



**Statistics Assignment 3**

1. Write the Gaussian Distribution empirical formula.

**Answer:**

Gaussian distribution is a bell shaped curve and it is assumed that during any measurement values will follow normal distribution with an equal number of measurements above and below the mean value. If distribution is normal the value of mean, median and mode value are the same.

Gaussian distribution follows the Empirical Rule as below.

If your distribution is normal -

* 68.2% of data values contain within Mean +/- One Standard deviation
* 95.5% of data values contain within Mean +/- Two Standard deviation
* 99.7% of data values contain within Mean +/- Three Standard deviation

(𝝻 +

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

**Answer:**

Z score also called as standard score gives you an idea of how far from the mean a data point is. Its measure of data point of population means how many standard deviations below and above the mean. The Z score range is -3 standard deviation upto +3 standard deviation. To measure Z score we must know population mean and standard deviation.

The formula of Z Score is :- **z = (x – μ) / σ**

It is important because it not only tells you the value but also where the value lies in the distribution.

3. What is an outlier, exactly?

**Answer:**

Outlier is those values in the data set which lies extremely high or low relative to the nearest data values, that standout far from overall patterns of the data set.

To find acceptable range and determine outlier position in the data set below the statistical method can be used.

Q1 = 25th Percentile - All data points that fall below 25% of data set values.

Q2 = 75th Percentile - All data points that fall below 75% of data set values.

IQR = Q3-Q1(The values falls within 25th and 75th quartile)

Outlier < Q1-1.5(IQR)

This means that as a data point needs to fall more than 1.5 times the IQR range below the first quartile(Q1) to be considered a low outlier.

Outlier > Q3+1.5(IQR)

This means that as a data point needs to fall more than 1.5 times the IQR range above the third quartile(Q3) to be considered a high outlier.

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

**Answer:**

Outliers in the data set could be anomalies, and may occur because of human error.. If it is present in a smaller amount then it is a better method to drop from the data set.

Outliers usually misinterpret the central tendency of the data set. If data is with no outliers, the mean will be accurate.

If outliers are present in the data set then better it is input with the median value of the data set.

If 50% of data are equivalent value then it is preferable to input the outlier with mode.

Before that we have detect the outliers in data set using below method-

* IQR
* Box Plot
* Z-Score

Handling outliers with below method

* Trimming or removing them from data set if they are in small in amount
* If they are in considerably in volume then impute them with Median or Mean value
* Capping them with close value which is within min max range. This call completely depends on the nature of data or characteristic of data and assigned target.

5. Write the sample and population variances equations and explain Bessel Correction.

**Answer:**

Population Variance :-

Population variance calculated on population dataset, in this no of observations denoted by capital N and formula is as below.

∑ni=1(xi−μ)2n ∑ i = 1 n ( x i − μ ) 2 n

Sample variance :-

Sample Variance calculated on sample dataset, in this no of observations denoted by small n and formula is as below.

**∑ni=1(xi−μ)2n−1 ∑ i = 1 n ( x i − μ ) 2 n − 1**

Bessel Correction :- In estimating the population variance from a sample when population mean is unknown, the uncorrected sample variance is the mean of the squares of deviations of sample values from the sample mean (i.e. using a multiplicative factor 1/n). In this case, the sample variance is a biased estimator of the population variance.

Multiplying the uncorrected sample variance by the factor gives an unbiased estimator of the population variance. This factor has called Bessel Correction or Degree of Freedom in statistics.