Q1.. Write the Gaussian Distribution empirical formula.

Answer: Under this rule, 68% of the data falls within one standard deviation, 95% percent within two standard deviations, and 99.7% within three standard deviations from the mean.

Q2. What is the Z-score, and why is it important?

Answer: A measure of how many standard deviations below or above the population mean a raw score is called z score. It will be positive if the value lies above the mean and negative if it lies below the mean.

Formula

The equation is given by z = (x – μ)/ σ.

μ = mean

σ = standard deviation

x = test value

Q3. What is an outlier, exactly?

Answer: outlier is an extremely high or extremely low data point relative to the nearest data point and the rest of the neighboring co-existing values in a data

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

Answer:

Transforming outliers

A third option for dealing with outliers is to transform them using mathematical functions or techniques that can alter the scale or shape of the data. Logarithmic, square root, and inverse transformations can reduce the skewness or asymmetry of the data, making it more normal or symmetrical.

Replacing method: o handle outliers is to replace them with more reasonable values. This can be done by utilizing different methods of imputation, such as using the mean, median, mode

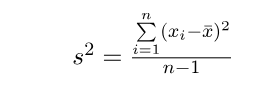
Removing outliers

One option to handle outliers is to remove them from the data. This can be done by applying a threshold or a rule to filter out the outliers.

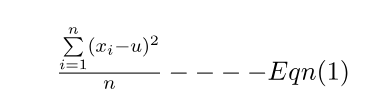
5. Write the sample and population variances equations and explain Bessel Correction

Bessel correction refers to the n-1 part used as the denominator in the formula of sample variance.

sample variance:



Population variance:





Since u is not known so sample-mean proves to be the best estimator that can be used. Hence

Problem:

Subtracting sample-mean in Eqn(2) makes this sum as small as it possibly could be, roughly the sample mean must fall near the center of the observations whereas population-mean could be any value. So the sum in Eqn(2) is going to be smaller than the sum in the Eqn(1), hence Eqn(2) tends to underestimate the true value of the population variance.

Solution:

To compensate for that, dividing by n-1 makes the sample variance a little bigger than it would be if divide by n. It turns out that mathematically it properly compensates for the problem.