Understanding Distributions in Statistics

1. What is a Distribution?

- A distribution in statistics describes how the values of a variable are spread out or distributed.
- It tells us what values are possible and how often (or how likely) each value occurs.
- Distributions help us understand the behavior of data, perform statistical analysis, and make predictions.

2. Two Definitions or Views of Distribution:

a) Mathematical View (Theoretical Distribution):

- A distribution can be defined as a mathematical function (like a PDF or PMF).
- It maps each possible value (x) to a probability (discrete) or density (continuous).
- Examples include Normal distribution, Binomial distribution, Poisson distribution, etc.

b) Data-Based View (Empirical Distribution):

- Based on actual observed data.
- Constructed by calculating the frequency or relative frequency of values in the dataset.
- Represented using a histogram, bar chart, or empirical cumulative distribution function (ECDF).

3. Why Do We Fit a Theoretical Distribution to Data?

- Real-world data is just a sample. It may contain noise, variability, and imperfections.
- Fitting a known distribution (like Normal) helps generalize our understanding to the entire population.
- Theoretical distributions allow us to compute probabilities, make predictions, run simulations, and apply statistical tests.
- Example: If data looks bell-shaped, we may fit a Normal distribution using its formula with estimated mean and standard deviation.

4. Understanding Histogram vs PDF (Probability Density Function):

Histogram:

- Built from data by dividing the range into bins and counting values in each bin.
- Shows relative frequency (proportion of data in each bin).
- Gives approximate probability by: relative frequency = count in bin / total count.

PDF (for continuous distributions):

- A smooth curve that defines the density of data.
- PDF value at a point is not a probability. To get probability, we must integrate the PDF over a range.
- For example, $P(55 < X < 65)$ = Area under PDF curve from 55 to 65.

5. Key Difference Between Relative Frequency and Density:

- Relative Frequency: Computed directly from data. It is the proportion of observations in a bin.
- Density: Found in a histogram when using 'density=True'. It is relative frequency divided by bin width.
- To get probability from density: Probability approximately Density × Bin Width
- In a fitted PDF, probability = Area under the curve = Integral of f(x) from a to b.

6. Summary Table:

Feature	Empirical Distribution	Theoretical Distribution
Source	Observed data	Mathematical model (PDF/PMF)
Y-axis shows	Relative frequency or count	Density (continuous) or probability (discrete)
Use of probability	Count or proportion	Integrate (continuous) or lookup (discrete)
Nature	Data-specific, finite	Generalized, population-level

7. Clarifying: Relative Frequency vs Density

- Relative Frequency: Proportion of data in a bin. It is an estimate of probability for that range.
- Density: Not a probability. It shows how concentrated the data is. Found in histograms when using density=True.

The Relationship:

Density = Relative Frequency / Bin Width

-> Relative Frequency = Density * Bin Width

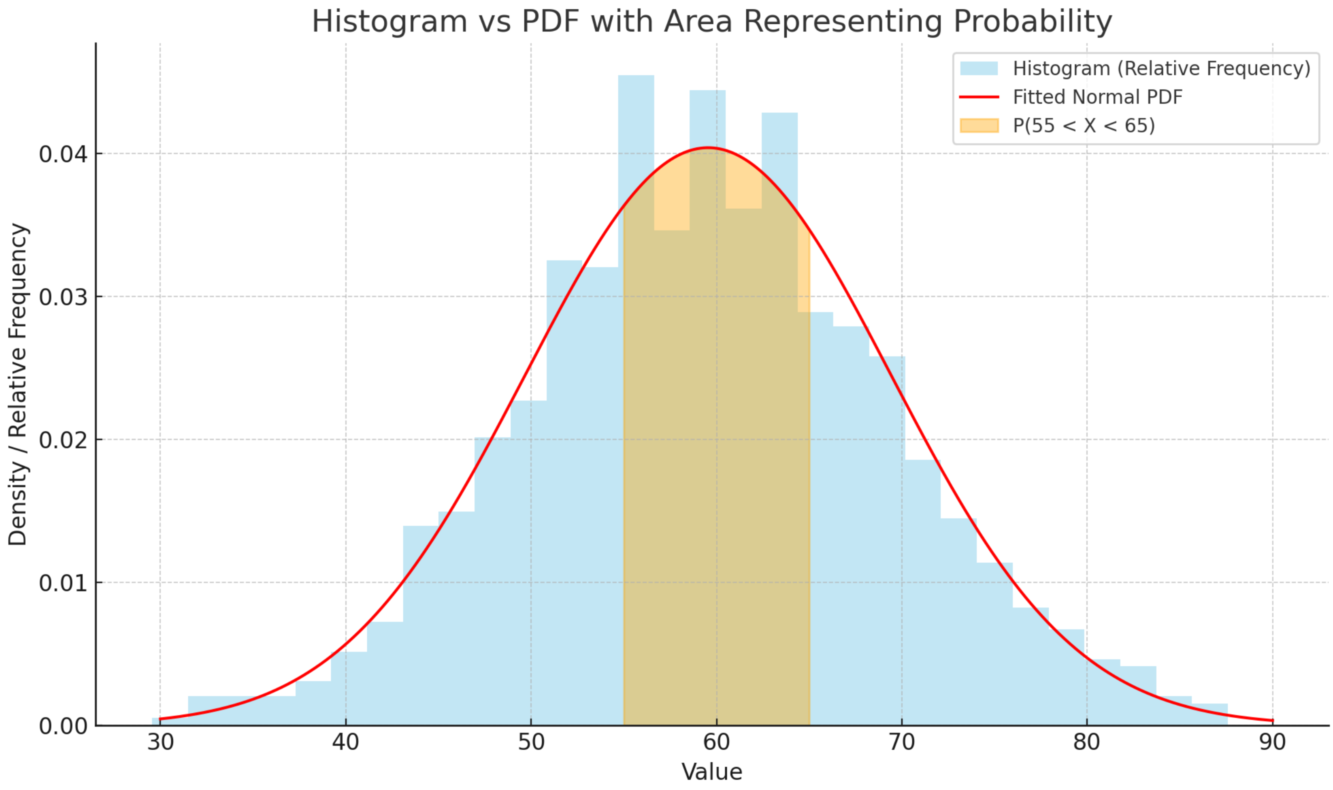
- In a regular histogram (density=False), bar height = relative frequency.
- In a density histogram (density=True), bar height = density. You need to multiply it by bin width to estimate probability.

Final Summary:

- Relative frequency is a direct estimate of probability from data.
- Density (PDF) is the rate of accumulation of probability per unit on the X-axis.
- To get true probability from PDF, integrate over a range.

8. Visual Explanation:

- The following chart shows how a histogram (empirical distribution) compares to a fitted Normal distribution (PDF).
- The shaded orange area between 55 and 65 represents the probability under the fitted curve for that range.


Histogram vs PDF with Area Representing Probability
Histogram (Relative Frequency)
Fitted Normal PDF
P(55 < X < 65)
Density / Relative Frequency
0.00
0.01
0.02
0.03
0.04
30
40
50
60
70
80
90
Value