Statistics Assignment 3

# **Write the Gaussian Distribution empirical formula.**

The 68-95-99.7 empirical formula, or the 3-sigma rule

Chart, histogram

Description automatically generated

# **What is the Z-score, and why is it important?**

Z-Scores are standardized values of a dataset that can be used to compare scores in different distributions.

Here is the formula for Z-score which can convert any normal distribution into a standard normal distribution:

Why do we use Z score?

* The z score converts any dataset to an equivalent dataset with
* Thus, is brings all datasets down to the same scale.
* This allows for comparison of datasets and makes possible running of various machine learning models on it.
* This technique is also called as standard scaling (or standardization) in machine learning.

# **What is an outlier, exactly?**

An outlier is an extremely high or extremely low data point relative to the nearest data point and the rest of the neighboring values in a data graph or dataset. Thus, outliers are extreme values that stand out greatly from the overall pattern of values in a dataset or graph.

A point beyond an inner fence on either side is considered a mild outlier. A point beyond an outer fence is considered an extreme outlier.

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

An outlier may occur due to the variability in the data, or due to experimental error/human error. They may indicate an experimental error or heavy skewness in the data.

Some of the visualization and mathematical techniques of detecting outliers are

* Boxplots, scatterplot & histogram
* Z-score

Some of the methods of treating the outliers are

* Removing outliers from the data. This may not be the right approach in many cases, but in some cases it may be meaningful. Therefore, it is essential to understand how outliers occur and whether they might happen again as a normal part of the process or study area. For example, outliers arising from data entry & measurement errors should be removed. If it is a natural part of the population you are studying, you should not remove it.
* Sampling problems can cause outliers; therefore, we should carefully define a population, and then draw a random sample from it.
* Quantile based flooring and capping
* Mean/Median imputation

# **Write the sample and population variances equations and explain Bessel Correction.**

The calculations for both the sample standard deviation and the sample variance both contain a little bias (or error). Bessel’s correction (i.e., subtracting 1 from your sample size) corrects this bias.

When we calculate the sample variance which is a sample statistic & infer the population variance from it which is a population parameter, there is going to be an inherent error because x̄ ≠ μ and any x-value in your random sample is going to be closer to x̄ than to μ. Therefore, the sum of squares for μ is going to be larger than the sum of squares for x̄. In other words, with sample calculations with n-1 in the denominator, you’re making a fairly good adjustment for the deflated sum of squares figure as long as n isn’t huge.

Degrees of Freedom Explanation:

While x̄ - μ ≠ 0, we will always have Σ(x - x̄) = 0

Thus, if you are trying to estimate the 'average' of the difference between the population mean and the set of values in the population, using sample residuals to do so will result in the last 'residual' not being free to vary over the 'full range' and it will just have to be whatever the value is that will allow the sum to add up to 0.

Thus, the final sample residual won't contribute the required quantity to the sum of (squared) residuals that would be needed to estimate the 'average' of n (squared) residuals, and so instead, we redistribute that final quantity to all the other residuals and divide the sum by n-1 instead.