Math for Data Science

Introduction to Linear Algebra

Linear algebra is a branch of mathematics that deals with vectors, matrices and linear transformations.
It provides tools to represent and manipulate data in multidimensional space, making it foundational for fields like data science, computer graphics, engineering and physics.

Representation of Data

In data science, data is often represented as vectors and matrices.
Linear algebra provides the tools to work with these structures.

Example 1

A dataset with n sample and m features is represented as a matrix X of size nxm, where each row is a sample (data point) and each column is a feature.

$$X = \begin{bmatrix} 5 & 3 & 2 \\ 6 & 7 & 8 \\ 4 & 5 & 9 \end{bmatrix}$$

Example 2 :

An Image is essentially a grid of pixels, and it can be represented as a matrix in linear algebra.
For grayscale images, each pixel's intensity is represented as a value between 0 (black) and 255 (White).

Here, X is a 3x3 matrix where each value represents the intensity of a pixel.

$$X = \begin{bmatrix} 155 & 153 & 255 \\ 62 & 75 & 28 \\ 4 & 5 & 0 \end{bmatrix}$$

- **Dimensionality Reduction**
  Linear algebra underpins techniques like Principal Component Analysis (PCA) which reduces dimensionality of data while preserving its variance.
- **Data Transformation**
  Data manipulation often involves operations like scaling, rotation or projection which rely on matrix operations.
- **Neural Networks**
  Linear algebra is at the core of neural networks. Each layers computation involves matrix multiplications:
  For a single layer
  $Z = WX + b$
  Here,
  W = weight, X = input matrix, b=bias vector, z = output

- **Optimization**
  Optimization algorithms like gradient descent use linear algebra for efficient computation of gradient and updates.

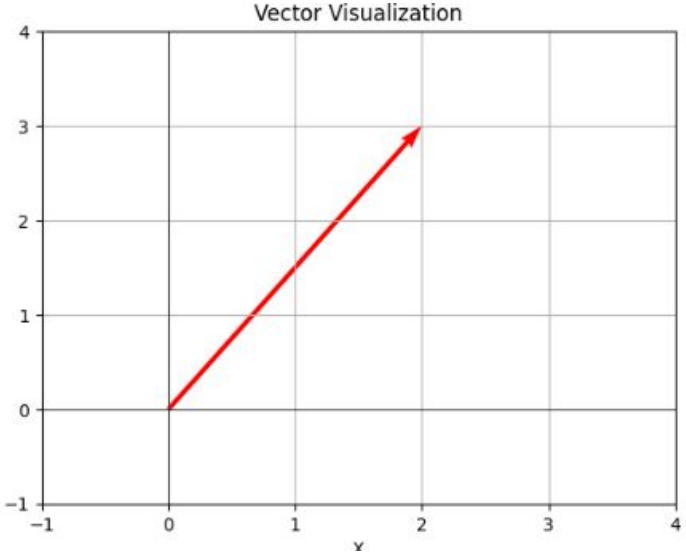
Vectors, Matrices and Matrix Factorization

- **Scalar:** A scalar is a single number. It represents a quantity with no direction. E.g. 5
- **Vectors:** A vector is a one-dimensional array of numbers (scalars) that represent quantities with both magnitude and direction.

  Geometrically, a vector can be visualized as an arrow in space.

E.g

$$X = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$



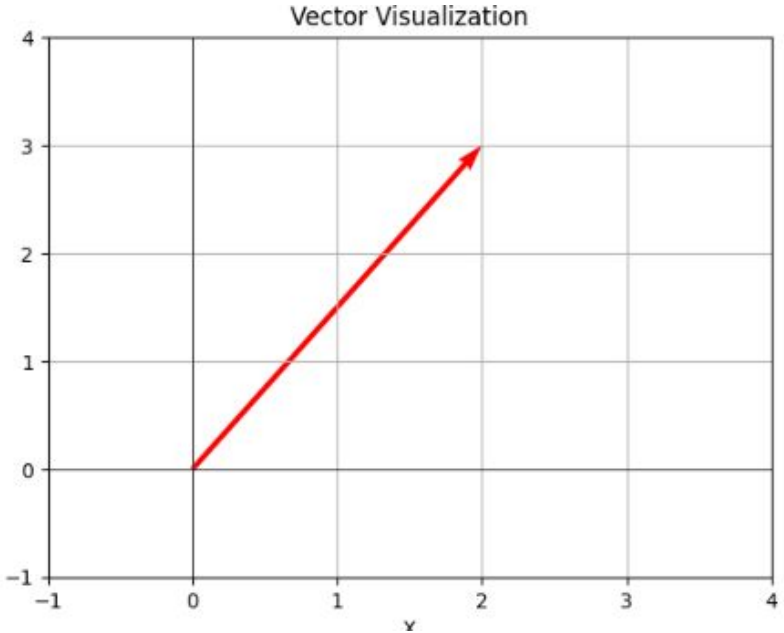
- **Matrices:**

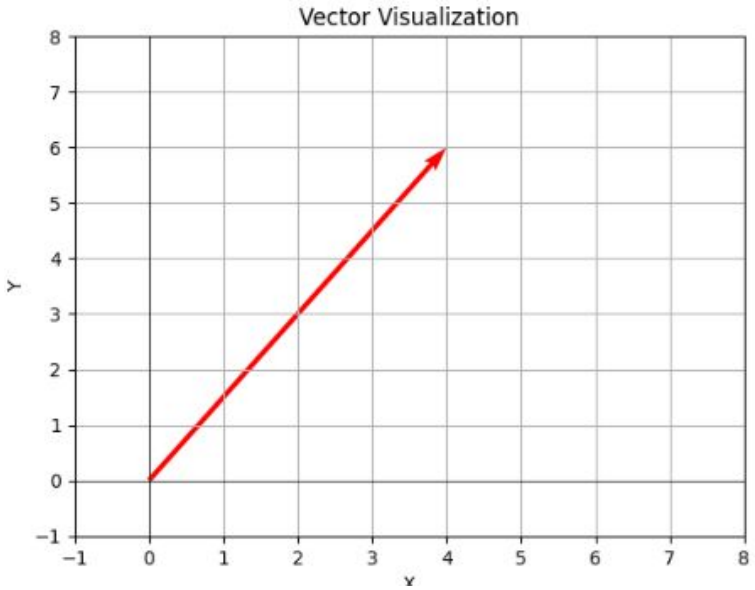
  A matrix is a two-dimensional array of numbers arranged in rows and columns.

$$X = \begin{bmatrix} 155 & 153 & 255 \\ 62 & 75 & 28 \\ 4 & 5 & 0 \end{bmatrix}$$

Linear Transformation

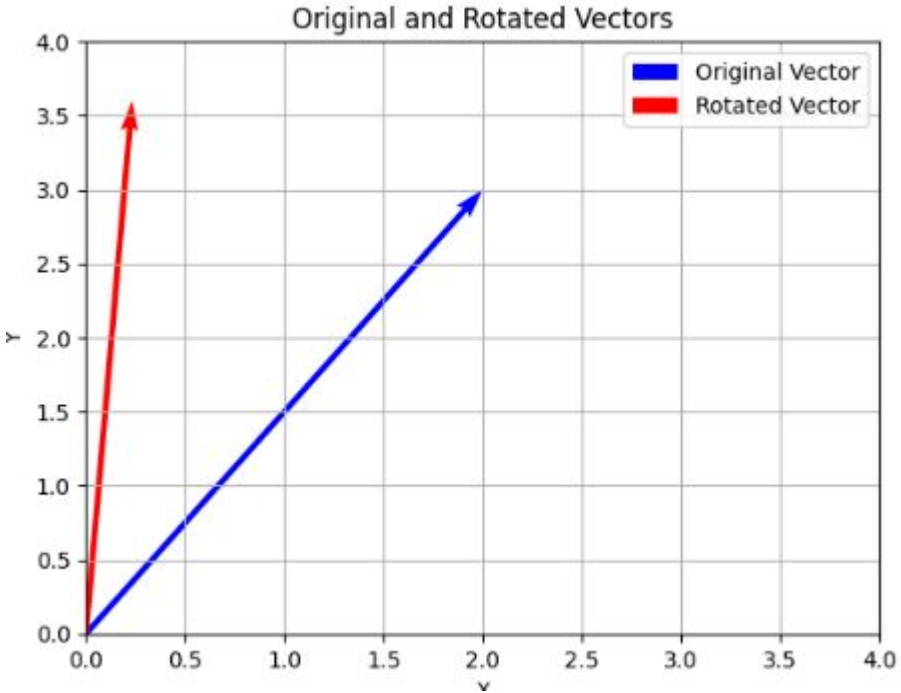
1. Scaling





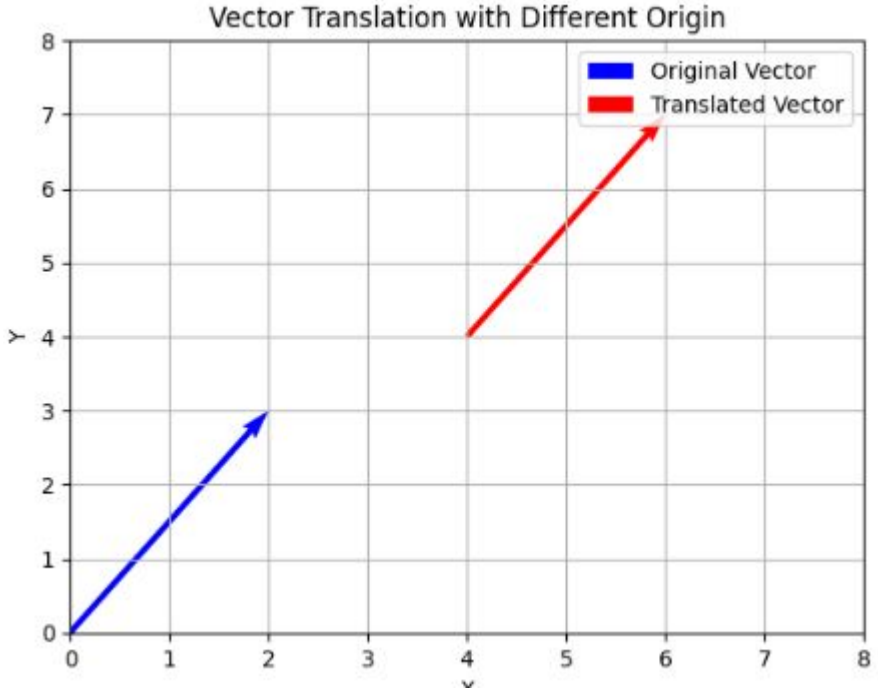
Linear Transformation

2. Rotation



Linear Transformation

3. Translation



Systems of linear equations

Linear algebra provides methods to solve systems of linear of equations such as

$2x+3y = 5$

$x-y = 1$

- Substitution method
- Matrix inversion
- Gaussian elimination
- LU Decomposition

Operations in Linear algebra

1. Addition and subtraction
2. Scalar multiplication
3. Dot Product
4. Matrix multiplication
5. Determinants and Inverse

Matrix Factorization

Matrix factorization is a mathematical technique in linear algebra where a matrix is decomposed into the product of two or more smaller matrices.

These smaller matrices capture latent structures or properties of the original matrix, making it easier to analyze, process or approximate

Key idea behind matrix factorization
Given a matrix A of size mxn, matrix factorization involves decomposing it into two or more matrices such that.

$A \approx P.\ Q$
Where,

- P is an mxk
- Q is a k x n
- K is the latent dimension (smaller that m and n)

Why matrix factorization?

- **Dimensionality Reduction**
  Reduces the size of data while retaining most of its meaningful structure.
- **Pattern Discovery**
  Extracts latent factors or hidden relationships in the data
- **Approximation**
  Provides a simple representation of a complex matrix

Types of matrix factorization

Singular value Decomposition (SVD)

Decompose a matrix A into three matrices

$A = U\sum V^T$

U : Orthogonal matrix of left singular vectors

$\sum$: Diagonal matrix of singular values

V^T : Orthogonal matrix of right singular vectors

Example

$$A = \begin{bmatrix} 4 & 0 \\ 3 & -5 \end{bmatrix}$$

$$U = \begin{bmatrix} -0.44 & -0.89 \\ -0.89 & 0.44 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 6.32 & 0 \\ 0 & 3.16 \end{bmatrix} \qquad V^T = \begin{bmatrix} -0.707 & 0.707 \\ -0.707 & -0.707 \end{bmatrix}$$

SVD used for image compression

Non Negative Matrix Factorization (NMF)

Factorizes A into P and Q with the constraint that all elements are non-negative

$A \approx PQ$

$P,Q \geq 0$

Application: Text mining, image processing and recommender systems.

Gradient Descent for Optimization

Gradient descent is an approximation algorithm used to minimize a function by iteratively moving to minimize a function by iteratively moving in the direction of the steeper descent, as defined by the negative of the gradient.

Key Idea:

1. Start with an initial guess for the parameters.
2. Compute the gradient of the loss function with respect to the parameters
3. Update the parameters in the direction opposite to the gradient to reduce the loss.
4. Repeat until convergence (the parameters stabilize or the loss stops decreasing significantly)

Formula

For a parameter $\boldsymbol{\theta}$ the update rule is

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta J(\theta_t)$$

Where:

$\boldsymbol{\theta}_t$: Parameter at iteration t

η : learning rate (step size)

$\nabla_\theta J(\theta_t)$: Gradient of the loss function J($\boldsymbol{\theta}$) with respect to the $\boldsymbol{\theta}$

Minimizing a Simple Quadratic Function

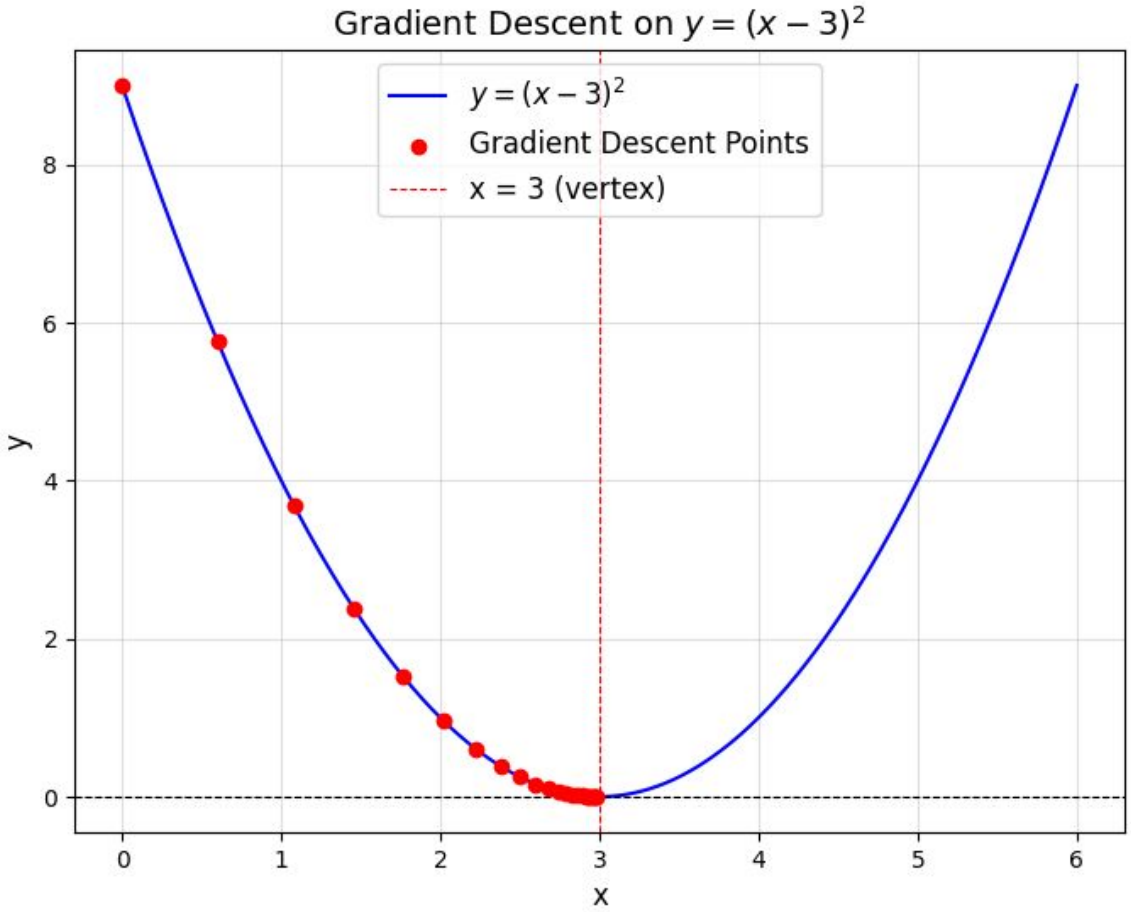
$J(\boldsymbol{\theta}) = (\boldsymbol{\theta} - 3)^2$

- Gradient calculation
  The gradient of $J(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ is:
  $$\nabla_\theta J(\theta) = 2(\theta - 3)$$
- Initialization
  Start with $\boldsymbol{\theta}_0 = 0$ (initial guess)
- Iteration (with learning rate $\eta = 0.1$)
  At each step update $\boldsymbol{\theta}$ using the formula

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta J(\theta_t)$$

```
theta: 0 func: 9
theta: 0.6 func: 5.76
theta: 1.08 func: 3.69
theta: 1.46 func: 2.37
theta: 1.77 func: 1.51
theta: 2.02 func: 0.96
theta: 2.22 func: 0.61
theta: 2.38 func: 0.38
theta: 2.5 func: 0.25
```

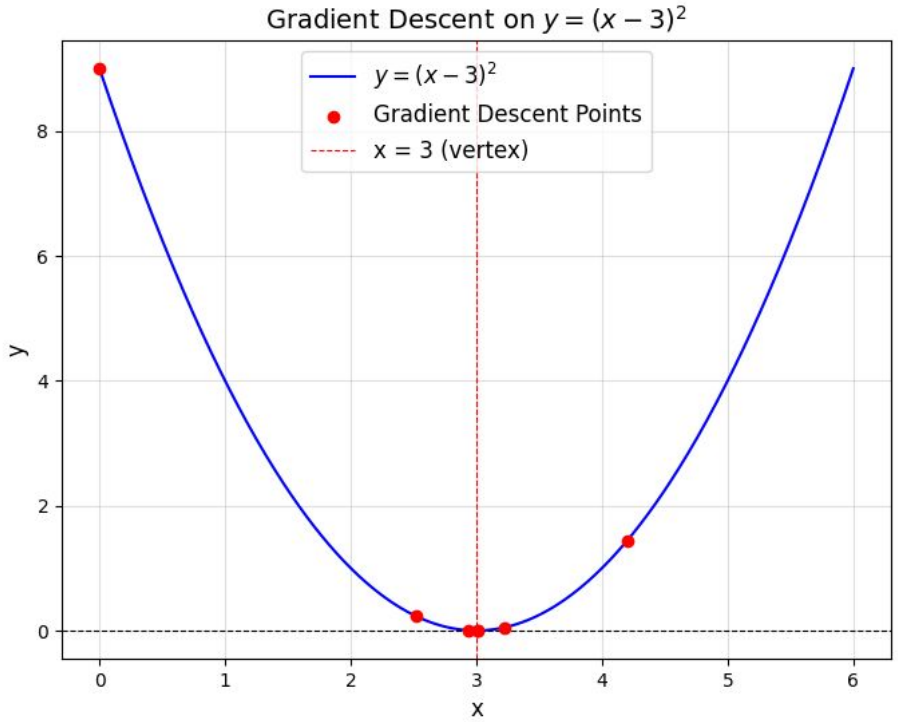
```
theta: 2.6 func: 0.16

theta: 2.68 func: 0.1

theta: 2.74 func: 0.07

theta: 2.79 func: 0.04

theta: 2.83 func: 0.03

theta: 2.86 func: 0.02

theta: 2.89 func: 0.01

theta: 2.91 func: 0.01

theta: 2.93 func: 0.0
```


Gradient Descent on $y = (x - 3)^2$
$y = (x - 3)^2$
Gradient Descent Points
x = 3 (vertex)
y
x
0
2
4
6
8
0
1
2
3
4
5
6

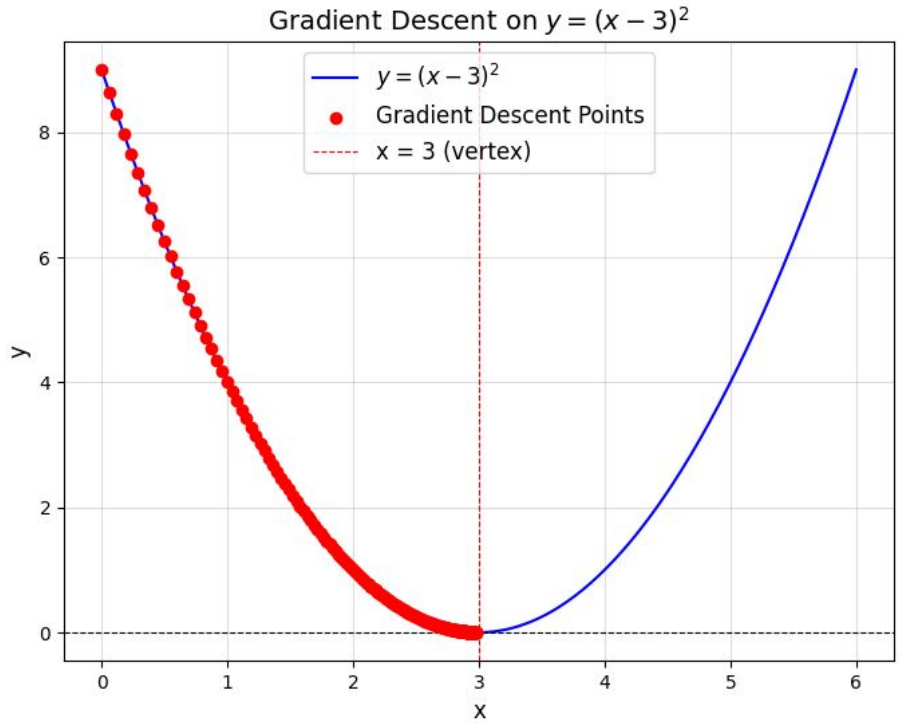
With learning rate = 0.7

```
theta: 0 func: 9

theta: 4.2 func: 1.44

theta: 2.52 func: 0.23

theta: 3.22 func: 0.05

theta: 2.94 func: 0.0

theta: 3.01 func: 0.0
```

Changing the learning rate to 0.7



Changing the learning rate to 0.01



Minimizing Mean Square Error Using gradient descent

Mean Square Error (MSE)

Use case: Regression problems

Formula: $$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

y_i : True value

$\hat{y}_i$ Predicted value

n : number of samples

For linear models : $$\hat{y}_i = wx_i + b$$

Gradient descent steps for MSE

1. Compute the gradient

   We differentiate the MSE loss with respect to the parameters w and b

- With respect to w

$$\frac{\partial \text{MSE}}{\partial w} = -\frac{2}{n}\sum_{i=1}^{n} x_i\,(y_i - \hat{y}_i)$$

- With respect to b

$$\frac{\partial \text{MSE}}{\partial b} = -\frac{2}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)$$

Gradient descent steps for MSE

2. Update Parameters

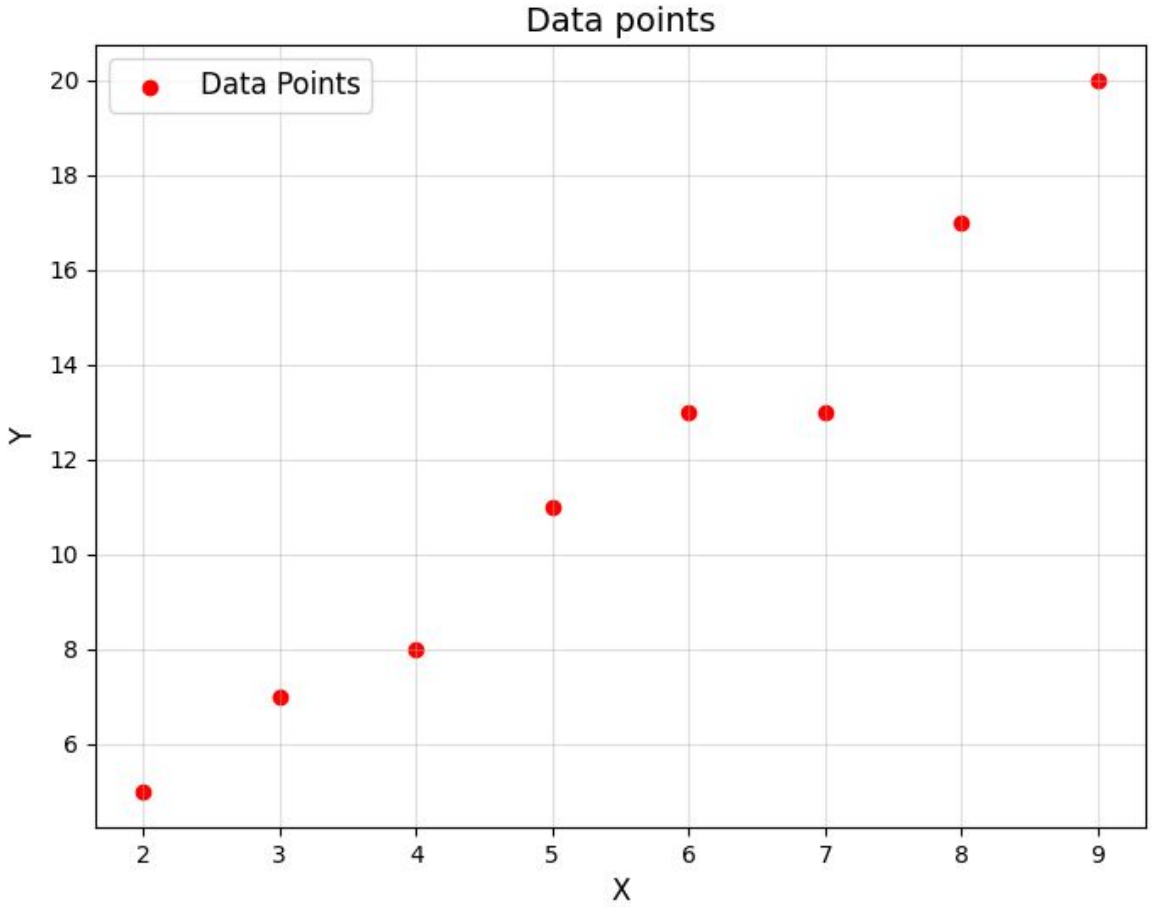
$$w = w - \eta \frac{\partial \text{MSE}}{\partial w}$$

$$b = b - \eta \frac{\partial MSE}{\partial b}$$

Example

Dataset

x	2	3	4	5	6	7	8	9
y	5	6	8	11	13	13	17	17


Data points
Data Points
20
18
16
14
12
10
8
6
Y
2
3
4
5
6
7
8
9
X

Using gradient descent to optimize MSE

y = wx + b

Initial

w= 0

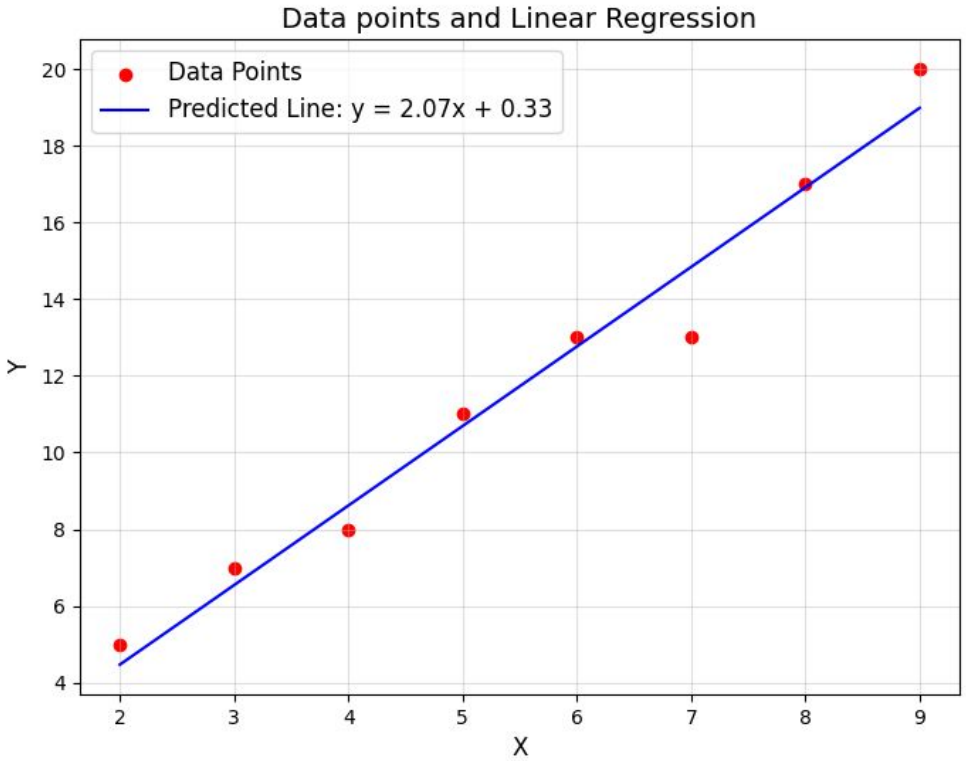
b = 0

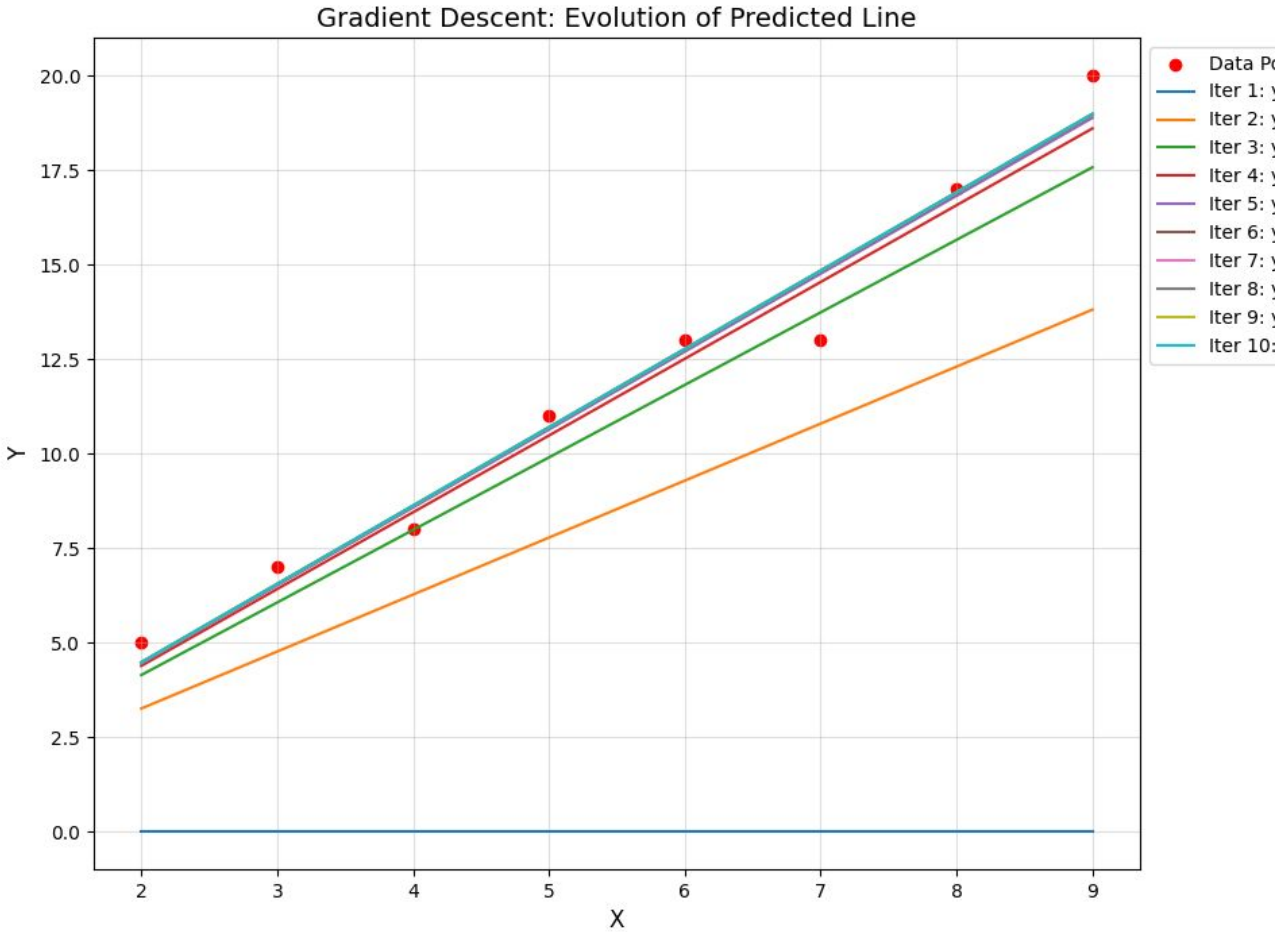
Learning rate: 0.01

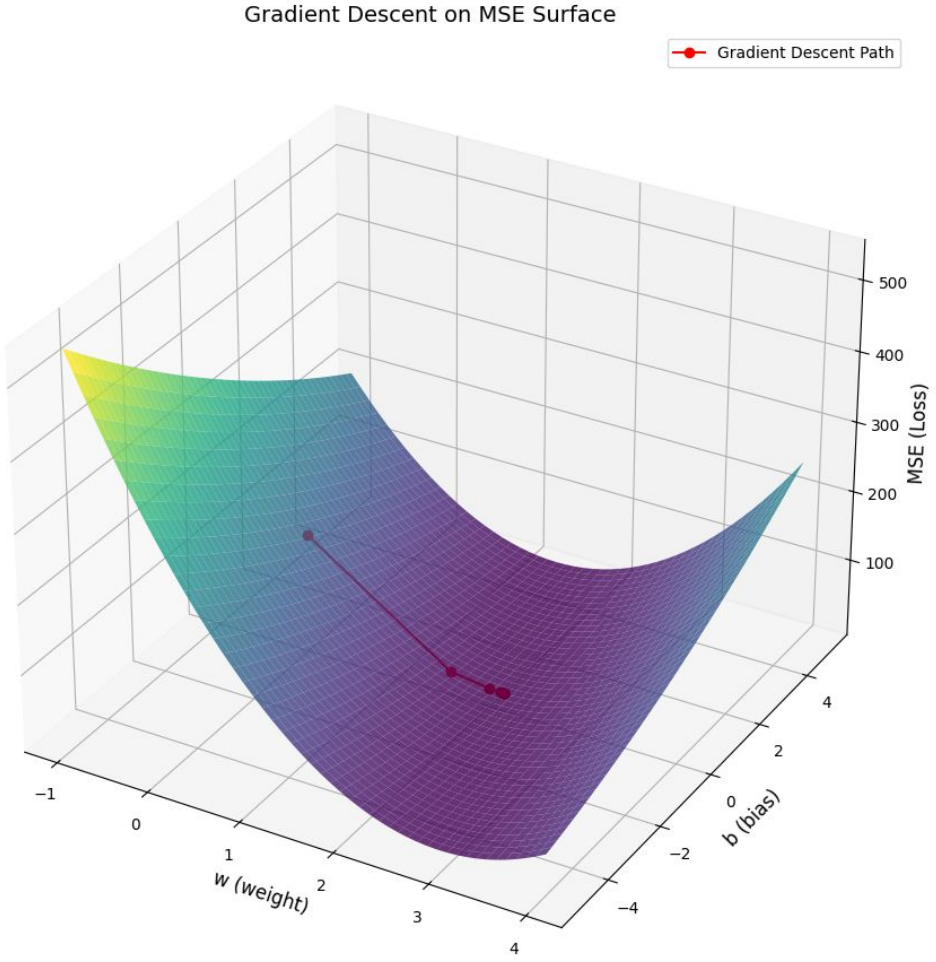
```
Iteration 1: w = 1.5075, b = 0.2350, MSE = 160.7500

Iteration 2: w = 1.9188, b = 0.2995, MSE = 12.5997

Iteration 3: w = 2.0310, b = 0.3174, MSE = 1.5673

Iteration 4: w = 2.0616, b = 0.3227, MSE = 0.7457

Iteration 5: w = 2.0699, b = 0.3244, MSE = 0.6845
```

```
Iteration 6: w = 2.0721, b = 0.3253, MSE = 0.6799
Iteration 7: w = 2.0726, b = 0.3258, MSE = 0.6795
Iteration 8: w = 2.0727, b = 0.3263, MSE = 0.6795
Iteration 9: w = 2.0727, b = 0.3268, MSE = 0.6795
Iteration 10: w = 2.0726, b = 0.3273, MSE = 0.6794
```

Fitting data points




Gradient Descent: Evolution of Predicted Line
Data Po
Iter 1: y
Iter 2: y
Iter 3: y
Iter 4: y
Iter 5: y
Iter 6: y
Iter 7: y
Iter 8: y
Iter 9: y
Iter 10:
Y
X
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
2
3
4
5
6
7
8
9


Gradient Descent on MSE Surface
Gradient Descent Path
MSE (Loss)
500
400
300
200
100
w (weight)
−1
0
1
2
3
4
b (bias)
4
2
0
−2
−4

Exercise

Find the value of w and b for which y is minimum

$y = (w-20)^2+(b-2)^2$

Introduction to Probability and Random Variable

Probability is a branch of mathematics that measures the likelihood of an event occurring. It quantifies uncertainty and assigns a numerical value, typically between 0 and 1, to the occurrence of an event:

- **0** indicates the event is impossible.
- **1** indicates the event is certain.

$$P(E) = \frac{\text{Number of favorable outcomes}}{\text{Total number of possible outcomes}}$$

Probability in Data Science

1. Spam Detection
    a. **Problem**: Classify emails as "spam" or "not spam."
    b. **Use of Probability**: Algorithms like Naïve Bayes calculate the probability of an email being spam given specific features (e.g., words like "discount" or "free").

$$P(Spam|Features) = \frac{P(Features)}{P(Features|Spam) \cdot P(Spam)}$$

Probability in Data Science

2. Diseases Prediction in HealthCare
    a. **Problem**: Diagnose or predict diseases based on patient data
    b. **Use of Probability**: Bayesian networks compute the probability of a disease given symptoms and test results. Example: Probability of a patient having diabetes given high glucose levels

$$P(Diabetes|HighGlucose) = \frac{P(HighGlucose)}{P(HighGlucose|Diabetes) \cdot P(Diabetes)}$$

Random Variable

In probability theory, a **random variable** is a mathematical function that assigns numerical values to the outcomes of a random process or experiment. It provides a way to quantify uncertain outcomes, making it easier to analyze and reason about them.

Key Features of a Random Variable:

1. **Domain**: The sample space (S) of the experiment, which is the set of all possible outcomes.
2. **Mapping**: The random variable maps each outcome in the sample space to a real number.

Types of Random Variables

Discrete Random Variable:

- Takes on a countable number of distinct values.
- Example: The result of rolling a six-sided die (X) where

  $X \in \{1,2,3,4,5,6\}$

Continuous Random Variable:

- Takes on an uncountable range of values, typically intervals of real numbers.
- Example: The time it takes for a car to complete a lap in a race (Y), where

  $Y \in [0,\infty)$.

Probability Distributions:

A random variable is associated with a **probability distribution**, which describes how probabilities are distributed over its possible values.

For Discrete Random Variables

Defined by a **probability mass function (PMF)**, which gives the probability that the random variable equals each specific value:

$P(X = x) = f_X(x)$

For Continuous Random Variables

Defined by a **probability density function (PDF)**, which describes the relative likelihood of the random variable taking on a value within a given range. The probability is given by integrating the PDF over an interval:

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

Example

1. Discrete Case

   Tossing a coin twice, define X as the number of heads:

   Sample Space: S={HH, HT, TH, TT}

   X(S): {0,1,2}where

   P(X=0)=0.25

   P(X=1)=0.5

   P(X=2)=0.25

Example

1. Continuous Case

   Measuring the height of students in a class, Y could take any real value within a range (e.g., 150 cm to 200 cm).

Few Probability Distributions

Discrete Probability Distributions:

1. **Bernoulli Distribution**:
   - Models a single trial with two possible outcomes (success or failure).
   - Example: Tossing a coin once.
   - $PMF : P(X = x) = p^x(1-p)^{(1-x)}, \quad x \in \{0, 1\}$
2. **Binomial Distribution**:
   - Models the number of successes in a fixed number of independent trials.
   - Example: Tossing a coin nnn times.
   - $$PMF : P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, 2, 3, . .$$

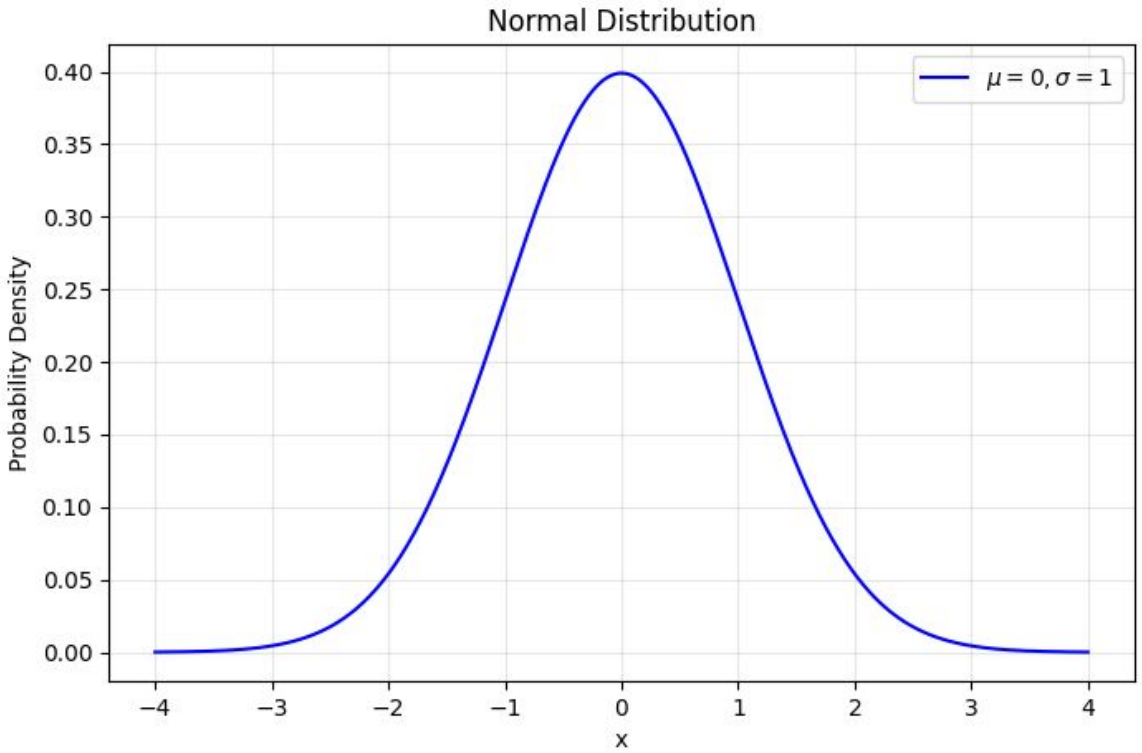
Discrete Probability Distributions:

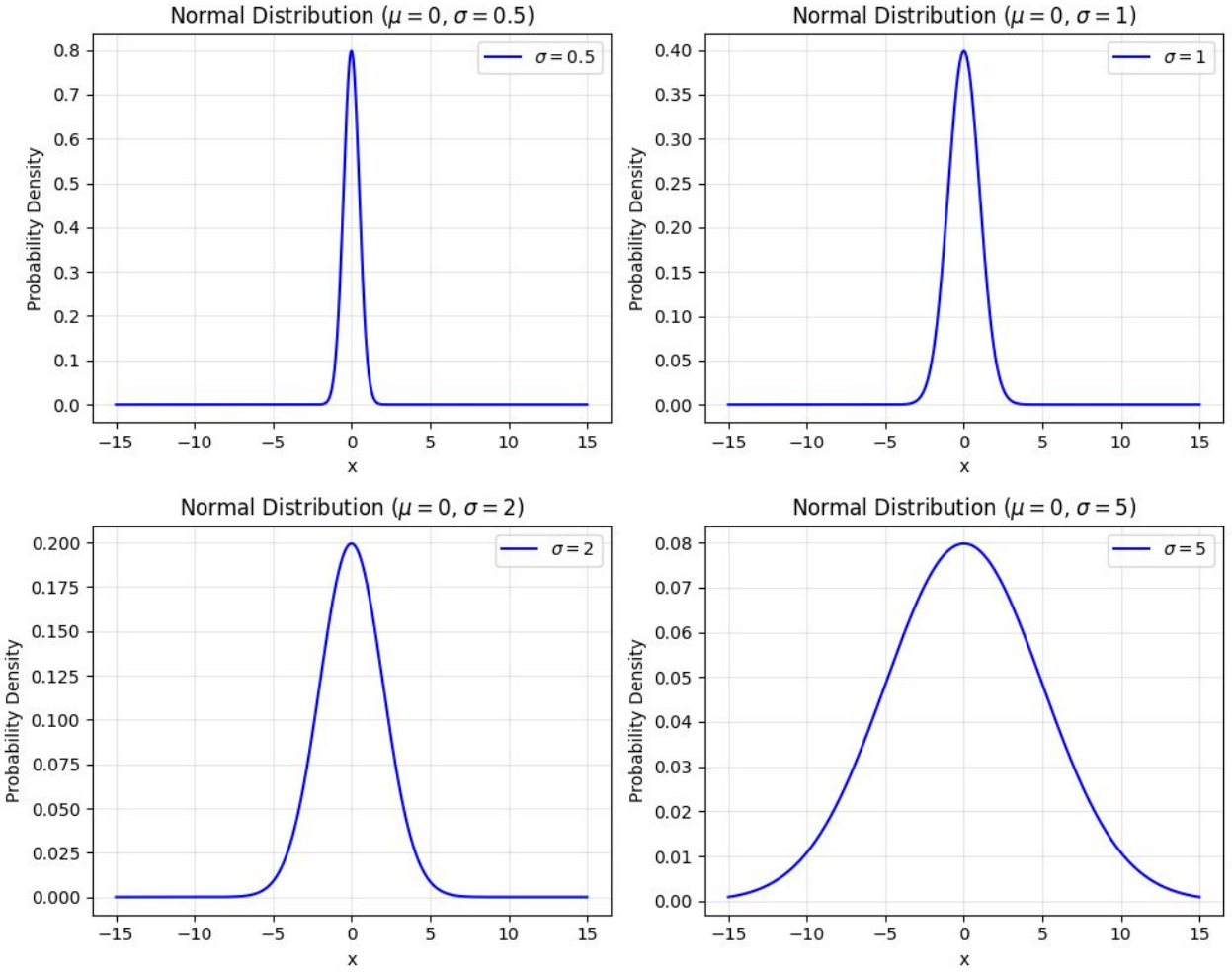
3. **Poisson Distribution**:
   - Models the number of events occurring in a fixed interval of time or space.
   - Example: Number of calls received by a call center in an hour.
   - $PMF: P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, 3, ..$
4. **Geometric Distribution**:
   - Models the number of trials needed to get the first success.
   - Example: Rolling a die until a 6 appears.
   - $PMF: P(X = k) = (1 - p)^{k-1} p, \quad k = 1, 2, 3, \ldots$

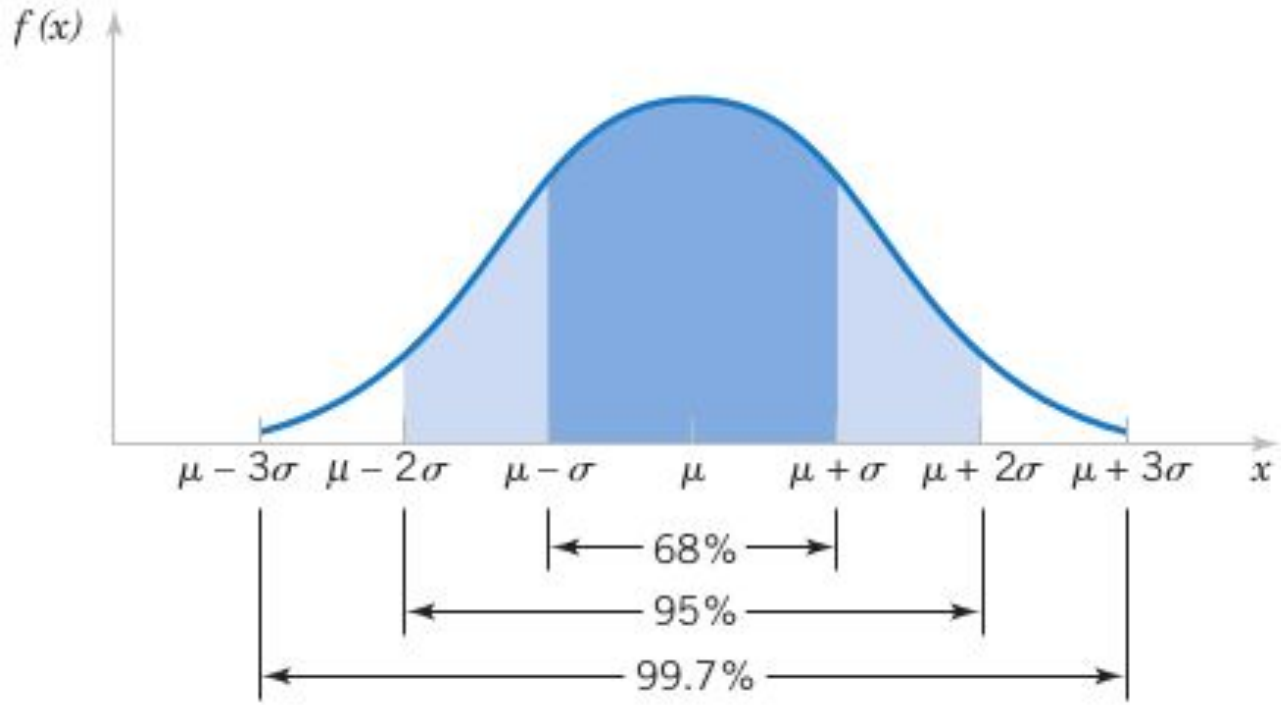
Continuous Probability Distribution

1. **Uniform Distribution**
   a. Models a situation where all outcomes in a range are equally likely.
   b. Example: Randomly picking a number between 0 and 1
   c. $PDF: f(x) = \frac{1}{b-a}, \quad a \leq x \leq b$
2. **Gaussian (Normal) Distribution**
   a. Models many natural phenomena like heights, test scores, etc.
   b. Example: Heights of people in a population.
   c. $PDF: f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in R$

Normal Distribution




Normal Distribution ($\mu = 0$, $\sigma = 0.5$)
$\sigma = 0.5$
Probability Density
x
Normal Distribution ($\mu = 0$, $\sigma = 1$)
$\sigma = 1$
Probability Density
x
Normal Distribution ($\mu = 0$, $\sigma = 2$)
$\sigma = 2$
Probability Density
x
Normal Distribution ($\mu = 0$, $\sigma = 5$)
$\sigma = 5$
Probability Density
x


f (x)
μ − 3σ
μ − 2σ
μ − σ
μ
μ + σ
μ + 2σ
μ + 3σ
x
68%
95%
99.7%

Example:

Assume that the current measurements in a strip of wire follow a normal distribution with a mean of 10 milliamperes and a variance of 4 (milliamperes)2. What is the probability that a measurement exceeds 13 milliamperes?

Solution:

$$\mu = 10 \quad \sigma^2 = 4 \quad or \quad \sigma = 2$$

Solution

Parameters:

- Mean (μ) = 10
- Variance = 4→ Standard Deviation (σ) = 2

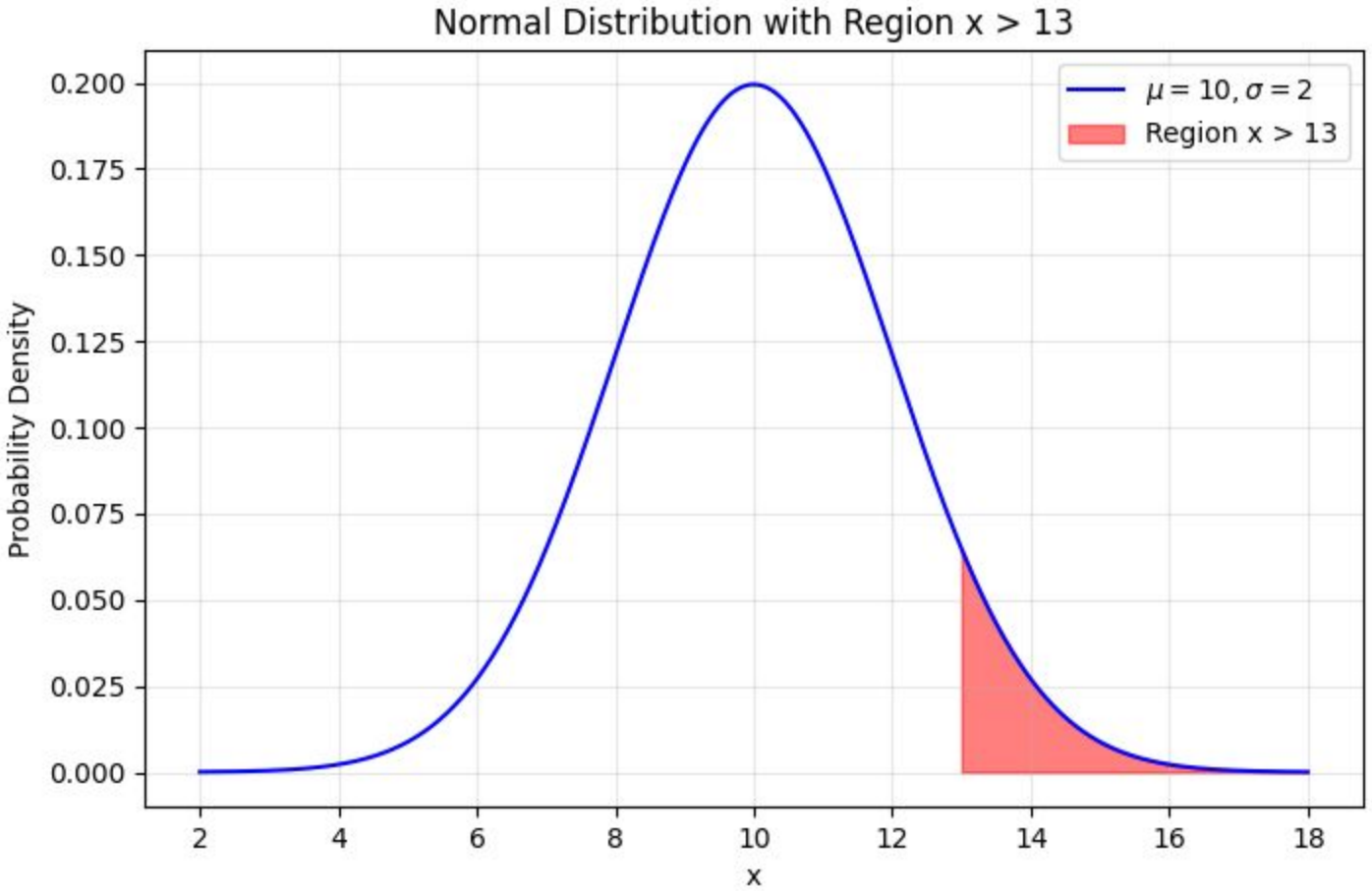
Standardizing the value:

Convert x=13 to a **z-score** using the formula: $$z = \frac{x - \mu}{\sigma}$$

Substituting:
z= 1.5
Find the probability: The probability of a measurement exceeding 13 is:
P(X>13)=1−P(Z≤1.5) = 1 - 0.933 = 0.0668

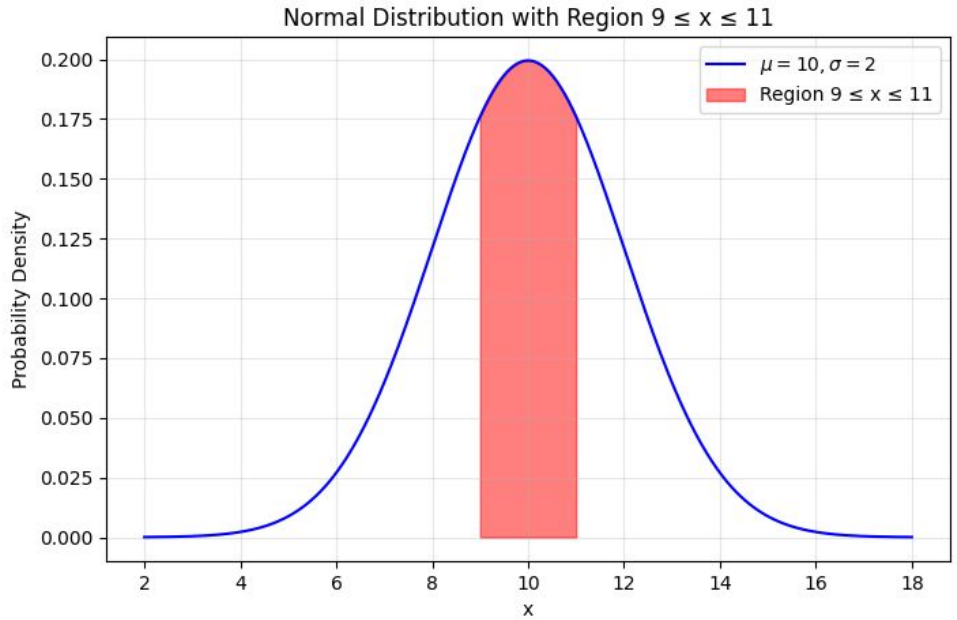
Python implementation

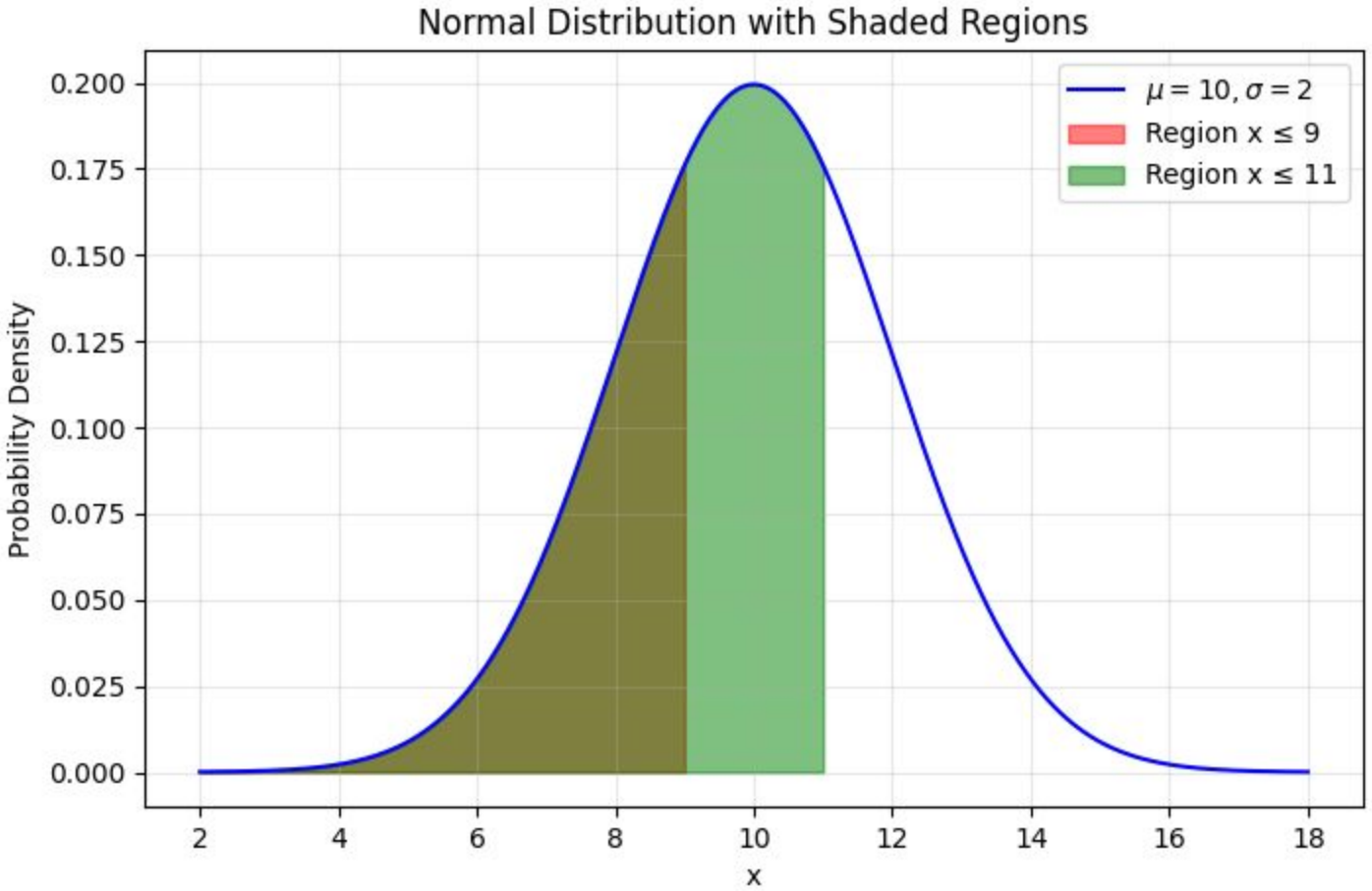
```
from scipy.stats import norm
# Given parameters
mu = 10  # mean
sigma = 2  # standard deviation
x = 13  # value to exceed
# Calculate the probability
p_exceed = 1 - norm.cdf(x, mu, sigma)
print(f"The probability that a measurement exceeds 13 milliamperes is {p_exceed:.4f}")
```


Normal Distribution with Region x > 13
$\mu = 10, \sigma = 2$
Region x > 13
Probability Density
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
2
4
6
8
10
12
14
16
18
x

Example

what is the probability that a current measurement is between 9 and 11 milliamperes?




Normal Distribution with Shaded Regions
$\mu = 10, \sigma = 2$
Region x ≤ 9
Region x ≤ 11
Probability Density
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
2
4
6
8
10
12
14
16
18
x

Solution

$$z_1 = \frac{x_1 - \mu}{\sigma} = \frac{9 - 10}{2} = -0.5$$

$$z_2 = \frac{x_2 - \mu}{\sigma} = \frac{11 - 10}{2} = 0.5$$

$$P(9 \leq x \leq 11) = P(-0.5 \leq z \leq 0.5)$$

$$P(9 \leq x \leq 11) = P(z \leq 0.5) - P(z \leq -0.5)$$

$$P(9 \leq x \leq 11) = 0.69146 - 0.30854 = 0.38292$$

Example

The compressive strength of samples of cement can be modeled by a normal distribution with a mean of 6000 kilograms per square centimeter and a standard deviation of 100 kilograms per square centimeter.

1. What is the probability that a sample's strength is less than

   6250 Kg/cm^2? Ans: 0.99378

2. What is the probability that a sample's strength is between 5800 and 5900 Kg/cm^2? Ans: 0.13591
3. What strength is exceeded by 95% of the samples?

What strength is exceeded by 95% of the samples?

To find the strength exceeded by 95% of the samples, we need to calculate the 5th percentile of the normal distribution.

This is the value x for which $P(X>x)=0.95P$

or equivalently, $P(X\leq x)=0.05$

Given Data:

- Mean (μ) = 6000 Kg/cm^2
- Standard deviation (σ) = 100 Kg/cm^2
- The cumulative probability $P(X\leq x)=0.05$

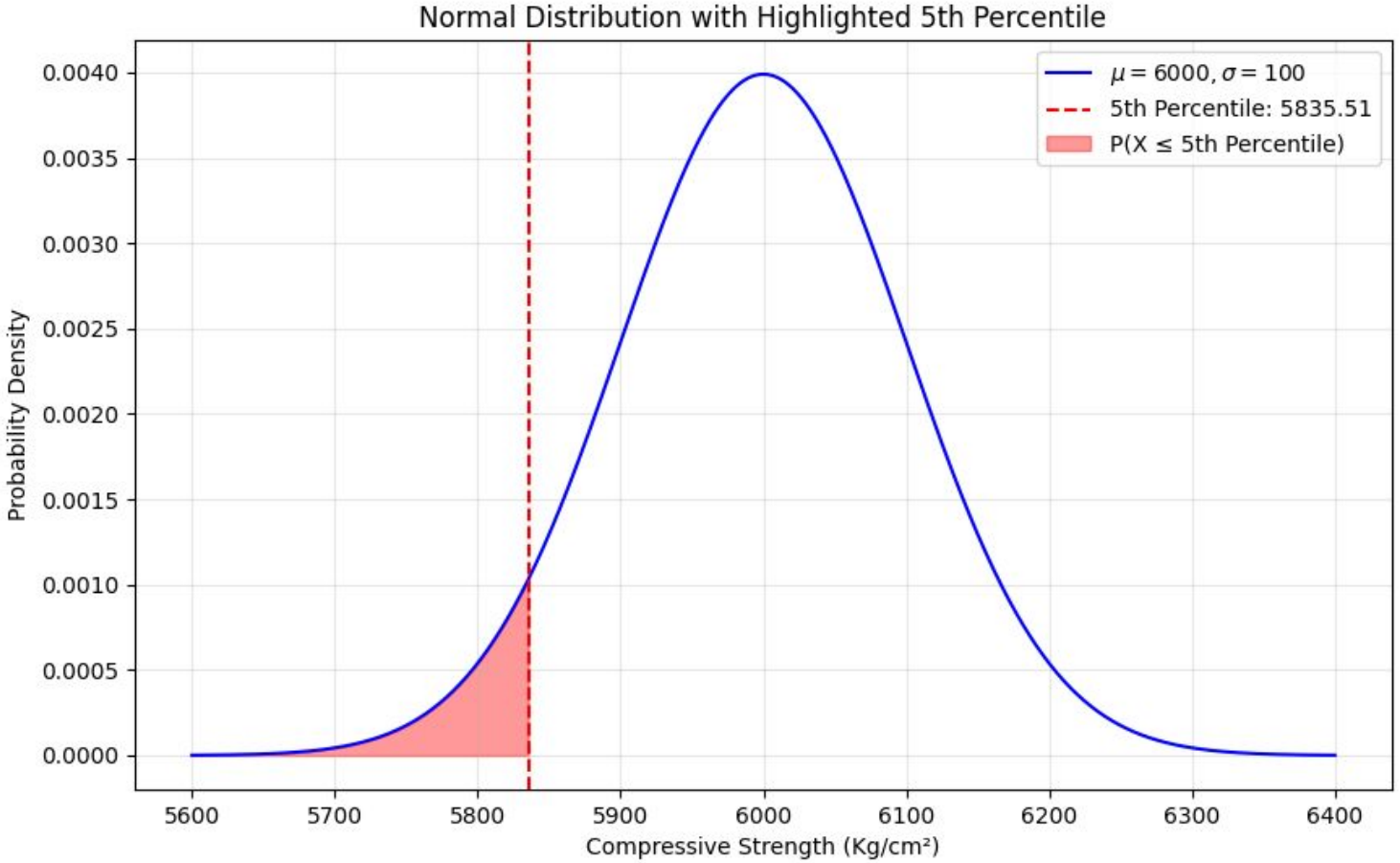
Using the inverse cumulative distribution function (also called the percentile point function, or `ppf`), we can compute x:

$x = \mu + z \cdot \sigma x$

Where z is the z-score corresponding to the cumulative probability 0.05.

From the standard normal table, the z-score corresponding to $P(Z \leq 0.05)$ is approximately -1.645

$x = 6000 + (-1.645) \cdot 100 = 6000 - 164.5 = 5835.5 Kg/cm^2$


Normal Distribution with Highlighted 5th Percentile
$\mu = 6000, \sigma = 100$
5th Percentile: 5835.51
P(X ≤ 5th Percentile)
Probability Density
0.0040
0.0035
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
5600
5700
5800
5900
6000
6100
6200
6300
6400
Compressive Strength (Kg/cm²)

Bernoulli Distribution

The **Bernoulli distribution** is a discrete probability distribution that describes the outcome of a single experiment with exactly two possible outcomes, typically referred to as **success** (1) and **failure** (0). It is named after the Swiss mathematician Jacob Bernoulli.

Key Properties

Random Variable: A Bernoulli random variable X can take one of two values:

.

$$X = \begin{cases} 1 & (\text{with probability } p) \\ 0 & (\text{with probability } 1-p) \end{cases}$$

where 0≤p≤10

Probability Mass Function (PMF): The PMF of the Bernoulli distribution is given by:

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

Key Properties

Mean: The expected value or mean of X is:

$$\mathbb{E}[X] = p$$

Variance: The variance of X is:

$$\mathrm{Var}(X) = p(1-p)$$

Email Spam Classification

Suppose you are building a spam classifier for emails. For each email, the classifier decides whether the email is **spam** (1) or **not spam** (0).

If your classifier predicts an email as spam with a probability p=0.85p = 0.85p=0.85, the distribution of this prediction is a Bernoulli distribution with p=0.85p = 0.85p=0.85. This means:

- The probability the email is classified as spam (success) is 0.850.850.85.
- The probability the email is classified as not spam (failure) is 0.150.150.15.

This binary classification problem aligns with the Bernoulli distribution because it involves a single trial with two possible outcomes.

Binomial Distribution

Binomial Distribution

The **binomial distribution** is a discrete probability distribution that describes the number of successes in a fixed number of independent trials, where each trial has exactly two possible outcomes (success or failure) and the probability of success remains constant.

Key Characteristics:

Parameters:

- n: Number of trials (a positive integer).
- p: Probability of success in a single trial ($0 \leq p \leq 1$).

Random Variable: The random variable X represents the number of successes in nnn trials.
Probability Mass Function (PMF): The probability of observing exactly k successes in n trials is:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Key Characteristics:

Mean

$$\mathbb{E}[X] = np$$

Variance

$$\mathrm{Var}(X) = np(1-p)$$

Special Case: When n=1, the binomial distribution becomes a Bernoulli distribution.

Example

Each sample of water has a 10% chance of containing a particular organic pollutant. Assume that the samples are independent with regard to the presence of the pollutant. Find the probability that in the next 18 samples, exactly 2 contain the pollutant.

Solution

Let X the number of samples that contain the pollutant in the next 18 samples analyzed. Then X is a binomial random variable with p 0.1 and n 18. Therefore,

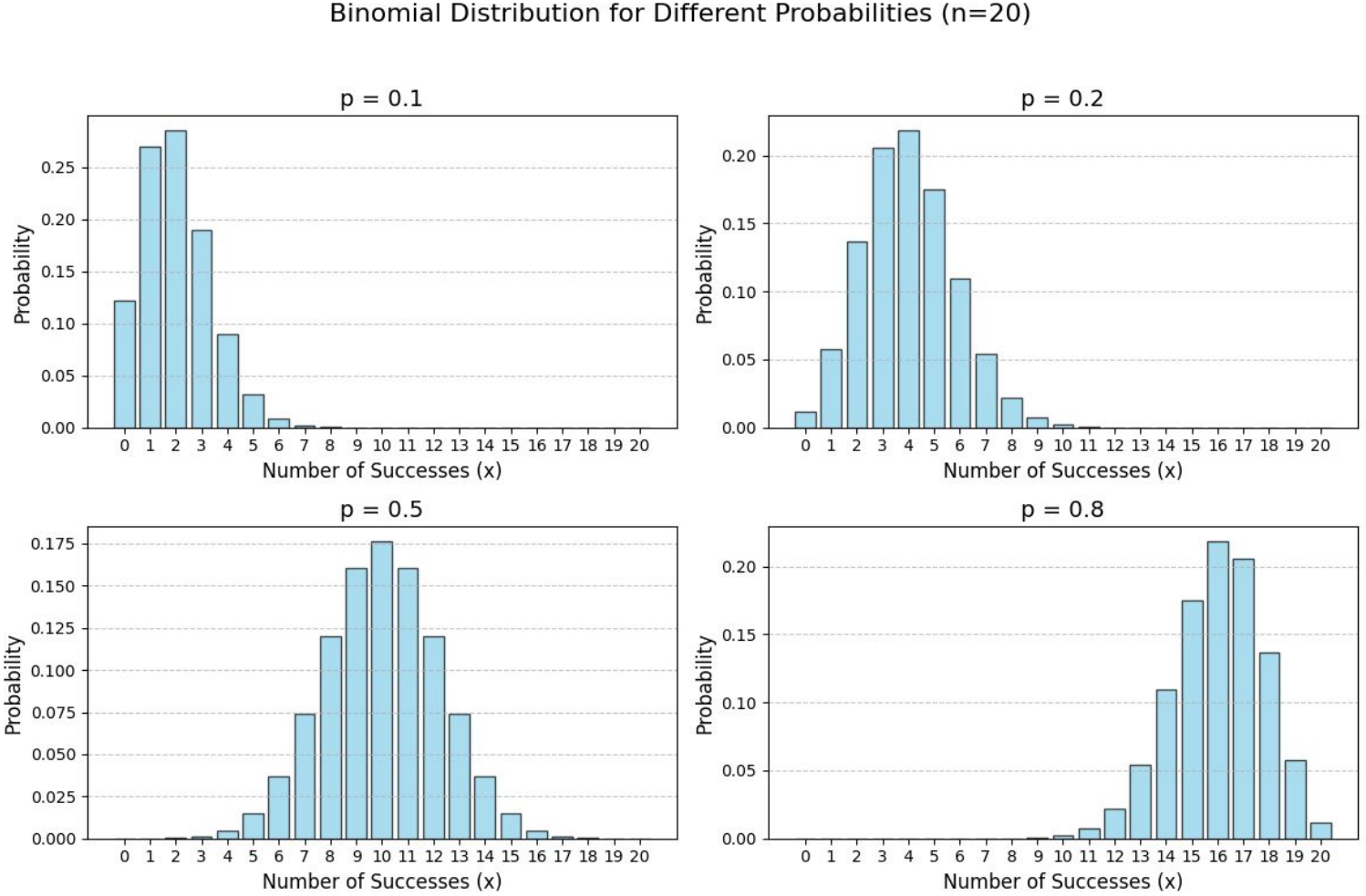
$$P(X = 2) = \binom{18}{2}(0.1)^2(0.9)^{16}$$

P(X= 2) = 0.284

Example

A particularly long traffic light on your morning commute is green 20% of the time that you approach it. Assume that each morning represents an independent trial.

1. Over five mornings, what is the probability that the light is green on exactly one day?=0.4096
2. Over 20 mornings, what is the probability that the light is green on exactly four days?=0.2182
3. Over 20 mornings, what is the probability that the light is green on more than four days? = 0.3707


Binomial Distribution for Different Probabilities (n=20)
p = 0.1
p = 0.2
p = 0.5
p = 0.8
Probability
Number of Successes (x)

Poisson Distribution

Poisson Distribution

A Poisson distribution is a discrete probability distribution. It gives the probability of an event happening a certain number of times (k) within a given interval of time or space.

The Poisson distribution has only one parameter, λ (lambda), which is the mean number of events.

PMF of Poisson Distribution

The Poisson probability of observing k events when the mean number of events is λ is given by:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

where:

- X: Number of events,
- λ: Expected (mean) number of events in the given interval,
- k: Number of occurrences (non-negative integer),
- e: Euler's number (≈2.718).

Characteristics

Parameter: $\lambda > 0$, the average rate of occurrences.

Support: $k = 0, 1, 2, \ldots$ (discrete non-negative integers).

Mean: $E[X] = \lambda$

Variance: $\mathrm{Var}(X) = \lambda$

Example

For the case of the thin copper wire, suppose that the number of flaws follows a Poisson distribution with a mean of 2.3 flaws per millimeter. Determine the probability of exactly two flaws in 1 millimeter of wire.

$$P(X = 2) = \frac{2.3^2 e^{-2.3}}{2!} = 0.265$$

Determine the probability of 10 flaws in 5 millimeters of wire. Let X denote the number of flaws in 5 millimeters of wire.

Then, X has a Poisson distribution with

E(X) = 5mm x 2.3 flaws/mm = 11.5 flaws

$$P(X = 10) = e^{-11.5} \frac{11.5^{10}}{10!} = 0.113$$

Determine the probability of at least one flaw in 2 millimeters of wire.

Let X denote the number of flaws in 2 millimeters of wire.

Then, X has a Poisson distribution with

E(X) = 2mm x 2.3 flaws/mm = 4.6 flaws

Therefore

$$P(X \geq 1) = 1 - P(X = 0) = 1 - e^{-4.6} = 0.9899$$

Example

The number of content changes to a Web site follows a Poisson distribution with a mean of 0.25 per day.

(a) What is the probability of two or more changes in a day? = 0.868

(b) What is the probability of no content changes in five days? = 0.0265

(c) What is the probability of two or fewer changes in five days? = 0.868

Descriptive and inferential statistics

Descriptive statistics summarize and organize data to make it understandable. It provides a clear overview of a dataset's main features without making predictions or generalizations.

Inferential statistics analyze data from a sample to make generalizations or predictions about a larger population.

Aspect	Descriptive Statistics	Inferential Statistics
Purpose	Summarizes data.	Make predictions or generalizations.
Focus	The data itself	The population represented by the dataset
Techniques	Mean, Median, Mode, range, graph etc	Hypothesis test, confidence interval, etc
Example	Average score of a	Predicting average

Sampling Distributions

A **sampling distribution** is the probability distribution of a statistic (such as the mean, variance, or proportion) that is calculated from a large number of samples taken from a population. It shows how the value of the statistic would vary across different samples.

Example

The mean fill volume in the population is required to be 300 milliliters. An engineer takes a random sample of 25 cans and computes the sample average fill volume to be $\bar{x}$ = 298 milliliters. The engineer will probably decide that the population mean is μ = 300 milliliters, even though the sample mean was 298 milliliters because he or she knows that the sample mean is a reasonable estimate of μ and that a sample mean of 298 milliliters is very likely to occur, even if the true population mean is μ = 300 milliliters. In fact, if the true mean is 300 milliliters, tests of 25 cans made repeatedly, perhaps every five minutes, would produce values of x that vary both above and below μ = 300 milliliters

Central Limit Theorem

The **Central Limit Theorem (CLT)** is a fundamental concept in statistics that explains how the sampling distribution of the sample mean (or sum) approaches a normal distribution as the sample size increases, regardless of the population's original distribution.

Key Points

If you take a large number of random samples from a population with any distribution (not necessarily normal) and calculate their means:

1. The distribution of those sample means will approach a normal distribution as the sample size becomes large.
2. The mean of the sample means will be approximately equal to the population mean (μ).
3. The standard deviation of the sample means will be equal to the population standard deviation (σ) divided by the square root of the sample size (n):

Conditions

Sample Size: The larger the sample size (n), the better the approximation to a normal distribution. A rule of thumb is $n \geq 30$ for the CLT to hold well.

Independence: Samples should be independent.

Random Sampling: Data must be collected randomly.

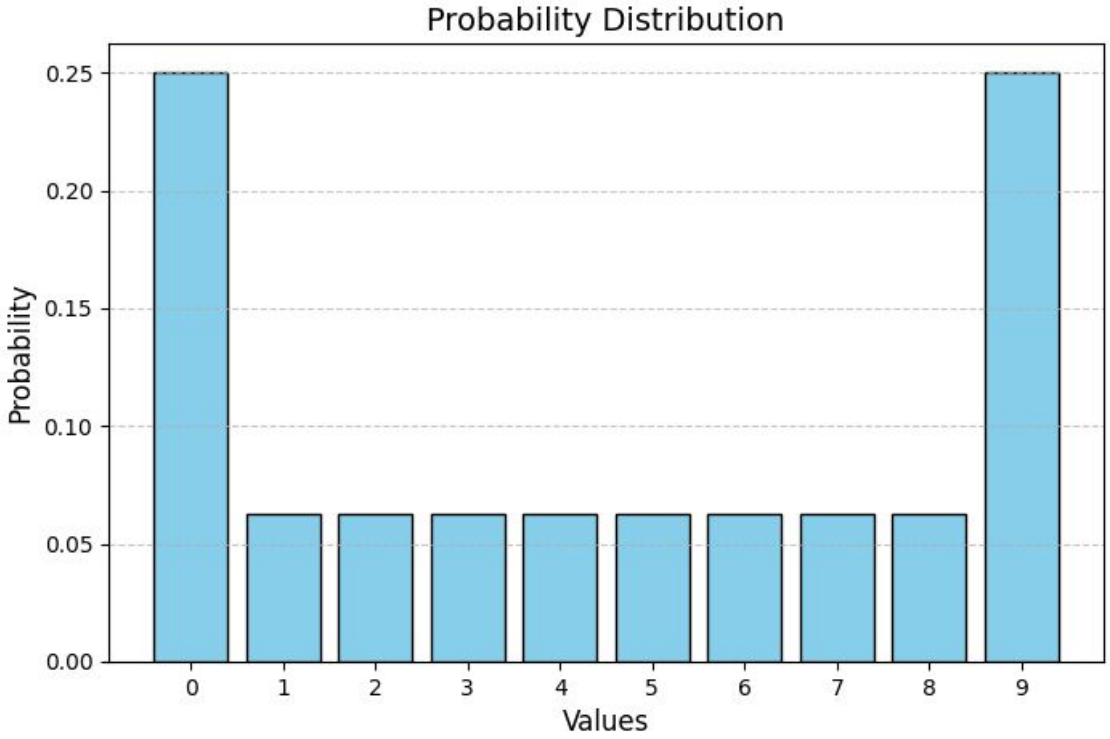
Why Is It Important?

- **Hypothesis Testing**: The CLT justifies using normal distribution-based methods for tests (e.g., z-test, t-test), even if the population is not normally distributed.
- **Confidence Intervals**: It allows for the estimation of population parameters using sample statistics.
- **Practical Applications**: In fields like finance, biology, and machine learning, it underpins many statistical and predictive models.

Example

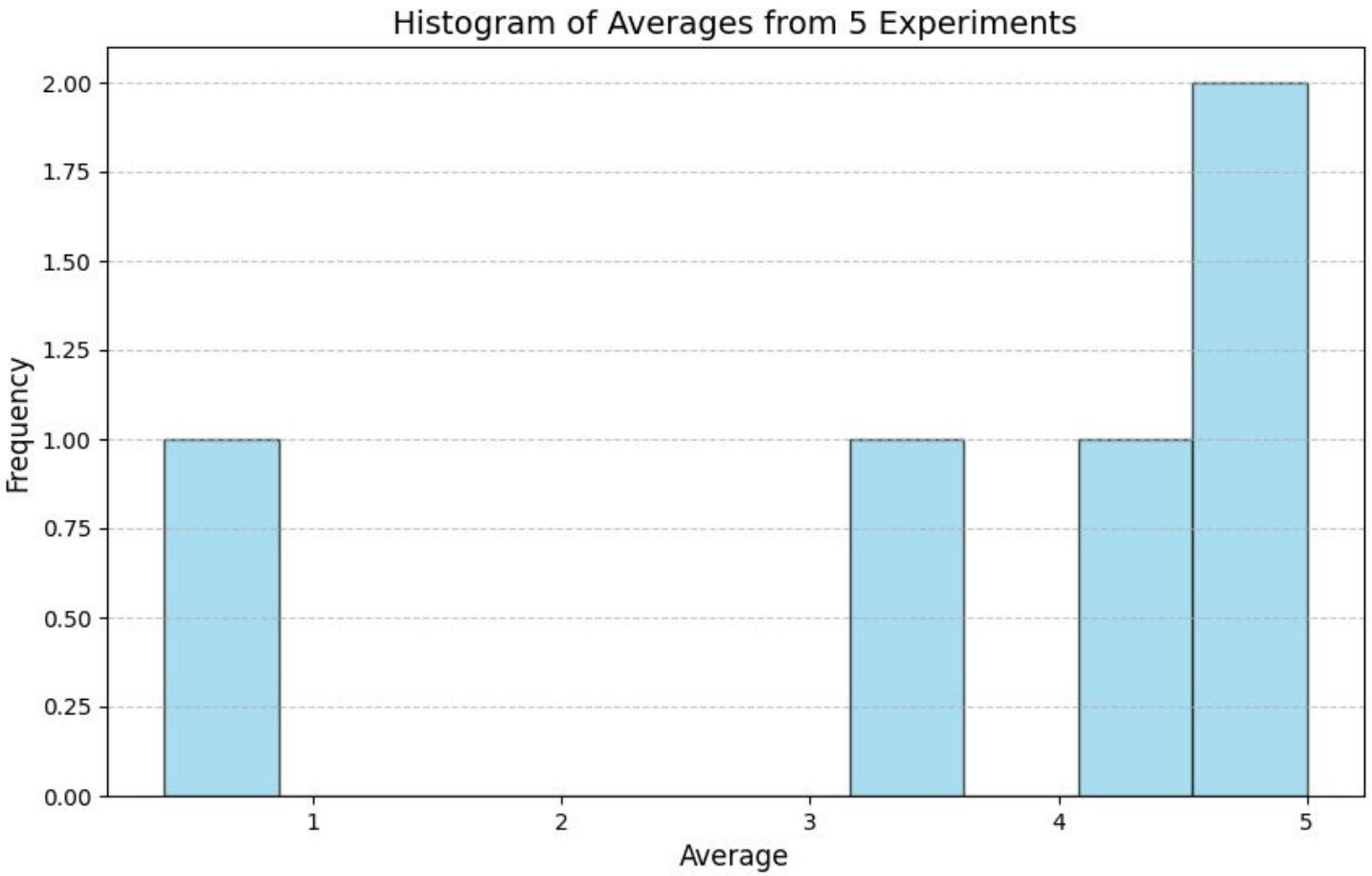
Let say we have a distribution where we can have integers from 0 to 9, the probability of 0 is 0.25 and probability of 9 is 0.25 and probability of 1 to 8 is 0.0625

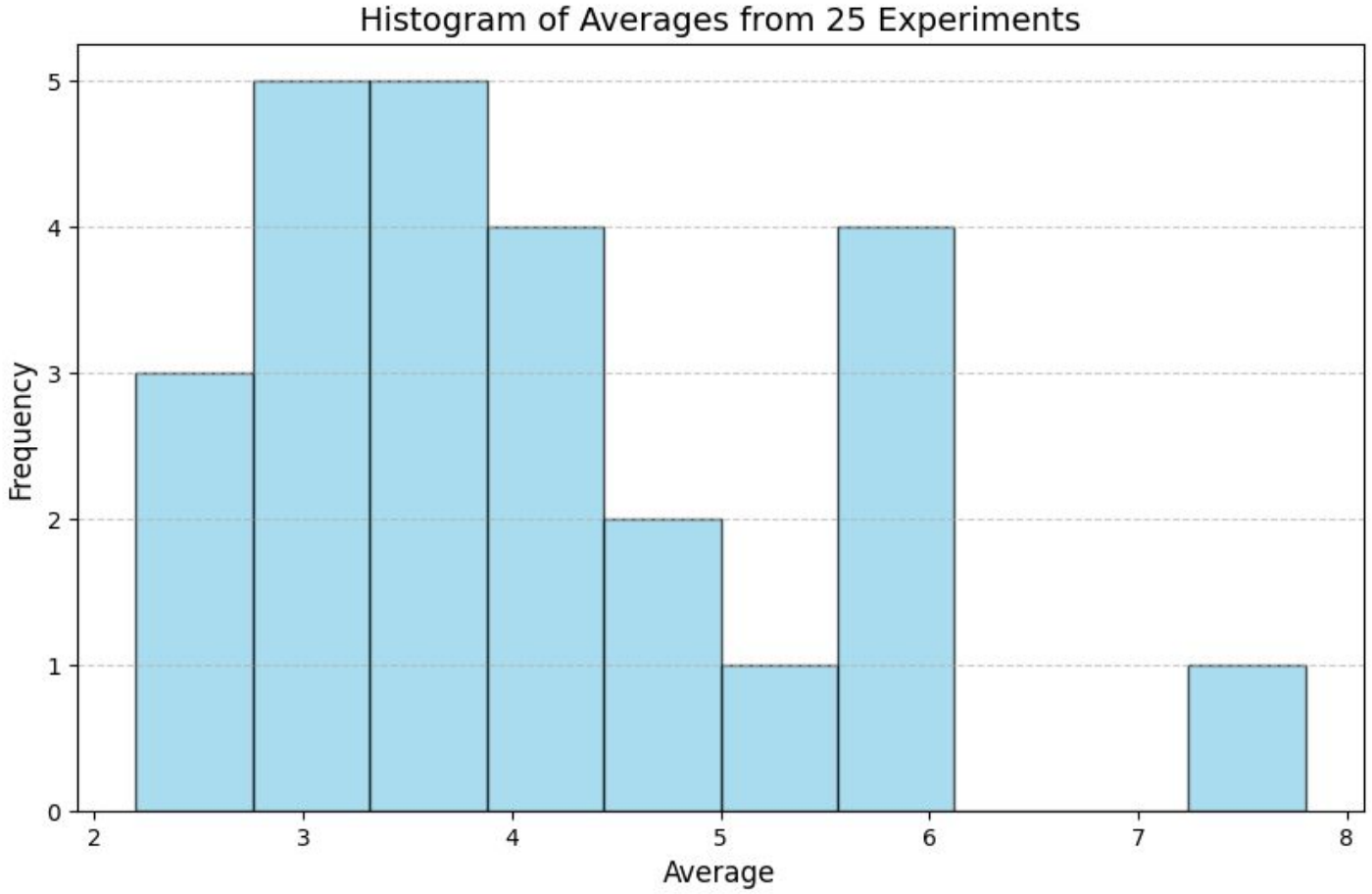
Plot

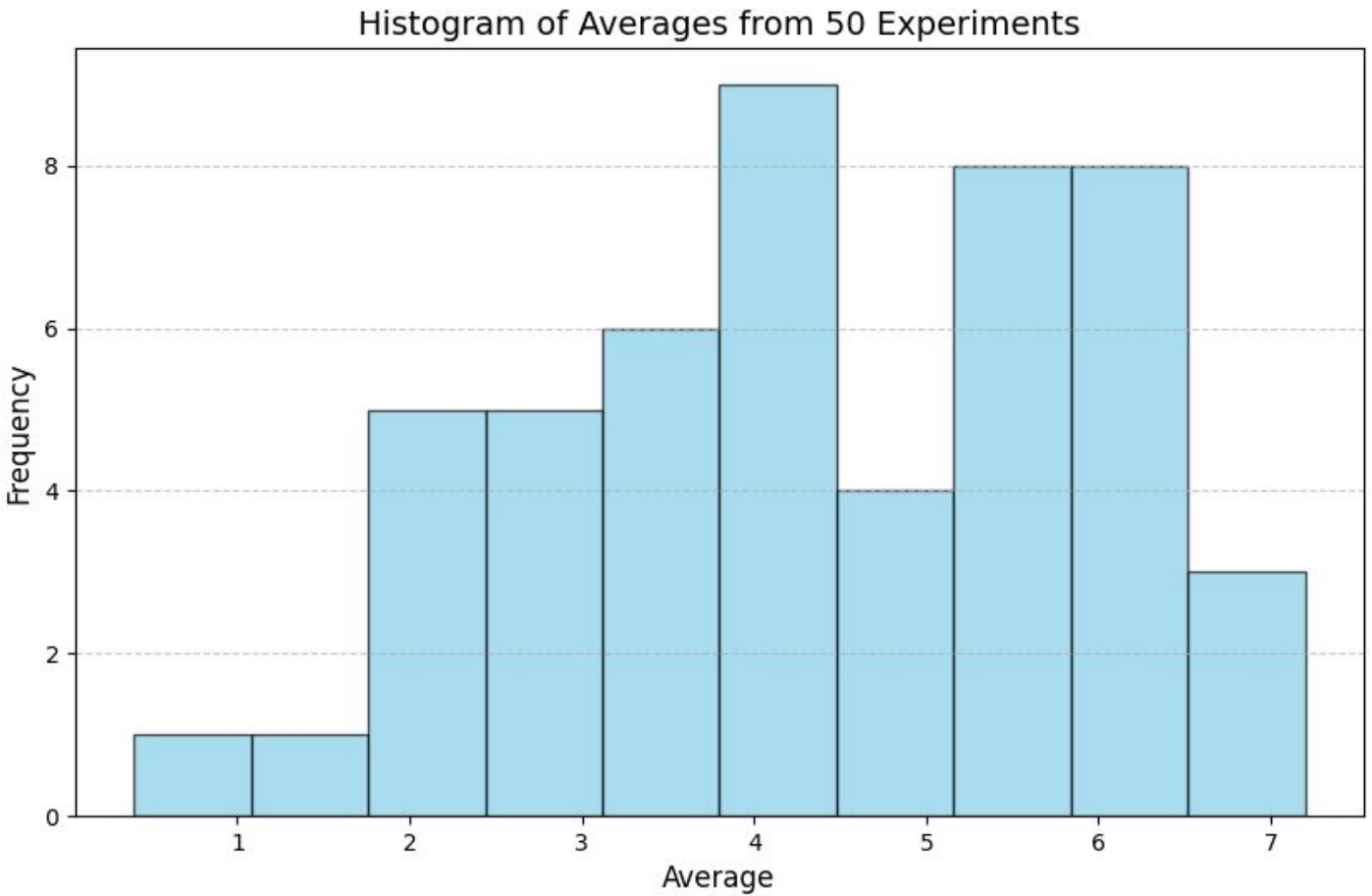

Probability Distribution
Probability
0.25
0.20
0.15
0.10
0.05
0.00
0
1
2
3
4
5
6
7
8
9
Values

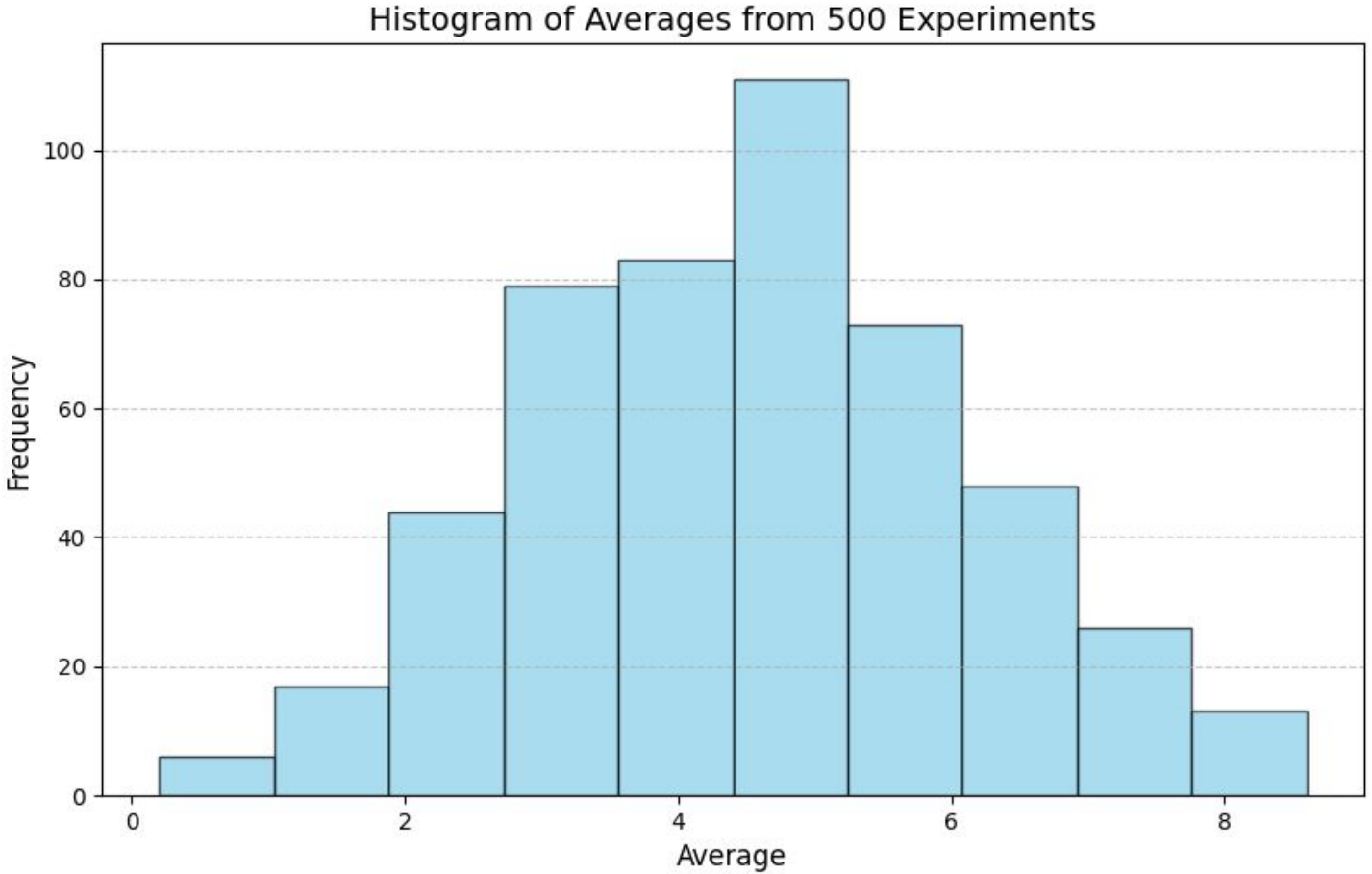
From above distribution generate 5 sample and calculate the average.

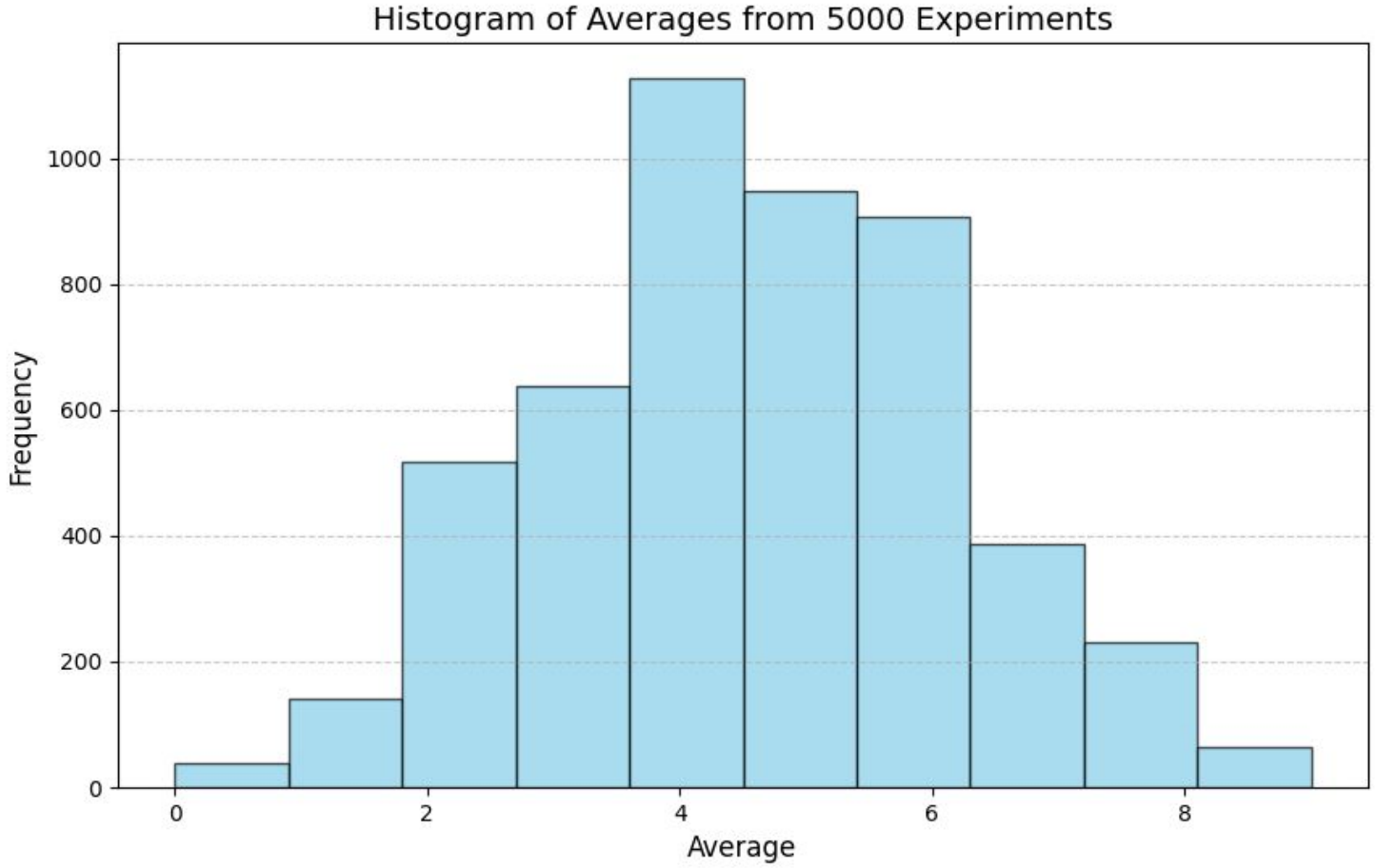
Repeat above experiment for 5 times, 25 times, 50 times, 500 times and 5000 times .


Histogram of Averages from 5 Experiments
Frequency
Average
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
1
2
3
4
5


Histogram of Averages from 25 Experiments
Frequency
Average
0
1
2
3
4
5
2
3
4
5
6
7
8


Histogram of Averages from 50 Experiments
Frequency
Average
0
2
4
6
8
1
2
3
4
5
6
7


Histogram of Averages from 500 Experiments
Frequency
Average
0
20
40
60
80
100
0
2
4
6
8


Histogram of Averages from 5000 Experiments
Frequency
1000
800
600
400
200
0
0
2
4
6
8
Average

Hypothesis Testing

What is hypothesis testing?

Hypothesis testing is a statistical method used to evaluate and make decisions about a population parameter based on sample data. It involves formulating two competing hypotheses—the **null hypothesis (H_0)**, which represents the status quo or no effect, and the **alternative hypothesis (H_a)**, which represents the claim or effect being tested.

Key Components of Hypothesis Testing

.**Null Hypothesis (H_0)**: Assumes no difference or effect (e.g., "The mean is equal to a specific value").

Alternative Hypothesis (H_a): Contradicts H_0 (e.g., "The mean is greater than or not equal to a specific value").

Test Statistic: A calculated value (e.g., z-value or t-value) used to evaluate H_0.

Significance Level (α): The threshold probability for rejecting H_0, typically set at 0.05 or 0.01.

Critical Value: The threshold value that separates the acceptance and rejection regions.

Key Components of Hypothesis Testing

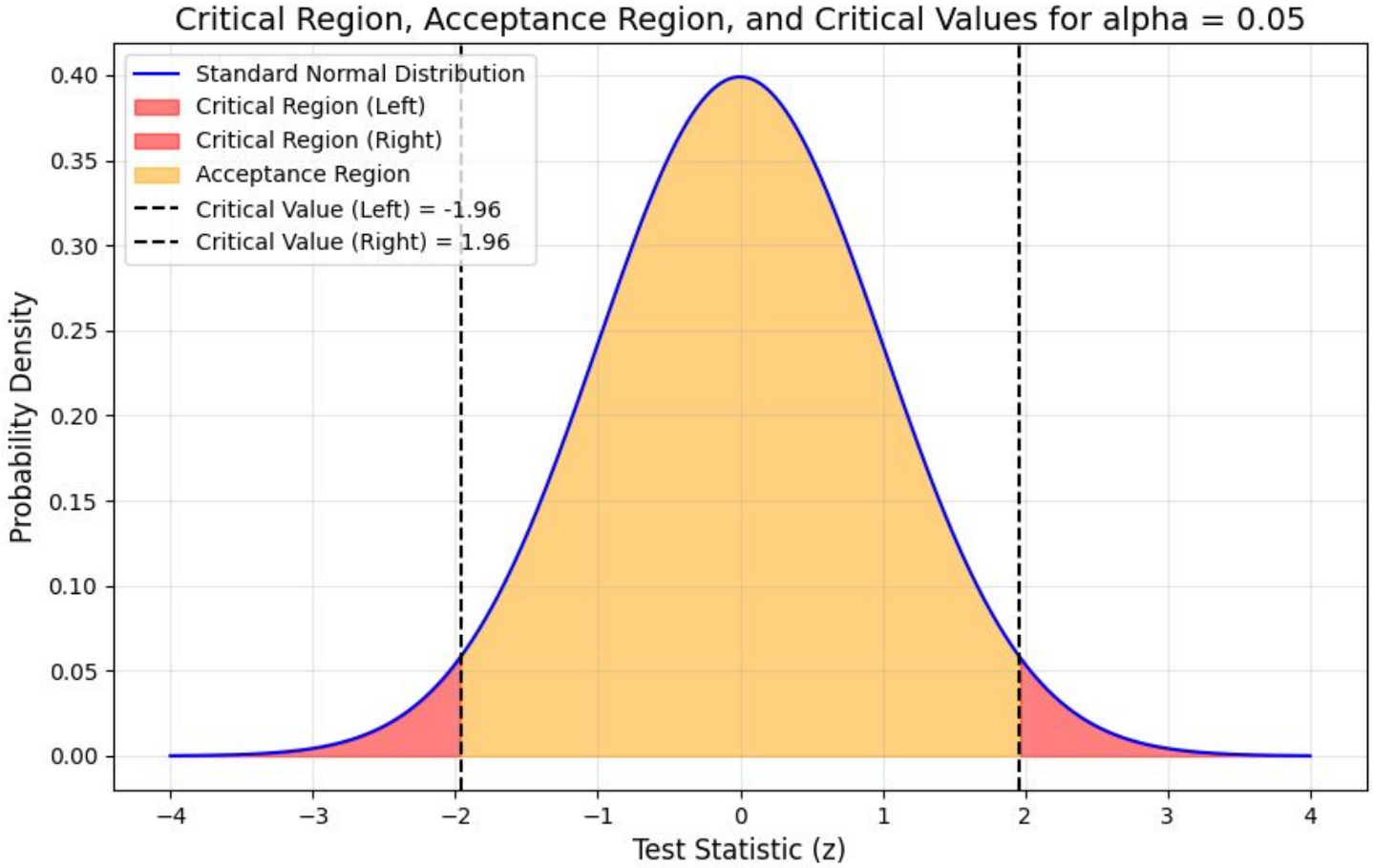
Critical Region: The range of values where H_0 is rejected.

Acceptance Region: The range of values where H_0 is not rejected.

P-Value: The probability of observing a test statistic as extreme as the sample result, given H_0 is true.

Type I Error: Rejecting H_0 when it is true.

Type II Error: Failing to reject H_0 when it is false.


Critical Region, Acceptance Region, and Critical Values for alpha = 0.05
Standard Normal Distribution
Critical Region (Left)
Critical Region (Right)
Acceptance Region
Critical Value (Left) = -1.96
Critical Value (Right) = 1.96
Probability Density
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−4
−3
−2
−1
0
1
2
3
4
Test Statistic (z)

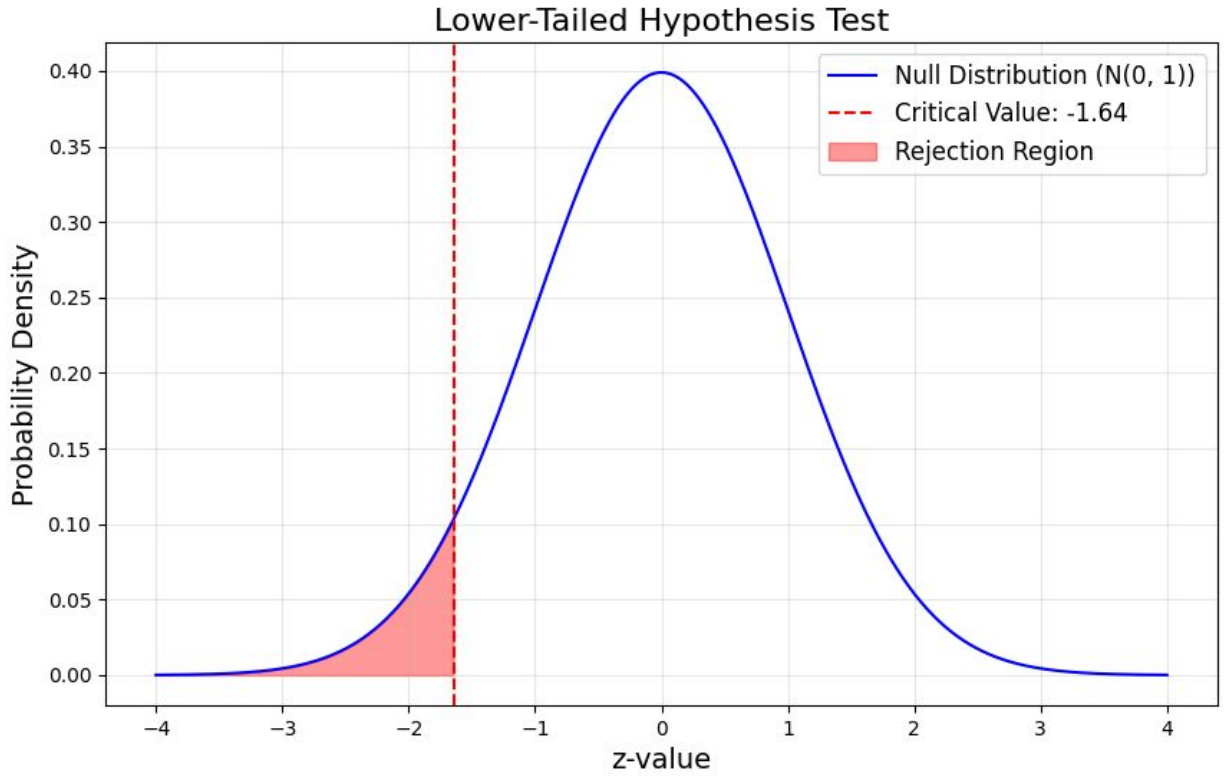
Steps in Hypothesis Testing

- Formulate H_0 and H_a
- Choose the significance level (α).
- Select the appropriate test (e.g. z-test, t-test)
- Calculate the test statistic
- Determine the critical value(s) or calculate the the p-value
- Make a decision:
  - Reject the null hypothesis if the statistic falls in the critical region or if p-value is less than alpha
  - Fail to reject the null hypothesis

Z-test

- Used when the population variance is known or the sample size is large (n >=30)
- Formula for test statistic
- Types of Z-test
  - Left-tailed test: (Ha: $\mu<\mu_0$): Critical region is $z < -z_\alpha$
  - Right-tailed test: (Ha: $\mu>\mu_0$): Critical region is $z > z_\alpha$
  - Two-tailed test: (Ha: $\mu \neq \mu_0$): Critical region is $z < -z_\alpha$ or $z>z_\alpha$
- Single Sample Z-test: Test mean of one sample against a known population mean.
- Two-Sample Z-test: Compare means of independent samples

Left Tailed Test



Example

A factory claims that the average weight of its product is $\mu_0=50$ grams. A quality control analyst suspects the average weight is less than 50 grams. To test this claim, she randomly selects a sample of $n=49$ products and finds the sample mean ($\bar{X}$) weight content to be 49 grams, with a population standard deviation $\sigma=8$ grams.

At a 1% significance level ($\alpha=0.01$), test whether the weight of product is less than 50 grams.

Steps to solve

1. State the null and alternative hypotheses
   a. Ho : μ≥ 50
   b. Ha : μ <50
2. Calculate the test statistic

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

$Z_{calculated}$= -1.75

3. Determine the Critical value

   For one tailed test at $\alpha = 0.01$ the critical value from the z-table is

   $Z_{critical} = -2.33$

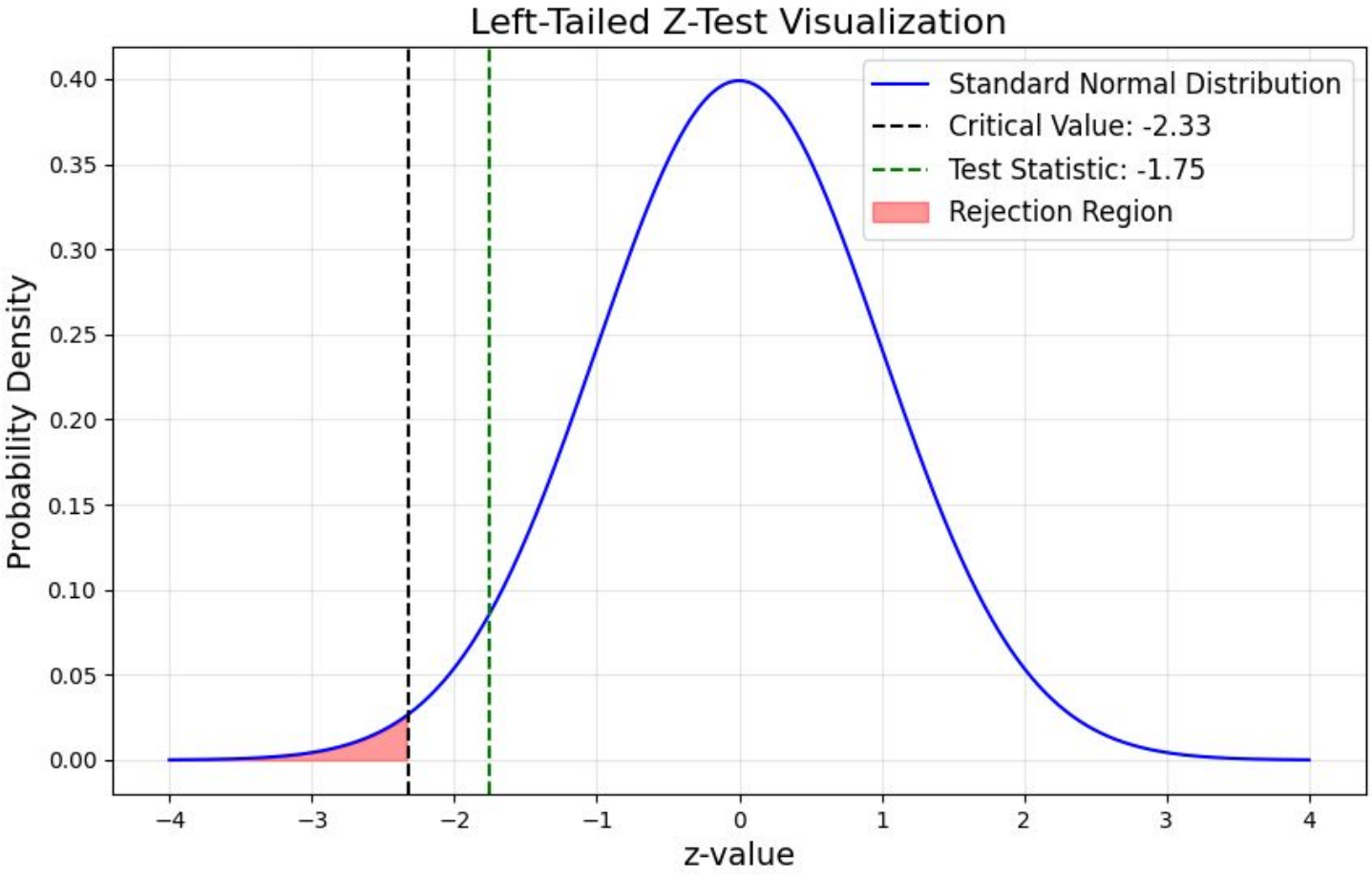
4. Compare the test static to the critical value

   $Z_{calculated} > Z_{critical}$

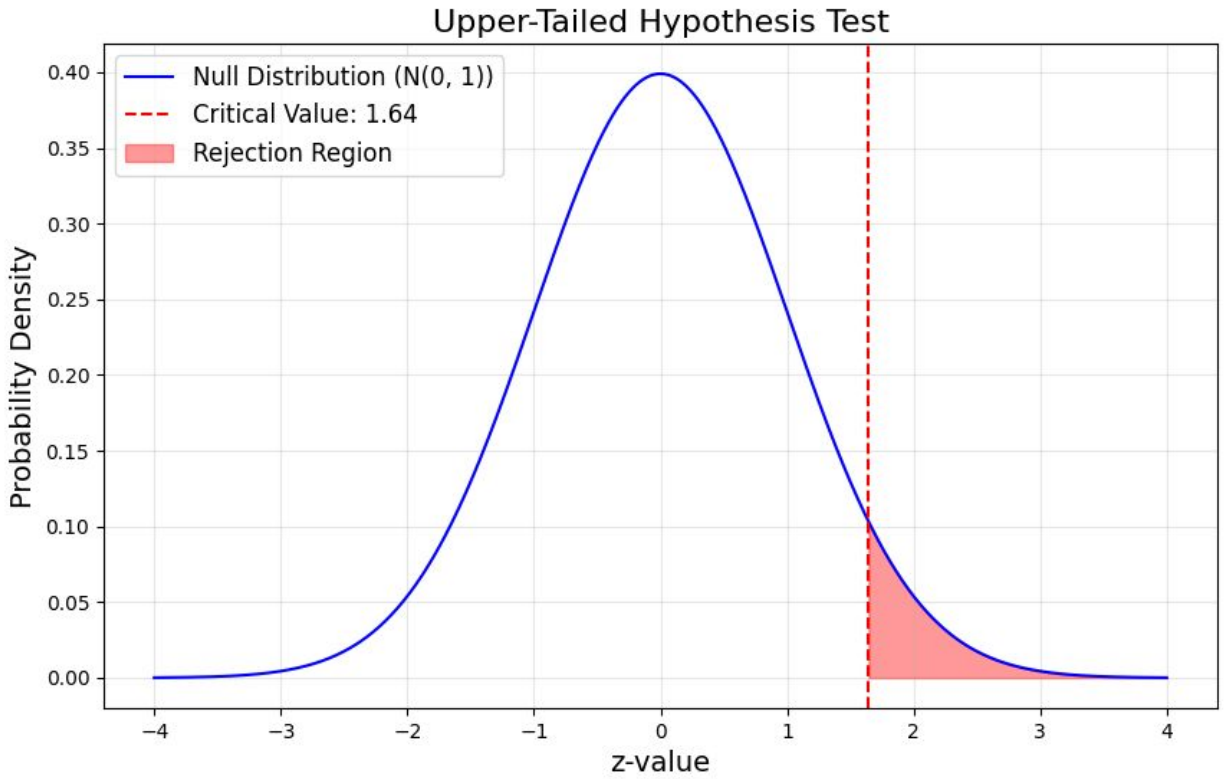
   $-1.75 > -2.33$

Conclusion

At 1% significance level, there is no sufficient evidence to reject the null hypothesis.


Left-Tailed Z-Test Visualization
Standard Normal Distribution
Critical Value: -2.33
Test Statistic: -1.75
Rejection Region
Probability Density
z-value
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−4
−3
−2
−1
0
1
2
3
4

Right Tailed Test



Example:

A cigarette manufacturer claims that the average nicotine content μ of brand B cigarettes is at most 1.5mg. To test this claim, a health organization randomly selects a sample of n=36 cigarettes from this brand and finds the sample mean ($\overline{X}$) nicotine content to be 1.6 mg, with a population standard deviation σ=0.3 mg.

At a 5% significance level (α=0.05), test whether the nicotine content of brand B cigarettes is greater than 1.5 mg.

Steps to solve

1. State the null and alternative hypotheses
   a. Ho : $\mu \leq 1.5$
   b. Ha : $\mu > 1.5$
2. Calculate the test statistic

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

$Z_{calculated} = 2.0$

Steps to solve

3. Determine the Critical value

   For one tailed test at $\alpha = 0.05$ the critical value from the z-table is

   $Z_{critical} = 1.645$

4. Compare the test static to the critical value

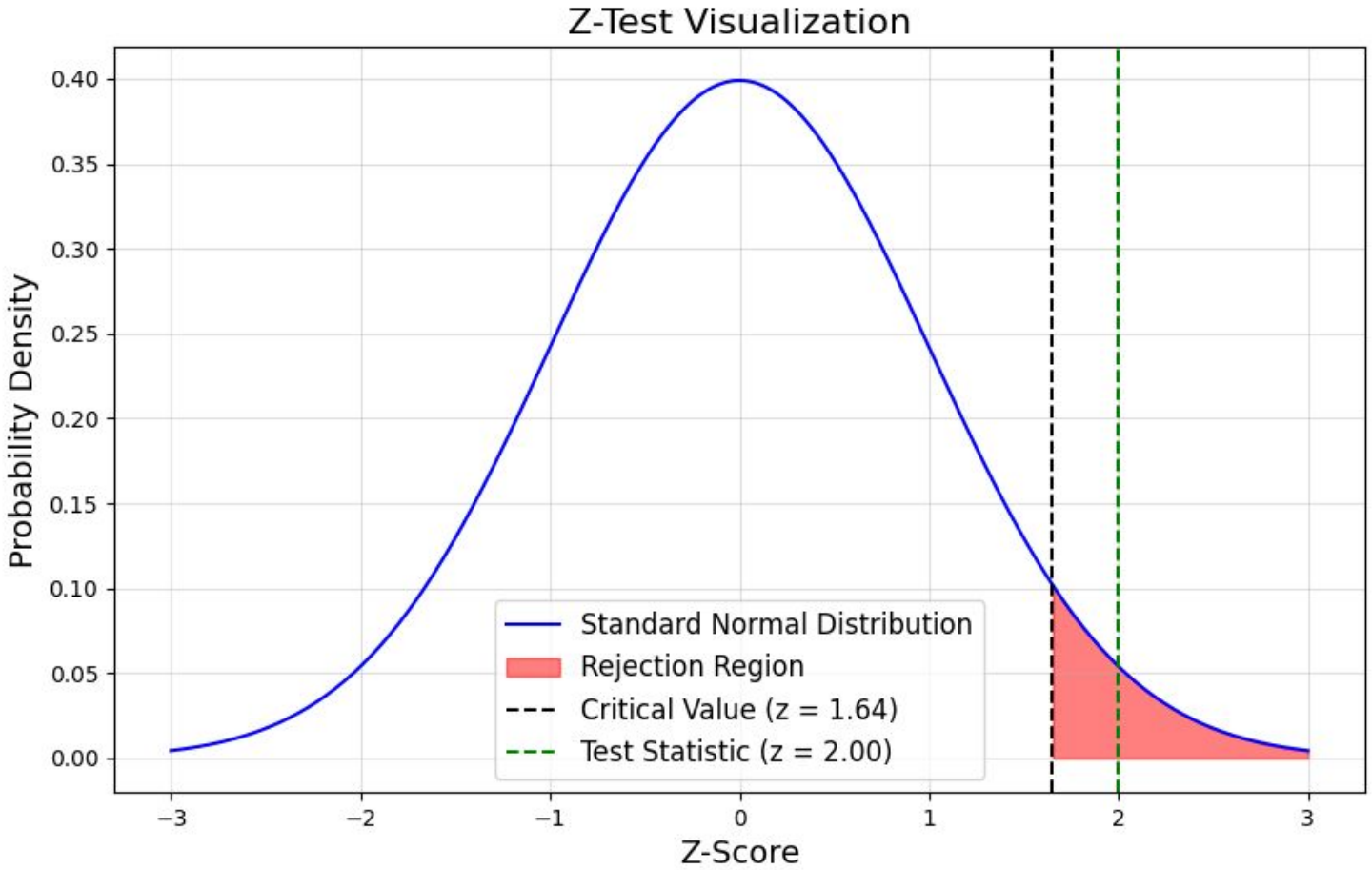
   $Z_{calculated} > Z_{critical}$

   $2.0 > 1.645$

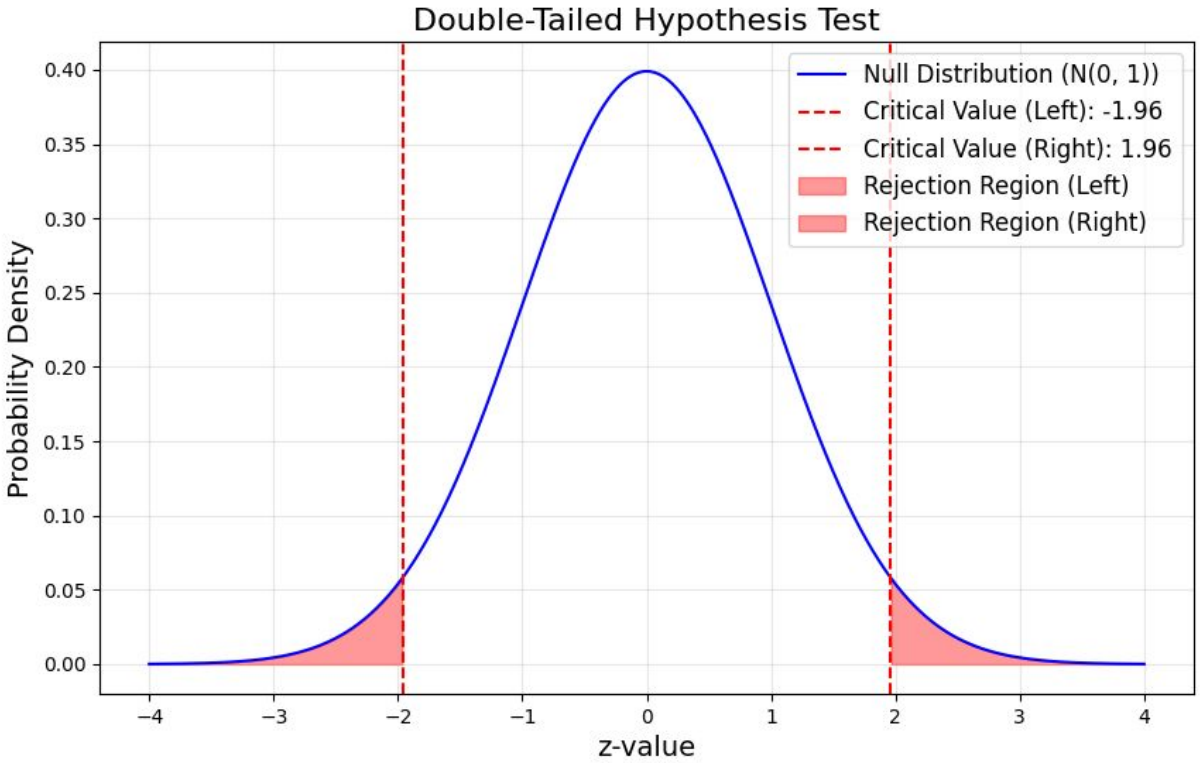
Conclusion

At 5% significance level, there is sufficient evidence to conclude that the average evidence to conclude that the average nicotine content of brand B cigarettes is greater than 1.5 mg.

Now try changing the the value of n to 55, 85, $\bar{X}$ to 1.55 or 1.52 and value of σ to 0.4 or 0.2


Z-Test Visualization
Probability Density
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−3
−2
−1
0
1
2
3
Z-Score
Standard Normal Distribution
Rejection Region
Critical Value (z = 1.64)
Test Statistic (z = 2.00)

Two Tailed Test



Example

A pharmaceutical company claims that a drug reduces blood pressure by an average (μ) of 10 mmHg. A researcher believes the actual average reduction may be different from 10 mmHg. To test this claim, a random sample of 40 patients was selected, and the average reduction in blood pressure was found to be 9.2 mmHg. The population standard deviation of the reduction in blood pressure is known to be 2.5 mmHg. Conduct a two-tailed z-test at the $\alpha=0.05$ significance level to determine if the average reduction in blood pressure is significantly different from 10 mmHg.

Steps to solve

1. State the null and alternative hypotheses
   a. Ho : μ= 10
   b. Ha : μ ≠ 10
2. Calculate the test statistic

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

$Z_{calculated}$= -2.02

3. Determine the Critical value

For a two-tailed test with α=0.05, divide α by 2 (α/2=0.025):

$Z_{critical} = \pm 1.96$

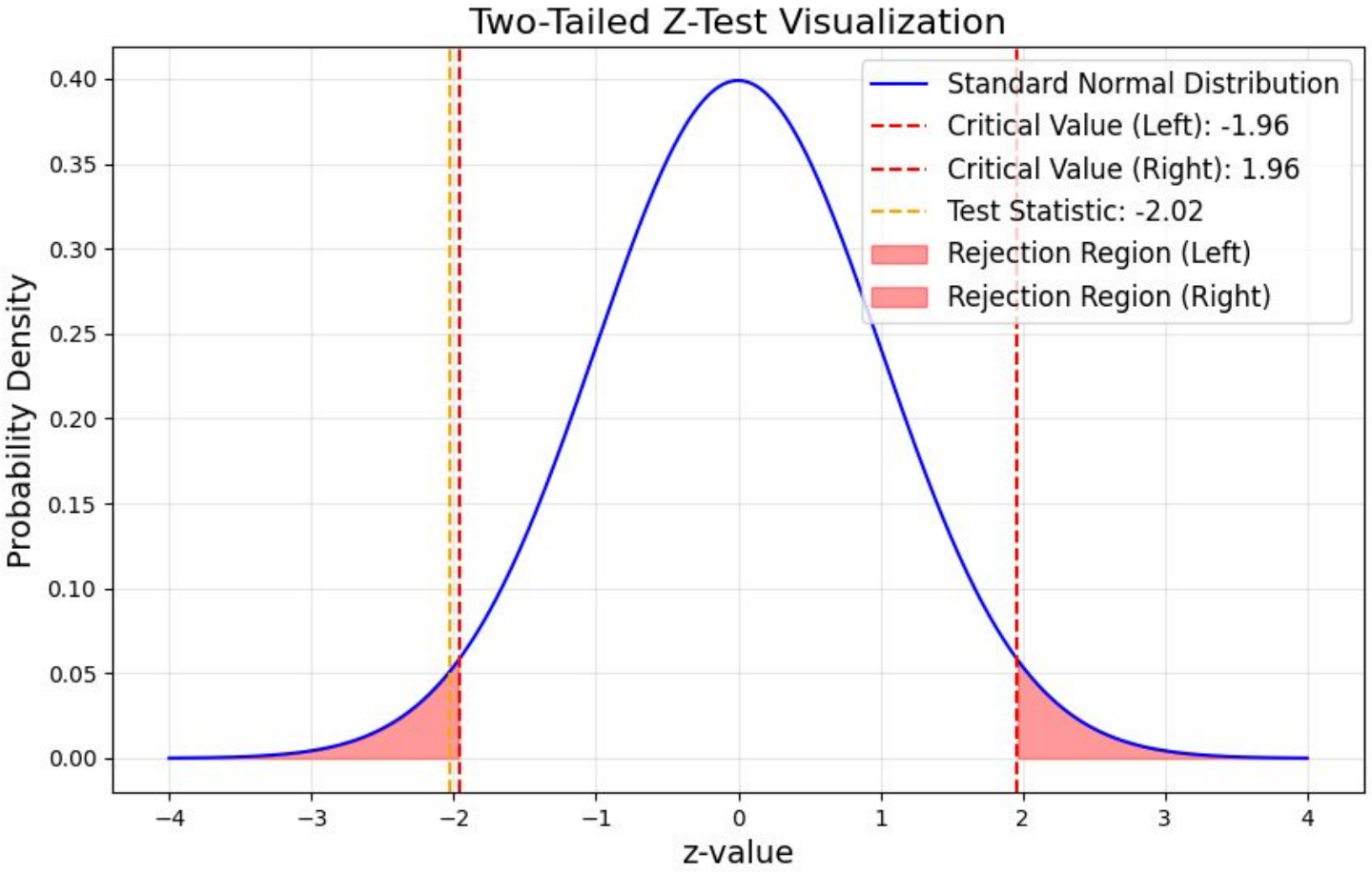
4. Compare the test static to the critical value

$Z_{calculated} < Z_{critical}$

$-2.02 < -1.96$

Conclusion

Since, Zcalculated -2.02 does not lie in the region of -1.96 to +1.96. We reject the null hypothesis.


Two-Tailed Z-Test Visualization
Standard Normal Distribution
Critical Value (Left): -1.96
Critical Value (Right): 1.96
Test Statistic: -2.02
Rejection Region (Left)
Rejection Region (Right)
Probability Density
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−4
−3
−2
−1
0
1
2
3
4
z-value

Z-test for differences between two populations means

A company wants to compare the average daily sales between two stores (Store A and Store B). They collected the following data:

- Store A:
  - Sample size (n1) = 40
  - Sample mean ($\bar{x}1$) = $200
  - Sample standard deviation (s1) = $20
- Store B:
  - Sample size (n2) = 35
  - Sample mean ($\bar{x}2$) = $190
  - Sample standard deviation (s2) = $25

Steps to solve

1. State the null and alternative hypotheses
   a. Ho : μ1=μ2 (No difference)
   b. Ha : μ1 ≠ μ2 (There is a difference)
2. Calculate the test statistic

$$z = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$Z_{calculated}$ = 1.89

Steps to solve

3. Determine the Critical value

   For a two-tailed test with α=0.05, divide α by 2 (α/2=0.025):

   $Z_{critical} = \pm 1.96$

4. Compare the test static to the critical value

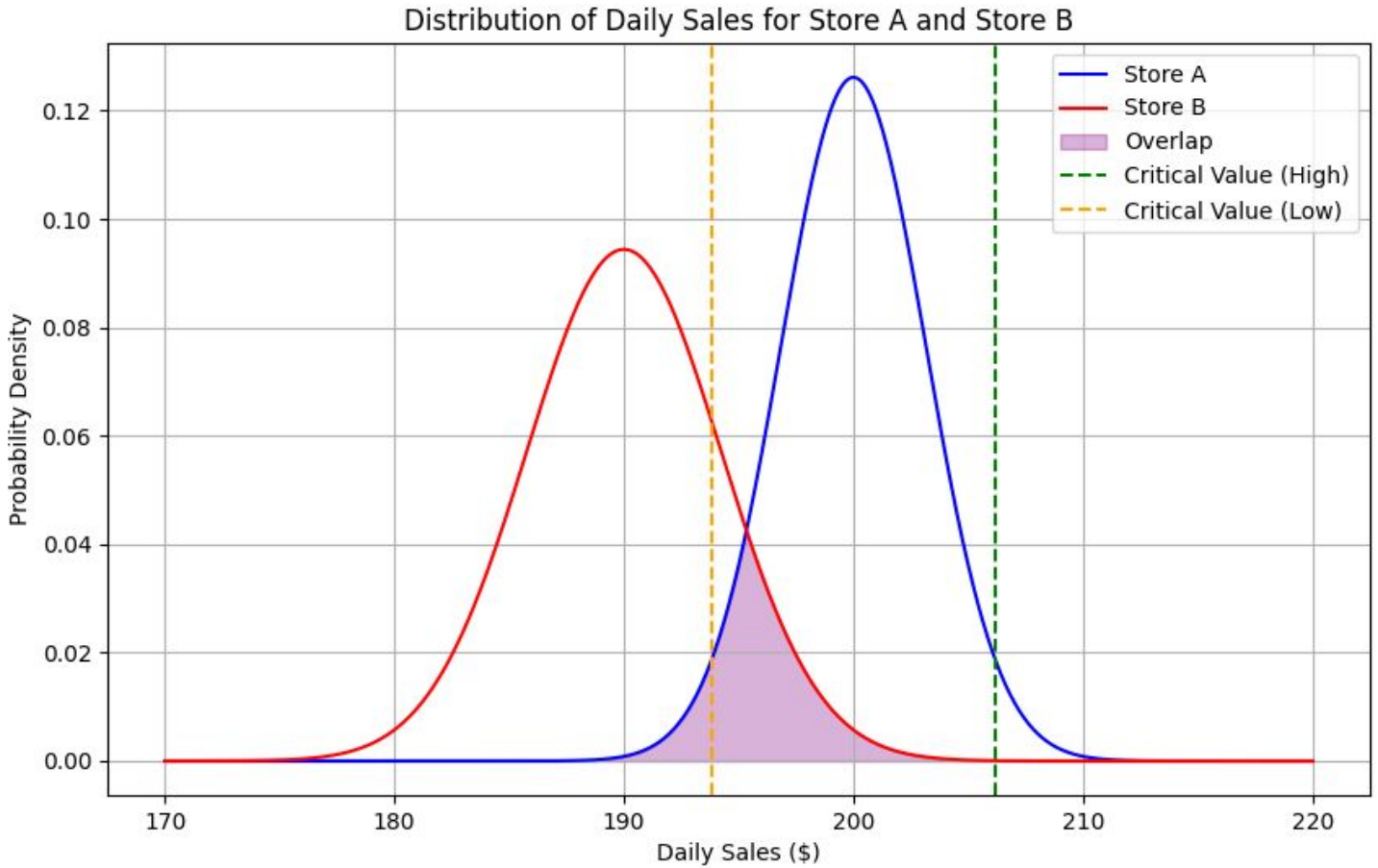
   $Z_{calculated} < Z_{critical}$

   $1.89 < 1.96$

Conclusion

We fail to reject the null hypothesis.

There is not enough evidence to suggest a significant difference in average daily sales between Store A and Store B at the 5% significance level.


Distribution of Daily Sales for Store A and Store B
Store A
Store B
Overlap
Critical Value (High)
Critical Value (Low)
Probability Density
0.00
0.02
0.04
0.06
0.08
0.10
0.12
170
180
190
200
210
220
Daily Sales ($)

T-test hypothesis test

Used when the population variance is unknown, and the sample size is small ($n<30$), though it can also be applied to larger samples when variance is unknown.

Example left tailed T-test

A company claims that its new training program improves employee productivity. Historically, the average time taken to complete a task was **50 minutes**. After the training, a random sample of **20 employees** completed the task, with an average time of **48 minutes** and a sample standard deviation of **4 minutes**. Test whether the training program has reduced the average task time at a **5% significance level ($\alpha=0.05$)**.

Solution

Step 1: Define Hypotheses

- Null Hypothesis (H0): $\mu \geq 50$ (The training program does not reduce the task time.)
- Alternative Hypothesis (Ha): $\mu < 50$ (The training program reduces the task time.)

Step 2: Determine the Test Statistic

Since the population standard deviation is unknown, we use the t-test formula:

$t_{calulated}$ = -2.24

$$t_{calculated} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Step 3: Find the Critical Value

Degrees of freedom (df) = $n-1=20-1=19$

For a one-tailed t-test at $\alpha=0.05$ the critical value ($t\alpha$) from the t-table is approximately **-1.729**.

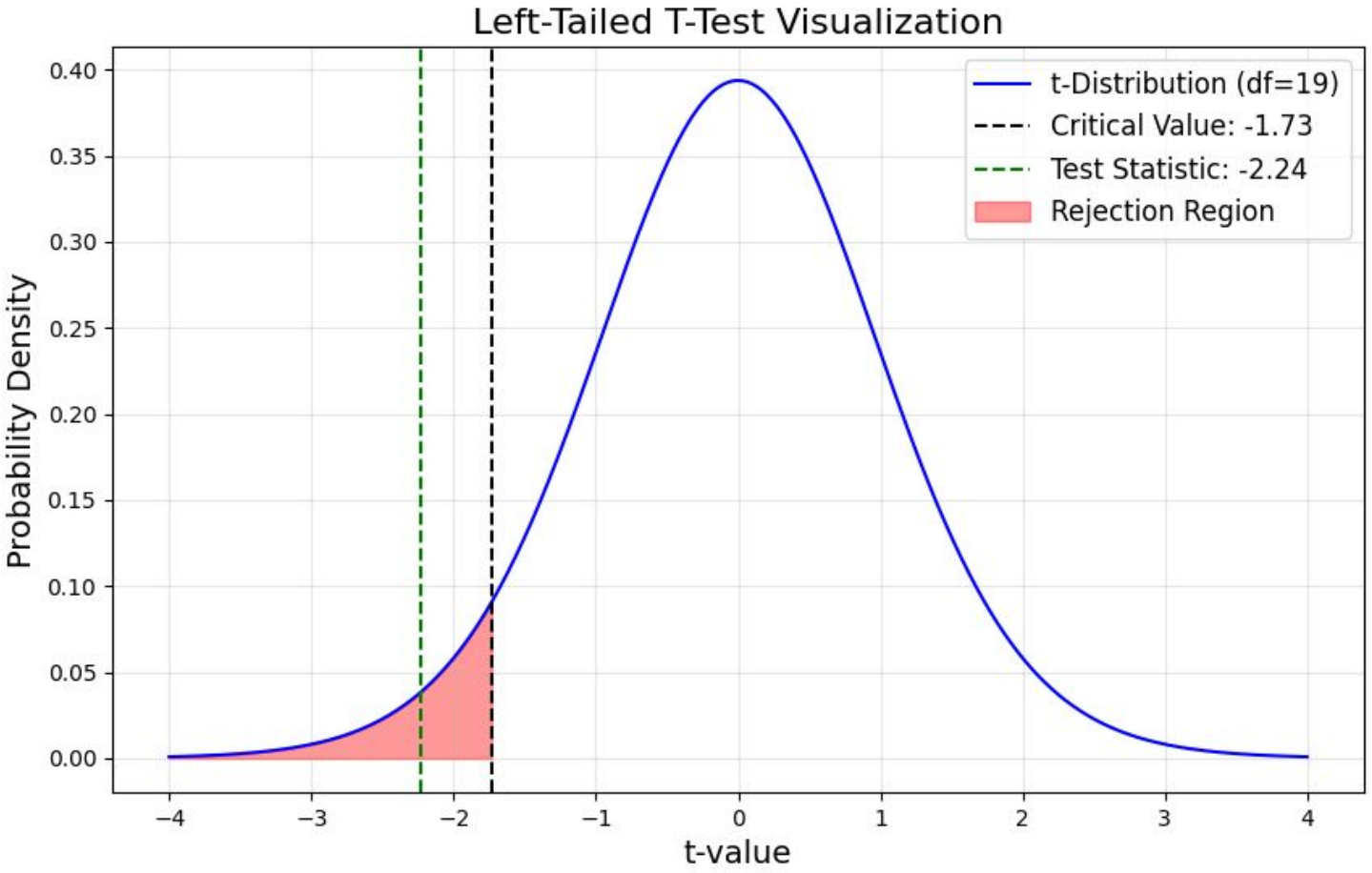
Step 4: Decision Rule

If t is less than $-t\alpha$, reject H0

Conclusion

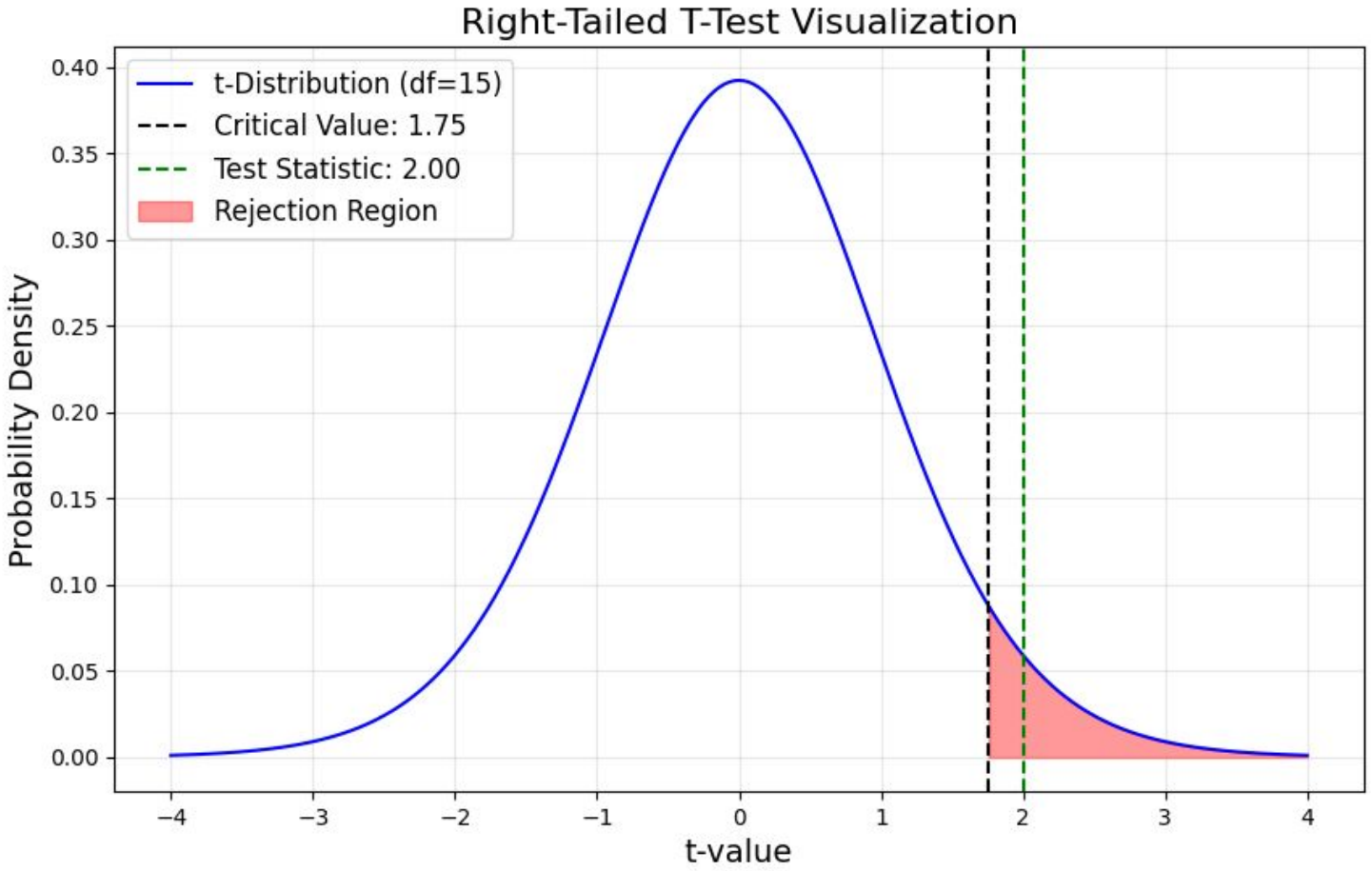
The calculated t-value is **-2.24**, which is less than the critical value of **-1.729**.

Thus, we **reject the null hypothesis (H0)** and conclude that the training program significantly reduces the average task time.


Left-Tailed T-Test Visualization
t-Distribution (df=19)
Critical Value: -1.73
Test Statistic: -2.24
Rejection Region
Probability Density
t-value
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−4
−3
−2
−1
0
1
2
3
4

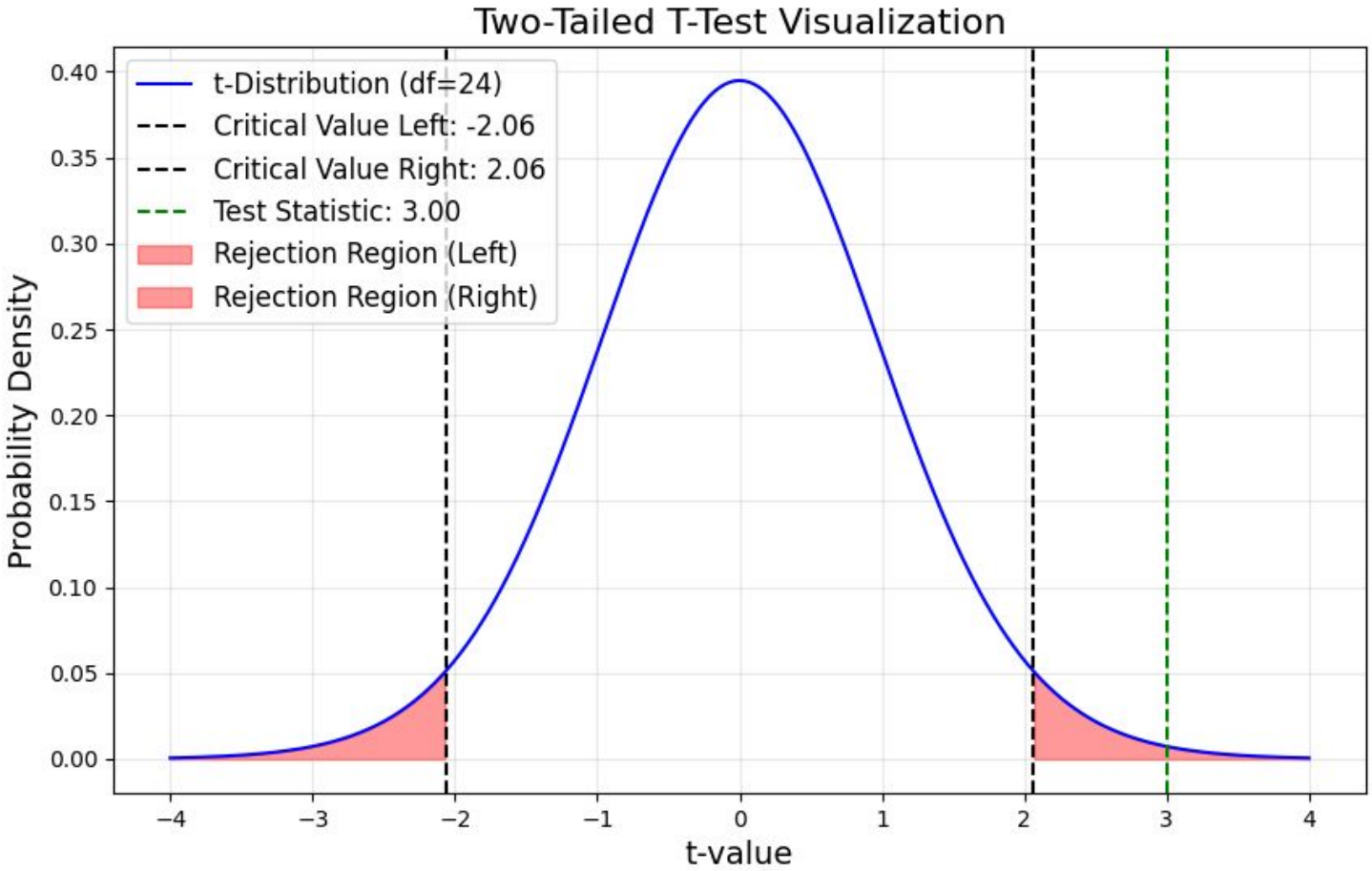
Example right tailed T-test

A factory claims that the average strength of its product is **200 units**. A competitor suspects the actual strength is higher. A random sample of **16 products** has a mean strength of **205 units**, with a sample standard deviation of **10 units**. Test the claim at a **5% significance level (α=0.05)**.


Right-Tailed T-Test Visualization
t-Distribution (df=15)
Critical Value: 1.75
Test Statistic: 2.00
Rejection Region
Probability Density
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−4
−3
−2
−1
0
1
2
3
4
t-value

Example of Two-Tailed T-Test

A school claims that the average score of its students in mathematics is **75**. A researcher wants to verify this claim. A random sample of **25 students** has a mean score of **78**, with a sample standard deviation of **5**. Test this claim at a **5% significance level (α=0.05)**.


Two-Tailed T-Test Visualization
t-Distribution (df=24)
Critical Value Left: -2.06
Critical Value Right: 2.06
Test Statistic: 3.00
Rejection Region (Left)
Rejection Region (Right)
Probability Density
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−4
−3
−2
−1
0
1
2
3
4
t-value

Two-Sample T-Test (Independent Samples)

A researcher wants to compare the average test scores of students from two different teaching methods.

- Group 1: $n_1=30, \bar{x}_1=85, s_1=5$
- Group 2: $n_2=35, \bar{x}_2=80, s_2=6$

Test at a **5% significance level ($\alpha=0.05$)** whether there is a significant difference in the means of the two groups.

Solution

Step 1: Define Hypotheses

- Null Hypothesis (H0): μ1=μ2 (The means are equal.)
- Alternative Hypothesis (Ha): μ1≠μ2 (The means are not equal.)

Step 2: Determine the Test Statistics

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Step 3: Find the Critical Values

Degrees of freedom (df) = $n1+n2-2=30+35-2=63$

For a two-tailed t-test at $\alpha=0.05$, the critical values from the t-table are approximately ± 2.000

Step 4: Decision Rule

If t is outside the range $[-t\alpha/2, +t\alpha/2]$, reject H0.

Conclusion

The calculated t-value is **3.45**, which is greater than the critical value 2.0.

Thus, we **reject the null hypothesis (H0)** and conclude that there is a significant difference in the means of the two teaching methods.

ANOVA

Overview

Analysis of Variance (ANOVA) is a statistical method used to compare the means of three or more groups to determine if there are statistically significant differences between them. Unlike a t-test, which compares only two groups, ANOVA can handle multiple groups simultaneously. It is commonly used in experiments and research to test hypotheses.

Types of ANOVA

1. **One-Way ANOVA**: Compares means across multiple groups based on a single independent factor.
2. **Two-Way ANOVA**: Compares means across groups based on two independent factors and can evaluate interaction effects.
3. **Repeated Measures ANOVA**: Used when the same subjects are measured under different conditions.

Assumptions

1. **Independence**: Observations are independent of one another.
2. **Normality**: Data in each group should be approximately normally distributed.
3. **Homogeneity of Variance**: The variance within each group is approximately equal.

Example

Three randomly selected groups of chickens are fed on three different diets. Each group consists of five chickens. Their weight gains during a specified period of time are as follows.

Diet I	4	4	7	7	8
Diet II	3	4	5	6	7
Diet II	6	7	7	7	8

Test the hypothesis that mean gains of weights due to the three diets are equal.

Level of Significance (α): 5% = 0.05

Solution

Step 1: Set up hypotheses

Null Hypothesis:

H0: $\mu_1=\mu_2=\mu_3$ (There is no significant difference in mean weight gains due to different diets.)

Alternative Hypothesis:

H1: $\mu_1\neq\mu_2\neq\mu_3$ (There is significant difference in mean weight gains due to different diets.)

Step 2

$$F = \frac{MST_r}{MSE} = \frac{\frac{SST_r}{k-1}}{\frac{SSE}{N-k}}$$

Where:

k = number of samples (treatments)

N = Total number of observations

MSTr = Mean Square of Treatments (This represent variance between group means)

MSE = Mean Square Error (This represent the variance within the Groups)

SSTr = Sum of Square for Treatments

SSE = Sum of squares for error

Given:

n1 = n2 = n3 =5, N = n1+n2+n3 = 15

$\bar{y}_1$ =6, $\bar{y}_2$= 5, $\bar{y}_3$ = 7, k = 3

Diet y1	**$(y1-\bar{y}1)^2$**	**Diet y2**	**$(y2-\bar{y}2)^2$**	**Diet y3**	**$(y3-\bar{y}3)^2$**
4	4	3	4	6	1
4	4	4	1	7	0
7	1	5	0	7	0
7	1	6	1	7	0
8	4	7	4	8	1
$\sum y1 = 30$	**$\sum(y1-\bar{y}1)^2 =14$**	**$\sum y2 = 25$**	**$\sum(y2-\bar{y}2)^2 =10$**	**$\sum y3 = 35$**	**$\sum(y3-\bar{y}3)^2 =2$**

$$\bar{y} = \frac{T}{N} = \frac{\sum_{i=1}^{k} n_i \bar{y}_i}{N}$$

$$SST_r = \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2 = n_1(\bar{y}_1 - \bar{y})^2 + n_2(\bar{y}_2 - \bar{y})^2 + n_3(\bar{y}_3 - \bar{y})^2$$

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

$$\bar{y} = \frac{n_1 \bar{y}_1 + n_2 \bar{y}_2 + n_3 \bar{y}_3}{15} = \frac{30 + 25 + 35}{15} = 6$$

SSTr = $5x(6-6)^2 + 5x(5-6)^2 + 5x(7-6)^2 = 10$

SSE = 14+10+2 = 26

Source of Variation	Degree of Freedom	Sum of Square	Mean Square	F-ratio
Between Samples	K-1	SSTr	MSTr	$F = \frac{MSTr}{MSE}$
Within Samples	N-k	SSE	MSE	
Total	N-1	SST		

Source of Variation	Degree of Freedom	Sum of Square	Mean Square	F-ratio
Between Samples	2	10	5	2.304
Within Samples	12	26	2.17	
Total	14	36		

Step 3

Critical value of F at for level of Significance (α): 5% = 0.05 and degree of freedom (2,12) is Fcritical = 3.89

Step 4

Since

F<Fcritical i.e 2.304<3.89

So, H0 is accepted and we conclude that there is no significant difference in mean weight gains due to different diets.