

# Probability

## OVERVIEW & PURPOSE

In this session, participants will have an Introduction to probability for data sciences

## OBJECTIVE

- Introduction to Probability
- Basic Probability Rules
- Conditional Probability and Independence
- Discrete Probability Distribution
- Continuous Probability Distributions
- Central Limit Theorem
- Descriptive Statistics and Data Visualization
- Applications in Data Science

## Introduction to Probability:

**Definition:** Probability quantifies the uncertainty or variability of outcomes. Mathematically, it's the ratio of favorable outcomes to all possible outcomes.

**Example:** If you flip a fair coin, the probability of getting heads is 0.5 because there is 1 favorable outcome (heads) and 2 possible outcomes (heads or tails).

**Importance:** In data science, we deal with uncertainty often. Probability provides a foundation to model, measure, and predict uncertain events.

### **Basic Terminology:**

**Outcome:** A single result of an experiment.

**Event:** One or more outcomes of interest.

**Sample Space:** The set of all possible outcomes.

**Example:** If you roll a die, the sample space is  $\{1,2,3,4,5,6\}$ . Getting an even number is an event, with outcomes  $\{2,4,6\}$ .

**Importance:** To correctly model uncertainty, you need to understand the elements you are working with and how they relate to each other.

**Fundamental Principle:** The probability of any event lies between 0 (impossible event) and 1 (certain event).

**Importance:** This helps ensure that our probability values make logical sense and keeps our predictions and models grounded.

## **Basic Probability Rules:**

**Addition Rule:**

$$P(A \cup B) =$$

$$P(A) + P(B) - P(A \cap B)$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

**Example:** Probability of drawing a heart or a king from a deck of cards.

**Importance:** Helps determine combined probabilities, crucial for evaluating multiple events or conditions.

**Multiplication Rule:**

$$P(A \cap B) = P(A) \cdot P(B | A)$$

$$P(A \cap B) = P(A) \times P(B | A)$$

**Example:** Probability of drawing two aces in a row from a deck.

**Importance:** Essential when dealing with dependent events in sequences or processes.

**Complementary Events:**

The complement of  $A$ , denoted by  $A'$ ,  $A$  or  $A^c$ , consists of all the outcomes in which the event  $A$  does not occur.

$$P(A) + P(A') = 1$$

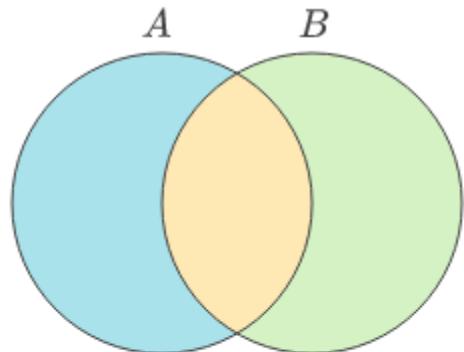
$$P(A) = 1 - P(A')$$

$$P(A') = 1 - P(A)$$

**Importance:** Useful when it's easier to calculate the probability of something not happening.

## Conditional Probability and Independence:

### Conditional Probability:



- $P(A)$
- $P(B)$
- $P(A \cap B)$

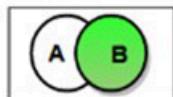
Conditional Probability Formula

$$P(A | B) = \frac{P(A \cap B)}{P(B)}$$

Probability that  $A$  occurs given that  $B$  has already occurred

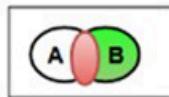
**Example:** Probability of someone being a smoker given that they have lung cancer.

$$P(A|B) = P(\text{A given B has occurred})$$



If B has already occurred  
then our sample space  
must be somewhere within B

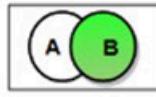
Now A can occur only  
within sample space B



P(A|B) is the ratio of Red  
space divided by Green space

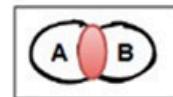
$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(B|A) = P(\text{B given A has occurred})$$



If A has already occurred  
then our sample space  
must be somewhere within A

Now B can occur only  
within sample space A



P(B|A) is the ratio of Red  
space divided by White space

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

**Importance:** Key for understanding dependencies and relationships between events.

**Bayes' Theorem:**

$$P(A|B) = \frac{(P(B|A) * P(A))}{P(B)}$$

# Bayes' Theorem

[bās 'thē-ə-rəm]

A mathematical formula  
for determining  
conditional probability.

**Importance:** Essential in many machine learning algorithms and for updating probabilities with new information.

**Independence:** If

### Probability of Compound Events

#### Independent Events

$$P(A \text{ and } B) = P(A) \times P(B)$$

#### Dependent Events

$$P(A \text{ and } B) = P(A) \times P(B | A)$$

#### Mutually Exclusive

$$P(A \text{ or } B) = P(A) + P(B)$$

#### Mutually Inclusive

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

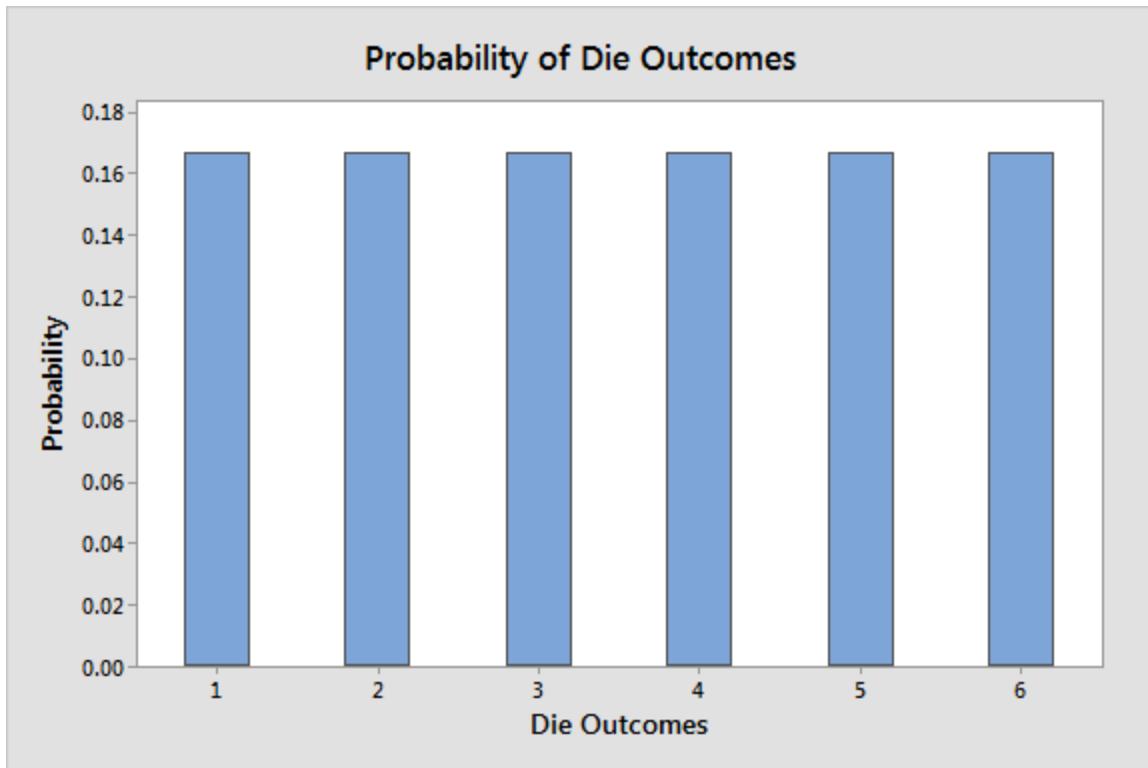
$P(A \cap B) = P(A) \times P(B)$ , A and B are independent.

- Example:
- Tossing a coin and rolling a number cube are independent events.

**Importance:** Helps simplify complex probability calculations and understand relationships.

## Discrete Probability Distributions:

**Uniform Distribution:** Each outcome is equally likely.



**Importance:** Serves as a baseline model for comparison.

### Binomial Distribution:

$$P(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{(n-x)!x!} p^x q^{n-x}$$

where

$n$  = the number of trials (or the number being sampled)

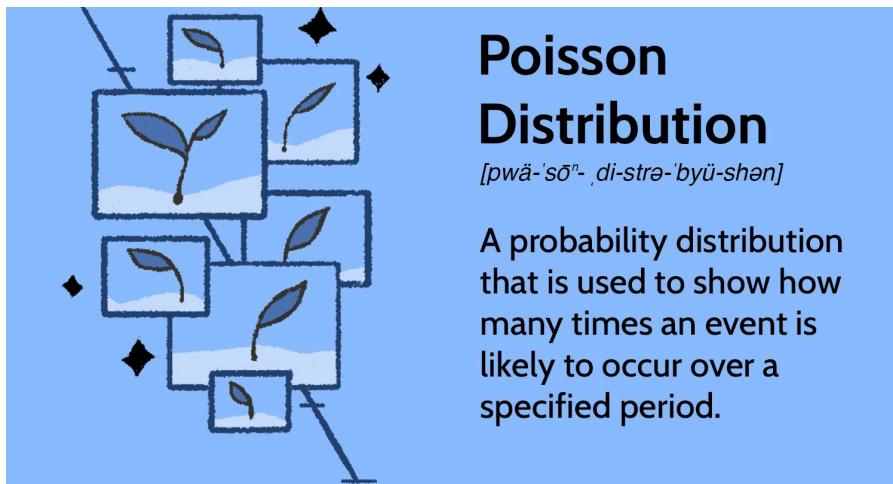
$x$  = the number of successes desired

$p$  = probability of getting a success in one trial

$q = 1 - p$  = the probability of getting a failure in one trial

**Importance:** Useful in situations with two outcomes, like pass/fail, win/lose.

### Poisson Distribution:



### Poisson Distribution

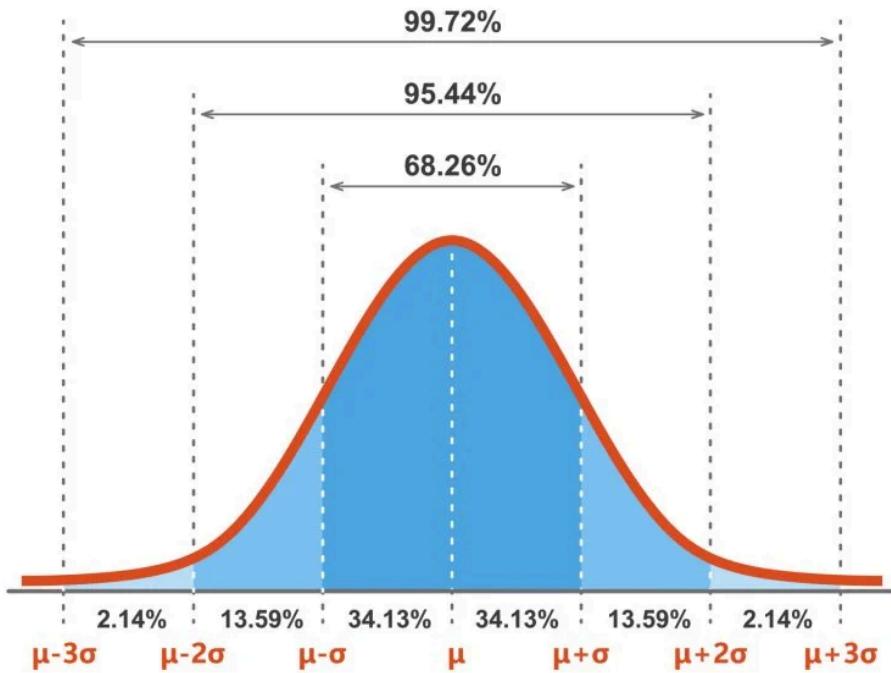
[pwä-'sōn-, di-strə-'byü-shən]

A probability distribution that is used to show how many times an event is likely to occur over a specified period.

**Importance:** Great for modeling rare events over time or space.

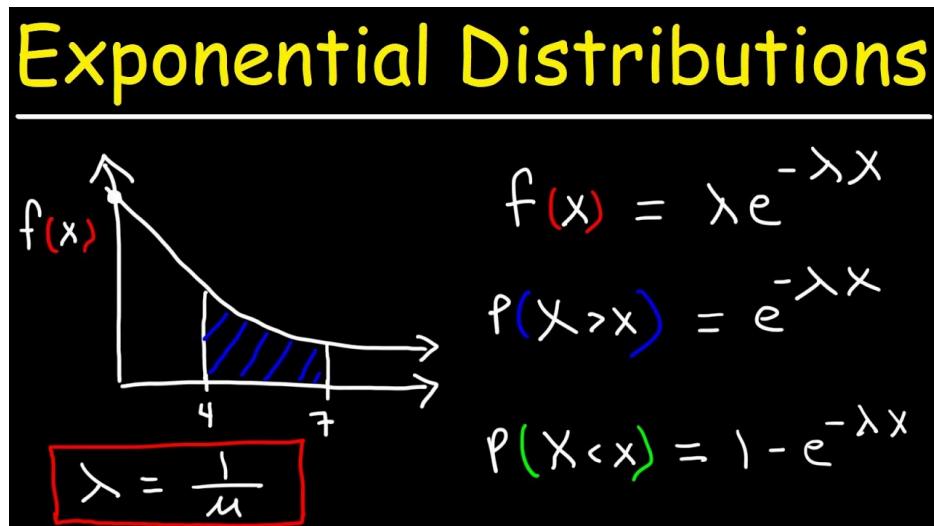
## Continuous Probability Distributions:

**Normal Distribution:** Characterized by mean  $\mu$  and standard deviation  $\sigma$ .



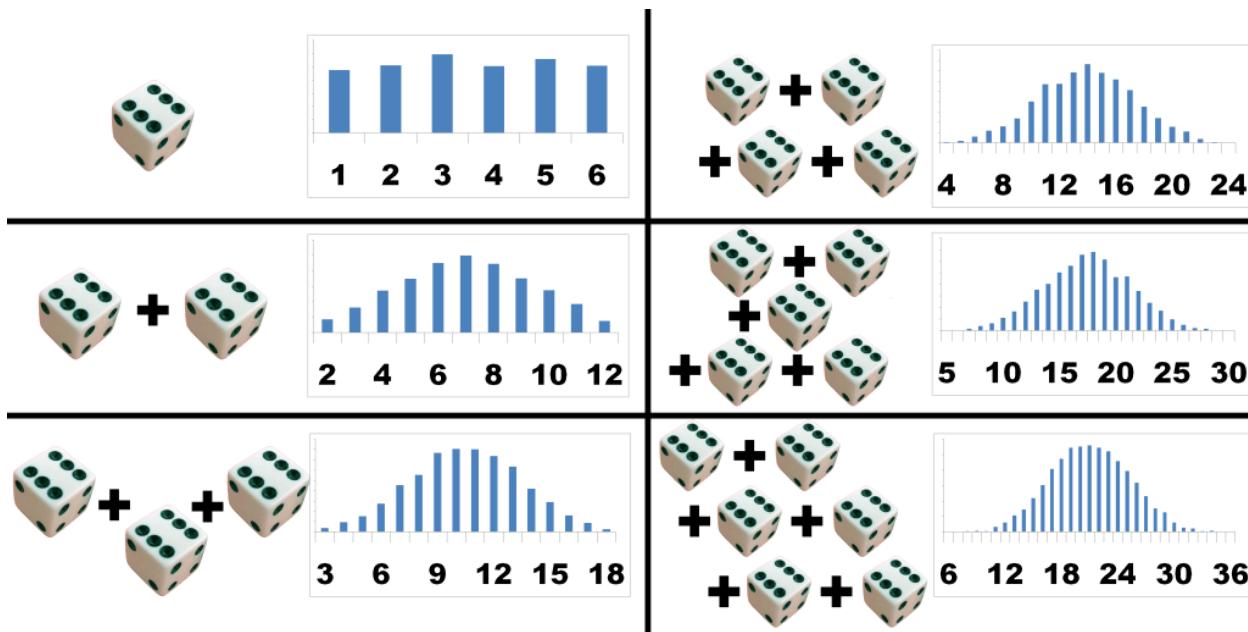
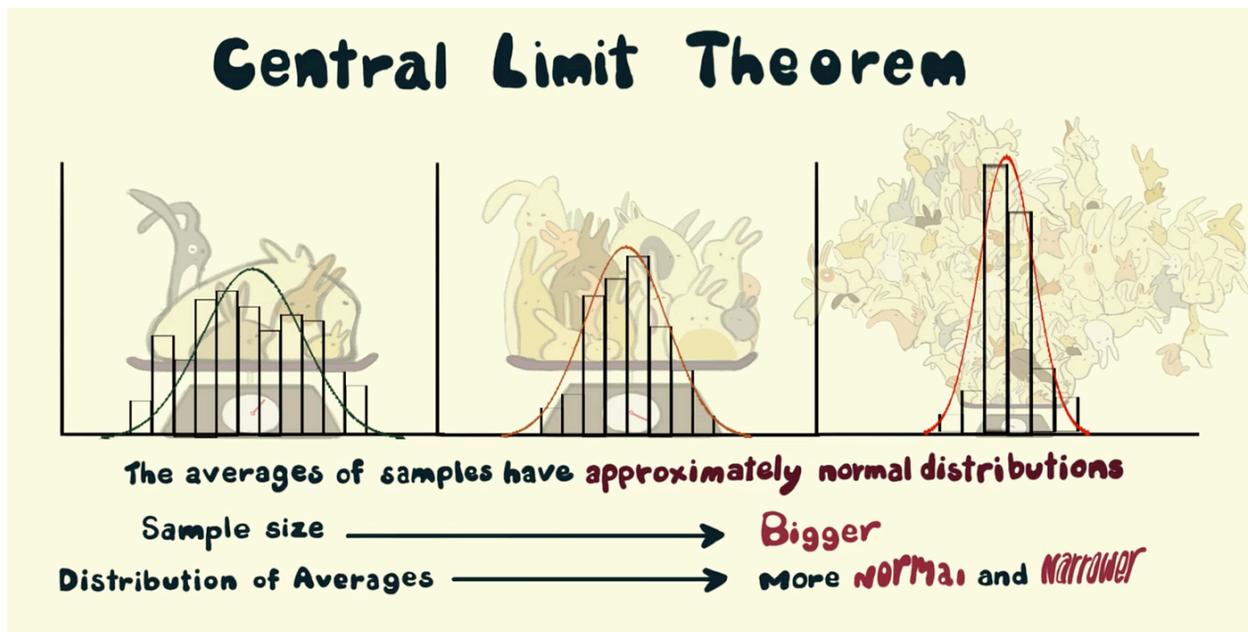
**Importance:** The foundation for many statistical methods and hypothesis tests.

**Exponential Distribution:** Models time between events.

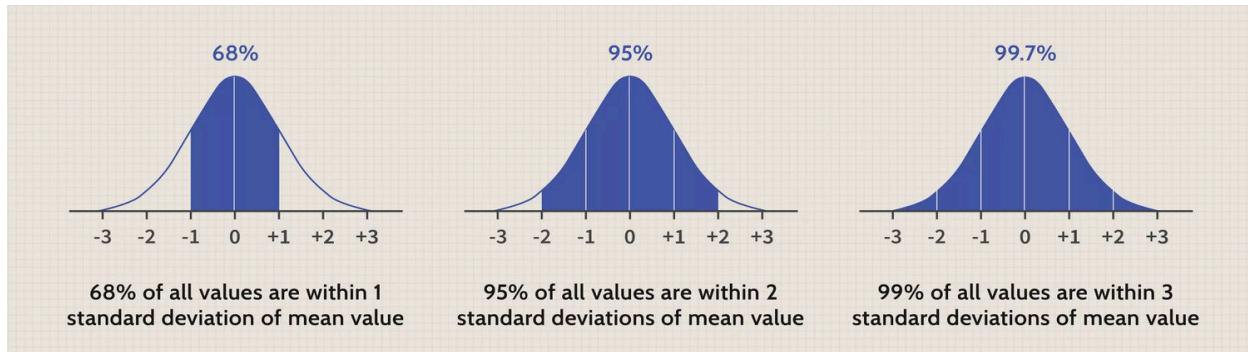


**Importance:** Useful for modeling lifetimes or decay.

## Central Limit Theorem:



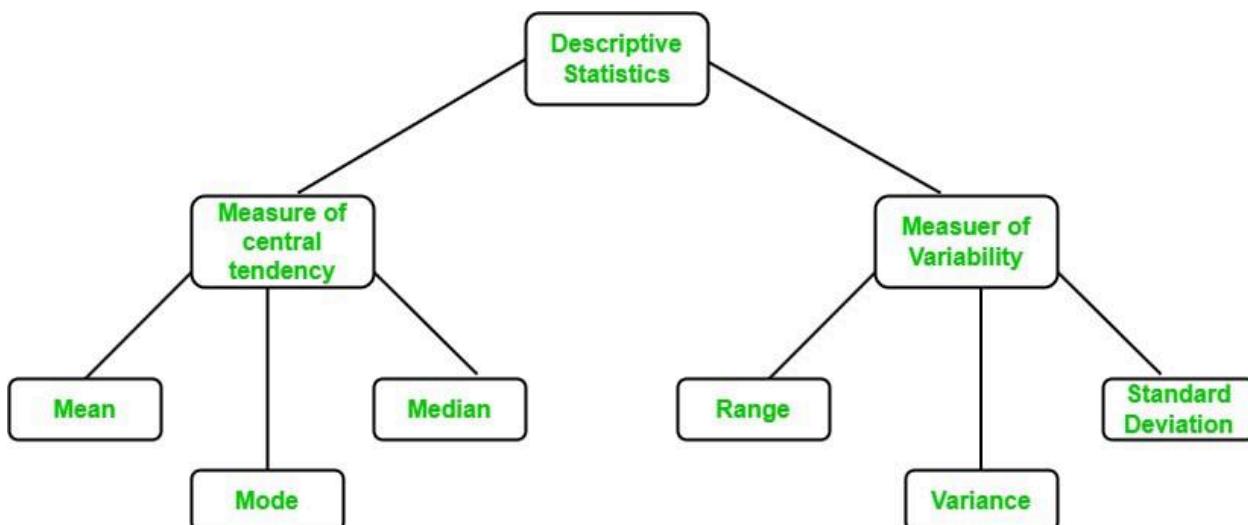
Even if initial data isn't normal, the distribution of sample means becomes more normal as sample size increases.



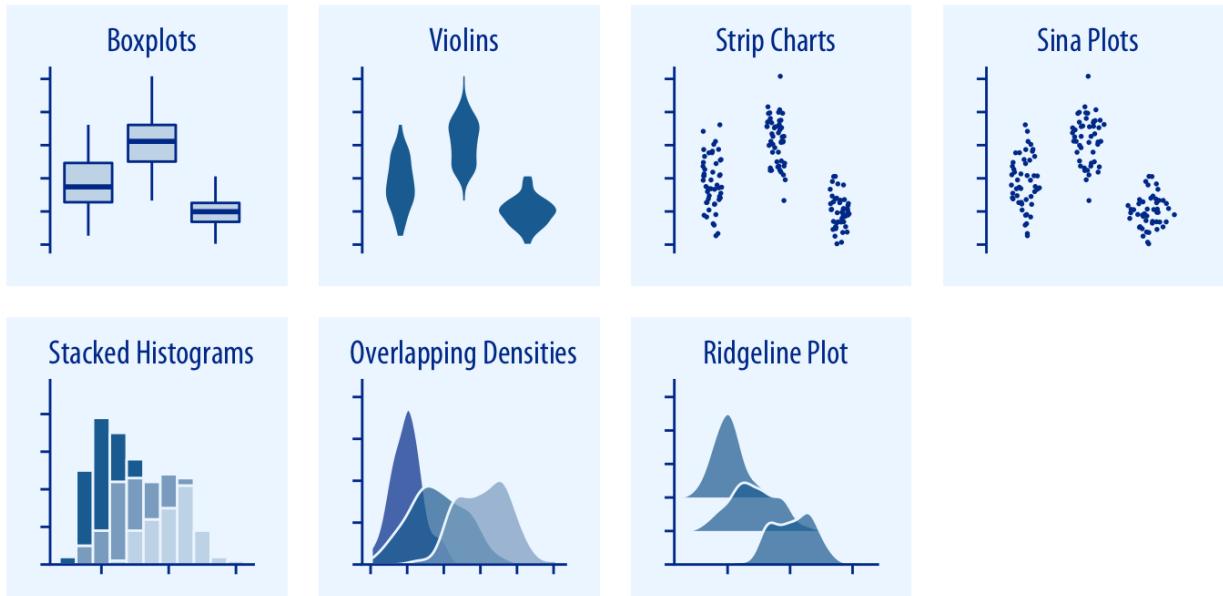
**Importance:** Underpins many statistical methods and hypothesis tests.

## Descriptive Statistics and Data Visualization:

**Metrics** like mean, median, mode, variance, and standard deviation provide summary statistics.



**Histograms and Box Plots** help visualize and understand data distributions.



**Importance:** Provides initial insights, helps spot trends, outliers, and anomalies.

## Applications in Data Science:

**Hypothesis Testing:** Test assumptions or hypotheses about a dataset.

**Importance:** Crucial for making informed decisions based on data.

**Regression Analysis:** Understand relationships and predict outcomes.

**Importance:** Core technique in predictive modeling.

**Bayesian Thinking:** Combine prior knowledge with observed data.

**Importance:** Offers a flexible framework for modeling and prediction, especially with limited data.

Understanding these concepts in depth provides a strong foundation for anyone diving into the realm of data science and statistical analysis.