Statistical Concepts: A Comprehensive Overview

Types of Variables

Discrete Random Variable

A **discrete random variable** takes a **countable** number of distinct values.

Characteristics:

✅ Takes only specific, separate values (e.g., integers).

✅ Usually counted, not measured.

✅ Has a **Probability Mass Function (PMF)**, which gives probabilities for exact values.

Examples:

- 🎲 **Rolling a die** → Possible outcomes: {1, 2, 3, 4, 5, 6}.
- 👶 **Number of children in a family** → Possible outcomes: {0, 1, 2, 3, ...}.
- 📧 **Number of emails received per day** → Countable values: {0, 1, 2, ...}.

The sum of probabilities must be **1**:

Continuous Random Variable

A **continuous random variable** takes an **infinite** number of values within a given range.

Characteristics:

✅ Takes values from a **continuous range**.

✅ Usually measured, not counted.

✅ Has a **Probability Density Function (PDF)** instead of a PMF.

✅ Probability of a single value is **zero** → We calculate probability over an **interval** using integration.

Examples:

- 🌡️ **Temperature in a city** (e.g., 22.5°C, 22.51°C, 22.512°C... infinite values).
- 🕒 **Time taken to finish a race** (e.g., 9.58s, 9.581s, 9.5812s... infinite precision).
- 📏 **Height of students** (e.g., 160.1 cm, 160.15 cm, ...).

Since there are infinite possible values, the probability of any single exact height (e.g.,
$P(X=165)$ is zero.)

Key Differences Between Discrete and Continuous Variables

Feature	Discrete Random Variable	Continuous Random Variable
Possible Values	Countable (finite or infinite)	Infinite within a range
Example	Number of students in a class (1, 2, 3...)	Temperature (22.1°C, 22.15°C...)
Probability	Uses PMF	Uses PDF
Probability of a single value	Can be nonzero	Always zero
Calculation of probability	Summation of probabilities	Integration of PDF

Probability functions

Probability Mass Function (PMF)

A **PMF** assigns probabilities to **specific discrete values**. It must satisfy:

$P(X = x) \geq 0$ for all values of x.

The total probability must sum to 1:

$$\sum P(X = x) = 1$$

Probability Density Function (PDF)

A **PDF** is used for **continuous random variables.** Instead of assigning probabilities to specific values, it represents a **smooth probability curve**.

Since continuous values are **uncountable**, the probability of one exact value is **always 0**. Instead, we calculate probabilities **over an interval** using integration:

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

Key Differences Between PMF and PDF

Feature	PMF (Discrete)	PDF (Continuous)
Variable type	Discrete (countable values)	Continuous (infinite values)
Probability of a single value	P(X = x) can be nonzero	P(X = x) = 0, must use integration
Representation	Bar graph	Smooth curve
Example	Rolling a die, Number of students in a class	Height of people, Temperature

Summary

- **PMF** is used for **discrete** random variables (e.g., die rolls 🎲).
- **PDF** is used for **continuous** random variables (e.g., heights 📏).
- **PMF assigns probability to exact values, PDF uses integration over a range**.

Degrees of Freedom (DOF)

Degrees of Freedom (DOF) refers to the number of values in a dataset that are **free to vary** while estimating a statistical parameter (like mean, variance, regression coefficients, etc.).

Think of it as **the number of independent choices you have before restrictions come into play.**

🔷 **Example:**

You have 5 exam scores with an average of **80**. If the first four scores are **75, 85, 90, and 70**, the last score is already **fixed**:

X_5= 80 × 5 − (75+85+90+70) ⇒ 80

So, **only 4 numbers are free to vary** → **Degrees of Freedom = 4**.

Degrees of Freedom in Statistics

A. Degrees of Freedom in Sample Variance

When calculating **sample variance**, we estimate the mean first. Since the mean is **fixed**, only $n-1$ values are free to vary.

$$s^2 = \frac{\sum(X_i - \bar{X})^2}{n-1}$$

So, **DOF =** $n-1$ for variance calculation.

B. Degrees of Freedom in Regression

In **linear regression**, we estimate coefficients $(\beta_0, \beta_1, ...)$, which reduces the number of independent data points.

For a model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

The **Degrees of Freedom for residuals** is:

$$DOF = n - (k+1)$$

where:

- n = number of data points
- k = number of independent variables
- $+1$ accounts for the intercept β_0

If you have **100 data points** and **3 predictors**, the residual DOF is:

$$DOF = 100 - (3+1) = 96$$

Why Are Degrees of Freedom Important?

✔️ **Used in statistical tests** (t-test, chi-square, F-test).

✔️ **Affects confidence intervals** and hypothesis testing.

✔️ **Determines model flexibility** in regression.

Empirical Rule (68-95-99.7 Rule)

The **Empirical Rule** is a guideline that applies to **normal (bell-shaped) distributions**. It tells us how much data falls within **one, two, and three standard deviations** of the mean.

The Rule in Numbers

For a normal distribution:

- **68%** of the data falls within **1 standard deviation** (σ) of the mean (μ).
- **95%** of the data falls within **2 standard deviations** (σ) of the mean (μ).
- **99.7%** of the data falls within **3 standard deviations** (σ) of the mean (μ).

Example: Heights of People 📏

Let's say human heights follow a **normal distribution** with:

- Mean **μ = 170 cm**
- Standard deviation **σ = 10 cm**

Applying the empirical rule:

- **68%** of people have heights between **160 cm and 180 cm**.
- **95%** of people have heights between **150 cm and 190 cm**.
- **99.7%** of people have heights between **140 cm and 200 cm**.

Correlation vs. Covariance

Both **correlation** and **covariance** measure relationships between two variables, but they differ in **scale and interpretation**.

Concept	Definition	Range	Interpretation
Covariance	Measures **how two variables move together**.	$-\infty$ to $+\infty$	**Positive**: Both increase together. **Negative**: One increases, the other decreases.
Correlation	Measures the **strength and direction** of the relationship (standardized).	-1 to $+1$	**+1**: Perfect positive relation. **-1**: Perfect negative relation. **0**: No relation.

Covariance – Measures the Relationship's Direction

Covariance tells us **if two variables move together** but does **not** tell us **how strong** the relationship is.

$$Cov(X,Y) = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

🔷 **Example:**

- If height and weight **both increase together**, covariance is **positive**.
- If study time increases and exam mistakes **decrease**, covariance is **negative**.
- The value of covariance depends on the units of measurement, so it's **not standardized**.

Correlation – Standardized Measure of Relationship

Correlation is **scaled** between −1 and +1, making it easier to interpret.

$$Correlation(r) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

where σ_X and σ_Y are standard deviations.

✅ **Always between -1 and 1:**

- **r = +1** → Perfect positive relationship (e.g., height vs. weight).
- **r = -1** → Perfect negative relationship (e.g., study time vs. mistakes).
- **r = 0** → No relationship (e.g., height vs. favorite color).

🔷 Example:
Even if we measure height in cm or inches, the correlation stays the same because it's unitless. But covariance would change.

Summary

🔷 **Covariance** shows **direction** (positive or negative).

🔷 **Correlation** shows **strength** and **direction** (always between −1 and +1).

🔷 Correlation is more useful because it's **standardized**.
