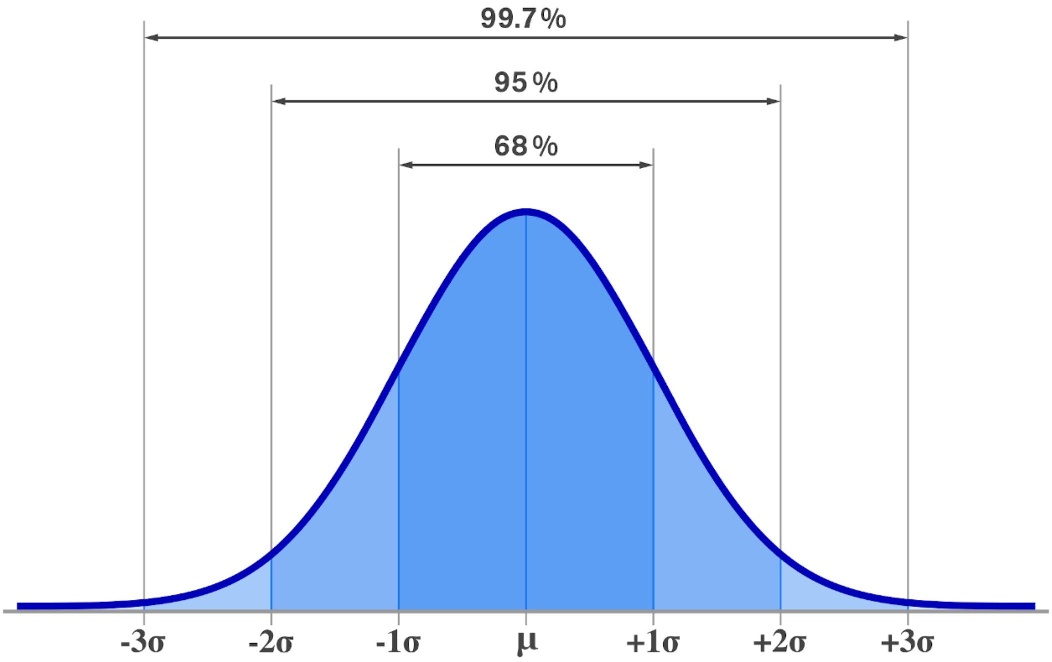
Statistics – Empirical Rule and Chebyshev Rule

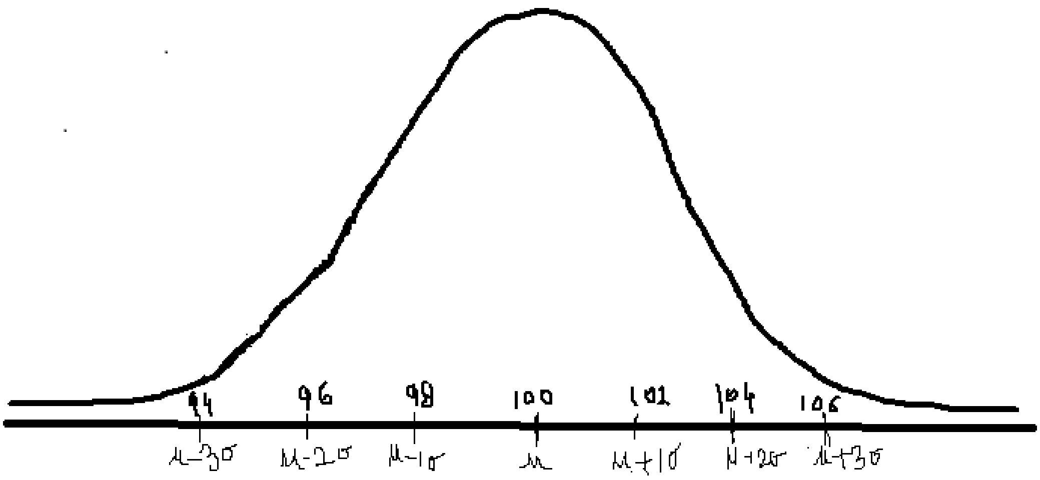
Empirical Rule:

- Empirical Rule is 68% - 95% - 99.7% rule
- It is a statistical rule that applies to normal distribution
- It is used to understand the distribution of data within standard deviations of the mean.
- We can understand it by below diagram.



- Standard Deviation = σ
- Mean = μ
- The above diagram follows normal distribution
- We have three observations from above diagram
- 1) **68% of the data** will cover between **μ - 1σ to μ + 1σ**
- 2) **95% of the data** will cover between **μ - 2σ to μ + 2σ**
- 3) **99.7% of the data** will cover between **μ - 3σ to μ + 3σ**
- The maximum data coverage will happen only in between: - **μ - 3σ to μ + 3σ** only.
- This rule is useful for quickly estimating the spread of data in a normal distribution and for identifying outliers.

- <u>Use Case:</u>
- In India, the average petrol rates are 100Rs and it varies state by state by 2Rs.
- Standard Deviation = σ = 2 Rs
- Mean = μ = 100 Rs



- For 68% of data = $\mu - 1\sigma$ to $\mu + 1\sigma$ = 100 – 1(2) to 100 + 1(2) **= 98 to 102**
- For 95% of data = $\mu - 2\sigma$ to $\mu + 2\sigma$ = 100 – 2(2) to 100 + 2(2) **= 96 to 104**
- For 99.7% of data = $\mu - 3\sigma$ to $\mu + 3\sigma$ = 100 – 3(2) to 100 + 3(2) **= 94 to 106**

- <u>Analysis:</u>
- There are 68% of states in India with petrol rates between 98 to 102
- There are 95% of states in India with petrol rates between 96 to 104
- There are 99.7% of states in India with petrol rates between 94 to 106

- Minimum petrol rates in India is 94Rs
- Maximum petrol rates in India is 106Rs
- If data does not follow normal distribution, then we will have to use another rule

Chebyshev's Inequality:

- Chebyshev's Inequality is a statistical theorem that applies to all data distributions, regardless of their shape
- It provides a lower bound on the proportion of observations that fall within a certain number of standard deviations from the mean.

- For Empirical Rule -> **μ ± kσ**
- Where k = 1,2,3,4

- For this Rule, **μ ± kσ, but the data percentage is = 1 - $\frac{1}{k^2}$**

- **For k = 1, μ ± 2σ, so data percentage will be zero**

- For k = 2, μ ± 2σ, so data percentage is
- = 1 - $\frac{1}{k^2}$ = 1 - $\frac{1}{2^2}$ = 1 - $\frac{1}{4}$ = $\frac{3}{4}$ = 75%

- **For 2 standard deviation data coverage will be 75%**

- For k = 3, μ ± 3σ, so data percentage is
- = 1 - $\frac{1}{k^2}$ = 1 - $\frac{1}{3^2}$ = 1 - $\frac{1}{9}$ = $\frac{8}{4}$ = 88.8 = 89%
- **For 3 standard deviation data coverage will be 89%**

- **For Chebyshev's inequality data coverage start from 2 standard deviation k>=2**

- Unlike the Empirical Rule, which is specific to normal distributions, Chebyshev's Inequality applies to any distribution, making it a more general but less precise tool.

Normal (Empirical) **(68-95-99.7)**	**Chebyshev's Inequality** $(1 - \frac{1}{k^2})$
μ - 1σ to μ + 1σ	**μ - 1σ to μ + 1σ: - 0% (Not Valid)**
μ - 2σ to μ + 2σ	**μ - 2σ to μ + 2σ: - 75%**
μ - 3σ to μ + 3σ	**μ - 3σ to μ + 3σ :- 89%**