

Assignment - 3

Ques 1 Write the Gaussian Distribution Empirical Formula.

Ans ① If data is Gaussian Distributed we can say that 1st region $(\mu - \sigma)$ to $(\mu + \sigma)$ will contain 68% of data.

② 2nd region that is $(\mu - 2\sigma)$ to $(\mu + 2\sigma)$ will contain 95% of data.

③ 3rd region that is $(\mu - 3\sigma)$ to $(\mu + 3\sigma)$ will contain 99.7% of data.

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

Ans Z-score describes the position of a raw score in terms of its distance from the mean when measured in standard deviation units.

Z score +ve if value lies above mean & -ve if below the mean.

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

Important because

- ① Enables us to compare two scores that arise from different samples.
- ② Allows us to calculate the probability of score occurring within SND.

Ques 3

What is an outlier, exactly?

Ans

An outlier is an observation that lies an abnormal distance from other values in sample or population.

Ques 4

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

Ans

We can use following methods to detect & deal with outliers.

- Detection {
- ① Box plot (~~S-Summary S-number Summary~~)
 - ② Histogram
 - ③ Inter Quartile Range (IQR)
 - ④ Z-Score.
- Dealing {
- ① Trimming using S-number Summary.
 - ② Mean / Median Imputation

Ques 5

Write Sample & population variance equations and explain Bessel Correction.

Ans

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

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

Bessel Correction \rightarrow It is the use of $(n-1)$ instead of n in the formula for Sample variance & Standard deviation.

This method corrects the bias in the estimation of population Variance.