Introduction to Statistics for DS

**Descriptive Statistics**

* Central tendency
* Dispersion / Spread
* Distribution of Variables
* Symmetry and Shape of Variables

Measure of Location (Central tendency):

Mean, Median, Mode, Percentiles, Quartiles.

Measure of Spread (Dispersion / Spread):

Variance, STD, Coefficient of Variance, Mean Absolute Error, Inter Quartile Range, Median Absolute Deviation from Median (MAD)

Measure of Symmetry:

Skewness

Measure of Shape:

Kurtosis

**Inferential Statistics**

* Strength of Association(correlation)
* Hypothesis testing

***Strength of Association:***

Pearson’s correlation

Spearman’s Rank Correlation

Cramer’s V

***Hypothesis Testing:***

Test for Normality:

Shapiro Wilk Test

Anderson-Darling Test

***Test for Association:***

Numeric Variables-

Pearson’s

Spearman’s

***Categorical Variables:***

Chi-Square Test

***Numeric – Categorical Variables:***

One-way ANOVA: In categorical column unique value is (>=3).

T-test: In categorical column unique value is 2.

Kruskal-Wallis Test

***Steps Involved in Hypothesis:***

H0(True): Null Hypothesis

H1(False): Alternate Hypothesis

Alpha: significance / strictness level (as usuall 5%)

Confidence Interveral: alpha - 1

P-value: Calculate the evidence against H0

If p-value <= alpha Rejected H0 ,else Fail to reject H0

NB: Popular technique to calculate p-value: T-test, Chisquare, Anova

**Plots**

**Univariate**

***Numeric***:

Histogram, KDE Plot, Rug Plot, Box Plot, Violin Plot, Q-Q Plot

***Categorical***:

Count Plot, Pie Plot

***Time Related:***

Line Plot, Aggregated Line Plot

**Bivariate**

***Numeric vs Numeric***

Scatter Plot, Hexagonal Bin Plot, Contour Density Plot

***Numeric vs Categorical:***

Bar Plot, Box Plot, KDE Plot, Violin Plot

***Categorical vs Categorical:***

Bar Plot, Stacked Bar Plot, Frequency Heatmap

**Multivariate**

Pair Plot

Correlation Heatmap:

Pearson, Spearman’s Rank, Cramer’s V

Facet Grid (Seaborn)