STA 111: Descriptive Statistics

LECTURE 4

Topic: Charts for Quantitative Data

By: E. E. E. Akarawak

SCHOOL OF
SCIENCE AND
TECHNOLOGY
PAN-ATLANTIC UNIVERSITY

Lecture Outline

- Cross tabulation for Categorical data
- Multiple and component Bar Charts
- Presentation of Quantitative data
  - Histogram
  - Ogive
  - Stem and Leaf Display
  - Box Plots
- Summary Measures of Location
  - Mean
  - median
  - Mode

Expected Learning Outcome

<u>**Learning Outcomes:**</u>

- At the end of this week 5 lecture, the students should be able to:

i. Represent bivariate qualitative data using Cross tabulation and bar charts

ii. Represent quantitative data using histogram, ogive, box plot, stem and leaf

iii. Interpret various charts in real life situations.

iv. Compute summary measures of location

Charts for Quantitative Data

Charts and plots for presenting quantitative data to be considered include:

•Line Graph

•Histogram

•Frequency Polygon

•Cumulative frequency curve

•Stem and leave

•Box and Whiskers plot

We will consider them one by one with examples

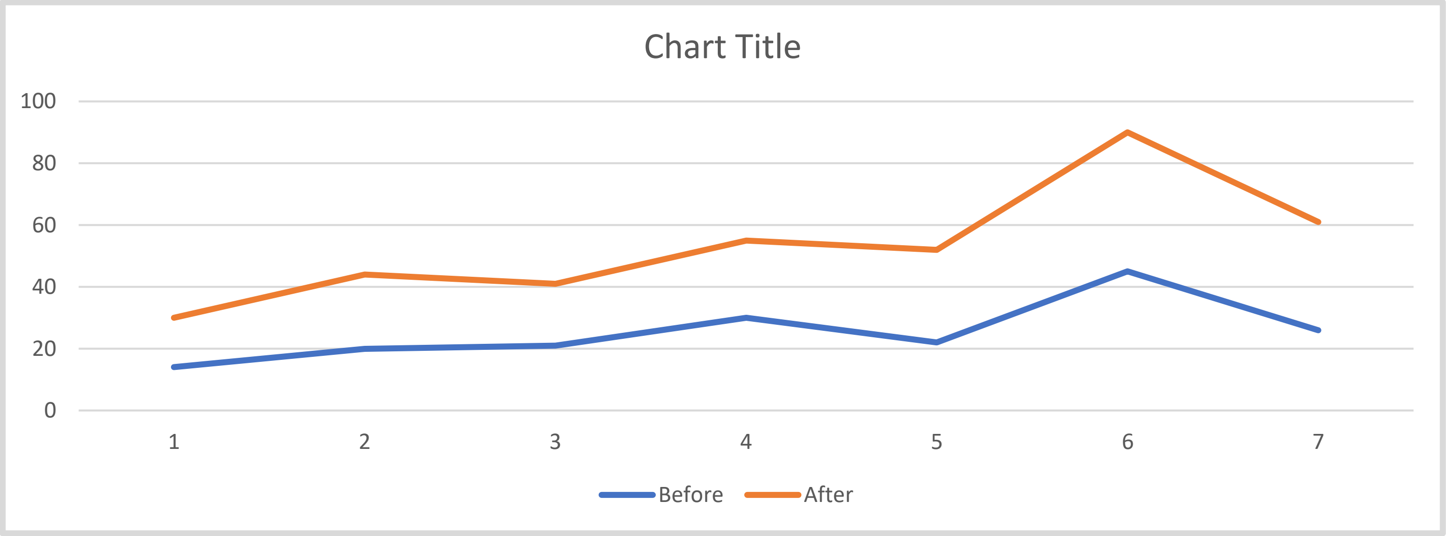
Line Graph

- Line graphs can be used to represent the relationship between two variables X and Y. The co-ordinate points of these variables are joined together to have the line graph.
- Line graph is especially useful for the study of some variables according to the passage of time. This kind of line graph is referred as Time Plot. The time, in days, weeks, months or years is marked along the horizontal axis; and the value of the quantity is marked on the vertical axis. Example could be the concentration of acetone in an hourly test sample of output product
- The line graph is suitable for predicting trend of a series over a long period.

Line Graph: Example 1

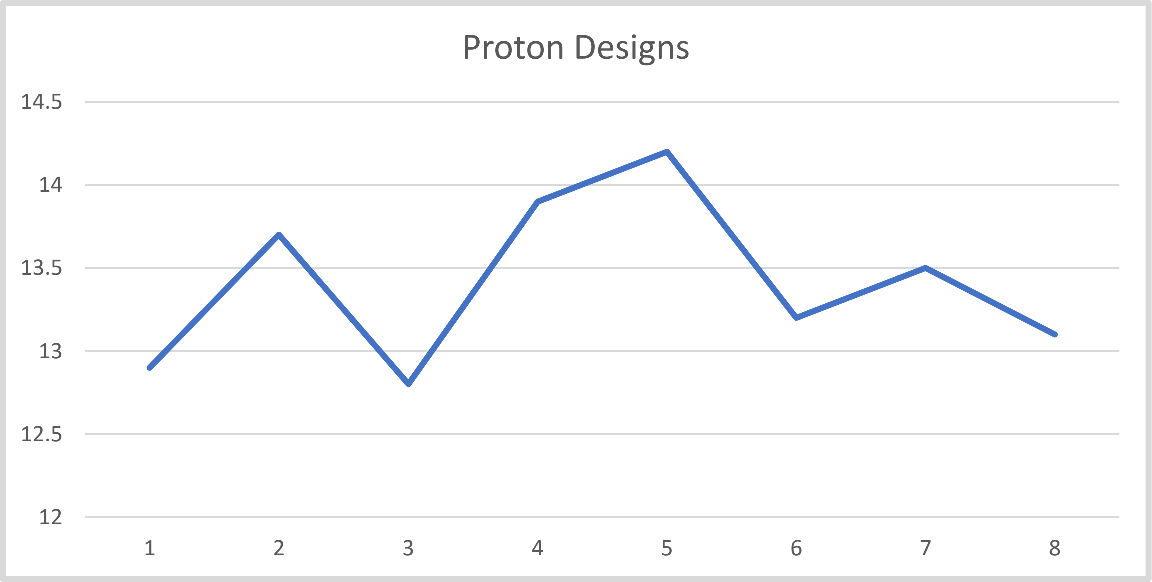
- Draw a line graph to represent the information below on concentration of chemical impurities in water before and after treatment.

Before	14	20	21	30	22	45	26
After	16	24	20	25	30	45	35



Line Graph: Example 2

- The observed pull-off force measurements at equal time interval from 8 prototype designs are 12.9, 13.7, 12.8, 13.9, 14.2, 13.2, 13.5, 13.1.
- Draw a Line graph (Time plot) for the data.



Histogram

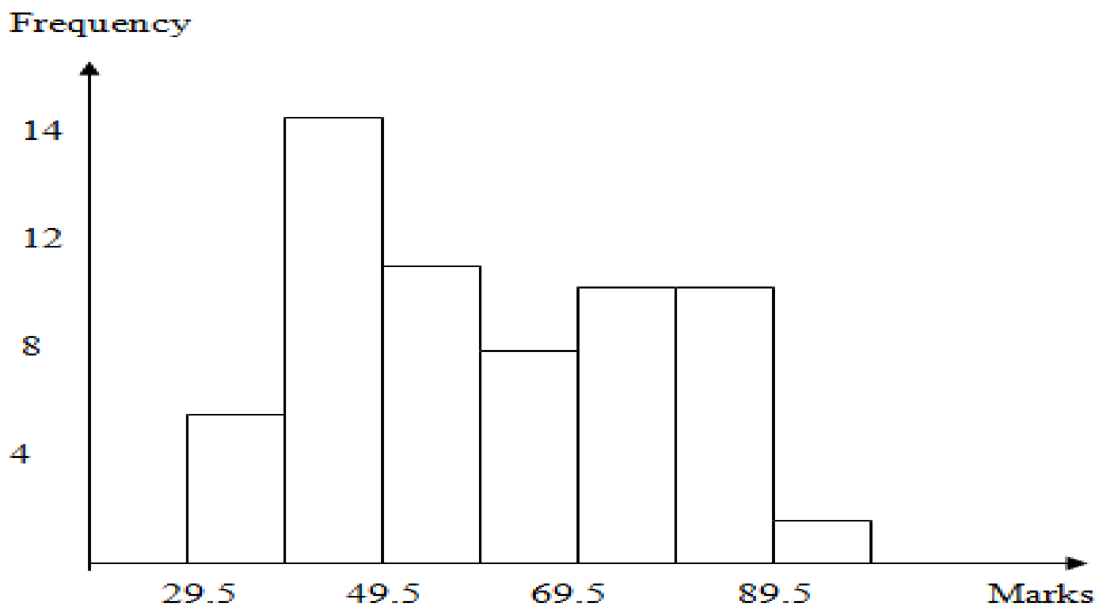
- A histogram is the graph of the frequency distribution of quantitative continuous data against the class boundaries. It is constructed on the basis of the following principles:
- a) The horizontal axis is a continuous scale running from one extreme end of the distribution to the other. It should be labelled with the name of the variable and the units of measurement.
- b) For each class in the distribution a vertical rectangle is drawn with
  - (i) its base on the horizontal axis extending from one class boundary of the class to the other class boundary, there will never be any gap between the histogram rectangles.
  - (ii) the bases of all rectangles will be determined by the width of the class intervals. If a distribution with unequal class-interval is to be presented by means of a histogram, it is necessary to make adjustment for varying magnitudes of the class intervals.
- Values for the class boundaries or class limits may be labeled along the x-axis.

Histogram: Example

Using the data below, plot a histogram

Class	Class limits	Frequency	Class boundaries	Cum Freq
1	30 – 39	5	29.5 – 39.5	5
2	40 – 49	13	39.5 – 49.5	18
3	50 – 59	9	49.5 – 59.5	27
4	60 – 69	6	59.5 – 69.5	33
5	70 – 79	8	69.5 – 79.5	41
6	80 – 89	8	79.5 – 89.5	49
7	90 – 99	1	89.5 – 99.5	50

Histogram: Solution



- Histogram can be used to estimate the mode and shape of the distribution (Check class notes)

Frequency Polygon

- If we join the midpoints of the tops of the adjacent rectangles of a histogram with line segments a frequency polygon is obtained.
- Note that it is not essential to draw histogram in order to obtain frequency polygon.
- Frequency polygon is actually the plot of the frequencies against the class midpoints.
- Let the plot touch the x axis

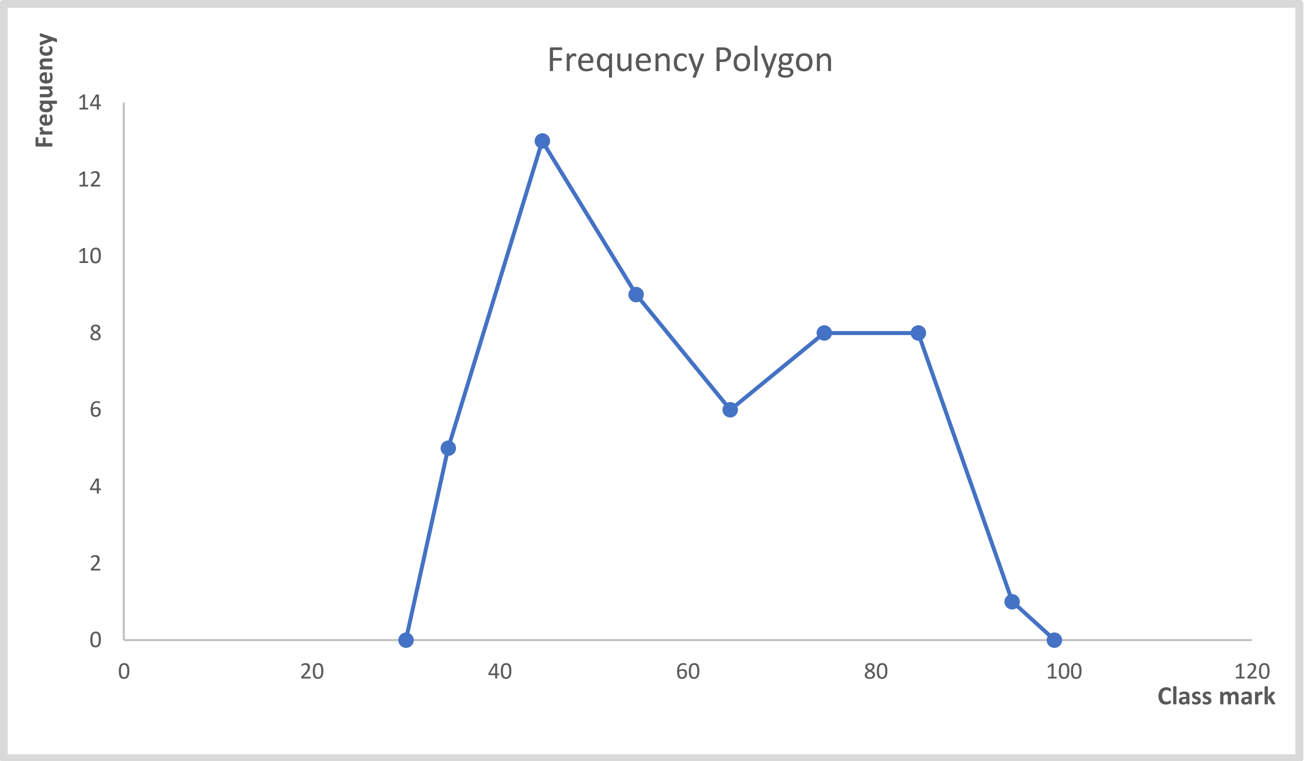
Example: Frequency Polygon

- Draw a frequency polygon for the data in the table.

Class	Class limits	Frequency	Class boundaries	Class mark
1	30 – 39	5	29.5 – 39.5	34.5
2	40 – 49	13	39.5 – 49.5	44.5
3	50 – 59	9	49.5 – 59.5	54.5
4	60 – 69	6	59.5 – 69.5	64.5
5	70 – 79	8	69.5 – 79.5	74.5
6	80 – 89	8	79.5 – 89.5	84.5
7	90 – 99	1	89.5 – 99.5	94.5

Frequency polygon: Solution

Class Mark	Freq
30	0
34.5	5
44.5	13
54.5	9
64.5	6
74.5	8
84.5	8
94.5	1
99	0



Cumulative Frequency curve (Ogive)

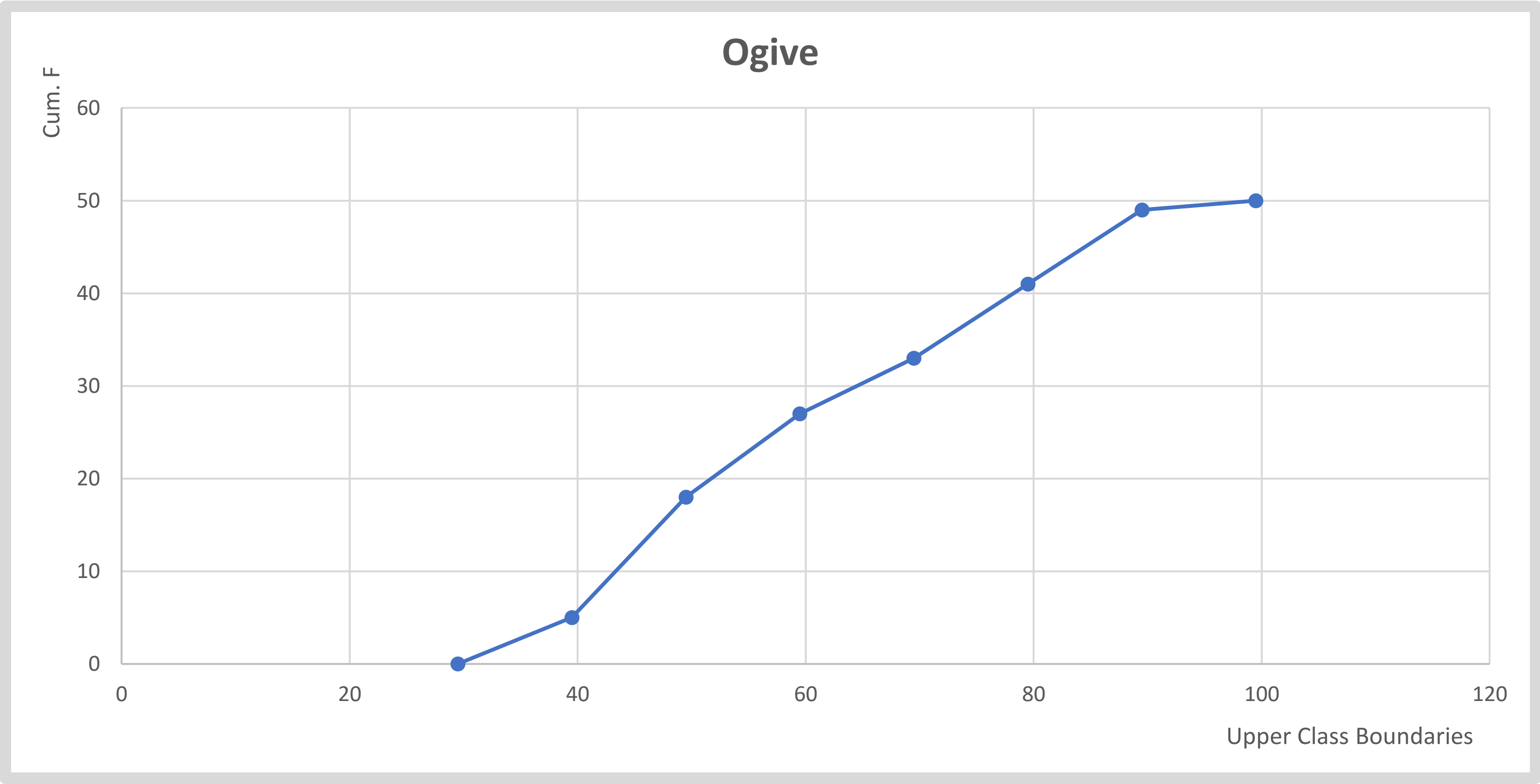
- A cumulative frequency curve is a graph of the cumulative frequency against the class boundaries. The curve is also called Ogive.
- Until the upper class boundary of a class has been reached, you cannot be sure you have accumulated all the data in that class. Therefore, the horizontal scale for an Ogive presents the upper class boundaries.
- The cumulative frequencies can sometimes be replaced by the cumulative relative frequencies or their percentages.
- Quartiles and percentiles can be obtained from an Ogive

Ogive: Example

- Plot an Ogive from the data in the table below

Class	Class limits	Frequency	Class boundaries	Cum Freq
1	30 – 39	5	29.5 – 39.5	5
2	40 – 49	13	39.5 – 49.5	18
3	50 – 59	9	49.5 – 59.5	27
4	60 – 69	6	59.5 – 69.5	33
5	70 – 79	8	69.5 – 79.5	41
6	80 – 89	8	79.5 – 89.5	49
7	90 – 99	1	89.5 – 99.5	50

Cumulative Frequency curve (Ogive)



Interpreting Ogive

- **Example:** If the data in the Ogive are exam marks,

i) Give the estimates of the quartiles

ii) Find the median

iii) Estimate the 30th and 70th percentiles

iv) Obtain the Interquartile range and semi interquartile range

v) What number of students scored marks between 60% and 80%?

(vi) What will be the pass mark if 60% of the students failed?

Solution

All explanations given in class:

i) Quartiles:

Q_1 = 25^{th} percentile = 46.5

Q_2 = 50^{th} percentile = 59.5

Q_3 = 75^{th} percentile = 72

ii) Median is the 50^{th} percentile and it is equal to 59.5

iii) 30^{th} percentile = 49.5

70^{th} percentile = 68

Ogive Solution *(contd)*

iv) Interquartile range $= Q_3 - Q_1$

$= 72 - 46.5 = 25.5$

Semi-Interquartile range $= (Q_3 - Q_1)/2$

$= (72 - 46.5)/2$

$= 25.5/2 = 12.75$

Ogive: Solution

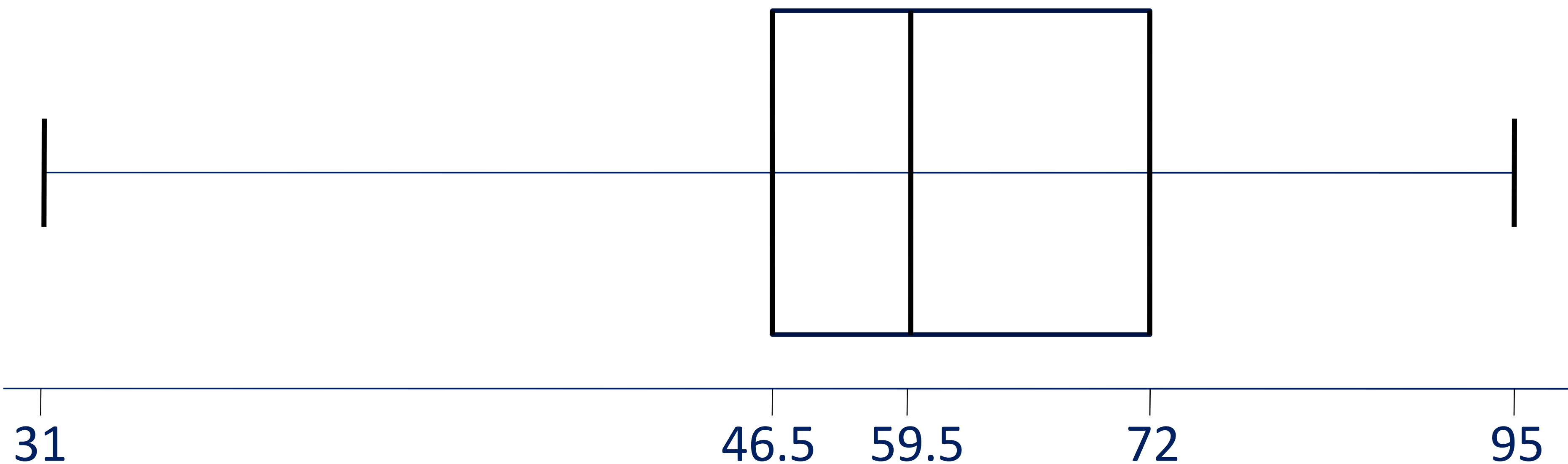
- (vi) At 60% mark this intercept the curve at cumulative frequency of 25 students and at 80% mark this intercept the curve at cumulative frequency of 43. Therefore, the number of students that scored between 60% and 80% mark are 43 – 25 = 18 students
- (vii) If 60% of the students failed, the pass mark will be from the 60^{th} percentile mark ($.6 \times 50 = 30$). Trace this to the curve and the pass mark will be 67.

Box-and-Whisker Diagram

- Box plots (in short) are extremely useful graphical devices for describing quantitative data.
- This plot is based on the five number summary of a set of data: Minimum, Q1, Q2 (median), Q3 and the Maximum.
- The three values used – Q_1, Q_2 and Q_3 – are sometimes called hinges and is also useful for identifying outliers. The ends of the box are the lower sample quartile (Q1) and upper sample quartile (Q3) and the length of the box is the IQR for the variable.
- The median (Q2) is marked by a line inside the box.
- The lines extending from the box up to the minimum and maximum are within the interval (Q_1-1.5IQR, Q_3+1.5IQR). These lines are the whiskers
- Points that are within the interval (Q_1-3IQR, Q_1-1.5IQR) are negative outliers and points that are within the interval (Q_3+1.5IQR, Q_3+3IQR) are positive outliers while those outside the interval (Q_1-3IQR, Q_3+3IQR) are extreme outliers.

Example: Box Plot

- Using the information from the Ogive example, we have the five number summary:
- Min. = 31; Q_1 = 46.5; Q_2 = 59.5; Q_3 = 72; Max. = 95



Interpretation

- Box plot can be used to infer the shape of the distribution and presence or absence of outlier.
- For instance, if the mean (M) of the data is found, it can be placed on the line. If it is to the left of the median, the distribution is skewed to the left, if on right of the median the distribution is right skewed and if equal to the median the distribution is symmetric.

Stem and Leaf Display (S&L)

- It is graphical display of data used to show how data are distributed and where concentrations exist. It is a tool to visualize data
- A stem-and-leaf display organizes data into groups (called stems) so that the values within each group (the leaves) branch out to the right on each row.
- **Method:** Separate the sorted data series into leading digits (the **stems) and** the trailing digits (the **leaves)**

Rules for Drawing Stem and Leaf

Rules for Stem:

i. There must be 6-13 stem
ii. Stems are consecutive (or repeated consecutive) numbers
iii. Stems numbers can be repeated 1, 2, or 5 times (or less at the ends) for all stems with No missing stem. That is, number of repeats should be the same for all stems, except possibly at the ends.
iv. Units for the stem must be indicated, to depict the place velue.
v. The first and last stem must have at least 1 leaf.

Rules for Drawing Stem and Leaf

Rules for Leaf:

i. Leaf is next digit after stem – you truncate (ignore) the rest of the digits.

ii. Write in ascending order, evenly spaced with no commas.

Illustration

- Draw stem and leave display for the data: 16 17 17 18 18 18 19 19 20 20 21 22 22 25 27 32 38 42

Stem (10)	**Leaf**
1	67788899
2	0012257
3	28
4	2

- The S&L above is not correct – rule 1: number of stems is less than 6
- So, we can repeat stems (twice) to obtain the following.

Illustration (contd)

STEMS (10)	LEAVES
1	67788899
2	00122
2	57
3	2
3	8
4	2

Notes on the Illustration

- Each stem is repeated twice
- The ends are trimmed off, top stem with 1 and bottom stem with 4, because they do not have leaf.
- Other Rules:
- When stems in the middle do not have leaf, they remain.
- When stems at the ends do not have leaf, they are trimmed off (as in above illustration).
- When repeating the stem twice, the first has leaf of digits 0 – 4 and the second stem has leaf digit 5 – 9.
- When repeating the stem 5 times, the leaf follow the order: 0-1, 2-3, 4-5, 6-7, 8-9

Example

- Draw stem and leaf display for the waitress's tips data:

1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.05
3.15	3.25	3.35	3.35	3.35	3.35	4.00	4.00	4.05	4.15
4.25	4.25	4.30	4.30	4.33	4.38	4.50	4.50	4.50	4.50
4.50	4.50	4.50	4.50	4.95	4.98	5.00	5.55	8.56	8.99

Solution

- Key: Stem × 1, leaf × 0.1

STEM	| LEAF
1	0000000000
2	
3	123333
4	0001223335555555599
5	05
6	
7	
8	59

Interpretation

- A stem and leaf can help us to reveal the shape of the distribution of data and area of concentration.
- For instance, from above example, it can be seen that the data is skewed to the right, with most waitress's tips concentrating between 1.0 and 4.9.