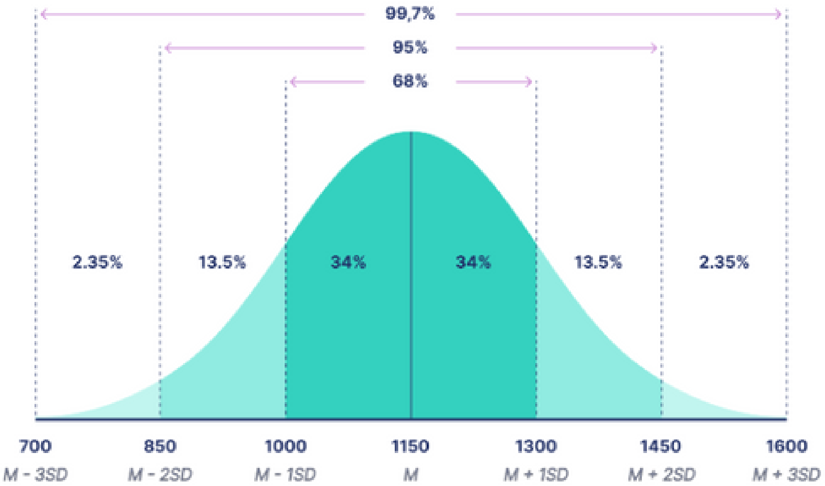
Data Analyst Perspective

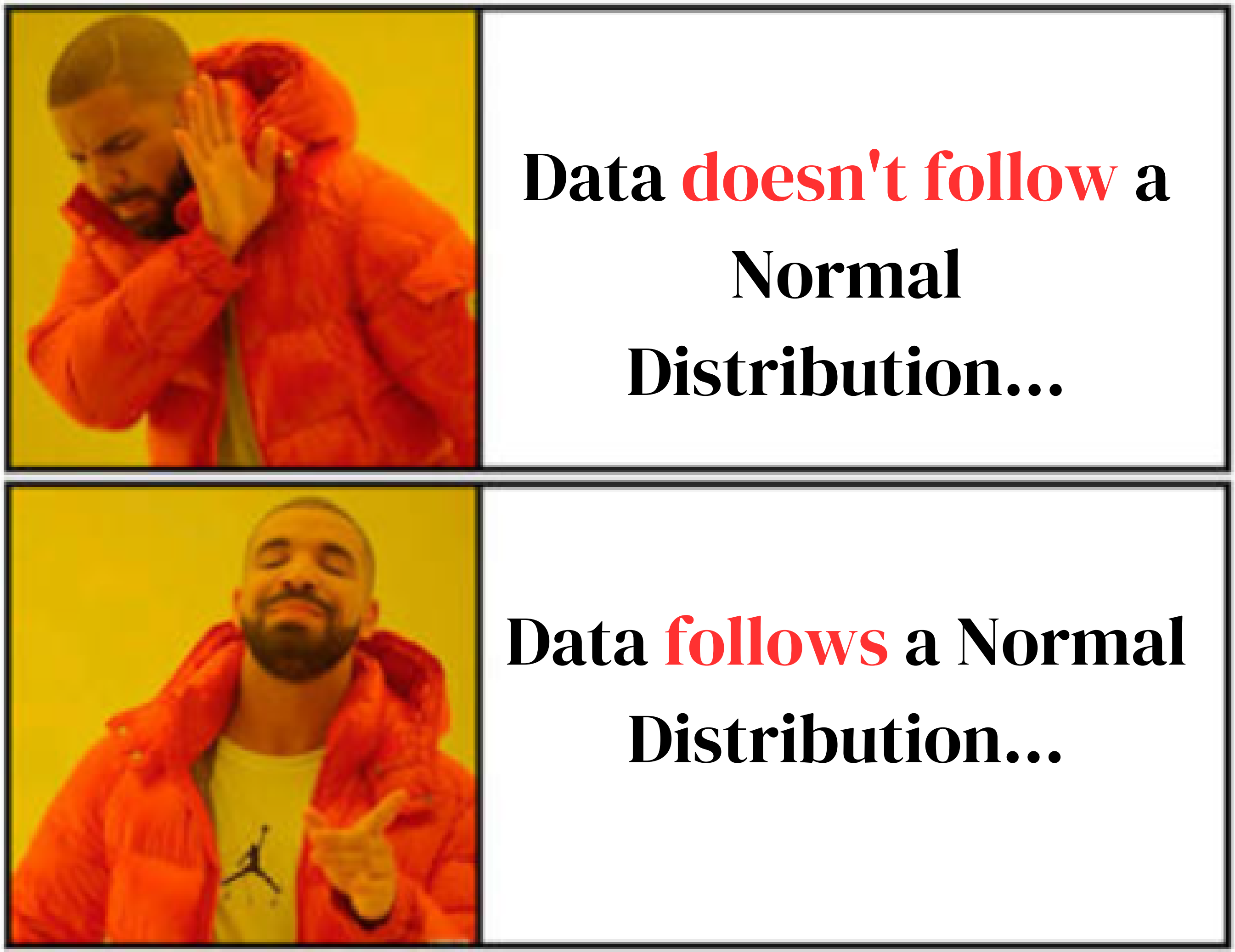
STATISTICS

DESCRIPTIVE STATISTICS

Normal Distribution



Data Scientist/Data Analysts

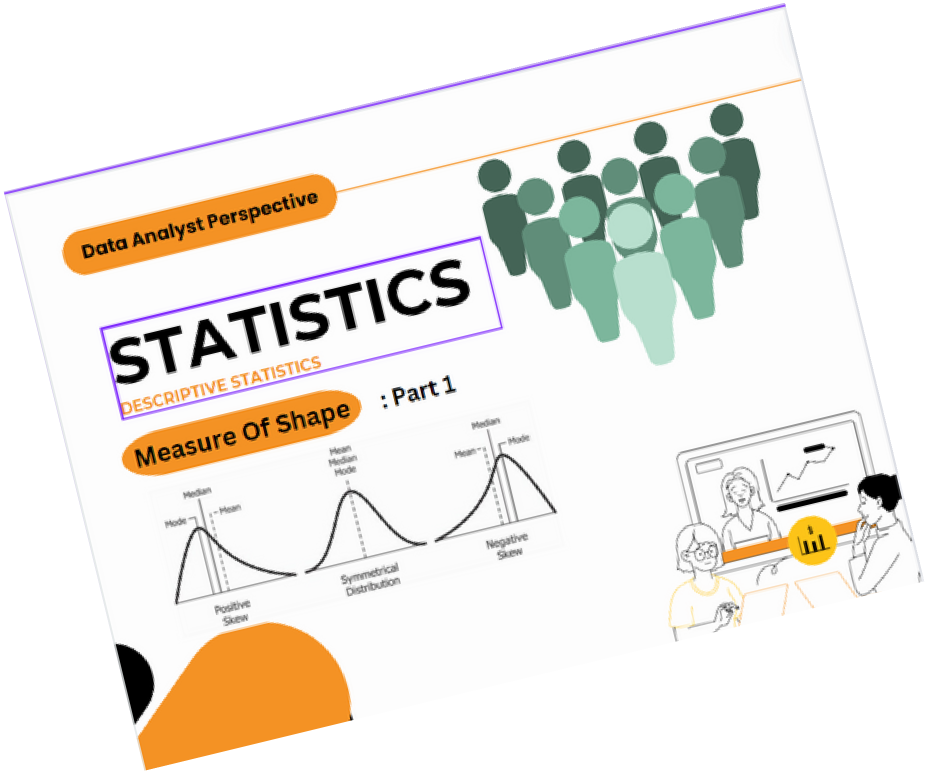


Data doesn't follow a Normal Distribution...

Data follows a Normal Distribution...

Agenda

- INTRODUCTION TO NORMAL DISTRIBUTION
- EMPIRICAL RULE
- APPLICATIONS OF NORMAL DISTRIBUTION
- METHODS TO CHECK NORMALITY



Data Analyst Perspective

STATISTICS

DESCRIPTIVE STATISTICS

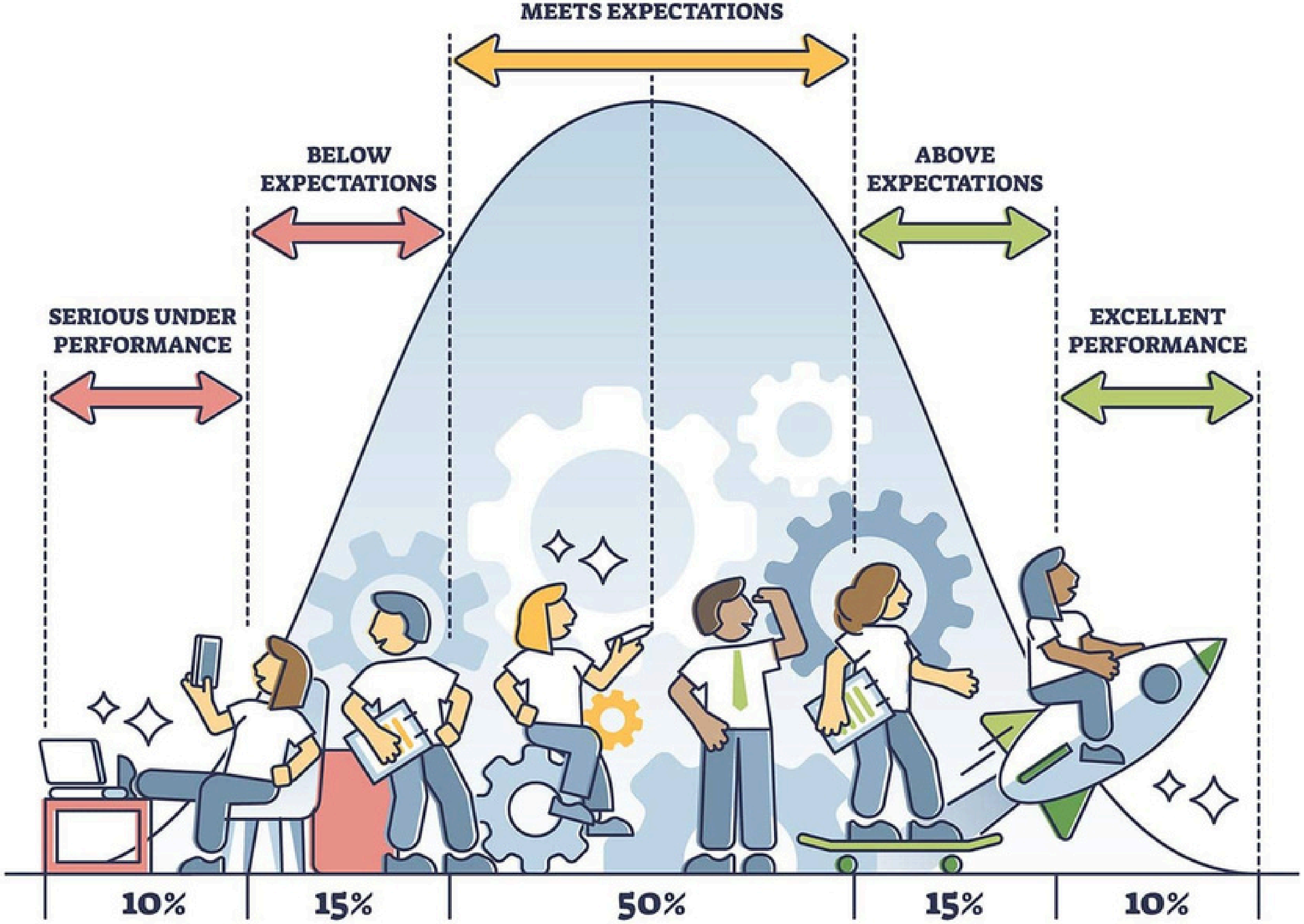
Measure Of Shape : Part 1

In this post , I explained what a data a distribution and skewness, feel free to have a quick revision for **better understanding of Normal Distribution.**

WHAT
IS NORMAL DISTRIBUTION

- Normal Distribution is a bell-shaped curve that is **symmetric about the mean.**
- In a normal distribution, **the mean, median, and mode are all equal** and located at the center of the distribution.
- It’s a fundamental concept in statistics because many natural phenomena follow this pattern.
- Normal distributions are also called **Gaussian distributions** or **bell curves** because of their shape.
- Many **Machine Learning Algorithms** works on assumption , that the data is normally distributed ,that’s why it is important.

BELL CURVE

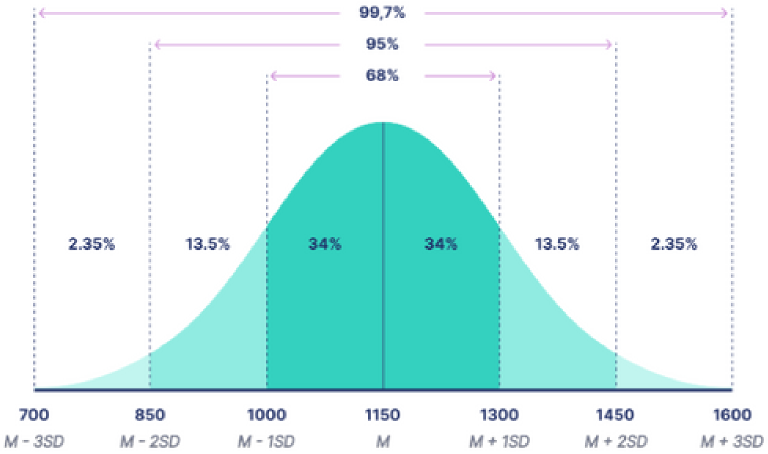

MEETS EXPECTATIONS
BELOW EXPECTATIONS
ABOVE EXPECTATIONS
SERIOUS UNDER PERFORMANCE
EXCELLENT PERFORMANCE
10%
15%
50%
15%
10%

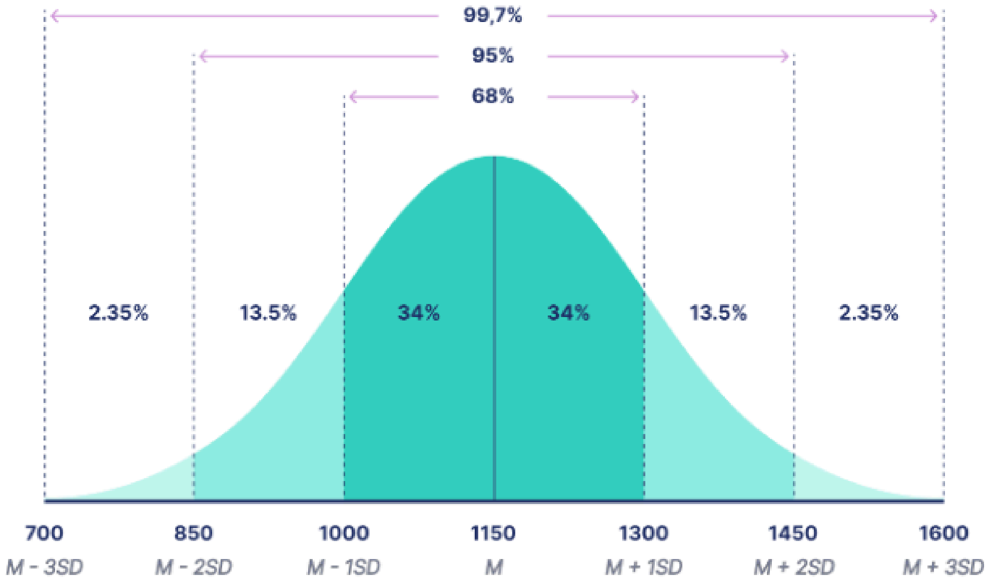
EMPIRICAL RULE

The **empirical rule,** or the **68-95-99.7 rule**, tells you where most of your values lie in a normal distribution:

- Around 68% of values are within 1 standard deviation from the mean.
- Around 95% of values are within 2 standard deviations from the mean.
- Around 99.7% of values are within 3 standard deviations from the mean.

- SAT scores from students in a new test preparation course. The **data follows a normal distribution with a mean score (M) of 1150 and a standard deviation (SD) of 150.**





Following the empirical rule:

- Around **68% of scores are between 1,000 and 1,300**, 1 standard deviation above and below the mean.
- Around **95% of scores are between 850 and 1,450**, 2 standard deviations above and below the mean.
- Around **99.7% of scores are between 700 and 1,600**, 3 standard deviations above and below the mean.

APPLICATION
OF NORMAL DISTRBUTION

Statistical Inference:

- **Hypothesis Testing:** The normal distribution is used in tests such as the t-test, z-test, and ANOVA to determine if observed data significantly deviates from what is expected under a **null hypothesis.**
- **Confidence Intervals**: It helps in estimating the range within which a population parameter is likely to fall with a certain **level of confidence.**

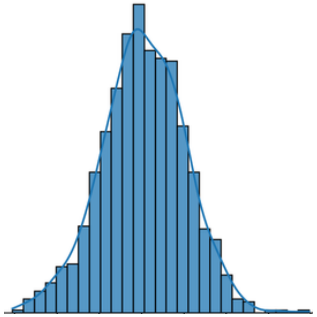
Machine Learning:

- **Feature Scaling**: In some algorithms (like **Principal Component Analysis (PCA) or Gaussian Naive Bayes)**, normalizing features to have a normal distribution can improve performance.
- **Data Modeling:** Algorithms that **assume normally distributed errors**, such as linear regression, can provide more accurate predictions if the data fits this assumption.
- **Data Transformation:** If data doesn't initially follow a normal distribution, **data scientists can often transform it (e.g., using logarithms) to approximate normality**, allowing them to apply powerful statistical tools more effectively.

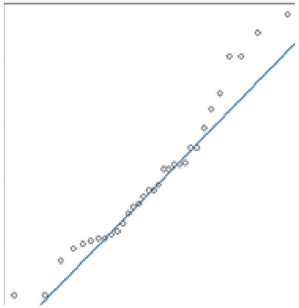
METHODS
TO CHECK NORMALITY

Visual Inspection of Plots/Charts:

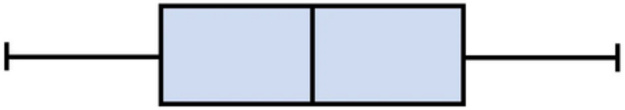
Histogram/KDE Plot : Look for a symmetric, bell-shaped curve.



Q-Q Plot (Quantile-Quantile Plot): Deviations from the line indicate deviations from normality.



Box Plot: A symmetric box plot with median close to the center of the box suggests normality.



METHODS
TO CHECK NORMALITY

Statistical Tests:

Shapiro-Wilk Test:

- Description: Tests the null hypothesis that the data is normally distributed.
- Interpretation: A p-value less than the chosen significance level (e.g., 0.05) indicates that the data significantly deviates from normality.

Kolmogorov-Smirnov Test:

- Description: Compares the sample distribution with a normal distribution.
- Interpretation: A small p-value indicates that the sample distribution differs significantly from a normal distribution.

Anderson-Darling Test:

- Description: Tests if a sample comes from a specific distribution, including the normal distribution.
- Interpretation: A p-value below the threshold suggests a deviation from normality.

Jarque-Bera Test:

- Description: Tests the skewness and kurtosis of the data to assess normality.
- Interpretation: A large test statistic or small p-value indicates that the data may not be normally distributed.

p-value, hypothesis testing ,confidence interval, anderson, shapiro ......
what are these man ????

- Things are getting complicated now ??? I Understand these things are complicated .

- Utilize our 2 buddies, ChatGPT and Google.

- In future I’ll cover all the concepts ,stay tuned.

THANK YOU

Share your thoughts and feedback !!

👍 🔁