BWT TASK 3

1. Probability
Probability means how likely an event is to occur.

2. Mean, median and mode

- Mean is calculated by dividing the sum of all values by the total number of values.
- Median is the average of two middle points.
- Mode is the most frequently occurring value in the dataset. Mode is the value at which distribution has the highest probability.

Mean, median and mode describe the center of the distribution.

3. Variance and standard deviation
Variance tells the degree of the spread in the dataset.
If the variance is high, it means the data points are spread out a lot.

Standard deviation is the square root of variance.
It tells on average, how far each value lies from the mean.

4. Percentiles and Quartiles
Percentiles divide the dataset into 100 equal parts.
Each part contains 1% of the data.

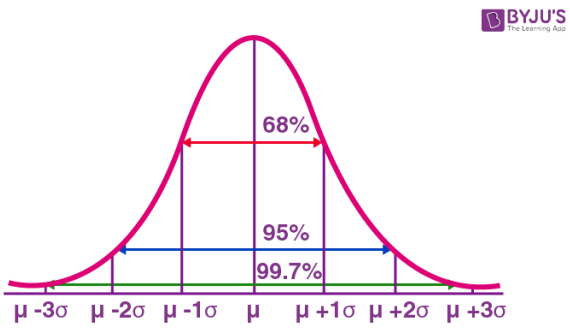
Quartiles divide the dataset into 4 equal parts.
Each part contains 25% of the data.

Interquartile range is the difference between 75^{th} and 25^{th} percentile or third quartile and first quartile (Q3 – Q1)

25^{th} percentile = First quartile (Q1) -> 25 percent of the data lies below this point.
50^{th} percentile = Median (Q2) -> 50 percent of the data lies below this point.
75^{th} percentile = Third quartile (Q3) -> 75 percent of the data lies below this point.

5. Normal distribution

A normal distribution is a bell shaped frequency distribution curve of a continuous random variable.



Source: Byju's

- 68% of the distribution's data lies between +1 and -1 std. deviation.
- 95% of the distribution's data lies between +2 and -2 std. deviation.
- 99.7% of the distribution's data lies between +3 and -3 std. deviation.

People's heights, weights, I.Q follows a normal distribution.
Many models in machine learning are designed under the assumption that the variables follow a normal distribution.

6. Confidence interval

A confidence interval is a range of values that you expect your estimate to fall between a certain percentage of the time if you run your experiment again or re sample the population in the same way.
Confidence interval= sample mean ± Margin of error

7. Hypothesis testing

Hypothesis testing is used to verify whether the results of an experiment are valid or not by using the null and alternate hypotheses.

Null Hypothesis (Ho): There's no effect in the population.
Alternate hypothesis (Ha): There's an effect in the population.

Example:
Ho: There is no difference in the salary of factory workers based on gender.
Ha: Male workers have a higher salary than female factory workers.

8. Law of large numbers
As the sample size increases, the average of the sample gets closer to the average of the entire population.

Conditions:

- Sample is randomly drawn.
- Sample size must be sufficiently large.
- Observations must be independent.

9. Central limit theorem
It states that the sampling distribution of the means will always be normally distributed, as long as the sample size is large enough.
Regardless of whether the population has a binomial, normal, Poisson or any other distribution the sampling distribution of the mean will be normal.

10. Covariance
It is a measure of how much two random variables vary together.
Covariance values are sensitive to the scale of data.

- If covariance > 0 ; variables move in the same direction
- If covariance = 0 ; No relation between variables
- If covariance < 0 ; variables move in the opposite direction

11. Correlation
Correlation measures the strength and direction of the linear relationship between 2 variables.

It is represented by a “correlation coefficient ” (r). Its value ranges from -1 to +1.

- Positive correlation ($0 < r \leq 1$): Both variables move in the same direction
- Negative correlation ($-1 \leq r < 0$): Variables move in opposite direction
- No correlation ($r \approx 0$): There is no consistent linear relationship between variables.

If the value is close to 1, it indicates a strong positive correlation.
If the value is close to -1, it indicates a strong negative correlation