SNS COLLEGE OF TECHNOLOGY

An Autonomous Institution
Coimbatore-35

DEPARTMENT OF ARTIFICIAL INTELLIGENCE & DATA SCIENCE

23ADT202 – FUNDAMENTALS OF DATA SCIENCE AND ANALYTICS

II YEAR IV SEM

UNIT II – z scores, correlation and scatter plots

What is Z-Score?

- Z-Score in statistics is a measurement of how much standard deviations away a data point is from the mean of a distribution.
- A z-score of 0 indicates that the data point's score is the same as the mean score.
- A positive z-score indicates that the data point is above average, while a negative z-score indicates that the data point is below average.

- It helps compare data values from different data sets by converting them to a common scale, even if the data sets have different averages and spreads.
- Z-score is a statistical measure that describes a value's position relative to the mean of a group of values.
- It is expressed in terms of standard deviations from the mean.
- The Z-score indicates how many standard deviations an element is from the mean.

Z-Score Formula

To calculate the z- score for any given data we need the value of the element along with the mean and standard deviation. A z-score can be calculated using the following Z- score formula.

$$z = (X - \mu) / \sigma$$

where,

z = Z-Score

X = Value of Element

μ = Population Mean

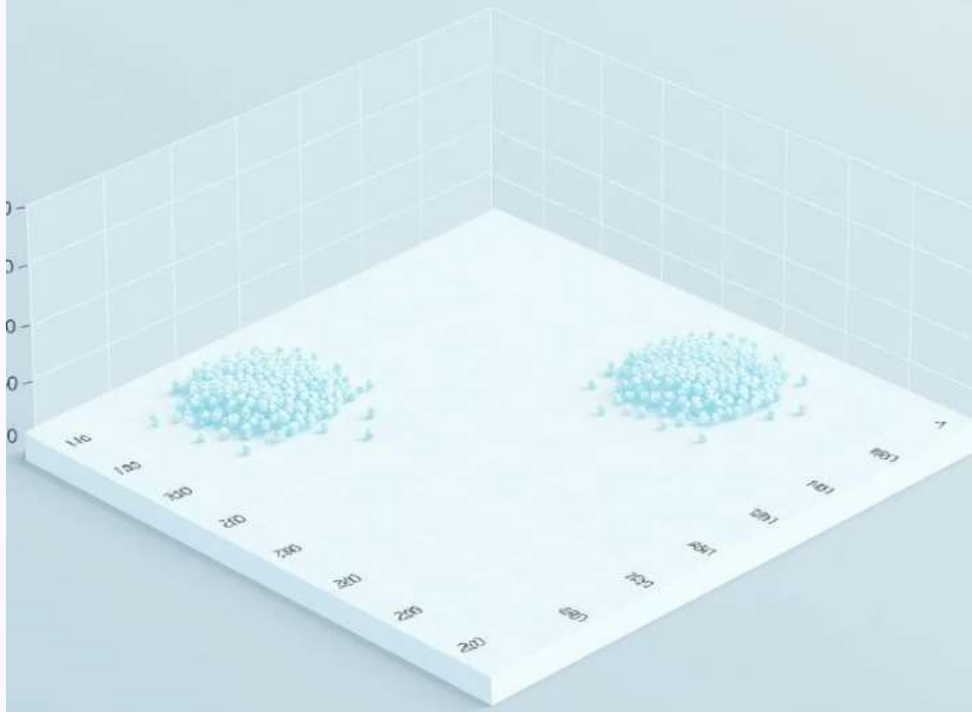
σ = Population Standard Deviation

EMPATHY:

How to Calculate Z-Score?

- We are given the population mean (μ), the population standard deviation (σ), and the observed value (x) by substituting these values in the z- score formula we can calculate the Z-Score value.

- Depending upon whether the given Z-Score is positive or negative, we can use positive Z-Table or negative Z-Table available online or on the back of your statistics textbook in the appendix.



Example 1: You take the GATE examination and score 500. The mean score for the GATE is 390 and the standard deviation is 45. How well did you score on the test compared to the average test taker?

Solution:
Following data is readily available in the above question statement

Raw score/observed value = X = 500

Mean score = $\mu = 390$

Standard deviation = $\sigma = 45$

By applying the z-score formula ,

$z = (X - \mu) / \sigma$
$z = (500 - 390) / 45$
$z = 110 / 45 = 2.44$

- This means that your z-score is 2.44.
- Since the Z-Score is positive 2.44, we will make use of the positive Z-Table.
- Now let's take a look at Z Table to know how well you scored compared to the other test-takers.
- Follow the instruction below to find the probability from the table.

Here, z-score = 2.44, which indicates that the data point is 2.44 standard deviations above the mean.

- Firstly, map the first two digits 2.4 on the Y-axis.
- Then along the X-axis, map 0.04
- Join both axes. The intersection of the two will provide you the cumulative probability associated with the Z-score value you're looking for
- [This probability represents the area under the standard normal curve to the left of the Z-score]

Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.9608	0.96164
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926
1.9	0.97128	0.97193	0.97257	0.9732	0.97381	0.97441	0.975	0.97558
2	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.9803	0.98077
2.1	0.98214	0.98257	0.983	0.98341	0.98382	0.98422	0.98461	0.985
2.2	0.9861	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.9884
2.3	0.98928	0.98956	0.98983	0.9901	0.99036	0.99061	0.99086	0.99111
2.4	0.9918	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324
2.5	0.99379	0.99396	0.99413	0.9943	0.99446	0.99461	0.99477	0.99492
2.6	0.99534	0.99547	0.9956	0.99573	0.99585	0.99598	0.99609	0.99621
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.9972

- As a result, you will get the final value which is 0.99266.

- Now, we need to compare how our original score of 500 on the GATE examination compares to the average score of the batch.

- To do that we need to convert the cumulative probability associated with the Z-score into a percentage value.

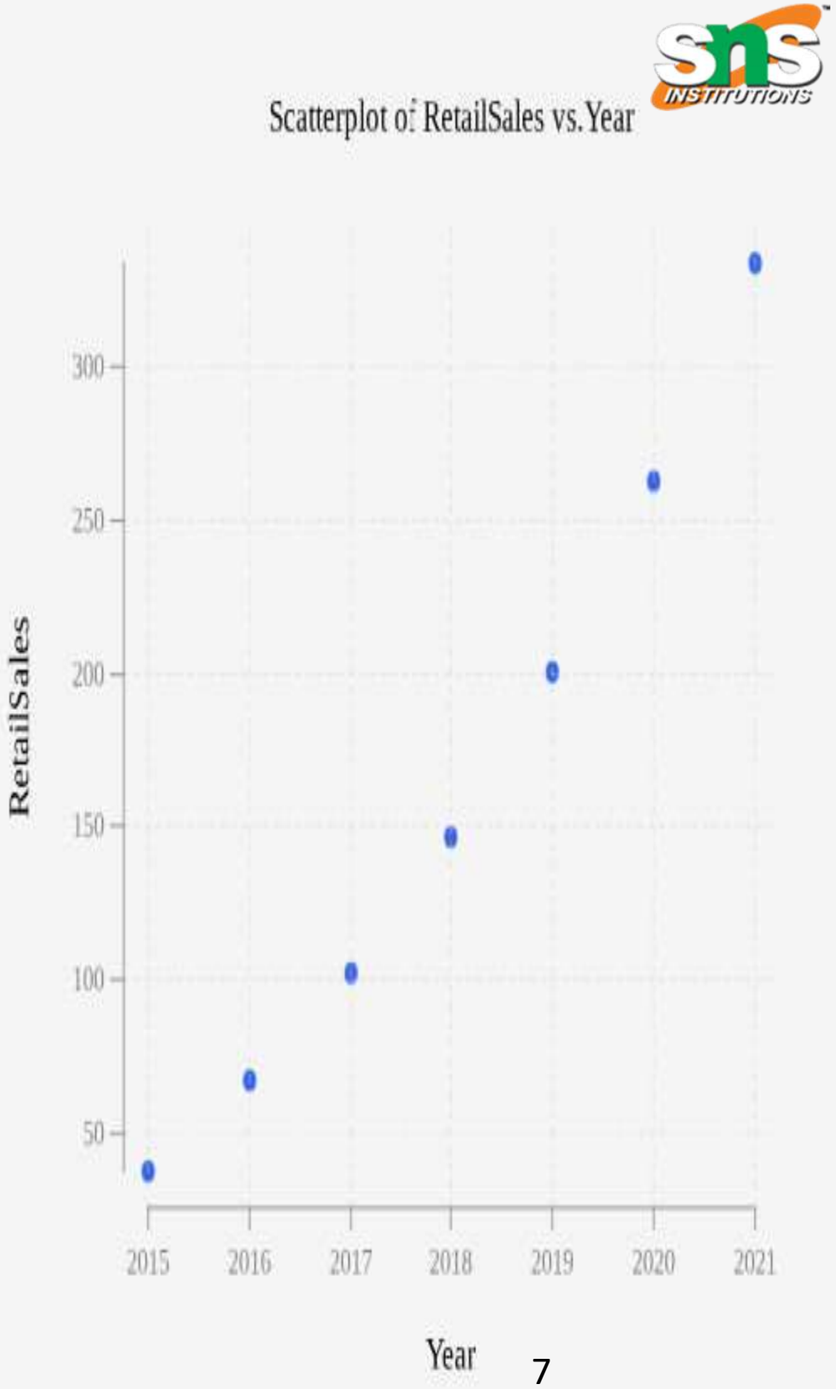
- 0.99266 × 100 = 99.266%

- Finally, you can say that you have performed well than almost 99% of other test-takers.

DEFINE:

Scatter plot

- A scatter plot shows the direction of a relationship between the variables. The direction can be positive or negative.
- A positive association happens when there are high values of one variable occurring with high values of the other variable and low values of one variable occurring with low values of the other variable.
- This was the case in Example 1, with the year and retail sales.
- Another way the direction of the relationship can show up is to see high values of one variable occurring with low values of the other variable.
- This would be a negative association.



- You can determine the strength of the relationship by looking at the scatter plot and seeing how close the points are to a linear function, a power function, an exponential function, a sinusoidal function, or to some other type of function.

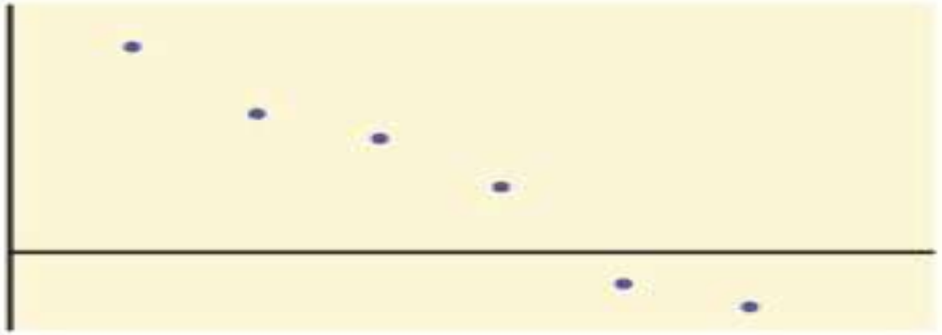
- The relationship can be strong, weak, or there can be no discernable relationship at all.

- When looking at a scatterplot, notice the overall pattern and any deviations from the pattern.

- The following six scatterplot examples in Figure 1 illustrate these concepts.

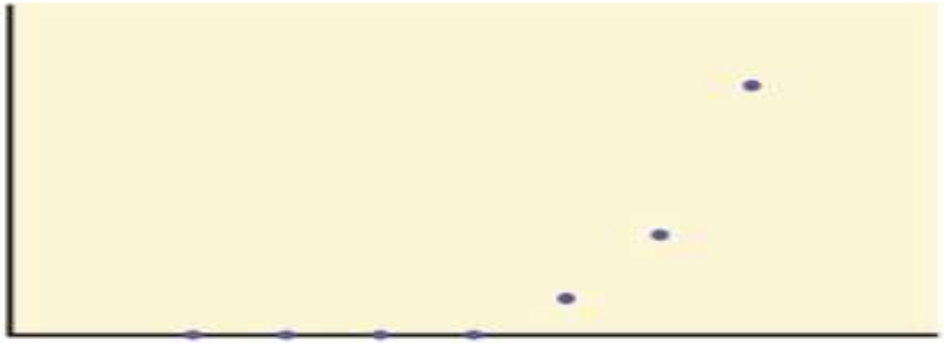
(a) Positive Strong Linear Pattern

(b) Positive Linear Pattern with a Deviation

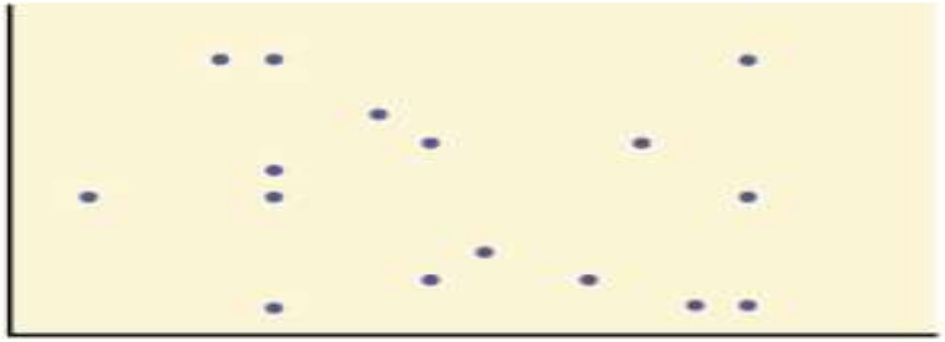


(c) Negative Strong Linear Pattern

(d) Negative Weak Linear Pattern



(e) Positive Strong Exponential Pattern



(f) No Pattern

IDEATE:

The Correlation Coefficient

- The correlation coefficient, r, developed by Karl Pearson in the early 1900s, is single number which provides a measure of strength and direction of the linear association between the independent variable x and the dependent variable y.

- There are several equivalent formulas for the correlation coefficient.

- In general, we will use a statistical program to calculate the correlation but it is worth taking a close look at the formula to get a sense of what it is calculating.

One formula for the correlation coefficient gives insight into what the statistic is calculating, where n is the number of data points.

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

PROTOTYPING & TESTING:

Correlation Reminders:

- The correlation coefficient r is a measure of the strength and direction of a linear association between two quantitative variables.
- The correlation coefficient is a number between −1 and 1.
- The closer r is to 1 or −1, the closer the data are to having a perfect linear association.
- The value of r alone cannot tell you if an association between two variables exists. A value of r close to 0 does not mean that there is no association. Before interpreting a correlation coefficient, we must first look at the scatterplot.

- The correlation coefficient measures association, but not causation.
- A strong correlation between two variables is evidence that there is a statistical relationship between the variables.
- Only the results of an experiment using random selection can establish a causal connection between two variables.