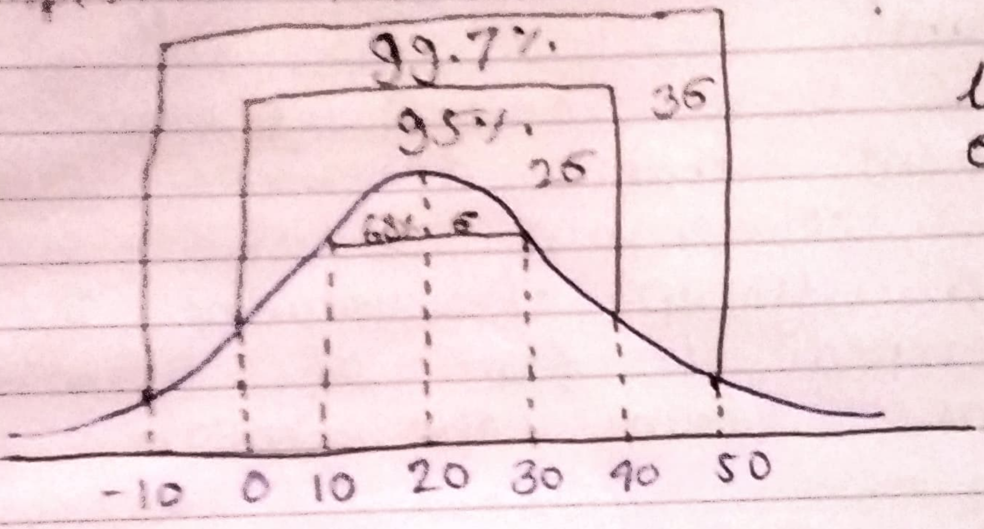
1. write the Gaussian Distribution emperical formula.



$\mu = 20$
$\sigma = 10$

If a data is in Gaussian Distribution the first standard deviation (σ) will contain 68% of data.

For Second Standard deviation (2σ) will contain 95% of all data available.

For Third standard deviation (3σ) will contain 99.7% of all data.

(REPUBLIC DAY) WEDNESDAY 26

Emperical formula: 68-95-99.7

2. What is z-score, and why is it important?

→ A Z-score is a numerical measurement that describes a value's relationship to the mean of a group of values. Z-score is measured in term of standard deviations from the mean.

If a Z-score is 0, it indicates that the data point's score is identical to the mean.

A Z score of 1.0 would indicate a value indicate a value that is one standard deviation from a mean.

Z-score may be positive or negative, with a positive value indicating the score is above the mean and a negative score indicating it is below the mean.

$$Z = \frac{x_i - \mu}{\sigma}$$

3. What is an outlier, Exactly?

-> An outliers is an observation that lies an abnormal distance from other values in a random sample from a population. In a sence, this definition leaves it up to the analyst to decide what will be considered abnormal. Before abnormal observations can be ~~sign~~ Singled out, it is necessary to ~~chater~~ characterize normal observations.

We take use five ~~of~~ number Summary for finding the outliers.

Dataset :-

Dataset :- 1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27

ii) calculate the median

Next, we need to calculate the median. The median is the center of the data if the data set has an odd number of data points, then the mean is centermost number. On the other hand, if the data set has an even number of values, then we will need to take the arithmetic average of the two centermost values. We will calculate the average by adding the two numbers together and then dividing that number by two.

Q_2 (median) – 10th position value
– 5

iii) calculate upper & lower limit :-

Q_1 (25%) $= \frac{\text{Percentile}}{100} \times (n+1)$

$= \frac{25}{100} \times (19+1)$

$= \frac{25}{100} \times 20$ = 5th index
= 3 (value)

Q_3 (70%) $= \frac{\text{Percentile}}{100} \times (n+1)$

$= \frac{75}{100} \times (19+1)$

$= \frac{75}{100} \times 20$

$=$ 15th index

$\Rightarrow 7$

iv) calculate the difference:-

$IQR = Q_3 - Q_1$

$= 7 - 3$

$= 4$

Lower fence:- $Q_1 - 1.5 \times IQR$

$= 3 - 1.5 \times 4$

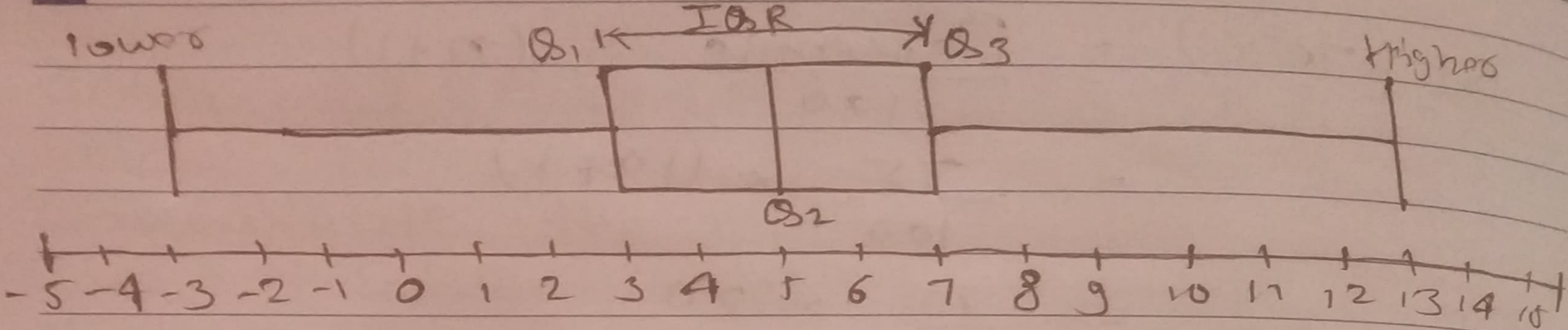
$= 3 - 6$

$= -3$

Higher fence $= Q_3 + 1.5 \times IQR$

$= 7 + 1.5 \times 4$

$= 7 + 6 = 13$



If any value goes beyond lower fence and higher fence. it is considered as outliers.
in the given dataset 27 is considered as outliers.

4. What are our options for dealing with outliers in our dataset?

→ One of the most important steps as part of data preprocessing is detecting and treating the outliers as they can negatively affect the statical analysis and the training process of a machine learning algorithm resulting in lower accuracy.

An outlier may occur due to the experimental error/ human error. They may indicate an experimental error or heavy skewness in data.

If the dataset is small, we can just look at it and find out the outliers, but if it is big then we need to have the mathematical and visualization technique. These techniques can be as follows:

30 SUNDAY

5. write the Sample and population variances equation and explain Bessel Correction.

→

Sample variance:-

$$\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n-1}$$

Population variance:-

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

Bassel's Correction explains that "the use of n-1 instead of n in the formula for Sample variance and Standard variance, where n is the number of data in Sample. This corrects the bias of the estimation of the population variance.