Chapter 2: Methods of Describing Data Sets

- Describing Data Set Using Frequency Table
- Describing Data Set Using Graphs

(I) Qualitative Data

(A) Frequency Table

Example 1 (Student College)

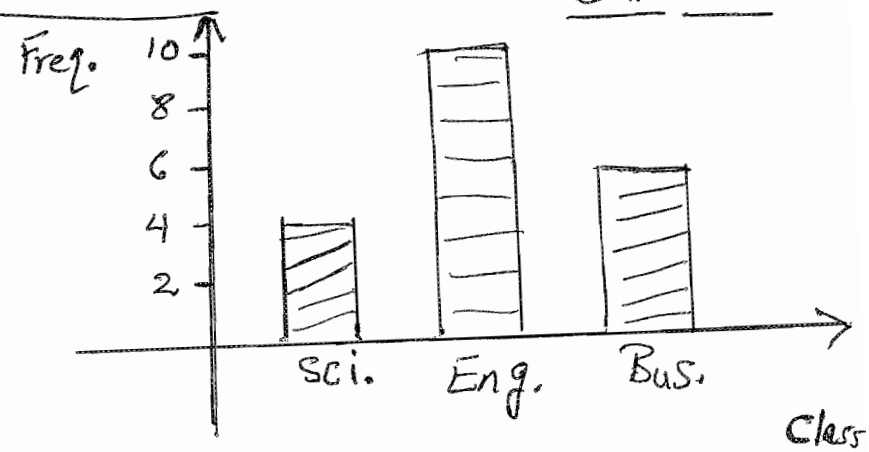
Sci., Eng., Bus., Bus., Eng., Sci., Eng.
Eng., Bus., Sci., Bus., Sci., Eng., Eng.
Eng., Bus., Eng., Bus., Eng., Eng.,

Class	Frequency	Relative Freq.	Class Percentage
Sci.	4	$\frac{4}{20} = 0.20$	20%
Eng.	10	$\frac{10}{20} = 0.50$	50%
Bus.	6	$\frac{6}{20} = 0.30$	30%
	20	1	100%

(B) Graphical Representation

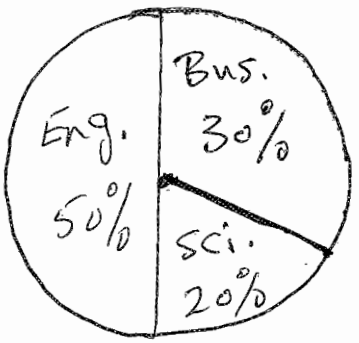
- Bar Chart

Bar chart



- Pie Chart

Class	Relative Freq.	Angle
Sci.	0.20	$0.20 \times 360^\circ = 72$
Eng.	0.50	$0.50 \times 360^\circ = 180$
Bus.	0.30	$0.30 \times 360^\circ = 108$



(2) Quantitative (Continuous) Data

Example 2: Grades of 20 students

13, 15, 22, 24, 28, 29, 31, 32

36, 36, 37, 38, 39, 39

43, 47, 48, 52, 54, 65

Note

(1) Minimum Grade = 13

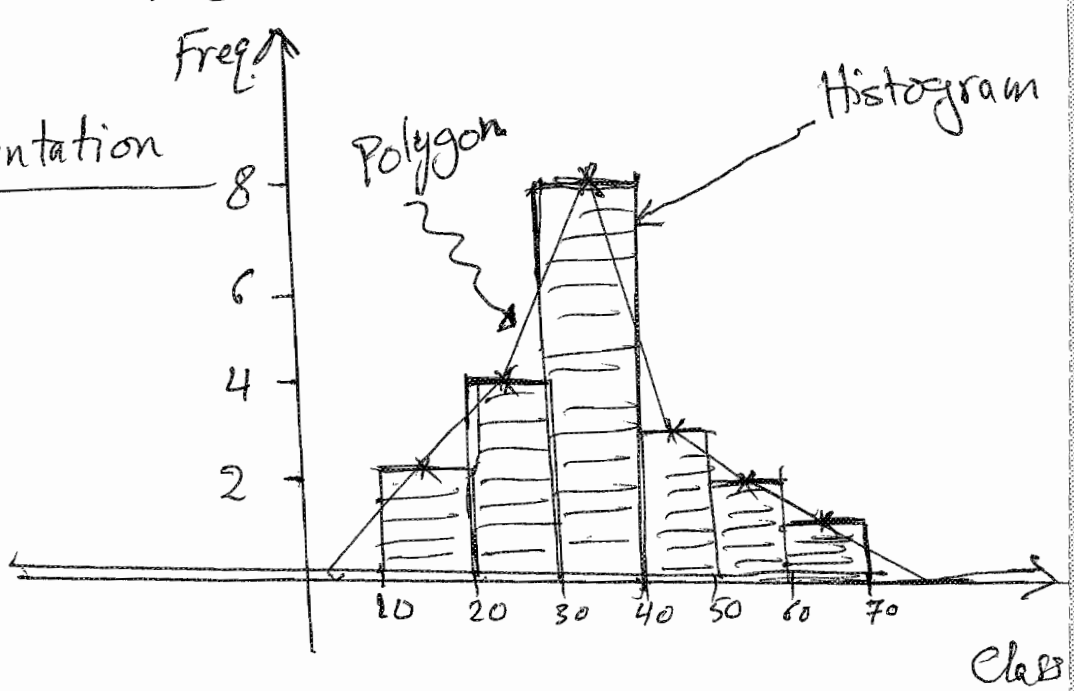
(2) Maximum Grade = 65

(3) # of observations = 20

Classes	Tally	Frequency	Relative Freq.	Cumulative Freq.	Cum. Rel. Freq.
10 - 19	//	2	0.10	2	0.10
20 - 29	////	4	0.20	6	0.30
30 - 39	//////	8	0.40	14	0.70
40 - 49	///	3	0.15	17	0.85
50 - 59	//	2	0.10	19	0.95
60 - 69	/	1	0.05	20	1
		20	1		

Graphical Presentation

(1) Histogram
Polygon



(2) Stems & Leafs

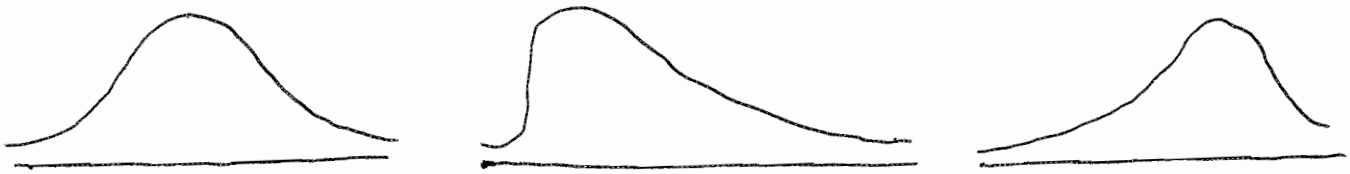
Stem unit = 10 , Leaf unit = 1

Stem	Leaf
1	3 5
2	2 4 8 9
3	1 2 6 6 7 8 9 9
4	3 7 8
5	2 4
6	5

Stem

Leaf

The distribution is skewed to right



Symmetric Distribution

Skewed to right

Skewed to left

Numerical Measures

Here we describe the data numerically.

* Sample data sets: $x_1, x_2, \cdots, x_n$

n = Sample size

$$\text{Sample mean} = \bar{X} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

* Population Data : $x_1, x_2, \cdots, x_N$

N = Population size

$$\text{Population mean} = \mu = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N}$$

(1) Measures of Central Tendency

* Mean

Data: 1, 2, 3, 4, 5

$$\bar{X} = \frac{1+2+3+4+5}{5} = \frac{15}{5} = \boxed{3}$$

* Median : Middle value

If n is odd, Median = value whose rank = $\frac{n+1}{2}$ (rank = position)

If n is even, Median = Average of values with ranks $\frac{n}{2}$ & $\frac{n}{2}+1$

Example 3 (Weights of new born babies)

3, 3.2, 2.5, 2.9, 3.4, 3.3, 3.1

$n = 7 \Rightarrow$ Position $= \frac{n+1}{2} = \frac{8}{2} = 4$

Ordering data $\Rightarrow$ 2.5, 2.9, 3, 3.1, 3.2, 3.3, 3.4

$\Rightarrow$ Median = 3.1

Example 4 : Same as Example 3 but with adding 2.9

Ordering data $\Rightarrow$ 2.5, 2.9, 2.9, 3, 3.1, 3.2, 3.3, 3.4

$n = 8 \Rightarrow$ ranks: $\frac{n}{2} = 4$, $\frac{n}{2}+1 = 5$

Median $= \frac{3.0+3.1}{2} = 3.05$

* Mode : Value with highest frequency

- Mode for data in Example 4 $= \boxed{2.9}$

- Modal Class : Midpoint of the class with highest frequency

Example 2 [Frequency Table]

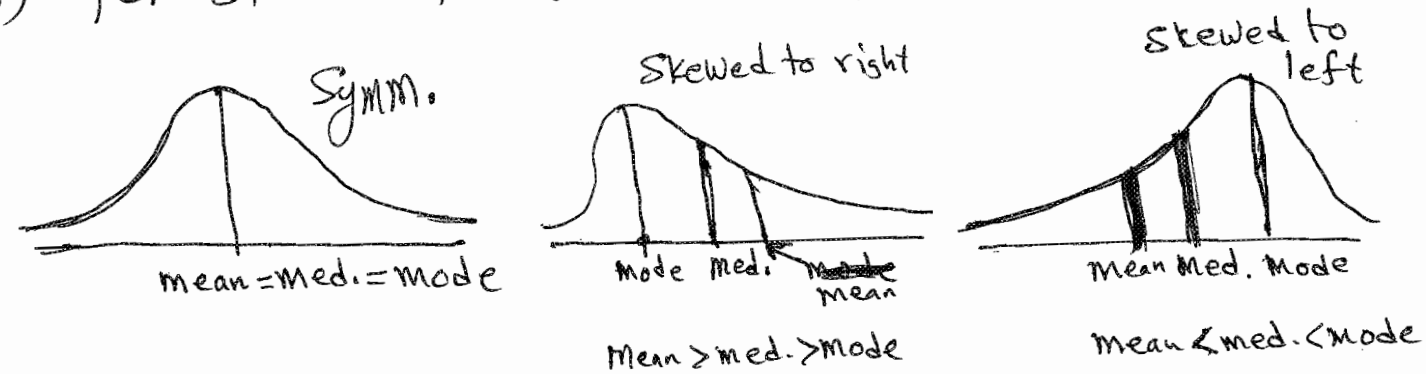
$$\text{Modal Class} = \frac{30+39}{2} = \frac{69}{2} = \boxed{34.5}$$

Notes

(1) For Symmetric data set, Mean = Median = Mode

(2) For skewed to right data set, Mean > Median > Mode

(3) For skewed to left data set, Mean < Median < Mode



Numerical Measures of Variation

(1) Range = Largest Value − Smallest Value

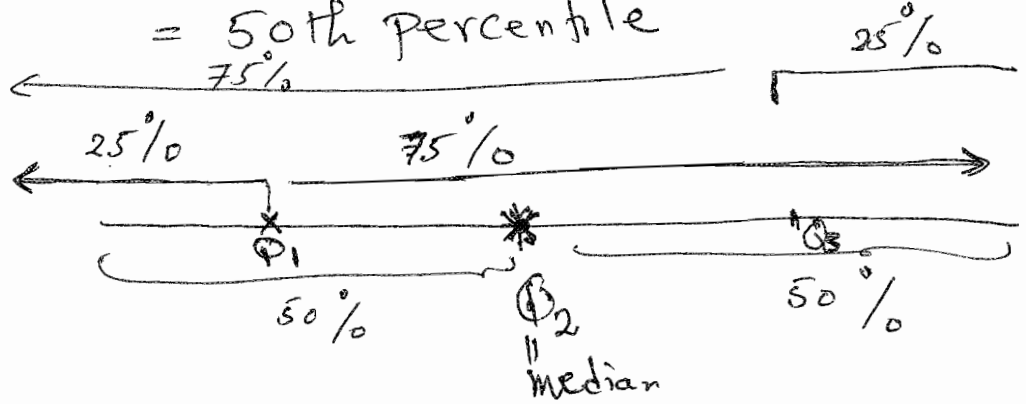
* Easy to compute
* Not informative

(2) Interquartile Range = $Q_U - Q_L$

upper quartile = 75th percentile (3rd quartile)

lower quartile = 25th percentile (First quartile)

Median = 2nd quartile = 50th percentile



Q_L = 1st quartile = value such that 25% of observations is below this value & 75% above

Q_U = 3rd quartile = value such that 75% of obs's is below this value & 25% above

IQ = Interquartile range = $Q_U - Q_L$

[Range for 50% of observations]

(3) Variance [measuring the spread of data]

σ^2 = population variance

s^2 = sample variance

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$ ← Avg. of squared deviations

$s = \sqrt{s^2}$ = sample standard deviation

= sample Std.

Data: 1, 2, 3, 4, 5 $\quad \bar{x} = \frac{15}{5} = 3$

x	$x-\bar{x}$	$(x-\bar{x})^2$
1	−2	4
2	−1	1
3	0	0
4	1	1
5	2	4
15	0	10

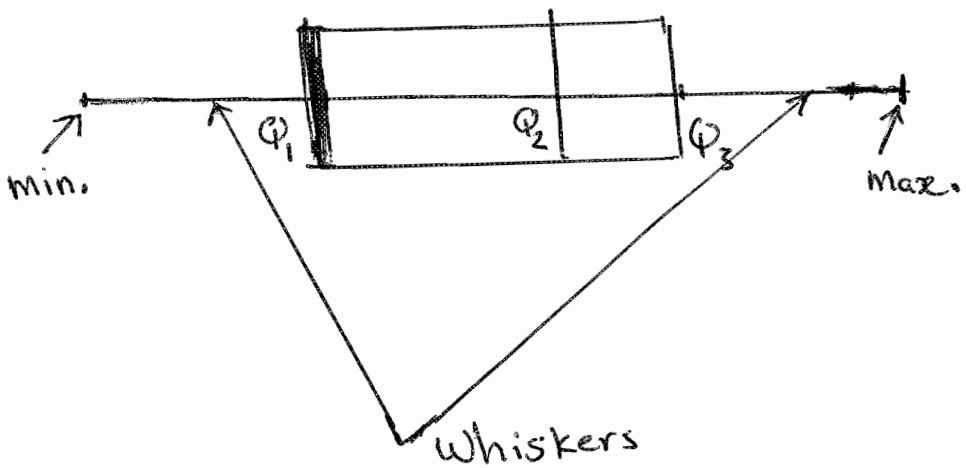
$$s^2 = \frac{\sum_{i=1}^{5}(x_i - \bar{x})^2}{4} = \frac{10}{4} = \boxed{2.5}$$

$$s = \sqrt{2.5} = \boxed{1.58}$$

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n} \qquad \sigma = \sqrt{\sigma^2}$$

Determining of Outliers

- Outlier is an observation that is unusually large or small relative to the other values.
- Reasons for Outliers
  - Entry Error
  - Observations come from different population
  - Rare event (small chance)
- Box plot of the Data Set



  - The size of the box indicates the size of variation
    Wide box $\Longrightarrow$ high variation
    small box $\Longrightarrow$ low variation
  - If the left whisker is longer $\Longrightarrow$ left skewed
    or if $Q_2 - Q_1 > Q_3 - Q_2 \Longrightarrow$ =
  - If the right whisker is longer $\Longrightarrow$ right skewed
    or if $Q_2 - Q_1 < Q_3 - Q_2 \Longrightarrow$ = =

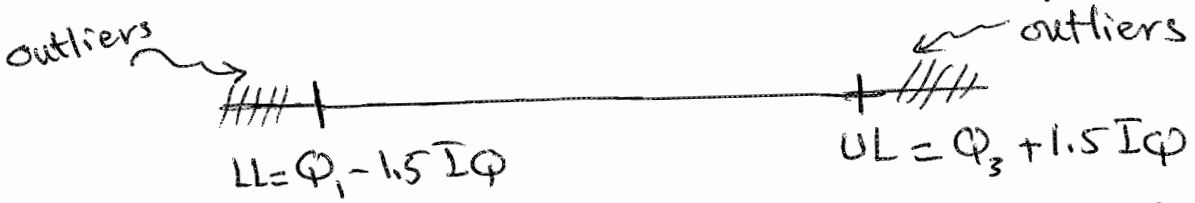
Determination of Outliers

Lower limit $= LL = Q_1 - 1.5\,IQ$

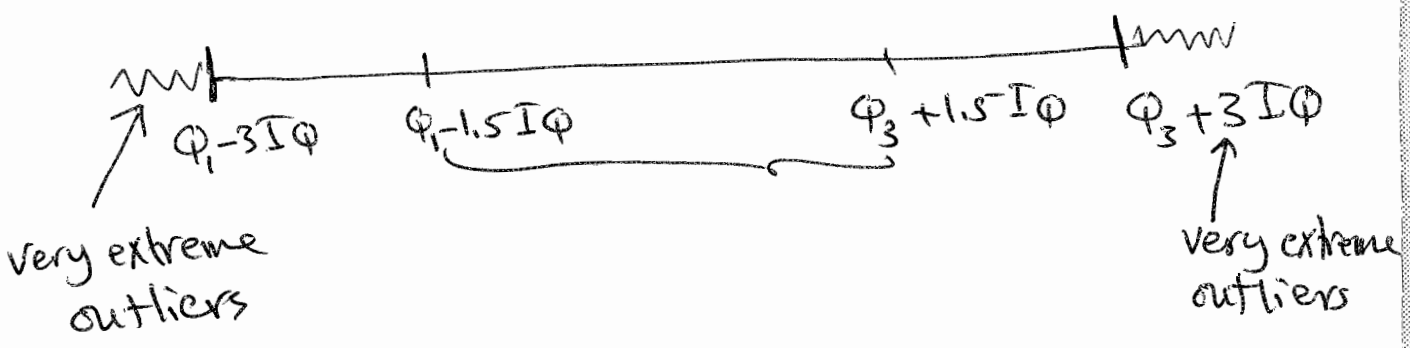
upper limit $= UL = Q_3 + 1.5\,IQ$

where $IQ = Q_3 - Q_1$

- The extreme values [extreme outliers] are the observations that are outside (LL, UL)



- The very extreme outliers are observations that are outside $(Q_1 - 3\,IQ, Q_3 + 3\,IQ)$



Example: Given the following data set

50 10 20 30 20 40 25

300 60 70

Determine whether the data contain outliers.

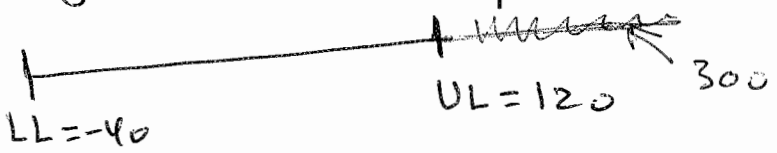
- ordering $\Rightarrow$ 10 20 20 25 30 40 50 60 70 300
- $n = 10 \Rightarrow$ positions for Median $\frac{n}{2} = 5$, $\frac{n}{2} + 1 = 6$

  $\Rightarrow Q_2 = \frac{30+40}{2} = \boxed{35}$
- For Q_1 & Q_3, $n = 5 \Rightarrow$ Position $= \frac{5+1}{2} = 3$

  $Q_1 = 20$, $Q_3 = 60$

$IQ = Q_3 - Q_1 = 60 - 20 = 40$

- $LL = Q_1 - 1.5\,IQ = 20 - (1.5)(40) = \boxed{-40}$
- $UL = Q_3 + 1.5\,IQ = 60 + (1.5)(40) = \boxed{120}$
- No small outliers since no observations below -40
  There is one large outlier $300 \notin (LL, UL)$



Another simple method [No available Quartiles]

- Compute Z-score $Z = \frac{X - \bar{X}}{S}$
- Any observation with $|Z| > 3$ is considered an outlier.

Example: The mean and standard deviation of the salaries of one of the companies are 4000 KD & 500 KD, resp. If one person is chosen & it is found that his salary is 2000 KD. Is his salary an outlier?

$\bar{X} = 4000$ KD & $S = 500$ KD

$Z\text{-score} = \frac{2000 - 4000}{500} = \frac{-2000}{500} = -4$

$|Z| = |-4| = 4 > 3$

$\Rightarrow$ Salary = 2000 is an outlier.