

Generalized linear model

A **Generalized Linear Model (GLM)** is a flexible extension of ordinary linear regression that lets you model **non-normal response variables** (like binary outcomes or counts) while still keeping a linear structure at heart.

Think of it as:

Linear regression, but upgraded to handle more types of data.

Why we need GLMs

Ordinary linear regression assumes:

- The response is **continuous**
- Errors are **normally distributed**
- Predictions can be any real number

But many real problems don't fit that:

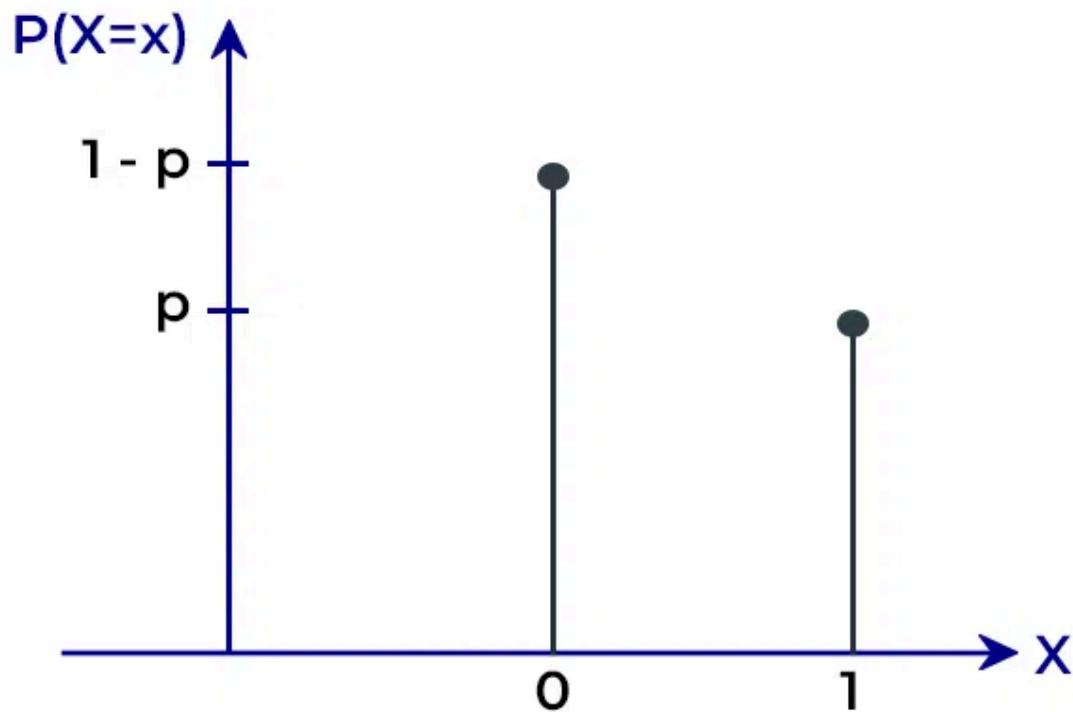
- Yes / No → **binary**
- Number of events → **counts**
- Proportions → **between 0 and 1**
--> So GLMs solve this.

The 3 key components of a GLM

1. Random component (Distribution)

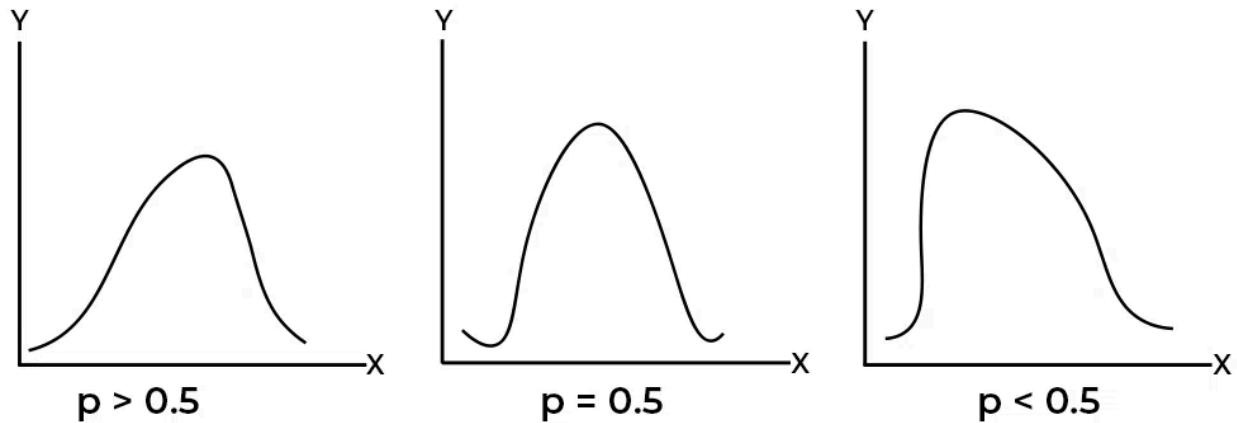
The response variable (Y) follows a distribution from the **exponential family**, such as:

- Normal
- Bernoulli



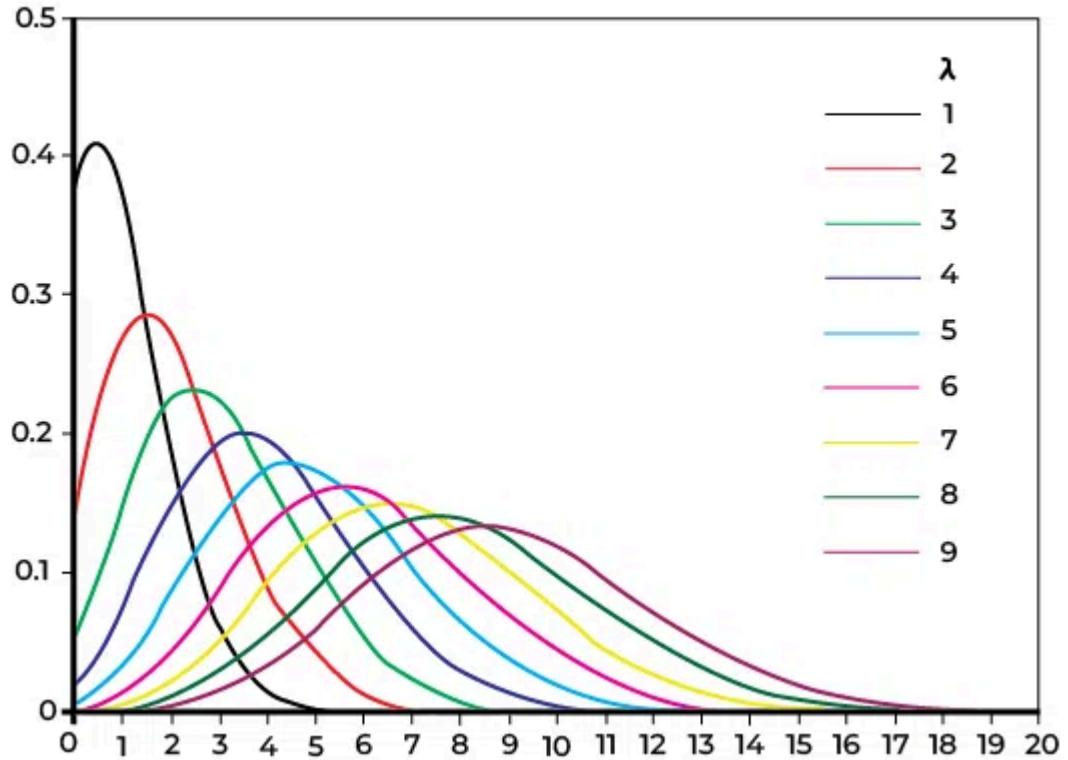
- Binomial

Shape of Binomial Distribution

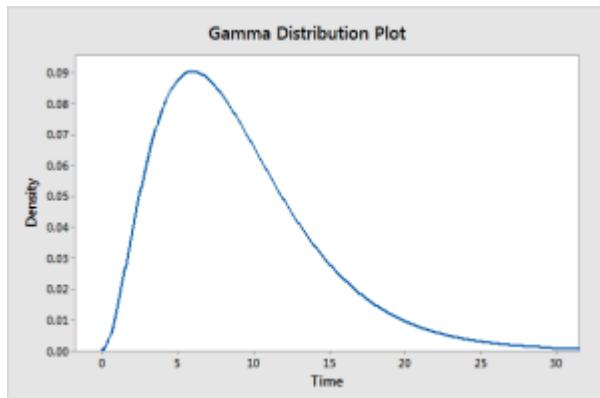


- Poisson

Poisson Distribution



- Gamma



2. Systematic component (Linear predictor)

Just like linear regression:

$$\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

3. Link function

Connects the **mean of (\mathbf{Y})** to the linear predictor:

$$g(\mathbb{E}[Y]) = \eta$$

Common GLMs (very important)

MODEL	RESPONSE TYPE	DISTRIBUTION	LINK FUNCTION
Linear regression	Continuous	Normal	Identity
Logistic regression	Binary (0/1)	Bernoulli	Logit
Poisson regression	Counts	Poisson	Log
Gamma regression	Positive continuous	Gamma	Log

👉 logistic regression is a GLM.

Example (intuition)

Logistic regression

- Outcome: pass / fail
- Mean must be between 0 and 1
- Model:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$

The logit link keeps probabilities valid.

Use a GLM when:

- The response is **not normally distributed**
- You need predictions with constraints (0–1, positive, integers)

- Linear regression assumptions are violated