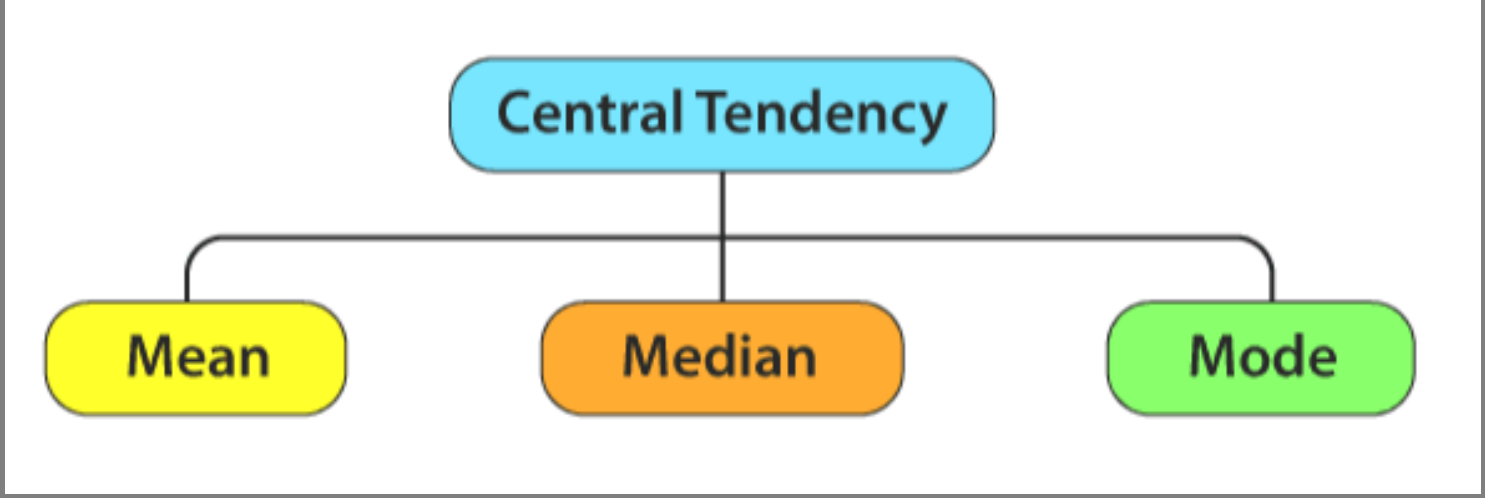
CSE303

Lecture 2: Measures of Central Tendency and Dispersion

CENTRAL TENDENCY

- The central tendency is stated as the statistical measure that represents the single value of the entire distribution or a dataset.
- These three values summarize the dataset using a single value.



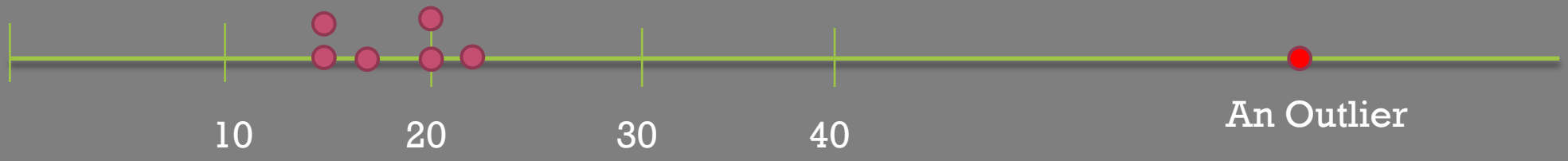
MEAN

- The sum of all values divided by the number of values.

$$\text{Mean} = \bar{x} = \frac{\sum_i^n x_i}{n}$$

- Mean is influenced by extreme values (Outliers).
- An outlier is a value or an element of a dataset that shows higher deviation from the rest of the values.

EXAMPLE – AN OUTLIER



TRIMMED MEAN

- The average of all values after dropping a fixed number of extreme values from both ends.

$$\text{Trimmed mean} = \bar{x} = \frac{\sum_{i=p+1}^{n-p} x_{(i)}}{n - 2p}$$

- Preferable to use instead of ordinary mean as it can negate the effect of extreme values (Outliers).

MEDIAN

- The median is the middle number on a sorted list of the data.
- If there is an even number of data values, the middle value is one that is not actually in the data set, but rather the average of the two values that divide the sorted data into upper and lower halves.

- If n is odd, $\frac{n+1}{2}$ th value → 10, 15, 18, 22, 25, 28, 33, 43, 50
- If n is even, $\frac{\frac{n}{2}\text{ th value} + (\frac{n}{2}+1)\text{ th value}}{2}$ → 10, 15, 18, 22, 25, 28, 33, 43, 50, 55 → (25+28)/2 = 26.5
- Not influenced by the extreme values (Outliers).

PRACTICE PROBLEM - 1

- A group of 10 students appeared in a test. Their obtained marks are given below.

60, 70, 80, 75, 65, 70, 80, 70, 65, 65

Sorted Order

60 65 65 65 70 70 70 75 80 80

Median = 70

Mean = 70

PRACTICE PROBLEM - 2

- A group of 10 students appeared in a test. Their obtained marks are given below.

60, 70, 80, 75, 65, 70, 100, 70, 65, 65

Mean = 720/10 = 72
Without considering the 100, mean = 620/9 = 68.89

For Trimmed mean,
Sorting the dataset: ~~60~~ 65 65 65 70 70 70 75 80 ~~100~~
For p = 1 → 60 and 100 will be discarded.
Then trimmed mean = 70

Median = 70

PRACTICE PROBLEM - 3

- A group of 10 students appeared in a test. Their obtained marks are given below.

60, 70, 80, 75, 65, 70, 10, 70, 65, 65

Mean = 630/10 = 63
Without considering the 10, mean = 620/9 = 68.89

For Trimmed mean,
Sorting the dataset: ~~10~~ 60 65 65 65 70 70 70 75 ~~80~~
For p = 1 → 10 and 80 will be discarded.
Then trimmed mean = 67.5

Median = 67.5

WEIGHTED MEAN

- It is calculated by multiplying each data value x_i by a weight w_i and dividing their sum by the sum of the weights (w_i).

$$\text{Weighted mean} = \bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i}^{n} w_i}$$

- Some values are intrinsically more variable than others, and highly variable observations are given a lower weight.

PRACTICE PROBLEM - 4

- In a software project, there could be several risks and these risks can be categorized according to their level of severity. Calculate the expected value of the Damage.
- $E(X) = \sum pixi$

Risk ID	Severity Level	Damage
1	Extreme (5)	250000
2	Significant (4)	150000
3	Moderate (3)	100000
4	Less (2)	50000
5	Insignificant (1)	10000

(5 * 250000 + 4 * 150000 + 3 * 100000 + 2 * 50000 + 1 * 10000) / (5+4+3+2+1) = 150666.6667

WEIGHTED MEDIAN

- Instead of the middle number, the weighted median is a value such that the sum of the weights is equal for the lower and upper halves of the sorted list.

- X = 15 17 20 22 24
- 17 is the maximum value of the lower half
- 20 is the minimum value of the upper half
- Weighted Median = (17+20) / 2 = 18.5
- Median = 20

MODE

- The value that occurs most frequently.
- Not that much useful.

DISPERSION

- In statistics, dispersion (also called variability, scatter, or spread) is the extent to which a distribution is stretched or squeezed. Common examples of measures of statistical dispersion are the **variance, standard deviation,** and **interquartile range.**

- Dataset 1 → 30 40 50 60 70 → Mean = 50, Median = 50, Variance = lower
- Dataset 2 → 0 25 50 75 100 → Mean = 50, Median = 50, Variance = higher

MEAN ABSOLUTE DEVIATION

- The mean of the absolute value of the deviations from the mean.

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

- Dataset 1 → 30 40 50 60 70 → Mean = 50, Median = 50, Mean Absolute Deviation = 12
- Dataset 2 → 0 25 50 75 100 → Mean = 50, Median = 50, Mean Absolute Deviation = 30

(SAMPLE) VARIANCE

- The sum of squared deviations from the mean divided by n – 1 where n is the number of data values.
- Average of the squared deviations.

$$\text{Variance} = s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

- **Why the denominator is n-1 instead of n?** To obtain the true estimate of the variance with regard to the population, it is divided by n-1 so that the estimated value would be little larger. This value it is known as the true estimate of the variance.

PRACTICE EXAMPLE

- Dataset 1 → 30 40 50 60 70 → Mean = 50, Median = 50, Variance = lower
- Dataset 2 → 0 25 50 75 100 → Mean = 50, Median = 50, Variance = higher
- Dataset 1
  - $(30-50)^2 + (40-50)^2 + (50-50)^2 + (60-50)^2 + (70-50)^2 / 5-1 = 250$
  - Standard Deviation = $\sqrt{250} = 15.8113$
- Dataset 2
  - $(0-50)^2 + (25-50)^2 + (50-50)^2 + (75-50)^2 + (100-50)^2 / 5-1 = 1562.5$
  - Standard Deviation = $\sqrt{1562.5} = 39.5284$

STANDARD DEVIATION

- The square root of the variance.

$$\text{Standard deviation} = s = \sqrt{\text{Variance}}$$

- Standard deviation is preferred over the mean absolute deviation.

MEDIAN OF MEDIAN ABOSULTE DEVIATION

- The median of the absolute value of the deviations from the median.

$$\text{Median absolute deviation} = \text{Median}\left(|x_1 - m|, |x_2 - m|, \ldots, |x_N - m|\right)$$

- Not influenced by the extreme values (Outliers).
- Dataset 1 → 30 40 50 60 70 → Median diff (20, 10, 0, 10, 20) → Sort (0, 10, 10, 20, 20) = 10
- Dataset 2 → 0 25 50 75 100 →Median diff (50, 25 , 0, 25, 50) → Sort (0, 25, 25, 50, 50) = 25

SUMMARY OF DISPERSION MEASURES

	Mean	Median	Mean Abs. Dev.	Variance	Standard Dev.	Median of Median Abs. Dev
Dataset 1	50	50	12	250	15.8113	10
Dataset 2	50	50	30	1562.5	39.5284	25

PRACTICE EXAMPLE

- **Find the mean, median, mode, range, variance, standard deviation, mean absolute deviation and median of the median absolute deviation for the following list of values:**

8, 9, 10, 10, 10, 11, 11, 11, 12, 13

→ **Mean = 10.5**

→ **Median = 10.5**

→ **Mode = 10, 11 (Bi-modal dataset)**

→ **Range = 13-8 = 5**

→ **Variance = 2.055556**

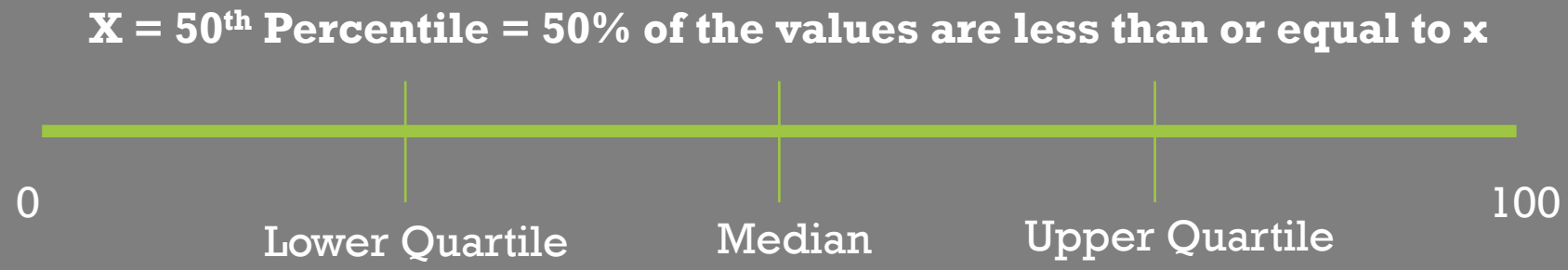
→ **Standard Deviation = 1.4337**

→ **Mean Absolute Deviation = 1.1**

→ **Median of the Median Absolute Deviation = 0.5**

PERCENTILE

- The value such that P percent of the values take on this value or less and (100–P) percent take on this value or more.
- The 25-th percentile is the lower quartile (Q1).
- The 50-th percentile is the median.
- The 75-th percentile is the upper quartile (Q3).
- Quantiles are the same as percentiles but are indexed by sample fractions rather than by sample percentage.
- 80^{th} Percentile = 0.8 Quantile



FINDING PERCENTILE

8, 9, 10, 10, 10, 11, 11, 11, 12, 13

- Total data points, n = 10
- 50^{th} percentile = 50% X 10 = 5^{th} value = 10
- Median = (10+11)/2 = 10.5

8, 9, 10, 10, 10, 11, 11, 11, 12, 13, 15, 16, 16, 17, 20

- Total data points, n = 15
- 50^{th} percentile = 50% X 15 = 7.5^{th} value = 11
- Median = 8^{th} value = 11
- 20^{th} percentile = 20% X 15 = 3^{rd} value = 10
- 75^{th} percentile = 75% X 15 = 11.25^{th} value = 15 + (16-15)X0.25 = 15.25
- 95^{th} percentile = 95% X 15 = 14.25^{th} value = 17 + (20-17)X0.25 = 17.75

INTERQUARTILE RANGE

- The difference between the 75th percentile and the 25th percentile.
- IQR = Q3 – Q1
- Range = max - min

ESTIMATING MEAN, MEDIAN FOR A GROUPED DATASET

Class Interval

1. Exclusive Interval
2. Inclusive Interval

Seconds	Frequency
51 - 55	2
56 - 60	7
61 - 65	8
66 - 70	4

$$\text{Estimated Median} = L + \frac{(n/2) - B}{G} \times w$$

EXAMPLE: EXCLUSIVE INTERVAL

85, 125, 210, 180, 160, 155, 135, 100 .....

Runs Scored in First Innings	Frequency
80-100	3
100-120	4
120-140	7
140-160	12
160-180	3
180-200	3
200-220	2

In case of Exclusive Interval, upper bound is excluded.

CONVERTING INCLUSIVE TO EXCLUSIVE

Seconds	Frequency
51 - 55	2
56 - 60	7
61 - 65	8
66 - 70	4

Seconds	Frequency
50.5 – 55.5	2
55.5 – 60.5	7
60.5 – 65.5	8
65.5 – 70.5	4

FINDING MEAN, MEDIAN, VARIANCE FOR A GROUPED DATASET

Seconds	Frequency (f)	Mid point (x)	Fi * Xi	Fi * $(Xi - X_{bar})^2$	CFi
51-55	2	53	106	138.8877	2
56-60	7	58	406	77.7762	9
61-65	8	63	504	22.2231	17
66-70	4	68	272	177.7795	21

Mean = ∑ (Fi * Xi) / n = 1288 / 21 = **61.3333**

Median = 61 + ((10.5-9)/8) * 5
= 61.9375

Variance = ∑ (Fi * $(Xi - X_{bar})^2$) / n-1
= 416.6665/20
= 20.8333

$$\text{Estimated Median} = L + \frac{(n/2) - B}{G} \times w$$

L = lower bound of the median class = 61
n/2 = 21/2 = 10.5
B = CFi of the class immediately before the Median class = 9
G = Fi of the median class = 8
W = width of the class = (max-min)+1 = 5

HOMEWORK

- Collect data from the given link.
  - https://www.espncricinfo.com/series/bangladesh-premier-league-2021-22-1296684/match-results
- Make the dataset like the table given below.

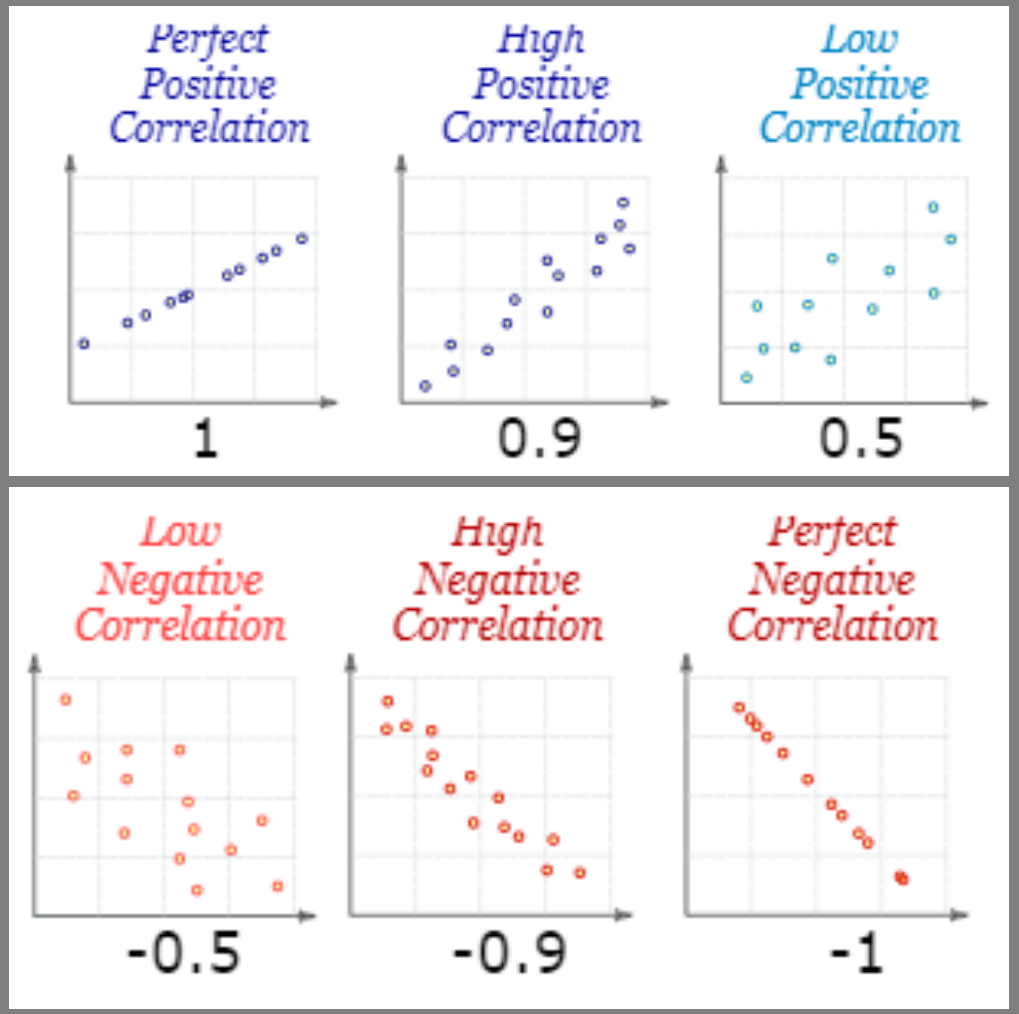
Match_id	First_innings_score	ground
1	125	Mirpur
2		
3		

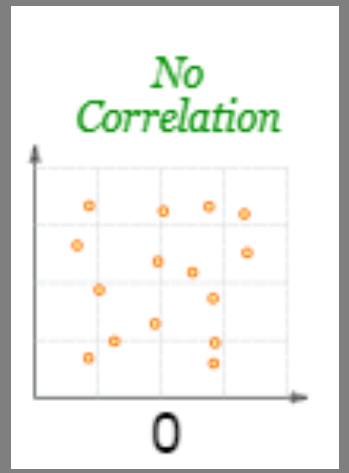
- Prepare a grouped dataset based on the above mentioned table for estimating mean, median and variance of the first innings score.
- The class interval must be within 10 to 20.
- You can use inclusive/exclusive interval.
- Also estimate mean, median and variance of the first innings score for each ground and then compare them

CORRELATION

- Correlation is a statistic that measures the degree to which two variables move in relation to each other.
- Correlation is Positive when the values increase together.
- Correlation is Negative when one value decreases as the other increases.

POSITIVE AND NEGATIVE CORRELATION





PEARSON'S CORRELATION COEFFICIENT

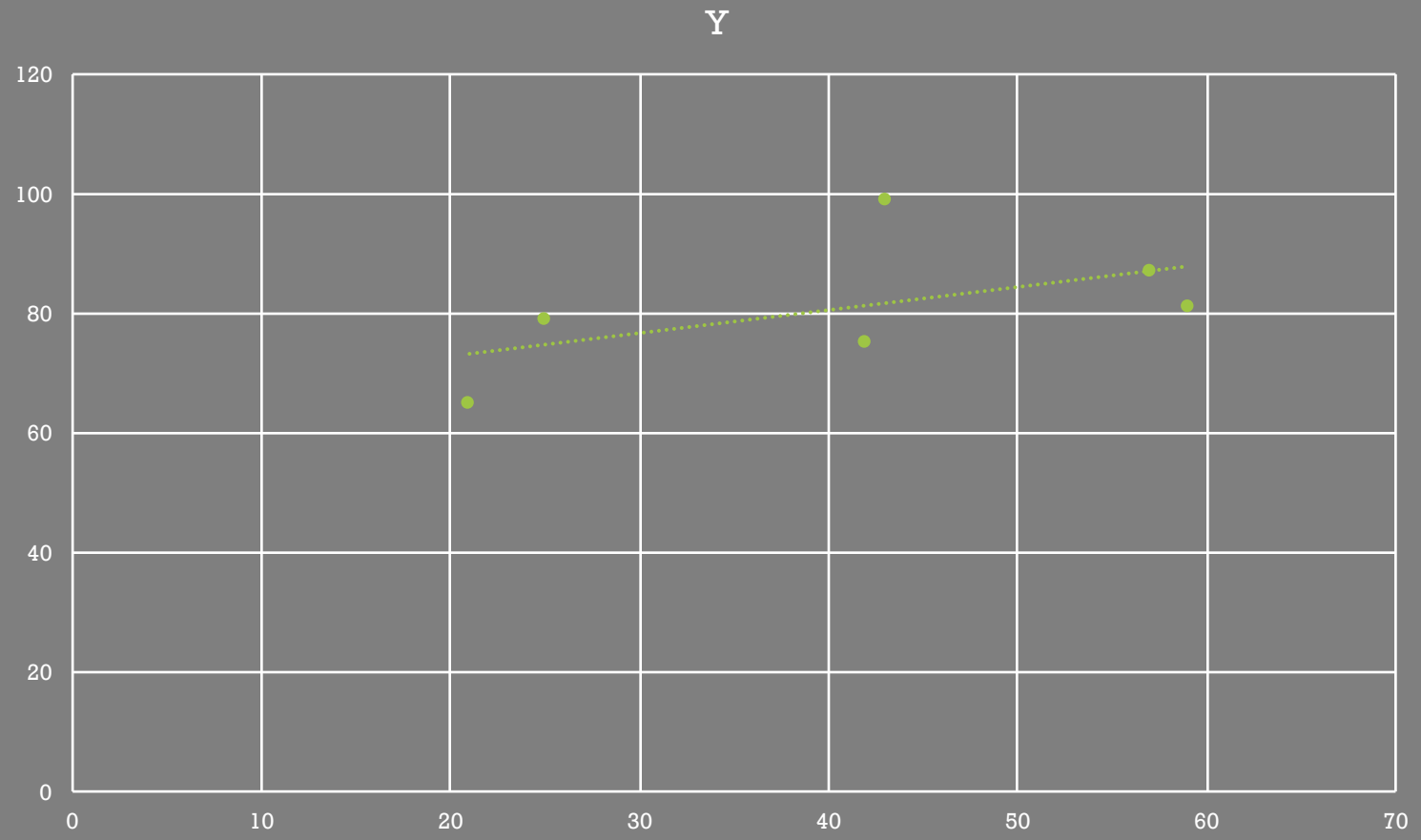
$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{(N-1)s_x s_y}$$

EXAMPLE

Example question: Find the value of the correlation coefficient from the following table:

SUBJECT	AGE X	GLUCOSE LEVEL Y
1	43	99
2	21	65
3	25	79
4	42	75
5	57	87
6	59	81

SCATTER PLOT ON DATASET



FINDING THE VALUE OF R

X	Y	Xi – Mean of X	Yi – Mean of Y	$(Xi - Xbar)(Yi - Ybar)$	(Xi – Mean of X)²	(Yi – Mean of Y)²
43	99	1.833333	18	33	3.361111	324
21	65	-20.1667	-16	322.6667	406.6944	256
25	79	-16.1667	-2	32.33333	261.3611	4
42	75	0.833333	-6	-5	0.694444	36
57	87	15.83333	6	95	250.6944	36
59	81	17.83333	0	0	318.0278	0

Mean of X = 41.1667
Mean of Y = 81

$$\sum (Xi - Xbar)(Yi - Ybar) = 478$$

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{(N-1)s_x s_y}$$

$\sum (Xi - Xbar)^2 = 1240.833$
Variance of X = 248.1667
Variance of Y = 131.2

$\sum (Yi - Ybar)^2 = 656$
Standard Deviation of X = 15.75331
Standard Deviation of Y = 11.45426

= 478/(5*15.75331*11.45426)

= 0.529809

COVARIANCE

$$Cov(A, B) = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n}.$$

$$r_{A,B} = \frac{Cov(A, B)}{\sigma_A \sigma_B},$$

USEFUL RESOURCES

- Chapter 1, Practical Statistics for Data Scientists by Bruce and Bruce
- https://www.geeksforgeeks.org/python-pandas-dataframe/
- https://www.tutorialspoint.com/python_pandas/python_pandas_descriptive_statistics.htm
- https://medium.com/swlh/statistical-functions-of-pandas-2862c290053a
- https://pandas.pydata.org/pandas-docs/stable/reference/frame.html

THANK YOU