**Statistics Assignment 3**

1. Write the Gaussian Distribution empirical formula.

Answer : -

- [68% - 95% - 99.7%].

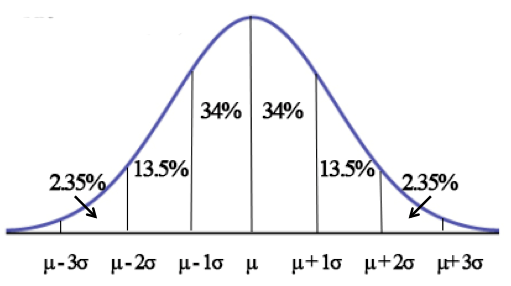
- 68% of data falls within 1 standard deviation from the mean - that means between μ - σ and μ + σ.

- 95% of data falls within 2 standard deviations from the mean - between μ – 2σ and

μ + 2σ.

- 99.7% of data falls within 3 standard deviations from the mean - between μ - 3σ and

μ + 3σ.



-----------------------------------------------------------------------------------------------------------------------

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

Answer : -

- A z-score describes the position of a raw score in terms of its distance from the mean, when measured in standard deviation units. The z-score is positive if the value lies above the mean, and negative if it lies below the mean.

- It is also known as a standard score, because it allows comparison of scores on different kinds of variables by standardizing the distribution. A standard normal distribution (SND) is a normally shaped distribution with a mean of 0 and a standard deviation (SD) of 1.

- The formula for calculating a z-score is z = (x-μ)/σ, where x is the raw score, μ is the population mean, and σ is the population standard deviation.

- It is useful to standardize the values (raw scores) of a normal distribution by converting them into z-scores because:

(a) It allows researchers to calculate the probability of a score occurring within a standard normal distribution.

(b) And enables us to compare two scores that are from different samples (which may have different means and standard deviations).

-----------------------------------------------------------------------------------------------------------------------

1. What is an outlier, exactly?

Answer: -

- An outlier is a data point that lies outside the overall pattern in a distribution.

- A commonly used rule says that a data point is an outlier if it is more than 1.5\*IQR above the third quartile or below the first quartile. Said differently, low outliers are below Q1 - 1.5 \* IQR and high outliers are above Q3 + 1.5 \* IQR.

- The formula for calculating a Lower Fence and Higher Fence are as follows:

Lower Fence = Q1 – 1.5 \* IQR

Higher Fence = Q3 + 1.5 \* IQR

IQR = Q3 – Q1

- Example:-

Dataset = { 1,2,2,2,3,3,4,5,5,5,6,6,6,6,7,8,8,9,27 }

Q1 (25%) = 25/100 \* 20 = 5 (index)

i.e. **Q1 = 3**

Q3 (75%) = 75/100 \* 20 = 15 (index)

i.e. **Q3 = 7**

IQR = Q3 – Q1

= 7 – 3

**IQR = 4**

Lower Fence = Q1 – 1.5 \* IQR

= 3 – 1.5 \* 4

= 3 – 6

**Lower Fence = -3**

Higher Fence = Q3 + 1.5 \* IQR

= 7 + 1.5 \* 4

= 7 + 6

**Higher Fence = 13**

- Data range should be in between [- 3 to 13].

- Therefore from the given dataset 27 is the outlier.

---------------------------------------------------------------------------------------------------------------------

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

Answer: -

- Remove or change outliers during post-test analysis.

- Change the value of outliers.

- Consider the underlying distribution.

- Consider the value of mild outliers.

---------------------------------------------------------------------------------------------------------------------------

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

Answer: -

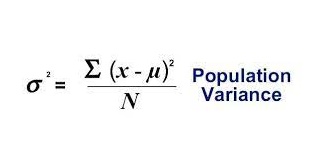
Variance:-

- Variance gives the dispersion or spread of a dataset.

Population Variances:-

- When the variance is calculated using the entire data, also known as the population, it gives the population variance.

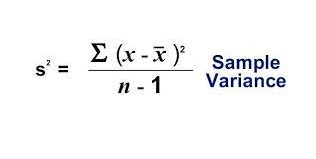
- Formula for Population Variance:



Sample Variances:-

- When the variance is calculated using the sample data it gives the sample variance.

- Formula for Sample Variance:



- In many real world scenarios, we do not really know what the population mean (μ) is, and therefore we can’t calculate the population standard deviation (σ). To circumvent this issue, we replace μ with its unbiased estimator (x̅ = the sample mean). However, when we do this, we “severely” underestimate the sum of squared differences and therefore, the estimated population variance. This “uncorrected sample standard deviation” cannot be used to estimate population standard deviation!

