypothesis H_0 with an alternative hypothesis H_1.
- Decide upon a level of significance α of the test, which is the probability of rejecting the null hypothesis when H_0 is true.
- Choose an appropriate test static.
- Calculate the value of test static from sample data.
- Determine the critical region depends upon the alternative hypothesis H_1.
- Make a conclusion. That is, if the value of test static falls in the critical region then rejectH_0 and if the value of test static falls in the accepting region then accept H_0.

Two Tailed Test

If the critical region is located equally in both tails of the sampling distribution of test statistic, the test is called two tailed test. It is also called two sided test.

One Tailed Test

If the critical region is located equally in only one tail of the sampling distribution of test statistic, the test is called one tailed test. It is also called one sided test.

Testing of Hypothesis using Z – test

- Formulate a hypothesis as

  $H_0 : \mu = \mu_0, H_1 : \mu \neq \mu_0$ (two sided or two tailed test)

  $H_0 : \mu \leq \mu_0, H_1 : \mu > \mu_0$ (one sided or one tailed test)

  $H_0 : \mu \geq \mu_0, H_1 : \mu < \mu_0$ (one sided or one tailed test)

- Decide level of significance $\alpha = 0.05$ (generally)
- Use formula $z = \dfrac{\bar{x} - \mu_0}{\dfrac{\sigma}{\sqrt{n}}}$
- Calculate the value of test static from sample data.
- **Critical Regions:**

Variance	Sample Size
$H_1 : \mu \neq \mu_0$	$\lvert Z \rvert \geq Z_{\frac{\alpha}{2}}$
$H_1 : \mu > \mu_0$	$Z > +Z_{\alpha}$
$H_1 : \mu < \mu_0$	$Z < -Z_{\alpha}$

- **Conclusions:**

  If value of test static falls in the critical region then reject H_0

  If value of test static falls in the acceptance region then accept H_0

Testing of Hypothesis using T– test

- Formulate a hypothesis as

  $H_0 : \mu = \mu_0, H_1 : \mu \neq \mu_0$ (two sided or two tailed test)

  $H_0 : \mu \leq \mu_0, H_1 : \mu > \mu_0$ (one sided or one tailed test)

  $H_0 : \mu \geq \mu_0, H_1 : \mu < \mu_0$ (one sided or one tailed test)
- Decide level of significance $\alpha = 0.05$ (generally)
- Use formula $t = \dfrac{\bar{x} - \mu_0}{\dfrac{\hat{s}}{\sqrt{n}}}$
- Calculate the value of test static from sample data.
- **Critical Regions:**

Variance	Sample Size
$H_1 : \mu \neq \mu_0$	$\lvert t \rvert \geq t_{\frac{\alpha}{2}(v)}$
$H_1 : \mu > \mu_0$	$t > +t_{\alpha(v)}$
$H_1 : \mu < \mu_0$	$t < -t_{\alpha(v)}$

- **Conclusions:**

  If value of test static falls in the critical region then reject H_0

  If value of test static falls in the acceptance region then accept H_0

Z – scores

The z-score for a given data value x is the number of standard deviations that x is above or below the mean of the data. The following formulas show how to calculate the z-score for a data value x in a population and in a sample.

Population: $z_x = \frac{x-\mu}{\sigma}$ Sample: $z_x = \frac{x-\bar{x}}{s}$

148.Must the z-score for a data value be a positive number?

Solution:

No. The z-score for a data value x is positive if x is greater than the mean, it is 0 if x is equal to the mean, and it is negative if x is less than the mean.

149.Raul has taken two tests in his chemistry class. He scored 72 on the first test, for which the mean of all scores was 65 and the standard deviation was 8. He received a 60 on the second test, for which the mean of all scores was 45 and the standard deviation was 12. In comparison with the other students, did Raul do better on the first test or the second test?

Solution: Find the z-score for each test.

$$z_{72} = \frac{72-65}{8} = 0.875 \qquad z_{60} = \frac{60-45}{12} = 1.25$$

Raul scored 0.875 standard deviation above the mean on the first test and 1.25 standard deviations above the mean on the second test. These z-scores indicate that in comparison with his classmates, Raul scored better on the second test than he did on the first test.

150.Cheryl has taken two quizzes in her history class. She scored 15 on the first quiz, for which the mean of all scores was 12 and the standard deviation was 2.4. Her score on the second quiz, for which the mean of all scores was 11 and the standard deviation was 2.0, was 14. In comparison with her classmates, did Cheryl do better on the first quiz or the second quiz?

Solution:

$$z_{15} = \frac{15-12}{2.4} = 1.25 \qquad z_{14} = \frac{14-11}{2.0} = 1.5$$

These z-scores indicate that in comparison with her classmates, Cheryl did better on the second quiz than she did on the first quiz.

151.A consumer group tested a sample of 100 light bulbs. It found that the mean life expectancy of the bulbs was 842 hours, with a standard deviation of 90. One particular light bulb from the Dura Bright Company had a z-score of 1.2.What was the life span of this light bulb?

Solution:

Substitute the given values into the z-score equation and solve for x.

$$z_x = \frac{x-\bar{x}}{s} \Rightarrow 1.2 = \frac{x-842}{90} \Rightarrow 108 = x-842 \Rightarrow 950 = x$$

The light bulb had a life span of 950 hours.

152. Roland received a score of 70 on a test for which the mean score was 65.5. Roland has learned that the z-score for his test is 0.6.Whatis the standard deviation for this set of test scores?

Solution:

$$z_x = \frac{x-\mu}{\sigma} \Rightarrow 0.6 = \frac{70-65.5}{\sigma} \Rightarrow \sigma = \frac{4.5}{0.6} = 7.5$$

The standard deviation for this set of test scores is 7.5.

153. The mean life time of electric bulbs produce by a company has in the past been 1120 hours with a standard deviation of 125 hours. A sample of 100 electric bulbs recently chosen from a supply of newly produced bulbs showed a mean life time of 1070 hours. Test the hypothesis that the mean life time of bulbs has not changed, using 5 % level of significance

Solution:

$H_0 = \mu = 1120$ and $H_1 = \mu \neq 1120$

Level of significance: $\alpha = 5\% = 0.05$

$\mu = 1120, \sigma = 125, n = 100, \bar{x} = 1070$

Test Statistic: $Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = -4$ implies $|Z| = 4$

Critical Region: Reject H_0 if $|Z| > Z_{\frac{\alpha}{2}} = 1.96$

Since the calculated value of Z lies in critical region, so we reject H_0 and conclude that the mean life time of bulbs has changed.

154. It has been found from experience that the mean breaking strength of thread is 9.63N with a standard deviation of 1.40N. Recently a sample of 36 pieces of thread showed a mean breaking strength of 8.93N. Can we conclude at 1% level of significance that thread has become inferior?

Solution:

$H_0 = \mu = 9.63$ and $H_1 = \mu < 9.63$

Level of significance: $\alpha = 1\% = 0.01$

$\mu = 9.63, \sigma = 1.40, n = 36, \bar{x} = 8.93$

Test Statistic: $Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = -3$ implies $|Z| = 3$

Critical Region: Reject H_0 if $|Z| < -Z_\alpha = -2.33$

Since the calculated value of Z lies in critical region, so we reject H_0 and conclude that the thread has become inferior.

155.The breaking strength of cables produced by a company has a mean of 1800 pounds and standard deviation of 100 pounds. By a new technique in production process, it is claimed that breaking strength can be increased. To test this claim, a random sample of 50 cables is tested and it is found that the mean breaking strength is 1850 pounds. Can we support claim at 0.01 significance level?

Solution:

$H_0 = \mu = 1800$ and $H_1 = \mu \neq 1800$

Level of significance: $\alpha = 1\% = 0.01$

$\mu = 1800, \sigma = 100, n = 50, \bar{x} = 1850$

Test Statistic: $Z = \dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}} = 3.54$

Critical Region: Reject H_0 if $|Z| > Z_{\frac{\alpha}{2}} = 2.33$

Since the calculated value of Z lies in critical region, so we reject H_0 and conclude that the claim should be supported.

156. A company claims that the average amount of coffee it supplies in jars is 6.0 oz with a standard deviation of 0.2 oz. a random sample of 100 jars is selected and average is found to be 5.9. Is the company cheating the customers? Use 5% level of significance.

Solution:

$H_0 : \mu \geq 6.0$ and $H_1 : \mu < 6.0$

Level of significance: $\alpha = 5\% = 0.05$

$\mu = 6.0, \sigma = 0.2, n = 100, \bar{x} = 5.9$

Test Statistic: $Z = \dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}} = -5$

Critical Region: $Z < -Z_{\alpha} = Z_{0.05} = -1.645$

Since the calculated value of Z lies in critical region, so we reject H_0 .

157.A timber company is interested in seeing if the number of board feet per tree has decreased since moving to a new location of timber. In the past, the company has an average of 93 board feet per tree. The company believes that the production has decreased since changing locations, a random sample of 25 trees yields $\bar{x} = 89$ and $\hat{s} = 20$. Assuming the normality of the data, test the hypothesis at a 10% level of significance.

Solution:

$H_0 : \mu \geq 93$ and $H_1 : \mu < 93$

Level of significance: $\alpha = 5\% = 0.05$

$\mu = 93, \hat{s} = 20, n = 25, \bar{x} = 89, v = n - 1 = 25 - 1 = 24$

Test Statistic: $t = \frac{\bar{x} - \mu}{\hat{s}/\sqrt{n}} = -1$

Critical Region: $t < -t_{\alpha(v)} = -t_{0.10(24)} = -1.318$

Since the calculated value of t lies in acceptance region, so we accept H_0 .

158.Ten cartons are taken at random from an automatic filling machine. The mean net weight of the 10 cartons is 15.90 oz and the sum of squared deviation from this mean is 0.276 $(oz)^2$. Does the sample mean differ significantly from intended weight of 16 oz?

Solution:

$H_0 : \mu = 16$ and $H_1 : \mu \neq 16$

Level of significance: $\alpha = 5\% = 0.05$

$\mu = 16, n = 10, \bar{x} = 15.90, v = n - 1 = 10 - 1 = 9$

$\sum (x - \bar{x})^2 = 0.276, \hat{s} = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}} = 0.18$

Test Statistic: $t = \frac{\bar{x} - \mu}{\hat{s}/\sqrt{n}} = -1$

Critical Region: $|t| \geq t_{\frac{\alpha}{2}(v)} = t_{0.025(9)} = 2.262$

Since the calculated value of t lies in acceptance region, so we accept H_0 .

Margin of error and Confidence Interval

Suppose you draw a single sample of size n from a large population and measure its sample proportion. The margin of error for 95% confidence is

margin of error ≈ $\frac{1}{\sqrt{n}}$

The 95% confidence interval is found by subtracting and adding the margin of error from the sample proportion. You can be 95% confident that the true population proportion lies within this interval. The margin of error decreases as the sample size increases.

159.Poll Margins: Find the margin of error and 95% confidence interval for the following surveys.

a. A survey of 500 people finds that 52% plan to vote for Smith for governor.

b. A survey of 1500 people finds that 87% support stricter penalties for child abuse.

Solution:

a. The proportion measured for the sample is 52%, or 0.52. For a poll of n = 500 people, the margin of error is approximately

$\frac{1}{\sqrt{n}} = \frac{1}{\sqrt{500}} \approx 0.045$

Adding and subtracting 0.045, or 4.5%, from the sample proportion of 52% produces a 95% confidence interval from 47.5% to 56.5%. We can be 95% confident that the true proportion of people who plan to vote for Smith lies in this interval.

b. The proportion measured for the sample is 87%, or 0.87. For a poll of n = 1500 people, the margin of error is approximately $\frac{1}{\sqrt{n}} = \frac{1}{\sqrt{1500}} \approx 0.026$

Adding and subtracting 0.026, or 2.6%, from the sample proportion of 87% produces a 95% confidence interval from 84.4% to 89.6%. We can be 95% confident that the true proportion of people who support the stricter penalties lies in this interval.

160. Unemployment rate: Suppose the Bureau of Labor Statistics finds 3420 unemployed people in a sample of n = 60,000 people. Estimate the population unemployment rate and give a 95% confidence interval.

Solution:

The sample proportion is the unemployment rate for the sample:

$$\frac{3420}{60,000} = 0.057$$

This proportion is likely to be close to the true population unemployment rate. The margin of error is approximately

$$\frac{1}{\sqrt{60,000}} \approx 0.004$$

We add and subtract the margin of error of 0.004 from the sample proportion of 0.057, yielding a 95% confidence interval from 0.053 to 0.061, or 5.3% to 6.1%. We can be 95% confident that the interval from 5.3% to 6.1% contains the true unemployment rate for the population.

Exercise

1) Marks obtained by 60 students of a class are given below;
60,50,46,28,58,64,36,20,50,18,42,56,20,38,40,34,24,64,64,42,46,52,50,44,36,0,
24,30,46,40,64,40,36,14,36,8,56,40,30,36,24,22,36,50,58,16,40,34,0,42,42,0,36,
18,18,68,30,46,38,16.
Make frequency distribution using appropriate class interval.

2) For data given below;
1.36,1.46,1.50,1.32,1.45,1.24,1.49,1.64,1.47,1.59,1.41,1.48,1.36,1.48,1.51,1.45,
1.26,1.38,1.76,1.63,1.19,1.56,1.65,1.54,1.61,1.73,1.60,1.50,1.45,1.76,1.67,1.35,
1.55,1.68,1.46,1.40,1.32,1.47,1.64,1.45
Make frequency distribution taking 0.05 as class interval and 1.19 as the lowest class limit.

3) Draw the histogram for the following frequency distribution;

Classes	frequency	Classes	frequency
110 – 119	2	160 – 119	18
120 – 129	4	170 – 179	13
130 – 139	17	180 – 189	6
140 – 149	28	190 – 199	5
150 – 159	25	200 – 209	2

4) Draw the histogram and a polygon for the following frequency distribution;

x	14	16	18	20	22	24
F	20	22	30	25	13	4

5) The heights of college students are given below;

Heights	14	16	18	20	22	24
Students	20	22	30	25	13	4

Draw a histogram and ogive.

6) The weights of 50 football players are listed below;
193 240 217 283 268 212 251 263 275 208
230 288 259 225 252 230 243 247 280 234
250 236 277 218 245 268 231 269 224 259
258 231 255 228 202 245 246 271 249 255
265 235 243 219 255 245 238 257 254 284
Make a stem and leaf display for the data and convert it to a frequency table with 10 class beginning with 190. Also make its frequency distribution.

7) The data are the distances (in kilometers) from a home to local supermarkets.
Create a stemplot using the data:1.1; 1.5; 2.3; 2.5; 2.7; 3.2; 3.3; 3.3; 3.5; 3.8; 4.0; 4.2; 4.5; 4.5; 4.7; 4.8; 5.5; 5.6; 6.5; 6.7; 12.3
Do the data seem to have any concentration of values?
HINT: The leaves are to the right of the decimal.

8) The following data show the distances (in miles) from the homes of off-campus statistics students to the college. Create a stemplot using the data and identify any outliers:
0.5; 0.7; 1.1; 1.2; 1.2; 1.3; 1.3; 1.5; 1.5; 1.7; 1.7; 1.8; 1.9; 2.0; 2.2; 2.5; 2.6; 2.8; 2.8; 2.8; 3.5; 3.8; 4.4; 4.8; 4.9; 5.2; 5.5; 5.7;5.8; 8.0

9) Draw simple bar chart to represent the production of commodity "A" during the years 2000 to 2008:

Years	Product	Years	Product
2000	115	2005	145
2001	113	2006	190
2002	110	2007	210
2003	135	2008	258
2004	100		

10) The following table shows disability in simple population;

Type of disability	Number of Persons
Blind	13
Deaf and dumb	26
Crippled	41
Other handicapped	33

Draw a simple bar chart.

11) The following table gives the birth and death rates per thousand of few countries. Represent this data by multiple bar chart;

Country	Birth rate	Death rate
India	33	24
Japan	32	19
Germany	16	10
Egypt	44	24
Australia	20	9
New Zealand	18	8
France	21	16
Russia	38	16

12) The table shows that quantities in hundreds of Kg of wheat, barley and oats produced on certain farm during the year 1971 – 75;

Years	Wheat	Barley	Oats
1971	34	18	27
1972	43	14	24
1973	43	16	27
1974	45	16	27
1975	50	13	34

Construct the percentage component bar chart to illustrate the data.
Also draw a multiple bar chart.

13) Represent the data by a Pie Chart;

Districts	LHR	MTN	RWP	DGK
Area	50	115	135	165

14) Represent the data by a Pie Chart;

Items	Food	Clothing	House Rent	Edu	Misc.
Expenditure	75	50	30	25	20

15) Obtain the equation of regression line Y on X between the given values;

X	78	77	85	88	83	83	82
Y	84	80	82	83	88	90	88
X	78	76	83	97	98		
Y	91	83	89	78	96		

16) Obtain the equation of regression line X on Y between the given values;

X	78	77	85	88	83	83	82
Y	84	80	82	83	88	90	88
X	78	76	83	97	98		
Y	91	83	89	78	96		

17) Price indices of cotton X and wool Y are given below for the 12 months of a year. Obtain the correlation coefficient between X and Y and obtain the equation of the lines of regression between indices.

X	78	77	85	88	83	83	82
Y	84	80	82	83	88	90	88
X	78	76	83	97	98		
Y	91	83	89	78	96		

18) Calculate the correlation coefficient between percentage of marks scored by 12 students in statistics X and economics Y.

x	50	54	56	59	60	61
y	22	25	34	28	26	30
x	62	65	67	71	71	74
y	32	30	28	36	36	60

19) Calculate the correlation coefficient between supply and demand from the following data;

x	400	200	700	100	500	300	600
y	60	30	70	10	40	20	52

20) Obtain D_3, D_7 and D_8 from the following data;

i. 127,113,132,128,125,130,119,117,121

ii. 121,115,79,52,102,126,81,65,109,119,115,121,103,75,59,110

21) Find2nd, 3^{rd}, 4^{th} , 5^{th}, 6^{th}, 7^{th} and 8^{th}deciles from the following data;

Class	**Frequency**	**Class**	**Frequency**
10 – 20	7	50 – 60	18
20 – 30	10	60 – 70	10
30 – 40	16	70 – 80	5
40 – 50	24	80 – 90	5

22) Obtain P_{38}, P_{45}, P_{67} and P_{86} from the following data;

i. 127,113,132,128,125,130,119,117,121

ii. 121,115,79,52,102,126,81,65,109,119,115,121,103,75,59,110

23) Find$47^{th}$and 83^{th}percentiles from the following data;

Class	**Frequency**	**Class**	**Frequency**
15 – 30	2	111 – 126	15
31 – 46	5	127 – 142	11
47 – 62	9	143 – 158	8
63 – 78	13	159 – 174	6
79 – 94	18	175 – 190	3
95 – 110	25		

24) The following data are number of pages in 40 books on a shelf. Construct a box plot
136; 140; 178; 190; 205; 215; 217; 218; 232; 234; 240; 255; 270; 275; 290; 301; 303; 315; 317; 318; 326; 333; 343;349; 360; 369; 377; 388; 391; 392; 398; 400; 402; 405; 408; 422; 429; 450; 475; 512

25) Graph a box-and-whisker plot for the data values shown.
0; 5; 5; 15; 30; 30; 45; 50; 50; 60; 75; 110; 140; 240; 330

26) The following data are the heights of 40 students in a statistics class.
59; 60; 61; 62; 62; 63; 63; 64; 64; 64; 65; 65; 65; 65; 65; 65; 65; 65; 66; 66; 67; 67; 68; 68; 69; 70; 70; 70; 70; 70; 71; 71; 72; 72; 73; 74; 74; 75; 77
Construct a box plot with the following properties; the calculator instructions for the minimum and maximum values as well as the quartiles follow the example.

- Minimum value = 59
- Maximum value = 77
- Q_1: First quartile = 64.5
- Q_2: Second quartile or median= 66
- Q_3: Third quartile = 70

27) A population consists of 2,3,6,8. Take all possible sample of size 2 with replacement. Form the sampling distribution of sample mean $\bar{x}$. Also obtain mean and standard deviation. Verify the results with population mean and standard deviation.

28) For the population consisting of 4,6,8,10,12. Take all possible sample of size 2 with replacement. Find the means of these samples and make frequency distribution of the sample means. Calculate the mean and variance of this frequency distribution and compare it with the mean and variance of the population.

29) A population consists of 1000 students has a height distribution with $\sigma = 3$. Find the standard error of mean height for a random sample of 50 students selected without replacement.

30) A population consists of 1000 students has a height distribution with $\sigma = 3$. Find the standard error of mean height for a random sample of 50 students selected with replacement.

31) A population consists of 6 members 2,4,6,8,10 and 12. Take all possible sample of size 2 with replacement and without replacement. Form the sampling distribution of means. Find its means and variance. Compare it with population mean and variance.

32) A random sample of n = 25 values given $\bar{x} = 83$. Can this sample be regarded as drawn from a normal population with mean $\mu = 80$ and $\sigma = 7$ at 5% level of significance.

33) Confidence Interval for Mean (Z – test): An auditor has selected a simple random sample of 100 accounts from the 8042 accounts receivable of a freight company to estimate the total audit amount of the receivable in the population. The sample mean is 33.19 and the sample standard deviation is $\hat{s} = 34.48$. Obtain the 95.44 percent confidence interval for the mean audit amount in the population. Hint: use $\bar{x} \pm Z_{\frac{\alpha}{2}} \frac{\hat{s}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$ with $Z_{\frac{\alpha}{2}} = 2.00$.

34) Confidence Interval for Mean (Z – test): Find a 90 percent confidence interval for the mean of a normal distribution with $\sigma = 3$ given the sample as 2.3, – 0.2, – 0.4, – 0.9. Hint: use $\bar{x} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ with $Z_{\frac{\alpha}{2}} = 1.645$.

35) Confidence Interval for Mean (t – test): A random sample of seven observations of a normal variable gave $\sum x = 35.90, \sum x^2 = 186.19.$ Calculate a 90% confidence interval for the population mean μ . Hint: use $\bar{x} \pm t_{\frac{\alpha}{2}(v)} \frac{\hat{s}}{\sqrt{n}}$.

36) Confidence Interval for Mean (t – test): A random sample of twelve measurements of the breaking strength of cotton threads gave a mean $\bar{x} = 209$ grams and a standard deviation $\hat{s} = 35$ grams. Calculate 95% and 99% confidence limits for the actual mean breaking strength. Hint: use $\bar{x} \pm t_{\frac{\alpha}{2}(v)} \frac{\hat{s}}{\sqrt{n}}$.

37) We wish to test the hypothesis that the mean weight of a population of people is 140 **lb**, using $\sigma = 15$ **lb**, $\alpha = 0.05$ **lb** and a sample of 36 people, find the values of $\bar{x}$ which would lead to rejection of the hypothesis.

38) A sample of 400 males students is found to have a mean height of 67.47 inches. Can it be regarded as a simple random sample from a large population with mean height 67.39 with standard deviation of 1.3 inches?

39) Injection of certain type of hormone into hens is said to increase the mean weight of eggs by 0.3 oz. A sample of 30 eggs has an arithmetic mean 0.4 oz above the pre injection mean and a value of $\hat{s}$ equal to 0.20. Is this enough reason to accept the statement that the mean increase is more than 0.3 oz?

40) A random sample of 25 hens from a normal population showed that the average laying is 272 eggs per year with a variance of 625 eggs. The company claimed that the average laying is at least 285 eggs per year. Test the claim of the company at $\alpha = 0.05$.

41) A random sample of size n is drawn from normal population with mean 5 and variance σ^2 . If n = 9, $\bar{x} = 2$ and $t = -2$ what is $\hat{s}$?

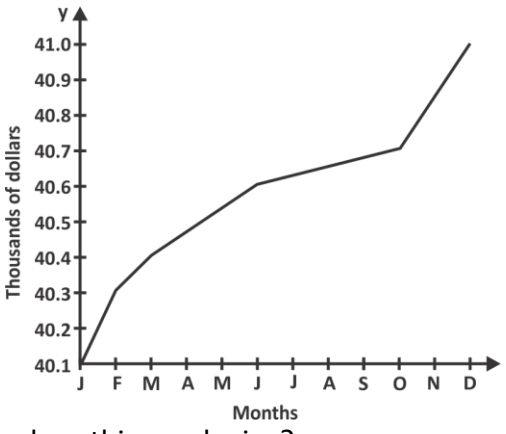
42) A random sample of size n is drawn from normal population with mean 5 and variance σ^2. If n = 25, $\hat{s} = 10$ and $t = 2$ what is $\bar{x}$?

43) Blood Pressure A blood pressure test was given to450 women ages 20 to 36. It showed that their mean systolic blood pressure was 119.4 mm Hg, with a standard deviation of 13.2 mm Hg.

a. Determine the z-score, to the nearest hundredth, for a woman who had a systolic blood pressure reading of 110.5 mm Hg.

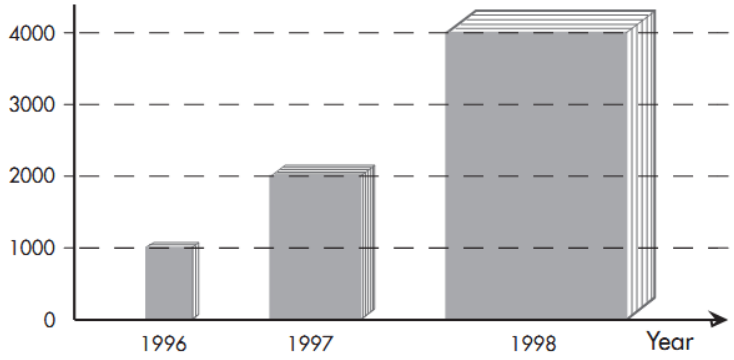
b. The z-score for one woman was 2.15.What was her systolic blood pressure reading?

44) The Brown Disc company presented the following graph to show its profit over the year.



a) What impression does this graph give?
b) What is misleading about this graph?
c) Why do you think the company would present this graph?
d) What does your version of the graph show?

45) This graph shows the sale of books in a store over three years.



a) In what way is this graph misleading?
b) Give the number of books sold in:1996, 1997, 1998

46) The heights of all of the players on a basketball team are shown in the table below. Calculate the standard deviation of the population.

Player	Height
Laura	183
Jamie	165
Deepa	148
Colleen	146
Ingrid	181
Justiss	178
Sheila	154

47) Felix and Melaniehave a job laying patio stones. Their boss in interested in whothe better worker is so randomly throughout the week he chooses a few hours to record how many stones each of the workers lays. The data is recorded in the table below:

Felix	34	41	40	38	38	45
Melanie	51	28	36	44	41	46

Calculate themean and standard deviation of each sample and compare use them to compare the two workers.

48) Vet Data. A small animal veterinarian reviews her records for the day and notes that she has seen eight dogs and eight cats with the following weights (in pounds):
Dogs: 12, 18, 26, 33, 41, 56, 74, 109
Cats: 4, 5, 9, 9, 12, 15, 21, 22
a. Before analyzing these data sets, make a conjecture about which set has the larger mean, median, and standard deviation. Explain your reasoning.
b. Compute the mean and median of each set.
c. Compute the standard deviation of each set.

49) Birth Weight Data. A nurse rotates between maternity wards in two different hospitals. During one shift at Healing Hospital and one shift at Healthy Hospital, she records the following weights (in pounds) of newborn babies.
Healing: 6.7, 7.5, 8.1, 8.6, 8.8, 9.0. 9.1
Healthy: 5.9, 6.6, 7.0, 7.2, 7.7, 8.0, 8.5
a. Find the mean and median of each data set.
b. Find the standard deviation of each data set.
c. Draw a boxplot for each data set, and give a possible explanation for the differences you observe.

50) Find the range of given observations: 32, 41, 28, 54, 35, 26, 23, 33, 38, 40.

51) Following are the marks of students in Mathematics: 50, 53, 50, 51, 48, 93, 90, 92, 91, 90. Find the range of the marks.

52) Calculate the range for the given frequency distribution.

Class Interval	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70	70 – 80
Frequency	2	3	14	8	3	8	2

53) Find the range of the following data.

CI	16 – 20	21 – 25	26 – 30	31 – 35	36 – 40	41 – 45	46 – 50	51 – 55
f	5	6	12	14	26	12	16	9