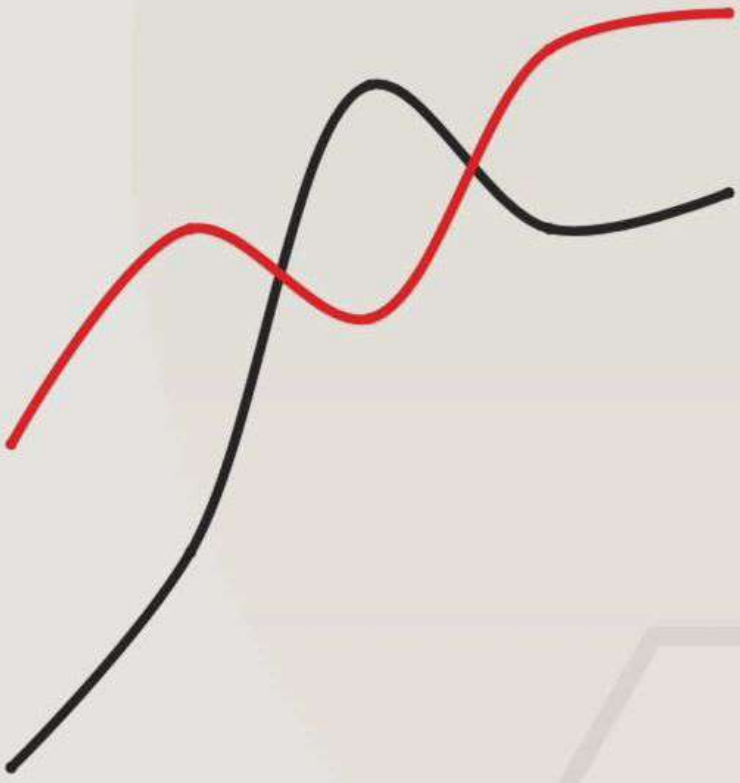
PROBABILITY FOR DATA SCIENCE



PROBABILITY FOR DATA SCIENCE

1. Basic Probability Concepts:

1. **Probability:** A measure of the likelihood of an event occurring, ranging from 0 to 1.
2. **Sample Space:** The set of all possible outcomes in a probability experiment.
3. **Event:** Any subset of outcomes from the sample space.
4. **Complementary Events:** The event that an event does not occur (denoted as A').
5. **Union of Events:** Probability that at least one of the events occurs (denoted as $P(A \cup B)$).
6. **Intersection of Events:** Probability that both events occur (denoted as $P(A \cap B)$).

2. Conditional Probability:

1. **Conditional Probability:** Probability of an event occurring given that another event has occurred (denoted $P(A \mid B)$).
2. **Bayes' Theorem: A method for updating the probability of a hypothesis based on new evidence.**
   $P(A \mid B) = P(B \mid A)P(A)/P(B)$

PROBABILITY FOR DATA SCIENCE

3. Random Variables:

1. **Random Variable:** A variable whose possible values are outcomes of a random phenomenon.
2. **Discrete Random Variable:** A random variable with countable outcomes.
3. **Continuous Random Variable:** A random variable with infinite possible outcomes within a range.
4. **Probability Distribution:** A function that provides the probabilities of occurrence of different possible outcomes for a random variable.
5. **Cumulative Distribution Function (CDF):** The probability that a random variable is less than or equal to a certain value.

4. Probability Distributions:

1. **Binomial Distribution:** Models the number of successes in a fixed number of independent trials (for discrete data).
2. **Normal Distribution:** A continuous probability distribution characterized by its bell shape, defined by mean (μ) and standard deviation (σ).
3. **Poisson Distribution:** Models the number of events occurring in a fixed interval of time/space, typically for rare events.
4. **Uniform Distribution:** All outcomes are equally likely, used for discrete or continuous data.
5. **Exponential Distribution:** Describes the time between events in a Poisson process.

PROBABILITY FOR DATA SCIENCE

5. Descriptive Statistics:

1. **Mean:** The average of a set of numbers.
2. **Median:** The middle value when data is sorted in order.
3. **Mode:** The most frequent value in the data set.
4. **Variance:** Measures how much data points deviate from the mean.
5. **Standard Deviation:** The square root of variance, representing spread or dispersion in the data.

6. Law of Large Numbers & CTL

1. **Law of Large Numbers:** States that as sample size increases, the sample mean will approach the population mean.
2. **Central Limit Theorem (CLT):** States that the distribution of the sample mean will approach a normal distribution as sample size increases, regardless of the population distribution.

7. Sampling & Estimation:

1. **Sampling Distribution:** The probability distribution of a sample statistic (e.g., sample mean).
2. **Point Estimation:** The use of sample data to estimate the value of an unknown population parameter.
3. **Confidence Interval:** A range of values, derived from the sample data, that is used to estimate a population parameter with a certain confidence level.

PROBABILITY FOR DATA SCIENCE

8. Correlation and Regression:

1. **Correlation:** A measure of the relationship between two variables, ranging from -1 (perfect inverse) to 1 (perfect positive).
2. **Covariance:** Measures the degree to which two variables change together.
3. **Linear Regression:** A method to model the relationship between a dependent variable and one or more independent variables.

9. Hypothesis Testing:

1. **Null Hypothesis (H0):** A statement asserting that there is no effect or difference.
2. **Alternative Hypothesis (HA):** A statement asserting that there is an effect or difference.
3. **p-value:** The probability of obtaining a test result at least as extreme as the one observed, assuming H0 is true.
4. **Test Statistic:** A value used to decide whether to reject H0, often derived from sample data.
5. **Significance Level (α):** The threshold for rejecting H0, typically 0.05.
6. **Confidence Interval:** A range of values within which the population parameter is expected to lie, with a given confidence level.

PROBABILITY FOR DATA SCIENCE

10. Common Statistical Tests:

1. **Z-Test:** A test for comparing the sample mean to the population mean when the population variance is known (large sample size).
2. **T-Test:** A test for comparing sample means when the population variance is unknown (small sample size).
3. **Chi-Square Test:** A test for the relationship between categorical variables or comparing observed frequencies with expected frequencies.
4. **ANOVA (Analysis of Variance):** A test for comparing the means of three or more groups.

11. Markov Chains:

1. **Markov Process:** A random process where the future state depends only on the current state and not on previous states.
2. **Transition Matrix:** A matrix representing the probabilities of moving from one state to another in a Markov process.

12. Monte Carlo Simulation:

1. **Monte Carlo Simulation:** A technique for estimating the probability of different outcomes using random sampling.

13. Bayes' Statistics:

1. **Prior Probability:** The initial probability of an event before new evidence is observed.
2. **Posterior Probability:** The updated probability of an event after observing new evidence.