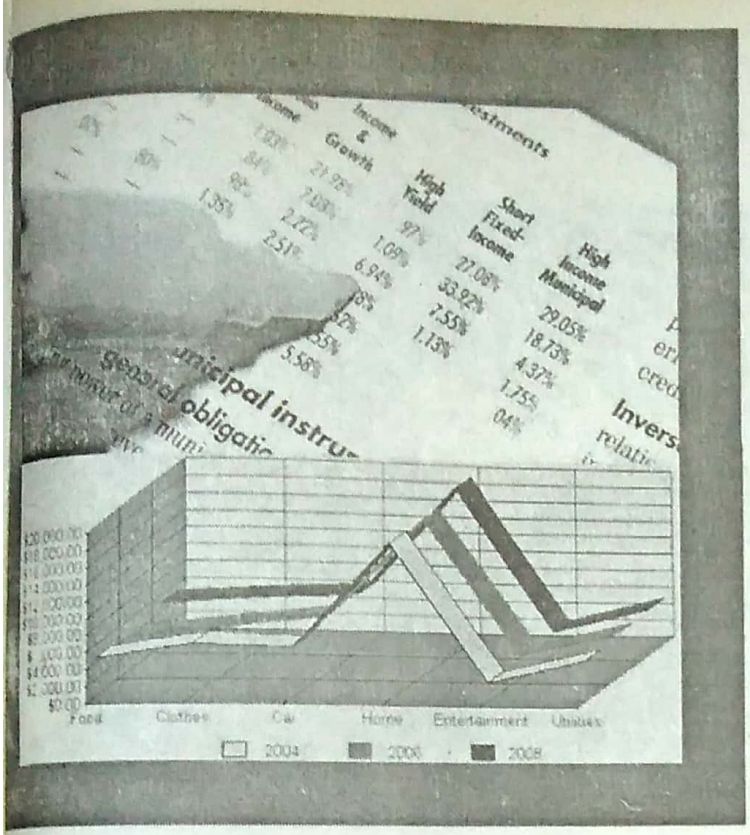


CHAPTER 1

The Nature of Probability and Statistics

Objectives

After completing this chapter, you should be able to

1. Demonstrate knowledge of statistical terms.
2. Differentiate between the two branches of statistics.
3. Identify types of data.
4. Identify the measurement level for each variable.
5. Identify the four basic sampling techniques.
6. Explain the difference between an observational and an experimental study.
7. Explain how statistics can be used and misused.
8. Explain the importance of computers and calculators in statistics.

Outline

Interesting Fact

Every day in the United States about 120 golfers claim that they made a hole-in-one.

game, or the number of hits a baseball player gets in a season. In other areas, such as public health, an administrator might be concerned with the number of residents who contract a new strain of flu virus during a certain year. In education, a researcher might want to know if new methods of teaching are better than old ones. These are only a few examples of how statistics can be used in various occupations.

Furthermore, statistics is used to analyze the results of surveys and as a tool in scientific research to make decisions based on controlled experiments. Other uses of statistics include operations research, quality control, estimation, and prediction.

Statistics is the science of conducting studies to collect, organize, summarize, analyze, and draw conclusions from data.

Historical Note

A Scottish landowner and president of the Board of Agriculture, Sir John Sinclair introduced the word *statistics* into the English language in the 1798 publication of his book on a statistical account of Scotland. The word *statistics* is derived from the Latin word *status*, which is loosely defined as a statesman.

Students study statistics for several reasons:

1. Like professional people, you must be able to read and understand the various statistical studies performed in your fields. To have this understanding, you must be knowledgeable about the vocabulary, symbols, concepts, and statistical procedures used in these studies.
2. You may be called on to conduct research in your field, since statistical procedures are basic to research. To accomplish this, you must be able to design experiments; collect, organize, analyze, and summarize data; and possibly make reliable predictions or forecasts for future use. You must also be able to communicate the results of the study in your own words.
3. You can also use the knowledge gained from studying statistics to become better consumers and citizens. For example, you can make intelligent decisions about what products to purchase based on consumer studies, about government spending based on utilization studies, and so on.

These reasons can be considered some of the goals for studying statistics.

It is the purpose of this chapter to introduce the goals for studying statistics by answering questions such as the following:

What are the branches of statistics?

What are data?

How are samples selected?

1–1 Descriptive and Inferential Statistics

Objective 1

Demonstrate knowledge of statistical terms.

To gain knowledge about seemingly haphazard situations, statisticians collect information for *variables,* which describe the situation.

A **variable** is a characteristic or attribute that can assume different values.

Data are the values (measurements or observations) that the variables can assume. Variables whose values are determined by chance are called **random variables.**

Suppose that an insurance company studies its records over the past several years and determines that, on average, 3 out of every 100 automobiles the company insured were involved in accidents during a 1-year period. Although there is no way to predict the specific automobiles that will be involved in an accident (random occurrence), the company can adjust its rates accordingly, since the company knows the general pattern over the long run. (That is, on average, 3% of the insured automobiles will be involved in an accident each year.)

A collection of data values forms a **data set.** Each value in the data set is called a **data value** or a **datum.**

Objective 2
Differentiate between the two branches of statistics.

Data can be used in different ways. The body of knowledge called statistics is sometimes divided into two main areas, depending on how data are used. The two areas are

1. Descriptive statistics
2. Inferential statistics

Descriptive statistics consists of the collection, organization, summarization, and presentation of data.

Historical Note
The origin of descriptive statistics can be traced to data collection methods used in censuses taken by the Babylonians and Egyptians between 4500 and 3000 B.C. In addition, the Roman Emperor Augustus (27 B.C.–A.D. 17) conducted surveys on births and deaths of the citizens of the empire, as well as the number of livestock each owned and the crops each citizen harvested yearly.

In *descriptive statistics* the statistician tries to describe a situation. Consider the national census conducted by the U.S. government every 10 years. Results of this census give you the average age, income, and other characteristics of the U.S. population. To obtain this information, the Census Bureau must have some means to collect relevant data. Once data are collected, the bureau must organize and summarize them. Finally, the bureau needs a means of presenting the data in some meaningful form, such as charts, graphs, or tables.

The second area of statistics is called *inferential statistics.*

Inferential statistics consists of generalizing from samples to populations, performing estimations and hypothesis tests, determining relationships among variables, and making predictions.

Here, the statistician tries to make inferences from *samples* to *populations.* Inferential statistics uses **probability,** i.e., the chance of an event occurring. You may be familiar with the concepts of probability through various forms of gambling. If you play cards, dice, bingo, or lotteries, you win or lose according to the laws of probability. Probability theory is also used in the insurance industry and other areas.

It is important to distinguish between a sample and a population.

A **population** consists of all subjects (human or otherwise) that are being studied.

Most of the time, due to the expense, time, size of population, medical concerns, etc., it is not possible to use the entire population for a statistical study; therefore, researchers use samples.

A **sample** is a group of subjects selected from a population.

Historical Note
Inferential statistics originated in the 1600s, when John Graunt published his book on population growth, *Natural and Political Observations Made upon the Bills of Mortality.* About the same time, another mathematician/astronomer, Edmund Halley, published the first complete mortality tables. (Insurance companies use mortality tables to determine life insurance rates.)

If the subjects of a sample are properly selected, most of the time they should possess the same or similar characteristics as the subjects in the population. The techniques used to properly select a sample will be explained in Section 1–3.

An area of inferential statistics called **hypothesis testing** is a decision-making process for evaluating claims about a population, based on information obtained from samples. For example, a researcher may wish to know if a new drug will reduce the number of heart attacks in men over 70 years of age. For this study, two groups of men over 70 would be selected. One group would be given the drug, and the other would be given a placebo (a substance with no medical benefits or harm). Later, the number of heart attacks occurring in each group of men would be counted, a statistical test would be run, and a decision would be made about the effectiveness of the drug.

Statisticians also use statistics to determine *relationships* among variables. For example, relationships were the focus of the most noted study in the 20th century, "Smoking and Health," published by the Surgeon General of the United States in 1964. He stated that after reviewing and evaluating the data, his group found a definite relationship between smoking and lung cancer. He did not say that cigarette smoking actually causes lung cancer, but that there is a relationship between smoking and lung cancer. This conclusion was based on a study done in 1958 by Hammond and Horn. In this study, 187,783 men were observed over a period of 45 months. The death rate from

Only one-third of crimes committed are reported to the police.

80 to 90% of the time usually received a B or C in the class. Students who attended class less than 80% of the time usually received a D or an F or eventually withdrew from the class.

Based on this information, attendance and grades are related. The more you attend class, the more likely it is you will receive a higher grade. If you improve your attendance, your grades will probably improve. Many factors affect your grade in a course. One factor that you have considerable control over is attendance. You can increase your opportunities for learning by attending class more often.

1. What are the variables under study?
2. What are the data in the study?
3. Are descriptive, inferential, or both types of statistics used?
4. What is the population under study?
5. Was a sample collected? If so, from where?
6. From the information given, comment on the relationship between the variables.

See page 33 for the answers.

1–2 Variables and Types of Data

Objective 3
Identify types of data.

As stated in Section 1–1, statisticians gain information about a particular situation by collecting data for random variables. This section will explore in greater detail the nature of variables and types of data.

Variables can be classified as qualitative or quantitative. **Qualitative variables** are variables that can be placed into distinct categories, according to some characteristic or attribute. For example, if subjects are classified according to gender (male or female), then the variable *gender* is qualitative. Other examples of qualitative variables are religious preference and geographic locations.

Quantitative variables are numerical and can be ordered or ranked. For example, the variable *age* is numerical, and people can be ranked in order according to the value of their ages. Other examples of quantitative variables are heights, weights, and body temperatures.

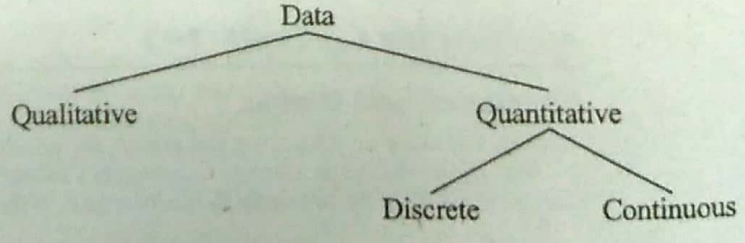
Quantitative variables can be further classified into two groups: discrete and continuous. *Discrete variables* can be assigned values such as 0, 1, 2, 3 and are said to be countable. Examples of discrete variables are the number of children in a family, the number of students in a classroom, and the number of calls received by a switchboard operator each day for a month.

Discrete variables assume values that can be counted.

Continuous variables, by comparison, can assume an infinite number of values in an interval between any two specific values. Temperature, for example, is a continuous variable, since the variable can assume an infinite number of values between any two given temperatures.

Continuous variables can assume an infinite number of values between any two specific values. They are obtained by measuring. They often include fractions and decimals.

The classification of variables can be summarized as follows:



Date

Frequency Distributions and Graphs

1. Introduction

When conducting a statistical study, the researcher must gather data for the particular variable under study. To describe situations and draw conclusions, the researcher must organize the data in some meaningful way. The most convenient method is to construct a freq. distt.
After organizing the data the researcher must present them so they can be understood easily. The most useful method of presenting the data is by constructing statistical charts and graphs. There are many different types of charts and graphs, and each one has a specific purpose.

2. Frequency Distribution

A frequency distribution is the organization of raw data in table form, using classes and frequencies.

3. Categorical Frequency Distribution

The categorical freq. distt. is used for data that can be placed in specific categories, such as nominal or ordinal level data. For example data such as political affiliation, religious affiliations, grades etc could use categorical freq. distt.

4. Example

Twenty-five army inductees were given a blood test to determine their blood type. The data set is

A	B	B	AB	O
O	O	B	AB	B
B	B	O	A	O
A	O	O	O	AB
AB	A	O	B	A

Construct a frequency distt. for this data.

Solution

Since the data are categorical, discrete classes can be used. There are four blood types. These types will be used as the classes for the distt.

A class	B Tally	C Frequency	D Percent
A	𝍸	5	$\% = \frac{f}{n} \times 100 = \frac{5}{25} \times 100 = 20$
B	𝍸 ||	7	$\frac{7}{25} \times 100 = 28$
O	𝍸 ||||	9	$\frac{9}{25} \times 100 = 36$
AB	||||	4	$\frac{4}{25} \times 100 = 16$
Total		25	100

EX- 2-1

Q7. A survey was taken on how much trust people place in the information they read on the internet. Construct a categorical freq. distt. for the data

A: trust in everything they read.
M: " " most of what they read.
H: " " one-half " " " ...
S: " " small portion " " ...

M M M A H M S M H M
S M M M M A M M A M
M M H M M M H M H M
A M M M H M M M M M

Soln. Discrete classes for this categorical data are A, M, H and S.

class	Tally	Frequency	Percent.
A	IIII	4	$\frac{4}{40} \times 100 = 10$
M	++++ ++++ ++++ ++++ IIII	28	$\frac{28}{40} \times 100 = 70$
H	++++ I	6	$\frac{6}{40} \times 100 = 15$
S	II	2	$\frac{2}{40} \times 100 = 5$
		$\Sigma f = 40$	100

Example Miles Run per week

Construct a histogram, frequency polygon, and ogive using relative frequencies for the distribution of miles that 20 randomly selected runners ran during a given week.

Class Boundaries	Frequency
5.5 – 10.5	1
10.5 – 15.5	2
15.5 – 20.5	3
20.5 – 25.5	5
25.5 – 30.5	4
30.5 – 35.5	3
35.5 – 40.5	2
	20

Soln. Step 1 Convert each frequency to a relative freq. (Divide freq of each class by total observation)

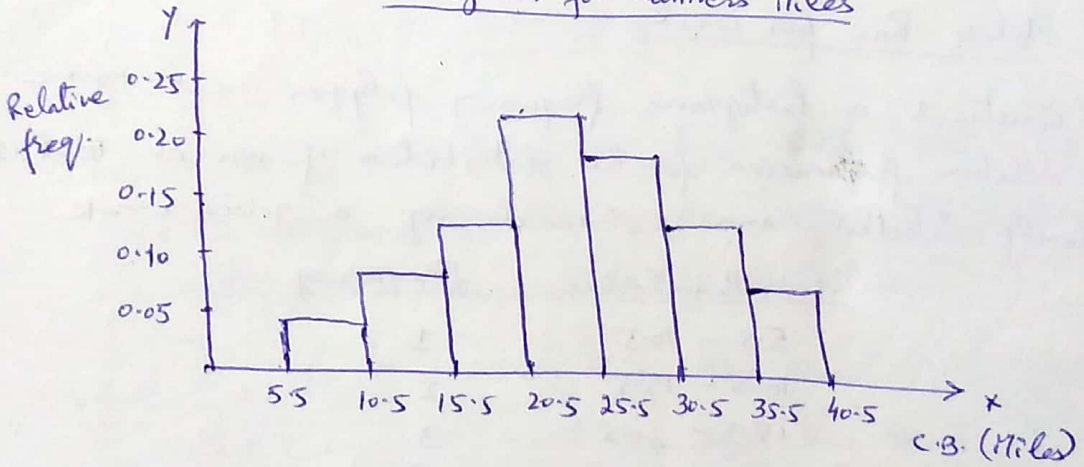
Class Boundaries	Midpoints	Relative Freq.
5.5 – 10.5	8	$1 \div 20 = 0.05$
10.5 – 15.5	13	$2 \div 20 = 0.10$
15.5 – 20.5	18	$3 \div 20 = 0.15$
20.5 – 25.5	23	$= 0.25$
25.5 – 30.5	28	$= 0.20$
30.5 – 35.5	33	$= 0.15$
35.5 – 40.5	38	$= 0.10$
		1.00

Step 2 Find cumulative relative frequencies.

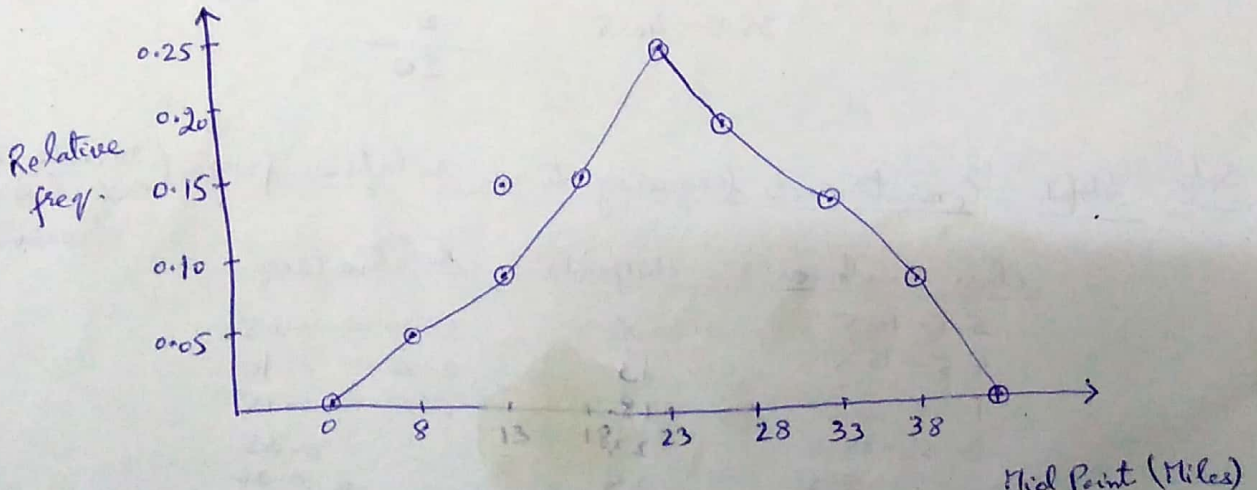
	Cumulative frequency	Cumulative relative frequency
Less than 5.5	0	0.00
Less than 10.5	1	0.05
Less than 15.5	3	0.15
Less than 20.5	6	0.30
Less than 25.5	11	0.55
Less than 30.5	15	0.75
Less than 35.5	18	0.90
Less than 40.5	20	1.00

Step 3 For histogram and ogive, use class boundaries along x-axis.
For freq. polygon use the midpoints on the x-axis.

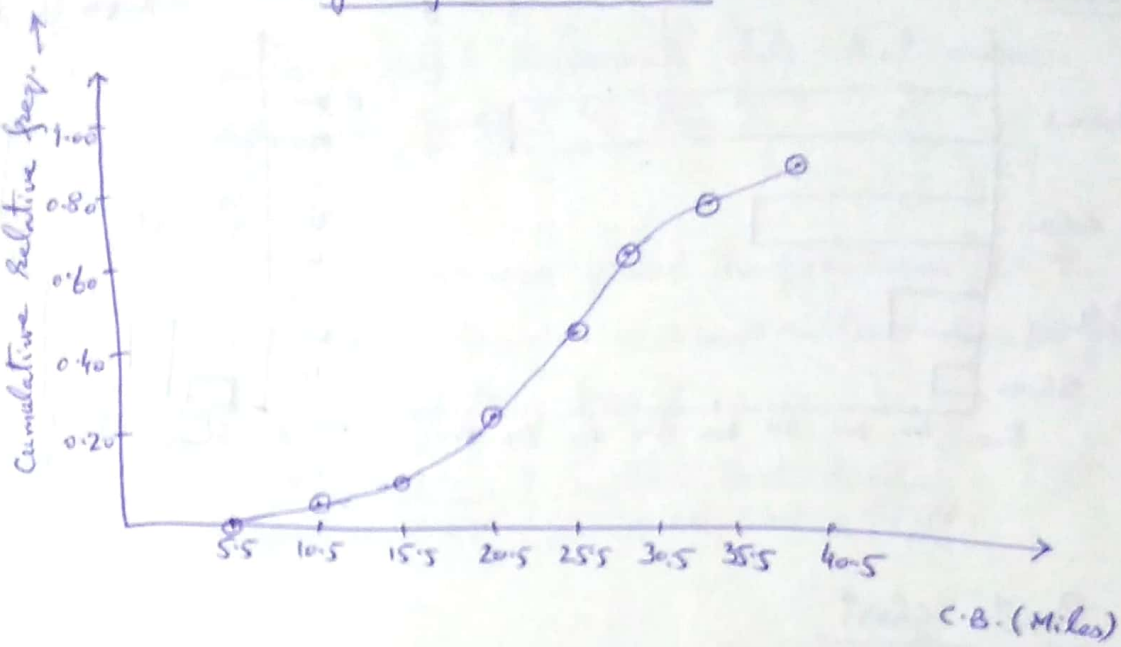
Histogram for Runners Miles



Frequency Polygon for Runners Miles



Ogive for Runner's Miles



Other Types of Graphs

Bar Graph

1. A bar graph represents the data by using vertical or horizontal bars whose heights or lengths represent the frequencies of the data.

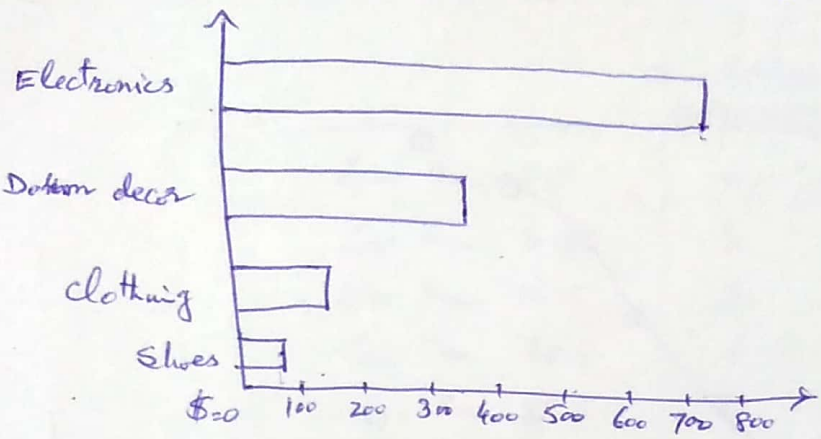
Example

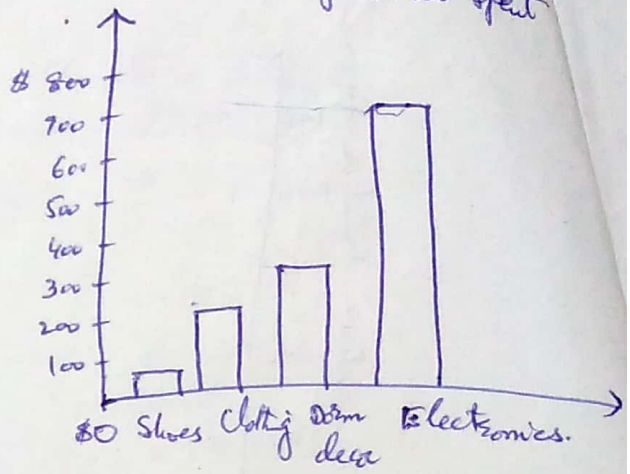
The table shows the average money spent by the first-year college students. Draw a horizontal and vertical bar graph for the data.

Electronics	$ 728
Dorm decor	344
Clothing	141
Shoes	72

~~1st year~~

Solution.





2. Pareto chart

A Pareto chart is used to reprent a frequency distribution for a categorical variable. The frequencies are displayed by the heights of the vertical bars, which are arranged in order from highest to lowest.

3. Example

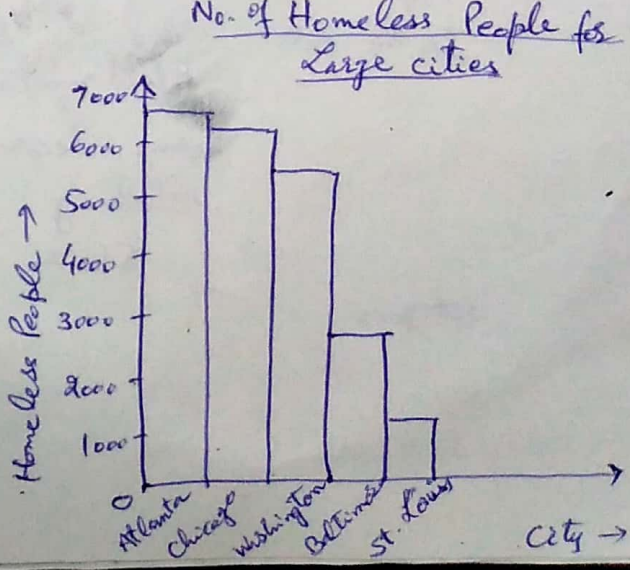
The data shown here consist of the number of homeless people for a sample of selected cities. Construct and analyse a Pareto chart for the data.

City	Number
Atlanta	6832
Baltimore	2904
Chicago	6680
St. Louis	1485
Washington	5518

4. Soln

Arrange the data from largest to smallest according to frequency.

City	Number
Atlanta	6832
Chicago	6680
Washington	5518
Baltimore	2904
St. Louis	1485



5. **The Time Series Graph**

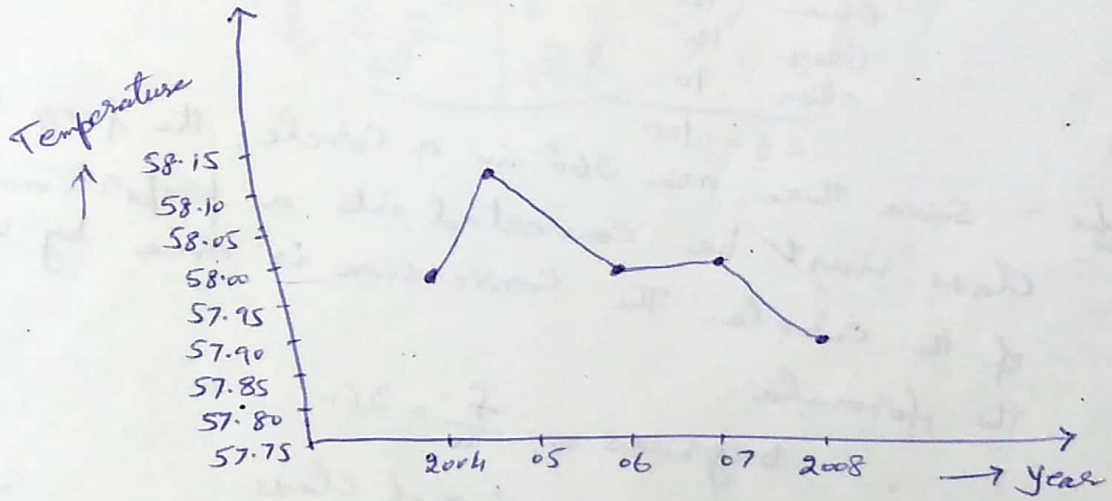
A time series graph represents data that occur over a specific period of time.

6. **Example** (Ex - 2.3 Q.8.)

The ~~number of~~ h average global temperatures for the following years are shown. Draw a time series graph and comment on the trend.

Year	2004	2005	2006	2007	2008
Temperatures	57.98	58.11	57.99	58.01	57.88

7. Soln.



After a slight increase in 2005, the average temperature has declined in the next years.

8. The Pie Graph

A pie graph is a circle that is divided into sections according to the percentage of frequencies in each category for the distribution.

EX- 2-3

Q#2 The popular vehicle car colors are shown. Construct a Pie graph for the data..

White	19%
Silver	18
Black	16
Red	13
Blue	12
Gray	12
other	10
	$\Sigma f = 100$

Solve Since there are 360° in a circle, the freq. of each class must be converted into a proportional part of the circle.. The conversion is done by using the formula

$$\text{Degrees} = \frac{f}{n} \times 360$$

where f = freq. of each class
n = Sum of the frequencies

Each frequency must be converted into %.

$$\% = \frac{f}{n} \times 100$$

White = $\frac{19}{100} \times 100 = 19\%$
Silver = $\frac{18}{100} \times 100$ 18%
Black = 16%
Red = 13%
Blue = 12%
Grey = 12%
others = 10%

White cars = $\frac{19}{100} \times 360 = 68.4^\circ$

Silver Cars = $\frac{18}{100} \times 360 = 64.8^\circ$

Black = $\frac{16}{100} \times 360 = 57.6^\circ$

Red = $\frac{13}{100} \times 360 = 46.8^\circ$

Blue = $\frac{12}{100} \times 360 = 43.2^\circ$

Grey = " " = 43.2°

other = $\frac{10}{100} \times 360 = 36^\circ$

Sum of the Portions = $\Sigma = 360^\circ$

Now using a protector and a compass, draw the graph using appropriate degrees found earlier.

Popular Vehicle Colors