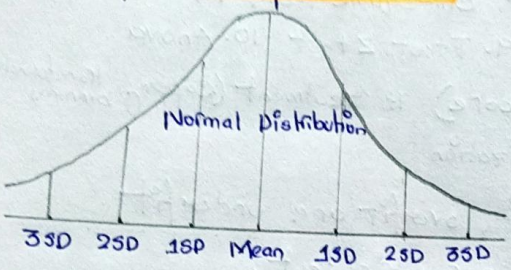
Z Score for Outlier Detection

- Z score is also called standard score.
- This score helps to understand if a data value is greater or smaller than mean and how far away it is from mean.
- More specifically, Z score tells how many standard deviation away a data point is from mean.



- A normal distribution is necessary.
- 68% of the data lies between +/-1SD.
- 95% of the data lies between +/-2SD.
- 99.7% of the data lies between +/-3SD.

$$\text{Z score} = \frac{x - \text{mean}}{\text{standard deviation}}$$

Z score and Outliers —

If the zscore of a data point is more than 3, it indicates that the data point is quite different from other data points. Such point is Outlier.

For example,

In a survey, it was asked how many children a person had.

1, 2, 2, 2, 3, 1, 1, 15, 2, 2, 2, 3, 1, 1, 2. → Clearly 15 is outlier.

mean of dataset = 2.66.

std deviation of dataset = 3.35.

$$\text{Z score (15)} = \frac{15 - 2.6}{3.3}$$

$$= 4.13.$$

If we take threshold = 3 which is 99.7%.

15 is outside 99.7% of data.

Disadvantage → Z score is sensitive to outlier because mean itself is sensitive to outlier/extreme values.