Chapter Five: Frequency Distributions

When you completed this chapter, you will be able to:
- ✓ organise and present the collected data in the most effectively way;
- ✓ construct a Frequency Distribution;
- ✓ distinguish histogram from bar chart;
- ✓ recognise the uses of Ogive; and
- ✓ identify the difference and similarity of Histogram and Stem-and-Leaf Display.

Reference(s): Mason Chapter 2, Berenson Chapter 3
Exercise(s): Seminars 7 and 8, Mason Chapter2 Exercises 5, 7, 11, 17 and 22

In recent years, the collection of statistical data has grown at such a rate that it would be impossible to manage unless this *information is arranged in summarised form*. The whole matter of putting large masses of data into a usable form is becoming easier, because of the development of electronic computer, which made it possible to accomplish in minutes while it would have taken months or years in previous.

Using Ordered Data Array

It arranges values in ***ascending*** or ***descending*** *order* to give a more meaningful pattern.

e.g.1 Given the data array as follows:

Chlorine levels in ppm (parts per million) of 30 gallons of treated water

16.2	15.8	15.8	15.8	16.3	15.6
15.7	16.0	16.2	16.1	16.8	16.0
16.4	15.6	15.9	15.9	15.9	16.8
15.4	15.2	15.9	16.0	16.3	16.0
16.4	16.6	15.6	15.6	16.9	16.3

Re-arrange values in *ascending order* :

Chlorine levels in ppm of 30 gallons of treated water

15.2	15.6	15.9	16.0	16.2	16.4
15.4	15.7	15.9	16.0	16.3	16.6
15.6	15.8	15.9	16.0	16.3	16.8
15.6	15.8	15.9	16.1	16.3	16.8
15.6	15.8	16.0	16.2	16.4	16.9

Frequency Distributions

We may compress the raw data by using *frequency distribution, relative frequency distribution, cumulative frequency distribution, or cumulative relative frequency distribution.*

Frequency Distribution:

A frequency distribution is a tabular summary of a set of data that shows the ***frequency*** or number of data items that fall in each of several distinct classes. A frequency distribution is also known as frequency table.

Frequency Distribution Table

Chlorine levels samples of treated water with *0.3 ppm class interval.*

Class in ppm	Frequency
15.2-15.4	2
15.5-15.7	5
15.8-16.0	11
16.1-16.3	6
16.4-16.6	3
16.7-16.9	3
	30

Relative Frequency Distribution:

The frequency of each value is expressed as a ***fraction*** or a ***percentage*** of the total number of observations. The sum of all the relative frequencies equals to ***1.00*** *or* ***100%***.

Relative Frequency Distribution Table

Chlorine levels samples of treated water with *0.3 ppm class interval.*

Class in ppm	Frequency	Relative Frequency
15.2-15.4	2	0.06 (0.666666…)
15.5-15.7	5	0.17 (0.166666…)
15.8-16.0	11	0.37 (0.366666…)
16.1-16.3	6	0.20
16.4-16.6	3	0.10
16.7-16.9	3	0.10
	30	1.00

Cumulative Frequency Distribution:

It enables us to see how many observations lie ***below*** *or* ***above*** certain values, rather than merely recording the number of items within intervals.

"Less Than" Cumulative Frequency Distribution

Class in ppm	Frequency	**Class in ppm**	**Cumulative Frequency**
		less than 15.2	***0***
15.2-15.4	2	**less than 15.5**	**2**
15.5-15.7	5	**less than 15.8**	**7**
15.8-16.0	11	**less than 16.1**	**18**
16.1-16.3	6	**less than 16.4**	**24**
16.4-16.6	3	**less than 16.7**	**27**
16.7-16.9	3	**less than 17.0**	**30**
	30		

OR, "More Than" Cumulative Frequency Distribution

Class in ppm	Frequency	**Class in ppm**	**Cumulative Frequency**
		over 15.1	***30***
15.2-15.4	2	**over 15.4**	**28**
15.5-15.7	5	**over 15.7**	**23**
15.8-16.0	11	**over 16.0**	**12**
16.1-16.3	6	**over 16.3**	**6**
16.4-16.6	3	**over 16.6**	**3**
16.7-16.9	3	**over 16.9**	**0**
	30		

*(**Class boundaries** are used for the less or more than cumulative frequency distribution)*

Cumulative Relative Frequency Distribution:

It enables us to see what is the ***cumulative fractions*** *or* ***percentages*** of observations lie ***below*** or ***above*** certain values, rather than recording the percentages of items within intervals

"Less Than" Cumulative Relative Frequency Distribution

Class in ppm	Relative Frequency	**Class in ppm**	**Cumulative Relative Frequency**
		less than 15.2	***0.00***
15.2-15.4	0.06	**less than 15.5**	**0.06**
15.5-15.7	0.17	**less than 15.8**	**0.23**
15.8-16.0	0.37	**less than 16.1**	**0.60**
16.1-16.3	0.20	**less than 16.4**	**0.80**
16.4-16.6	0.10	**less than 16.7**	**0.90**
16.7-16.9	0.10	**less than 17.0**	**1.00**
	1.00		

Classification of Class

1. Quantitative Class

Quantitative classes are class that can be measured on a ***numerical scale***.

e.g.2 the above chlorine levels in ppm of 30 gallons of treated water are *quantitative class*.

2. Qualitative class

Qualitative classes are class that classifies information according to ***qualitative characteristics***.

e.g.3 sex, race, and religion do not fall naturally into numerical categories.

3. Open-ended Class

It consists of either the ***upper*** or the ***lower*** end of a quantitative classification scheme, and it is ***limitless***. This class normally attempts to cover all other possibilities not covered in other classes.

e.g.4 classes for age distribution :
under 20, 20 to under 30, 30 to under 40, 40 to under 50, *50 and over*

4. Discrete Class

Discrete Classes are separate entities that progress from one class to the next *with a* ***break***.

e.g.5 no. of children per family (classes :0-1, 2-3, 4-5, ..), we wouldn' t have 1.2 children.

5. Continuous Class

Continuous Classes progress from one class to the next ***without break***. They involved numerical measurement such as weight, height, and volume.

Continuous data can be expressed in either *fractions* or *whole numbers*.

e.g.6 classes for age distribution :
under 20, 20 to under 30, 30 to under 40, 40 to under 50, 50 and over

Stem-and-Leaf Display

A stem-and-leaf display separates data entities into "***leading*** *digits*" and "***trailing*** *digits*".

e.g.7 Daily expense of 37 HKTC students

35	65	79	119
42	68	78	120
54	67	80	125
54	60	89	136
51	69	89	135
52	69	88	149
59	71	90	160
58	72	94	
68	76	104	
65	71	109	

since the daily expense for the students are all two- or three- digit numbers, either the tens column or the hundreds column would be the leading digit and the remaining column would be the trailing digit.

i.e. 68 has a leading digit of 6 and a trailing digit of 8
136 has a leading digit of 13 and a trailing digit of 6

Stem-and-leaf display of daily expense:

3	5	
4	2	
5	441298	
6	85587099	
7	126198	
8	0998	
9	04	
10	49	
11	9	
12	05	
13	65	
14	9	
15		
16	0	N=37

Stem : column of numbers to the left of the vertical line (leading digits)

Leaves : right side of the vertical line (trailing digits)

Advantages of stem-and-leaf display

1. ***summarising*** important features of a data set

2. ***actual*** data values are presented

3. present data both in ***tabular*** and ***chart*** form

To assist us in further examining the data, we may wish to *rearrange the leaves within each of the stems by placing the digits in **ascending** order, row-by-row*.

Revised stem-and-leaf display of the 37 daily expense:

3	5
4	2
5	124489
6	05578899
7	112689
8	0899
9	04
10	49
11	9
12	05
13	56
14	9
15	
16	0

N=37

Class Exercise 1:

Present the following 50 data in *Revised Stem-and-leaf display*:

23	7	96	66	89	43	9	15	89	53
45	27	77	44	47	67	50	55	83	100
18	32	70	62	57	78	50	39	34	72
41	39	50	41	45	59	68	68	91	61
30	48	59	45	57	62	27	104	89	66

Read Mason p.33 for the Minitab output of stem-and-leaf display

Terminology associated with frequency distribution:

Class Limits: *lower* and *upper* values of the classes

e.g.8i class 10 - 19 where 10 is the lower class limit of the class
and 19 is the upper class limit of the class

Class Boundaries: lower and upper values of a class that mark ***common points*** between classes.

e.g.8ii	class 10 - 19	Class boundaries	9.5 - 19.5
	class 20 - 29	Class boundaries	19.5 - 29.5

Note : sometimes class limits and boundaries will ***coincide***, as in the structure : "30 and up to 40", "40 and up to 50", and etc.
The "30", "40" and "50" are both the class limits and class boundaries.

Class Width/Interval: *Upper class boundaries - Lower class boundaries*

e.g.8iii	Class	Class boundaries	Class interval
	10 to 19	9.5 to 19.5	19.5 - 9.5 = 10

Class Mid-points: ***midway*** *between upper and lower class boundaries.*

e.g.8iv	Class	Class boundaries	Class Mid-point
	10 to 19	9.5 to 19.5	(9.5+19.5)/2 = 14.5

Constructing a Frequency Distribution

To construct a frequency distribution, we must determine the followings:

1. the number of classes,
2. the class interval or width of the class,
3. the class boundaries or the values that form the interval for each one of the class,
4. the frequency for each class.

1. Determine the number of classes

The number of classes depends on the number of ***data points*** and the range of the data collected.

However, it is often better to use a rule for selecting the number of classes that is designed to summarise the data effectively as follows:

Approximate number of classes = 1 + 3.322 log(number of data)

e.g.9i Suppose we have 105 data in the data set,

the number of classes = 1 + 3.322 log(**105**) = **7.7144** ≅ **8**

*As a rule, statisticians rarely use fewer than **five** or more than **twenty** classes.*

2. Determine the class interval

$$\textbf{Approximate class width/interval} = \frac{\text{Largest data value - Smallest data value}}{\text{Number of classes}}$$

e.g.9ii suppose the largest and the smallest value of the 105 data are 19 and 4, then

$$\text{the approximate class width/interval} = \frac{19-4}{8} = 1.875$$

To use a class interval of 1.875 is very *inconvenient*, we restrict the class interval to one decimal place, i.e., we will use an interval of 1.9, or restrict the class interval to an integer, i.e., use an interval of 2.

You can only ***round up*** the approximate class interval.

3. Determine the class limits for each class

Class limits are selected so that the *classes cover* ***all the data values*** and so that each value *falls in a* ***distinct*** *class*.

Make sure that the first class covers the smallest data value, and the last class covers the largest data value.

The lower boundary of the first class is ***arbitrary***, *but* ***convenient*** value below the lowest value. It is suggested, but not necessary, that the lower limit of the first class should be an ***even multiple*** of the class interval.

e.g. 9iii suppose that the class interval of 2 is used, then the lower limit of the first class would be (2x2=) 4, (2x4=)8, or (2x6=)12, etc. The 4, 8, and 12 are the *even multiple* of the class interval "2".

For example, if 4 is used as the lower limit of the first class, then the class limits are:

4 - 5	6 - 7	8 - 9	10 - 11
12 - 13	14 - 15	16 - 17	18 - 19

And the respective class boundaries are:

3.5 - 5.5	5.5 - 7.5	7.5 - 9.5	9.5 - 11.5
11.5 - 13.5	13.5 - 15.5	15.5 - 17.5	17.5 - 19.5

4. Determine the frequency for each class

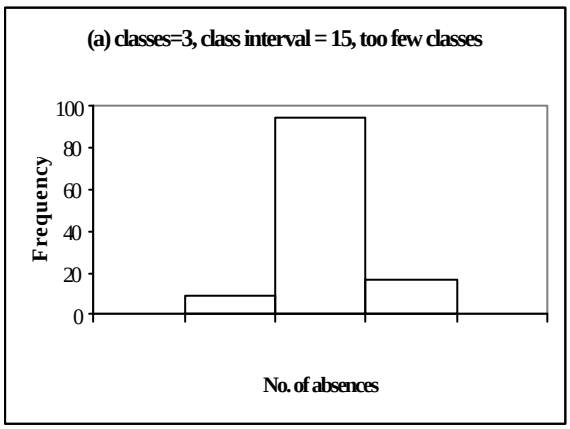
Every data point fits into ***one and only one*** class interval. After all data are fitted into the class, the frequency of each class is counted.

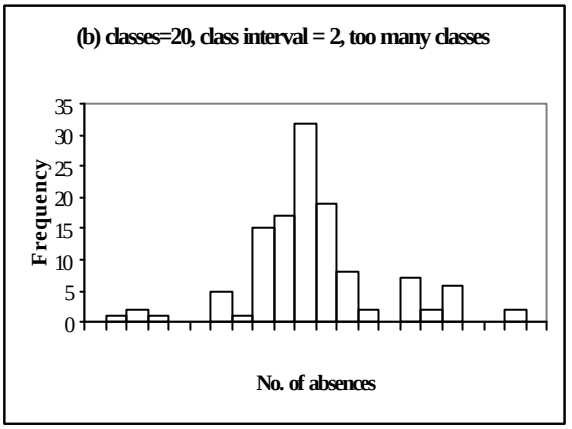
Remark: When raw data are grouped into classes, a *certain amount of information is lost*, since no distinction is made between observations falling in the same class.

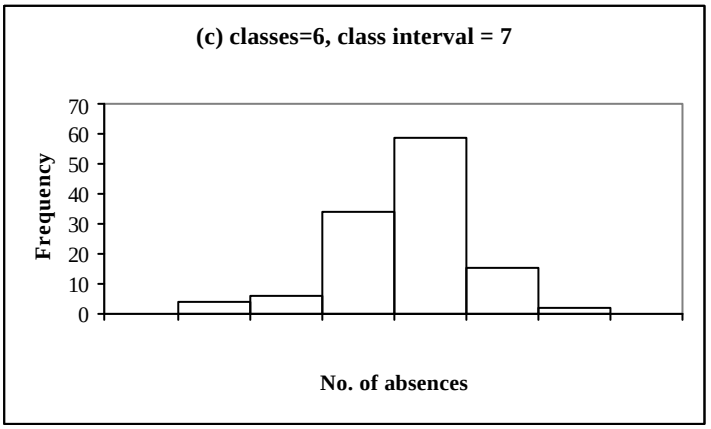
The larger the class interval is, the greater is the amount of information lost. The smaller the class interval is, the little is the amount of information lost, but the presentation of information is somewhat misleading because the small irregularities in the histogram merely reflect the accidents of sampling.

e.g.10

Class Interval = 2		*Class Interval = 7*		*Class Interval = 15*	
No. of Absences	No. of Observations	No. of Absences	No. of Observations	No. of Absences	No. of Observations
120-121	1	119-125	4	117-131	9
122-123	2	126-132	6	132-146	94
124-125	1	133-139	34	147-161	17
126-127	0	140-146	59		
128-129	0	147-153	15		
130-131	5	154-160	2		
132-133	1				
134-135	15				
136-137	17				
138-139	32				
140-141	19				
142-143	8				
144-145	2				
146-147	0				
148-149	7				
150-151	2				
152-153	6				
154-155	0				
156-157	0				
158-159	2				
Total	**120**	Total	**120**	Total	**120**







Presentation of the frequency distribution can be in tabular form, or in graphical form that is called *Histogram*.

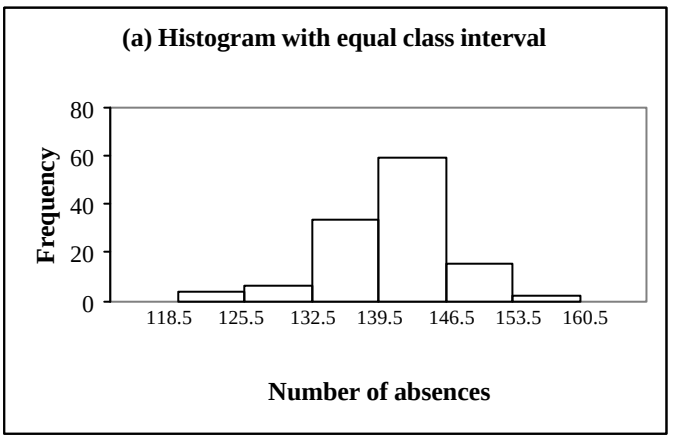
Histogram

A graphic presentation of a frequency distribution and is constructed by ***erecting*** bars on the class intervals.

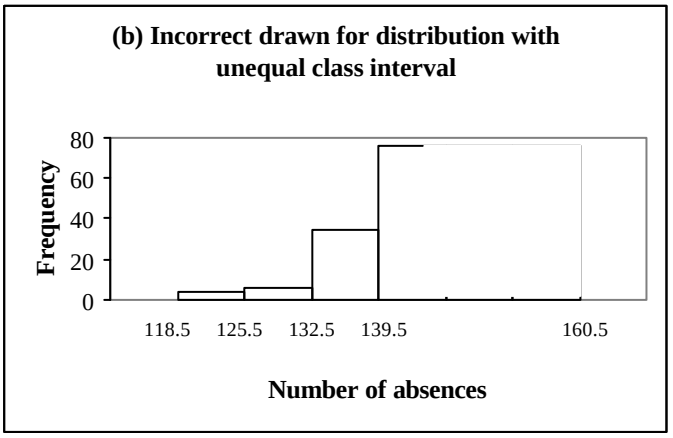
If we have equal class intervals, we erect over each class a rectangle whose ***height*** is proportional to the frequency of that class.

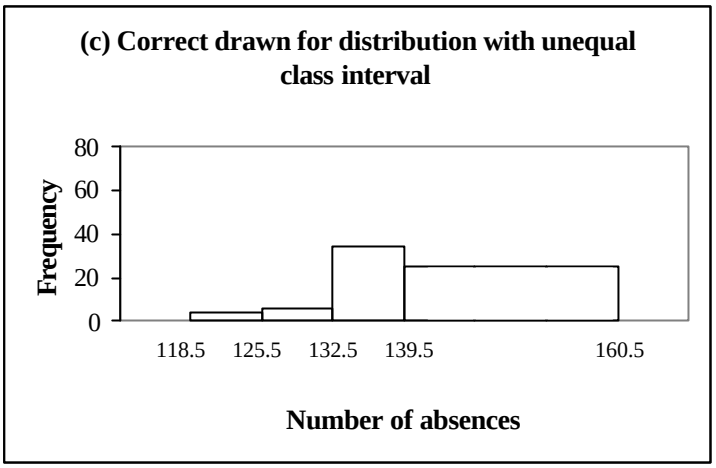
e.g.11

No. of Absences	No. of Observations	No. of Absences	No. of Observations
119-125	4	119-125	4
126-132	6	126-132	6
133-139	34	133-139	34
140-146	59	***140-160***	***76***
147-153	15		
154-160	2		
Total	120	Total	120



If we have unequal class interval, the ***area of the bar*** over a class interval must be proportional to the frequency of the class.





Class Exercise 2:

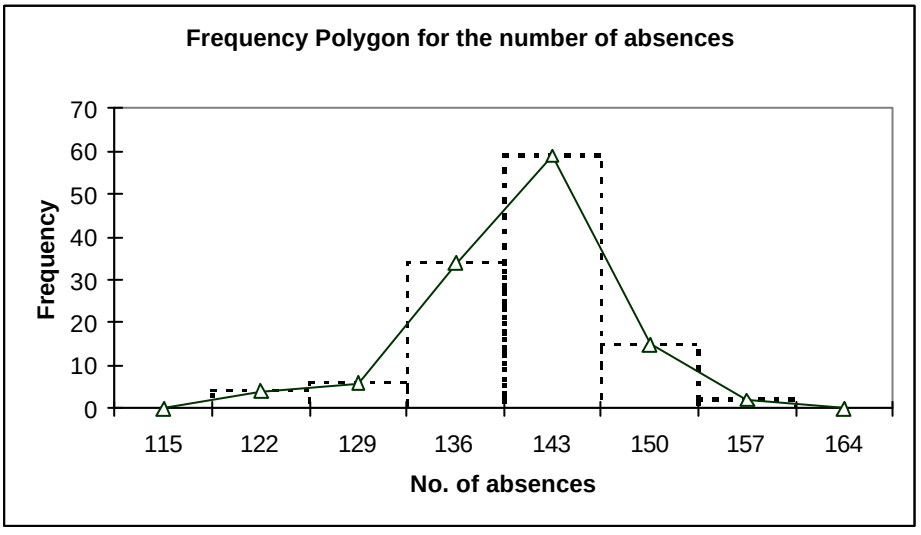
Present the data set, in Class Exercise 1, in *Histogram with equal class interval*

Frequency Polygon

It is constructed by plotting the class frequencies versus the ***class mid-points*** and then connecting the points with straight lines.

The polygon should *touch the horizontal axis at both ends* of the distribution by *adding a class (with zero frequency) to each end* and including their mid-points in the point connection process.

e.g.12



Frequency polygon allows ***comparing*** directly two or more frequency distributions.

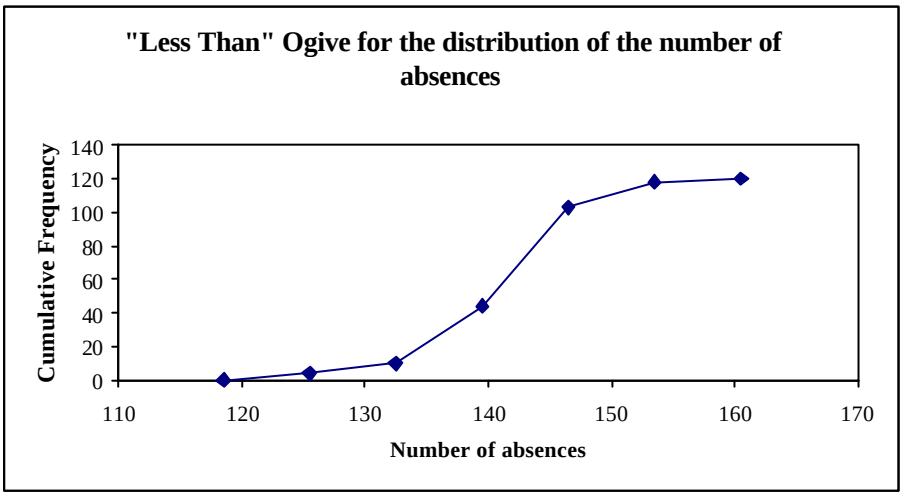
Sometimes, we require information on the *number of observations whose numerical value is less than a given value*, this information is contained in the cumulative frequency distribution, and a cumulative frequency distribution is represented graphically by *Ogives.*

1. A "Less Than" Ogive

We can use a table recording the cumulative "less than" frequencies for the above example (e.g.11).

No. of Absences	No. of Observations
less than 118.5	0
less than 125.5	4
less than 132.5	10
less than 139.5	44
less than 146.5	103
less than 153.5	118
less than 160.5	120

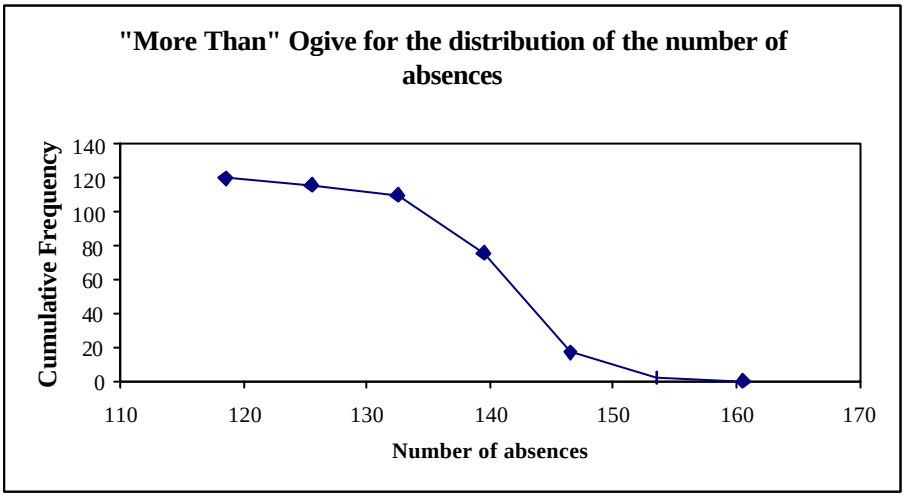
The accumulative frequencies are plotted against <u>Lower class boundaries</u>.



2. A "More Than" Ogive

We can use a table recording the cumulative "more than" frequencies for the above example (e.g.11).

No. of Absences	No. of Observations
more than 118.5	120
more than 125.5	116
more than 132.5	110
more than 139.5	76
more than 146.5	17
more than 153.5	2
more than 160.5	0



Class Exercise 3:

Construct a "Less Than" Ogive for the data set in Class Exercise 1. Hence, find the number of data that are greater than 50.

Class : _____ **Name** : _____ **No.** : _____

Class Exercise 1 (Solution):

Stem-and-leaf display of the 50 data:

0	79
1	85
2	377
3	02994
4	518415753
5	097790053
6	62728816
7	7082
8	9939
9	61
10	40

N = 50

Revised Stem-and-leaf display of the 50 data:

0	79
1	58
2	377
3	02499
4	113455578
5	000357799
6	12266788
7	0278
8	3999
9	16
10	04

N = 50

Class : _____ **Name** : _____ **No.** : _____

Class Exercise 2 (solution):

Step 1 : *Determine the number of class*

Approximate number of class = 1 + 3.322 log(50) = 6.643978 ≅ 7

Step 2 : *Determine the class interval*

Approximate class interval = (104 - 7) /7 = 13.86 ≅ 14

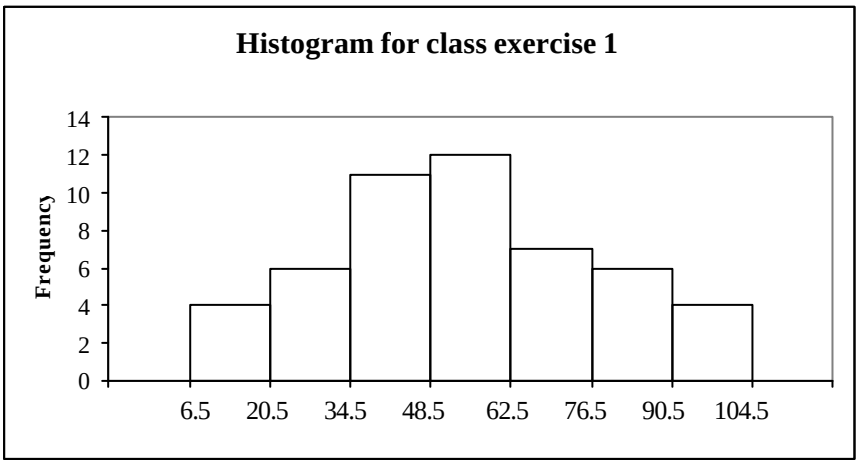
Step 3 : *Determine the class boundaries*

Since the first class must cover the smallest data, 7, the arbitrary and convenient value, 6.5, is chosen as the lower boundary of the first class.

so the boundaries of the classes are :
6.5 to 20.5, 20.5 to 34.5, 34.5 to 48.5, 48.5 to 62.5, 62.5 to 76.5, 76.5 to 90.5, and 90.5 to 104.5

Step 4 : Determine the frequency for each class

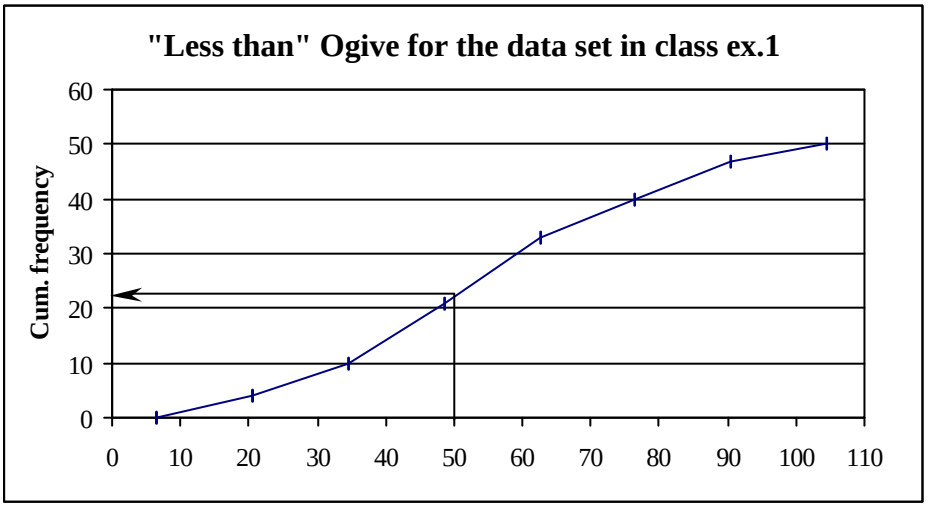
Class	Frequency
6.5 to 20.5	4
20.5 to 34.5	6
34.5 to 48.5	11
48.5 to 62.5	12
62.5 to 76.5	7
76.5 to 90.5	6
90.5 to 104.5	4
Total	50



Class : _____ **Name :** _____ **No. :** _____

Class Exercise 3 (Solution):

Class	Frequency
less than 6.5	0
less than 20.5	4
less than 34.5	10
less than 48.5	21
less than 62.5	33
less than 76.5	40
less than 90.5	46
less than 104.5	50



About 22 data are smaller than 50, so about (50 - 22=) 28 data are greater than 50.