SWAYAM NPTEL COURSE ON

MINE AUTOMATION AND DATA ANALYTICS

By

Prof. Radhakanta Koner

Department of Mining Engineering

Indian Institute of Technology (Indian School of Mines) Dhanbad

Module 07: Descriptive Statistics

Lecture 17 B : Introduction - II

CONCEPTS COVERED

1. Compute and Interpret numerical Summaries of Data

- Compute and Interpret measures of dispersion: Variance, Standard Deviation
- Compute and Interpret percentiles and Interquartile Range (IQR)
- Compute and Interpret five number summary

2. Association between two variables

- Understanding the association between numerical variables through a scatter plot
- Compute and interpret Covariance and Correlation.

Variance

- In contrast to the range, the variance considers all the observations.
- One way of measuring the variability of a data set is to consider the deviations of the data values from a central value.

Population Variance and Sample Variance

- When we refer to a dataset from a population, we assume the dataset has N observations. In contrast, when referring to a data set from a sample, we assume the data set has n observations.
- The variation is computed using the following formulae

Population Variance $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$

Sample Variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Units of Variance

- **The sample variance is expressed in units of square units of the original variable.**

Example

	Data	Deviation from Mean $(x_i - \bar{x})$	Squared Deviation $(x_i - \bar{x})^2$
1	9	3.2	10.24
2	5	-0.8	0.64
3	8	2.2	4.84
4	3	-2.8	7.84
5	4	- 1.8	3.24
Total	29	0	26.8

Sample Variance = $\frac{26.8}{4}$ = 6.7

Population Variance = $\frac{26.8}{5}$ = 5.36

Adding/Multiplying a constant

- Let $y_i = x_i + c$ where c is a constant then $new\ variance = old\ variance$
- *In general, adding a constant does not change dataset variability. Hence, it is the same.*
- Let $y_i = x_i * c$ where c is a constant then $new\ variance = c^2 * old\ variance$

Standard Deviation

Another handy measure of dispersion is the standard deviation.

Definition

The quantity, the square root of the sample variance, is the sample standard deviation.

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Example

	Data	Deviation from Mean $(x_i - \bar{x})$	Squared Deviation $(x_i - \bar{x})^2$
1	9	3.2	10.24
2	5	-0.8	0.64
3	8	2.2	4.84
4	3	-2.8	7.84
5	4	- 1.8	3.24
Total	29	0	26.8

Sample Variance = $\frac{26.8}{4}$ = 6.7

Sample Standard Deviation = $\sqrt{6.7}$ = 2.58

Population Variance = $\frac{26.8}{5}$ = 5.36

Population Standard Deviation = $\sqrt{5.36}$ = 2.31

Units of Standard Deviation

- **The sample standard deviation is measured in the same units as the original data.**

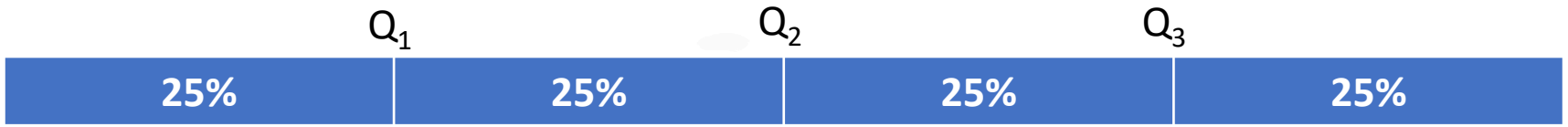
Adding/Multiplying a constant

- Let $y_i = x_i + c$ where c is a constant then $new\ standard\ deviation = old\ standard\ deviation$
- *In general, adding a constant does not change dataset variability. Hence, it is the same.*
- Let $y_i = x_i * c$ where c is a constant then $new\ standard\ deviation = c * standard\ deviation$

Quartiles

Definition

The sample 25th percentile is called the first quartile. The sample 50th percentile is called the median or the second quartile. The sample 75th percentile is called the third quartile.



In other words, the quartiles break a data set into four parts, with about 25 percent of the data values being less than the first (lower) quartiles, about 25 percent being between the first and second quartiles, about 25 percent being between the second and third (upper) quartiles, and about 25 percent being larger than the third quartile

Interquartile Range (IQR)

Definition

The interquartile range (IQR) is the difference between the first and third quartiles.

$$IQR = Q_3 - Q_1$$

The Five Number Summary

i) Minimum

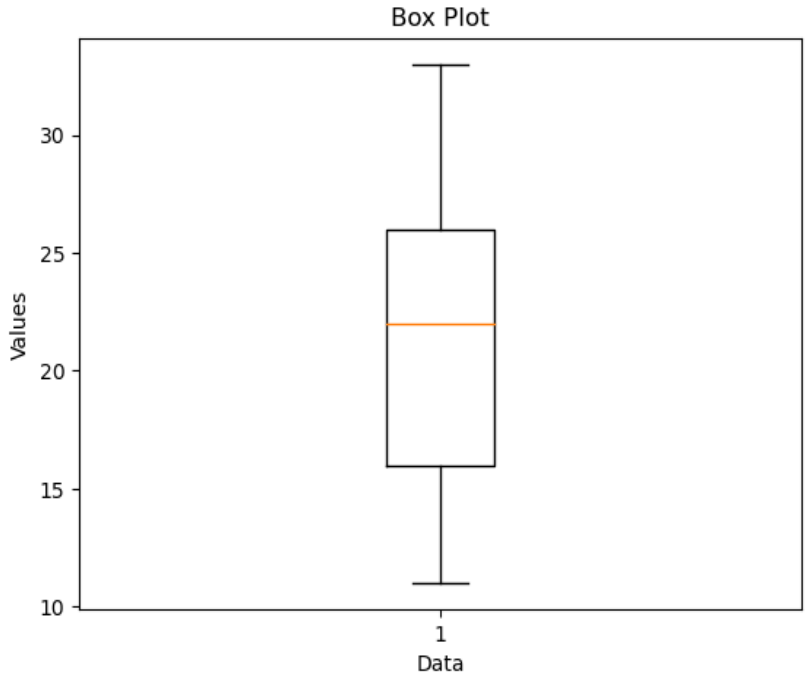
ii) Q_1: First Quartile or Lower Quartile

iii) Q_2: Second Quartile of Median

iv) Q_3: Third Quartile or upper quartile

v) Maximum

Data = [11, 22, 15, 29, 33, 15, 17, 22, 19, 25, 27]

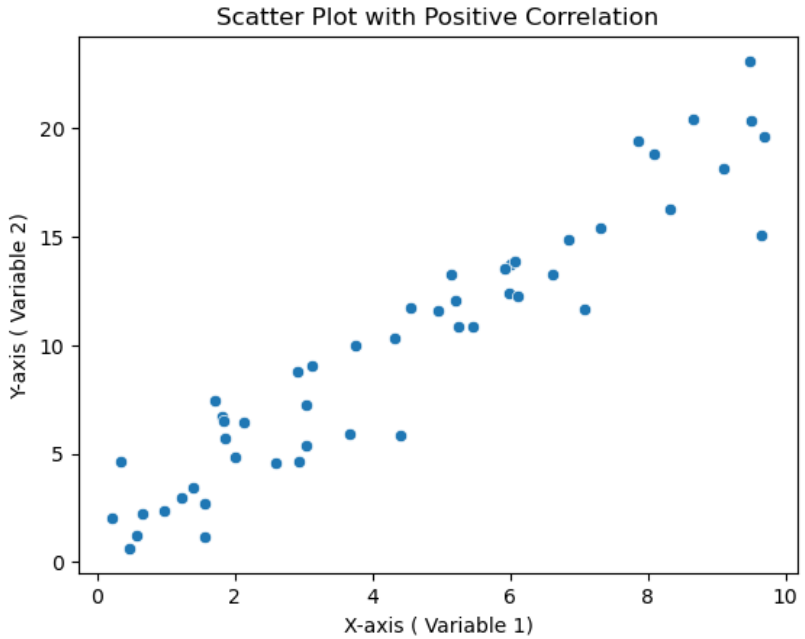


Association

The pattern of data in one variable occurs in a particular manner related to the pattern of data in one or several other variables.

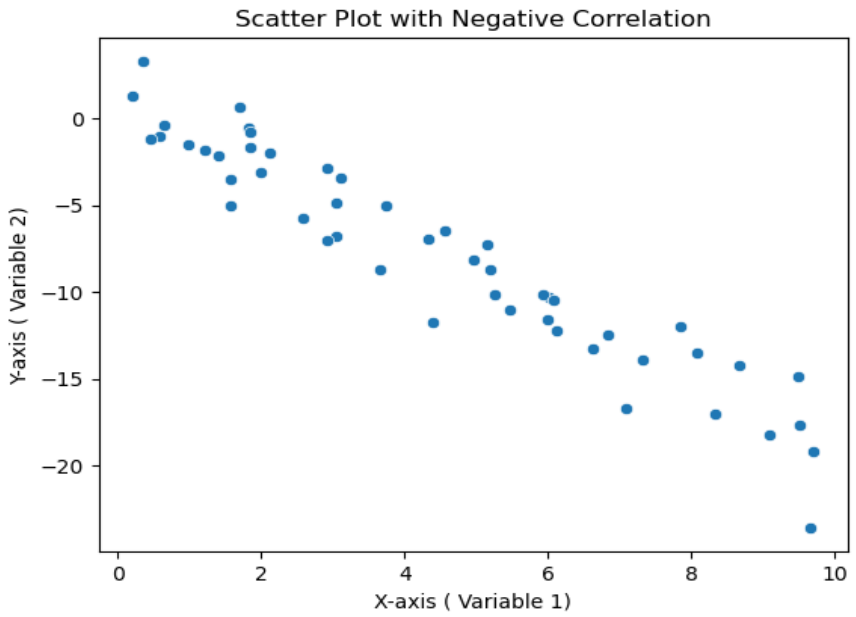
Scatter Plot

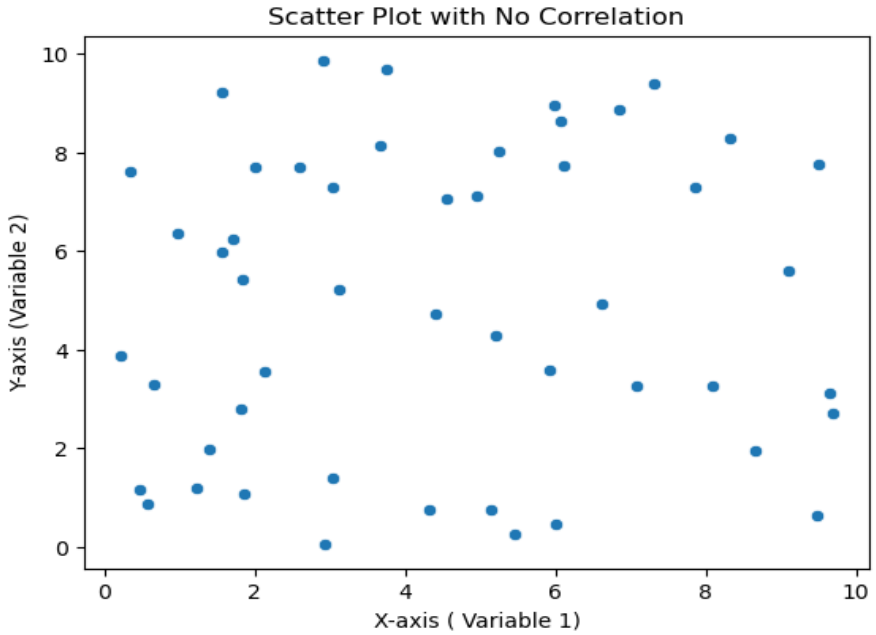
- A scatter plot is a graphical representation of the relationship between two numerical variables.
- It allows you to visually inspect the pattern of the data points and understand the association or correlation between the variables.



Positive Relationship:
If the points on the scatter plot generally form an upward-sloping pattern from left to right, it indicates a positive correlation. This means that as one variable increases, the other also tends to increase.

Scatter Plot





Negative Relationship:
Conversely, if the points on the scatter plot form a downward-sloping pattern from left to right, it indicates a negative correlation. This means that as one variable increases, the other tends to decrease.

No Relationship:
If the points appear randomly scattered without any clear pattern, it suggests no solid linear relationship between the variables. However, other types of relationships might still exist, such as non-linear or complex associations.

Measures of association

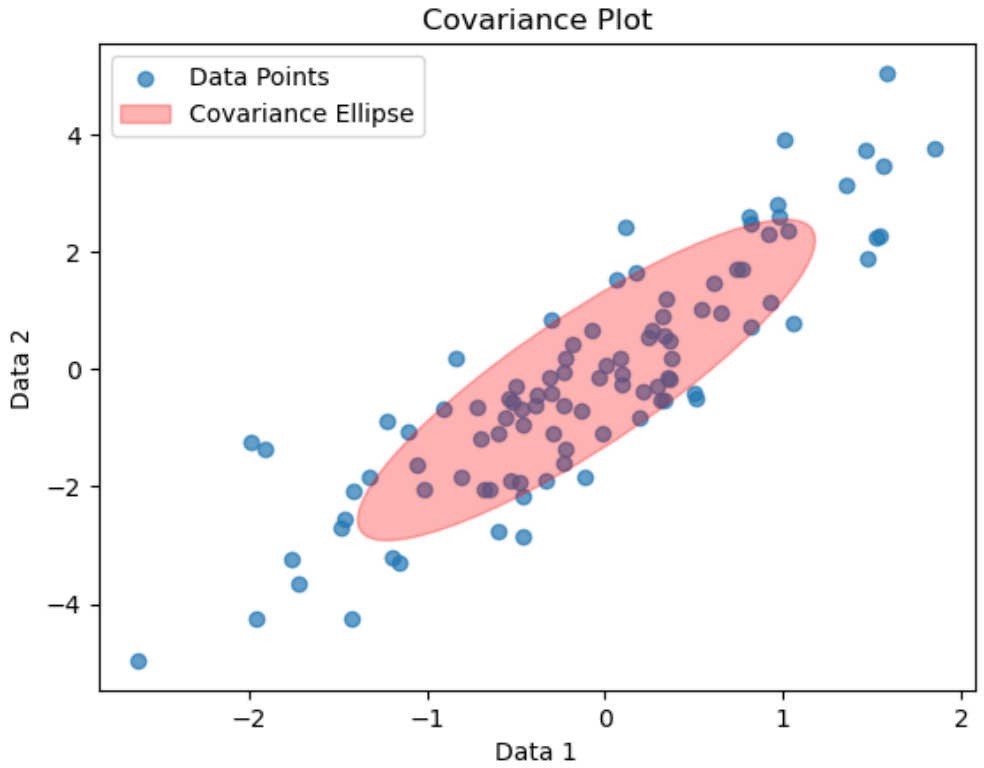
How do we measure the strength of association between two variables i.e. from earlier plots ?

1. Covariance

2. Correlation

Covariance

- **Covariance quantifies the strength of the linear association between two numerical variables.**



Covariance

- **Definition**
- ***Let xi denote the i^{th} observation of variable x and yi is the i^{th} observation of variable y. Let (x_i, y_i) be the i^{th} paired observation of a population (sample) dataset having N(n) observations. The covariance between the variables x and y is given by***

Population Variance : $Cov(x, y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{N}$

Sample Variance : $Cov(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$

Example

- Recall , the association between variable x and variable y of a person

x	y	Deviation of x $(x_i - \bar{x})$	Deviation of y $(y_i - \bar{y})$
2	5	2 – 17.67	5 – 25.5
8	12	8 – 17.67	8 – 25.5
18	18	18 – 17.67	18 – 25.5
20	23	20 – 17.67	20 – 25.5
28	45	28 – 17.67	28 – 25.5
30	50	30 – 17.67	30 – 25.5

Example

x	y	Deviation of x $(x_i - \bar{x})$	Deviation of y $(y_i - \bar{y})$
2	5	2 – 17.67	5 – 25.5
8	12	8 – 17.67	12 – 25.5
18	18	18 – 17.67	18 – 25.5
20	23	20 – 17.67	23 – 25.5
28	45	28 – 17.67	45 – 25.5
30	50	30 – 17.67	50 – 25.5

Number of observations = 6
Mean of X = 17.67
Mean of Y = 25.5

Cov (X, Y) =
(⅙) [(2 – 17.67)(5 – 25.5) + (8 – 17.67)(12 – 25.5) + (18 – 17.67)(18 – 25.5) + (20 – 17.67)(23 – 25.5) + (28 – 17.67)(45 – 25.5) + (30 – 17.67)(50 – 25.5)]
= 157.83

Units of Covariance

- The size of the covariance, however, is difficult to interpret because the covariance has units.
- The units of the covariance are those of the x variable times of the y variable.

Correlation

- A more easily interpreted measure of linear association between two numerical variables in correlation.
- It is derived from covariance
- To find the correlation between two numerical variables, x and y, divide the covariance between x and y by the product of the standard deviations of x and y.
- The Pearson correlation coefficient (r) between x and y is given by

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\ \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{Cov\,(x\,,y)}{s_x s_y}$$

Units of Correlation

- The units of the standard deviation cancel out the units of covariance
- It can be shown that the correlation measure always lies between -1 and +1

Example

- Recall , the association between x variable and y variable

x	y	Deviation of x	Deviation of y
20	60	20 – 37.5	60 – 67.5
25	60	25 – 37.5	60 – 67.5
30	70	30 – 37.5	70 – 67.5
40	73	40 – 37.5	73 – 67.5
50	67	50 – 37.5	67 – 67.5
60	75	60 – 37.5	75 – 67.5

Example

- Recall , the association between x variable and y variable

x	y	Deviation of x $(x_i - \bar{x})$	$(x_i - \bar{x})^2$	Deviation of y $(y_i - \bar{y})$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
20	60	-17.5	306.25	-7.5	56.25	131.25
25	60	-12.5	156.25	-7.5	56.25	93.5
30	70	-7.5	56.25	2.5	6.25	-18.75
40	73	-2.5	6.25	5.5	30.25	13.75
50	67	12.5	156.25	-0.5	0.25	-6.25
60	75	22.5	506.25	7.5	56.25	168.75
			= 1187.5		= 205.5	382.5

$$Cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$Cov(x,y) = \frac{382.5}{5}$ = 76.5

Variance of x = $\frac{1187.5}{5}$ = 237.5 , $S_x = \sqrt{237.5}$ = 15.41

Variance of y = $\frac{205.5}{5}$ = 41.1 , $S_y = \sqrt{41.1}$ = 6.41

$r = \frac{Cov(x,y)}{s_x s_y} = \frac{76.5}{15.41 * 6.41}$ = 0.774

REFERENCES

- Introduction to Probability and Statistics for Engineers and Scientists, Sixth Edition, Sheldon M. Ross
- Statistical Methods Combined Edition (Volume I& II), N G Das

CONCLUSION

1. Numerical Summaries
   - ❖ Variance
   - ❖ Standard Deviation
   - ❖ Interquartile Range (IQR)
   - ❖ Interpret five number summary using Box Plot
2. Association between two variables
   - ❖ Scatter plot
   - ❖ Covariance
   - ❖ Correlation


THANK YOU