Chapter 1: An Introduction to Probability and Statistics

Prof. Alok Shukla
with help from Gemini and Chat GPT

Department of Physics
IIT Bombay, Powai, Mumbai 400076

Course Name: AI and Data Science (PH 227):
Part 1 (Data Science)

1. **Introduction to Probability**
2. **Axioms of Probability**
3. **Conditional Probability and Bayes' Theorem**
4. **Introduction to Random Variables**
5. **Discrete Random Variables**
6. **Continuous Random Variables**
7. **Common Distributions**
8. **Derivations of the Normal Distribution**
   - Central Limit Theorem
   - Maximum Entropy Derivation
   - Setting the Constraints
   - Using Lagrange Multipliers
   - Solving for the PDF
   - Finding the Constants
9. **Measures of Central Tendencies**
10. **Measures of Variability**
11. **Measures of Distribution**
12. **Moments**

What is Probability?

- Probability is the measure of the likelihood that an event will occur.
- It quantifies uncertainty.
- Expressed as a number between 0 and 1, inclusive.
  - 0: Event is impossible.
  - 1: Event is certain.

Key Concepts: Sample Space

- **Sample Space (Ω or S)**: The set of all possible outcomes of a random experiment.
- **Examples**:
  - Flipping a coin: $\Omega = \{\text{Heads, Tails}\}$
  - Rolling a die: $\Omega = \{1, 2, 3, 4, 5, 6\}$
  - Tossing two coins: $\Omega = \{\text{HH, HT, TH, TT}\}$

Key Concepts: Events

- **Event (E)**: Any subset of the sample space. It's a collection of outcomes.
- **Examples (from rolling a die)**:
  - Event A: Rolling an even number $\implies A = \{2, 4, 6\}$
  - Event B: Rolling a number greater than 4 $\implies B = \{5, 6\}$
  - Event C: Rolling a 7 $\implies C = \emptyset$ (impossible event)

Kolmogorov's Axioms

For any event E in a sample space Ω:

1. **Non-negativity**: The probability of an event is a non-negative real number.
$$P(E) \geq 0$$
2. **Normalization**: The probability of the entire sample space (a certain event) is 1.
$$P(\Omega) = 1$$
3. **Additivity (for mutually exclusive events)**: For any sequence of disjoint (mutually exclusive) events $E_1, E_2, E_3, \ldots$
$$P(E_1 \cup E_2 \cup E_3 \cup \ldots) = P(E_1) + P(E_2) + P(E_3) + \ldots$$
(i.e., if $E_i \cap E_j = \emptyset$ for $i \neq j$)

Conditional Probability

- The probability of an event A occurring, given that another event B has already occurred.
- Denoted as $P(A|B)$.
- Formula:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad \text{provided } P(B) > 0$$

- $P(A \cap B)$ is the probability of both A and B occurring.

Bayes' Theorem

- Relates conditional probabilities of two events.
- Useful for updating beliefs based on new evidence.
- Formula:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- Can be expanded using the law of total probability for $P(B)$:

$$P(B) = P(B|A)P(A) + P(B|A^c)P(A^c)$$

Where A^c is the complement of event A.

What is a Random Variable?

- A **random variable** is a function that maps outcomes from a random experiment (sample space) to real numbers.
- It is denoted by a capital letter, e.g., X, Y, Z.
- **Example**: Tossing two coins.
  - Sample Space $\Omega = \{\text{HH, HT, TH, TT}\}$
  - Let X be the number of heads.
  - $X(\text{HH}) = 2$
  - $X(\text{HT}) = 1$
  - $X(\text{TH}) = 1$
  - $X(\text{TT}) = 0$

Types of Random Variables

- **Discrete Random Variable**:
  - Takes on a finite or countably infinite number of values.
  - Examples: Number of heads in coin tosses, number of cars passing a point.
- **Continuous Random Variable**:
  - Takes on any value within a given range (uncountably infinite).
  - Examples: Height, weight, temperature, time.

Probability Mass Function (PMF)

- For a discrete random variable X, the **Probability Mass Function (PMF)**, denoted by $P_X(x)$ or $f_X(x)$, gives the probability that X takes on a specific value x.

$$P_X(x) = P(X = x)$$

- **Properties**:
  - $0 \leq P_X(x) \leq 1$ for all x.
  - $\sum_x P_X(x) = 1$ (sum over all possible values of x).

Expectation (Mean) of a Discrete RV

- The **Expected Value** or **Mean** of a discrete random variable X, denoted $E[X]$ or μ_X, is the weighted average of all possible values, where the weights are their probabilities.

$$E[X] = \sum_x x \cdot P_X(x)$$

- Represents the long-run average value of the random variable.

Variance of a Discrete RV

- The **Variance** of a discrete random variable X, denoted $Var(X)$ or σ_X^2, measures the spread or dispersion of its values around the mean.

$$Var(X) = E[(X - E[X])^2] = \sum_x (x - \mu_X)^2 P_X(x)$$

- An alternative formula:

$$Var(X) = E[X^2] - (E[X])^2$$

- The **Standard Deviation** is $\sigma_X = \sqrt{Var(X)}$.

Probability Density Function (PDF)

- For a continuous random variable X, the **Probability Density Function (PDF)**, denoted by $f_X(x)$, describes the relative likelihood for the random variable to take on a given value.
- **Note**: $P(X = x) = 0$ for any single value x. We talk about probabilities over intervals.
- **Properties**:
  - $f_X(x) \geq 0$ for all x.
  - $\int_{-\infty}^{\infty} f_X(x)dx = 1$.
- Probability of X being in an interval $[a, b]$:

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

Cumulative Distribution Function (CDF)

- The **Cumulative Distribution Function (CDF)**, denoted by $F_X(x)$, gives the probability that the random variable X takes on a value less than or equal to x.

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)dt$$

- **Properties**:
  - $0 \leq F_X(x) \leq 1$.
  - $F_X(x)$ is non-decreasing.
  - $\lim_{x \to -\infty} F_X(x) = 0$
  - $\lim_{x \to \infty} F_X(x) = 1$

Expectation (Mean) of a Continuous RV

- The **Expected Value** or **Mean** of a continuous random variable X:

$$E[X] = \int_{-\infty}^{\infty} x \cdot f_X(x) dx$$

Variance of a Continuous RV

- The **Variance** of a continuous random variable X:

$$Var(X) = E[(X - E[X])^2] = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x) dx$$

- Alternative formula:

$$Var(X) = E[X^2] - (E[X])^2 = \int_{-\infty}^{\infty} x^2 f_X(x) dx - (E[X])^2$$

Bernoulli Distribution

- Models a single trial with two possible outcomes: success (1) or failure (0).
- Parameter: p (probability of success).
- PMF:

$$P(X = x) = p^x(1-p)^{1-x} \quad \text{for } x \in \{0, 1\}$$

- $E[X] = p$
- $Var(X) = p(1-p)$

Binomial Distribution

- Models the number of successes in a fixed number (n) of independent Bernoulli trials.
- Parameters: n (number of trials), p (probability of success in each trial).
- PMF: is the probability of getting success in k out of n trials

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for } k \in \{0, 1, \ldots, n\}$$

- $E[X] = np$
- $Var(X) = np(1-p)$

Normal (Gaussian) Distribution

- One of the most important continuous distributions.
- Bell-shaped, symmetric curve.
- Parameters: μ (mean), σ^2 (variance).
- PDF:

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

- $E[X] = \mu$
- $Var(X) = \sigma^2$

Plots of a Few Normal Distributions

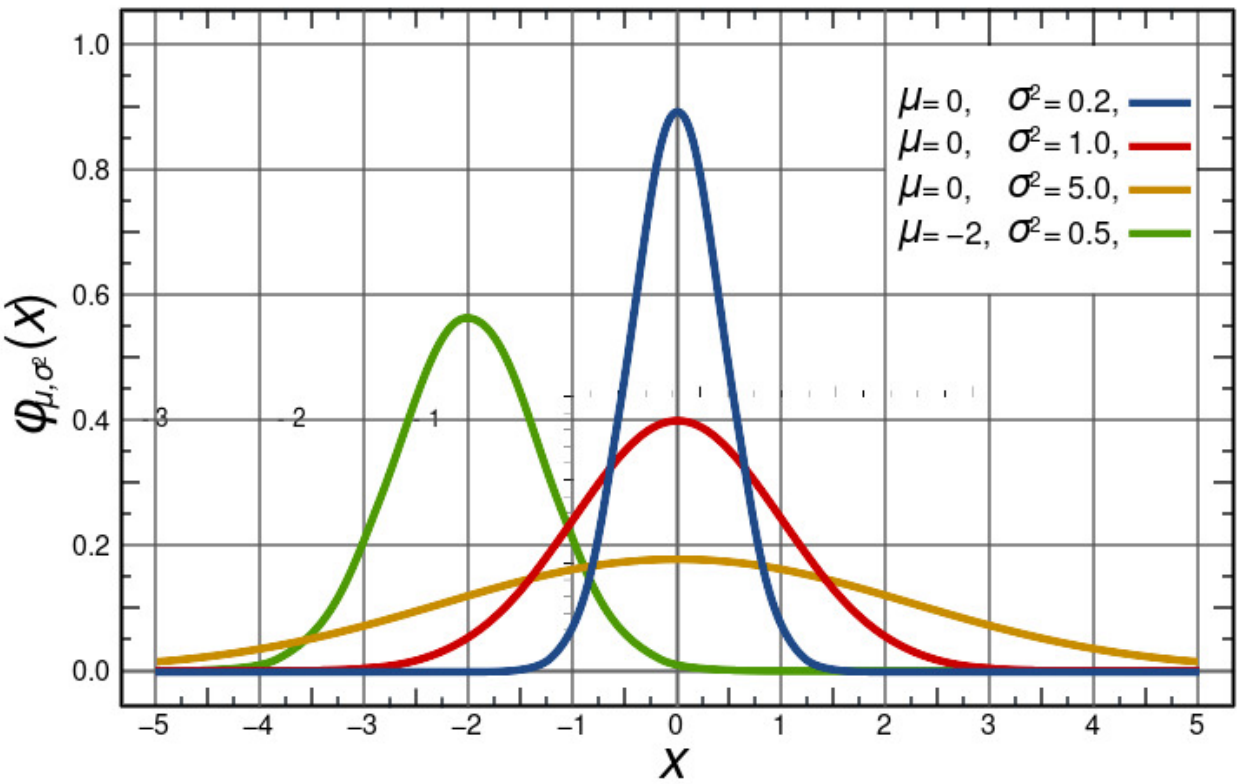


Figure: Plots of normal distributions for a few values of mean and variance (courtesy Wikipedia)

Calculation of Mean and Variance of the Normal Distribution

To calculate these quantities, the following Gaussian and related integrals are useful

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi} \tag{1}$$

Using this integral, the following useful integrals can be derived

$$\int_{-\infty}^{\infty} e^{-a(x+b)^2} dx = \sqrt{\frac{\pi}{a}}$$

$$\int_{-\infty}^{\infty} e^{-(ax^2+bx+c)} dx = \sqrt{\frac{\pi}{a}} e^{\frac{b^2}{4a}-c}$$

$$\int_{-\infty}^{\infty} x^{2n} e^{-x^2/a^2} dx = \sqrt{\pi} \frac{a^{2n+1}(2n-1)!!}{2^n}$$

$$\int_{-\infty}^{\infty} x^{2n+1} e^{-x^2/a^2} dx = 0$$

Calculation of the basic Gaussian integral

Let the basis Gaussian integral be $I = \int_{-\infty}^{\infty} e^{-x^2} dx$. Clearly

$$I^2 = \left(\int_{-\infty}^{\infty} e^{-x^2} dx \right) \left(\int_{-\infty}^{\infty} e^{-y^2} dy \right) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(x^2+y^2)} dxdy.$$

This double integral is over the entire xy plane. Let us change the variables from the Cartesian coordinates to the plane polar coordinates (r, θ) by making the substitution $x = r\cos\theta$ and $y = r\sin\theta$, so that $dxdy = rdrd\theta$, with $0 \leq r \leq \infty$ and $0 \leq \theta \leq 2\pi$. With this

$$I^2 = \left(\int_0^{\infty} drre^{-r^2} \right) \left(\int_0^{2\pi} d\theta \right).$$

The radial integral can be computed easily by making the substitution $r^2 = t$, while the θ integral is trivial, leading to

$$I^2 = \pi \implies I = \sqrt{\pi}$$

Checking the normalization of the normal distribution

Any normalized PDF must satisfy

$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

Let us verify it for the normal distribution $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ by performing this integral

$$I = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx.$$

We make the substitution $x - \mu = \sigma\sqrt{2}t \implies dx = \sigma\sqrt{2}dt$, leading to

$$I = \frac{\sigma\sqrt{2}}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-t^2} dx = \frac{\sigma\sqrt{2\pi}}{\sigma\sqrt{2\pi}} = 1$$

Calculation of Mean

The mean can be calculated as

$$\bar{x} = \int_{-\infty}^{\infty} x f(x) dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

here we make a change of variables $x - \mu = \sigma\sqrt{2}t \implies x = t\sigma\sqrt{2} + \mu$, and $dx = \sigma\sqrt{2}dt$, leading to

$$\bar{x} = \frac{\mu\sigma\sqrt{2}}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-t^2} dt + \frac{\sigma\sqrt{2}}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} t e^{-t^2} dt$$

$$\bar{x} = \frac{\mu\sigma\sqrt{2}}{\sigma\sqrt{2\pi}} \left(\sqrt{\pi}\right) + 0 = \mu$$

$$\boxed{\bar{x} = \mu}$$

Calculation of Variance

We calculate the variance from its definition

$$\mathrm{Var}(X) = \langle x^2 \rangle - \langle x \rangle^2,$$

where the symbol $\langle \cdots \rangle$ implies the mean of the quantity. We need to calculate only $\langle x^2 \rangle$

$$\langle x^2 \rangle = \int_{-\infty}^{\infty} x^2 f(x) dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

again making the substitution $x - \mu = \sigma\sqrt{2}t$, the integral becomes

$$\begin{aligned} \langle x^2 \rangle &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (\sigma\sqrt{2}t + \mu)^2 e^{-\frac{1}{2}\left(\frac{t}{\sigma}\right)^2} dx \\ &= \frac{\sigma\sqrt{2}}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \left(2\sigma^2 t^2 + 2\sqrt{2}\mu\sigma t + \mu^2\right) e^{-t^2} dt \end{aligned}$$

Calculation of Variance...

Using the values of the integrals given above, we obtain

$$\begin{aligned}\langle x^2 \rangle &= \frac{1}{\sqrt{\pi}} \left(\frac{2\sigma^2 \sqrt{\pi}}{2} + 0 + \mu^2 \sqrt{\pi} \right) \\ &= \sigma^2 + \mu^2\end{aligned}$$

As a result

$$\text{Var}(x) = \langle x^2 \rangle - \langle x \rangle^2 = \sigma^2 + \mu^2 - \mu^2 = \sigma^2$$

and the standard deviation (SD)

$$\text{SD} = \sqrt{\text{Var}} = \sigma$$

Thus σ is nothing but the standard deviation of the normal distribution

Derivations of the Normal Distribution

The justification/derivation of the normal distribution is normally based on two approaches:

- Central Limit Theorem
- Maximum-Entropy Approach

We briefly discuss both next.

What is the Central Limit Theorem?

The Big Idea

The Central Limit Theorem (CLT) states that the distribution of sample means approximates a normal distribution as the sample size becomes larger, regardless of the shape of the original population distribution.

- This is one of the most powerful and important concepts in all of statistics.
- It provides the theoretical foundation for many inferential statistical procedures.

Key Conditions

For the Central Limit Theorem to apply, a few conditions must be met:

1. **Random Sampling:** The samples must be independent and identically distributed (i.i.d.).
2. **Sufficiently Large Sample Size:** The sample size, denoted by n, should be large enough. A common rule of thumb is $n \geq 30$.

he larger the sample size, the more closely the sampling distribution of the mean will resemble a normal distribution.

The Sampling Distribution of the Mean

When the CLT applies, the sampling distribution of the mean, $\bar{x}$, will have the following properties:

- **Mean:** The mean of the sample means ($\mu_{\bar{x}}$) will be equal to the population mean (μ).

$$\mu_{\bar{x}} = \mu$$

- **Standard Deviation:** The standard deviation of the sample means, known as the **Standard Error** ($\sigma_{\bar{x}}$), will be equal to the population standard deviation (σ) divided by the square root of the sample size (n).

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Why is it so important?

- It allows us to apply the properties of the normal distribution to analyze and make inferences about population parameters, even when the original population data is not normally distributed.
- This is fundamental for hypothesis testing and creating confidence intervals.
- It explains why a simple random sample can be used to represent a much larger population.

The Principle of Maximum Entropy

- The principle states that the probability distribution that best represents the current state of knowledge is the one with the largest entropy, subject to known constraints.
- Entropy is a measure of uncertainty or randomness. A higher entropy means less information.
- We seek the distribution $f(x)$ that maximizes entropy:

$$H(f) = -\int_{-\infty}^{\infty} f(x) \ln f(x)\, dx$$

- Subject to the constraints we have on the system.

Constraints for a Normal Distribution

We assume we know the following about the distribution of a random variable X:

1. **Normalization:** The total probability must be equal to one.

$$\int_{-\infty}^{\infty} f(x)\, dx = 1$$

2. **Fixed Mean:** The expected value (mean) is a known constant, μ.

$$\int_{-\infty}^{\infty} xf(x)\, dx = \mu$$

3. **Fixed Variance:** The expected value of the squared deviation from the mean is a known constant, σ^2.

$$\int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx = \sigma^2$$

Applying Lagrange Multipliers

We can set up a function $\mathcal{L}$ to maximize, incorporating the constraints with Lagrange multipliers $\lambda_0, \lambda_1, \lambda_2$:

$$\begin{aligned}\mathcal{L} = &- \int f(x) \ln f(x)\, dx \\ &- \lambda_0 \left(\int f(x)\, dx - 1 \right) \\ &- \lambda_1 \left(\int x f(x)\, dx - \mu \right) \\ &- \lambda_2 \left(\int (x-\mu)^2 f(x)\, dx - \sigma^2 \right)\end{aligned}$$

We now take the functional derivative with respect to $f(x)$ and set it to zero.

Solving for the Probability Density Function

The derivative yields:

$$\frac{\partial \mathcal{L}}{\partial f(x)} = -1 - \ln f(x) - \lambda_0 - \lambda_1 x - \lambda_2 (x-\mu)^2 = 0$$

Rearranging the terms, we get:

$$\begin{aligned} \ln f(x) &= -1 - \lambda_0 - \lambda_1 x - \lambda_2 (x-\mu)^2 \\ f(x) &= \exp\left(-1 - \lambda_0 - \lambda_1 x - \lambda_2 (x-\mu)^2\right) \\ &= C \exp\left(-Ax^2 - Bx\right) \end{aligned} \tag{2}$$

where above $C = \exp\left(-(1 + \lambda_0 + \lambda_2 \mu^2)\right)$, $A = \lambda_2$, $B = (\lambda_1 - 2\mu\lambda_2)$.

Determining the Multipliers

Next, we use the relations satisfied by $f(x)$ of Eq. 2 to determine the multipliers

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

$$C\int_{-\infty}^{\infty} \exp\left(-Ax^2 - Bx\right) = 1$$

$$C\sqrt{\frac{\pi}{A}}\exp\left(\frac{B^2}{4A}\right) = 1 \tag{3}$$

Determining the multipliers...

And

$$\int_{-\infty}^{\infty} xf(x)dx = \mu$$

$$C\int_{-\infty}^{\infty} x\exp\left(-Ax^2 - Bx\right)dx = \mu$$

$$C\exp\left(\frac{B^2}{4A}\right)\int_{-\infty}^{\infty} x\exp\left(-A\left(x+\frac{B}{2A}\right)^2\right)dx = \mu$$

Above we used the result

$$-Ax^2 - Bx = -A\left(x+\frac{B}{2A}\right)^2 + \frac{B^2}{4A}$$

By making the substitution $t = x + \frac{B}{2A}$

$$C\exp\left(\frac{B^2}{4A}\right)\int_{-\infty}^{\infty} (t-\frac{B}{2A})\exp\left(-At^2\right)dt = \mu$$

Determining the multipliers...

leading to

$$-C\left(\frac{B}{2A}\right)\exp\left(\frac{B^2}{4A}\right)\sqrt{\frac{\pi}{A}}=\mu$$

Using Eq. 3 on the LHS of the previous equation, we obtain

$$\mu=-\frac{B}{2A}$$

Therefore

$$f(x)=C\exp\left(-Ax^2-Bx\right)=C\exp\left(\frac{B^2}{4A}\right)\exp\left(-A\left(x+\frac{B}{2A}\right)^2\right)$$

$$=C\exp\left(\frac{B^2}{4A}\right)\exp\left(-A\left(x-\mu\right)^2\right)$$

Determining the multipliers...

On using Eq. 3, we obtain a simplified expression

$$f(x) = \sqrt{\frac{A}{\pi}} \exp\left(-A\,(x-\mu)^2\right) \tag{4}$$

Now we impose the third Lagrange multiplier condition

$$\int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx = \sigma^2$$

$$\sqrt{\frac{A}{\pi}} \int_{-\infty}^{\infty} (x-\mu)^2 \exp\left(-A\,(x-\mu)^2\right) dx = \sigma^2$$

On making the substitution $t = (x-\mu)\sqrt{A}$

$$\sqrt{\frac{A}{\pi}} A^{-3/2} \int_{-\infty}^{\infty} t^2 e^{-t^2} dt = \sigma^2$$

$$\frac{1}{A\sqrt{\pi}} \left(\frac{\sqrt{\pi}}{2}\right) = \sigma^2 \implies \boxed{A = \frac{1}{2\sigma^2}}$$

The Normal Distribution

This, on substitution in Eq. 4, leads to the final expression

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

Thus, we have shown that the function that maximizes the entropy under the constraints of fixed normalization, mean and variance is the normal distribution.

Measures of Central Tendency

The "Center" of the Data

- These measures describe the central position of a dataset.
- They are often referred to as the "average" of the data.
- The three most common measures are:
  1. **Mean** (Arithmetic Average)
  2. **Median** (Middle Value)
  3. **Mode** (Most Frequent Value)

The Mean

The Arithmetic Average

- The **mean** is the sum of all values divided by the number of values.
- It is sensitive to outliers.
- **Formula:**

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

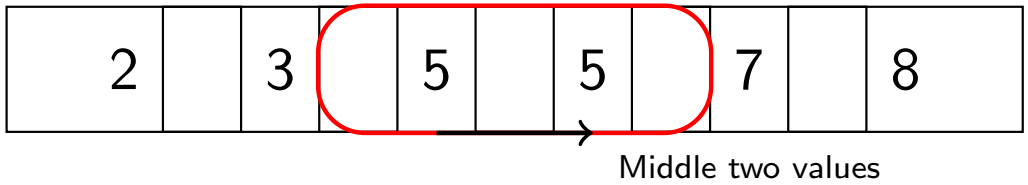
- **Example:** Dataset $= \{2, 3, 5, 5, 7, 8\}$

$$\text{Mean} = \frac{2+3+5+5+7+8}{6} = \frac{30}{6} = 5$$

The Median

The Middle Value

- The **median** is the middle value of a dataset that is ordered from least to greatest.
- It is not affected by outliers.
- **How to find it:**
  - **Odd number of values:** The middle value.
  - **Even number of values:** The average of the two middle values.
- **Example:** Dataset $= \{2, 3, 5, 5, 7, 8\}$



$$\text{Median} = \frac{5+5}{2} = 5$$

The Mode

The Most Frequent Value

- The **mode** is the value that appears most often in a dataset.
- A dataset can have one mode (**unimodal**), more than one mode (**multimodal**), or no mode at all.
- **Example:** Dataset $= \{2, 3, 5, 5, 7, 8\}$
  - The number 5 appears twice. All other values appear once.
  - Therefore, the mode is 5.

Measures of Variability

The "Spread" of the Data

- These measures describe how spread out or dispersed the data is.
- A low measure of variability indicates that the data points tend to be very close to the mean.
- A high measure indicates that the data points are spread out over a wide range.
- Common measures include:
  1. **Range**
  2. **Variance**
  3. **Standard Deviation**

The Range

Maximum minus Minimum

- The **range** is the difference between the highest and lowest values in a dataset.
- It is the simplest measure of variability, but can be misleading due to outliers.
- **Formula:**

$$\text{Range} = \text{Maximum value} - \text{Minimum value}$$

- **Example:** Dataset $= \{2, 3, 5, 5, 7, 8\}$

$$\text{Range} = 8 - 2 = 6$$

Variance and Standard Deviation

Average Deviation from the Mean

- **Variance (s^2)** measures the average of the squared differences from the mean.
- **Standard Deviation (s)** is the square root of the variance. It is a more interpretable measure because it is in the same units as the original data.
- **Formulas:** With Bessel correction ($n-1$, instead of n)

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} \quad \text{and} \quad s = \sqrt{s^2}$$

Measures of Distribution

The Shape of the Data

- These measures describe the shape of the data's distribution.
- They help us understand whether the data is symmetrical or skewed.
- The two most common measures are:
  1. **Skewness**
  2. **Kurtosis**

Skewness

Symmetry of the Distribution

- **Skewness** measures the asymmetry of the data distribution.
- **Negative Skew:** The tail on the left is longer; mean $<$ median.
- **Positive Skew:** The tail on the right is longer; mean $>$ median.
- **Zero Skew:** A symmetric distribution; mean $=$ median.

Skewed Distributions Examples

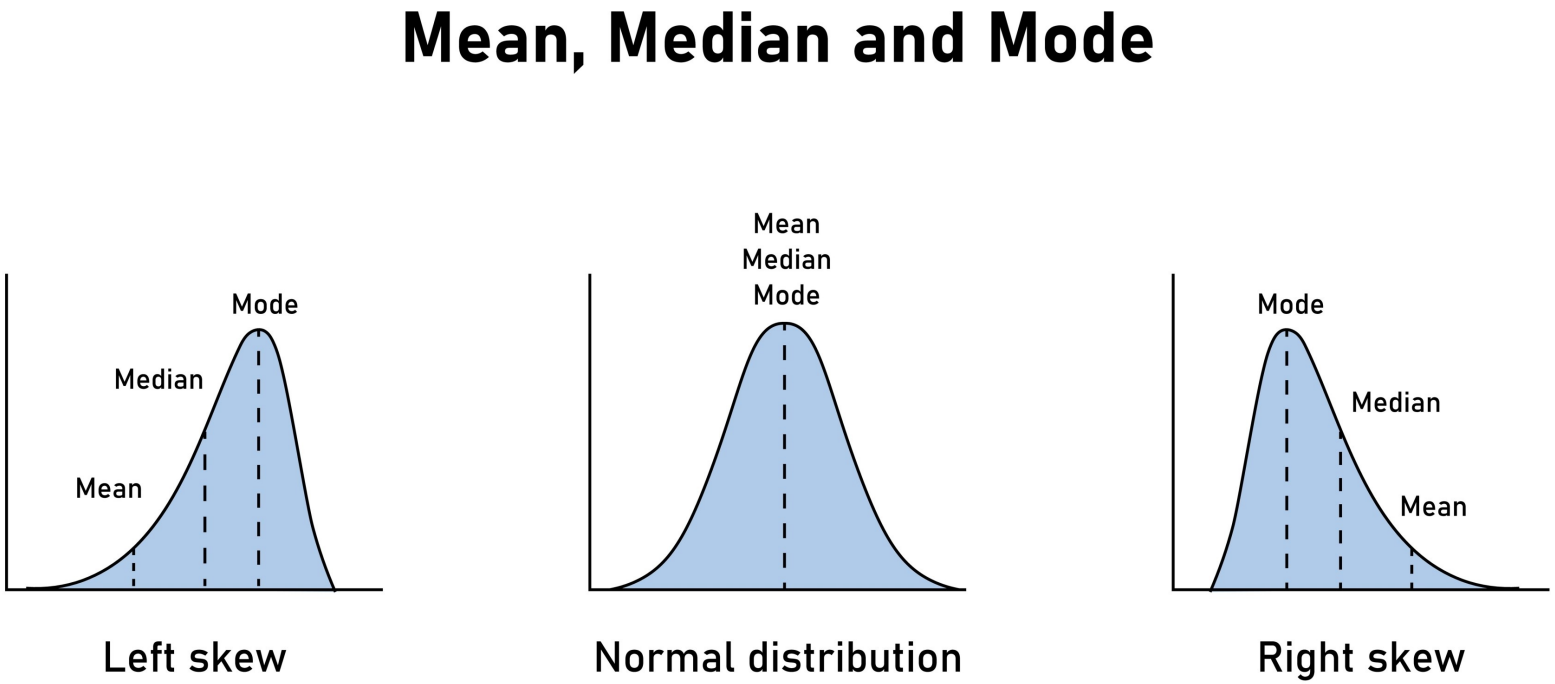


Figure: Examples of left-skewed, normal, and right-skewed distributions, with mean, median and mode indicated. (Courtesy, Google Gemini)

Types of Moments

- Moments are a set of statistical measures that describe the **shape and characteristics** of a probability distribution.
- They provide a way to summarize the key features of a dataset or a random variable.
- ****Raw Moments** (Moments about the origin)**
  - The k-th raw moment for a random variable X is defined as:

$$\mu'_k = E[X^k]$$

    - The first raw moment (μ'_1) is clearly the **mean** ($E[X]$).

Types of Moments...

- ****Central Moments** (Moments about the mean)**
  - The k-th central moment for a random variable X is defined as:

$$\mu_k = E[(X - \mu)^k]$$

  - The first central moment (μ_1) is always 0.
  - The second central moment (μ_2) is the **variance** ($E[(X - \mu)^2]$).
- **The first four moments are the most commonly used:**
  - 1st Moment: Mean (μ)
  - 2nd Moment: Variance (σ^2)
  - 3rd Moment: Skewness
  - 4th Moment: Kurtosis

Some Central Moments for the Normal Distribution

The normal distribution is given by

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

For the normal distribution, a given central moment μ_k is given by

$$\mu_k = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-\mu)^k e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx.$$

Let us make the substitution $t = \frac{(x-\mu)}{\sigma\sqrt{2}} \implies dx = \sigma\sqrt{2}dt$, leading to

$$\mu_k = \frac{\sigma^k 2^{k/2}}{\sqrt{\pi}} \int_{-\infty}^{\infty} t^k e^{-t^2} dt. \tag{5}$$

Moments of the Normal Distribution

- Clearly, all odd moments are zero because for odd values of k, the integrand is an odd function of t. Therefore

$$\mu_1 = 0$$
$$\mu_3 = 0 \text{ (which means skewness=0)}$$
$$\text{etc.}$$

- One can easily compute using Eq. 5, and various Gaussian integrals

$$\mu_0 = 1$$
$$\mu_2 = \sigma^2$$
$$\mu_4 = 3\sigma^4 \tag{6}$$

Formulas for Skewness

- Skewness is the third standardized moment and measures the asymmetry of a distribution.
- **Population Skewness (γ_1):**

$$\gamma_1 = \frac{E[(X-\mu)^3]}{\sigma^3} = \frac{\mu_3}{\sigma^3}$$

  - Clearly for the normal distribution $\gamma_1 = 0$, because $\mu_3 = 0$
- For a discrete distribution with a population of N values, it is calculated as

$$\gamma_1 = \frac{1}{\sigma^3} \sum_{i=1}^{n} \frac{(x_i - \bar{x})^3}{N},$$

where $\bar{x}$ is the sample mean, and σ is the sample standard deviation defined as

$$\sigma = \left[\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{N} \right]^{1/2}$$

Skewness...

- For a discrete distribution with a sample of n values ($n \ll N$), it is calculated as

$$g_1 = \frac{1}{s^3} \sum_{i=1}^{n} \frac{(x_i - \bar{x})^3}{n},$$

where $\bar{x}$ is the sample mean, and s is the sample standard deviation defined as

$$s = \left[\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} \right]^{1/2},$$

and $n-1$ in the denominator is called Bessel's correction to account for small values of n

- $\gamma_1/g_1 = 0 \Rightarrow$ symmetric distribution
- $\gamma_1/g_1 > 0 \Rightarrow$ right-skewed
- $\gamma_1/g_1 < 0 \Rightarrow$ left-skewed

Formulas for Kurtosis

- Kurtosis is the fourth standardized moment and measures the "tailedness" of a distribution.
- **Population Kurtosis (γ_2):**

$$\gamma_2 = \frac{E[(X-\mu)^4]}{\sigma^4} = \frac{\mu_4}{\sigma^4}$$

  - Using Eq. 6 above, we find that for the normal distribution $\gamma_2 = 3$
- For a discrete distribution with the population N, the kurtosis is defined as

$$\gamma_2 = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^4}{\sigma^4}$$

- On the other hand, for a sample of n values

$$g_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4}{s^4}$$

- $\gamma_2 = 3$ for normal distribution
- $\gamma_2/g_2 > 3 \Rightarrow$ leptokurtic (peaked)
- $\gamma_2/g_2 < 3 \Rightarrow$ platykurtic (flat)

Kurtosis: Summary Discussion

- Kurtosis is a statistical measure that describes the **shape of a distribution's tails** relative to its center.
- It quantifies how much of a distribution's variance comes from **extreme values** (outliers).
- A higher kurtosis value means more of the distribution's data is located in the tails.

Kurtosis: The baseline

- **Mesokurtic** ($\gamma_2 = 3$)
  - This is the benchmark for kurtosis.
  - A mesokurtic distribution has tails of a moderate length and thickness.
  - The **standard normal distribution** is the classic example, with a kurtosis of exactly 3.

Because $\gamma_2 = 3$ is considered to be the baseline, many authors define the so-called "excess kurtosis" as

$$\text{excess kurtosis} = \gamma_2 - 3$$

Types of Kurtosis: Beyond the Baseline

- **Leptokurtic** ($\gamma_2 > 3$)
  - Latin for "slender."
  - Characterized by a **sharp peak** and **heavy tails**.
  - Indicates a higher probability of extreme outcomes or outliers.
- **Platykurtic** ($\gamma_2 < 3$)
  - Latin for "broad" or "flat."
  - Characterized by a **flatter peak** and **light tails**.
  - Indicates a lower probability of extreme outcomes.

Kurtosis: An Illustration

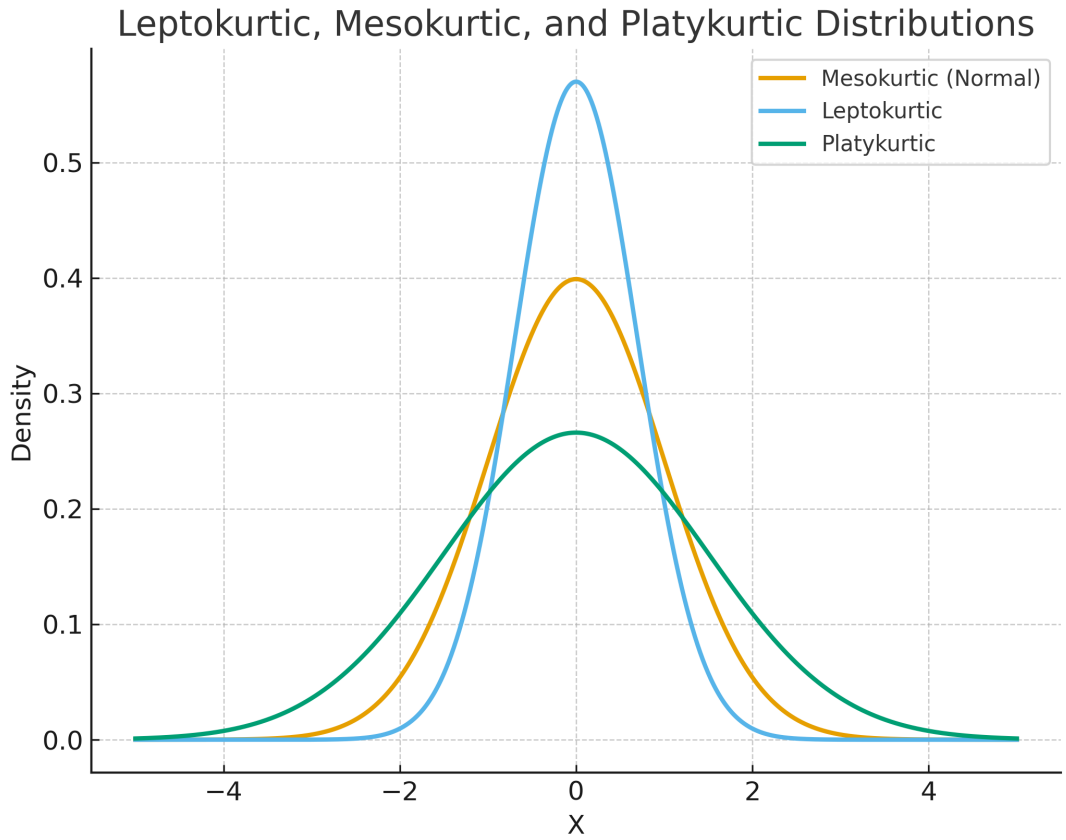


Figure: Distributions with various types of Kurtosis

Example 1: Skewness Calculation

Data: 2, 4, 6, 8, 10

- Mean: $\bar{x} = \frac{2+4+6+8+10}{5} = 6$
  - Deviations: $-4, -2, 0, 2, 4$
  - Variance: $\frac{16+4+0+4+16}{5} = 8$
  - Standard Deviation: $\sqrt{8} \approx 2.828$
  - Third central moment: $\frac{(-4)^3+(-2)^3+0^3+2^3+4^3}{5} = 0$

$$\gamma_1 = \frac{0}{(8)^{3/2}} = 0$$

Conclusion: Data is symmetric.

Example 2: Skewness (Asymmetric Data)

Data: 1, 2, 2, 3, 9

- Mean: $\bar{x} = \frac{17}{5} = 3.4$
  - Deviations: $-2.4, -1.4, -1.4, -0.4, 5.6$
  - Variance: $\frac{5.76+1.96+1.96+0.16+31.36}{5} = 8.24$
  - Std Dev: $\sqrt{8.24} \approx 2.87$
  - Third moment: $\frac{(-2.4)^3+(-1.4)^3+(-1.4)^3+(-0.4)^3+(5.6)^3}{5} = 41.92$

$$\gamma_1 = \frac{41.92}{(8.24)^{3/2}} \approx 1.60$$

Conclusion: Data is positively skewed.

Example 3: Kurtosis

Data: 2, 4, 6, 8, 10

- Mean $= 6$
  - Variance $= 8$
  - Fourth moment: $\frac{256+16+0+16+256}{5} = 108.8$

$$\gamma_2 = \frac{108.8}{(8)^2} = \frac{108.8}{64} = 1.7$$

Conclusion: Since $\gamma_2 < 3$, the distribution is platykurtic (flatter than normal).

Data: 5, 5, 6, 6, 7

- Mean $= 5.8$
  - Variance $= \frac{(-0.8)^2+(-0.8)^2+0.2^2+0.2^2+1.2^2}{5} = 0.56$
  - Fourth moment $= \frac{0.41+0.41+0.0016+0.0016+2.07}{5} = 0.58$

$$\gamma_2 = \frac{0.58}{(0.56)^2} \approx 1.85$$

Conclusion: Still platykurtic, but closer to normal.

Important Formulas: summary

	Population	Sample
Mean	$\mu = \frac{1}{N}\sum x_i$	$\bar{x} = \frac{1}{n}\sum x_i$
Variance	$\sigma^2 = \frac{1}{N}\sum(x_i - \mu)^2$	$s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$
Skewness	$\gamma_1 = \frac{\frac{1}{N}\sum(x_i-\mu)^3}{\sigma^3}$	$g_1 = \frac{\frac{1}{n}\sum(x_i-\bar{x})^3}{s^3}$
Kurtosis	$\gamma_2 = \frac{\frac{1}{N}\sum(x_i-\mu)^4}{\sigma^4}$	$g_2 = \frac{\frac{1}{n}\sum(x_i-\bar{x})^4}{s^4}$