PROBABILITY

Course: Foundations of Statistics for Data Analytics and Machine Learning Using Excel

@ **DV Data Analytics, Bangalore**

Instructor:
Dr. Ujwal Deep Kadiyam

Counting

Counting

- Counting is a fundamental concept in combinatorics.
- Helps determine possible outcomes, optimize decisions, and solve real-world problems.
- Two key methods: **Permutations** and **Combinations**.
- Used in probability, statistics, Engineering and computer science.

The Counting Principle

- It states that if there are m ways to perform one action and n ways to perform another, then there are $m \times n$ ways to perform **both** actions.
- This principle is used to calculate the total number of possibilities when multiple events or choices are involved.
- Extends to multiple events: $m_1 \times m_2 \times \cdots \times m_k$.

Problem: A restaurant offers 3 choices of starters and 4 choices of main courses. How many different meal combinations can be made?

- Each appetizer choice can be paired with any main course choice.
- Using the multiplication principle there are $12(3 \times 4)$ different meal combinations.

Example: Counting Principle (Multiplication Principle)

Problem: How many different passwords can be created if a password consists of 3 letters (A-Z) followed by 2 digits (0-9)?

- Each letter has 26 choices.
- Each digit has 10 choices.
- Since each choice is independent, we multiply:

$$26 \times 26 \times 26 \times 10 \times 10 = 26^3 \times 10^2$$

$$= 17,576 \times 100 = 1,757,600$$

Answer: There are 1,757,600 possible passwords.

Rules for Counting in Sets

Addition Principle:

If event A can occur in m ways and event B (which is independent of A) can occur in n ways, then the number of ways **either** event can occur is:

$$m + n$$

Example: A student should select one project and if that student can choose a project from two different categories:

- 3 AI projects
- 2 Robotics projects

Since the student cannot choose from both categories at the same time, the total number of choices is:

$$3 + 2 = 5$$

Limitations of addition and Multiplication rule: Choosing a Committee with Restrictions

A school is forming a 2-member committee from a pool of 4 students:

$$A, B, C, D$$

- Using Multiplication rule, we will get $4 \times 3 = 12$ different committees.
- The following are the committees AB, AC, AD, BC, BD, CD.
- Multiplication Rule Fails.

Incorrect Counting Example

Three officers—a president, a treasurer, and a secretary—are to be chosen from four people: Ann, Bob, Cyd, and Dan.

- Ann cannot be president.
- Either Cyd or Dan must be secretary.

A mistaken approach:

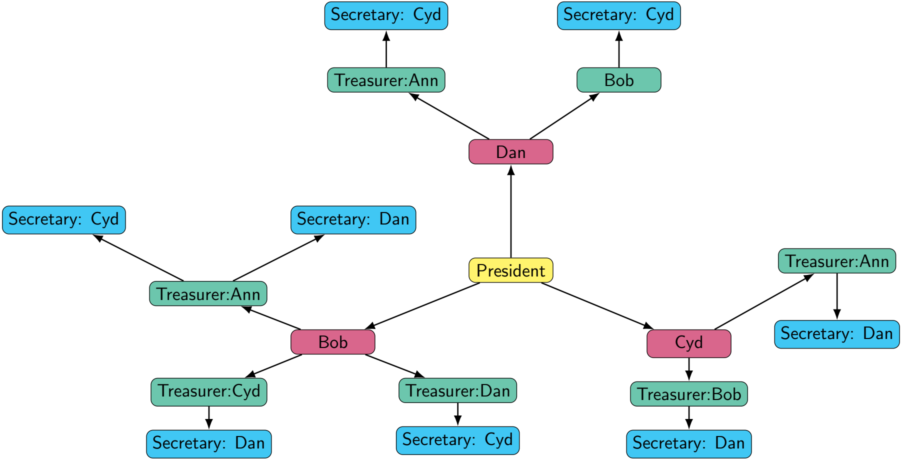
1. Choose a president from $\{$Bob, Cyd, Dan$\} \to 3$ choices.
2. Choose a treasurer from the remaining 3 people $\to 3$ choices.
3. Choose a secretary from $\{$Cyd, Dan$\} \to 2$ choices.

Applying the multiplication principle incorrectly:

$$3 \times 3 \times 2 = 18$$

This answer is incorrect because choosing a treasurer should be from only 2 remaining candidates.

Officer Selection



Permutations

Permutations

Definition: The number of ways to arrange r objects from n distinct objects.

$$^{n}P_r = P(n,r) = \frac{n!}{(n-r)!}$$

- Order matters.
- Example: Arranging 2 books from a set of 4.

All permutations of A,B,C,D:

- AB
- AC
- AD
- BA
- BC
- BD
- CA
- CB
- CD
- DA
- DB
- DC

$$P(5,3) = \frac{5!}{3! \cdot 2!} = \frac{120}{6 \cdot 2} = \boxed{10}$$

DVAnalytics Transforming You

Permutations with Repetition

Definition: The number of ways to arrange r objects from n objects when repetition is allowed.

$$P_{rep}(n, r) = n^r$$

- Order matters.
- Objects can be repeated.
- Example: Creating a 4-digit PIN using 10 digits (0-9):

$$P_{rep}(10, 4) = 10^4 = 10,000$$

Combinations

Combinations

Definition: The number of ways to choose r objects from n distinct objects without considering order.

$$^{n}C_r = C(n, r) = \binom{n}{r} = \frac{n!}{r!(n-r)!}$$

- Order does not matter.
- Example: Selecting 3 students from a group of 5.

All possible selections:

- ABC
- ABD
- ABE
- ACD
- ACE
- ADE
- BCD
- BCE
- BDE
- CDE

$$C(5, 3) = \frac{5!}{3! \cdot 2!} = \frac{120}{6 \cdot 2} = \boxed{10}$$

Combinations with Repetition

Definition: The number of ways to choose r objects from n distinct objects when repetition is allowed.

$$C_{rep}(n, r) = \binom{n+r-1}{r} = \frac{(n+r-1)!}{r!(n-1)!}$$

- Order does not matter.
- Objects can be repeated.
- Example: Choosing 3 scoops of ice cream from 5 flavors (where repeats are allowed):

$$C_{rep}(5, 3) = \binom{5+3-1}{3} = \binom{7}{3} = 35$$

Key Differences

- Permutations consider order; combinations do not.
- $P(n, r) > C(n, r)$ for $r > 1$.
- Use permutations when arranging, and combinations when selecting.

Example: Counting

Problem: A password consists of 4 distinct letters chosen from the English alphabet (26 letters). How many such passwords are possible?

Example: Counting

Problem: A password consists of 4 distinct letters chosen from the English alphabet (26 letters). How many such passwords are possible?

- Since order matters, we use permutations:

$$P(26,4) = \frac{26!}{(26-4)!} = \frac{26!}{22!}$$

- Expanding the factorial:

$$26 \times 25 \times 24 \times 23 = 358800$$

- Thus, there are **358,800** possible passwords.

Summary Table with Excel Functions

	Formula	Without Repetition	With Repetition
Permutations	$P(n, r)$	$\frac{n!}{(n-r)!}$	n^r
Excel Formula		`PERMUT(n, r)`	`PERMUTATIONA(n, r)`
Combinations	$C(n, r)$	$\frac{n!}{r!(n-r)!}$	$\frac{(n+r-1)!}{r!(n-1)!}$
Excel Formula		`COMBIN(n, r)`	`COMBINA(n, r)`

Excel Functions for Counting and Probability

Function	Description	Example	Output
`RAND()`	Generates a random number between 0 and 1	`=RAND()`	0.423 (varies)
`RANDBETWEEN`	Generates a random integer between two bounds	`=RANDBETWEEN(1, 6)`	4 (varies)
`FACT`	Returns the factorial of a number	`=FACT(5)`	120
`PERMUT`	Returns the number of permutations (order matters)	`=PERMUT(5, 3)`	60
`PERMUTATIONA`	Returns the number of permutations allowing repetitions	`=PERMUTATIONA(5, 3)`	125
`COMBIN`	Returns the number of combinations (order does not matter)	`=COMBIN(5, 3)`	10
`COMBINA`	Returns the number of combinations allowing repetitions	`=COMBINA(5, 3)`	35
`ROUND`	Rounds a number to a specified number of digits	`=ROUND(0.123456, 2)`	0.12

Probability

Introduction to Probability

- **Probability** of an event is a measure (number) of the chance with which we can expect the event to occur.
- It is quantified as a number between 0 and 1.
- Defined as:

$$P(A) = \frac{\text{Number of favorable outcomes}}{\text{Total number of outcomes}}$$

- Probability values range from 0 (impossible event) to 1 (certain event).

Basic Concepts:

- **Outcome:** A single possible result of a random experiment. It is an element of the sample space S.
- **Sample Space** (S): The set of all possible outcomes.
- **Event** (E): A subset of the sample space.
- **Probability of an event**: $P(E) = \frac{|E|}{|S|}$ (for equally likely outcomes).

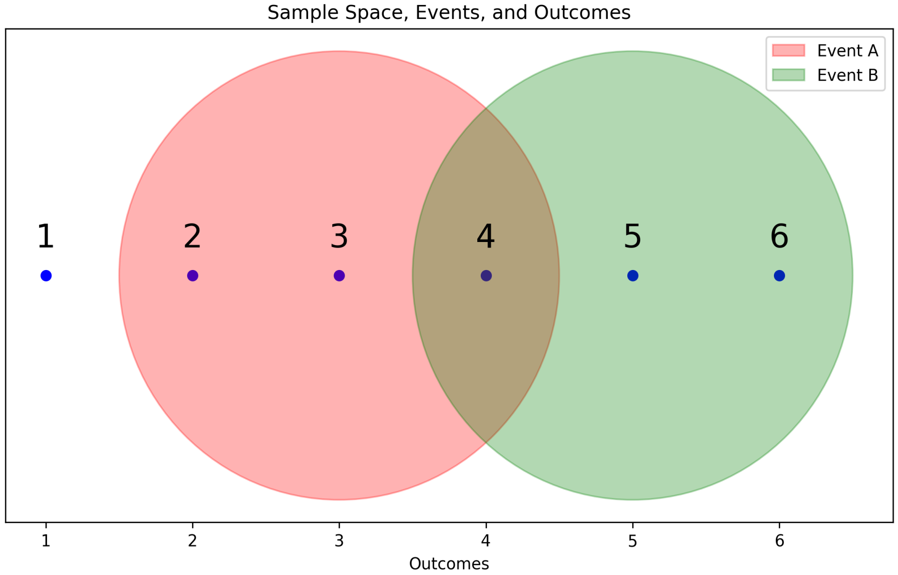
Example: Rolling a Die

- Sample space: $S = \{1, 2, 3, 4, 5, 6\}$.
- Outcome: Rolling a 3 (i.e., outcome $= 3$).
- Event A: Rolling an even number, $A = \{2, 4, 6\}$.

Key Difference:

- An outcome is a single result.
- An event consists of multiple possible outcomes.

Outcome, Sample Space and Event



Types of Probability: Classical

Classical Probability (Theoretical Probability)

- Based on equally likely outcomes.
- Formula: $P(E) = \frac{M}{N}$, where M is the number of favorable outcomes and N is the total outcomes.
- **Example:** Rolling a fair six-sided die. The probability of rolling a 4:

$$P(4) = \frac{1}{6}$$

Types of Probability: Empirical

Empirical Probability (Experimental Probability)

- Based on observed data from experiments.
- Formula: $P(E) = \frac{\text{Number of times } E \text{ occurs}}{\text{Total number of trials}}$.
- **Example:** If a coin is flipped 100 times and heads appear 55 times:

$$P(H) = \frac{55}{100} = 0.55$$

Types of Probability: Subjective

Subjective Probability

- Based on intuition, personal judgment, or expert opinions.
- Used when no historical data is available.
- **Example:** A doctor estimates a 70% chance of recovery for a patient based on symptoms and experience.

When Does an Event Occur?

Question: If one outcome in an event happens, does the event occur?

Answer: Yes! If at least one outcome from the event occurs, then the event has occurred.

Example: Rolling a Die

- Event A: Rolling an even number, $A = \{2, 4, 6\}$.
- If the outcome is 4, then since $4 \in A$, event A occurs.
- If the outcome is 3, then since $3 \notin A$, event A does not occur.

Conclusion: An event occurs if any of its outcomes happens.

Why Permutations and Combinations in Probability?

- Probability is about comparing:

$$\text{Probability} = \frac{\text{Number of favorable outcomes}}{\text{Total number of possible outcomes}}$$

- To find these counts, we often need:
  - **Permutations** – when *order matters*
  - **Combinations** – when *order doesn't matter*

Example 1: Permutation (Order Matters)

Problem: A password consists of 3 different digits from 0–9. What is the probability that you guess it correctly on the first try?

Solution:

- Total outcomes = Number of ways to arrange 3 digits from 10:

$$^{10}P_3 = \frac{10!}{(10-3)!} = 720$$

- Favorable outcomes = 1 (only one correct password)

$$\text{Probability} = \frac{1}{720}$$

Example 2: Combination (Order Doesn't Matter)

Problem: From a group of 8 students, a committee of 3 is selected. What is the probability that a specific group of 3 students is chosen?

Solution:

- Total combinations:

$$\binom{8}{3} = \frac{8!}{3! \cdot 5!} = 56$$

- Favorable outcomes $= 1$ (only one specific group)

$$\text{Probability} = \frac{1}{56}$$

DV Analytics
Transforming You

Axioms and Rules of Probability

Axioms of Probability

Given a sample space S and an event E, probability satisfies:

1. Non-negativity: $P(E) \geq 0$ for all events E.
2. Normalization: $P(S) = 1$ (Sum of probabilities of all events = 1).
3. Additivity: If $E_1, E_2, \ldots$ are disjoint events, then $P(E_1 \cup E_2 \cup \ldots) = P(E_1) + P(E_2) + \ldots$

Example: Rolling a fair die

- $S = \{1, 2, 3, 4, 5, 6\}$
- $P(\text{rolling a 2}) = \frac{1}{6}$
- $P(\text{rolling an even number}) = P(2) + P(4) + P(6) = \frac{3}{6} = \frac{1}{2}$

Basic Probability Rules

- **Addition Rule:** If events A and B are mutually exclusive:

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

- **Multiplication Rule:** If events A and B are independent:

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

- **Complement Rule:** The probability of an event not occurring:

$$P(A^c) = 1 - P(A)$$

Basic Probability Rules

- *Example:* Rolling a die — what's the probability of getting a 2 or a 5?

Basic Probability Rules

- *Example:* Rolling a die — what's the probability of getting a 2 or a 5? Since 2 and 5 can't happen at the same time:

$$P(2 \text{ or } 5) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6}$$

- *Example:* Tossing a coin and rolling a die. Probability of getting heads **and** a 4:

Basic Probability Rules

- *Example:* Rolling a die — what's the probability of getting a 2 or a 5? Since 2 and 5 can't happen at the same time:

$$P(2 \text{ or } 5) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6}$$

- *Example:* Tossing a coin and rolling a die. Probability of getting heads **and** a 4:

$$P(H \text{ and } 4) = \frac{1}{2} \times \frac{1}{6} = \frac{1}{12}$$

- *Example:* Probability of not getting a 6 when rolling a die:

Basic Probability Rules

- *Example:* Rolling a die — what's the probability of getting a 2 or a 5? Since 2 and 5 can't happen at the same time:

$$P(2 \text{ or } 5) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6}$$

- *Example:* Tossing a coin and rolling a die. Probability of getting heads **and** a 4:

$$P(H \text{ and } 4) = \frac{1}{2} \times \frac{1}{6} = \frac{1}{12}$$

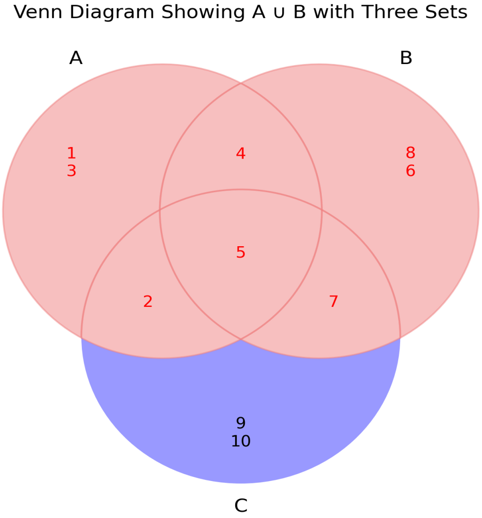
- *Example:* Probability of not getting a 6 when rolling a die:

$$P(\text{not } 6) = 1 - \frac{1}{6} = \frac{5}{6}$$

Union of sets: Addition Rule

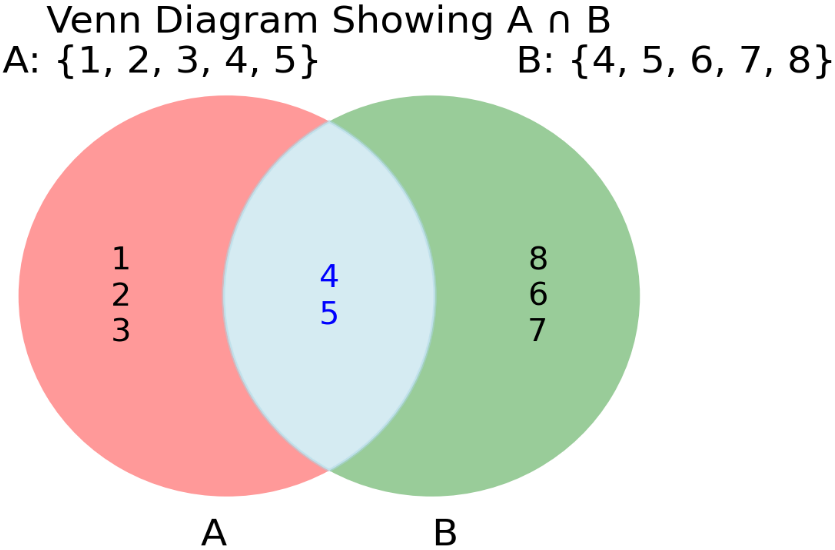
A: {1, 2, 3, 4, 5}

B: {4, 5, 6, 7, 8}



C: {2, 5, 7, 9, 10}

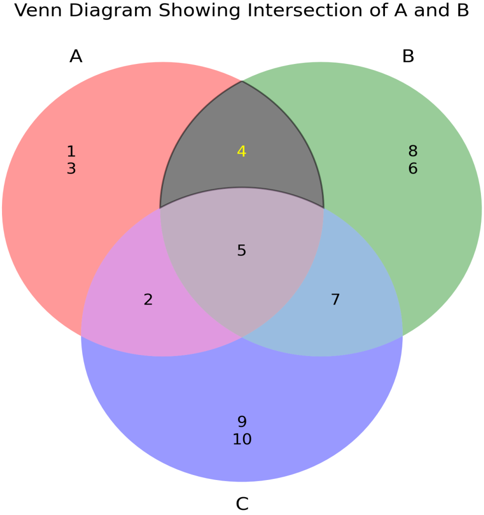
Intersection of sets: Multiplication Rule



Intersection of sets

A: {1, 2, 3, 4, 5}

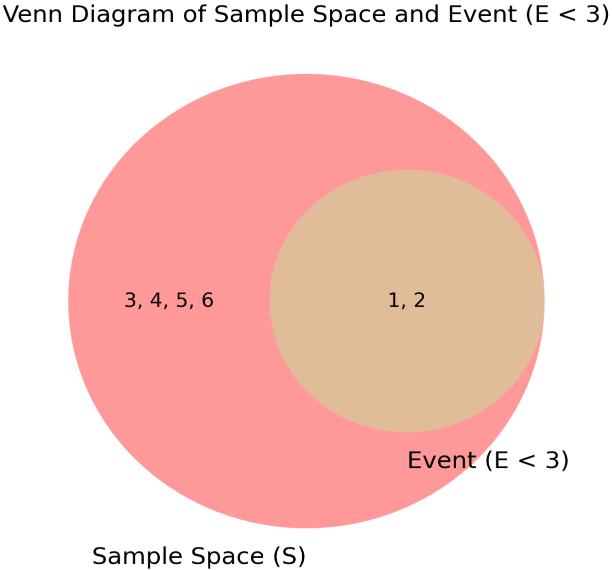
B: {4, 5, 6, 7, 8}



C: {2, 5, 7, 9, 10}

Sample Space and Event

- Event is a subset of Sample Space.



Mutually Exclusive Events

Mutually Exclusive Events

Definition: Two events A and B are said to be **mutually exclusive** if they cannot occur at the same time.

Key Characteristics:

- If one event happens, the other cannot.
- Their intersection is empty: $A \cap B = \emptyset$.
- The probability of both occurring together is zero.

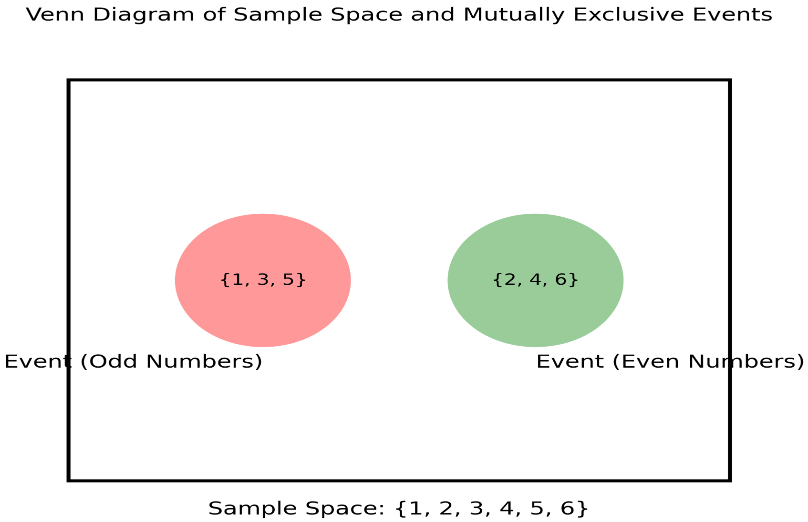
$$P(A \cap B) = 0.$$

Example: Rolling a Die

- Let A be the event of rolling an even number: $A = \{2, 4, 6\}$.
- Let B be the event of rolling an odd number: $B = \{1, 3, 5\}$.
- Since no number can be both even and odd, A and B are mutually exclusive.
- Mathematically, $P(A \cap B) = 0$.

Sample Space and Mutually Exclusive Events

- Events that cant occur at the same time.



Conditional Probabilty

Introduction to Conditional Probability

Definition: The conditional probability of an event A given that event B has occurred is defined as:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \quad P(B) > 0.$$

Key Points:

- Measures the likelihood of A occurring given that B has occurred.
- The probability space is reduced to B.
- Requires that $P(B) > 0$.

Example Problem

Example: A box contains 5 red balls and 7 blue balls. If a ball is drawn at random, what is the probability that it is red given that it is not blue?
Solution:

- Let A be the event that the ball drawn is red.
- Let B be the event that the ball drawn is not blue.
- Since the only non-blue balls are red, we have $P(A \cap B) = P(A)$.

Using the conditional probability formula:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{5/12}{5/12} = 1.$$

Example Problem

Example: A bag contains 4 green, 3 red, and 5 yellow marbles. If a marble is randomly selected, what is the probability that it is red given that it is not yellow?

Solution:

- Let A be the event that the marble drawn is red.
- Let B be the event that the marble drawn is not yellow.
- The total number of marbles is $4 + 3 + 5 = 12$.
- The number of non-yellow marbles is $4 + 3 = 7$.
- The number of red marbles is 3.

Using the conditional probability formula:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{3/12}{7/12} = \frac{3}{7}.$$

Example Problem

Example: Suppose we have a deck of 52 playing cards. What is the probability that a card drawn is an Ace given that it is a Spade?

Solution:

- Let A be the event that the card drawn is an Ace.
- Let B be the event that the card drawn is a Spade.
- There are 13 Spades and only 1 Ace of Spades.

Using the conditional probability formula:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{1/52}{13/52} = \frac{1}{13}.$$

DVAnalytics Transforming You

Independence

Independent Events

Definition: Two events A and B are said to be **independent** if the occurrence of one does not affect the probability of the other. Mathematically,

$$P(A \cap B) = P(A)P(B).$$

Equivalent Conditions:

- $P(A \mid B) = P(A)$, meaning knowing that B occurred does not change the probability of A.
- $P(B \mid A) = P(B)$, meaning knowing that A occurred does not change the probability of B.

Example: Suppose we roll a fair die and flip a fair coin. Define:

- A: Getting a 6 on the die.
- B: Getting heads on the coin.

Since rolling the die does not affect the coin flip,

$$P(A \cap B) = \left(\frac{1}{6}\right) \times \left(\frac{1}{2}\right) = \frac{1}{12},$$

which matches $P(A)P(B)$, proving independence.

Examples and Misconceptions

Example: Drawing Cards from a Deck

- Let A be the event of drawing a red card.
- Let B be the event of drawing a spade.
- Since a card cannot be both red and a spade, $P(A \cap B) = 0$.
- However, $P(A)P(B) = \left(\frac{26}{52} \times \frac{13}{52}\right) \neq 0$.
- Since $P(A \cap B) \neq P(A)P(B)$, A and B are **not independent**.

Common Misconceptions:

- **Independence is not the same as mutual exclusivity**: If A and B are mutually exclusive, then $P(A \cap B) = 0$, so they cannot be independent unless $P(A) = 0$ or $P(B) = 0$.
- **Independence is not about equal probabilities**: Events can be independent even if they have different probabilities.

Understanding $P(A \cap B)$: Intersection of Events

Case	**Formula** and **Explanation**
Independent Events	$P(A \cap B) = P(A) \cdot P(B)$ A and B do not affect each other (e.g., coin toss and die roll)
Dependent Events	$P(A \cap B) = P(A) \cdot P(B \mid A)$ Probability of B depends on A happening (e.g., drawing 2 cards without replacement)
Mutually Exclusive Events	$P(A \cap B) = 0$ A and B cannot happen at the same time (e.g., getting 2 and 5 on a single die roll)

Law of Total Probability

Law of Total Probability

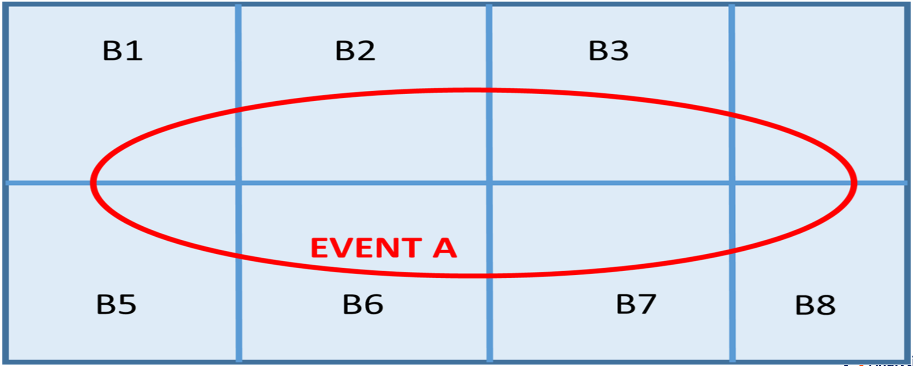
- The **law of total probability** is a fundamental rule in probability theory.
- The law of total probability provides a way to compute the probability of an event by considering all possible ways that event can occur.
- It provides a way to compute the probability of an event by considering partitioning of the sample space.
- Useful when dealing with conditional probabilities.
- If $B_1, B_2, \ldots, B_n$ form a partition of the sample space and $P(B_i) > 0$ for all i, then for any event A:

$$P(A) = \sum_{i=1}^{n} P(A \mid B_i) P(B_i)$$

Illustration

Consider a sample space partitioned into disjoint events $B_1, B_2, B_3 \ldots B_8$:

- Each B_i represents a different way an event can happen.
- The probability of A is the sum of its conditional probabilities weighted by the probability of each B_i.



Example 1: Defective Items

A factory has two machines:

- Machine 1 produces 60% of items, and Machine 2 produces 40%.
- Machine 1 has a 5% defect rate, while Machine 2 has a 10% defect rate.

What is the probability that a randomly chosen item is defective?

Solution

Define events:

- D: Item is defective
- B_1: Item from Machine 1, $P(B_1) = 0.6$
- B_2: Item from Machine 2, $P(B_2) = 0.4$

Given:

$$P(D \mid B_1) = 0.05, \quad P(D \mid B_2) = 0.10$$

Using the law of total probability:

$$P(D) = P(D \mid B_1)P(B_1) + P(D \mid B_2)P(B_2)$$

$$P(D) = (0.05)(0.6) + (0.10)(0.4) = 0.03 + 0.04 = 0.07$$

Thus, the probability of selecting a defective item is 0.07 (or 7%).

Bayes Rule

Bayes Theorem

- **Bayes Theorem** is a fundamental result in probability theory.
- It describes how to update our belief about an event based on new evidence.
- Commonly used in medical diagnosis, spam filtering, and machine learning.

Derivation of Bayes' Theorem:

- Consider a sample space partitioned into two mutually exclusive events B_1 and B_2, such that $B_1 \cup B_2 = S$.
- From the definition of conditional probability:

$$P(A \mid B_1) = \frac{P(A \cap B_1)}{P(B_1)}$$

$$P(B_1 \mid A) = \frac{P(A \cap B_1)}{P(A)}$$

- Rearranging both equations:

$$P(A \cap B_1) = P(A \mid B_1)P(B_1)$$

$$P(A \cap B_1) = P(B_1 \mid A)P(A)$$

- Equating the two expressions for $P(A \cap B_1)$:

$$P(A \mid B_1)P(B_1) = P(B_1 \mid A)P(A)$$

Derivation

- Solving for $P(B_1 \mid A)$ gives:

$$P(B_1 \mid A) = \frac{P(A \mid B_1)P(B_1)}{P(A)}$$

- This is Bayes' Theorem.
- Using the Law of Total Probability:

$$P(A) = P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2)$$

Theorem Statement

Bayes' Theorem: Given an event A and a partition of the sample space $B_1, B_2, \ldots, B_n$:

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{P(A)}$$

where

$$P(A) = \sum_{i=1}^{n} P(A \mid B_i)P(B_i)$$

LTP: If $B_1, B_2, ..., B_n$ form a partition of the sample space, then for any event A:

$$P(A) = \sum_{i=1}^{n} P(A \mid B_i)P(B_i)$$

- This rule helps compute the probability of an event by considering all possible ways it can occur.
- It is used as the denominator in Bayes' Theorem.

Example 1: Spam Filtering

- A spam filter detects spam emails based on keywords.
- Suppose 20% of emails are spam.
- Given that an email contains the word "free", how likely is it to be spam?

Let:

$$P(S) = 0.20, \quad P(S^c) = 0.80$$
$$P(F \mid S) = 0.50, \quad P(F \mid S^c) = 0.05$$

Using Bayes' Theorem:

$$P(S \mid F) = \frac{P(F \mid S)P(S)}{P(F)}$$

$$P(F) = (0.50 \times 0.20) + (0.05 \times 0.80) = 0.14$$
$$P(S \mid F) = \frac{0.50 \times 0.20}{0.14} = 0.714$$

The probability that an email is spam given that it contains "free" is **71.4%**.

Example 2: Medical Testing

- A disease affects 1% of a population.
- A test detects the disease with 95% sensitivity and 90% specificity.
- What is the probability that a person who tests positive actually has the disease?

Using Bayes' Theorem:

$$P(D \mid T^+) = \frac{P(T^+ \mid D)P(D)}{P(T^+)}$$

$$\begin{aligned}
P(T^+ \mid D) &= 0.95, \quad P(D) = 0.01 \\
P(T^+ \mid D^c) &= 1 - 0.90 = 0.10, \quad P(D^c) = 0.99 \\
P(T^+) &= (0.95 \times 0.01) + (0.10 \times 0.99) = 0.1045
\end{aligned}$$

$$P(D \mid T^+) = \frac{0.95 \times 0.01}{0.1045} \approx 0.091$$

Only 9.1% of those who test positive actually have the disease.

Bayes' Theorem and its Relationship with LTP

- **LTP** computes the total probability of an event.
- **Bayes' Theorem** inverts this relationship to find the probability of causes given evidence.
- It allows us to update beliefs based on new information.

Law of Total Probability vs Bayes' Rule

Big Picture: How They Differ
Pizza Delivery Example

- 60% orders from A (20% late), 40% from B (50% late)

Law of Total Probability:

$$P(\text{Late}) = P(\text{Late} \mid \text{A}) \cdot P(A) + P(\text{Late}|\text{B}) \cdot P(B)$$

$\rightarrow$ *What's the chance your pizza is late overall?*

Bayes' Rule:

$$P(\text{B} \mid \text{Late}) = \frac{P(\text{Late}|\text{B}) \cdot P(B)}{P(\text{Late})}$$

$\rightarrow$ *Given that it's late, what's the chance it came from B?*

Summary

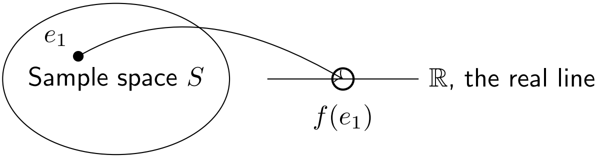
	Law of Total Probability
Think of it as...	Looking **forward** from causes to outcomes
What it helps with	Calculates overall probability of an outcome from multiple scenarios

	Bayes' Rule
Think of it as...	Looking **backward** from outcomes to causes
What it helps with	Updates the probability of a cause given that an outcome has occurred

Random Variables

Introduction to Random Variables

- A **random variable** is a numerical description of the outcome of a statistical experiment. It is a way to assign numbers to outcomes of a random experiment
- It is a function defined on a sample space, S, that associates a real number with each outcome in S.



- Two main types:
  - **Discrete Random Variables**: Take countable values (e.g., number of heads in a coin toss).
  - **Continuous Random Variables**: Take an infinite number of values within a range (e.g., height of individuals).
- If each event is associated with its probability in random variable, it is called as **Probability distribution**.

Example: Coin Toss

- Experiment: Toss a coin
- Let X be a random variable such that:
  - $X = 1$ if Heads
  - $X = 0$ if Tails
- X is a discrete random variable.

What is a Distribution?

- A **distribution** tells us the probability of each value of a random variable.
- Shows how likely each outcome is.
- A distribution describes how values of a random variable are spread.
- It helps in understanding the likelihood of different outcomes.
- Examples: Normal, Uniform, Binomial, Poisson, etc.
- Can be a:
  - **Probability distribution** (for discrete variables)
  - **Probability density function** (for continuous variables)

Creating a Distribution:

- Consider rolling a six-sided die multiple times.
- We record the outcomes and plot their frequency.

Coin Toss: Distribution Example

Outcome	Value (X)	Probability
Heads	1	0.5
Tails	0	0.5

This is the probability distribution of X.

Dice Roll Example

- Experiment: Roll a fair 6-sided die
- Random variable X = number on the die

Value of X	Probability
1	$1/6$
2	$1/6$
3	$1/6$
4	$1/6$
5	$1/6$
6	$1/6$

This is a uniform distribution.

Summary

Concept	Explanation
Random Variable	A variable that assigns numbers to outcomes of a random process
Distribution	Describes how probabilities are assigned to each value of the random variable

Example

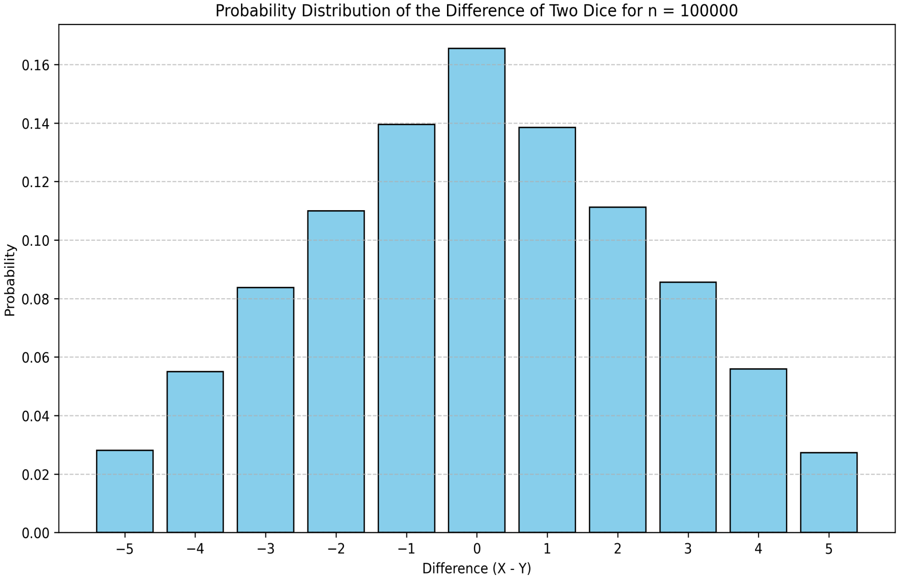
Let X and Y be the outcomes of two fair 6-sided dice.
Define a new **random variable**:

$$D = X - Y$$

Possible values of the **random variable** D are:

$$\{-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5\}$$

The Distribution



Tabulated Distribution (Example)

D	Probability	Cumulative
-5	0.0278	0.0278
-4	0.0556	0.0834
-3	0.0833	0.1667
-2	0.1111	0.2778
-1	0.1389	0.4167
0	0.1667	0.5834
1	0.1389	0.7223
2	0.1111	0.8334
3	0.0833	0.9167
4	0.0556	0.9723
5	0.0278	1.0000

- Observe the Cumulative Probability

Motivating a Formal Function

Question: How do we represent a distribution mathematically?

Answer: Using a function that tells us the probability of each value:

- For discrete random variables: **Probability Mass Function (PMF)**
- For continuous random variables: **Probability Density Function (PDF)**

PMF: Probability Mass Function

Definition: For a discrete random variable X, the PMF is defined as:

$$p(x) = P(X = x)$$

Example: Let $D = X - Y$ be the difference of two dice rolls. Then:

$$p(0) = 0.1667, \quad p(1) = 0.1389, \quad \ldots$$

Properties:

- $0 \leq p(x) \leq 1$
- $\sum_x p(x) = 1$

DV Analytics
Transforming You

PDF: Probability Density Function

Definition: For a continuous random variable X, the PDF is a function $f(x)$ such that:

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx$$

Properties:

- $f(x) \geq 0$
- $\int_{-\infty}^{\infty} f(x)\,dx = 1$
- It means The total area under the PDF curve is always 1

Note: $f(x)$ is not a probability — only areas under the curve give probabilities.

PMF vs PDF

- **PMF**: Used for discrete variables (e.g., dice outcomes)
- **PDF**: Used for continuous variables (e.g., height, weight)
- Both describe how probability is distributed over outcomes

Why CDF?

Question: What is the probability that a variable is *less than or equal to* a value?

Answer: We use the **Cumulative Distribution Function (CDF)** to capture that.

Cumulative Distribution Function (CDF)

Definition: For a random variable X, the CDF is defined as:

$$F(x) = P(X \leq x)$$

Discrete: $F(x) = \sum_{t \leq x} p(t)$
Continuous: $F(x) = \int_{-\infty}^{x} f(t)\, dt$

Properties of CDF

- $F(x)$ is non-decreasing
- $\lim_{x \to -\infty} F(x) = 0$
- $\lim_{x \to \infty} F(x) = 1$
- For continuous X, $F'(x) = f(x)$

Example: CDF of $D = X - Y$

- We already saw the PMF for D.
- The CDF at a point is the sum of all probabilities up to that point.

Example:

$$F(0) = P(D \leq 0) = p(-5) + p(-4) + \cdots + p(0) = 0.5834$$

Cumulative Distribution Function (CDF)

- The Cumulative Distribution Function (CDF) gives the probability that a random variable X is less than or equal to a certain value x.
- It is defined as:

$$F(x) = P(X \leq x)$$

- For a discrete random variable:

$$F(x) = \sum_{t \leq x} P(X = t)$$

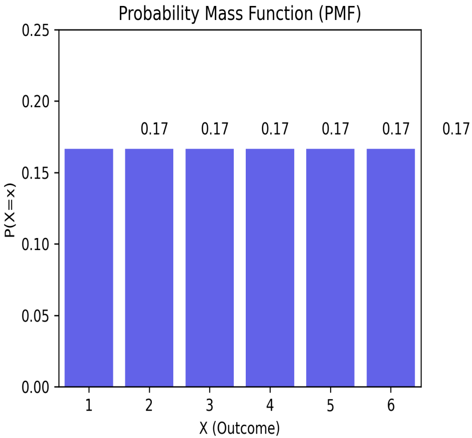
- For a continuous random variable:

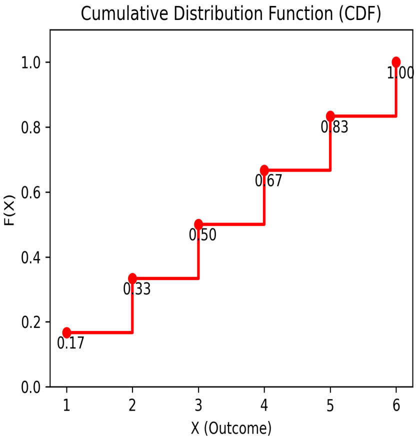
$$F(x) = \int_{-\infty}^{x} f(t)dt$$

- The CDF is a non-decreasing function that ranges from 0 to 1.
- Example: For a standard normal variable, the CDF is given by:

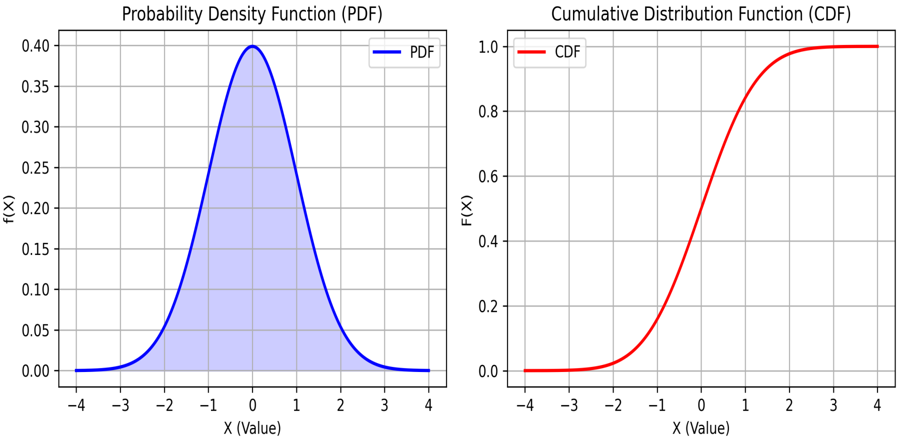
$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}dt$$

PMF and CDF


Probability Mass Function (PMF)
0.25
0.20
0.15
0.10
0.05
0.00
P(X=x)
0.17
0.17
0.17
0.17
0.17
0.17
1
2
3
4
5
6
X (Outcome)


Cumulative Distribution Function (CDF)
1.0
0.8
0.6
0.4
0.2
0.0
F(X)
0.17
0.33
0.50
0.67
0.83
1.00
1
2
3
4
5
6
X (Outcome)

PDF and CDF


Probability Density Function (PDF)
PDF
f(X)
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−4
−3
−2
−1
0
1
2
3
4
X (Value)
Cumulative Distribution Function (CDF)
CDF
F(X)
1.0
0.8
0.6
0.4
0.2
0.0
X (Value)

Comparison of PMF, PDF, and CDF

- **Probability Mass Function (PMF):**
  - Used for discrete random variables.
  - Assigns probabilities to specific values.
- **Probability Density Function (PDF):**
  - Used for continuous random variables.
  - Represents density, not direct probabilities.
  - The area under the curve represents probability.
- **Cumulative Distribution Function (CDF):**
  - Used for both discrete and continuous random variables.
  - Gives cumulative probability up to a given value.
  - Always non-decreasing and ranges from 0 to 1.

Summary of Applications

To summarize, here are the main uses of these functions:

- **PMF**: Used for discrete random variables to calculate probabilities for specific outcomes.
- **CMF**: Useful for calculating the cumulative probability in discrete distributions.
- **PDF**: Used for continuous random variables to model probability densities and calculate probabilities over intervals.
- **CDF**: Essential for continuous variables to calculate the cumulative probability up to a given value.

These functions are widely applied in fields like:

- Statistics and data analysis.
- Risk management and decision-making.
- Machine learning algorithms (e.g., Gaussian Naive Bayes classifier).
- Quality control and reliability engineering.

Example: Rolling a Die

- Consider rolling a fair six-sided die.
- The outcome X is a discrete random variable with possible values: $\{1, 2, 3, 4, 5, 6\}$.
- Probability Mass Function (PMF):

$$P(X = x) = \frac{1}{6}, \quad x = 1, 2, 3, 4, 5, 6$$

- Expected Value:

$$E[X] = \sum xP(X = x) = \frac{1 + 2 + 3 + 4 + 5 + 6}{6} = 3.5$$

Random Variables

- A **random variable** is a function that assigns a numerical value to each outcome in a sample space.
- Can be:
  - **Discrete**: Countable outcomes (e.g., number of heads in coin tosses).
  - **Continuous**: Uncountable outcomes (e.g., height of students).

Joint Distribution: Definition

Definition

The **joint distribution** of two random variables X and Y gives the probability that X takes a specific value and Y takes a specific value at the same time.

Notation: $P(X = x, Y = y)$

Joint Probability Table (Discrete Case)

Example: Joint Distribution Table

X\Y	1	2	3
1	0.1	0.1	0.1
2	0.2	0.1	0.1
3	0.05	0.1	0.15

Each cell gives $P(X = x, Y = y)$.

Marginal Distribution: Definition

Definition

The **marginal distribution** of a random variable is the probability distribution of that variable alone, obtained by summing (or integrating) the joint distribution over the other variable.

For discrete variables:

$$P(X = x) = \sum_y P(X = x, Y = y)$$

$$P(Y = y) = \sum_x P(X = x, Y = y)$$

Finding Marginal Distributions (Example)

From the previous table:

Marginal of X:

$$P(X = 1) = 0.1 + 0.1 + 0.1 = 0.3$$

$$P(X = 2) = 0.2 + 0.1 + 0.1 = 0.4$$

$$P(X = 3) = 0.05 + 0.1 + 0.15 = 0.3$$

Marginal of Y is found similarly by adding down columns.

Joint and Marginal Densities (Continuous Case)

Joint Probability Density Function (pdf): $f(x, y)$

- $f(x, y)$ satisfies:
$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1$$
- Probability that (X, Y) lies in a region A:

$$P((X, Y) \in A) = \iint_A f(x, y)\, dx\, dy$$

Marginal Densities:

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx$$

DV Analytics Transforming You

Visualizing Joint Distributions

Joint distributions can be visualized as 3D surfaces or contour plots.

Key Points

- **Joint Distribution** shows relationship between two random variables.
- **Marginal Distribution** looks at one variable alone.
- In the discrete case, sum probabilities; in the continuous case, integrate.
- Helps in finding conditional distributions and checking independence.

Quick Check!

Question:

Given the joint table below:

X\Y	0	1
0	0.3	0.2
1	0.1	0.4

Find:

- $P(X = 0)$
- $P(Y = 1)$

Solution

Marginal for $X = 0$:

$$P(X = 0) = 0.3 + 0.2 = 0.5$$

Marginal for $Y = 1$:

$$P(Y = 1) = 0.2 + 0.4 = 0.6$$

Independence of Random Variables

Definition

Two random variables X and Y are **independent** if and only if:

$$P(X = x, Y = y) = P(X = x) \times P(Y = y) \quad \text{for all } x, y$$

Otherwise, X and Y are said to be **dependent**.

Idea Behind Checking Independence

- Start from the **joint probability** $P(X = x, Y = y)$.
- Compute the **marginal probabilities** $P(X = x)$ and $P(Y = y)$.
- Check if:

$$P(X = x, Y = y) = P(X = x) \times P(Y = y)$$

for all possible pairs (x, y).

Example: Joint Distribution Table

X**Y**	0	1
0	0.3	0.2
1	0.1	0.4

This table gives $P(X = x, Y = y)$ for each pair (x, y).

Step 1: Marginal Probabilities

Compute Marginal of X:

$$P(X=0)=0.3+0.2=0.5$$

$$P(X=1)=0.1+0.4=0.5$$

Compute Marginal of Y:

$$P(Y=0)=0.3+0.1=0.4$$

$$P(Y=1)=0.2+0.4=0.6$$

Step 2: Check for Independence

Check if $P(X = x, Y = y) = P(X = x) \times P(Y = y)$ for all x, y.
Example:

$$P(X = 0, Y = 0) = 0.3$$

$$P(X = 0) \times P(Y = 0) = 0.5 \times 0.4 = 0.2$$

Since $0.3 \neq 0.2$, X and Y are NOT independent.

Conclusion

- If $P(X = x, Y = y) \neq P(X = x)P(Y = y)$ for any (x, y), variables are **dependent**.
- Independence implies that knowing X tells us nothing about Y, and vice versa.

Checking joint vs marginal probabilities is the key!

Independence and Conditional Distributions

Independence of Random Variables

Two random variables X and Y are **independent** if:

$$P(X = x, Y = y) = P(X = x) \times P(Y = y) \quad \text{for all } x, y$$

Interpretation: Knowing X does not affect beliefs about Y.

Effect on Conditional Distributions

If independent:

$$P(Y = y \mid X = x) = P(Y = y)$$

(Conditioning on X does not change the distribution of Y.)

If not independent:

$$P(Y = y \mid X = x) \neq P(Y = y)$$

(Knowing X *does* change the distribution of Y.)

Independence vs Dependence: A Comparison

Property	Independent	Dependent
Joint Probability	$P(X, Y) = P(X)P(Y)$	$P(X, Y) \neq P(X)P(Y)$
Conditional Probability	$P(Y\|X) = P(Y)$	$P(Y\|X)$ changes with X
Interpretation	X and Y do not affect each other	X and Y influence each other

Key Point:

Independence means the conditional distribution is the same as the marginal distribution.

Population and Sample

Population

- **Population** refers to the entire set of individuals, items, or data points that we are interested in studying.
- The population can be finite or infinite, and it contains all members from which we wish to draw conclusions.

Key metrics for a population include:

- **Total Population size** (N) ; Size of the population.
- **Population Mean** (μ): The average of all the values in the population.
- **Population Variance** (σ^2): Measures the spread or dispersion of the values in the population.
- **Population Standard Deviation** (σ): The square root of the population variance, representing the average distance of the data points from the population mean.

Sample

- A **sample** is a subset of the population that is selected for the purpose of statistical analysis.
- It is often impractical or impossible to collect data from an entire population.
- so samples are used to make inferences about the population.

Key metrics for a sample include:

- **Total Sample size** (n; Size of the sample.
- **Sample Mean** ($\bar{x}$): The average of all the values in the sample.
- **Sample Variance** (s^2): Measures the spread or dispersion of the values in the sample.
- **Sample Standard Deviation** (s): The square root of the sample variance, representing the average distance of the sample points from the sample mean.

DVAnalytics Transforming You

Comparison: Population vs. Sample

The population and sample have similar metrics, but the key difference is that the population includes all members, while the sample is just a subset.

Metric	Population	Sample
Mean	μ	$\bar{x}$
Variance	σ^2	s^2
Standard Deviation	σ	s
Proportion	p	$\hat{p}$

While the population parameters are fixed, the sample statistics are subject to sampling variability and serve as estimates of the population parameters.

Population Parameter vs. Sample Statistic

In statistics, we often distinguish between two key concepts:

- **Population Parameter**
- **Sample Statistic**

These concepts are fundamental to understanding how we make inferences about a population based on a sample of data.

Population Parameter

- A **population parameter** is a numerical value that summarizes a characteristic of an entire population.
- These parameters are fixed and unchanging.
- But they are often unknown because it's usually impractical to measure an entire population.

Examples of Population Parameters:

- **Population Mean** (μ)
- **Population Variance** (σ^2)
- **Population Standard Deviation** (σ)
- **Population Proportion** (p)

Key Point: Population parameters are fixed values, but they are often difficult or impossible to measure directly due to the size or accessibility of the population.

Sample Statistic

- A **sample statistic** is a numerical value calculated from the data of a sample, which is a subset of the population.
- Sample statistics are used to estimate population parameters

.

Examples of Sample Statistics:

- **Sample Mean** $(\bar{x})$
- **Sample Variance** (s^2)
- **Sample Standard Deviation** (s)
- **Sample Proportion** $(\hat{p})$.

Key Point: Sample statistics are variable, meaning they can change from sample to sample, even when taken from the same population.

Key Differences: Population Parameter vs. Sample Statistic

Aspect	Population Parameter
Definition	A numerical value that summarizes a characteristic of the entire population.
Notation	μ (mean), σ^2 (variance), σ (standard deviation), p (proportion)
Fixed or Variable?	Fixed and constant (if the population is known).
Accessibility	Often unknown because it's difficult to measure the entire population.
Usage	Used to describe the entire population.

Aspect	Sample Statistic
Definition	A numerical value calculated from the data of a sample.
Notation	$\bar{x}$ (mean), s^2 (variance), s (standard deviation), $\hat{p}$ (proportion).
Fixed or Variable?	Variable; changes depending on the sample.
Accessibility	Known and easy to calculate because it's based on the sample data.
Usage	Used to estimate the corresponding population parameter.

Inferential Statistics

Sample Statistic: Its behaviour

- A **sample statistic** (e.g., sample mean $\bar{X}$, sample variance S^2) is computed from a sample drawn from a population.
- Each time we draw a sample, the values in the sample may change.
- Since the sample changes randomly, the statistic computed from it also changes.
- **Therefore, the statistic depends on random outcomes** — it is a function of random variables.
- **Conclusion:** A sample statistic is a **random variable** because its value varies across different samples.
- It means Sample statistics will have a distribution.

Inferential Statistics: Estimating Population Parameters

Since it's often impractical to measure an entire population, we rely on **inferential statistics** to make predictions or inferences about population parameters based on sample statistics.

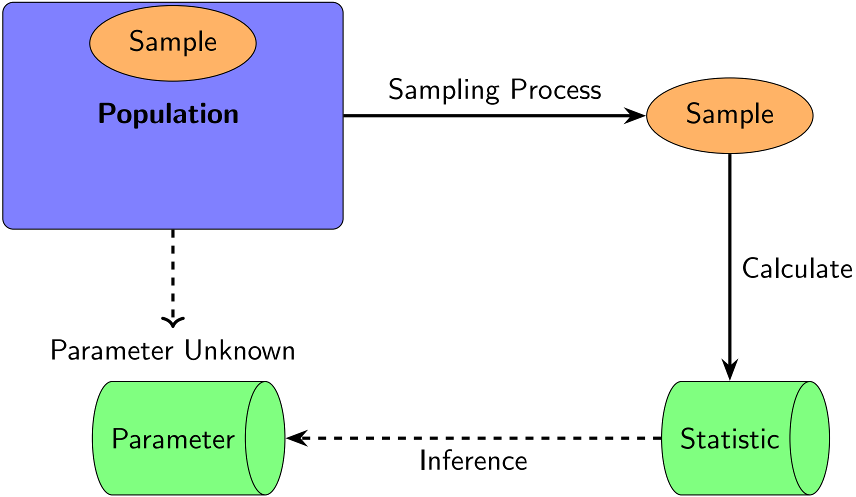
For example:

- If we want to know the average height of all high school students in a country, we might take a random sample of 500 students and calculate the sample mean height ($\bar{x}$).
- The sample mean ($\bar{x}$) can be used to estimate the population mean (μ).
- We can also compute confidence intervals and conduct hypothesis testing to make more precise inferences about the population parameter.

Key Point:

- Sample statistics are used as estimates of population parameters.
- There is always some degree of uncertainty associated with these estimates.
- This can be quantified using confidence intervals, margin of error etc.

Sampling and Inference



Research Question

What is the average number of cups of coffee per day consumed by college students in the U.S.?

- Population: All college students in the U.S.
- Parameter of interest: Average daily coffee consumption

Why We Sample

- It's impractical to survey the entire population.
- We take a random sample to estimate the population mean.

Example:

Randomly select 30 students and ask how many cups of coffee they drink per day.

Sample Data (Example) [2, 0, 1, 3, 2, 1, 2, 4, 0, 1, 2, 3, 1, 2, 2, 1, 5, 1, 0, 3, 2, 1, 2, 0, 4, 3, 2, 1, 2, 1]

- Sample size: $n = 30$
- Sample mean: (computed visually)

Each Response is a Random Variable

- Let X_i = cups of coffee consumed by the i-th student.
- Each X_i is a random variable:
  - Depends on which student is randomly chosen
  - Value varies with each sample
- The sample mean $\bar{X}$ is also a random variable.

Summary Table:

Concept	**Example**
Population	All U.S. college students
Random Variable X_i	Coffee cups consumed by student i
Sample	30 values like [2, 0, 1, ..., 1]
Random selection	Selecting students at random
Why It's Random	Different students give different values

Probability Distributions: Binomial Distributions

Binomial Distribution

- A **Bernoulli trial** is an experiment or process that results in a binary outcome: success (usually coded as 1) or failure (usually coded as 0).
- The **binomial distribution** describes the number of successes in a fixed number of independent Bernoulli trials, each with the same probability of success.
- The binomial distribution can be defined by the following parameters:
  - n : The number of trials.
  - p : The probability of success on a single trial.
  - k : The number of successes in n trials.
- The probability mass function (PMF) of the binomial distribution is given by:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

  where:
  - $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ is the binomial coefficient.
  - p^k is the probability of having k successes.
  - $(1 - p)^{n-k}$ is the probability of having $n - k$ failures.

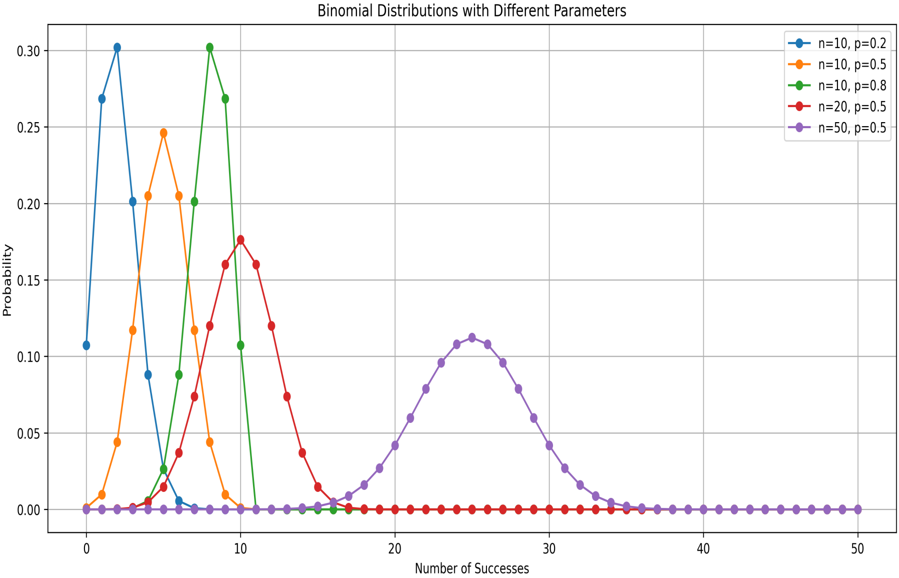
Properties of the Binomial Distribution

Some important properties of the binomial distribution:

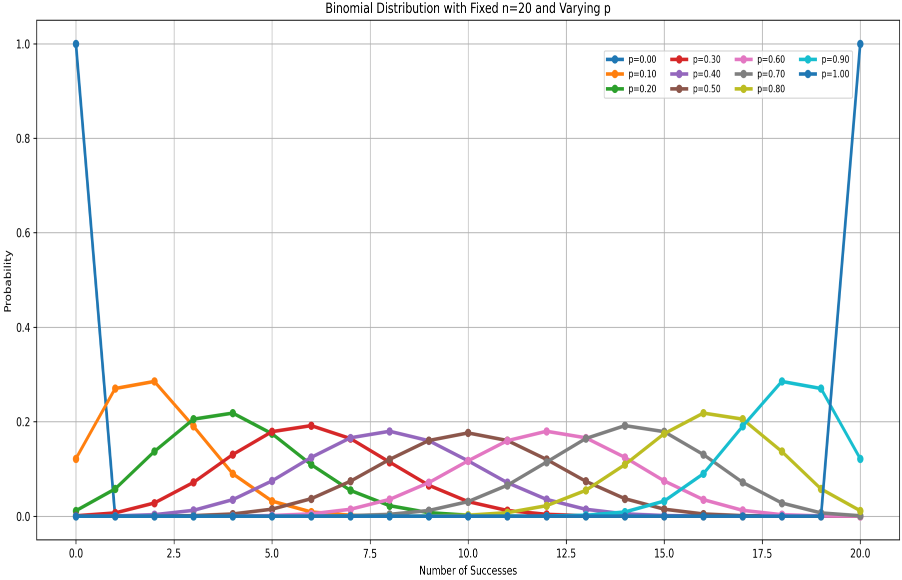
- **Mean:** $\mu = n \cdot p$
- **Variance:** $\sigma^2 = n \cdot p \cdot (1 - p)$
- **Standard Deviation:** $\sigma = \sqrt{n \cdot p \cdot (1 - p)}$

These properties help in understanding the central tendency and spread of the distribution.

Binomial Distribution for different parameters



Binomial Distribution for different parameters



Example Problem

Consider a scenario where a fair coin is flipped 10 times. What is the probability of getting exactly 6 heads?

We know the following:

- $n = 10$ (number of flips),
- $p = 0.5$ (probability of getting a head),
- $k = 6$ (number of heads we are interested in).

The probability mass function for the binomial distribution is:

$$P(X = 6) = \binom{10}{6} (0.5)^6 (0.5)^{10-6} = \binom{10}{6} (0.5)^{10}$$

Now, calculating the binomial coefficient:

$$\binom{10}{6} = \frac{10!}{6!4!} = 210$$

Therefore, the probability is:

$$P(X = 6) = 210 \times (0.5)^{10} = 210 \times \frac{1}{1024} \approx 0.205$$

Industrial Example: Quality Control in Manufacturing

In a factory producing electronic components, the quality control department tests randomly selected parts for defects. Suppose the probability of a part being defective is 0.02 (2%).

The company inspects 100 parts from a production batch. What is the probability that exactly 3 parts are defective?

We have:

- $n = 100$ (number of parts inspected),
- $p = 0.02$ (probability of a part being defective),
- $k = 3$ (number of defective parts we are interested in).

The probability mass function is:

$$P(X = 3) = \binom{100}{3} (0.02)^3 (0.98)^{97}$$

First, calculate the binomial coefficient:

$$\binom{100}{3} = \frac{100!}{3!(100-3)!} = 161700$$

DVAnalytics Transforming You

Example

Now, calculate the probability:

$$P(X = 3) = 161700 \times (0.02)^3 \times (0.98)^{97}$$

- This gives the probability of having exactly 3 defective parts in the batch.
- useful for manufacturers to predict and manage the likelihood of defective items and take corrective actions accordingly

Binomial Distribution Functions in Excel

Excel provides two main functions for the Binomial Distribution:

- `BINOM.DIST(x, n, p, cumulative)`
  - `x`: Number of successes
  - `n`: Number of trials
  - `p`: Probability of success
  - `cumulative`: TRUE for cumulative distribution function (CDF), FALSE for probability mass function (PMF)
- `BINOM.DIST.RANGE(n, p, x1, [x2])`
  - Computes the probability of `x1` to `x2` successes (or exactly `x1` if `x2` omitted)

Example: `=BINOM.DIST(3, 10, 0.5, FALSE)` gives $P(X = 3)$
Example: `=BINOM.DIST.RANGE(10, 0.5, 3, 5)` gives $P(3 \leq X \leq 5)$

Excel Binomial Functions: Examples

1. **Using** `BINOM.DIST`

Question: A coin is flipped 8 times. What is the probability of getting exactly 5 heads?
Solution: `=BINOM.DIST(5, 8, 0.5, FALSE)` $\Rightarrow P(X = 5)$

2. **Using** `BINOM.DIST.RANGE`

Question: A multiple-choice quiz has 10 questions, each with 4 options. What is the probability of guessing between 3 and 5 questions correctly?
Solution: `=BINOM.DIST.RANGE(10, 0.25, 3, 5)` $\Rightarrow P(3 \leq X \leq 5)$

BINOM.INV in Excel

Function: `BINOM.INV(trials, probability_s, alpha)`

- Returns the smallest value x such that:

$$P(X \leq x) \geq \alpha$$

- Useful for finding the inverse of the cumulative binomial distribution.

Example Question: In a quality check of 20 items with a defect rate of 10%, what is the smallest number of defective items such that the cumulative probability is at least 90%?

Solution: `=BINOM.INV(20, 0.1, 0.9)` $\Rightarrow$ Smallest x such that $P(X \leq x) \geq 0.9$

Summary: Binomial Functions in Excel

Function	**Purpose**	**Typical Use Case**
`BINOM.DIST(x, n, p, FALSE)`	PMF	Probability of exactly x successes
`BINOM.DIST(x, n, p, TRUE)`	CDF	Probability of up to x successes
`BINOM.DIST.RANGE(n, p, x1, [x2])`	Range of PMF	Probability that $x1 \leq X \leq x2$ (or exactly $x1$ if $x2$ omitted)
BINOM.INV(n, p, α)	Inverse CDF	Smallest x such that $P(X \leq x) \geq \alpha$

Tip: Use `BINOM.DIST.RANGE` for probability intervals, and `BINOM.INV` to find thresholds based on cumulative probability.

Probability Distributions: Poisson Distributions

What is the Poisson Distribution?

The Poisson distribution is a probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space, given the average number of times the event occurs over that interval.
It is typically used for rare events that happen independently of each other.

Characteristics:

- The events are independent.
- The rate (λ) of occurrence is constant.
- The number of events in non-overlapping intervals are independent.

The Poisson distribution is often used in situations like:

- Number of phone calls at a call center in an hour.
- Number of accidents at a traffic intersection in a day.
- Number of customers arriving at a store in a given time period.

Poisson Distribution Formula

The probability mass function (PMF) of the Poisson distribution is given by:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots$$

where:

- X is the random variable representing the number of events.
- λ is the average number of events in the given time period (also called the rate parameter).
- k is the number of events observed.
- e is the Euler's number, approximately 2.71828.

This formula gives the probability of observing exactly k events when the expected number of events is λ.

Mean and Variance of Poisson Distribution

For a Poisson distribution with parameter λ:

- The mean is:

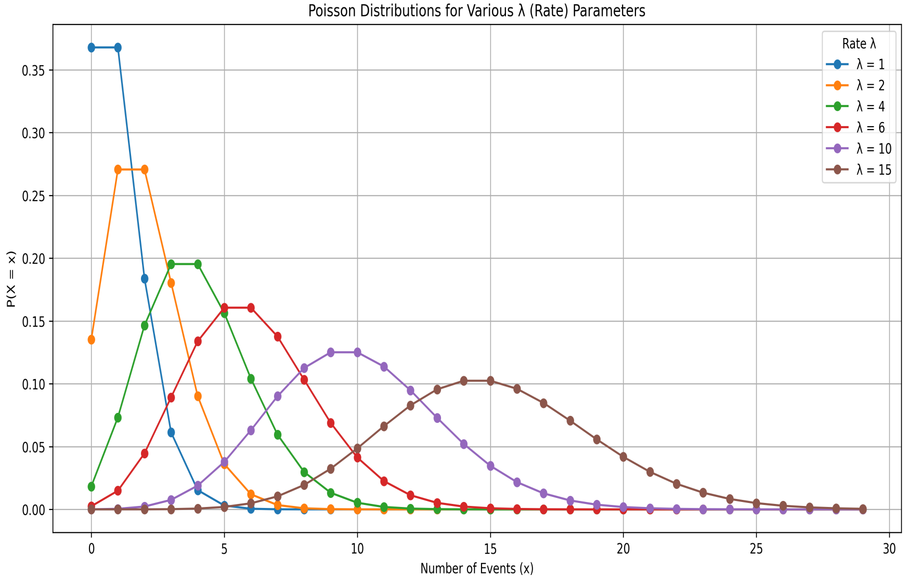
$$\mu = \lambda$$

- The variance is:

$$\sigma^2 = \lambda$$

Notice that the mean and variance of a Poisson distribution are both equal to λ, which is an interesting property.

Poisson Distribution for different parameters



Example of Poisson Distribution

Let's say the average number of cars passing through a toll booth per hour is 5. We want to calculate the probability of exactly 3 cars passing through the booth in one hour.
Here, the rate $\lambda = 5$ (the average number of cars per hour), and we want to find the probability when $k = 3$.
Using the Poisson formula:

$$P(X = 3) = \frac{5^3 e^{-5}}{3!}$$

$$P(X = 3) = \frac{125e^{-5}}{6}$$

$$P(X = 3) \approx 0.1404$$

So, the probability of exactly 3 cars passing through the toll booth in an hour is approximately 0.1404, or 14.04%.

Applications of Poisson Distribution

The Poisson distribution is widely used in various fields, such as:

- **Telecommunications:** Number of calls or messages in a given time frame.
- **Healthcare:** Number of patients arriving at an emergency room in a given hour.
- **Traffic Flow:** Number of cars passing a point in a road during a fixed period.
- **Physics:** Number of radioactive decay events in a time period.

The Poisson distribution is particularly useful when events occur randomly and independently, and we are interested in counting how often they happen in a specific interval.

Poisson Distribution Function in Excel

Function: `POISSON.DIST(x, mean, cumulative)`

- `x`: Number of events
- `mean`: Expected number of events (λ)
- `cumulative`:
  - TRUE – Cumulative distribution function $P(X \leq x)$
  - FALSE – Probability mass function $P(X = x)$

Example: `=POISSON.DIST(3, 4, FALSE)` returns $P(X = 3)$
Example: `=POISSON.DIST(3, 4, TRUE)` returns $P(X \leq 3)$

Poisson Function: Examples

1. PMF with `POISSON.DIST`

Question: A call center receives 4 calls per hour on average. What is the probability it receives exactly 3 calls in the next hour?
Solution: `=POISSON.DIST(3, 4, FALSE)` $\Rightarrow P(X = 3)$

2. CDF with `POISSON.DIST`

Question: What is the probability the call center receives at most 3 calls?
Solution: `=POISSON.DIST(3, 4, TRUE)` $\Rightarrow P(X \leq 3)$

Summary: Poisson Function in Excel

Function	Purpose	Typical Use Case
`POISSON.DIST(x, mean, FALSE)`	PMF	Probability of exactly x events in a fixed interval
`POISSON.DIST(x, mean, TRUE)`	CDF	Probability of up to x events (cumulative)

Tip: Use Poisson when modeling counts of rare events over time, space, or area.

Probability Distributions: Normal Distributions

What is the Normal Distribution?

The Normal distribution is a continuous probability distribution that is symmetric about the mean. It is one of the most important distributions in statistics due to its natural occurrence in many real-world phenomena.

Key Characteristics:

- It has a bell-shaped curve.
- The distribution is symmetric around the mean (μ).
- The mean, median, and mode of the distribution are all the same.
- The standard deviation (σ) controls the spread of the distribution.
- It is defined by two parameters: the mean (μ) and the standard deviation (σ).

The Normal distribution is used to model many natural and social phenomena such as heights, test scores, and errors in measurements.

Normal Distribution Formula

The probability density function (PDF) of the Normal distribution is given by:

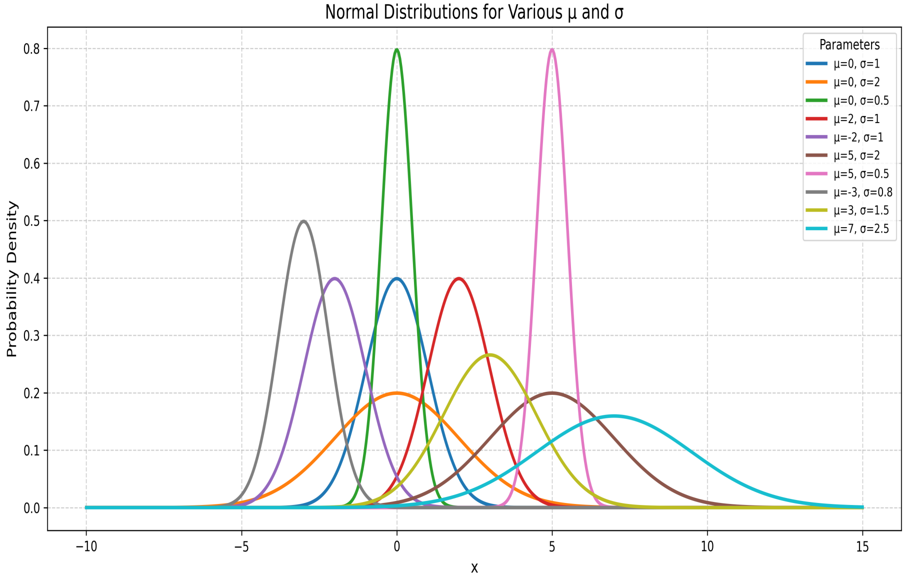
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where:

- $f(x)$ is the probability density at x.
- μ is the mean (average) of the distribution.
- σ is the standard deviation (a measure of the spread of the distribution).
- x is the variable of interest.
- e is Euler's number (approximately 2.71828).

The Normal distribution is fully described by its mean and standard deviation.

Normal Distribution for different parameters

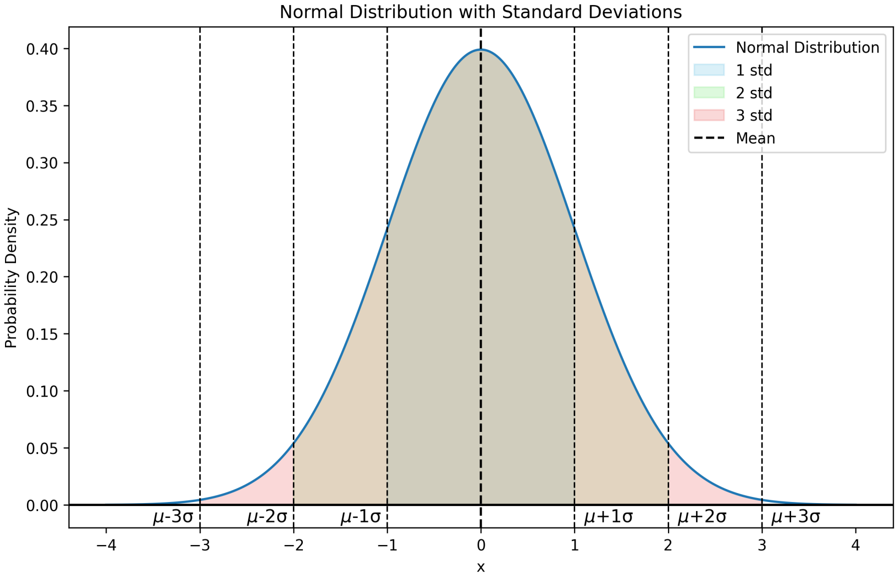


Properties of the Normal Distribution

Some key properties of the Normal distribution:

- **Symmetry:** The distribution is symmetric around the mean (μ).
- **68-95-99.7 Rule:**
  - About 68% of the data falls within one standard deviation of the mean.
  - About 95% falls within two standard deviations.
  - About 99.7% falls within three standard deviations.
- **Asymptotic:** The tails of the Normal distribution approach, but never touch, the horizontal axis.
- **Area under the Curve:** The total area under the Normal curve is 1.

Normal dist and standard deviation



Applications of Normal Distribution

The Normal distribution is used in a wide range of fields and applications:

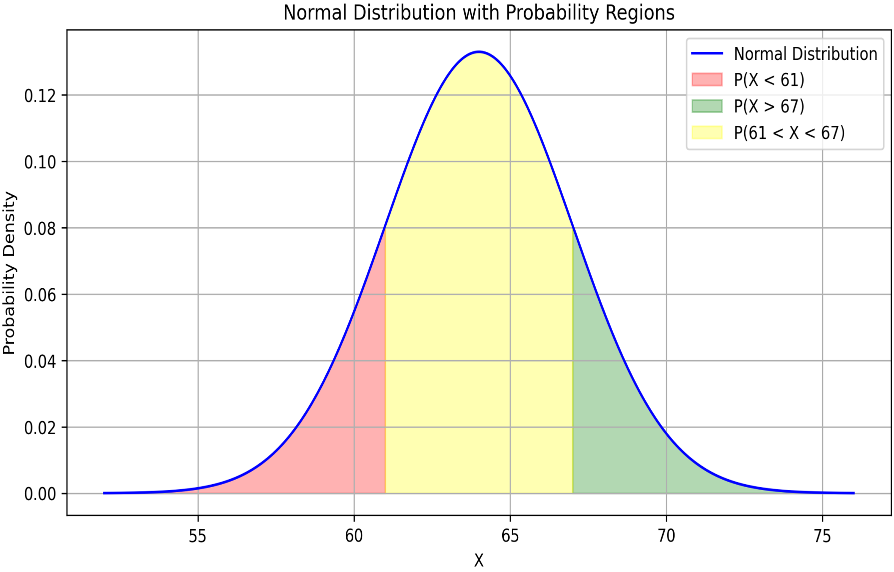
- **Psychology and Education:** Test scores (e.g., IQ scores, SAT scores).
- **Finance:** Stock returns, asset prices.
- **Healthcare:** Measurements of physical traits like height, weight, and blood pressure.
- **Manufacturing:** Quality control, measurement of defects.
- **Natural Sciences:** Errors in scientific measurements, physical phenomena.

The normal distribution is often used in hypothesis testing, confidence intervals, and regression analysis.

Summary

- The Normal distribution is a continuous, symmetric distribution defined by its mean μ and standard deviation σ.
- It has a bell-shaped curve and is used to model many natural and social phenomena.
- The 68-95-99.7 Rule helps in understanding the spread of data in a Normal distribution.
- The Standard Normal distribution has $\mu = 0$ and $\sigma = 1$, and Z-scores are used to standardize data.
- It is widely applied in statistics, finance, healthcare, and more.

Normal Distribution



Probability Distributions: Standard Normal Distributions

What is the Standard Normal Distribution?

- A special case of the Normal distribution.
- It is a Normal distribution with a mean of 0 and a standard deviation of 1.
- Mathematically, the Standard Normal distribution is denoted by:

$$Z \sim N(0,1)$$

  where:
  - Z is the random variable.
  - The mean (μ) is 0.
  - The standard deviation (σ) is 1.

The probability density function (PDF) of the Standard Normal distribution is:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

where z is the standard score (also called the Z-score).

DVAnalytics
Transforming You

Z-Score:

$$z = \frac{x - \mu}{\sigma}$$

Applications:

- **Standardization:**
  - Many datasets do not follow a Standard Normal distribution.
  - Transform them into a Standard Normal form (using Z-scores).
- **Z-Scores:**
  - Gives how many standard deviations a data point is from the mean.
  - Used for identifying outliers and comparing data points from different distributions.
- **Simplicity in Calculation:**
  - Well tabulated, simplifies calculations in hypothesis testing, confidence intervals, and many other statistical methods.
- **Central Limit Theorem (CLT):**
  - The sampling distribution of the sample mean of a large enough sample from any population will be approximately normally distributed, regardless of the original population's distribution.

Z-Scores and Their Importance

The Z-score is a key component of the Standard Normal distribution. It measures the relative position of a data point within a distribution. The Z-score is computed as:

$$z = \frac{x - \mu}{\sigma}$$

where:

- x is the data point.
- μ is the mean of the distribution.
- σ is the standard deviation of the distribution.

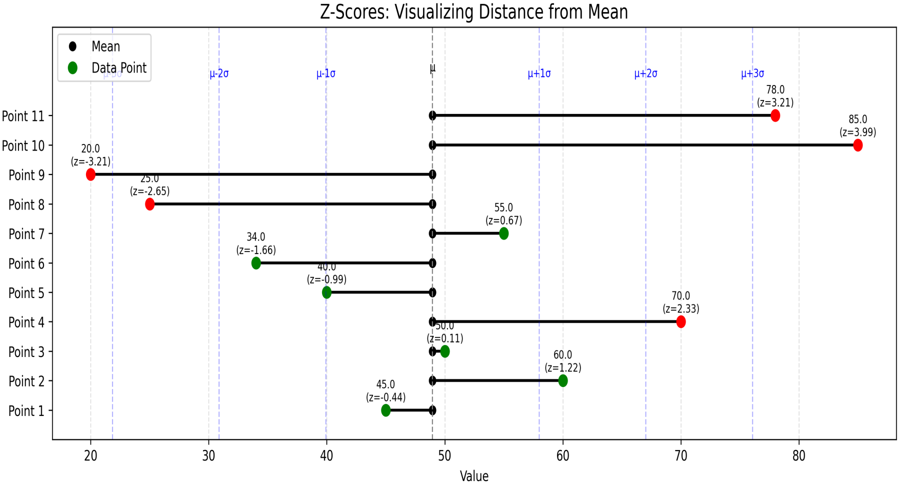
Interpretation of Z-scores:

- A Z-score of 0 means that the data point is exactly at the mean.
- A positive Z-score indicates the data point is above the mean.
- A negative Z-score indicates the data point is below the mean.

Z-scores are essential for:

- Comparing scores from different normal distributions.
- Identifying outliers.

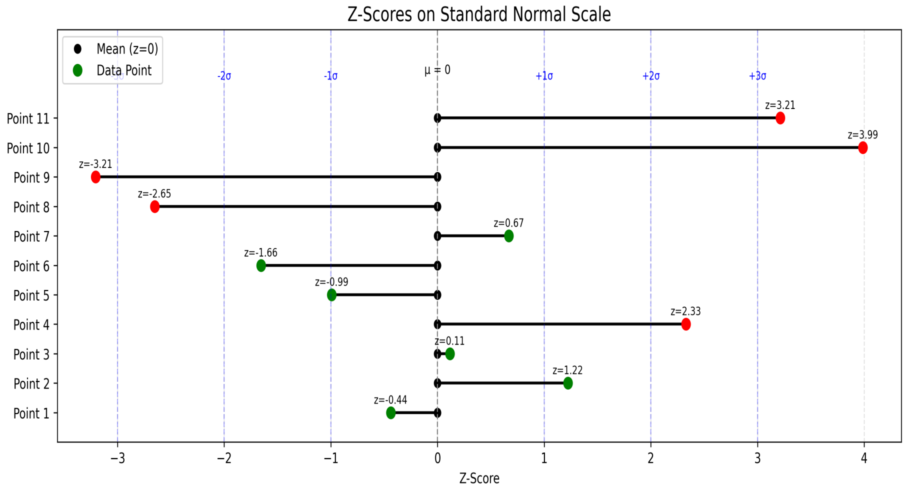
Normal Distribution



Observations from Z-Score Visualization

- **Mean (μ)** is shown as a black dot on each line and marked with a vertical dashed line for reference.
- Each data point is plotted on its own horizontal line, showing its distance from the mean.
- **Z-scores** are annotated beside each point, indicating how many standard deviations away the point is from the mean.
- Points within $\mu \pm 2\sigma$ are marked in green, indicating they fall within the typical range of a normal distribution.
- Points beyond $\mu \pm 2\sigma$ are marked in red, indicating they are **statistical outliers**.
- Vertical dashed lines mark standard deviation levels: $\mu \pm 1\sigma$, $\mu \pm 2\sigma$, and $\mu \pm 3\sigma$, helping to contextualize where each point lies.
- A point near 85 lies just beyond $\mu + 3\sigma$, a rare event (occurs in less than 0.15% of data in a standard normal distribution).
- Similarly, points like 20 and 25 fall below $\mu - 2\sigma$, also considered rare in a normal distribution.

Standard Normal Distribution



Observations from Standard Normal Z-Score Plot

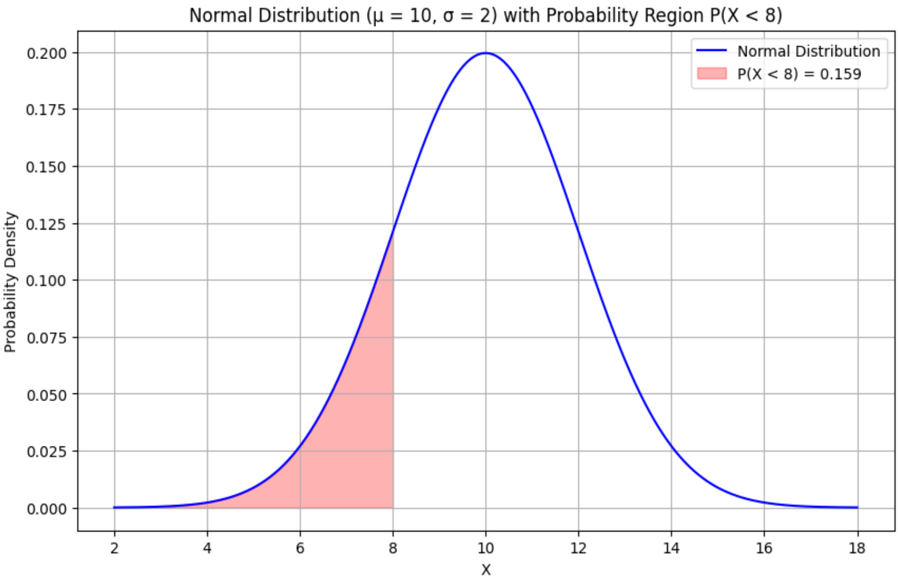
- The mean (μ) of the standard normal distribution is at $z = 0$, marked with a vertical line.
- Each data point's position is shown as a z-score, indicating how far it lies from the mean in units of standard deviation.
- **Green points** represent values within $\pm 2\sigma$, considered typical or common.
- **Red points** indicate data that lies beyond $\pm 2\sigma$, considered outliers in a standard normal distribution.
- Vertical dashed lines mark $z = \pm 1, \pm 2, \pm 3$, helping students understand how data is spread around the mean.
- The standard normal scale provides a universal reference to compare different datasets, regardless of their original units.
- The transformation helps visualize the **relative rarity** of extreme points.
- For example, the point at $z > +3$ is extremely rare, appearing in less than 0.15% of standard normal data.

Comparison: Normal vs Standard Normal Representation

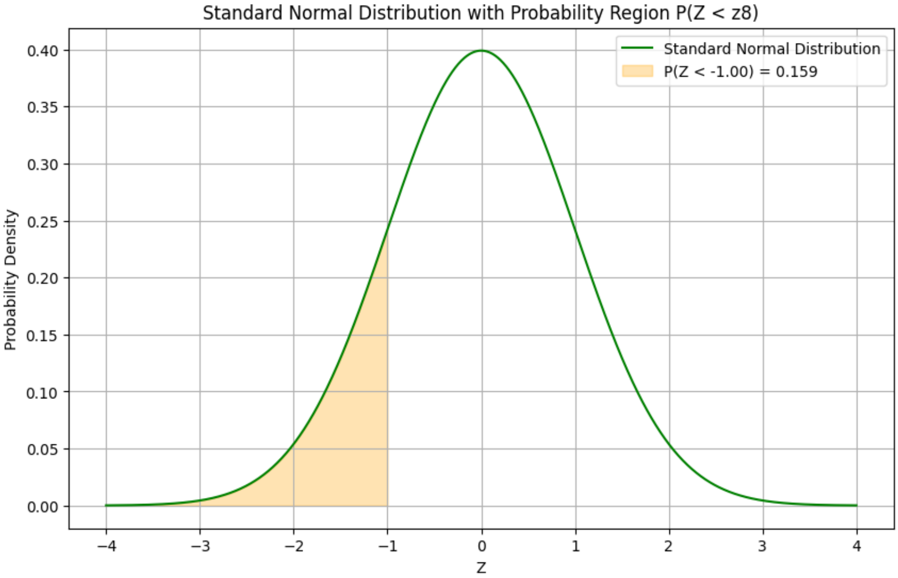
Feature	Normal Distribution Plot	Standard Normal Plot
X-Axis Scale	Original values (e.g., 45, 60, 85)	Standardized z-scores (z)
Mean Location	Varies with data (e.g., $\mu \approx 50$)	Fixed at $z = 0$
Interpretation of Distance	Contextual: depends on unit scale	Universal: measured in σ
Outliers Detection	Based on absolute value range	Based on $\lvert z \rvert > 2$ or $\lvert z \rvert > 3$
Standard Deviation Lines	$\mu \pm 1\sigma$, $\mu \pm 2\sigma$, etc.	$z = \pm 1, \pm 2, \pm 3$
Visualization Purpose	Shows actual data spread and units	Shows relative spread across datasets
Color Coding	Based on deviation from mean	Same (based on z magnitude)

DV Analytics Transforming You

Normal distribution curve



Standard Normal distribution curve



Standard Normal Table (Z-Table)

The Standard Normal distribution has been extensively tabulated, with values for the cumulative distribution function (CDF) at different Z-scores. These values give the probability that a random variable is less than or equal to a given value in the Standard Normal distribution.
For example, the Z-score table gives us:

$$P(Z < 1.96) = 0.9750$$

This means that the probability of a random variable being less than 1.96 in the Standard Normal distribution is 97.5%.

Z-tables are essential for:

- Finding probabilities for areas under the Normal curve.
- Hypothesis testing and confidence interval calculations.
- Determining critical values for statistical tests.

Application of the Standard Normal Distribution

The Standard Normal distribution is used in various areas of statistics and data analysis:

- **Hypothesis Testing:** In hypothesis testing, we standardize the test statistic (such as the Z-test) to compare it against the Standard Normal distribution.
- **Confidence Intervals:** The Z-distribution is used to calculate confidence intervals for population parameters (e.g., mean) when the population standard deviation is known.
- **Central Limit Theorem (CLT):** The Standard Normal distribution is a reference distribution in CLT for approximating the sampling distribution of the sample mean.
- **Data Transformation:** Converting a dataset into a Z-score (Standardizing) allows comparison across different datasets or variables.
- **Error Detection:** Z-scores are useful for identifying outliers in data, i.e., data points that are significantly higher or lower than the mean.

Example:

Let's consider the heights of adult women in a population, which are normally distributed with a mean height of 64 inches and a standard deviation of 3 inches. We want to calculate the probability that a randomly selected woman has a height between 61 and 67 inches.
First, we calculate the Z-scores for 61 inches and 67 inches:

$$z_1 = \frac{61 - 64}{3} = -1 \quad \text{and} \quad z_2 = \frac{67 - 64}{3} = 1$$

Now, we look up the cumulative probabilities for these Z-scores in the Standard Normal distribution table (or use a calculator). The probability for $z_1 = -1$ is 0.1587 and for $z_2 = 1$ is 0.8413.
The probability that a woman's height lies between 61 and 67 inches is:

$$P(61 \leq X \leq 67) = P(z_2) - P(z_1) = 0.8413 - 0.1587 = 0.6826$$

Thus, the probability is approximately 68.26%.

Example: Standardizing a Dataset

Suppose you have the following dataset representing the scores of students in two different classes:

$$\text{Class A:} \quad 70, 75, 80, 85, 90$$

$$\text{Class B:} \quad 55, 60, 65, 70, 75$$

The means and standard deviations for each class are:

$$\mu_A = 80, \quad \sigma_A = 7.9$$

$$\mu_B = 65, \quad \sigma_B = 7.9$$

Class A: Marks and Z-Scores

Table: Class A (Mean = 80, SD = 7.07)

Marks	Z-Score
70	-1.41
75	-0.71
80	0
85	0.71
90	1.41

Table: Class B (Mean = 65, SD = 7.07)

Marks	Z-Score
55	-1.41
60	-0.71
65	0
70	0.71
75	1.41

The Z-scores are identical, indicating that a score of 85 in Class A and a score of 70 in Class B are equally above their respective means.

Summary

- The Standard Normal distribution is a special case of the Normal distribution with mean 0 and standard deviation 1.
- It is essential for standardizing data and comparing different datasets with varying means and standard deviations.
- Z-scores represent the number of standard deviations a data point is from the mean and are used in hypothesis testing, confidence intervals, and error detection.
- Z-tables help compute cumulative probabilities and critical values for statistical tests.
- The Standard Normal distribution is widely used in fields like statistics, hypothesis testing, data analysis, and quality control.

Summary: Normal Distribution Functions in Excel

Function	**Purpose**	**Typical Use Case**
`NORM.DIST(x, mean, sd, TRUE)`	CDF	Cumulative probability $P(X \leq x)$
`NORM.DIST(x, mean, sd, FALSE)`	PDF	Height of normal curve at x
`NORM.S.DIST(z, TRUE)`	Standard Normal CDF	$P(Z \leq z)$ from z-table
`NORM.S.DIST(z, FALSE)`	Standard Normal PDF	Height at z on standard normal curve
`NORM.INV(prob, mean, sd)`	Inverse CDF	Returns x such that $P(X \leq x) =$ prob
`NORM.S.INV(prob)`	Inverse Std Normal	Returns z such that $P(Z \leq z) =$ prob
`STANDARDIZE(x, mean, sd)`	Z-Score	Converts raw score x to $z = \frac{x-\mu}{\sigma}$

Tip: Use `STANDARDIZE` to normalize values and apply standard normal functions easily.

Probability Distributions: Chi-Square Distribution

Introduction to the Chi-Square Distribution

- The Chi-Square distribution is a continuous probability distribution.
- It is used primarily in hypothesis testing, especially in goodness-of-fit tests.
- It is defined by the sum of the squares of independent standard normal random variables.

Properties of the Chi-Square Distribution

- The Chi-Square distribution is parameterized by the degrees of freedom (df), which is typically denoted as k.
- The distribution is skewed to the right, especially for smaller degrees of freedom.
- Mean $= k$ (where k is the degrees of freedom).
- Variance $= 2k$.

Chi-Square Distribution Formula

The probability density function (PDF) of the Chi-Square distribution is given by:

$$f(x;k) = \frac{x^{(k/2)-1}e^{-x/2}}{2^{k/2}\Gamma(k/2)}, \quad x \geq 0,\ k > 0$$

Where:

- x is the random variable.
- k is the degrees of freedom.
- Γ is the Gamma function.

Derivation of the Chi-Square Distribution

The Chi-Square distribution can be derived from the normal distribution. Let $Z_1, Z_2, \ldots, Z_k$ be independent standard normal variables. The Chi-Square random variable X is defined as:

$$X = Z_1^2 + Z_2^2 + \cdots + Z_k^2$$

Derivation Steps

- Each Z_i follows the standard normal distribution, i.e., $Z_i \sim N(0, 1)$.
- The square of a standard normal variable Z_i^2 follows a Chi-Square distribution with 1 degree of freedom.
- The sum of squares of independent standard normal variables follows a Chi-Square distribution with k degrees of freedom.

DV Analytics
Transforming You

Example 1: Deriving Chi-Square for $k = 2$

Let's consider the case where $k = 2$, so we have two independent standard normal variables Z_1 and Z_2. Then the Chi-Square random variable X is:

$$X = Z_1^2 + Z_2^2$$

Each $Z_i \sim N(0, 1)$, so we now have two independent variables squared and summed.
The distribution of Z_1^2 and Z_2^2 is known as a Chi-Square distribution with 1 degree of freedom.
Thus, $X = Z_1^2 + Z_2^2 \sim \chi_2^2$ (Chi-Square distribution with 2 degrees of freedom).

Example 2: Deriving Chi-Square for $k = 3$

Consider the case where $k = 3$. Now, we have three independent standard normal variables Z_1, Z_2, Z_3. The Chi-Square random variable X is:

$$X = Z_1^2 + Z_2^2 + Z_3^2$$

Each $Z_i \sim N(0, 1)$, so the sum of their squares follows a Chi-Square distribution with 3 degrees of freedom. That is:

$$X \sim \chi_3^2$$

This is a Chi-Square distribution with 3 degrees of freedom.

Chi-Square Distribution with 2 and 3 Degrees of Freedom

For $k = 2$:

$$f(x; 2) = \frac{e^{-x/2}}{2}, \quad x \geq 0$$

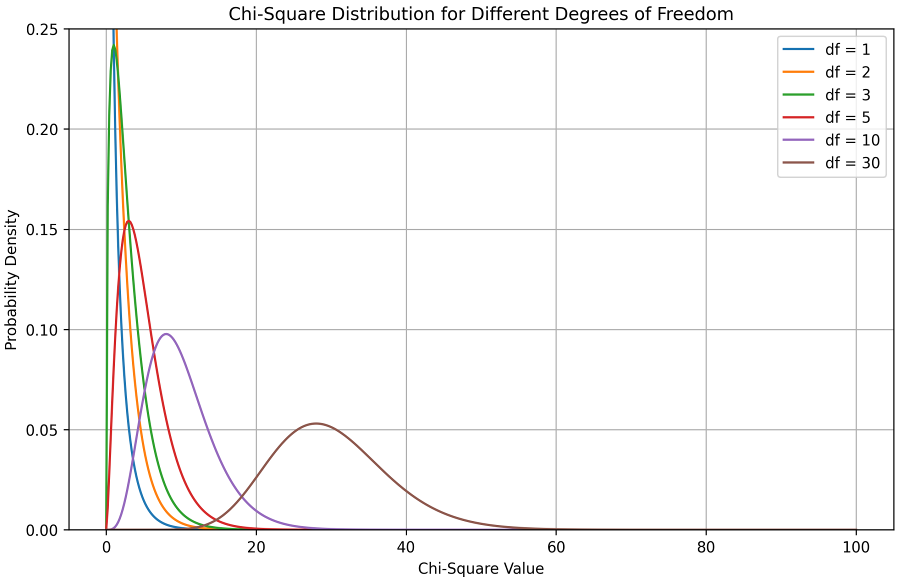
For $k = 3$:

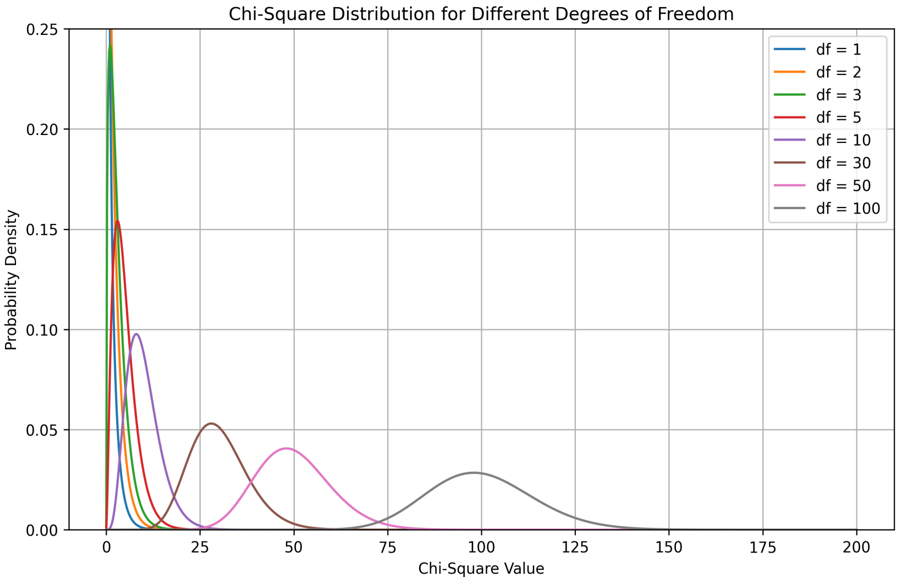
$$f(x; 3) = \frac{x^{(3/2)-1} e^{-x/2}}{2^{3/2}\Gamma(3/2)}, \quad x \geq 0$$

We can observe that as the degrees of freedom increase, the distribution becomes more symmetric.

Effect of Degrees of Freedom on the Shape of the Distribution

- The Chi-Square distribution is positively skewed when k is small.
- As the degrees of freedom increase, the distribution becomes more symmetric.
- For very large degrees of freedom, the distribution approximates a normal distribution.


Chi-Square Distribution for Different Degrees of Freedom
Probability Density
Chi-Square Value
0.00
0.05
0.10
0.15
0.20
0.25
0
20
40
60
80
100
df = 1
df = 2
df = 3
df = 5
df = 10
df = 30



DV Analytics
Transforming You

When to Use the Chi-Square Distribution

The Chi-Square distribution is used in the following cases:

- **Goodness-of-fit tests**: To determine how well observed data fits a specific distribution (e.g., testing whether a die is fair).
- **Test of independence**: Used in contingency table analysis to test if two categorical variables are independent.
- **Test for variance**: In cases where we want to test if the variance of a population is equal to a specified value (e.g., testing if the variance of a sample matches the variance of the population).

Chi-Square Distribution Functions in Excel

1. `CHISQ.DIST(x, deg_freedom, cumulative)`
   - `x` – Test statistic
   - `deg_freedom` – Degrees of freedom
   - `cumulative` – TRUE for CDF $P(X \leq x)$, FALSE for PDF
2. `CHISQ.INV(probability, deg_freedom)`
   - Returns the value of x such that $P(X \leq x) =$ probability
3. `CHISQ.INV.RT(probability, deg_freedom)`
   - Returns x such that $P(X \geq x) =$ probability — commonly used for hypothesis testing

Summary: Chi-Square Functions in Excel

Function	Purpose	Typical Use Case
`CHISQ.DIST(x, df, TRUE)`	CDF	Cumulative left-tail probability $P(X \leq x)$
`CHISQ.DIST(x, df, FALSE)`	PDF	Height of the chi-square curve at value x
`CHISQ.DIST.RT(x, df)`	Right-tail probability	Returns $P(X \geq x)$, used in right-tailed hypothesis tests
`CHISQ.INV(prob, df)`	Inverse CDF	Returns x such that $P(X \leq x) =$ prob
`CHISQ.INV.RT(prob, df)`	Inverse Right-Tail	Returns x such that $P(X \geq x) =$ prob, for critical values
`CHISQ.TEST(actual_range, expected_range)`	p-value from test	Returns $P(X \geq x)$ for a test of independence or goodness-of-fit

DVAnalytics Transforming You

Probability Distributions: t-Distribution

Introduction to t-Distributions

- The t-distribution is a family of probability distributions that arise when estimating population parameters from a small sample.
- It is used in hypothesis testing, particularly for small sample sizes.
- It is similar to the normal distribution but has heavier tails.
- The distribution is defined by a single parameter: the degrees of freedom (df).

t-Distribution vs Normal Distribution

- The t-distribution approaches the standard normal distribution as the sample size increases.
- For large sample sizes (df > 30), the t-distribution and the normal distribution are nearly identical.
- The t-distribution has thicker tails, which allows for more variability in smaller samples.
- This characteristic helps account for the additional uncertainty in estimating the population mean from a small sample.

Degrees of Freedom (df)

- Degrees of freedom (df) refer to the number of independent pieces of information used to estimate a statistical parameter.
- In the context of the t-distribution, df is typically $n - 1$, where n is the sample size.
- As df increases, the t-distribution approaches the normal distribution.

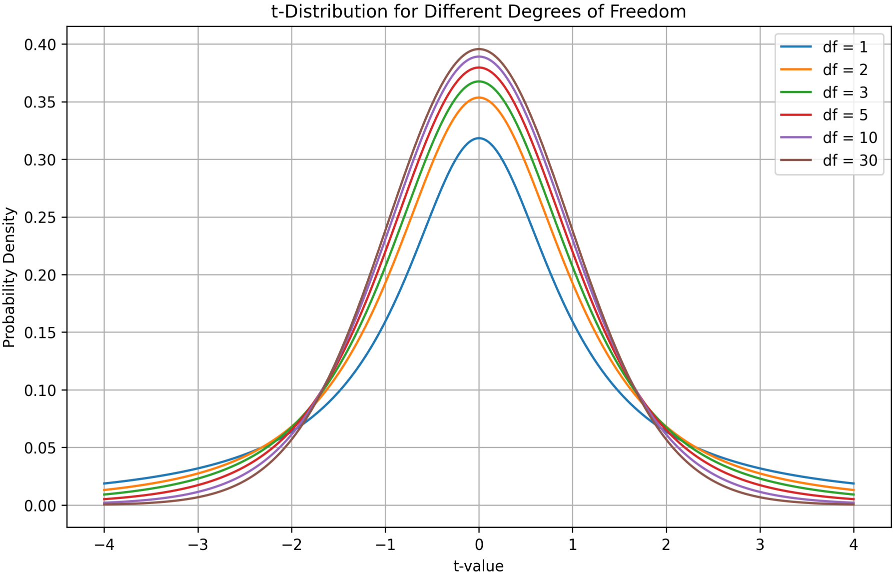
Probability Density Function (PDF) of the t-Distribution

The probability density function (PDF) of the t-distribution is given by:

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Where:

- ν is the degrees of freedom,
- Γ is the Gamma function.


t-Distribution for Different Degrees of Freedom
Probability Density
t-value
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
−4
−3
−2
−1
0
1
2
3
4
df = 1
df = 2
df = 3
df = 5
df = 10
df = 30

Summary

- The t-distribution is used for small sample sizes and is particularly important in hypothesis testing.
- As the sample size increases, the t-distribution approaches the normal distribution.
- The degrees of freedom determine the shape of the t-distribution.
- The t-statistic is used to perform hypothesis tests, and the p-value helps make a decision.

Summary: t-Distribution Functions in Excel

Function	Purpose	Typical Use Case
`T.DIST(x, df, TRUE)`	CDF (left-tail)	Returns $P(T \leq x)$
`T.DIST.RT(x, df)`	Right-tail probability	Returns $P(T \geq x)$; common for one-tailed tests
`T.DIST.2T(x, df)`	Two-tail probability	Returns $P(\lvert T\rvert \geq \lvert x\rvert)$; common for two-tailed tests
`T.INV(prob, df)`	Inverse CDF (left-tail)	Returns x such that $P(T \leq x) =$ prob
`T.INV.2T(prob, df)`	Inverse two-tail	Returns x such that $P(\lvert T\rvert \geq \lvert x\rvert) =$ prob
`T.TEST(array1, array2, tails, type)`	Hypothesis testing	Returns the p-value for a t-test between two samples

Tip: Use `T.TEST` to compare two means; use `T.DIST.RT` or `T.DIST.2T` if manually computing test statistics.

Probability Distributions:F - Distribution

Introduction to F-Distributions

- The F-distribution is a continuous probability distribution that arises from the ratio of two independent chi-squared variables.
- It is used primarily in hypothesis testing, especially for comparing variances in two populations.
- The F-distribution is not symmetric, and it only takes positive values.
- The distribution is determined by two degrees of freedom: one for the numerator and one for the denominator.
- Define the random variable F as the ratio of scaled chi-squared variables:

$$F = \frac{X_1/df_1}{X_2/df_2}$$

F-Distribution vs Other Distributions

- The F-distribution is used in analysis of variance (ANOVA) and regression analysis.
- Unlike the normal distribution or t-distribution, the F-distribution is not symmetric and is right-skewed.
- It has two parameters, the degrees of freedom for the numerator (df_1) and the denominator (df_2).
- As df_1 and df_2 increase, the distribution becomes more symmetric.

Degrees of Freedom (df) in F-Distribution

- The degrees of freedom are crucial in determining the shape of the F-distribution.
- The numerator degrees of freedom (df_1) typically represent the number of groups or treatments.
- The denominator degrees of freedom (df_2) often represent the number of observations within each group or the error degrees of freedom.
- The shape of the distribution depends on both df_1 and df_2. Larger degrees of freedom lead to a more symmetric distribution.

F-Distribution

- The F-distribution is a continuous probability distribution.
- It arises in comparing two sample variances (e.g., ANOVA, regression).
- It is defined by two parameters: degrees of freedom:

$$F_{m,n} \text{ where } m = \text{df}_{\text{numerator}}, \; n = \text{df}_{\text{denominator}}$$

- The distribution is right-skewed and non-negative.
- A large F-statistic suggests a significant difference between variances or group means.

Critical F-value and Alpha

- α is the significance level (e.g., 0.05).
- Understand α as area under the curve.
- It defines the probability of rejecting the null hypothesis when it's true (Type I error).
- The critical value is denoted:

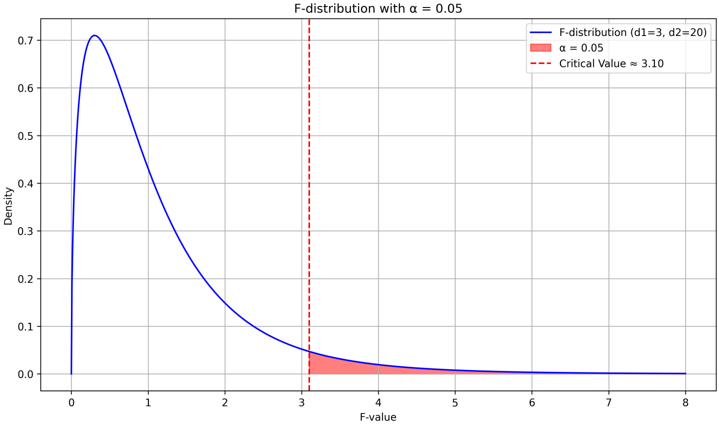
$$F_{\alpha,m,n}$$

- This value cuts off the right tail of the distribution such that:

$$P(F > F_{\alpha,m,n}) = \alpha$$

F-Distributions

- The red area under the curve represents α.
- $F_{\alpha,m,n}$ is the value such that the area to its right equals α.
- Prob of $F_{3,20} < 3.10 = 1 - F_{0.05,3,20}$.



Applications of the F-Distribution

The F-distribution is widely used in several applications, including:

- **Analysis of Variance (ANOVA)**: Used to compare the means of three or more groups. The F-test is used to determine if the group means are significantly different.
- **Regression Analysis**: The F-distribution is used to test the overall significance of a regression model, comparing the explained variance to the unexplained variance.
- **Testing Variance Ratios**: Used to test the hypothesis that two populations have equal variances.
- **Model Comparisons**: The F-distribution can be used to compare the fit of two competing models in statistics, such as nested models.

F-Distribution Reciprocal Property

- The F-distribution arises when comparing two sample variances:

$$F = \frac{s_1^2}{s_2^2} \sim F_{d_1,d_2}$$

- Used in ANOVA and hypothesis tests for comparing population variances.
- Typically, we use the **right-tail** of the distribution for significance tests.
- The F-distribution is asymmetric and right-skewed. However, it exhibits a special reciprocal relationship between the degrees of freedom.

Reciprocal Identity of F-distributions

Fundamental Property:

$$F \sim F_{d_1,d_2} \quad \Rightarrow \quad \frac{1}{F} \sim F_{d_2,d_1}$$

Implication on tail probabilities:

$$P(F_{d_1,d_2} > f) = P\left(F_{d_2,d_1} < \frac{1}{f}\right)$$

- A **right-tail** probability in one F-distribution becomes a **left-tail** probability in the reciprocal.
- Especially useful when test statistic is reversed.

Summary

- The F-distribution is not symmetric — but it has a reciprocal symmetry:

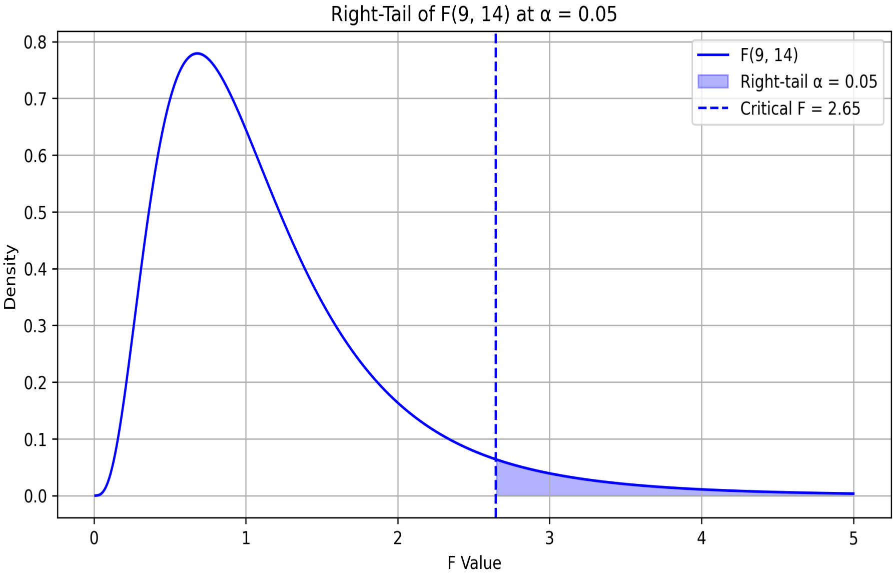
$$F \sim F_{d_1,d_2} \Rightarrow \frac{1}{F} \sim F_{d_2,d_1}$$

- This allows us to:
  - Transform right-tail areas into left-tail areas
  - Understand the behavior of reversed F-ratios
  - Simplify computations across software and tables

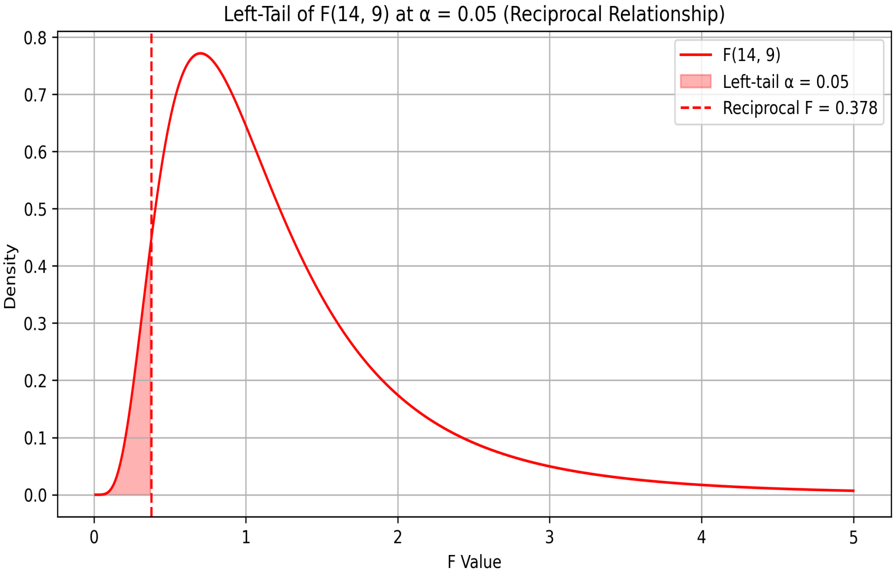
Summary

- The F-distribution is used to compare variances between two or more populations.
- It is commonly applied in hypothesis testing (e.g., comparing variances, ANOVA).
- The distribution is determined by two degrees of freedom: numerator (df_1) and denominator (df_2).
- A large F-statistic suggests a significant difference between variances or group means.
- The F-distribution is derived from the ratio of two chi-squared distributions, and it plays a key role in regression analysis, model comparisons, and hypothesis testing.

F-distribution



Reciprocal F-distribution



Summary: F-Distribution Functions in Excel

Function	Purpose	Typical Use Case
`F.DIST(x, df1, df2, TRUE)`	CDF (left-tail)	Returns $P(F \leq x)$; cumulative probability from the left
`F.DIST.RT(x, df1, df2)`	Right-tail probability	Returns $P(F \geq x)$; used in right-tailed F-tests
`F.INV(prob, df1, df2)`	Inverse CDF (left-tail)	Returns x such that $P(F \leq x) =$ prob
`F.INV.RT(prob, df1, df2)`	Inverse right-tail	Returns x such that $P(F \geq x) =$ prob; critical value for right-tailed test
`F.TEST(array1, array2)`	p-value for F-test	Returns the right-tailed p-value for comparing variances of two samples

Tip: Use `F.TEST` for variance comparisons. Use `F.INV.RT` to find the critical value from a significance level.

When to Use Each Distribution

- **Binomial**: Binary outcomes, fixed trials (e.g., coin flips, pass/fail tests).
- **Poisson**: Count of events in a fixed interval (e.g., arrivals per hour).
- **Normal**: Continuous data, central limit theorem applies (e.g., IQ scores).
- **Chi-Square**: Hypothesis testing for categorical data (e.g., independence test).
- **t-Distribution**: Small sample means, unknown variance (e.g., student test scores).
- **F-Distribution**: Comparing variances or ANOVA (e.g., multiple group comparisons).

Central Limit Theorem (CLT)

Introduction

- The Central Limit Theorem (CLT) is a fundamental theorem in probability theory.
- It states that, under certain conditions, the sum (or average) of a large number of independent and identically distributed (i.i.d.) random variables follows a normal distribution, regardless of the original distribution.

Formal Definition

Let $X_1, X_2, \ldots, X_n$ be a sequence of i.i.d. random variables with mean μ and variance σ^2. Define the sample mean as:

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Then, as $n \to \infty$,

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1).$$

This means the standardized sample mean converges in distribution to a standard normal distribution.

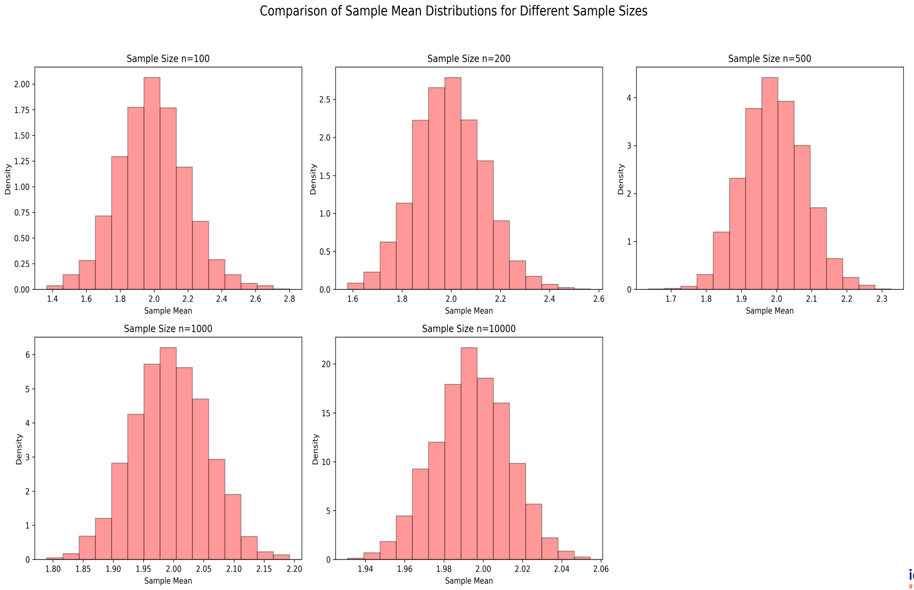
Why is CLT Important?

- Enables statistical inference when population distribution is unknown.
- Justifies the normality assumption in many practical applications.
- Forms the basis for confidence intervals and hypothesis testing.

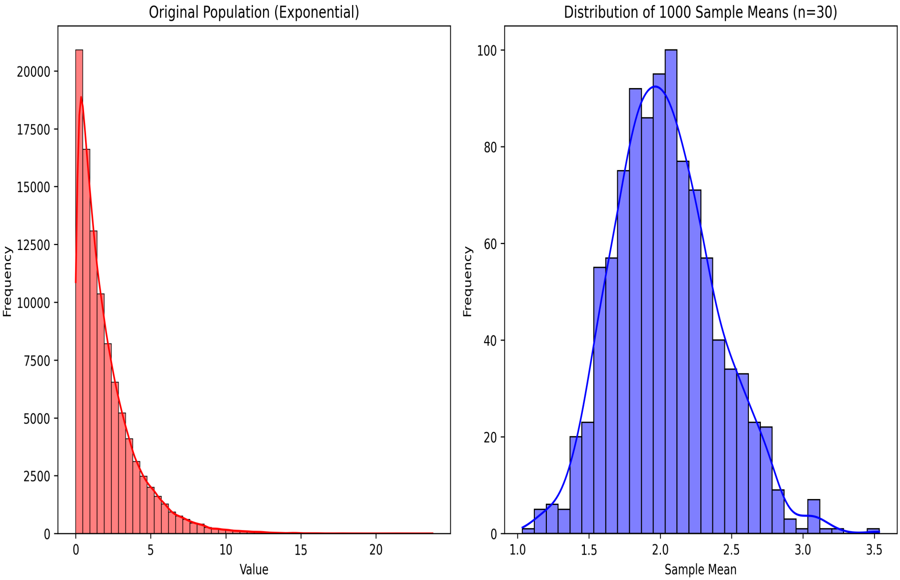
Illustration: CLT in Action

- Consider different distributions (e.g., uniform, exponential, binomial).
- Take samples of increasing size and compute sample means.
- Observe how the distribution of sample means approaches a normal distribution.

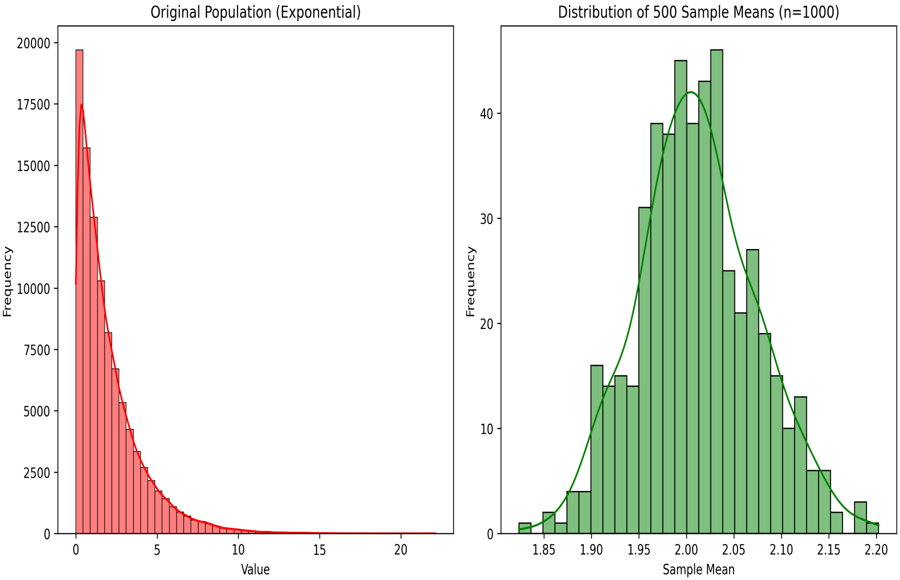
For 3000 sample means



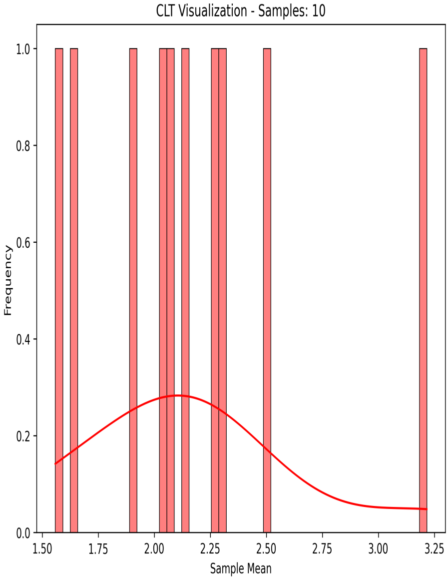
For sample size 30 and 1000 samples

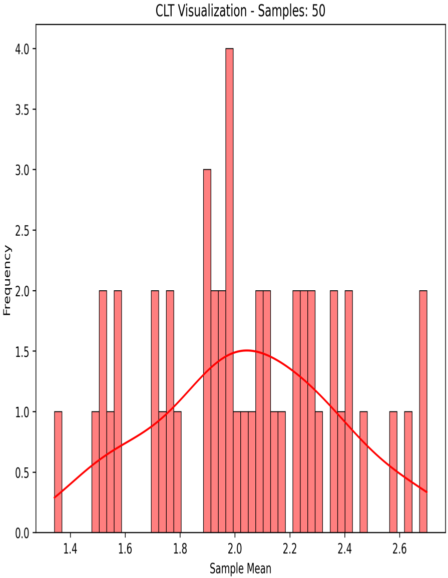


For sample size 1000 and 500 samples

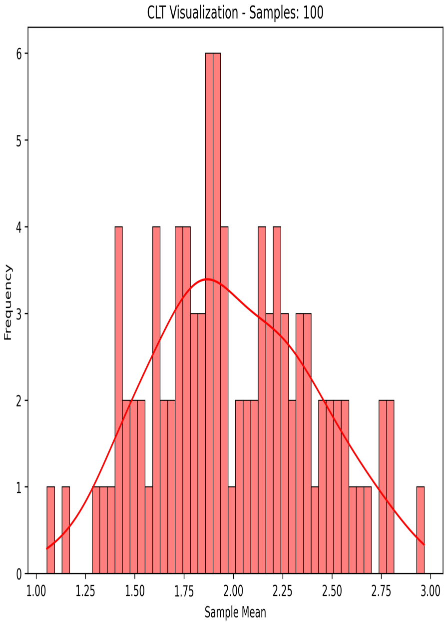


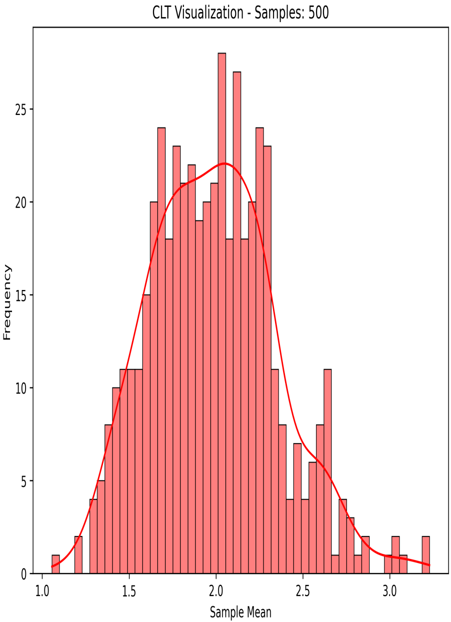
For Fixed sample size



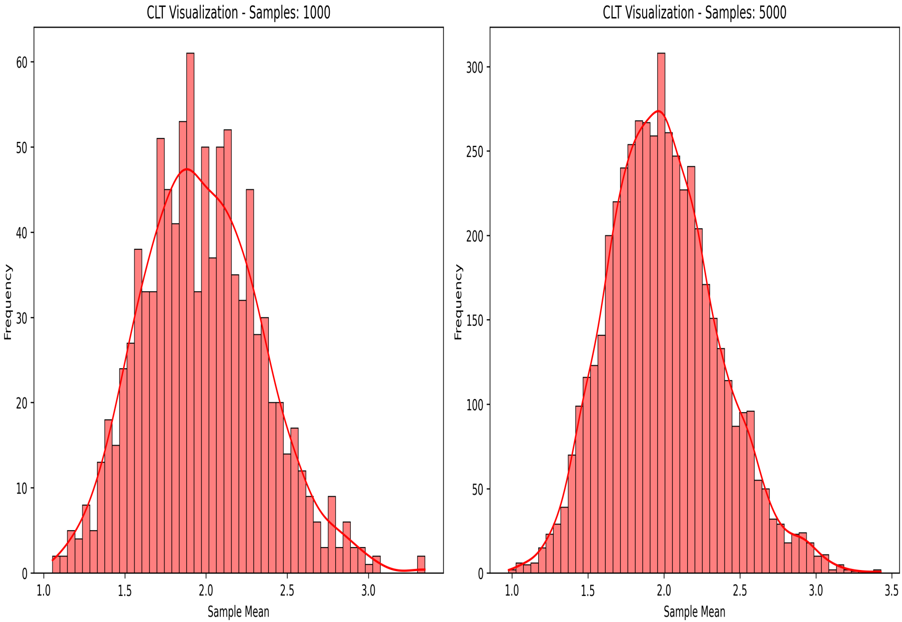


For Fixed sample size

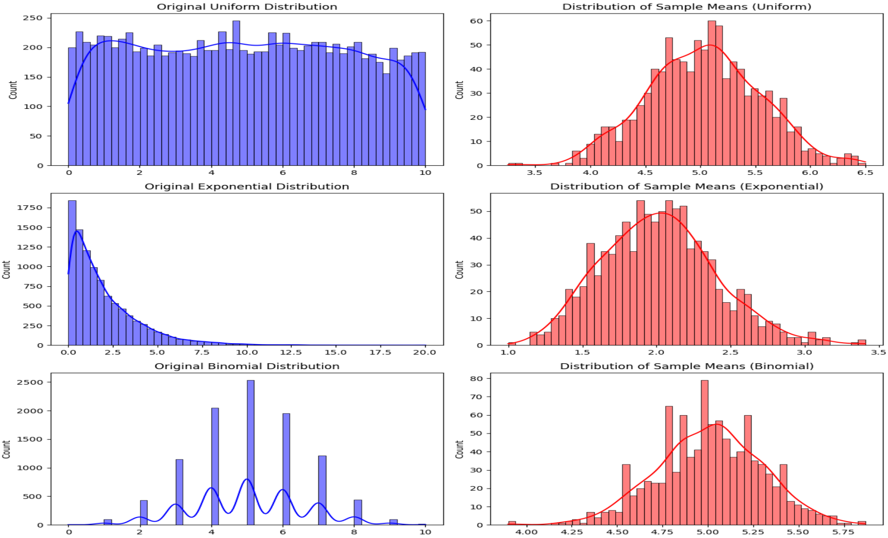




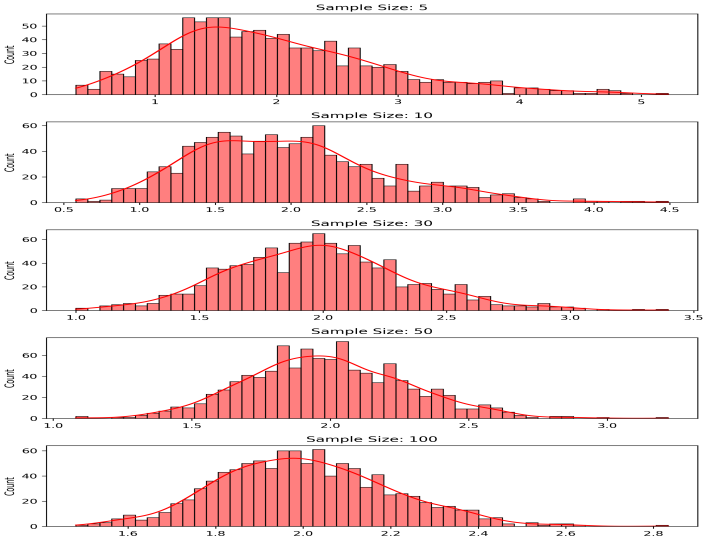
For Fixed sample size



For any distribution



For different sample sizes



Example Application

Suppose the heights of students in a university are not normally distributed. However:

- If we randomly select 30 students and compute the sample mean,
- The sample mean follows approximately a normal distribution.
- This allows us to make statistical inferences using normality-based methods.

Limitations and Assumptions

- CLT requires a sufficiently large sample size ($n \geq 30$ is a common rule of thumb).
- Assumes independence of random variables.
- Works best when individual observations have finite variance.

CONFIDENCE INTERVALS

Introduction to Confidence Intervals

- A confidence interval (CI) is a range of values, derived from sample statistics, that is likely to contain the true population parameter.
- It provides an estimate along with an indication of the estimate's reliability.
- Confidence intervals are widely used in inferential statistics to make predictions about population parameters based on sample data.
- The confidence level represents the probability that the confidence interval contains the true population parameter.
- Common confidence levels: **90%, 95%, 99%**.
- A higher confidence level means a wider interval, increasing the certainty of capturing the population parameter.
- Example: "We are 95% confident that the true mean lies between 50 and 60."

Formula for Confidence Interval

- For a population mean μ with known standard deviation σ, the confidence interval is given by:

$$\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

  where:

  - $\bar{X}$ = sample mean,
  - $Z_{\alpha/2}$ = critical value from the standard normal distribution,
  - σ = population standard deviation,
  - n = sample size.

Interpreting Confidence Intervals

- A 95% confidence interval means that if we were to take 100 different samples and compute a CI for each, about 95 of them would contain the true population parameter.
- It does not mean that there is a 95% probability that the true parameter is within the given interval.

Effect of Sample Size on CI

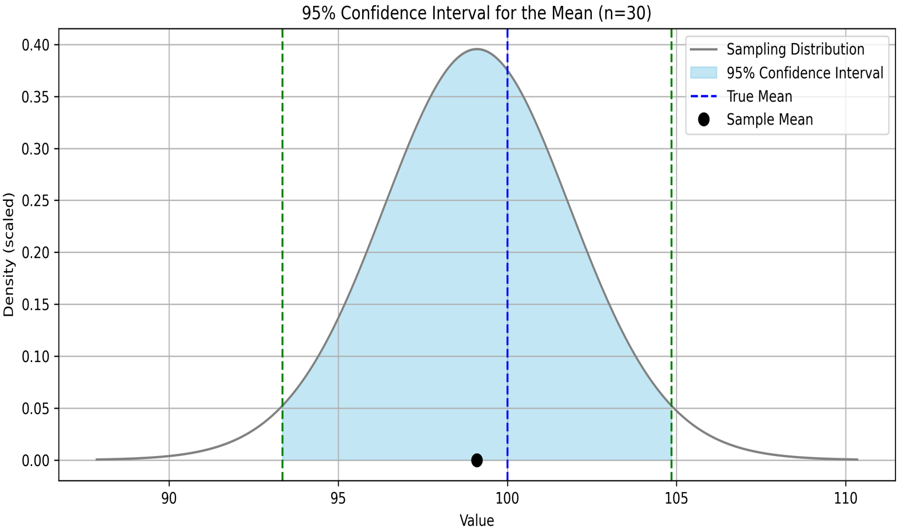
- Larger sample sizes lead to narrower confidence intervals, increasing precision.
- Smaller sample sizes lead to wider confidence intervals, reflecting greater uncertainty.

Effect of Confidence Level on CI

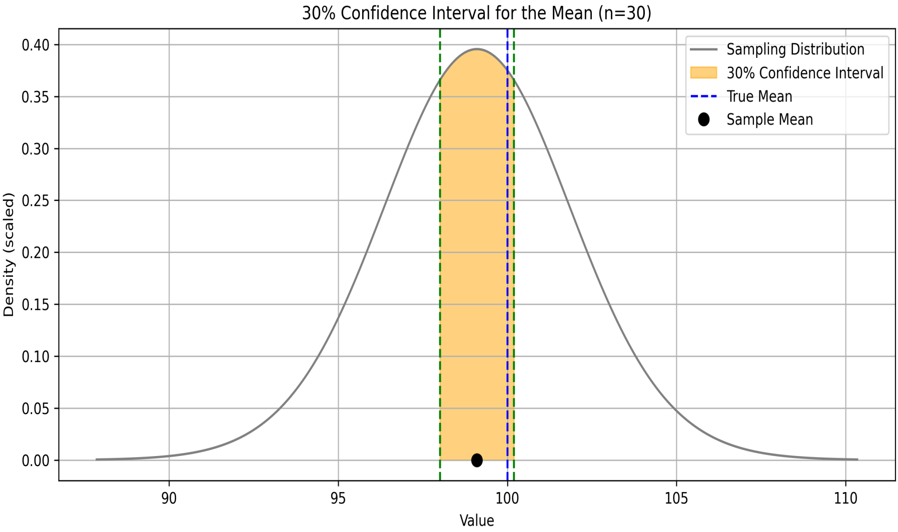
- Increasing the confidence level (e.g., from 95% to 99%) makes the interval wider to ensure greater certainty.
- Decreasing the confidence level (e.g., from 95% to 90%) makes the interval narrower but with less certainty.

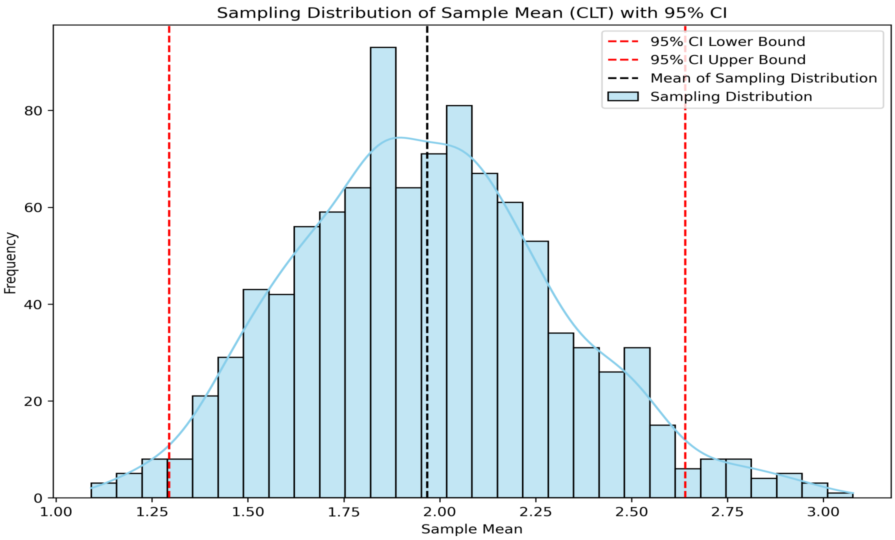
Conclusions from Different Confidence Intervals

- **Wide Confidence Interval**: Indicates more variability in data or smaller sample size.
- **Narrow Confidence Interval**: Suggests more precise estimates due to a larger sample size or lower variability.
- **Overlapping Intervals**: If two confidence intervals overlap, it suggests no significant difference between compared parameters.
- **Non-Overlapping Intervals**: Suggests a statistically significant difference between parameters.

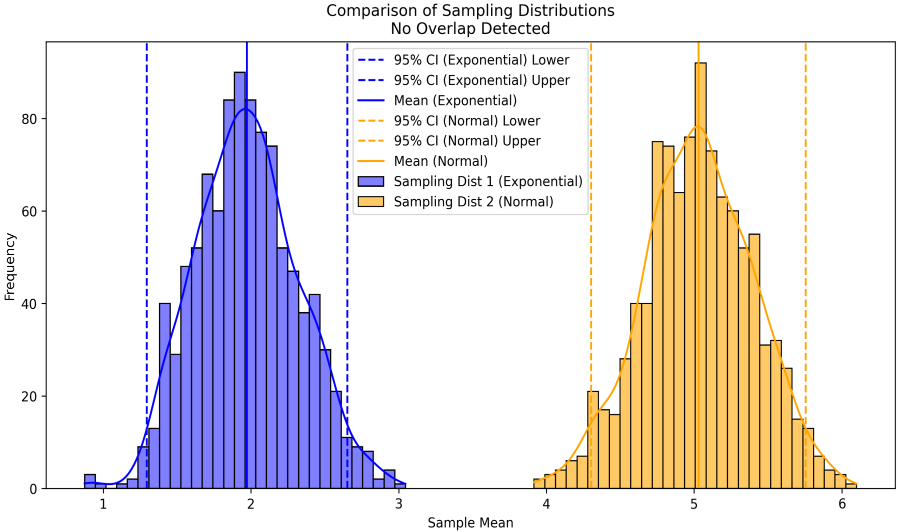


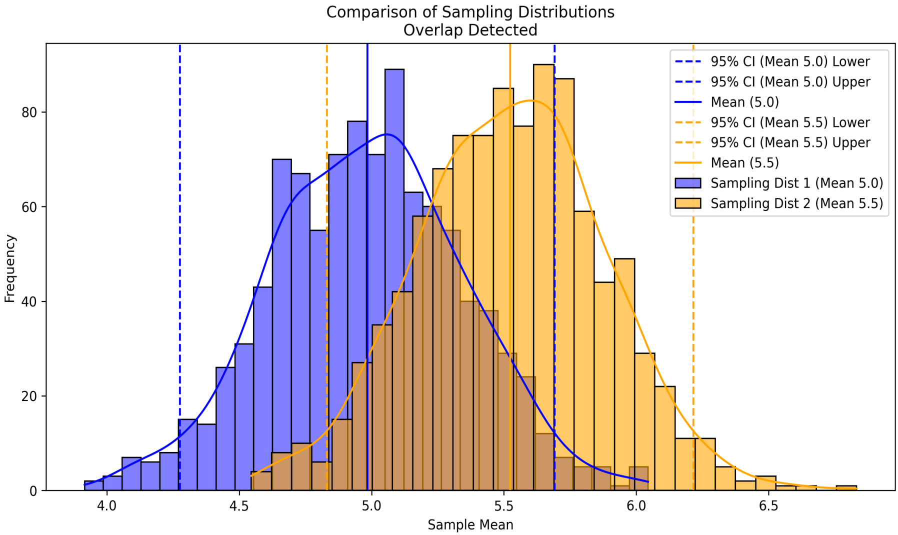
DV Analytics
Transforming You


30% Confidence Interval for the Mean (n=30)
Sampling Distribution
30% Confidence Interval
True Mean
Sample Mean
Density (scaled)
Value
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
90
95
100
105
110


Sampling Distribution of Sample Mean (CLT) with 95% CI
95% CI Lower Bound
95% CI Upper Bound
Mean of Sampling Distribution
Sampling Distribution
Frequency
0
20
40
60
80
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
Sample Mean

Confidence Intervals of Two Samples





DV Analytics
Transforming You

Real-World Applications

- Confidence intervals are used in medical research to determine the effectiveness of treatments.
- In finance, they help estimate stock market trends.
- In quality control, they assess manufacturing consistency.

Limitations of Confidence Intervals

- Assumes random sampling; biased samples affect validity.
- Relies on the assumption that data follow a particular distribution.
- Wider intervals may not always provide useful conclusions.

Conclusion

- Confidence intervals provide a way to estimate population parameters with a degree of certainty.
- The width of the interval depends on sample size and confidence level.
- Proper interpretation is crucial for making informed statistical decisions.

THANK YOU