MAT 3103: Computational Statistics and Probability

Chapter 2: Descriptive Statistics

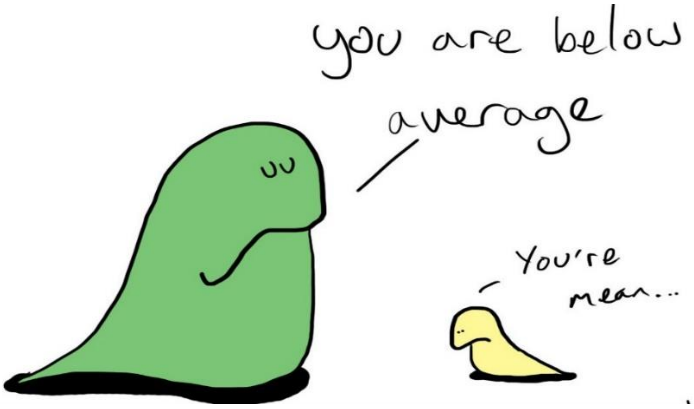


Image source: http://www.redpaper.in/genre/genre-entertainment/are-you-intimidated-by-the-idea-of-being-average/

Descriptive statistics:

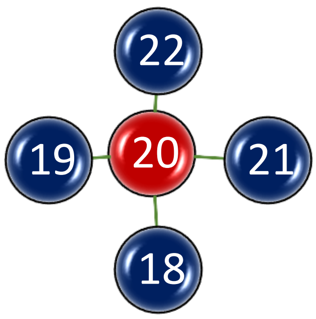
Descriptive statistics are brief descriptive measures that summarize a given data set, which can be either a representation of the entire or a sample of a population. Descriptive statistics are broken down into

i) **Measures of central tendency**
ii) **Measures of dispersion**

Descriptive statistics are very important because if we simply present our raw data it would be hard to visualize what the data is showing, especially if there is a lot of it. Descriptive statistics therefore enables us to present the data in a more meaningful way, which allows simpler interpretation of the data.

Central tendency:

Central tendency is a descriptive summary of a dataset through a single value that reflects the center of the data distribution. As we can see, average of the values is 20 and all the values have a tendency to nearing 20 (center).



Measures of Central Tendency: Measures of central tendency are statistical constants which enable us to comprehend in a single effort the significance of the whole. A number which represents the entire list of numbers. A value which describes some attributes of the population. It helps us to condense data in a single value. A score that indicates where the center of the distribution tends to be located.

There are different types of measures of central tendency; each has its own advantages and disadvantages. These are -

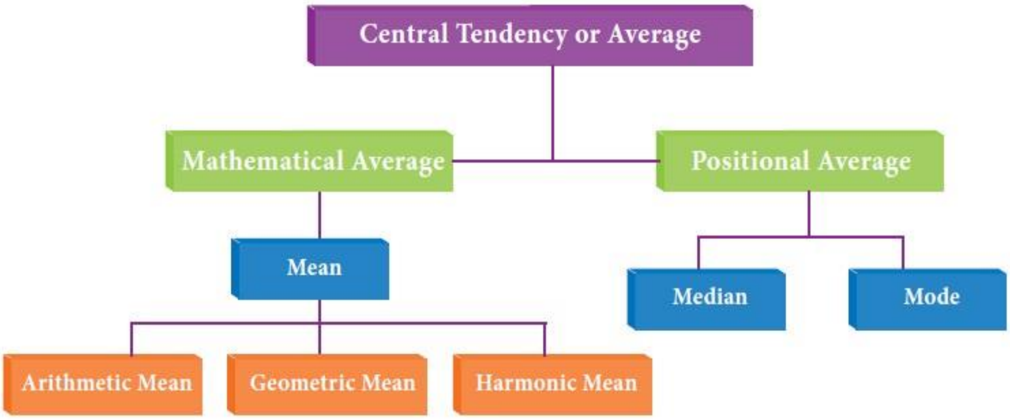
i. Mean **a) Arithmetic mean**

b) Geometric mean

c) Harmonic mean

ii. Median

iii. Mode



Source: https://www.brainkart.com/article/Various-measures-of-central-tendency 35079/

Arithmetic mean:

It is generally known as average. To find the mean, add up all the numbers and divide by the number of numbers. It is denoted as $\bar{x}$ or AM.

Calculation procedures:

For ungrouped data: Let $x_1, x_2, \dots, x_n$ are n variates, then, the arithmetic mean is defined by

$$\text{AM} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

For grouped data: Let $x_1, x_2, \dots, x_n$ are n variates with frequencies $f_1, f_2, f_3, \dots, f_n$ then, the arithmetic mean is defined by

$$\text{AM} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} f_i x_i,$$

where x_i = midpoints of groups, f_i = frequency, $n = \sum_{i=1}^{n} f_i$.

Advantages of the arithmetic mean:

- It is rigidly defined.
- It is easy to calculate and simple to follow.
- It is based on all the observations.
- It is determined for almost every kind of data.

Disadvantages of the arithmetic mean:

- The arithmetic mean is highly affected by extreme values.
- It cannot average the ratios and percentages properly.
- It is not an appropriate average for highly skewed distributions.
- It cannot be computed accurately if any item is missing.

Geometric mean:

Geometric mean is relevant when several quantities multiply together to produce a product or there is geometric progression in data. It is sometimes preferred for averaging ratios of two variables: rates of population growth, rates of interest, and rates of depreciation. The geometric mean of a set of n values of a variable is the nth root of their product.

Suppose you have an investment which earns 10% the first year, 50% the second year, and 30% the third year. What is its average rate of return? It is not the arithmetic mean, because what these numbers mean is that on the first year your investment was multiplied (not added to) by 1.10, on the second year it was multiplied by 1.60, and the third year it was multiplied by 1.20. The relevant quantity is the geometric mean of these three numbers.

Calculation procedures:

For ungrouped data: Let a variable *x* assumes *n* values $x_1, x_2, \ldots, x_n$. Then, geometric mean defined as,

$$\text{GM} = \bar{x}_G = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n} = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}}.$$

For grouped data: Let a variable *x* assumes *n* values $x_1, x_2, \ldots, x_n$ with respective frequencies as $f_1, f_2, \ldots, f_n$. Then,

$$\text{GM} = \bar{x}_G = \sqrt[n]{(x_1^{f_1} \cdot x_2^{f_2} \cdot \ \ldots \cdot x_n^{f_n})} = \left(\prod_{i=1}^{n} x_i{}^{f_i}\right)^{\frac{1}{n}},$$

$$\text{alternatively, } \bar{x}_G = \text{Antilog}\left(\frac{1}{n}\sum_{i=1}^{n} f_i log x_i\right),$$

where x_i = midpoints of groups, f_i = frequency, $n = \sum_{i=1}^{n} f_i$.

Advantages of the geometric mean:

- The geometric mean is rigidly defined.
- The geometric mean is directly based on all the observations.
- Extremely small or large values has no considerable effect on geometric mean.

Disadvantages of the geometric mean:

- It is difficult to compute.
- If a single value of a variable is zero, then the geometric mean becomes zero, irrespective of the magnitudes of the other values.
- It may be imaginary if some values are negative.

Harmonic mean:

Harmonic mean is helpful when dealing with datasets of rates or ratios (i.e. fractions) over different lengths or periods. For example, in first test a typist types 400 words in 50 minutes, in second test he types the same words (400) in 40 minutes and in third test he takes 30 minutes to type the 400 words. Then average time of typing can be calculated by harmonic mean.

Calculation procedures:

For ungrouped data: Let a variable *x* assumes n values $x_1, x_2, \ldots, x_n$. Then, harmonic mean

$$\text{HM} = \bar{x}_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}.$$

For grouped data: Let a variable *x* assumes *n* values $x_1, x_2, \ldots, x_n$ with respective frequencies as $f_1, f_2, \ldots, f_n$. Then,

$$\text{HM} = \bar{x}_H = \frac{n}{\sum_{i=1}^{n} \frac{f_i}{x_i}},$$

where x_i = midpoints of groups, f_i = frequency, $n = \sum_{i=1}^{n} f_i$.

Advantages of the harmonic mean:

- It is based on all observations.
- It is capable of algebraic treatment.
- It is an appropriate average for averaging ratios and rates.
- It does not give much weight to the large items

Disadvantages of the harmonic mean:

- It gives high weight-age to the small items.
- It cannot be calculated if any one of the items is zero.
- It is usually a value which does not exist in the given data.

Example 2.1: The quiz scores of 5 randomly selected students in a section of Mathematics course at AIUB are recorded as: 15, 14, 13, 17, and 15. Calculate arithmetic mean, geometric mean and harmonic mean for the given scores of the students.

Solution: Let first arrange the values in order as: 13, 14, 15, 15, 17. Then,

Arithmetic Mean (AM): $\bar{x} = \frac{1}{5}(13 + 14 + 15 + 15 + 17) = 14.80$

Geometric Mean (GM): $\bar{x}_G = \sqrt[5]{13 \times 14 \times 15 \times 15 \times 17} = 14.74$

Harmonic Mean (HM): $\bar{x}_H = \frac{5}{\frac{1}{13}+\frac{1}{14}+\frac{1}{15}+\frac{1}{15}+\frac{1}{17}} = 14.68$

Example 2.2: The distribution of number of faded signals of a day sent from different stations:

Class Interval of faded signals	1 - 3	3 - 5	5 - 7	Total
No. of stations (*f*)	2	5	3	10

Calculate arithmetic mean, geometric mean and harmonic mean for the distribution.

Solution:

Class	Frequency (*f*)	Mid value(*x*)	*fx*	$f \log x$	$\frac{f}{x}$
1 – 3	2	2	4	2log2	1
3 – 5	5	4	20	5log4	1.25
5 – 7	3	6	18	3log6	0.5
Total	n = 10		42	5.95	2.75

Arithmetic Mean (AM): $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} f_i x_i = \frac{42}{10} = 4.20$

Geometric Mean (GM): $\bar{x}_G$ = Antilog $(\frac{1}{n}\sum_{i=1}^{n} f_i log x_i)$ = Antilog $(\frac{5.95}{10})$ = 3.9

Harmonic Mean (HM): $\bar{x}_H = \frac{n}{\sum_{i=1}^{n}\frac{f_i}{x_i}} = \frac{10}{2.75} = 3.63$

Example 2.3: Show by examples that, AM ≥ GM ≥ HM.

Solution: Let a set of data *x:* 2, 3, 4, 5, 6.

AM = $\frac{2+3+4+5+6}{5} = \frac{20}{5} = 4$

GM = $\sqrt[5]{2 \times 3 \times 4 \times 5 \times 6} = \sqrt[5]{720} = 3.7279$

HM = $\frac{5}{\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\frac{1}{5}+\frac{1}{6}} = 3.4483$

In this case, AM > GM > HM

If data set is as *x*: 4, 4, 4, 4, 4.

Then, AM = GM = HM = 4

Note: As we can see, AM > GM > HM. If all the values are same, then AM = GM = HM. We must not use GM and HM for a dataset having a value zero (0) as no matter what the remaining values are, the result will always be 0 due to that single 0.

Example 2.4: The following is the age distribution of 1000 persons working in a large industrial house:

Age Group	20-25	25-30	30-35	35-40	40-45	45-50	50-55	55-60	60-65
No. of person	30	160	210	180	145	105	70	60	40

Calculate mean for the above data.

Solution:

Age group	**No. of person (*f*)**	**Mid Value (*x*)**	***fx***
20-25	30	22.5	675
25-30	160	27.5	4400
30-35	210	32.5	6825
35-40	180	37.5	6750
40-45	145	42.5	6162.5
45-50	105	47.5	4987.5
50-55	70	52.5	3675
55-60	60	57.5	3450
60-65	40	62.5	2500
Total	1000		39425

Mean, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} f_i x_i = \frac{39425}{1000} = 39.425$

Median:

The median is the number that is halfway into the dataset. It overcomes the limitation of arithmetic means' inability to deal with outliers (extreme values). The middle number; found by ordering all data points and picking out the one in the middle (or if there are two middle numbers, taking the mean of those two numbers) is the median.

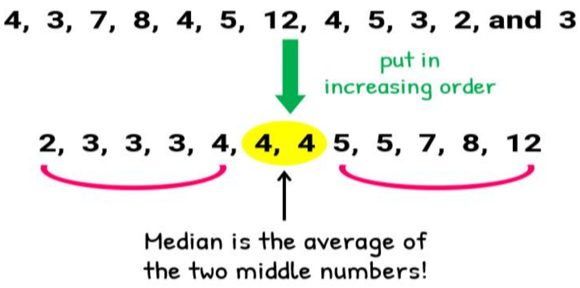


Image source: https://www.khanacademy.org/math/probability/data-distributions-a1/summarizing-center-distributions/a/choosing-the-best-measure-of-center

Calculation procedures:

Median is the middle value in the arrayed data (data arranged in either ascending or descending order).

For ungrouped data,

$$Me = \text{The value of } \frac{1}{2}(n+1)th\ observation \qquad [n\ odd]$$

$$= \text{The value of } \frac{1}{2}\left(\frac{n}{2}th\ observation + (\frac{n}{2}+1)th\ observation\right) \qquad [n\ even]$$

For grouped data,

$$\text{Median} = L + \frac{\frac{n}{2} - c}{f} \times h$$

L = lower limit of median class

h = size of median class

f = frequency of median class

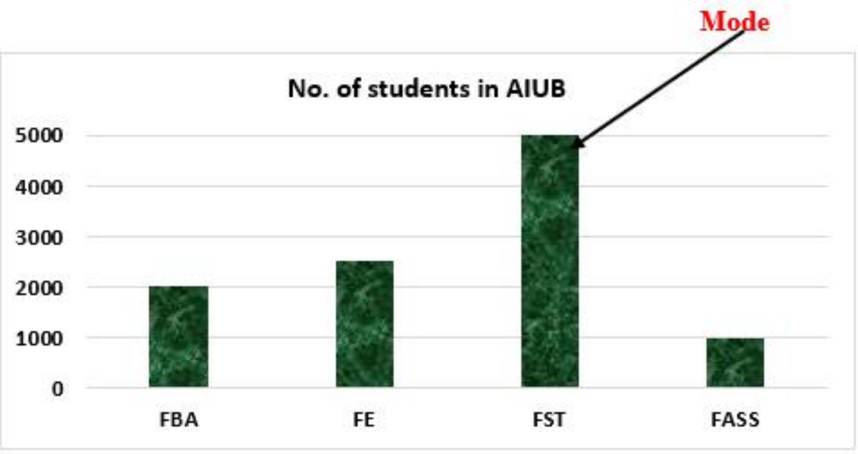
c = cumulative frequency of previous class of median class

Median class is that class for which cumulative frequency, $cf \geq \frac{n}{2}$.

Mode:

Mode is the most frequently occurring number found in a dataset. It is mainly used in situations where the variable under consideration is qualitative in nature. There can be more than one mode in a data set. If the two values are tied for being the most common values in the set, the data set can be said to be *bimodal*, whereas if three values are tied, the set is *trimodal*, and so on.

We can use bar diagram to find the mode for qualitative data.



Example 2.5:

The data set is given as x: 2, 3, 3, 4, 6. Mode = 3.

The data set is given as x: 2, 3, 3, 4, 6, 6. Mode = 3 and 6.

The data set is given as x: 2, 3, 3, 4, 4, 4, 6. Mode = 4

The data set is given as x: 2, 3, 4, 6. Mode = No mode.

For grouped data, Mode = $L + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times h$

L = lower limit of modal class

h = size of modal class

f_m = frequency of modal class

f_1 = frequency of previous class of modal class

f_2 = frequency of next class of modal class

Modal class is that class for which frequency is the highest one.

Example 2.6: Suppose the annual salaries in Taka of 7 employees in a small company are 17000, 20000, 28000, 18000, 1000, 120000 and 24000. Find the median salary of the employees.

Solution: array of x: 1000, 17000, 18000, 20000, 24000, 28000, 120000

n = 7 (an odd number)

$$Me = \text{The value of } \frac{1}{2}(n+1)\text{th number observation}$$

$$= \text{The value of 4th number observation}$$

$$= \text{Tk. 20000 per month}$$

So the median salary of the employee is 20,000 per month.

Example 2.7: In a football game, 10 players scored the following number of goals: 4, 1, 2, 6, 10, 8, 11, 3, 9, 7. Find the median of the data.

Solution: array of x: 1, 2, 3, 4, 6, 7, 8, 9, 10, 11

n = 10 (an even number)

$$Me = \text{The value of } \frac{1}{2}\left(\frac{n}{2}\text{th number observation} + \left(\frac{n}{2}+1\right)\text{th number observation}\right)$$

$$= \text{ The value of } \frac{1}{2}(\text{5th number observation} + \text{6th number observation})$$

$$= \frac{6+7}{2} \quad = 6.5$$

So, the median score is 6.5.

Example 2.8: A school district wants to research the class sizes in its elementary schools. The district gathers class sizes from each of its 20 elementary classes:

24, 25, 24, 30, 26, 25, 19, 29, 28, 29, 23, 31, 22, 22, 26, 21, 24, 25, 26, 26

Solution: The values of the data set in ascending order is: 9, 21, 22, 22, 23, 24, 24, 24, 25, 25, 25, 26, 26, 26, 26, 28, 29, 29, 30, 31

Mode = 26, since it occurs four times in the data set.

Example 2.9: The following data give the speeds (in miles per hour) of eight cars that were stopped on I-95 speed violations.

77, 82, 74, 81, 79, 84, 74, 78, 82

Find the mode.

Solution: The values of the data set in ascending order is: 74, 74, 77, 79, 81, 82, 82, 84

Since 74 and 82 occur twice, and each of the remaining values occurs only once. So,

Mode = 74 and 82 miles per hour.

Example 2.10: The following frequency distribution refers to the data for marks obtained by students in a paper. Compute mean, median of the frequency distribution.

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
No. of students	5	7	8	10	10	10	4	4	2

Solution: The calculation is shown in the table below:

Marks	Frequency, f	Mid Value, x	fx	Cumulative frequency, cf
0-10	5	5	25	5
10-20	7	15	105	12
20-30	8	25	200	20
30-40	10	35	350	30
40-50	10	45	450	40
50-60	10	55	550	50
60-70	4	65	260	54
70-80	4	75	300	58
80-90	2	85	170	60
Total	$n = 60$		$fx = 2410$	

Mean, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} f_i x_i$ = $\frac{2410}{60}$ = 40.16

Here, $\frac{n}{2}$ = $\frac{60}{2}$ = 30, for the class $30 - 40$, cf = 30 = 30. So, the median class is 30-40.

$$Me = L + \frac{\frac{n}{2} - c}{f} \times h \ = 30 + \frac{\frac{60}{2} - 20}{10} \times 10 = 40$$

Example 2.11: A survey on the heights (in cm) of 50 girls of class X was conducted at a school and the following data was obtained:

Height (in cm)	120-130	130-140	140-150	150-160	160-170	Total
No. of girls (*f*)	2	8	12	20	8	50

Calculate median and mode for the distribution.

Solution:

Class Interval	Number of Girls (*f*)	*cf*
120-130	2	2
130-140	8	10
140-150	12	22
150-160	20	42
160-170	8	50
Total	n = 50	

Here, $\frac{n}{2}$ = $\frac{50}{2}$ = 25. For the class (150 – 160), we get cf = 42 > 25. Hence, this is our median class.

Median, $Me = L + \frac{\frac{n}{2} - c}{f} \times h = 150 + \frac{25 - 22}{20} \times 10$ = 151.5

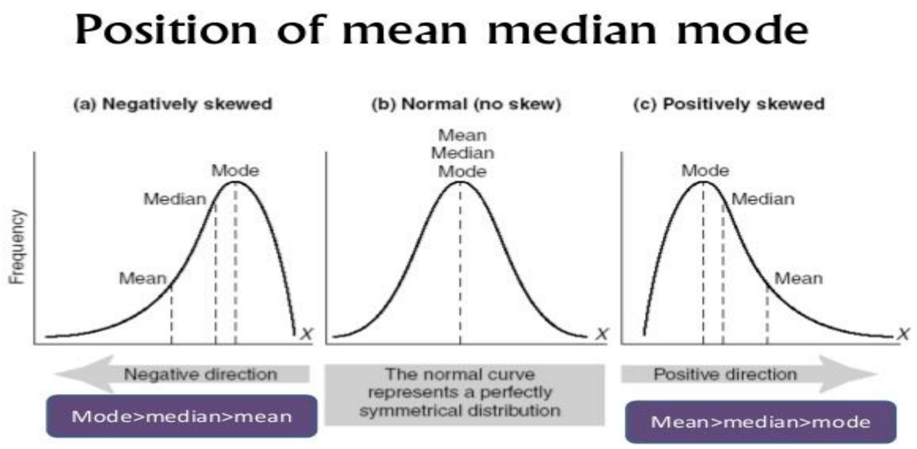
For the class (150 – 160), we have the highest frequency 20. Hence, this is our modal class.

Mode, $Mo = L + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times h = 150 + \frac{20 - 12}{40 - 12 - 8} \times 10$ = 154

Note: If f_m is there in the first class, then f_1 = 0. If f_m is there in the last class, then f_2 = 0. If there are two or more f_m, then separate modes should be calculated considering separate f_m.

Application of central tendency measures in describing shape of distributions:

Central tendency measures can help us in describing the shape of distributions through a measure called **skewness**. It measures the lack of symmetry in data distribution.



A **symmetrical** distribution will have a skewness of zero. If data is **positively skewed**, then it will have a much longer right tail than the left tail. If data is **negatively skewed**, then it will have a much longer left tail than the right tail.

Calculation and interpretation rules for skewness:

Measure of skewness	Interpretation rule
Mean = Median = Mode Mean > Median > Mode Mode > Median > Mean	the distribution is symmetric the distribution is positively skewed the distribution is negatively skewed
SK = mean – median Or SK = mean – mode	If SK = 0, the distribution would be symmetrical If SK > 0, the distribution would be positively skewed If SK < 0, the distribution would be negatively skewed

Example 2.12: Comment on the skewness of the distribution as given in example 2.10.

Solution: Based on the results we found in example 2.10, SK = mean – median = 40.16 – 40 = 0.16. So, the distribution given by example 2.10 is positively skewed.

Comment on the skewness of the given distributions:

1. 4, 6, 9, 12, 5; mean = 7.2; median = 6; mode = no mode
2. 7, 13, 4, 7; mean = 7.75; median = 7; mode = 7
3. 10, 3, 8, 15; mean = 9; median = 9; mode = no mode
4. 9, 9, 9, 9, 8; mean = 8.8; median = 9; mode = 9
5. 300, 24, 40, 50, 60; mean = 96.8; median = 50; mode = no mode
6. 23, 23, 12, 12; mean = 17.5; median = 17.5; mode = 12, 23

Dispersion:

Dispersion is a way of describing how spread out a set of data is. When a data set has a large value, the values in the set are widely scattered; when it is small the items in the set are tightly clustered.

Measures of central tendency might not always be helpful to describe data. Two datasets having same average could be entirely different in pattern. Let us illustrate it by a practical example. The average midterm scores of 2 courses of 5 students of CSE in AIUB are given as:

Courses	Scores					Average
Math	46	48	50	52	54	50
Statistics	10	40	50	60	90	50

In both courses, average scores are equal. But in Math, the observations are concentrated on the center. All students have almost the same level of performance. We say that there is consistence in the observations. In Statistics, the observations are not closed to the center. One observation is as small as 10 and one observation is as large as 90. Thus, there is greater dispersion in Statistics.

Measures of dispersion:

Measures of variability help communicate this by describing the shape and spread of the data set. Mean deviation, Variance, Standard deviation, coefficient of variation and quartiles are all examples of measures of variability.

Range:

Range is the simplest method of studying variation. It is defined as the difference between the value of the smallest observation and the value of the largest observation include in the distribution.

Range= Largest Value- Smallest Value

Mean deviation:

Mean deviation is the average distance between each observed value and the mean. Mean deviation is used frequently by engineers to show the variability of their data, although it is usually not the best choice. Its advantage is that it is simpler to calculate than other measures of dispersion.

Calculation procedures:

For ungroup data: Let a variable *x* assumes values $x_1, x_2, \dots, x_n$. Then,

$$\text{Mean deviation, MD} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

For grouped data: Let a variable *x* assumes values $x_1, x_2, \dots, x_n$ with respective frequencies as $f_1, f_2, \dots, f_n$. Then,

$$\text{Mean deviation, MD} = \frac{1}{n}\sum_{i=1}^{n} f_i|x_i - \bar{x}|$$

Variance:

Variance is the mean of the squares of the deviations of each observations from their mean. Note that variance has units of the quantity squared, for example m^2 or s^2 if the original quantity was measured in meters or seconds, respectively. Standard deviation overcomes this problem.

Calculation procedures:

For ungroup data: Let a variable *x* assumes values $x_1, x_2, \dots, x_n$. Then,

$$\text{Variance, } \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

For grouped data: Let a variable *x* assumes values $x_1, x_2, \dots, x_n$ with respective frequencies as $f_1, f_2, \dots, f_n$. Then,

$$\text{Variance, } \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} f_i(x_i - \bar{x})^2$$

Standard deviation:

Standard deviation is the positive square root of variance. Thus, it has the same units as the original data and is a representative of the deviations from the mean. Since the variance is the mean square of the deviations from the mean, the standard deviation is the root-mean-square deviation from the mean.

$$\text{Standard deviation (SD), } \sigma = \sqrt{\text{Variance}}$$

Root-mean-square quantities are crucial in describing the alternating current of electricity. An analogy can be drawn between the standard deviation and the radius of gyration encountered in applied mechanics.

Coefficient of variation:

Coefficient of variation is the ratio between the standard deviation and the mean for the same set of data, expressed as a percentage. When its value is 20%, it means that the observations vary, on an average, 20% with respect to mean. It is a unit free measurement, used to compare different sets of data having different units of measurement.

$$\text{Coefficient of variation, CV} = \frac{\text{Standard deviation}}{Mean} \times 100\% = \frac{\sigma}{\bar{x}} \times 100\%$$

Example 2.13: The following are the prices of a company from Sunday to Thursday

Day	Sunday	Monday	Tuesday	Wednesday	Thursday
Price (Tk)	200	210	208	160	250

Find the range.

Solution: Here,

Largest Value = 250

Smallest Value = 160

Range = Largest Value − Smallest Value

= 250-160

= 90Tk

Example 2.14: The following data relate to the marks obtained by nine students in a class test. Compute the mean deviation from this data set.

3, 4, 4, 5, 5, 7, 7, 7, 12

Solution: $Mean,\ \bar{x} = 6$

Mean deviation, $MD = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}| = \frac{|3-6|+|4-6|+\cdots\ldots+|12-6|}{8} = 2$

Example 2.15: The following frequency distribution gives the pattern of overtime work per week by 100 employees of a company. Calculate mean deviation of the following distribution:

Overtime (in hour)	10-15	15-20	20-25	25-30	30-35
Number of employees	3	5	7	4	2

Solution:

Overtime	x_i	f_i	$x_i f_i$	$\lvert x_i - \bar{x}\rvert$	$f_i \lvert x_i - \bar{x}\rvert$
10-15	12.5	3	37.5	9.286	27.858
15-20	17.5	5	87.5	4.286	21.43
20-25	22.5	7	157.5	0.714	4.998
25-30	27.5	4	110	5.714	22.856
30-35	32.5	2	65	10.714	21.428
Total		21	457.5		98.57

$$\bar{x} = \frac{\sum f_i x_i}{n} = \frac{457.5}{21} = 21.786$$

$$MD = \frac{1}{n}\sum_{i=1}^{n} f_i \lvert x_i - \bar{x}\rvert = \frac{98.57}{21} = 4.69$$

Example 2.16: A company of tea prices at 6 randomly selected grocery stores in an area of Sylhet city showed increase of 2, 4, 8, 6, 10, 12 Tk. per kilogram from the previous month. Find the variance, standard deviation and coefficient of variation of the price increases.

Solution:

x_i	$(x_i - \bar{x})^2$
2	$(2-7)^2=25$
4	$(4-7)^2=9$
8	$(8-7)^2=1$
6	$(6-7)^2=1$
10	$(10-7)^2=9$
12	$(12-7)^2=25$
$\sum x_i=42$	$\sum (x_i - \bar{x})^2=70$

$$\bar{x} = \frac{\sum x}{n} = \frac{42}{6} = 7$$

$$\sigma^2 = \frac{\sum (x-\bar{x})^2}{n} = \frac{70}{6} = 11.67$$

$$\sigma = \sqrt{11.67} = 3.42$$

$$CV = \frac{\sigma}{\bar{x}} \times 100 = 48.86\%$$

Example 2.17: The following frequency distribution is the data of days to maturity 40 short term investment. Calculate standard deviation and coefficient of variation.

Class Interval	30-40	40-50	50-60	60-70	70-80	80-90	90-100
Frequency	3	1	8	10	7	7	4

Solution:

Class Interval	f_i	x_i	$f_i\, x_i$	$(x-\bar{x})^2$	$f(x-\bar{x})^2$
30-40	3	35	105	1122.25	3366.75
40-50	1	45	45	552.25	552.25
50-60	8	55	440	182.25	1458
60-70	10	65	650	12.25	122.5
70-80	7	75	525	42.25	295.75
80-90	7	85	595	272.25	1905.75
90-100	4	95	380	702.25	2809
Total	40		2740		10510

$\bar{x} = \frac{\sum f_i x_i}{n} = \frac{2740}{40} = 68.5 \approx 67$

$\sigma^2 = \frac{\sum f_i (x_i - \bar{x})^2}{n} = \frac{10510}{40} = 262.75$

$\sigma = \sqrt{262.75} = 16.21$

Coefficient of Variation, $CV = \frac{\sigma}{\bar{x}} \times 100 = 23.66\%$

Example 2.18: The quiz scores of a student in a Math course at AIUB are recorded as: 10, 0, and 20. Calculate mean deviation, variance, standard deviation and coefficient of variation for the given scores of that student.

Solution:

<table>
<tr><th>Quiz</th><th>x</th><th>$\bar{x}$</th><th>x_i - $\bar{x}$</th><th>$|x_i - \bar{x}|$</th><th>$(x_i - \bar{x})^2$</th></tr>
<tr><td>1</td><td>10</td><td rowspan="4">30/3 = 10</td><td>10 – 10 = 0</td><td>0</td><td>0</td></tr>
<tr><td>2</td><td>0</td><td>0 – 1 0= -10</td><td>10</td><td>100</td></tr>
<tr><td>3</td><td>20</td><td>20 – 10= 10</td><td>10</td><td>100</td></tr>
<tr><td>Total</td><td>30</td><td>0</td><td>20</td><td>200</td></tr>
<tr><td colspan="4">MD = $\frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}| = \frac{20}{3} = 6.67$</td><td colspan="2">$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{200}{3} = 66.67$</td></tr>
<tr><td colspan="4">SD = σ = $\sqrt{\text{Variance}} = \sqrt{66.67} = 8.16$</td><td colspan="2">CV = $\frac{\mathbf{8.16}}{\mathbf{1}}$ × 100% = 81.65%</td></tr>
</table>

Example 2.19: The distribution of number of faded signals of a day sent from different stations is as:

Class Interval of faded signals	1 - 3	3 - 5	5 - 7	7 - 9	Total
No. of stations (*f*)	4	3	2	1	10

Calculate mean deviation, variance, standard deviation and coefficient of variance for the distribution.

Solution:

Class	f	x	fx	$\bar{x}$	x_i - $\bar{x}$	$f_i\|x_i - \bar{x}\|$	$f_i(x_i - \bar{x})^2$
1 – 3	4	2	8	40/10 = 4	2 – 4 = −2	8	16
3 – 5	3	4	12		4 – 4 = 0	0	0
5 – 7	2	6	12		6 – 4 = 2	4	8
7 – 9	1	8	8		8 – 4 = 4	4	16
Total	n = 10		40		0	16	40
Mean deviation, MD $= \frac{1}{n}\sum_{i=1}^{n} f_i\|x_i - \bar{x}\| = \frac{16}{10}$					Variance, $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} f_i(x_i - \bar{x})^2 = \frac{40}{10} = 4$		
Standard deviation, $\sigma = \sqrt{\text{Variance}} = \sqrt{4} = 2$					Coefficient of variation, CV $= \frac{2}{4} \times 100\% =$ 50%		

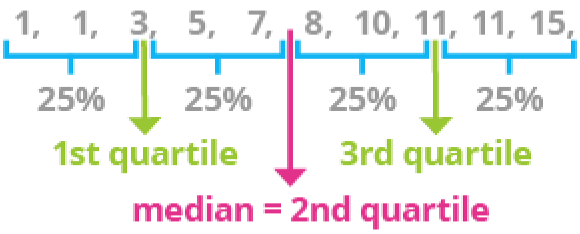
Quartiles:

Quartiles are the values that break down the dataset into quarters, or quartiles:

The first quartile (Q_1) is the point below which a quarter of the data lies. It is sometimes called the lower quartile.

The second quartile (Q_2) is the point below which half of the data lies. We already know this point by the name of median. It is also called the middle quartile.

The third quartile (Q_3) is the point below which three-quarters of the data lies. It is called the upper quartile.



Application of quartiles:

Quartiles are used to summarize a group of numbers. Instead of looking a big list of numbers, you are looking at just a few numbers that give you a picture of what's going on in the big list. Quartiles are great for reporting on a set of data and for making box and whisker plots. Quartiles are good for NOT symmetrically distributed data set, or a data set that has outliers.

Calculation procedure:

For ungrouped data: Let a variable x takes n values $\boldsymbol{x_1, x_2, \ldots, x_n}$. After arranging the observations in an ascending order, the quartiles are evaluated as:

$$Q_i = The\ value\ of\ \frac{i(n+1)}{4} th\ observation \qquad [n\ odd];\ \ i = 1,2,3$$

$$= The\ value\ of\ \frac{1}{2}\left[\frac{in}{4} th\ observation + (\frac{in}{4}+1)\ th\ observation\right] \qquad [n\ even];\ \ i = 1,2,3$$

For grouped data: Quartiles, $Q_i = L + \frac{\frac{in}{4}-c}{f} \times h$; i = 1, 2, 3

L = lower limit of quartile class

h = size of quartile class

f = frequency of quartile class

c = cumulative frequency of previous class of quartile class

Quartile class is that class for which cumulative frequency, $\boldsymbol{cf \geq \frac{in}{4}\ ; i = 1, 2, 3}$

Example 2.20: A data set of the no. of OPD patients in a clinic in a week is given as,

x: 2, 5, 3, 6, 7, 4, 9

Calculate the value of Q_1, Q_2 and Q_3.

Solution: Make an array, x: 2, 3, 4, 5, 6, 7, 9

Here, n = 7 (odd)

$$Q_i = The\ value\ of\ \frac{i(n+1)}{4} th\ observation$$

$$Q_1 = \ The\ value\ of\ \frac{(n+1)}{4} th\ observation$$

$$= The\ value\ of\ 2nd\ observation$$

$$= 3$$

$$Q_2 = \ The\ value\ of\ \frac{2(n+1)}{4} th\ observation$$

$$= The\ value\ of\ 4th\ observation$$

$$= 5$$

$$Q_3 = \ The\ value\ of\ \frac{3(n+1)}{4} th\ observation$$

$$= The\ value\ of 6th\ observation$$

$$= 7$$

Example 2.21: A data set of the no. of cancer patients in a hospital is given as.

x: 2, 5, 3, 6, 7, 4, 9,13

Calculate the value of $Q_{1,}Q_2$ and Q_3.

Solution: Make an array, x: 2, 3, 4, 5, 6, 7, 9, 13

Here, n = 8 (even)

$$Q_i = The\ value\ of\ \frac{1}{2}\left[\frac{in}{4} th\ observation + (\frac{in}{4} + 1)\ th\ observation\right]$$

$$Q_1 = The\ value\ of\ \frac{1}{2}\left[\frac{n}{4} th\ observation + (\frac{n}{4} + 1)\ th\ observation\right]$$

$$= The\ value\ of\ \frac{1}{2}[2nd\ observation + 3rd\ th\ observation]$$

$$= \frac{1}{2}(3 + 4)$$

$$= 3.5$$

$$Q_2 = The\ value\ of\ \frac{1}{2}\left[\frac{2n}{4} th\ observation + (\frac{2n}{4} + 1)\ th\ observation\right]$$

$$= The\ value\ of\ \frac{1}{2}[4th\ observation + 5th\ th\ observation]$$

$$= \frac{1}{2}(5 + 6)$$

$$= 5.5$$

$$Q_3 = The\ value\ of\ \frac{1}{2}\left[\frac{3n}{4} th\ observation + (\frac{3n}{4} + 1)\ th\ observation\right]$$

$$= The\ value\ of\ \frac{1}{2}[6th\ observation + 7th\ th\ observation]$$

$$= \frac{1}{2}(7 + 9)$$

$$= 8$$

Example 2.22: The distribution of temperature of 20 cities is given below. Find the value of Q_1, Q_2 and Q_3 from these given data.

Temperature	No of cities	Cumulative frequency
0-10	2	2
10-20	3	5
20-30	5	10
30-40	2	12

40-50	6	18
50-60	2	20
Total	20	

Solution: $Q_i = L + \frac{\frac{in}{4} - c}{f} \times h$

For $i = 1$, $\frac{\mathrm{n}}{4} = \frac{20}{4} = 5$. In the class (10 – 20), *cf* = 5. Hence, this is our class for Q_1.

$$Q_1 = L + \frac{\frac{n}{4} - \mathrm{c}}{f} \times h = 10 + \frac{5-2}{3} \times 10 = 20$$

For $i = 2$, $\frac{2\mathrm{n}}{4} = \frac{40}{4} = 10$. In the class (20 – 30), *cf* = 10. Hence, this is our class for Q_2.

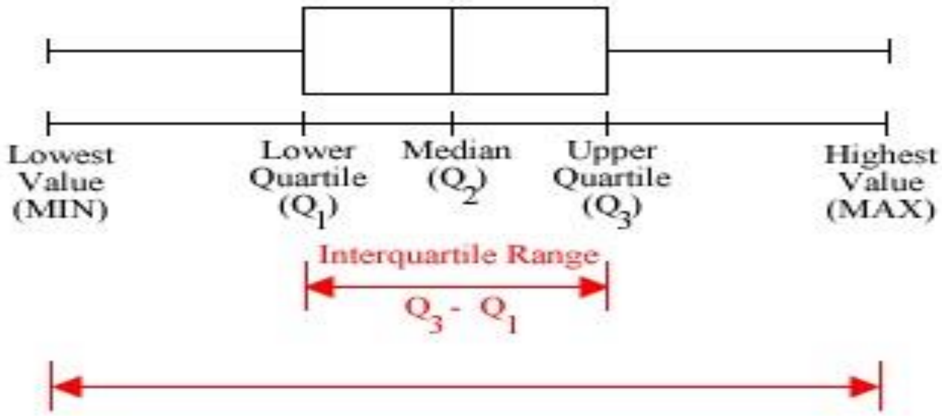
$$Q_2 = L + \frac{\frac{2n}{4} - \mathrm{c}}{f} \times h = 20 + \frac{10-5}{3} \times 10 = 36.67$$

For $i = 3$, $\frac{3\mathrm{n}}{4} = \frac{60}{4} = 15$. In the class (40 – 50), *cf* = 18 > 15. Hence, this is our class for Q_3.

$$Q_3 = L + \frac{\frac{3n}{4} - \mathrm{c}}{f} \times h = 40 + \frac{15-12}{6} \times 10 = 45$$

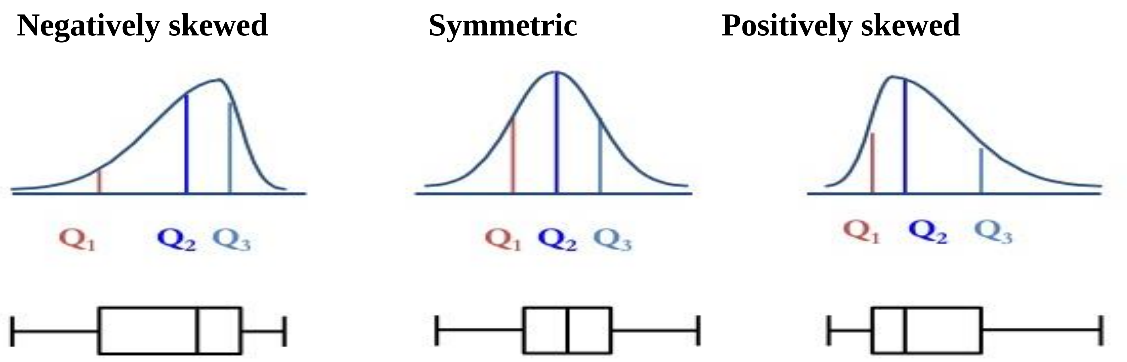
Box-and-whisker plot:

Boxplots are a standardized way of displaying the distribution of data based on a five-number summary (minimum, first quartile (Q_1), median, third quartile (Q_3), and maximum).



Application of box-and-whisker plot:

Box plots are useful as they show outliers within a data set. Box plots also show the average score of a data set. The median is the average value from a set of data and is shown by the line that divides the box into two parts. Half the scores are greater than or equal to this value and half are less. The box plot shape will show if a statistical data set is normally distributed or skewed.



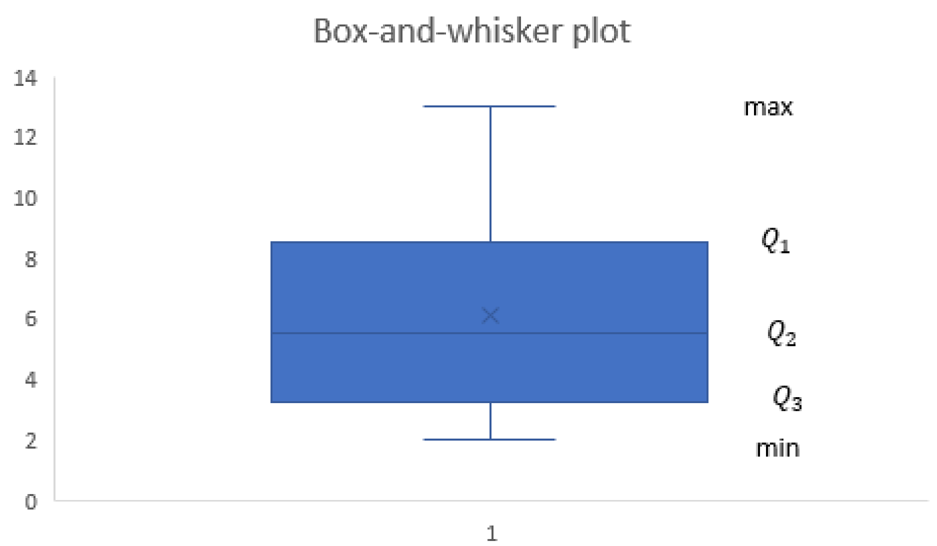
Example 2.23: : A data set of the no. of cancer patients in a hospital is given as.

x: 2, 5, 3, 6, 7, 4, 9,13

Draw a box-and-whisker plot.

Solution: In the given data set (from Example 2.18)

Min = 2, Q_1 = 3.5, Q_2 = 5.5, Q_3 = 8, Max = 13.



Screening of the Data: Data screening should be carried out prior to any statistical procedure. The screening of the data after collection and before analysis is probably the most important part of data analysis. Often data screening procedures are so tedious that they are skipped. Then, after an analysis produces unanticipated results, the data are scrutinized. This step is, however, of utmost importance as it provides the foundation for any subsequent analysis and decision-making which rests on the accuracy of the data. This procedure performs a screening of data in a database, reporting on the:

1. Detect and correct data errors
2. Type of data (discrete or continuous)
3. Missing-value patterns
4. Data transformations & standardizations
5. Presence of outliers

Detect and correct data errors: Examine summary statistics (e.g., n, mean, min, max) and check for irregularities.

Where did all the data go?

	variable	nobs	min	max	mean	median	sum	sd	cv	zeros	pct.zeros	missing	pct.missing	se	se.ratio
1	AMGO	32	1	1	1.000	1	1	NA	NA	0	0.000	31	$\boxed{96.875}$	NA	NA
2	AMRO	32	0	5	0.625	0	20	1.129	180.640	21	65.625	0	0.000	0.200	32.000
3	BCCH	32	0	2	0.125	0	4	0.421	336.800	29	90.625	0	0.000	0.074	59.200
4	BEKI	32	0	0	0.000	0	0	0.000	NaN	32	100.000	0	0.000	0.000	NaN
5	BEWR	32	0	1	0.062	0	2	0.246	396.774	30	93.750	0	0.000	0.043	69.355
6	BGWA	32	0	2	1.379	2	40	0.942	68.310	9	28.125	3	9.375	0.175	12.690
7	BHGR	32	1	$\boxed{250}$	10.875	5	348	43.680	401.655	0	0.000	0	0.000	7.722	71.007

Unrealistic value?

Action:

- Correct errors in the raw data.
- Fixing or removing incorrect data.
- Identify unusual cases.
- Distinguish duplicate cases.
- Manual check for other anomalies.

Missing-value patterns: Evaluate volume and pattern of missing data and take corrective action, if needed:

e.g., median replacement

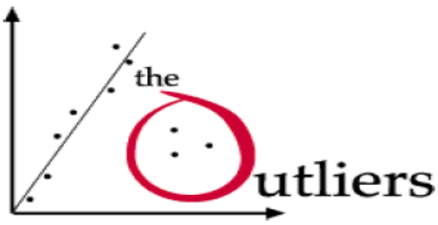
	AMGO	AMRO	BCCH	BEKI	BEWR	BGWA	BHGR
1	0	1	0	5	0	NA	25
2	NA	0	0	0	0	0	4
3	0	5	2	0	0	2	1
4	0	0	0	0	0	2	1
5	0	3	0	0	1	2	1
6	1	1	0	0	0	2	1
7	0	0	0	0	0	2	1
8	0	0	0	0	0	2	1
9	0	0	0	0	0	2	1
10	0	0	0	0	0	2	1
11	0	1	1	0	0	2	1
12	0	2	0	0	0	2	1
13	0	0	1	0	0	0	5
14	0	2	0	0	1	0	5
15	0	1	0	0	0	NA	5
16	0	1	0	0	0	NA	5

	AMGO	AMRO	BCCH	BEKI	BEWR	BGWA	BHGR
1	0	1	0	5	0	$\boxed{2}$	25
2	$\boxed{0}$	0	0	0	0	0	4
3	0	5	2	0	0	2	1
4	0	0	0	0	0	2	1
5	0	3	0	0	1	2	1
6	1	1	0	0	0	2	1
7	0	0	0	0	0	2	1
8	0	0	0	0	0	2	1
9	0	0	0	0	0	2	1
10	0	0	0	0	0	2	1
11	0	1	1	0	0	2	1
12	0	2	0	0	0	2	1
13	0	0	1	0	0	0	5
14	0	2	0	0	1	0	5
15	0	1	0	0	0	$\boxed{2}$	5
16	0	1	0	0	0	$\boxed{2}$	5

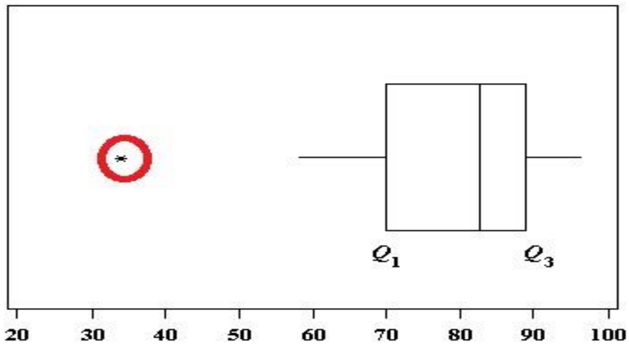
Action: Replace with prior knowledge; insert means or medians; use regression to estimate values.

Finding outliers:

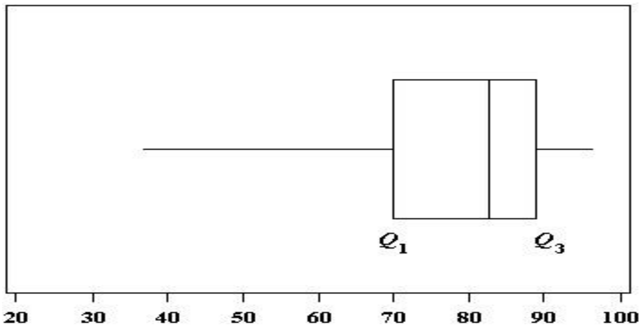
To ease the discovery of outliers, we have plenty of methods in statistics, but we will only be discussing two of them. Mostly we will try to see visualization methods (easiest ones) rather mathematical.



An outlier is an observation that is numerically distant from the rest of the data, that is, it lies outside the other values in the set. Box plots are useful as they show outliers within a data set. The whiskers of a box-and-whisker chart reach out to include outliers.



Some boxplots may not show outliers. For example, this chart has whiskers that reach out to include outliers:



Therefore, don't rely on finding outliers from a box and whiskers chart. That said, box and whiskers charts can be a useful tool to display them after calculating what the outliers are. The most effective way to find all of your outliers is by using the interquartile range (IQR). The IQR contains the middle bulk of the data.

If a data point is below **Low value** or above **High value**, it will be defined as an outlier.

Low $= Q_1 -$ 1.5* IQR

High $= Q_3 +$ 1.5* IQR

IQR = $Q_3 - Q_1$

Example 2.24: For the given data set identify is there any outlier if there find the value and remove this.

x: 32, 14, 70, 29, 22, 19, 3, 36, 49, 10

Solution:

$Q_1 = 14$

$Q_2 = 25.5$

$Q_3 = 36$

IQR = 22

High = 36 + 1.5*22 = 69

Low = 14 – 1.5*22 = – 19

The data set in order

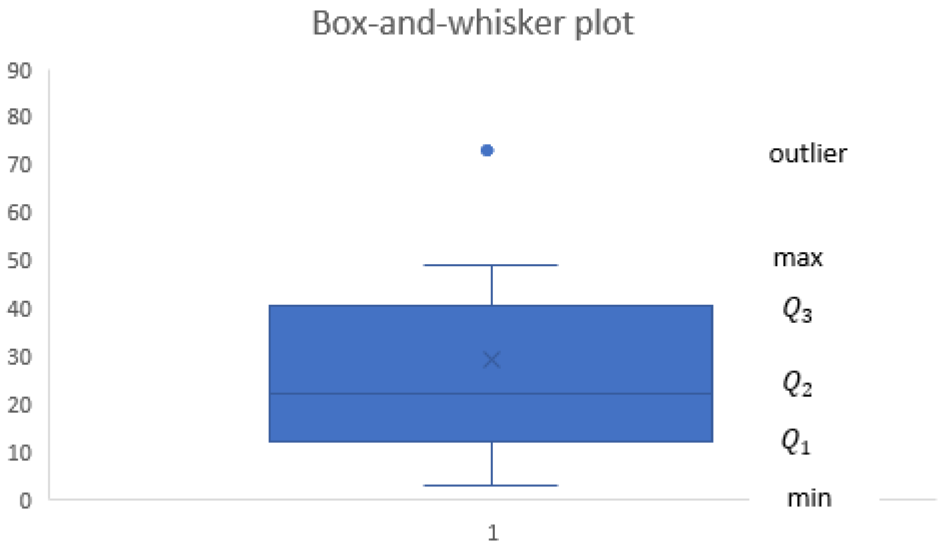
3, 10, 14, 19, 22, 29, 32, 36, 49, 70

Insert low and high values into the data set, in order

-19, 3, 10, 14, 19, 22, 29, 32, 36, 49, **69**, 70

So, outlier is 70. For further analysis remove the value.

Visualization of outlier by box-and-whisker plot.



So, bad data, wrong calculation, these can be identified as outliers and should be dropped but at the same time you might want to correct them too, as they change the level of data i.e. mean which cause issues when we model our data. For example 5 people get salary of 10K, 20K, 30K, 40K and 50K and suddenly one of the person start getting salary of 100K. Consider this situation as, you are the employer, the new salary update might be seen as biased and you might need to increase other employee's salary too, to keep the balance. So, there can be multiple reasons you want to understand and correct the outliers.

MATLAB code

Sum computes the sum of the columns of a matrix

```
>> x=[1 2 3; 4 5 6]
x =
1 2 3
4 5 6
```

- column sum

```
>> sum(x)
ans = 5 7 9
```

- row sum

```
>> sum(x')
```

ans = 6 15

mean(x) computes the mean of a vector or matrix

```
>> x=[1 2 3; 4 5 6]
x =
1 2 327
4 5 6
>> mean(x)
ans = 2.5000 3.5000 4.5000
```

var computes the sample variance of a vector or matrix

```
>> x=[1 2 3; 4 5 6]
x =
1 2 3
4 5 6
>> var(x)
ans = 4.5000 4.5000 4.5000
```

cov computes the sample covariance of a vector or matrix

```
x =
1 2 3
4 5 6
>> cov(x)
ans =   4.5000      4.5000      4.5000
        4.5000      4.5000       4.5000
```

std computes the sample standard deviation of a vector or matrix (column-by-column)

```
>> x=[1 2 3; 4 5 6]
x =  1 2 3
     4 5 6
>> std(x)
ans = 2.1213     2.1213     2.1213
```

sort orders the values in a vector or the rows of a matrix from smallest to largest

```
>> x=[1 5 2; 4 3 6]
x =
1 5 2
4 3 6
>> Y=sort(x)
ans = 1 3 2
      4 5 6
>> med = median(Y)
>> mod = mode(x)
```

skewness computes the sample skewness of a vector or matrix (column-by-column)

```
>> x=[1 2 3; 4 5 6]
x =
1 2 3
4 5 6
>> skewness(x)
ans = 0 0 0
```

kurtosis computes the sample kurtosis of a vector or matrix

```
>> x=[1 2 3; 4 5 6]
x =
1 2 3
4 5 6
>> kurtosis(x)
ans = 1 1 1
```

Quartile computes the sample quartiles of a vector or matrix

```
>> x = [1 2 3 4 7 10; 2 5 6 10 11 13]
x =
1 2 3 4 7 10
2 5 6 10 11 13
>>  quartile(x)
ans = 2.2500 5.2500
      3.5000 8.0000
      6.2500 10.7500
```

<u>Exercise 2</u>

1. Show, by examples, that AM $\geq$ GM $\geq$ HM.

2. In a perfectly symmetrical frequency distribution, the mean of the data is 88. Find median and mode.

3. The mean weight of five complete computer stations is 167.2 pounds. The weights of four of the computer stations are 158.4 pounds, 162.8 pounds, 165 pounds, and 178.2 pounds respectively. What is the weight of the fifth computer station?

4. A student recorded her scores on weekly math quizzes that were marked out of a possible 10 points. Her scores were as follows: 8, 5, 8, 5, 7, 6, 7, 7, 5, 7, 5, 5, 6, 6, 9, 8, 9, 7, 9, 9, 6, 8, 6, 6, 7 What is the mode of her scores on the weekly math quizzes?

4. A group committed to quality television has been concerned about a new talk show. For two weeks, they decide to count the number of words that must be "bleeped" as too obscene

for television and the number of physical altercations. They hope that after recording this data that they will be able to argue that the show is inappropriate for television particularly during the day. The data for number of words censored is provided below.

342, 267, 321, 157, 33, 349, 254, 166, 132, 289

(i) Find the Mean, Median, and Mode for the above data.

(ii) What does this information tell you about the talk show?

(iii) Is this data skewed?

5. For any two numbers, AM = 10 and GM = 8. Find out the numbers.

6. A train moves 1st 80 km at speed 75 km/h, 2nd 70 km at speed 85 km/h, 3rd 85 km at speed 66 km/h and 4th 55 km at speed 50 km/h, find the average speed throughout the journey.

7. The populations (in millions) in 2000 on each of the six inhabited continents were 803, 487, 348, 3686, 730, and 31. Which measure of central tendency best represents the data? Explain.

8. Alam took five tests in a class and had scores of 92, 75, 95, 90, and 98. Find the mean deviation for her test scores.

9. A video shop owner wants to find out the performance sales of his two branch stores for the last five months. The table shows their monthly sales in thousands of pesos.

Branch A	20	18	18	19	17
Branch B	17	15	25	17	18

(i) Find the variance, mean deviation and coefficient of variation of the two branch stores.

(ii) Which store is performing consistently?

10. A company has two sections with 40 and 65 staffs respectively. Their average weekly wages are \$450 and \$350. The standard deviations are 7 and 9. Which section has larger variability in wages?

11. Following data represent the number of computer centers in different localities in Dhaka. Calculate the average computers per locality.

Class interval of centers	No. of localities		
10-20	8		
20-30	7		
30-40	5		
40-50	6		
50-60	4		

12. The following data represent the distribution of time (minutes) needed to develop computer programs for solving some mathematical problems.

Class interval of time	0-10	10-20	20-30	30-40	40-50	Total
No. of programs	6	8	9	8	9	40

Calculate the mean time of developing a computer program.

13. The following is the distribution of consumption of electricity (MW/locality) in different days:

Class	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
Frequency	3	12	8	7	15	26	5	3	1

Calculate harmonic mean of the distribution.

Class	frequency			
10-20	3			
20-30	12			
30-40	8			
40-50	7			
50-60	15			
60-70	26			
70-80	5			
80-90	3			
90-100	1			

14. The following table shows the salary structure of Statsville Plush Toys, Inc. Assume that salaries exactly on an interval boundary have been placed in the higher interval.

Salary range	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100	Total
No. of employees (*f*)	12	24	32	19	9	3	0	1	100

Calculate

(a) Arithmetic, geometric and harmonic mean

(b) Median and mode of the data set

(c) Skewness and comment.

(d) Mean deviation

(e) Variance and Standard deviation

(f) Coefficient of variation

15. The analysis of the sales of 500 firms in an industry give the following distribution

Sales (in thousand Tk.)	0-50	50-100	100-150	150-200	200-250	250-300
No. of firms	3	24	55	98	120	95

Calculate

(a) Arithmetic, geometric and harmonic mean

(b) Median and mode of the data set

(c) Skewness and comment.

(d) Mean deviation

(e) Variance and Standard deviation

(f) Q_1 and Q_3, also draw a box $-$ and $-$ whisker plot.

17. The following frequency distribution refers to the number of hours worked per month of 50 workers of a factory. Compute median and mode of the frequency distribution.

Number of hours worked	30-55	55-80	80-105	105-130	130-155	155-180	180-205
Number of workers	3	4	6	9	12	11	5

Calculate

(a) Arithmetic, geometric and harmonic mean
(b) Median and mode of the data set
(c) Skewness and comment.
(d) Mean deviation
(e) Variance and Standard deviation
(f) Q_1 and Q_3, also draw a box $-$ and $-$ whisker plot.

18. What are outliers? How can we screen the data if the outliers are present in the given data?

x: 84, 10, 32, 19, 21, 29, 33, 15, 49, 2, 7, 52.

19. How do you detect and correct the collected information?

20. How do you perform with the given data having missing values?

	Col 1	**Col 2**	**Col 3**	**Col 4**	**Col 5**
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

Sample MCQs

1. The wind speed (knot) 10 random days is: 6.2, 5.8, 7, 6.8, 4.8, 5.2, 5.8, 6.4, 6.7, 5.9. What will be the median?

a. 7 b. 6.4 c. 6.8 d. 6.05

2. The distribution of number of emails received by a person in different days are given as:

Class interval	4 - 6	6 - 8	8 - 10	10 - 12	12 - 14	Total
Frequency	15	18	7	8	2	50

Find mean number of emails received per day by the person.

a. 5.7 b. 6.4 c. 6.8 d. 7.56

3. The distribution of production of cement (in million tons) of a factory in different days is:

Class interval	4 - 6	6 - 8	8 - 10	10 - 12	12 - 14	Total
Frequency	24	62	8	4	2	100

Find variance of the distribution.

a. 1.71 b. 4.62 c. 3.84 d. 2.64

4. For a given data set *x*: 30, 11, 75, 29, 22, 20, 3, 35, 47, 9; the outlier is:

a. 75 b. 22 c. 49 d. 3

5. The distribution of number of emails received by a person in different days is given as:

Class interval	4 - 6	6 - 8	8 - 10	10 - 12	12 - 14	Total
Frequency	15	18	7	8	2	50

Find mode of the distribution.

a. 8.7 b. 6.6 c. 7.8 d. 6.4

6. The distribution of number of emails received by a person in different days is given as:

Class interval	4 - 6	6 - 8	8 - 10	10 - 12	12 - 14	Total
Frequency	15	18	7	8	2	50

Find Q_3 of the distribution.

a. 8.7 b. 6.6 c. 7.8 d. 9.3

7. The distribution of production of cement (in million tons) of a factory in different days is:

Class interval	4 - 6	6 - 8	8 - 10	10 - 12	12 - 14	Total
Frequency	24	62	8	4	2	100

Find coefficient of variation of the distribution.

a. 18.45 b. 34.62 c. 13.84 d. 23.41

8. The wind speed (knot) of 10 random days is: 6.2, 5.8, 7, 6.8, 4.8, 5.2, 5.8, 6.4, 6.7, 5.9. What will be the highest quartile (Q_3)?

a. 7.51 b. 7.62 c. 6.75 d. 5.43