**Literacy Rate Analysis: Normal Distribution & Outlier Detection**

Introduction*:*

In this analysis, we employed the normal distribution to model data related to the literacy rate in various districts in India. As part of the Department of Education in a large nation, our goal was to explore the distribution characteristics of district literacy rates and identify potential outliers using z-scores and derive conclusions.

# Libraries and Data Import:

We began by importing essential Python libraries, including Numpy, pandas, matplotlib, scipy.stats, and statsmodels. These libraries facilitated data manipulation, visualization, and statistical analysis. The dataset, named 'education\_districtwise.csv,' was loaded into a pandas DataFrame, and missing values were handled using the dropna() function.

A screenshot of a computer code

Description automatically generated

# Histogram Plot:

To understand the distribution of district literacy rates, we created a histogram using the 'OVERALL\_LI' column. The resulting plot exhibited a bell-shaped curve, indicative of a symmetric and approximately normal distribution. This observation prompted us to consider the normal distribution as a potential model for our data.

A graph with blue lines

Description automatically generated with medium confidence

# Empirical Rule Verification:

We verified the applicability of the empirical rule to our dataset, which states that for a normal distribution:

* 68% of values fall within +/- 1 standard deviation from the mean.
* 95% of values fall within +/- 2 standard deviations from the mean.
* 99.7% of values fall within +/- 3 standard deviations from the mean.

A diagram of a normal distribution with Ryugyong Hotel in the background

Description automatically generated

Using the mean and standard deviation of district literacy rates, we computed the actual percentages falling within each range and found close agreement with the empirical rule (66.4%, 95.4%, and 99.6%).

A white background with text

Description automatically generated

*Output:*

*A close-up of a white background

Description automatically generated*

*Z-Scores and Outlier Detection:* Next, we calculated z-scores for each district's literacy rate, providing a measure of how many standard deviations a data point is from the mean. This information was crucial for identifying outliers. Districts with z-scores smaller than -3 or larger than +3 were considered outliers. We identified two such outliers, namely DISTRICT461 and DISTRICT429, both exhibiting significantly lower literacy rates.

A diagram of a normal distribution

Description automatically generated

A computer screen shot of a computer code

Description automatically generated

Output :

A black text on a white background

Description automatically generated

Conclusion*:*

In conclusion, this analysis demonstrated the utility of the normal distribution in modeling district literacy rates. The agreement with the empirical rule and the identification of outliers using z-scores enhance our understanding of the distribution characteristics and outliers in the dataset. This information can be valuable for educational policymakers, allowing them to allocate resources more effectively, address specific districts with lower literacy rates, and contribute to overall educational improvement.