Statistics Assignment 3

1. Write the Gaussian Distribution empirical formula?

Answer:

Gaussian Distribution empirical formula is known as 68-95-99.7 rule. The Empirical formula states that 68% of distribution falls within the first standard deviation i.e. ($\mu \pm \sigma$), 95% of distribution falls within the second standard deviation i.e. ($\mu \pm 2\sigma$), and 99.7% of distribution falls within the third standard deviation i.e. ($\mu \pm 3\sigma$).

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

Answer:

Z-score gives a relationship between value and the mean of group of values.
Z-score helps to find how much standard deviation away a value is from mean.
Z-score = (Observed Value – Mean) / Standard Deviation = $(x - \mu) / \sigma$

Now why Z-score is important?
In dataset with multiple features, to bring each feature to a same scale so that our Machine Learning Algorithms preform in better way. This process to bring each feature to same scale is known as Standardization.
Z-scores are standardized values that can be used to compare scores in different distribution.

3. What is an outlier, exactly?

Answer:

Outlier is an extreme value which is very far from mean of the data or extremely low or extremely high from the other data values.
According to empirical rule (68-95-99.7 rule), any value which lies outside the 99.7% of the distribution is known as outliers.
According to IQR formula, if a value is either higher than Upperfence(Q3 + (1.5*IQR)) or lower than Lowerfence(Q1 - (1.5*IQR)) will be known as Outlier.
For example, dataset = {2,3,3,4,5,6,6,7,7,8,8,8,9,10,100,110}. In this dataset 100 and 110 will be outliers.

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

Answer:

We can deal with outliers in 3 ways:

I. Using Z-score:-
   If absolute value of Z-score of any data point is greater than the third standard deviation, then that data point will be outlier.

```
threshold = 3          # third standard deviation
mean = np.mean(data)
std = np.std(data)
for i in data:
   z_score = (i-mean)/std
   if np.abs(z_score)>threshold:
         outlier.append(i)
return outilers
```

II. Using Interquartile Range (IQR):-

In first step, we sort our dataset. Then we will find Q1, Q3 and IQR where Q1 and Q3 are 25 and 75 percentile respectively and IQR = (Q3 – Q1).
In next step we will calculate Lowerfence as Q1 – (1.5 * IQR) and Upperfence as Q3 + (1.5 * IQR)
Now any data point which is either lower than Lowerfence or greater than Upperfence will be added to our Outlier list.

III. Using Box plot:-

Using boxplot we can also find outliers which can be seen away from our boxplot in the graph.

```
import seaborn as sns
sns.boxplot()
```

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

Answer:

Sample Variance: $\mathbf{s^2 = \sum(x - \bar{x})^2 / (n - 1)}$

where, x = observed value, $\bar{x}$ = sample mean, and n = sample size

Population Variance: $\mathbf{\sigma^2 = \sum(x - \mu)^2/N}$

where, x = observed value, μ = population mean, and N = population size

Bessel Correction:

In the above given Sample Variance formula, in denominator we have (n-1) instead of n. This is nothing but Bessel Correction. We did it to make Sample Mean become much closer to Population Mean. So that an accurate Population Mean can be determined from the Sample Mean. If we use n instead of (n-1) then there can be huge possibility that the difference between Sample Mean and Population Mean can be large.