There are several approaches to determining the sample size. These include using a census for small populations, imitating a sample size of similar studies, using published tables, and applying formulas to calculate a sample size.

Using a census for small populations
One approach is to use the entire population as the sample. It's impractical for large populations. A census eliminates sampling error and provides data on all the individuals in the population. Finally, virtually the entire population would have to be sampled in small populations to achieve a desirable level of precision

Using a sample size of a similar study
Another approach is to use the same sample size as those of studies similar to the one you plan. Without reviewing the procedures employed in these studies you may run the risk of repeating errors that were made in determining the sample size for another study. However, a review of the literature in your discipline can provide guidance about "typical" sample sizes which are used.

Using published tables
One can also rely on published tables which provide the sample size for a given set of criteria. Yamane, 1967 Table 2.1 and Table 2.2 present sample sizes that would be necessary for given combinations of precision, confidence levels, and variability.
NB, i) these sample sizes reflect the number of *obtained* responses, and not necessarily the number of surveys mailed or interviews planned (this number is often increased to compensate for non-response). Ii) the sample sizes in Table 2.2 presume that the attributes being measured are distributed normally or nearly so. If this assumption cannot be met, then the entire population may need to be surveyed.

Using formulas to calculate a sample size
Yamane (1967) provides a simplified formula to calculate sample sizes. A 95% confidence level and P = .5 are assumed for this Equation. $n = \frac{N}{1+Ne^2}$ Where n is the sample size, N is the population size, and e is the level of precision

2. Data Presentation

2.1 Frequency Distributions Tables

Definitions
Raw Data: unprocessed data ie data in its original form.
Frequency Distribution: The organization of raw data in table form with classes and frequencies. Rather it's a list of values and the number of times they appear in the data set. We have grouped and ungrouped frequency distribution tables for large and small data sets respectively.

2.1.1 Construction of Ungrouped Frequency Distributions

- Note the largest and smallest observations in the data
- Starting with the smallest value, tally the observations of each quantity.
- Count the number of tallies for each quantity and record it as frequency.

Example: Construct an ungrouped frequency table for the data below
16 14 15 13 12 14 16 15 15 14 17 16 13 16 15 14 18 13 15 17

Solution

Value	12	13	14	15	16	17	18
Tally	/	///	////	𝍸	////	//	/
Frequency	1	3	4	5	4	2	1

Note ; when tallying 𝍸 is used for 5 counts and not /////

2.1.2 Construction of Grouped Frequency Distributions

When the number of observations is too large and/or when the variable of interest is continuous, it's cumbersome to consider the repetition of each observation. A quick and more convenient way is to group the range of values into a number of exclusive groups or classes and count the class frequency. The resulting table is called a grouped frequency distribution table.

A grouped frequency distribution consists of *classes* and their corresponding *frequencies.* Each raw data value is placed into a quantitative or qualitative category called a **class.** The **frequency** of a class then is the number of data values contained in a specific class.

Steps in construction

- Select the number of classes k. Choose the smallest integer k such that
  $2^k > n \Rightarrow k > \dfrac{\log n}{\log 2}$ Eg if n = 30 $k > \dfrac{\log 30}{\log 2} \Rightarrow k = 5$
- Identify the largest and smallest observation and compute the ranrge R=largest –smallest value.
- Identify the smallest unit of measurement (u) used in the data collection (ie the accuracy of the measurement.)

Eg for the data 10 30 20 50 30 60 u=10 For the data 12 15 11 17 13 u=1
For the data 1.6 3.2 2.8 5.6 3.5 1.6 u=0.1

- Estimate the class width/interval. i = round up $\left(\frac{R}{k}\right)$ to the nearest u
  Eg round up 3.13 t0 the nearest 0.1 is 3.2
- Pick the starting value (lower class limit of the 1st class (LCL_1)) as the smallest value used in the computation of R above. Successive LCL_s are got by adding I to the previous LCL
- Find the upper class limit of the 1st class (UCL_1)) by subtracting u from LCL_2. Successive UCL_s are got by adding I to the previous UCL
- If necessary find the class boundaries as follows $LCB = LCL - \frac{1}{2}u \quad and \quad UCB = UCL + \frac{1}{2}u$
- Tally the number of observations falling in each class and record the frequency.
  NB a value x fall into a class $LCL - UCL \quad if \quad LCB \le x \le UCB$

Example 1

Organize the data below into a grouped frequency table.

15.0 17.4 10.3 9.2 20.7 18.9 16.6 22.4 23.7 18.6 26.1 16.5 19.7 12.9 15.7
30.8 15.4 20.3 24.0 29.6 18.3 23.7 17.8 24.6 23.0 21.4 32.8 12.5 17.5 18.3
23.2 21.6 20.8 29.8 24.5 28.4 13.5 17.1 27.1 27.9

Solution

$u = 0.1 \quad n = 40 \Rightarrow k > \frac{\log 40}{\log 2} \Rightarrow k = 6$ Range = 32.8 - 9.2 = 23.6

i = round up $\left(\frac{23.6}{6}\right)$ to the nearest $0.1 = 4.0$

Now $LCL_1 = 9.2 \Rightarrow LCL_2 = 13.2 \quad LCL_3 = 17.2$ etc

$\Rightarrow UCL_1 = LCL_2 - u = 13.2 - 0.1 = 13.1, \quad UCL_2 = 17.1 \quad UCL_3 = 21.1$ etc

The frequency table is as shown belo

Class	Boundaries	tally	Freq	C.F
9.2 - 13-1	9.15 - 13-15	////	4	4
13.2 - 17.1	13.15 - 17.15	𝍸 //	7	11
17.2 - 21-1	17.15 - 21-15	𝍸 𝍸 /	11	22
21.2 - 25.1	21.15 - 25.15	𝍸 𝍸	10	32
25.2 - 29.1	25.15 - 29.15	////	4	36
29.2 - 33-	29.15 - 33-15	////	4	40

Sometimes due to convenience, it may be necessary to slightly lower the starting value. Eg in the above case we may use 9.0 in place of 9.2.

Example 2

These data represent the record high temperatures in degrees Fahrenheit (F) for each of the 50 states. Construct a grouped frequency distribution for the data.

112 100 127 120 134 118 105 110 109 112 110 118 117 116 118 122 114 114 105 109 107 112 114 115 118 117 118 122 106 110 116 108 110 121 113 120 119 111 104 111 120 113 120 117 105 110 118 112 114 114

Solution

$u = 1 \quad n = 50 \Rightarrow k > \frac{\log 50}{\log 2} \Rightarrow k = 6 \qquad \text{Range} = 134 - 100 = 34$

$\text{i} = \text{round up } \left(\frac{34}{6}\right) \text{to the nearest} 1 = 6$

$\text{Now } \text{LCL}_1 = 100 \Rightarrow \text{LCL}_2 = 106 \quad \text{LCL}_3 = 112 \text{ etc}$

$\Rightarrow U\text{CL}_1 = \text{LCL}_2 - u = 106 - 1 = 105, \quad \text{UCL}_2 = 111 \quad \text{UCL}_3 = 117 \text{ etc}$

The frequency table is as shown below

Class	Boundaries	tally	Freq	C.F
100 - 105	99.5 – 105.5	𝍸	5	5
106 - 111	105.5 - 111.5	𝍸 𝍸 ///	13	18
112 - 117	111.5 – 117.5	𝍸 𝍸 𝍸 //	16	34
118 - 123	117.5 -123.5	𝍸 𝍸 ////	14	48
124 - 129	124.5 - 129.5	/	1	49
130 - 135	129.5 - 135-5	/	1	50

Exercise

1) The data shown here represent the number of miles per gallon (mpg) that 30 selected four-wheel-drive sports utility vehicles obtained in city driving. Construct a frequency distribution, and analyze the distribution. 12 17 12 14 16 18 16 18 12 16 17 15 15 16 12 15 16 16 12 14 15 12 15 15 19 13 16 18 16 14
2) Suppose a researcher wished to do a study on the ages of the top 50 wealthiest people in the world. The researcher first would have to get the data on the ages of the people. In this case, these ages are listed in *Forbes Magazine*. 49 57 38 73 81 74 59 76 65 69 54 56 69 68 78 65 85 49 69 61 48 81 68 37 43 78 82 43 64 67 52 56 81 77 79 85 40 85 59 80 60 71 57 61 69 61 83 90 87 74 Organize the data into a grouped frequency table
3) The data represent the ages of our Presidents at the time they were first inaugurated. 57 61 57 57 58 57 61 54 68 51 49 64 50 48 65 52 56 46 54 49 50 47 55 55 54 42 51 56 55 54 51 60 62 43 55 56 61 52 69 64 46 54
   a) Were the data obtained from a population or a sample? Explain your answer.
   b) What was the age of the oldest and youngest President?
   c) Construct a frequency distribution for the data.

d) Are there any peaks in the distribution?
e) identify any possible outliers.

4) The state gas tax in cents per gallon for 25 states is given below. Construct a grouped frequency distribution for the data. 7.5 16 23.5 17 22 21.5 19 20 27.1 20 22 20.7 17 28 20 23 18.5 25.3 24 31 14.5 25.9 18 30 31.5
5) Listed are the weights of the NBA's top 50 players. Construct a grouped frequency distribution and analyze the results in terms of peaks, extreme values, etc.
240 210 220 260 250 195 230 270 325 225 165 295 205 230 250 210 220 210 230 202 250 265 230 210 240 245 225 180 175 215 215 235 245 250 215 210 195 240 240 225 260 210 190 260 230 190 210 230 185 260
6) The number of stories in each of the world's 30 tallest buildings is listed below. Construct a grouped frequency distribution and analyze the results in terms of peaks, extreme values, etc. 88 88 110 88 80 69 102 78 70 55 79 85 80 100 60 90 77 55 75 55 54 60 75 64 105 56 71 70 65 72
7) The average quantitative GRE scores for the top 30 graduate schools of engineering are listed. Construct a grouped frequency distribution and analyze the results in terms of peaks, extreme values, etc. 767 770 761 760 771 768 776 771 756 770 763 760 747 766 754 771 771 778 766 762 780 750 746 764 769 759 757 753 758 746
8) The number of passengers (in thousands) for the leading U.S. passenger airlines in 2004 is indicated below. Use the data to construct a grouped frequency distribution with a reasonable number of classes and comment on the shape of the distribution.
9) 91.570 86. 755 81.066 70.786 55.373 42.400 40.551 21.119 16.280 14.869 13.659 13.417 3.170 12.632 11.731 10.420 10.024 9.122 7.041 6.954 6.406 6.362 5. 930 5.585
10) The heights (in feet above sea level) of the major active volcanoes in Alaska are given here. Construct a grouped frequency distribution for the data. 4,265 3,545 4,025 7,050 11,413 3,490 5,370 4,885 5,030 6,830 4,450 5,775 3,945 7,545 8,450 3,995 10,140 6,050 10,265 6,965 150 8,185 7,295 2,015 5,055 5,315 2,945 6,720 3,465 1,980 2,560 4,450 2,759 9,430 7,985 7,540 3,540 11,070 5,710 885 8,960 7,015

2.2 Graphical Displays

After you have organized the data into a frequency distribution, you can present them in graphical form. The purpose of graphs in statistics is to convey the data to the viewers in pictorial form. It is easier for most people to comprehend the meaning of data presented graphically than data presented numerically in tables or frequency distributions. This is especially true if the users have little or no statistical knowledge.

Statistical graphs can be used to describe the data set or to analyze it. Graphs are also useful in getting the audience's attention in a publication or a speaking presentation. They can be used to discuss an issue, reinforce a critical point, or summarize a data set. They can also be used to discover a trend or pattern in a situation over a period of time.

The commonly used graphs in research are; the pie chart, bar chart, histogram, frequency polygon and the cumulative frequency curve (Ogive).

2.2.1 Pie Chart

It's a circular graph having radii divide a circle into sectors proportional in angle to the relative size of the quantities in the category being represented.

How to Draw

(i) Add up the given quantities and let s be the sum of the values

(ii) For each quantity x, calculate the representative angle and percentage as $\frac{x}{s}(360^0)$ and $\frac{x}{s}(100\%)$ respectively

(iii)Draw a circle and divide it into sectors using the angles calculated in step ii above

(iv)Label the sector by the group represented and indicate the corresponding percentage.

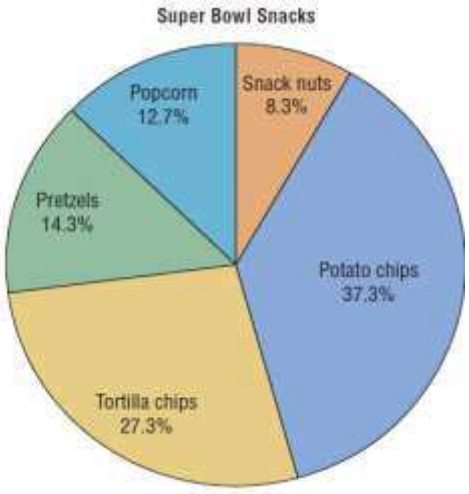
Example

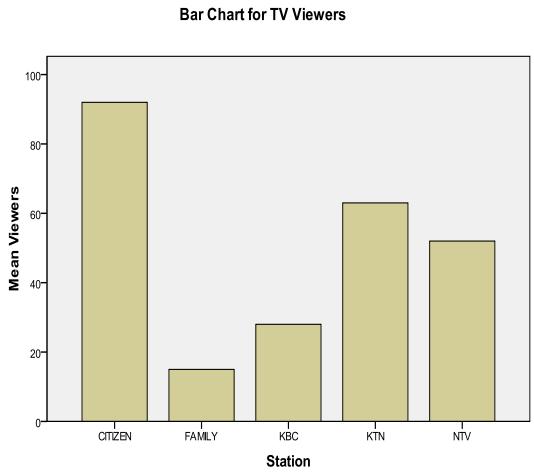
This frequency distribution shows the number of pounds of each snack food eaten during the Super Bowl. Construct a pie graph for the data.

Snack	Potato chips	Tortilla chips	Pretzels	Popcorn	Snack nuts
Pounds (in millions)	11.2	8.2	4.3	3.8	2.5

Solution

Snack	Potato chips	Tortilla chips	Pretzels	Popcorn	Snack nuts	Total
Pounds (in millions)	11.2	8.2	4.3	3.8	2.5	30.0
Representative Angle	134	98	52	46	30	360
Representative %age	37.3	27.3	14.3	12.7	8.3	99.9





.

2.2.2 Bar chart

A bar chart consists of a set of equal spaced rectangles whose heights are proportional to the frequency of the category /item being considered. The X axis in a bar chart can represent the number of categories.

Note: Bars are of uniform width and there is equal spacing between the bars.

Example

A sample of 250 students was asked to indicate their favourite TV channels and their responses were as follows.

TV station	KBC	NTV	CITIZEN	KTN	FAMILY
No. of viewers	28	52	92	63	15

Draw a bar chart to represent this information.

Solution see the bar chart above and on the right.

2.23 Pareto Charts

It consist of a set of continuous rectangles where the variable displayed on the horizontal axis is qualitative or categorical and the frequencies are displayed by the heights of vertical bars, which are arranged in order from highest to lowest. A **Pareto chart** is used to represent a frequency distribution for a categorical variable,
Points to note when drawing a Pareto Chart
i) Make the bars the same width.
ii) Arrange the data from largest to smallest according to frequency.
iii) Make the units that are used for the frequency equal in size.
When you analyze a Pareto chart, make comparisons by looking at the heights of the bars.

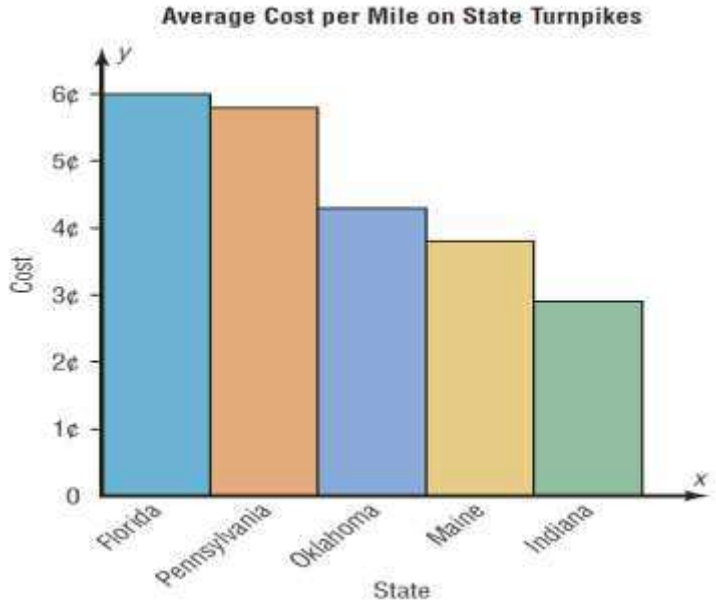
Example
The table shown here is the average cost per mile for passenger vehicles on state turnpikes. Construct and analyze a Pareto chart for the data.

State	Indiana	Oklahoma	Florida	Maine	Pennsylvania
Number	2.9	4.3	6.0	3.8	5.8

Solution
Arrange the data from the largest to smallest according to frequency.

State	Florida	Pennsylvania	Oklahoma	Maine	Indiana
Number	6.0	5.8	4.3	3.8	2.9

Draw and label the *x* and *y* axes and then the bars corresponding to the frequencies. The Pareto chart shows that Florida has the highest cost per mile. The cost is more than twice as high as the cost for Indiana.



2.2.4 Histogram
It consists of a set of continuous rectangles such that the areas of the rectangles are proportional to the frequency. For ungrouped data, the heights of each bar is proportional to frequency. For grouped data, the height of each rectangle is the relative frequency *h* and is given by $h = \frac{\text{frequency}}{\text{IClass Interval}}$. The width of the bars is determined by the class boundaries.

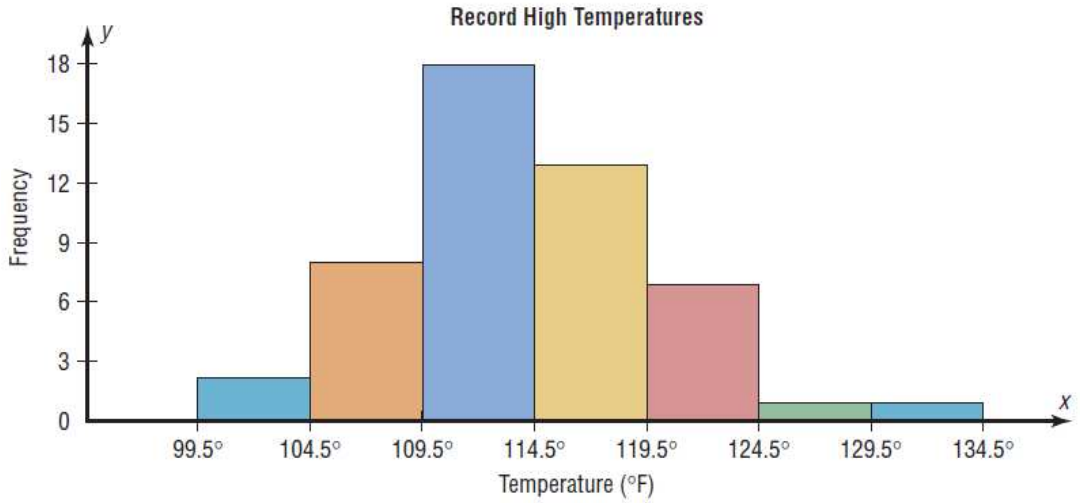
Example
Construct a histogram to represent the data shown below

Class	100-104	105-109	110 -114	115-119	120 - 24	125-129	130 -134

Freq	2	8	18	13	7	1	1

Solution

Boundaries	99.5-104.5	104.5-109.5	109.5 - 114.5	114.5-119.5	*119.5 – 124.5*	*124.5-129.5*	129.5 - 134.5
Heights	2	8	18	13	7	1	1



2.2.5 Frequency polygon

It's a plot of frequency against mid points joined with straight line segments between consecutive points. It can also be obtained by joining the mid point of the tops of the bars in a histogram. The gaps at both ends are filled by extending to the next lower and upper imaginary classes assuming frequency zero.

Example: Consider the following frequency distribution.

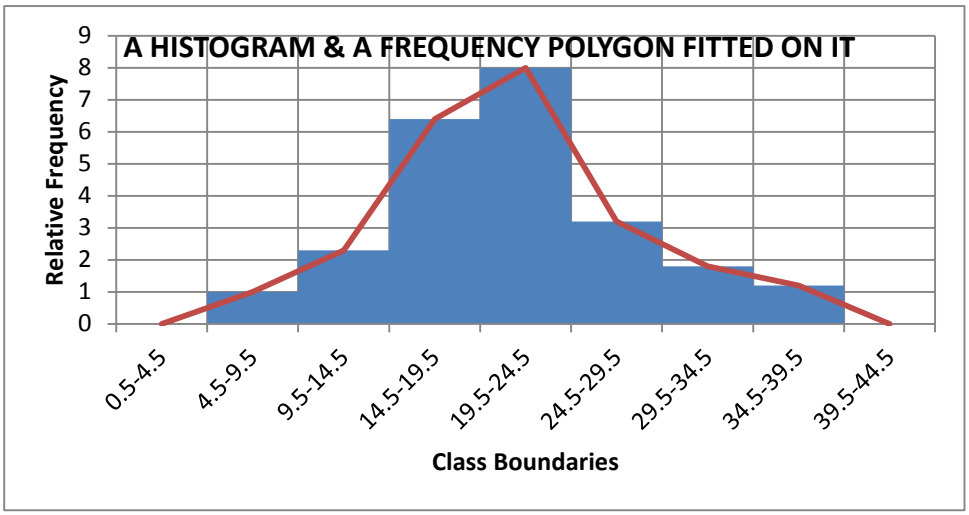
Class	5-9	10-14	15-19	20-24	25-29	30-34	35-39
freq	5	12	32	40	16	9	6

Draw a histogram to represent this information and fit a frequency polygon on it.

Solution

Boundaries	4.5-9.5	9.5-14.5	14.5-19.5	19.5-24.5	24.5-29.5	29.5-34.5	34.5-39.5
heights	1	2.4	6.4	8	3.2	1.8	1.2

The corresponding histogram is as shown below.



2.2.6 Cumulative frequency curve (ogive)

It is a plot of cumulative frequency against upper boundaries joined with a smooth curve. The gap on the lower end is filled by extending to the next lower imaginary class assuming frequency zero. This graph is useful in estimating median and other measures of location.

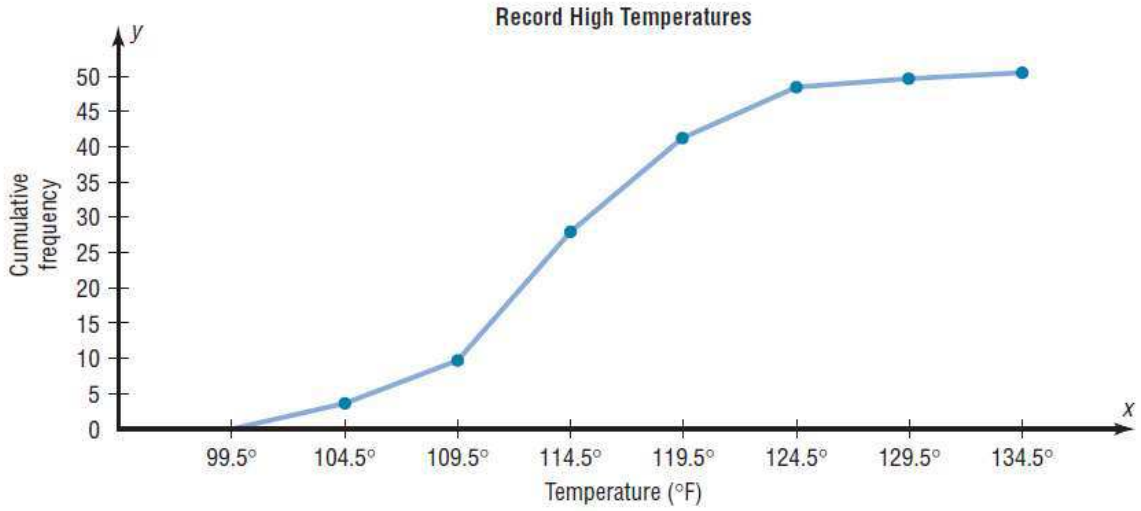
Example:

Construct an ogive to represent the data shown below

Class	100-104	105-109	110 -114	115-119	120 - 24	125-129	130 -134
Freq	2	8	18	13	7	1	1

Solution

Upper Boundaries	**99.5**	104.5	109.5	114.5	*119.5*	*124.5*	*129.5*	134.5
CF	**0**	2	10	28	41	48	49	50



Exercise

1) Construct a pie chart and a bar graph showing the blood types of the army inductees described in the frequency distribution is repeated here.

Blood group	A	B	AB	O
Frequency	5	7	4	9

2) The table below shows the average money spent by first-year college students. Draw a pie chart and a bar graph for the data.

Nature of Expense	Electronics	Dorm decor	Clothing	Shoes
Amount(in $)	728	344	141	72

3) The table shown here is the average cost per mile for passenger vehicles on state turnpikes. Draw a pie chart and a bar graph for the data.

State	Indiana	Oklahoma	Florida	Maine	Pennsylvania
Number	2.9	4.3	6.0	3.8	5.8

4) The following data are based on a survey from American Travel Survey on why people travel. Construct a pie chart a bar graph and a Pareto chart for the data and analyze the results.

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

5) The following percentages indicate the source of energy used worldwide. Construct a Pareto chart and a vertical pie chart, a bar graph and a Pareto graph for the energy used.

Energy Type	Petroleum	Coal	Dry natural gas	Hydroelectric	Nuclear	Others
percentage	39.8	23.2	22.4	7.0	6.4	1.2

6) The following elements comprise the earth's crust, the outermost solid layer. Illustrate the composition of the earth's crust with a pie chart and a bargraph for this data.

Element	Oxygen	Silicon	Aluminum	Iron	Calcium	Others
percentage	45.6	27.3	8.4	6.2	4.7	7.8

7) The sales of recorded music in 2004 by genre are listed below. Represent the data with an appropriate graph.

Rock	Country	Rap/hip-hop	R&B/urban	Pop	Religious	Children's	Jazz	Classical	Oldies	Soundtracks	New age	Others
23.9	13.0	12.1	11.3	10.0	6.0	2.8	2.7	2.0	1.4	1.1	1.0	8.9

8) The top 10 airlines with the most aircraft are listed. Represent these data with an appropriate graph.

American	Continental	United	Southwest	Northwest	American Eagle	U.S. Airways	Lufthansa (Ger.)
714	364	603	327	424	245	384	233

9) The top prize-winning countries for Nobel Prizes in Physiology or Medicine are listed here. Represent the data with an appropriate graph.

United States	Denmark	United Kingdom	Austria	Germany	Belgium	Sweden	Italy	France	Australia	Switzerland
80	5	24	4	16	4	8	3	7	3	6

10) Construct a histogram, frequency polygon, and an ogive for the distribution (shown here) of the miles that 20 randomly selected runners ran during a given week.

Class	6-10	11-15	16 -20	21-25	26 - 30	31-35	36 -40
Freq	1	2	3	5	4	3	2

11) For 108 randomly selected college applicants, the following frequency distribution for entrance exam scores was obtained. Construct a histogram, frequency polygon, and ogive for the data.

Class	90-98	99-107	108-116	117125	126-134
Freq	6	22	43	28	9

Applicants who score above 107 need not enrol in a summer developmental program. In this group, how many students do not have to enroll in the developmental program?

12) Thirty automobiles were tested for fuel efficiency, in miles per gallon (mpg). The following frequency distribution was obtained. Construct a histogram, a frequency polygon, and an ogive for the data.

Class	8-12	13-17	18-22	23-27	28-32
Freq	3	5	15	5	2

13) The salaries (in millions of dollars) for 31 NFL teams for a specific season are given in this frequency distribution.

Class	39.9-42.8	42.9-45.8	45.9-48.8	48.9-51.8	51.9-54.8	54.9-57.8
Freq	2	2	5	5	12	5

Construct a histogram, a frequency polygon, and an ogive for the data; and comment on the shape of the distribution.

14) In a study of reaction times of dogs to a specific stimulus, an animal trainer obtained the following data, given in seconds. Construct a histogram, a frequency polygon, and an ogive for the data; analyze the results.

Class	2.3-2.9	3.0-3.6	3.7-4.3	4.4-5.0	5.1-5.7	5.8-6.4
Freq	10	12	6	8	4	2

15) The animal trainer in question above selected another group of dogs who were much older than the first group and measured their reaction times to the same stimulus. Construct a histogram, a frequency polygon, and an ogive for the data.

Class	2.3-2.9	3.0-3.6	3.7-4.3	4.4-5.0	5.1-5.7	5.8-6.4
Freq	1	3	4	16	14	4

Analyze the results and compare the histogram for this group with the one obtained in the above question. Are there any differences in the histograms?

16) The frequency distributions shown indicate the percentages of public school students in fourth-grade reading and mathematics who performed at or above the required proficiency levels for the 50 states in the United States. Draw histograms for each, and decide if there is any difference in the performance of the students in the subjects.

Class	18-22	23-27	28-32	33-37	38-42	43-48
Reading Freq	7	6	14	19	3	1
Math Freq	5	9	11	16	8	1

Using the histogram shown here, Construct a frequency distribution; include class limits, class frequencies, midpoints, and cumulative frequencies. Hence answer these questions.

a) How many values are in the class 27.5–30.5?
b) How many values fall between 24.5 and 36.5?
c) How many values are below 33.5?
d) How many values are above 30.5?