Math Basics for Data science

How to ?

what things mean in data science – Terminology, purpose, methods and when to use

Descriptive Statistics

- Descriptive Statistics : Summarizes and describes the main features of the dataset
- Methods : mean, median, mode, range, variance, standard deviation, quartiles, interquartile range, histograms, box plots, frequency tables
- When to use ? To get basic characteristics of data to :
  - Identify outliers and visualize distributions

Inferential Statistics

Purpose : To make inferences about a "population" based on a sample

ex: I have 1200000 people information. I will select a subset called sample – 80000 -> Analyzing these will let me understand and generically describe the rest of the 1200000 population

Methods : Hypothesis testing – t-test, ANOVA, chi-square test

Confidence Intervals, correlation analysis, regression analysis

When to use : When we want to generalize findings from a smaller to a larger sample or "population", test hypothesis or model relationships between variables

SampleID	Morning First thing	Breakfast Time
1	Brush	9
2	Have Coffee	8
3	Brush	8
1200000	Don't brush at all	10

30% of Americans have high vulnerability to Diabetes

40% of Americans have High Cholesterol and Blood Pressure.
so use the FDA approved some drug that makes it worse

Probability - Quantify and Qualify

Quantify what ? – Uncertainty and makes predictions about future events

Methods used in Probability – Probability Distributions – normal, binomial, Poisson, conditional probability, Bayes Theorem

When to use this : When you want to understand or assess the events , make predictions or calculate risk.

Hypothesis Testing

- Purpose : Tests a claim or hypothesis about a population (Sample)
- Methods used : T-Test, ANOVA, chi-square, Z-test
- When to use : When there is a need to get a significant difference between groups or if there is a relationship between variables

30% of Americans have high vulnerability to Diabetes

40% of Americans have High Cholesterol and Blood Pressure. so use the FDA approved some drug that makes it worse

60% of the worlds Almonds come from California.

California Had the worst summer in 100 years with a severe water crisis.,
Cutting trees to make concrete jungles
increase in the conflict between humans and nature
global warming

FirstThing in the Morning vs Age group

Regression Analysis

- Models the relationship between a dependent variable and one or more independent variables

- Age Group -> Spend, Income

- Spend -> Age Group, Income

- Customers-> Transaction Amount by City

- When to use : To predict a numerical value based on other variables, understand the impact of variables on an outcome (Result) or model relationship between variables

Clustering

- Group similar data points together
- Methods : K-means, hierarchcial, DBSCAN (eps, Minpts)
- When to use: To Discover natural groupings in data, identify patterns or segmentation

Distance Between Points : K-means, Mean shift, Affinity propagation

DBSCAN – Distance Between Nearest points)

Gaussian Mixtures : Mahalanobis distance to centers, Spectral Clustering (graph distance)

Mahalanobis distance is the distance of the test point from the center of the mass divided by the width of the ellipsoid in the direction of the test point

Dimensionality Reduction

- Why : Reduce the number of variables in a dataset while keeping important information
- Methods : Principal Component Analysis, t-SNE, factor analysis

- When to use : Large number of variables and need to simplify analysis, visualization or better model performance.

Principal Component Analysis

- Lineraly transforms data onto a new co-ordinate system to identify directions that have the most variation in data

Regression Analysis Step by Step

- - To estimate the level effect on an independent variable on dependent variable – x, y => x = age y=>spending or education
- formula for regression
  - Y = a + bx , a and b are parameters and they remain constant as x and y change
  - a is the intercept and B is the slope
  - if we know a,b we can calculate value of Y for a given value of x

Age, spend – Regression

	Age	Spend	AGE*Spend	Age(2)	Spend(2)
	45	50000	220000	2025	2500000
Sigma	45	50000	220000	2025	2500000

b0 = Sig(Y)(Sig(X^{2}) – Sig(X)(Sig(XY)/n(SigX^{2})-Sig(X)^{2}

b1 =n Sig(XY) – Sig(X)Sig(Y) / n(SigX^{2})-Sig(X)^{2}

y = b0 + b1*x