Introduction to Probability and Statistics

Gaussian Distribution and Central Limit Theorem

Manuele Leonelli

School of Science and Technology, IE University

Why Probability and Statistics?

- **Modeling Real-Life Problems:** Many real-world problems involve uncertainty, variability, and randomness.
- **Random Variables:** In many cases, the outcomes or measurements we deal with are random variables.
- **Importance in Machine Learning:** Machine learning models often rely on probabilistic frameworks to make predictions and generalize from data.

What is a Random Variable?

- A **random variable** is a variable whose possible values are numerical outcomes of a random phenomenon.
- Random variables are used to quantify uncertain outcomes, such as the result of a dice roll, the height of a randomly chosen person, or the error in a machine learning model.
- Random variables can be **discrete** (e.g., number of heads in coin tosses) or **continuous** (e.g., temperature in a city on a given day).

Why Assume Normality?

- In many situations, we assume that random variables follow a **Normal (Gaussian) distribution**.
- This assumption is often justified by the **Central Limit Theorem** (CLT), which states that the sum (or average) of a large number of independent, identically distributed random variables tends toward a normal distribution, regardless of the original distribution.
- The normal distribution is easy to work with and has desirable properties that simplify analysis and inference.

Gaussian (Normal) Distribution

The **Gaussian distribution**, also known as the **Normal distribution**, is one of the most important probability distributions in statistics.

- It is symmetric and describes many natural phenomena.
- The probability density function (PDF) of the normal distribution is given by:

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

  where:

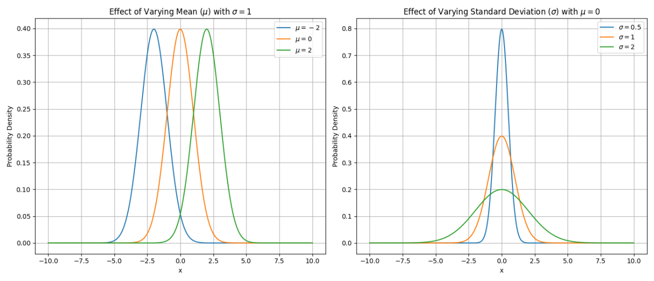
  - μ is the mean (center) of the distribution.
  - σ^2 is the variance, a measure of the spread.

Properties of the Gaussian Distribution

The Gaussian distribution has several key properties:

- **Symmetry:** The distribution is symmetric around the mean μ.
- **68-95-99.7 Rule:**
  - 68% of the data falls within 1 standard deviation (σ) of the mean.
  - 95% within 2 standard deviations.
  - 99.7% within 3 standard deviations.
- **Central Role in Statistics:** Due to the Central Limit Theorem, many statistics are normally distributed, even if the underlying data is not.

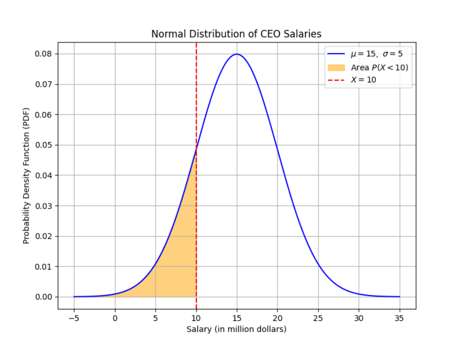
Visualizing the Gaussian Distribution



- The curve is bell-shaped and symmetric around the mean μ.
- The spread of the curve is determined by the standard deviation σ.

Practical Example: CEO Salaries in S&P 500 Companies

- Suppose the salaries of CEOs in S&P 500 companies are assumed to follow a **Normal distribution** with:
  - Mean salary $\mu = \$15$ million
  - Standard deviation $\sigma = \$5$ million
- Question: What is the probability that a randomly selected CEO earns less than $10 million?



Solving the Problem: Probability and the Normal Distribution

- We want to find $P(X < 10)$ where $X \sim \mathcal{N}(15, 5^2)$.
- Mathematically, this can be calculated as:

$$P(X < 10) = \int_{-\infty}^{10} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = 0.1587$$

- Perhaps surprisingly, this integral cannot be solved exactly, but only approximately!!

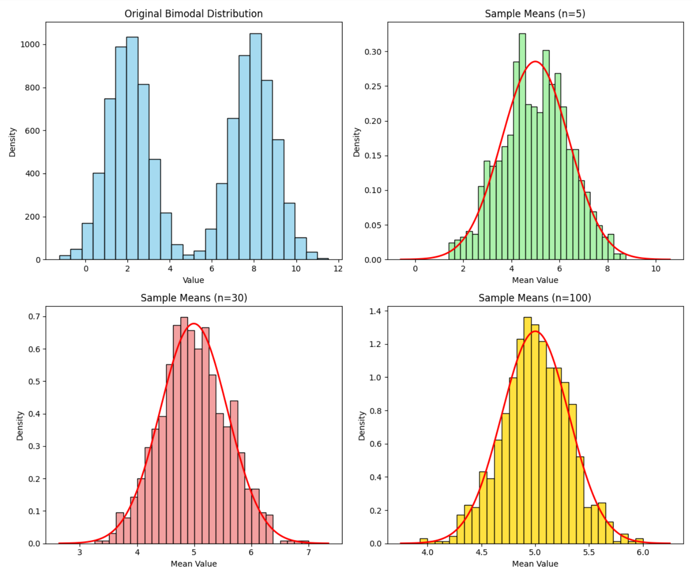
The Central Limit Theorem (CLT)

The **Central Limit Theorem** is a fundamental theorem in probability and statistics.

- The CLT states that the sum (or average) of a large number of independent, identically distributed (i.i.d.) random variables will be approximately normally distributed, regardless of the original distribution.
- Mathematically, if $X_1, X_2, \ldots, X_n$ are i.i.d. random variables with mean μ and variance σ^2, then as n becomes large:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

An Example



Original Bimodal Distribution
Density
Value
0
200
400
600
800
1000
0
2
4
6
8
10
12
Sample Means (n=5)
Density
Mean Value
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0
2
4
6
8
10
Sample Means (n=30)
Density
Mean Value
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
3
4
5
6
7
Sample Means (n=100)
Density
Mean Value
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
4.0
4.5
5.0
5.5
6.0

Importance of the Central Limit Theorem

The Central Limit Theorem is crucial for several reasons:

- **Enables Inference:** Allows us to make inferences about population parameters based on sample statistics.
- **Justifies the Use of Normal Models:** In many practical situations, we can assume normality, even if the underlying data is not normally distributed.
- **A/B Testing:** In hypothesis testing, especially in A/B testing, the CLT allows us to assume that the sampling distribution of the test statistic is approximately normal.