DMV_U1

Random Variable (RV) — The Big Picture

A **random variable** is like a bridge between a realworld random experiment and numbers we can work with mathematically.

- It **assigns a numerical value** to each possible outcome of a random process.
- Formally, it's a function from the *sample space* (all possible outcomes) to real numbers.

Example:

If we roll a die, the random variable XXX could be defined as "the number showing on the die." Possible outcomes $\rightarrow \{1, 2, 3, 4, 5, 6\}$.

1. Discrete Random Variable (DRV)

- **Definition:** Can only take **finite** or **countably infinite** values.
- **Key property:** We can list all possible values, even if the list is very long.
- Probability tool: Probability Mass Function (PMF) P(X=x)P(X = x)P(X=x) → gives the probability of each possible value.
- The sum of all probabilities is **1**.

Example: Number of students in a class.

• Possible values: {0, 1, 2, 3, ...}

• We can assign:

 $P(X = 30) = 0.25, \quad P(X = 31) = 0.15, \quad ...$

Common discrete distributions: Binomial, Poisson, Geometric.

2. Continuous Random Variable (CRV)

- **Definition:** Takes **uncountably infinite** values within an interval or multiple intervals.
- **Key property:** The probability of taking any *exact* single value is **zero**. We talk about **probability over intervals** instead.
- Probability tool: Probability Density Function (PDF) $f(x)f(x)f(x) \rightarrow$
 - The area under the curve of the PDF over an interval gives the probability.
 - \circ The total area under the curve = 1.

Example: Height of a person.

- Possible values: 150.000... cm, 150.001 cm, 150.002 cm, ...
- To find the probability that height is between 150 cm and 160 cm:

$$P(150 \leq X \leq 160) = \int_{150}^{160} f(x) \, dx$$

Common continuous distributions: Normal, Uniform, Exponential.

Independence for Random Variables

Just like **two events** are independent if one happening doesn't affect the other, **two random variables** are independent if knowing the value of one gives you **no information** about the value of the other.

Mathematical Definition

The mathematical definition of independence depends on the type of random variable.

 Discrete Case: For discrete variables, X and Y are independent if for all possible values x and y:

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

This means the probability of both events occurring together (the joint probability) is simply the product of their individual probabilities (their marginal probabilities).

 Continuous Case: For continuous variables, independence is defined using their probability density functions (PDFs). X and Y are independent if their joint PDF, f,X,Y(x,y), can be factored into the product of their individual (marginal) PDFs, fX(x) and fY(y):

$$fX,Y(x,y)=fX(x)\cdot fY(y)$$

Intuition and Examples

The core idea is that if X and Y are independent, knowing the value of X doesn't change the **probability distribution** of Y, and vice versa.

• Example 1 (Discrete): Consider two independent coin tosses. Let X be the result of the first toss (1 for heads, 0 for tails) and Y be the result of the second. The probability of getting heads on the first toss is P(X=1)=0.5, and the probability of getting heads on the

second is P(Y=1)=0.5. Since the tosses are independent, the joint probability of getting heads on both is P(X=1,Y=1)=0.25, which is exactly $P(X=1)\cdot P(Y=1)$. This confirms their independence.

Events vs. Random Variables

It's important to distinguish between these two concepts:

- **Event Independence** focuses on whether the **outcome** of one specific event (e.g., getting a 6 on a die) changes the probability of another specific event.
- Random Variable Independence is a more powerful concept that focuses on whether the entire probability distribution of one variable remains unchanged regardless of the value of the other.

Covariance: A Measure of Joint Variability

Covariance is a statistical tool that describes how two random variables change in relation to each other. It indicates the **direction** of the linear relationship between them.

Interpretation of Covariance

- **Positive Covariance** (>0): When one variable tends to increase, the other also tends to increase. Think of a positive trend.
 - Example: As daily temperatures rise (X), ice cream sales also tend to increase (Y).
- Negative Covariance (<0): When one variable increases, the other tends to decrease. This indicates an inverse relationship.
 - Example: The more hours a student spends playing video games (X), the lower their exam score tends to be (Y).
- **Zero Covariance** (=0): There's no consistent linear relationship between the two variables. Their movements are not predictably linked in a linear fashion.

Mathematical Definition

For two random variables, X and Y, the covariance is defined as the expected value of the product of their deviations from their respective means.

 $Cov(X,Y)=E[(X-\mu X)(Y-\mu Y)]$

Where:

- $E[\cdot]$ is the expected value operator.
- $\mu X = E[X]$ is the mean of X.
- $\mu Y = E[Y]$ is the mean of Y.

An alternative, often more practical, formula is:

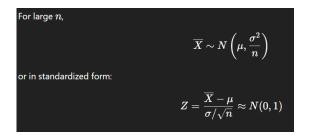
Cov(X,Y)=E[XY]-E[X]E[Y]

Covariance vs. Independence

It's crucial to remember that **zero covariance does not necessarily imply independence**. While independent variables always have a covariance of zero, a covariance of zero only means there is no **linear** relationship. There could still be a non-linear relationship between the variables.

The Central Limit Theorem (CLT) is a statistical principle that states that the distribution of sample means will be approximately normal for a large enough sample size, regardless of the original population's distribution. This powerful theorem is the cornerstone of many statistical methods.

Mathematical Formulation 🗗



Key Conditions and Implications

The CLT holds true under a few key conditions:

- The samples must be **independent** and **identically distributed** (i.i.d.).
- The population must have a **finite** variance.
- The sample size (n) must be **large enough** (a common rule of thumb is n≥30).

The primary implication of the CLT is that it allows us to use normal distribution-based techniques, such as **confidence intervals** and **hypothesis testing**, even when we don't know the original population's distribution.

Example Calculation

Consider a population with a mean of μ =50 and a standard deviation of σ =20. If we take a sample of n=100, we can determine the properties of the sampling distribution of the sample mean.

The standard error is:

 $SE=n\sigma=10020=2$

According to the CLT, the distribution of our sample means (X) will be approximately normal with a mean of 50 and a standard deviation (standard error) of 2. We can write this as $X \sim N(50,22)$. This means we can expect most sample means to fall within a predictable range around 50.

Advantages (Pros)

- 1. **Universality:** Works for any population shape if the sample size is large enough.
- 2. **Foundation for Statistics:** Justifies using normal-based methods for hypothesis testing, confidence intervals, etc.
- 3. **Simplifies Analysis:** Converts complex or unknown distributions into a predictable normal form.
- 4. **Practical Application:** Works well in realworld scenarios where exact population distribution is unknown.

Disadvantages (Cons)

- 1. **Sample Size Requirement:** Needs a sufficiently large nnn for accuracy, especially for skewed or heavy-tailed data.
- 2. **Not Always Exact:** For small samples from non-normal populations, the approximation can be poor.
- 3. **Finite Variance Requirement:** Does not hold for distributions with infinite variance (e.g., Cauchy).
- 4. **Independence Assumption:** Fails if data points are highly dependent or correlated.

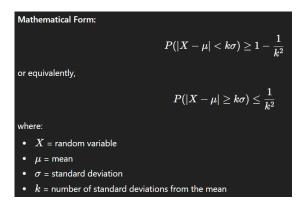
Chebyshev's Inequality

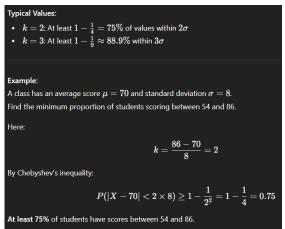
Chebyshev's Inequality states that for any dataset or probability distribution (with finite mean μ and finite standard deviation σ), the proportion of observations lying within k standard deviations of the mean is \boldsymbol{at} least $1{-}1/k^2$, where $k{>}1.$

Key Features:

1. **Distribution-free:** Works for *any* distribution shape (normal, skewed, uniform, etc.)

- 2. **Minimum guarantee:** Gives the smallest possible proportion of values within a range the actual proportion could be higher.
- 3. **Finite variance condition:** Requires σ2\sigma^2σ2 to exist.





Advantages:

- Works without knowing the exact distribution
- Useful for identifying outliers and spread in non-normal data

Limitations:

- Bound is conservative (actual proportion is often much higher)
- Requires finite variance fails for infinite-variance distributions.

Diverse Continuous and Discrete Distributions

A) Discrete Distributions

1. Bernoulli Distribution The Bernoulli distribution models a single random trial with two possible outcomes: **success** (1) or **failure** (0). The probability of success is denoted by p. A single coin toss is a perfect example, where p=0.5 for heads.

• **Probability Mass Function** (PMF): P(X=x)=px(1-p)1-x, for $x \in \{0,1\}$

2. Binomial Distribution

The Binomial distribution describes the number of successes in a fixed number, n, of **independent Bernoulli trials**. Each trial has the same probability of success, p. This is often used to model the number of times a certain outcome occurs in a series of identical experiments, such as counting the number of defective items in a batch of 20 products.

- **PMF:** P(X=k)=(kn)pk(1-p)n-k
- 3. Poisson Distribution The Poisson distribution counts the number of times an event occurs within a specific interval of time or space, assuming the events happen at a constant average rate, λ . This is useful for modeling rare events over a continuous period, like the number of emails a person receives in an hour.
 - **PMF:** $P(X=k)=k!\lambda ke-\lambda$, for k=0,1,2,...

B) Continuous Distributions

- **1. Uniform Distribution** The Uniform distribution models situations where all values within a given interval [a,b] are **equally likely**. This means the probability density is constant across the entire interval. An example is a random number generator that produces a value between 1 and 10, where every number has the same chance of being selected.
 - **Probability Density Function (PDF):** f(x)=b-a1, for $a \le x \le b$
- **2. Normal Distribution** The Normal distribution, also known as the "bell curve," is a symmetric, unimodal distribution defined by its **mean** (μ) and **standard deviation** (σ) .

It is one of the most important distributions in statistics because it describes many natural phenomena, such as human heights, blood pressure, and test scores.

- **PDF:** $f(x) = \sigma 2\pi 1e 2\sigma 2(x \mu)2$
- **3. Exponential Distribution** The Exponential distribution models the **time between events** in a Poisson process. It's often used to describe the waiting time until the next event occurs, assuming the events happen at a constant rate. For example, it can model the time between consecutive bus arrivals at a bus stop.

• **PDF:** $f(x) = \lambda e - \lambda x$, for $x \ge 0$

1. Descriptive Statistics:

Descriptive statistics are the first step in data analysis, providing a concise summary of a dataset. They help us understand the data's central location and how its values are spread out.

Main Measures

• Central Tendency:

- Mean (x⁻): The average value, calculated by summing all values and dividing by the count. It's sensitive to extreme values.
- Median: The middle value of an ordered dataset. It's a robust measure, unaffected by outliers.
- Mode: The value that appears most frequently in the dataset.

• Spread:

- Range: The difference between the maximum and minimum values, indicating the total span of the data.
- O Variance (σ2) & Standard Deviation (σ):
 These measure the average squared deviation (variance) and root mean square deviation (standard deviation) from the mean. A larger value indicates the data is more spread out. For an unbiased estimate of the population variance from a sample, the formula uses 1/(n-1) in the denominator.

Example: X={4,7,7,10,12}

- **Mean:** 54+7+7+10+12=540=8
- **Median:** The sorted data is $\{4,7,7,10,12\}$, so the median is 7.
- **Mode:** 7, because it appears twice.
- **Range:** 12–4=8
- Sample Variance (s2): Using deviations from the mean (-4,-1,-1,2,4), the sum of squares is 16+1+1+4+16=38. The sample variance is s2=5-138=9.5.
- Sample Standard Deviation (s): s=9.5 ≈ 3.082

2. Graphical Statistics:

Graphical statistics use visual tools to quickly reveal patterns, distributions, and relationships within data that may be difficult to discern from numbers alone.

Common Plots

- **Histogram:** Displays the frequency distribution of a numeric variable. It shows the data's shape, skewness, and modality.
- **Boxplot:** A concise summary of the data's quartiles. It shows the median (Q2), the first and third quartiles (Q1, Q3), the interquartile range (IQR), and potential outliers.
- Scatter Plot: Illustrates the relationship between two numeric variables, helping to identify correlation and trends.
- Q-Q Plot: Compares the quantiles of the data to the quantiles of a theoretical distribution (e.g., normal), providing a visual check for how well the data fits that distribution.

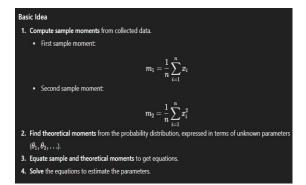
Example (using X={4,7,7,10,12}):

• A **boxplot** of this data would show the median at 7. The first quartile (Q1) would be the median of the lower half {4, 7}, which is 5.5. The third quartile (Q3) would be the median of the upper half {10, 12}, which is 11. The IQR is 11–5.5=5.5.

3. Method of Moments (MoM): Parameter Estimation

The **Method of Moments** is a statistical technique for estimating unknown parameters of a probability distribution by equating **sample moments** (calculated from data) with **theoretical moments** (derived from the distribution).

- **Moments** are numerical measures that describe the shape and characteristics of a distribution.
- The **r-th moment about the origin** is:
- μ^r=E[Xr]
- The **first moment** is the mean, the **second central moment** is the variance, etc.



Procedure

- 1. Compute the first k sample moments (e.g., mean, sample variance) from the data, where k is the number of parameters to be estimated.
- 2. Write the theoretical moments of the chosen distribution in terms of its parameters.
- 3. Set the sample moments equal to the theoretical moments and solve the resulting equations for the parameters.

Example: Estimating the Poisson parameter λ A Poisson distribution has only one parameter, λ , and its theoretical mean is μ = λ .

- 1. From our dataset $X=\{4,7,7,10,12\}$, the sample mean is x=8.
- Equating the sample mean to the theoretical mean, we get λ^=x⁻.
- 3. The MoM estimate for λ is λ^{\wedge} MoM=8.

4. Maximum Likelihood Estimation (MLE): Finding the Best Fit

Maximum Likelihood Estimation is a powerful method for estimating model parameters. It seeks to find the parameter values that make the observed data most probable.

Procedure

- 1. **Likelihood Function:** Write down the likelihood function, $L(\theta)$, which is the product of the probability (or density) of each data point given the parameter(s) θ .
- 2. **Log-Likelihood:** Take the natural logarithm of the likelihood function to simplify calculations: $\ell(\theta) = \log L(\theta)$.
- 3. **Optimization:** Differentiate the log-likelihood function with respect to the parameter(s), set the derivative to zero, and solve for the parameter estimate(s), θ^{\wedge} .

Example A — Bernoulli / Binomial (simple closed form):
$$\text{Data: } n \text{ independent trials, } k \text{ successes. Likelihood for } p \text{:}$$

$$L(p) = p^k (1-p)^{n-k}, \qquad \ell(p) = k \ln p + (n-k) \ln(1-p)$$

$$\text{Differentiate:}$$

$$\frac{d\ell}{dp} = \frac{k}{p} - \frac{n-k}{1-p} = 0 \ \Rightarrow \ \hat{p}_{\text{MLE}} = \frac{k}{n}.$$
 So if 7 heads in 10 tosses, $\hat{p} = 7/10 = 0.7$.

1. Subtypes and Supertypes

 Supertype → A general entity containing attributes common to multiple related entities.

- Subtype → A specialized entity that inherits all attributes of its supertype and may have extra attributes or behaviors.
- Similar to **inheritance** in object-oriented programming but used in database modeling.

Purpose

- 1. **Avoids redundancy** Common attributes are stored only once in the supertype.
- 2. **Encourages specialization** Unique features of subtypes can be stored separately.
- 3. **Improves clarity** Clearly shows relationships and differences between entities.
- 4. **Supports flexibility** New subtypes can be added without changing the supertype.
- 5. **Models real-world hierarchy** Many systems naturally have "general-specific" structures.

Example

- Supertype: Vehicle
 - o Attributes: VehicleID, Model, Manufacturer
- Subtypes:
 - o **Car** → Additional attributes: NumberOfDoors, FuelType
 - o **Bike** → Additional attribute: Type (e.g., mountain, road)

Real-World Use

- Transportation System:
 - o Supertype: Vehicle
 - o Subtypes: Car, Bus, Truck
- Hospital Database:
 - Supertype: Person
 - Subtypes: Patient, Doctor, Staff

Diagram Representation

Vehicle (Supertype)

/ | \

Car Bike Truck (Subtypes)

Implementation Approaches in Databases

- 1. **Single table** for all (with nullable subtype fields).
- 2. **Supertype table** + **separate subtype tables** (linked via primary–foreign key).

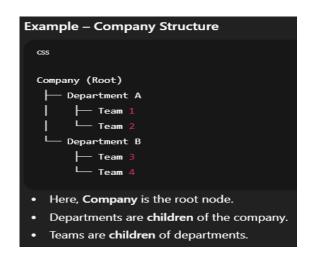
3. **Separate subtype tables only** (less common).

2. Hierarchical Data

- Hierarchical data is organized in a treelike structure where:
 - Each record/node has exactly one parent (except the root).
 - A record can have **zero or more children**.
- The **root** is the top-most node with no parent.
- Relationships are **one-to-many** (parent → multiple children).

Purpose

- 1. **Models natural hierarchies** like organizations, categories, or file systems.
- Efficient navigation when moving topdown or bottom-up in the hierarchy.
- 3. **Clear structure** for representing nested relationships.
- 4. **Logical grouping** of related data under a common parent.



Real-World Uses

- Organizational charts (e.g., CEO → Managers → Employees).
- **File systems** (e.g., Folder → Subfolder → Files).
- **Product categories** in e-commerce.
- **XML/JSON** data storage (tags/nodes form a hierarchy).

Advantages

• Intuitive structure for hierarchical relationships.

- Efficient for queries like "get all sub-items of X".
- Matches many real-world use cases.

Limitations

- Can be harder to update (moving nodes may require multiple updates).
- More complex queries compared to flat relational tables.

3. Recursive Relationships

- A recursive relationship is a relationship where an entity is related to itself.
- The **same table/entity** is used for both sides of the relationship.
- This is also called a **self-referencing relationship**.

Purpose

- 1. **Model self-referential structures** where an object relates to other objects of the same type.
- 2. **Represent hierarchies** within a single entity (e.g., manager-employee, parent-child).
- 3. **Avoid duplicate tables** for the same type of data.

Example – Employee Management

Entity: Employee (EmployeeID, Name, ManagerID)

• ManagerID is a **foreign key** referencing EmployeeID in the **same table**.

Table Example:

EmployeeID	Name	ManagerID
1	Alice	NULL
2	Bob	1
3	Carol	1
4	David	2

Meaning:

- Alice has no manager (top-level).
- Bob and Carol report to Alice.
- David reports to Bob.

Real-World Uses

- Organizational hierarchies (CEO → Managers → Employees).
- **Folder structures** (Folder contains subfolders).
- **Bill of Materials (BOM)** (product contains sub-components).
- Linked lists stored in databases.

Types of Recursive Relationships

- 1. **One-to-One (1:1):** An entity relates to exactly one other of the same type.
 - Example: One employee mentors exactly one other employee.
- 2. **One-to-Many** (1:N): One record relates to many others of the same type (most common).
 - Example: One manager manages many employees.
- 3. **Many-to-Many (M:N):** Many records relate to many others of the same type.
 - Example: Authors collaborating on multiple books with each other.

Advantages

- Reduces redundancy (only one table for the entity).
- Easy to expand the hierarchy to multiple levels.
- Supports complex, multi-level relationships naturally.

Limitations

- Queries can become complex (especially retrieving multiple hierarchy levels).
- Performance can drop for deep hierarchies without proper indexing.

4. Historical Data

- Historical data refers to the storage of past states of data for reference, analysis, compliance, or tracking changes over time.
- Instead of overwriting old values, new records are created with validity periods (start and end dates).
- Often used in **time-based queries** to retrieve data "as it was" at a certain point in the past.

Purpose

- 1. **Track changes over time** maintain a history of modifications.
- 2. **Enable time-travel queries** see data as it existed on a specific date.
- 3. **Compliance & auditing** fulfill legal or business record-keeping requirements.

4. **Trend analysis** – identify patterns by comparing past and current data.

Example – Employee Salary History

EmployeeID Salary StartDate EndDate

101 40000 01-01-2020 31-12-2021 101 45000 01-01-2022 NULL

Meaning:

- Employee 101 earned ₹40,000 between Jan 2020 and Dec 2021.
- From Jan 2022 onward, salary increased to ₹45,000 (EndDate = NULL means current record).

Real-World Uses

- **Data warehouses** store years of transactional history for analysis.
- **Financial records** maintain past account balances and transactions.
- **Healthcare systems** track patient medical history.
- **Retail sales** monitor price changes over time.

Types of Historical Data Storage

- 1. **Slowly Changing Dimensions (SCD)** (in data warehousing):
 - **Type 1:** Overwrite old data (no history).
 - Type 2: Keep history with start and end dates.
 - Type 3: Store limited history in extra columns.
- 2. **Audit Tables:** Separate tables just for historical logs.
- 3. **Temporal Tables:** Database feature to auto-maintain history (available in modern DBMS).

Advantages

- Enables trend analysis and forecasting.
- Essential for compliance and audits.
- Improves business intelligence decisionmaking.

Limitations

- Increases storage requirements.
- Requires careful indexing for performance.
- Can make queries more complex.