CSE - 2-1 2-2-13

Frequency Distributions

Ungrouped Frequency Distributions

A more efficient arrangement, and one that conveys more meaning, is to list the scores with their frequency of occurrence. This listing is called a frequency distribution.

A **frequency distribution** presents the score values and their frequency of occurrence. When presented in a table, the score values are listed in rank order, with the lowest score value usually at the bottom of the table.

The major purpose of a frequency distribution is to present the scores in such a way to facilitate ease of understanding and interpretation.

Table 3.1: Scores from statistics exam (N = 70)

95	57	76	93	86	89	80	72	88	84	70	83	93	76
65	79	60	56	72	82	70	82	96	87	69	89	77	81
67	79	71	77	52	76	68	87	65	77	72	56	78	78
58	54	82	66	73	79	86	81	63	46	62	99	93	82
82	92	75	76	90	74	67	76	76	63	74	94	96	77

Table 3.2: Scores from Table 3.1 organized into a frequency distribution

Score	Frequency	Score	Frequency	Score	Frequency	Score	Frequency
46	1	66	1	76	6	87	2
52	1	67	2	77	4	88	1
54	1	68	1	78	2	89	2
56	2	69	1	79	3	90	1
57	1	70	2	80	1	92	1
58	1	71	1	81	2	93	3
60	1	72	3	82	5	94	1
62	1	73	1	83	1	95	1
63	2	74	2	84	1	96	2
65	2	75	1	86	2	99	1

Grouped Distributions

Guidelines on constructing Frequency Distributions

The rules for the construction of frequency distributions are not absolute. For most data sets there is no single correct answer for the construction of a frequency distribution.

1. *The class intervals used in the frequency distribution should be equal.* Unequal class intervals present problems in graphically portraying the distribution. Unequal class intervals however may be necessary in certain situations in order to avoid a large number of empty or almost empty classes. Such case is presented in the following table. The revenue and tax department used unequal sized class intervals to report the adjusted gross income on individual tax returns. If revenue and tax department used an equal sized interval of $1000 more than 1000 classes would have been required to describe all the incomes. A frequency distribution with 1000 classes would be difficult to interpret. In this case Unequal sized class intervals would be more interpretable than equal sized class intervals.

Additional gross income	Number of returns (in thousand)
Under $ 2000	135
$ 2000 up to 3000	3399
$ 3000 up to 5000	8175
$ 5000 up to 10000	19740
$ 10000 up to 15000	15539
$ 15000 up to 25000	14944
$ 25000 up to 50000	4451
$ 50000 up to 100000	699
$ 100000 up to 500000	162
$ 500000 up to 1000000	3
$ 1000000 and over	1

2. *Determine the width of each class interval (i):*

$$i = \frac{Highest\ Value - Lowest\ Value}{Number\ of\ Classes} \quad \text{.....(i)}$$

The following formula also used to estimate the class interval based on the number of observations

$$i = \frac{Highest\ Value - Lowest\ Value}{1+3.322(\log\ of\ the\ Total\ Frequencies)} \quad \text{.....(ii) [Base 10]}$$

Our professional judgment can determine the Number of Classes. Too many classes or too few classes might not reveal the basic shape of the set of data. As a general rule, it is best to not use less than 5 nor more than 15 classes in the construction of frequency distribution.

The "2 to the k rule" also used to determine the number of classes. To estimate the number of classes we select the smallest integer (whole number) such that $2^k \geq n$ where n is the total number of observations.

3. The intervals must not overlap, otherwise there is a confusion concerning in which class an observation belongs.
4. For a continuous distribution there must be continuity from one class to the next, otherwise some observation may not fit the class.
5. Open end class might be avoided.

Relative Frequency, Cumulative Frequency, and Cumulative Percentage Distributions:

It is often desirable to express the data from a frequency distribution as a relative frequency, a cumulative frequency, or a cumulative percentage distribution.

A relative frequency distribution *indicates the proportion of the total number of scores that occur in each interval.* Relative Frequency = $\frac{f}{N}$

A cumulative frequency distribution *indicates the number of scores that fall below the upper real limit of each interval.*

A cumulative percentage distribution *indicates the percentage of scores that fall below the upper real limit of each interval.* Cumulative Percentage = $\frac{cum\ f}{N} \times 100$

Practice Problem:

118	68	55	33	72	80	35	55	62	42
102	65	104	51	100	74	45	60	58	62
92	44	122	73	65	78	49	61	65	86
83	76	95	55	50	82	51	138	73	94
83	72	89	37	63	95	109	93	65	52
75	24	60	43	130	107	72	86	71	106
128	90	48	22	67	76	57	86	114	30
33	54	64	82	47	81	28	79	85	117
98	58	32	68	77	28	69	46	53	38

Practice Problem:

1.4	2.9	3.1	3.2	2.8	3.2	3.8	1.9	2.5	4.7	1.6	2.8	2.8
1.8	3.5	2.7	2.9	3.4	1.9	3.2	2.4	1.5	1.6	3.0	2.9	2.8
2.5	3.5	1.8	2.2	4.2	2.4	4.0	1.3	3.9	2.7	3.1	3.2	3.5
2.5	3.1	3.1	4.6	3.4	2.6	4.4	1.7	4.0	3.3	2.7	0.8	3.7
1.9	0.6	1.7	5.0	1.0	1.5	2.8	3.7	4.2	4.0	2.2	3.2	2.9
2.8	1.3	3.6	2.2	3.5	3.5	3.1	3.2	3.5	2.7	3.1	3.2	3.5
3.8	2.9	3.4	0.9	0.8	1.8	2.6	3.7	1.6	4.8	1.4	2.9	1.0
3.5	1.9	2.2	2.8	3.8	3.7	1.8	1.1	2.5	1.4	3.0	2.6	4.1
3.7	3.5	4.0	1.9	3.3	2.2	4.6	2.5	2.1	3.4	4.4	2.2	4.4
1.7	4.6	3.1	2.1	4.2	4.2	1.2	4.7	4.3	3.7	3.3	3.6	2.2

Graphing Frequency Distributions

Frequency distributions are often displayed as graphs rather than tables. Since a graph is based completely on the tabled scores, the graph does not contain any new information. However, a graph presents the data pictorially, which often makes it easier to see important features of the data.

not begin at zero and is greatly expanded from that of part (a). The impressions conveyed by the two graphs are very different. Part (a) gives the correct impression of a very stable enrollment, whereas part (b) greatly distorts the data, making them seem as though there were large enrollment fluctuations.

5. Ordinarily, the intersection of the two axes is at zero for both scales. When it is not, this is indicated by breaking the relevant axis near the intersection. For example, in Figure 3.4, the horizontal axis is broken to indicate that a part of the scale has been left off.
6. Each axis should be labeled, and the title of the graph should be both short and explicit.

Four main types of graphs are used to graph frequency distributions: the *bar graph*, the *histogram*, the *frequency polygon*, and the *cumulative percentage* curve.

The Bar Graph

Frequency distributions of nominal or ordinal data are customarily plotted using a bar graph. This type of graph is shown in Figure 3.3. A bar is drawn for each category, where the height of the bar represents the frequency or number of members of that category. Since there is no numerical relationship between the categories in nominal data, the various groups can be arranged along the horizontal axis in any order. In Figure 3.3, they are arranged from left to right according to the magnitude of frequency in each category. Note that the bars for each category in a bar graph do not touch each other. This further emphasizes the lack of a quantitative relationship between the categories.

The Histogram

The histogram is used to represent frequency distributions composed of interval or ratio data. It resembles the bar graph, but with the histogram, a bar is drawn for each class interval. The class intervals are plotted on the horizontal axis such that each class bar begins and terminates at the real limits of the interval. The height of the bar corresponds to the frequency of the class interval. Since the intervals are continuous, the vertical bars must touch each other rather than be spaced apart as is done with the bar graph. Figure 3.4 shows the statistics exam scores (Table 3.4, p. 47) displayed as a histogram. Note that it is customary to plot the midpoint of each class interval on the abscissa. The grouped scores have been presented again in the figure for your convenience.

The Frequency Polygon

The frequency polygon is also used to represent interval or ratio data. The horizontal axis is identical to that of the histogram. However, for this type of graph, instead of using bars, a point is plotted over the midpoint of each interval at a height corresponding to the frequency of the interval. The points are then joined with straight lines. Finally, the line joining the points is extended to meet the horizontal axis at the midpoint of the two class intervals falling immediately beyond the end class intervals containing scores. This closing of the line with the horizontal axis forms a polygon, from which the name of this graph is taken.

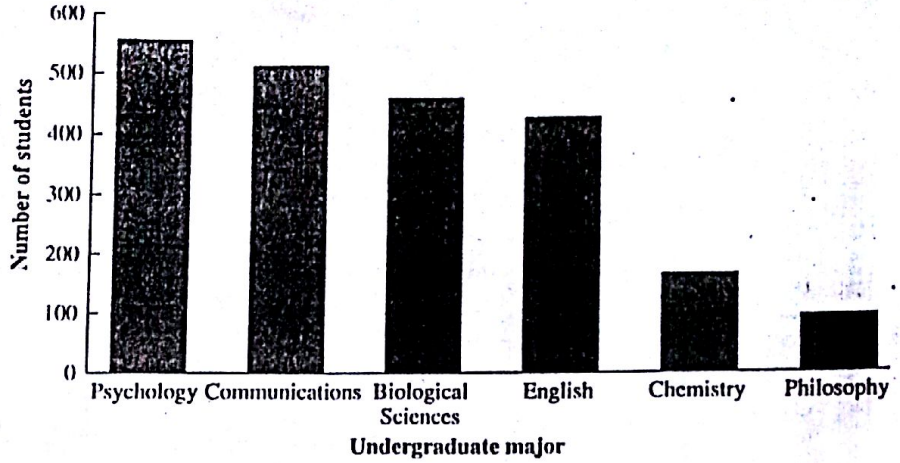


Figure 3.3 Bar graph: Students enrolled in various undergraduate majors in a college of arts and sciences.

Figure 3.5 displays the scores listed in Table 3.4 as a frequency polygon. The major difference between a histogram and a frequency polygon is the following: The histogram displays the scores as though they were equally distributed over the interval, whereas the frequency polygon displays the scores as though they were all concentrated at the midpoint of the interval. Some investigators prefer to use the frequency polygon when they are comparing the shapes of two or more distributions. The frequency polygon also has the effect of displaying the scores as though they were continuously distributed, which in many instances is actually the case.

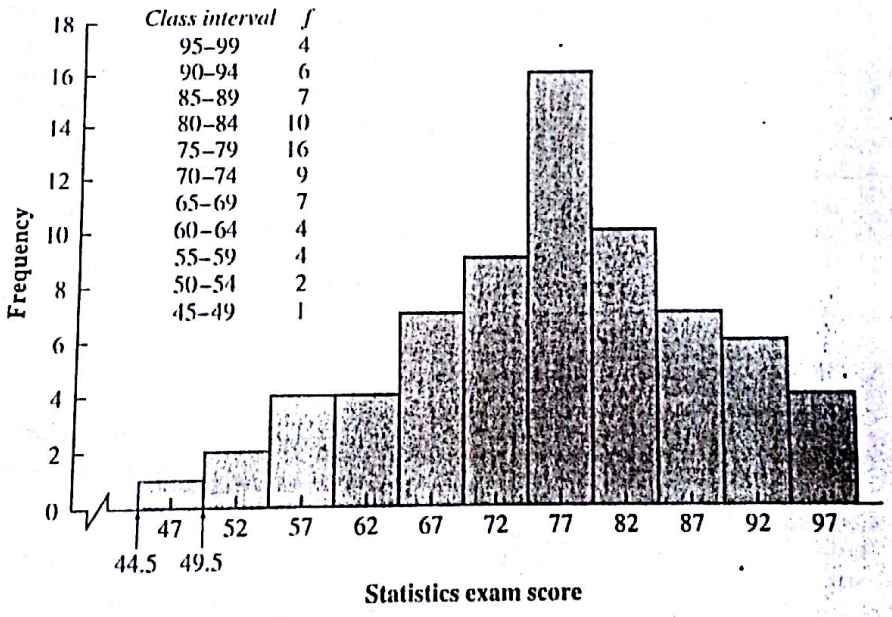


Figure 3.4 Histogram: Statistics exam scores of Table 3.4.

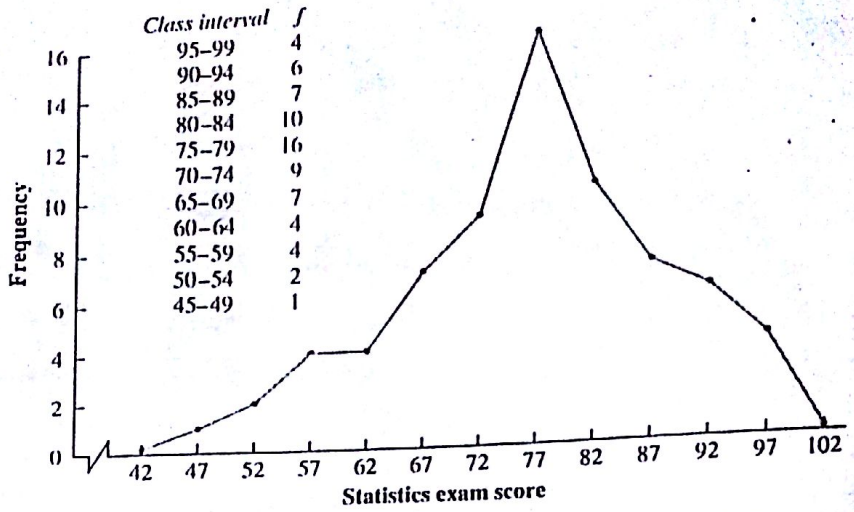


Figure 3.5 Frequency polygon: Statistics exam scores of Table 3.4.

The Cumulative Percentage Curve

Cumulative frequency and cumulative percentage distributions may also be presented in graphical form. We shall illustrate only the latter because the graphs are basically the same and cumulative percentage distributions are more often encountered. You will recall that the cumulative percentage for a class interval indicates the percentage of scores that fall below the upper real limit of the interval. Thus, the vertical axis for the cumulative percentage curve is plotted in cumulative percentage units. On the horizontal axis, instead of plotting points at the midpoint of each class interval, we plot them at the upper real limit of the interval. Figure 3.6 shows the scores of Table 3.7 (p. 50) displayed as a cumulative percentage curve. It should be obvious that the cumulative frequency curve would have the same shape, the only difference being that the vertical axis would be plotted in cumulative frequency rather than in cumulative percentage units. Both percentiles and percentile ranks can be read directly off the cumulative percentage curve. The cumulative percentage curve is also called an *ogive*, implying an *S* shape.

Shapes of Frequency Curves

Frequency distributions can take many different shapes. Some of the more commonly encountered shapes are shown in Figure 3.7. Curves are generally classified as *symmetrical* or *skewed*.

definition ■ *A curve is* **symmetrical** *if when folded in half the two sides coincide. If a curve is not symmetrical, it is* **skewed.**

The curves shown in Figure 3.7(a), (b), and (c) are symmetrical. The curves shown in parts (d), (e), and (f) are skewed. If a curve is skewed, it may be *positively* or *negatively skewed*.

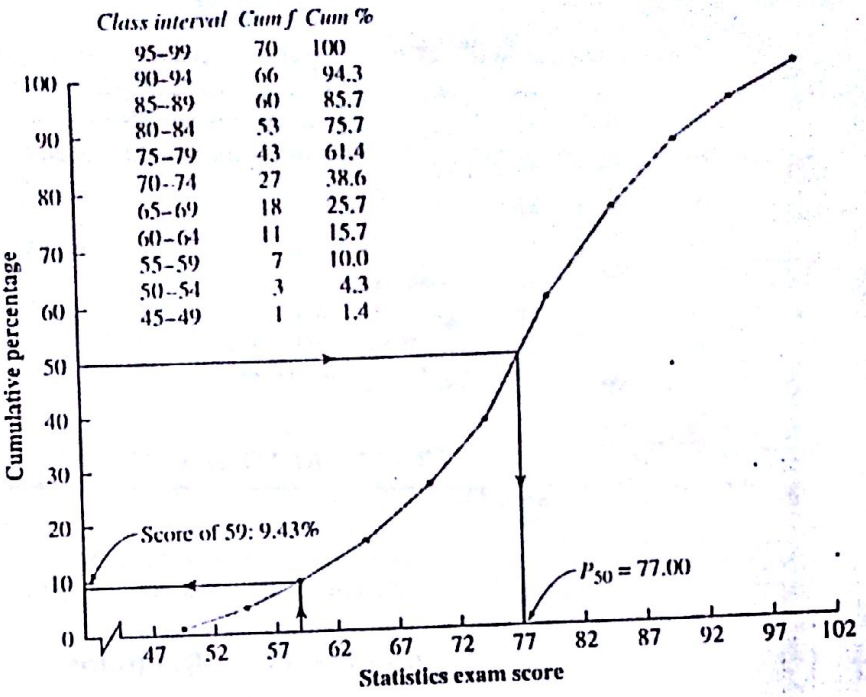


Figure 3.6 Cumulative percentage curve: Statistics exam scores of Table 3.7.

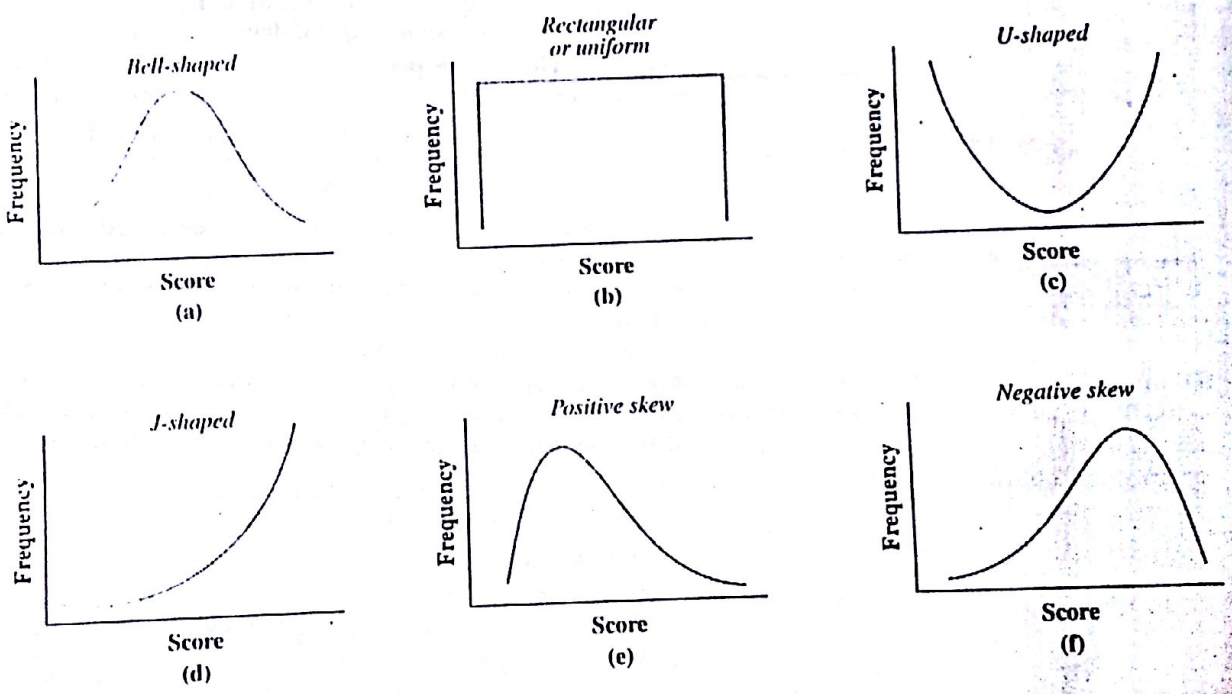


Figure 3.7 Shapes of frequency curves.

Definitions ■ *When a curve is* **positively skewed,** *most of the scores occur at the lower values of the horizontal axis and the curve tails off toward the higher end. When a curve is* **negatively skewed,** *most of the scores occur at the higher values of the horizontal axis and the curve tails off toward the lower end.*

The curve in part (e) is positively skewed, and the curve in part (f) is negatively skewed.

Frequency curves are often referred to according to their shape. Thus, the curves shown in parts (a), (b), (c), and (d) are, respectively, called *bell-shaped,* rectangular or uniform, *U-shaped,* and *J-shaped* curves.

EXPLORATORY DATA ANALYSIS

Exploratory data analysis is a recently developed procedure. It employs easy-to-construct diagrams that are quite useful in summarizing and describing sample data. One of the most popular of these is the *stem and leaf diagram.*

Stem and Leaf Diagrams

Stem and leaf diagrams were first developed in 1977 by John Tukey, working at Princeton University. They are a simple alternative to the histogram and are most useful for summarizing and describing data when the data set includes less than 100 scores. Unlike what happens with a histogram, however, a stem and leaf diagram does not lose any of the original data. A stem and leaf diagram for the statistics exam scores of Table 3.1 is shown in Figure 3.8.

In constructing a stem and leaf diagram, each score is represented by a *stem* and a *leaf.* The stem is placed to the left of the vertical line and the leaf to the right. For example, the stems and leafs for the first and last original scores are:

stem	leaf
9	5

stem	leaf
6	7

In a stem and leaf diagram, stems are placed in order vertically down the page, and the leafs are placed in order horizontally across the page. The leaf for each score is usually the last digit, and the stem is the remaining digits. Occasionally, the leaf is the last two digits depending on the range of the scores.

Note that in stem and leaf diagrams, stem values can be repeated. In Figure 3.8, the stem values are repeated twice. This has the effect of stretching the stem—that is, creating more intervals and spreading the scores out. A stem and leaf diagram for the statistics scores with stem values listed only once is shown here.

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

Original Scores

95	57	76	93	86	80	89
76	76	63	74	94	96	77
65	79	60	56	72	82	70
67	79	71	77	52	76	68
72	88	84	70	83	93	76
82	96	87	69	89	77	81
87	65	77	72	56	78	78
58	54	82	82	66	73	79
86	81	63	46	62	99	93
82	92	75	76	90	74	67

Stem and Leaf Diagram

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

Figure 3.8 Stem and leaf diagram: Statistics exam scores of Table 3.1.

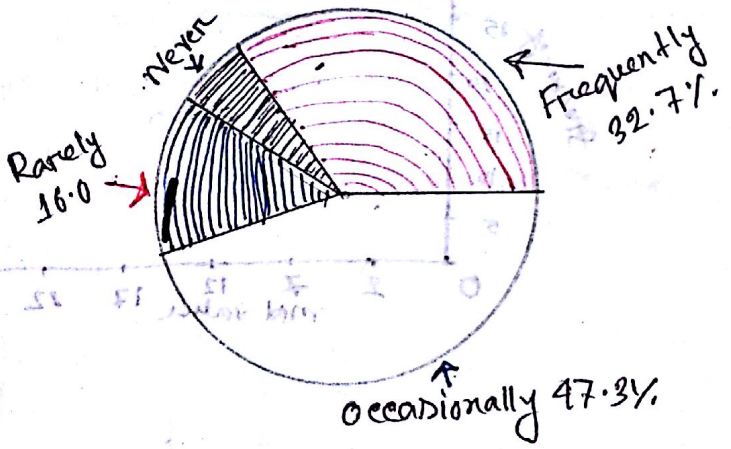
Listing stem values only once results in fewer, wider intervals, with each interval generally containing more scores. This makes the display appear more crowded. Whether stem values should be listed once, twice, or even more than twice depends on the range of the scores.

You should observe that rotating the stem and leaf diagram of Figure 3.8 counterclockwise 90°, such that the stems are at the bottom, results in a diagram very similar to the histogram shown in Figure 3.4. With the histogram, however, we have lost the original scores; with the stem and leaf diagram, the original scores are preserved.

Example: Health centre visit data for constructing pie diagram.

Response	Frequency	Percentrerelative frequeng	Angles of the rector
Frequent	49	32.7	117.6
Occasional	71	47.3	170.4
Rare	24	16.0	57.6
Never	6	4.0	14.4
Total = 150		Total = 100.0	Total = 360.0

Figure 2.5 : Simple pie diagram displaying the data in Table 2.14



3. Histogram: It is a graphical method of representationg a frequency distribution in which a frequency distribution can be shown in the form of a diagram. This diagram is known as histogram.

While constructing a histogram the horizontal axis is divided into segments corresponding to the class boundaries

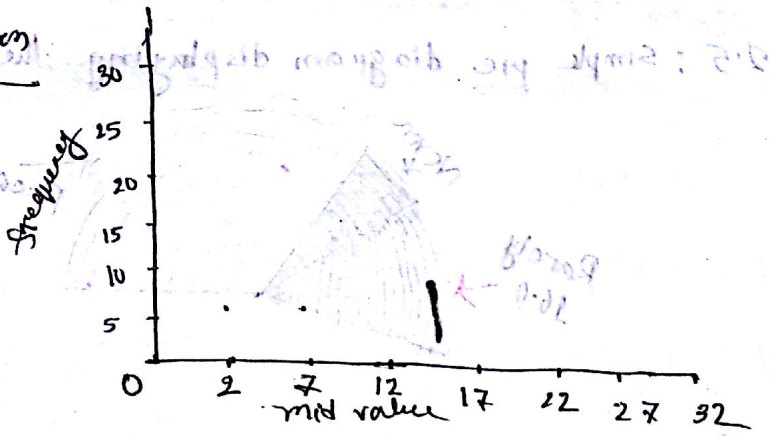
Frequency Polygons → Provides an alternative to a histograms as way of graphically presenting a distribution of a continuous variable.

A: 6, 8, 12, 14, 14, 15, 15, 16, 18, 19, 19, 23, 23, 24, 26, 26

Stem	Leaf
5	13
10	249
15	001349
20	334
25	11

Key: 5|1 means 6; 15|3 means 18

Frequency Polygon:

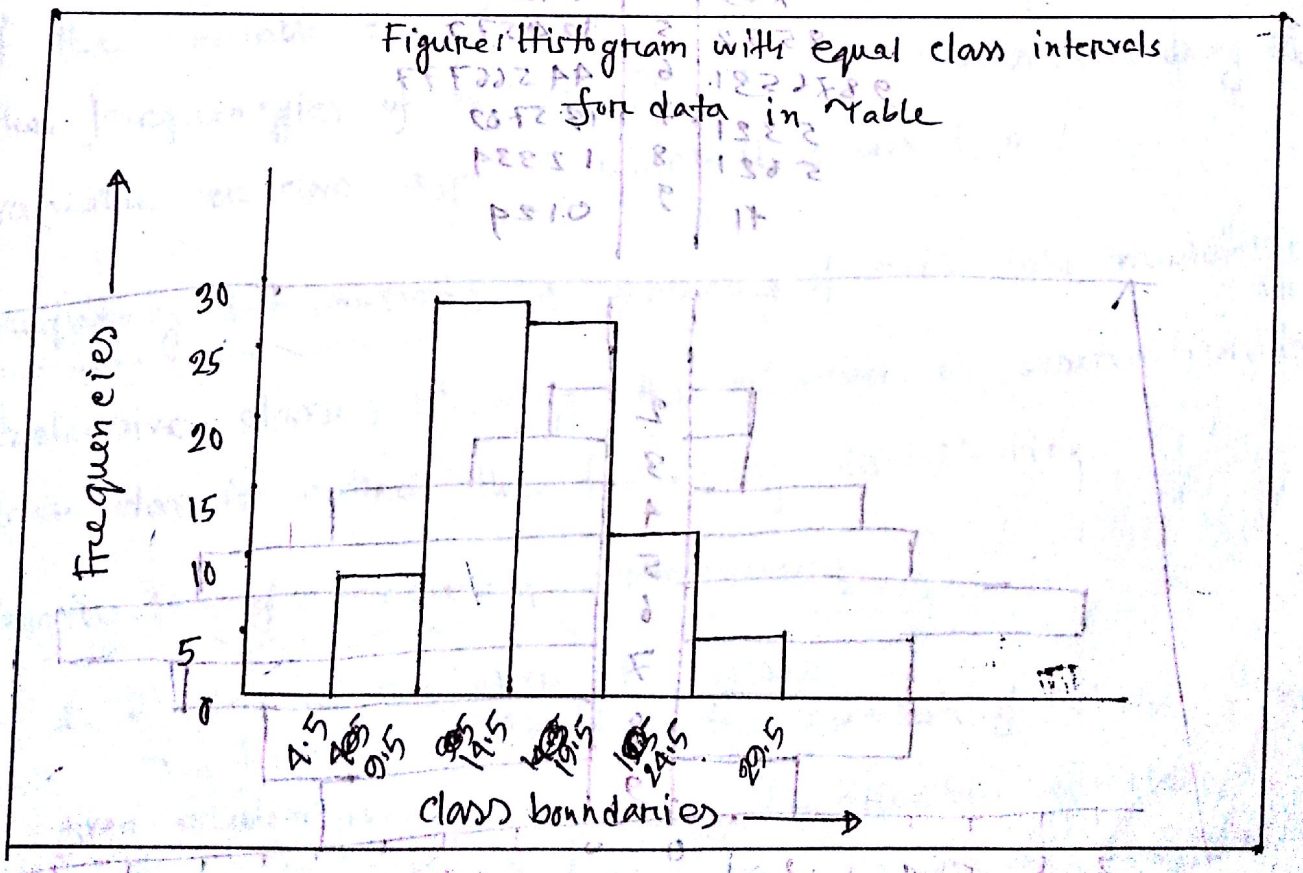


of the frequency distribution. And the vertical axis is divided into segments corresponding to frequencies.

Example: Data for constructing histogram with equal class widths.

Expenditure	class frequency	Height of the rectangle	class width
4.5 – 9.5	8	8	5
9.5 – 14.5	29	29	5
14.5 – 19.5	27	27	5
19.5 – 24.5	12	12	5
24.5 – 29.5	4	4	5
Total	80		

Figure: Histogram with equal class intervals for data in Table



Stem and leaf Ploat:

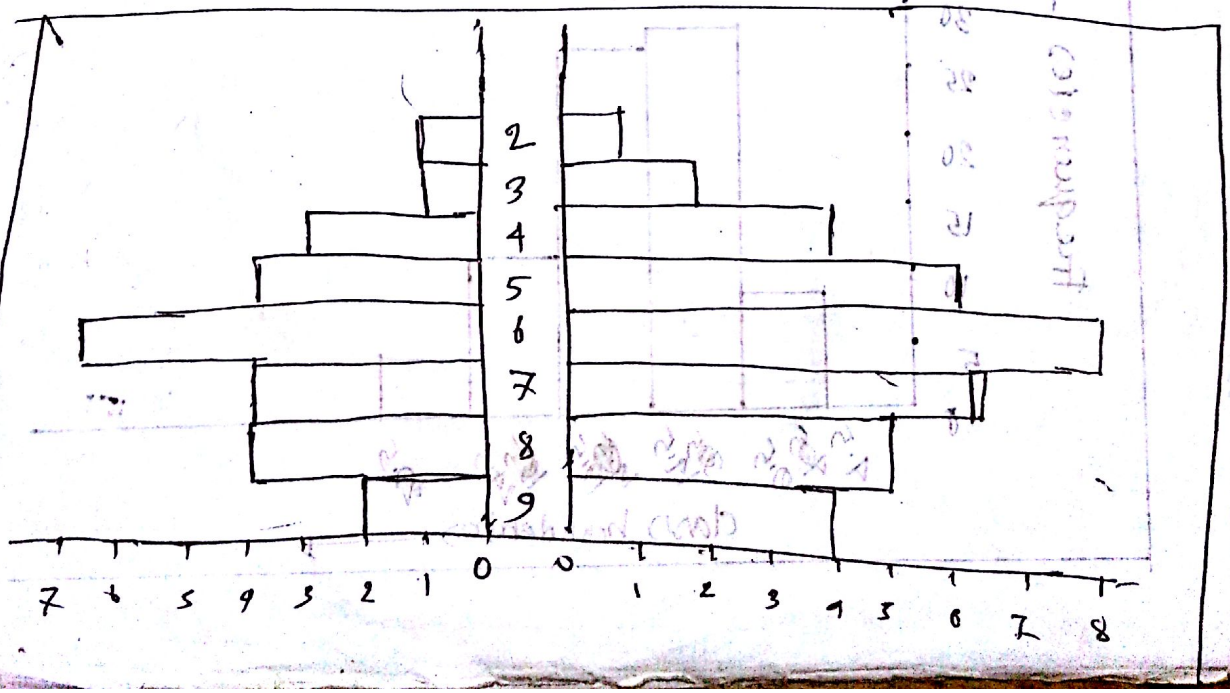
The Stem and leaf plot is a simple device to construct a histogram-like picture of a frequency distribution.

for example: A: 22, 35, 47, 55, 58, 63, 65, 66, 68, 69, 61, 71, 95, 96
52, 75, 73, 72, 81, 82, 85, 86, 94, 91, 67.

B: 90, 91, 02, 94, 21, 34, 34, 40, 41, 44, 45, 51, 52, 54, 55
57, 59, 64, 64, 65, 66, 66, 67, 67, 67, 78, 73, 75, 77, 79
71, 81, 82, 83, 83, 89

The diagram is follow:

leaf	stem	leaf
2	2	1
5	3	44
765	4	0145
8552	5	124577
98764531	6	445.66777
5321	7	13 5789
5621	8	12334
41	9	0129



Frequency: The number of observation in a particular variable or attribute is called the frequency. (আবৃত্তি বা ফ্রিকোয়েন্সি)

Class mid point or class Marks:

The central value of the class interval is called class mid-point. It lies half way between the lower limit and the upper limit of a class interval. (মধ্যবিন্দু)

$$\text{Mid point of a class} = \frac{(\text{upper} + \text{lower}) \text{ limit of a class}}{2}$$

Mean value: $\bar{x} = \sum_{i=1}^{n} \frac{f_i x_i}{\sum f_i}$ (গড়মান বা মধ্যমান)

Cumulative frequency: Cumulative frequency for any value of the variable or class is obtained by adding successively the frequencies of all the previous variables including the variable or class against which it is written.

Frequency distribution: A grouping of data into mutually exclusive classes showing the number of observations in each class is called the frequency distribution.

Construction of frequency distribution:

1. **Decide the number of classes.**

The formula to find the number of classes of a given problem is –

$2^K \geq n$ | Where K = number of class.
n = number of observation.

2. Determine the class interval:

The difference between the lower limit and upper limit of a class is known as the class interval. The formula to find the class interval of a given class is—

$$i \geq \frac{H-L}{k}$$

where
i = class interval
H = Highest observation value
L = Lowest observation value
k = number of class.

3. Set the individual class limit.
4. Tally the frequency into the classes.
5. Count the number of items in each class.

Formation of a frequency distribution:

The age of Santosh village people are given below - we will make a frequency distribution by using this following data—

10, 12, 14, 15, 19, 21, 23, 27, 31, 34, 36, 40, 42, 48.

Class	Tally	Frequency	Ralative frequency	Cumulative frequency	Percentage (%)
10-20	𝍸	5	5/14 = 0·36	5	5/14 × 100 = 35·714
20-30	III	3	3/14 = 0·214	8	3/14 × 100 = 21·428
30-40	IIII	4	4/14 = 0·285	12	4/14 × 100 = 28·571
40-50	II	2	2/14 = 0·142	14	2/14 × 100 = 14·285
		Total = 14	Sum = 1·00		Total = 100