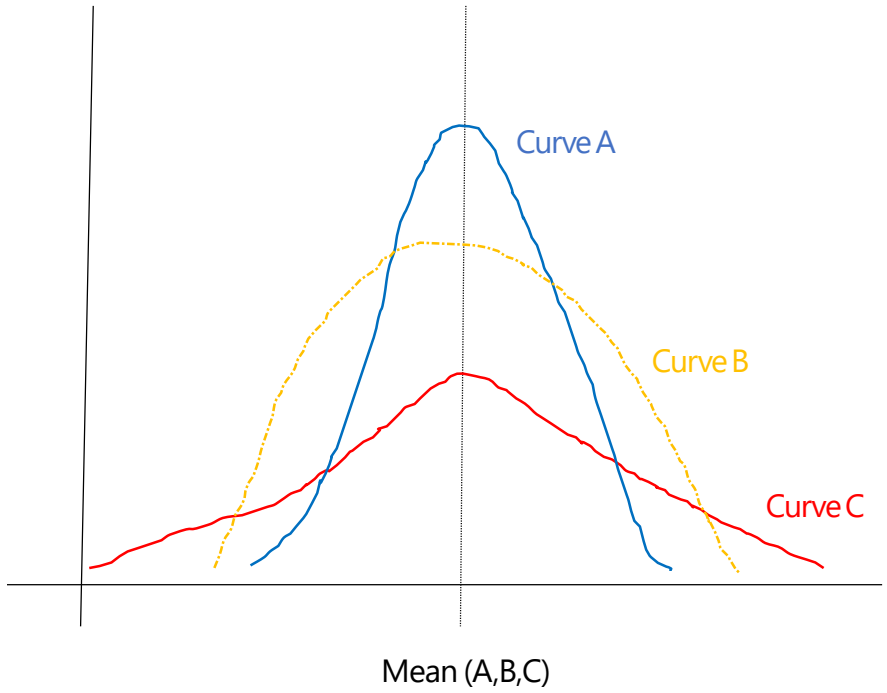
II. Measure of Dispersion

Dispersion measures the spread or variability of data

1. Range
2. Quartiles
3. Interquartile Range
4. Variance
5. Standard Deviation



1. Range

- Difference between the highest and the lowest observed values in a dataset
- Easy to understand and find
- Usefulness as a dispersion measure is limited – only 2 values are considered
- Heavily influenced by extreme values
- Range values may change from one sample to another
- For open-ended class, there is no *range*

Example

Values	Max	Min	range
22	90	6	84
49			
78			
6			
78			
76			
44			
90			
18			
63			
49			
62			

2. Quartiles

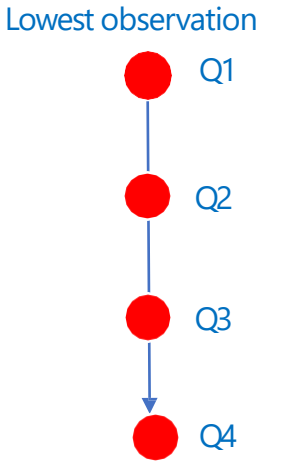
- Division of data into 4 segments according to the distribution of values
- The width of the four quartiles need not be the same
- Each part contains 25% data
- Quartiles are the highest values in each of the 4 parts
- Formula to calculate the quartiles:
  - $Q1 = [(n+1)/4]^{th}$ value, lower quartile
  - $Q2 = [(n+1)/2]^{nd}$ value, middle quartile
  - $Q3 = [3(n+1)/4]^{th}$ value, upper quartile

quartile	value	interpretation
1st quartile	18.75	25% values are <= 18.75
2nd quartile	27.5	50% values are <= 27.5
3rd quartile	35.25	75% values are <= 35.25
4th quartile	45	100% values are <= 45

Excel calculation

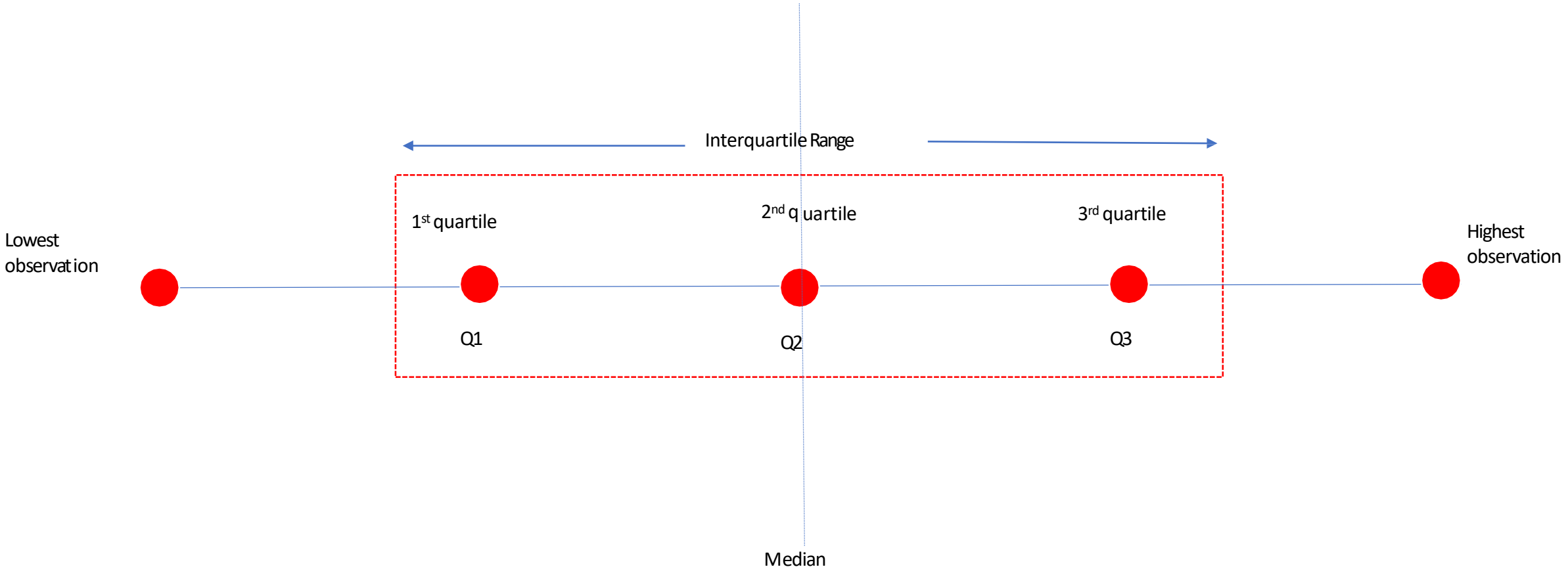
quartile	formula
1st quartile	= QUARTILES(<range>,1)
2nd quartile	= QUARTILES(<range>,2)
3rd quartile	= QUARTILES(<range>,3)
4th quartile	= QUARTILES(<range>,4)

	S.No	Data
Q1	1	10
	2	11
	3	14
	4	16
	5	17
	6	18
Q2	7	19
	8	21
	9	21
	10	23
	11	24
	12	26
Q3	13	29
	14	30
	15	32
	16	33
	17	34
	18	35
Q4	19	36
	20	37
	21	39
	22	40
	23	42
	24	45



3. Interquartile Range

- Approximately measures how far from the median on either side to include one-half of data
- IQR is the difference between the values of the first and third quartiles



4. Variance

- Average deviation from some measure of central tendency
- Every population / sample has variance
- Represented by the symbol σ^2
- Formula to calculate variance

  $\sigma^2 = (\sum(x - \mu)^2) / N$

  - σ^2 : population variance
  - x : observed value
  - μ : population mean
  - N : total number of items in population

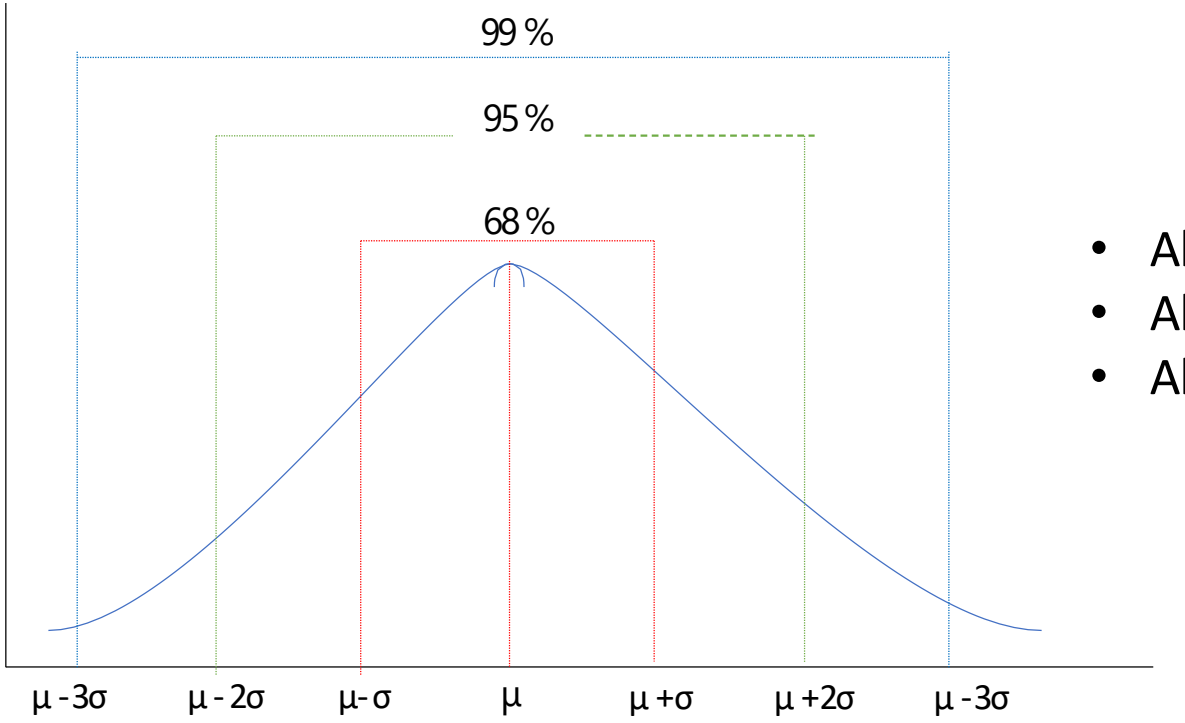
- Units of variance are *squares of units* of data – eg: squared miles, squared rupees etc.
- Not intuitively clear or interpreted in the right way

5. Standard Deviation

- Square root of the average of the squared distances of observation from the mean
- Represented by the symbol σ
- Formula to calculate Standard Deviation:

$$\sigma \quad \sqrt{\sigma^2} = \sqrt{(\sum(x - \mu)^2) / N}$$

- Units of SD are in the same units as that of the data
- SD enables to determine, with a high accuracy, the values of the frequency distribution in relation to the mean



- About ***68%*** data lies within ±1 SD from the mean
- About ***95%*** data lies within ±2 SD from the mean
- About ***99%*** data lies within ±3 SD from the mean

Difference between Standard Deviation and Variance

Standard deviation and variance are both measures of variability in a distribution, but they differ in a few ways:

- **Definition**

Variance is the average of the squared differences between each data point and the mean. Standard deviation is the square root of the variance.

- **Units**

Standard deviation is expressed in the same units as the original data, such as minutes or meters. Variance is expressed in larger units, such as meters squared.

- **Interpretation**

Standard deviation measures how far apart the numbers in a data set are. A small standard deviation means the data is tightly grouped around the mean, while a larger standard deviation means the data is more spread out. Variance gives a value to how much the numbers in a data set vary from the mean. A significant variance means the data points are far away from the mean.

- In practice, standard deviation is probably preferred over variance because it has the same units as the data. Variance is more often used in the background, such as in theory or deriving something else.

You’re interested in **calculating** the standard deviation of the exam scores of a national standardized test to see if many people scored close to the mean or not. Use the following dataset.

Test Taker	Score
1	20
2	40
3	60
4	60
5	75
6	80
7	70
8	65
9	70
10	90

$\sum x^2 - \bar{x}$

- In order to solve for the standard deviation, we have to follow the formula given earlier. Take a look at the **solution** below.

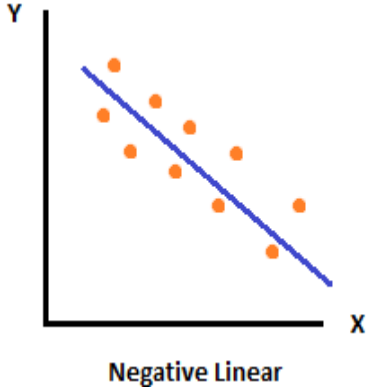
Test Taker	**Score**	$x - \bar{x}$	x^2
1	20	-43	1849
2	40	-23	529
3	60	-3	9
4	60	-3	9
5	75	12	144
6	80	17	289
7	70	7	49
8	65	2	4
9	70	7	49
10	90	27	729
$\bar{x}$	63	$\sum$	3660

$$sd = \sqrt{\frac{3660}{10-1}} = 20.2$$

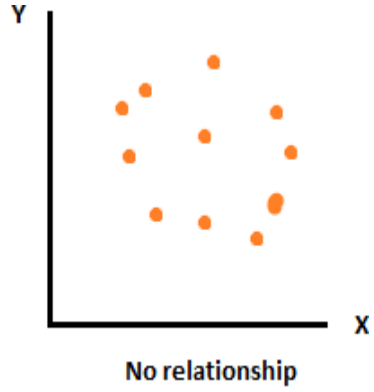
III. Measure of Association

Measures the relationship (degree and strength) between two variables thatare linearly related

Positive Linear



Negative Linear



No relationship

1. Covariance
2. Correlation
3. Coefficient of Variation

1. Covariance (+ -)

- Covariance is the joint variability of two random variables
- Measures the direction / sign of relationship only (+ or -) and not the strength
- How X and Y variables are linearly associated, working in tandem
  - ❑ Eg: Weight lifter training time vs Sprinter training time
    - Weight lifter trains more and lifts more weight (+)
    - Trainer trains more and runs in less time (-)
- Covariance measured as ***positive, negative*** or ***zero***
  - ❑ ***Positive:*** indicates direct or increase linear relationship
    - ➢ Xup - Yup
    - ➢ Xdown – Ydown
  - ❑ ***Negative:*** indicates indirect or decrease in linear relationship
    - ➢ Xup - Ydown
    - ➢ Xdown – Yup
- Covariance can be any number and not restricted to 0 and 1
- Formula
  - Sample CoV_{xy} $= \Sigma(x-x')(y-y') / n-1$
  - Population CoV_{xy} $= \Sigma(x-x')(y-\bar{y}) / n$

where
x and y are the 2 random variables
$\bar{x}$ and $\bar{y}$ are the means of the 2 random variables

2. Correlation (o)

- Measures the degree to which one variable is linearly related to the other
- 2 measures are used to describe correlation
  - ❑ Coefficient of Correlation (r)
    - $0 \leq r \leq -1$: Inverse relationship -> X-increases, Y-decreases
    - $0 \leq r \leq 1$: direct relationship -> X-increases, Y-increases
    - Measures the strength and direction
    - Formula (Karl Pearson's Coefficient of Correlation / Product moment) r = covariance of x and y / (SD of x)*(SD of y)

      $r = Cov(xy) / std(x).std(y)$

  - ❑ Coefficient of Determination (r^2)
    - $r^2 = r * r$
    - Measured in percentage
    - Eg: $r^2 = 0.83$ means 83% of variation in Y(dependent variable) accounted by X(independent variables)
- ***r*** does not mean anything, r^2 conveys the actual meaning

3. Coefficient of Variation

- Relative Standard Deviation
- Measured in %
- Shows variations with relation to the mean
- Does not have any units
- Smaller CoV is better represents better quality
- Formula

  $CoV = \sigma / \mu$

Example:

Last 15 days, trading of 2 stocks are as follows:

Stock A	Stock B
Average price: 135	Average price: 87.5
SD: 15.35	SD: 1.02

Which is more risky?

$Cov_A = 15.35/115 = 0.133$

$Cov_B = 1.02/87.5 = 0.011$

Stock A is more risky

Sum of Squares

Total sum of squares is used to denote the amount of variation in the dependent variable.

Mathematically, the difference between variance and SST is that we adjust for the degree of freedom by dividing by n–1 in the variance formula.

$$SST=\sum(yi-\bar{y}) \exp 2$$

Where:

- yi – observed dependent variable
- $\bar{y}$ – mean of the dependent variable

The SST tells us how close sample values are to the mean. As the SST increases, so does the variability of the data.

Example of Calculating the SST for a Sample with Low Variability

Calculate the SST for the following data:

{1, 2, 3}

Step 1: The mean of the sample can be calculated by adding up the values in the sample (1 + 2 + 3) and dividing this sum by the number of values (3). Thus, the mean of this sample is:

$$\bar{y} = (1+2+3)/3$$
$$= 6/3$$
$$= 2$$

Step 2: Subtract the calculated mean from each value, and square each difference.

$1-2 = -1$	$(-1) = 1$	$= 1$
$2-2 = 0$	$-0 = 0$	$= 0$
$3-2 = 1$	$-1 = 1$	$= 1$

Step 3: Sum the differences.

$$SST = 1+0+1 = 2$$

Thus, **the total sum of squares for the data {1, 2, 3} is 2.**

What is the Absolute Deviation Formula?

$M = \Sigma (x_i - \bar{x})/n$

where,

M is the average absolute deviation,

$\bar{x}$ is the mean of data set,

$\Sigma (x_i - \bar{x})$ is the summation of deviations from mean,

n is the number of values in data set.

What Is Average Deviation Formula?

The formula for average deviation is utilized to determine how much individual observations differ from the mean of a data set. Presented below is the formula for computing the average deviation across n observations:

$$\text{Average Deviation} = \Sigma|x_i - \bar{x}|/n$$

where x_i are the data points,

$\bar{x}$ is the mean, and

n is the number of data points.

Question : Find the Average Deviation for the data 10,25,30,14,39,18,17. (Use median to find central point)

Step 1: *Find median for the given data.*

To find median first we need to sort the given data either in ascending order or descending order.

Sorted data- 10,14,17,18,25,30,39

Here the size of data set is odd i.e., count=7.

So we have only one middle value18 which is median.

Step 2: *Find absolute deviations from data using median.*

abs(10-18) = 8

abs(14-18) = 4

abs(17-18) = 1

abs(18-18) = 0

abs(25-18) = 7

abs(30-18) = 12

abs(39-18) = 21

Step 3: *Sum of all deviations = 8+4+1+0+7+12+21*

=53

Step 4: *Find Average Deviation=sum of all deviations/count of values in data*

=53/7

=> 7.57

So Average Deviation within the given data is 7.57

Question : Find the Average Deviation for the data 10,20,30,40,50 (Use mean/median to find central point)

Step 1: *Find the center point for the given data.*

As data is already in sorted order it is preferred to use the median to find the central point.

Here the size of the data set is odd i.e., count=5.

So we have only one middle value 30 which is the median.

Step 2: *Find absolute deviations from data using the median.*

abs(10-30)=20

abs(20-30)=10

abs(30-30)=0

abs(40-30)=10

abs(50-30)=20

Step 3: *Sum of all deviations=20+10+0+10+20 =60*

Step 4: *Find Average Deviation=sum of all deviations/count of values in data*

=60/5

=>12

So Average Deviation within the given data is 12

Problem 1. Calculate the average absolute deviation of the data set, 2, 6, 7, 4, 1.

Solution:

The data set is 2, 6, 7, 4, 1.

Here, n = 5.

Mean of the data, $\bar{x} = (2 + 6 + 7 + 4 + 1)/5$

$= 20/5$

$= 4$

Using the formula we get,

$M = \Sigma (x_i - \bar{x})/n$

$= [|4 - 2| + |4 - 6| + |4 - 7| + |4 - 4| + |4 - 1|]/5$

$= (2 + 2 + 3 + 0 + 3)/5$

$= 10/5$

$= 2$