17-July-2025
Thursday

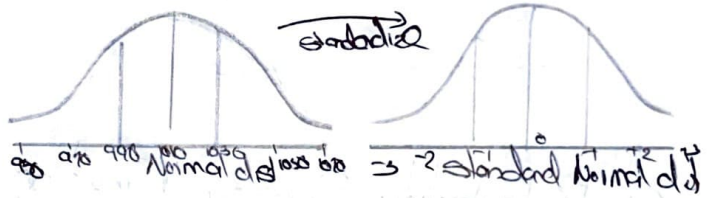
- Standard Normal Distribution

The standard normal distribution is a special case of normal distribution. It is the distribution that occurs when a normal random variable has a mean of zero and a SD of one

- Z-score / Z-value / Standard score

A z-score (aka, a standard score) indicates how many SD an element is above or below from the mean

$$Z = \frac{(X - \mu)}{\sigma}$$



- Calculating Z-score

$$\text{Z-Score} = \frac{x_i - \bar{x}}{s}$$

x_i = data point
$\bar{x}$ = mean
s = standard deviation

Example, $Z = \frac{231 - 130.1}{47.85} = 2.11$

Another Example,

Que. Score on a exam are normally dist with a mean of 65 and a SD of 9. Find the parent of scores a) less than 54

$\mu = 65$
$\sigma = 9$
$x = 54$

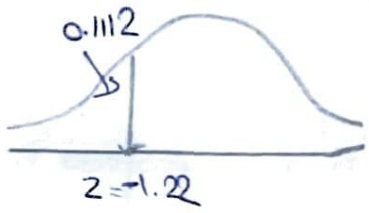
Ans. $$Z = \frac{x - \mu}{\sigma} = \frac{54 - 65}{9} = Z = -1.2222$$

Now look the z-score test the and find

Row = -1.2
Column = 0.02

∴ now we got 0.02
i.e. row 1.2 | 2 2 2 ~ column

$\mu = 65$ $\sigma = 9$



$P(x<54)$

$= P(z<-1.22)$

$= 0.1112$

This mean 11.12%

Therfore the probability that $x<54$ is 11.12%

Example:

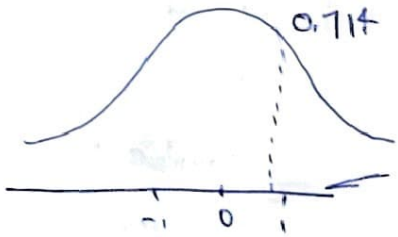
$x = 4$

$\mu = 3$

$\sigma = 1.4$

$$Z = \frac{4-3}{1.4} = 0.714$$

Z-score = 76.11%



⊙ Probability Density Function

Probability density is the relationship between observation and their Probability

⊙ Measure of Distance

- Euclidean Distance
- Manhattan Distance
- Minkowski distance.

- Euclidean Distance

It is a classical method to calculate the distance between two object X and Y in the Euclidean space (1- or 2- or n-dimension space). This distance can be calculated by traveling along the line connecting the points

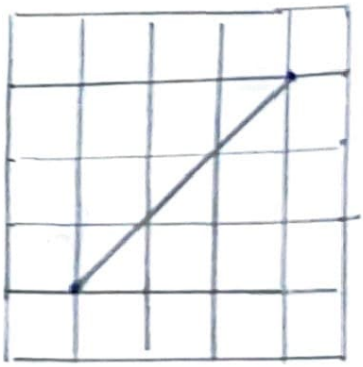
$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

→ Pythagorean Theorem to compute the distance.

$$= \sqrt{(6-4)^2 + (6-2)^2}$$

$$= \sqrt{(4)^2 + (4)^2}$$

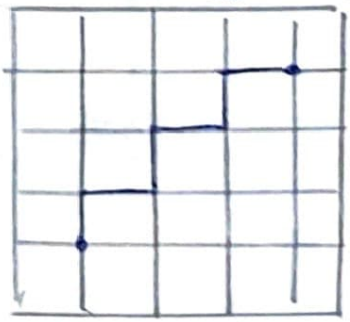
$$= \sqrt{32} = 5.65$$



- Manhattan Distance

It is similar to Euclidean Distance, but the distance (for example, two points, separated by building blocks in a city) is calculated by traversing vertical and horizontal lines in the grid-based system.

$$d_1 = |x_2 - x_1| + |y_2 - y_1|$$



- Minkowski Distance

It is a metric on the Euclidean space can be considered as a generalization of both the Euclidean and Manhattan distance

$$dist = \left(\sum_{k=1}^{n} |p_k - q_k|^r \right)^{\frac{1}{r}}$$

when $r=1$, it computes Manhattan

when $r=2$, it computes Euclidean

⊙ Covariance

∴ Covariance & Correlation

The Covariance measure how two random variable change together.

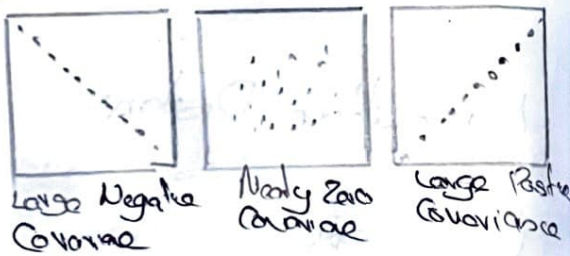
For population

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

For sample

$$\frac{}{(n-1)}$$

Covariance



Example.

x	y	$x-\bar{x}$	$x-\bar{y}$	$(x-\bar{x})\cdot(y-\bar{y})$
1	2	-2	-4	8
2	4	-1	-2	2
3	6	0	0	0
4	8	1	2	2
5	10	2	4	8

$\bar{x} = 3 \quad \bar{y} = 6 \quad \Sigma(x-\bar{x})\cdot(y-\bar{y}) = 20$

$$Covariance = \frac{20}{4} = 5$$

Here n-1 because Sample.

① Correlation

- It is the scaled version of Covariance
- It show how strong the relationship
- It has range -1 to $+1$

Formula:

$$Corr(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$\bar{x} = 3 \quad \bar{y} = 6$

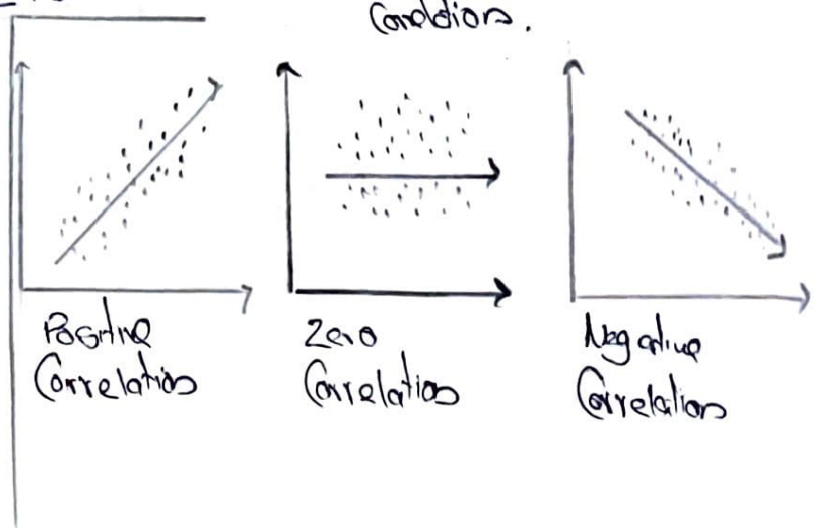
$(x-\bar{x})^2$	$(y-\bar{y})^2$
4	16
1	4
0	0
1	4
4	16
Σ 10	Σ 40

$$Corr = \frac{20}{\sqrt{400}}$$

$$= \frac{20}{20} = 1$$

+1 mean highly correlated

Condition.



- Based on the correlation coefficient, we can select the important features:
  - If the independent feature is not correlated with the target variable, we could remove that feature
  - Or if two independent features are highly correlated, we could remove any one feature

---

- Hypothesis Testing
  - Hypothesis is a statement, assumption or claim about the parameter (mean, variance, median etc)
  - A hypothesis is an educated guess about the something that could around you. It should be testable, either by experiment or observation

Example: if we make a statement that 'Dhoni is the best Indian Captain ever'. This is an assumption that we are making based on the average wins and loses team had under his captaincy. We can test this statement based on all the match data.

- Comparing and Analyzing Relationships
  - Does the treatment with new drug help more patients than the standard treatment with old drug?
  - Which of these four methods is the most efficient way of teaching machine learning?

- Types of Hypothesis

. When a hypothesis specifies an exact value of parameter, it is simple hypothesis. For eg. A motorcycle company claiming that a certain model gives an average mileage of 100 kms per litre, this is a ~~example~~ case of simple hypothesis.

. If a hypothesis specifies a range of values then it is called a composite hypothesis. For eg. Average age of students in a class is greater than 20. This statement is a composite hypothesis.

- Null Hypothesis (H_0)

. The null hypothesis is the hypothesis to be tested for possible rejection under the assumption that it is true

. The concept of the null is similar to innocent until proven guilty.

- Alternate Hypothesis (H_A)/(H_1)

. The alternative hypothesis complements the Null hypothesis

. It is opposite of the null hypothesis such that both the alternative and null hypothesis together cover all possible values of the population parameter.

- Hypothesis Testing - Case Discussion

Consider a court of law:

This null hypothesis is always begins with the assumption that the defendant is innocent.

. We require evidence to reject the null hypothesis (convict the defendant)

- When we collect evidence and try to reject null hypothesis, there are 2 errors that could potentially occur:

Type I or Type 2 error.

Decision	H0 True	H0 False
Reject H0	Type I error	Correct Decision
Do not reject H0	Correct Decision	Type II error