

# Range

## What is Range?

The **range** of a dataset is the **simplest measure of spread (variability)**.

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

---

## Example with Dataset

Dataset:

[12, 18, 25, 26, 30, 34, 40, 45, 50]

- **Minimum (min)** = 12
- **Maximum (max)** = 50

$$\text{Range} = 50 - 12 = 38$$

---

## How to Interpret Range

↳ A **small range** → all values are close together.

- A **large range** → data is more spread out.
- ⚠ **Range depends only on 2 values (min & max)**, so it is very sensitive to **outliers**.

**Example:**

If one extra score = **100** is added →

New Range =  $100 - 12 = 88$  This makes the spread look huge, even though most values are still between 12 and 50.

---

---

## Variance Formula (Simplified)

$$\text{Variance} = \frac{1}{N} \sum (x_i - \mu)^2$$

- $x_i$  : each data value
- $\mu$ : mean
- N : number of data points

It's the **average of the squared differences** from the mean.

---

# What is Standard Deviation?

**Standard Deviation (SD)** is the **square root of variance**.

It tells you **how much the data varies from the mean** in the **original units** (like marks, rupees, etc.).

$$\text{Standard Deviation} = \sqrt{\text{Variance}}$$

Example: score={10,20,30,40,50}

## ◆ Step 1: Calculate the Mean ( $\mu$ )

$$\text{Mean} = \frac{10 + 20 + 30 + 40 + 50}{5} = \frac{150}{5} = 30$$

So, the average score is **30**.

## ◆ Step 2: Find Deviations from the Mean

Score (X)	Deviation (X - $\mu$ )	Squared Deviation ((X - $\mu$ ) <sup>2</sup> )
10	10 - 30 = -20	(-20) <sup>2</sup> = 400
20	20 - 30 = -10	(-10) <sup>2</sup> = 100
30	30 - 30 = 0	0 <sup>2</sup> = 0
40	40 - 30 = 10	10 <sup>2</sup> = 100
50	50 - 30 = 20	20 <sup>2</sup> = 400

## ◆ Step 3: Calculate Variance ( $\sigma^2$ )

$$\text{Variance} = \frac{\sum (X - \mu)^2}{N} = \frac{400 + 100 + 0 + 100 + 400}{5} = \frac{1000}{5} = 200$$

✦ Variance tells us how **spread out** the data is (in squared units).

## ◆ Step 4: Calculate Standard Deviation ( $\sigma$ )

$$\text{Standard Deviation} = \sqrt{\text{Variance}} = \sqrt{200} \approx 14.14$$

✦ This means, on average, each score is about **14.14 points away** from the mean.

This way you'd see that:

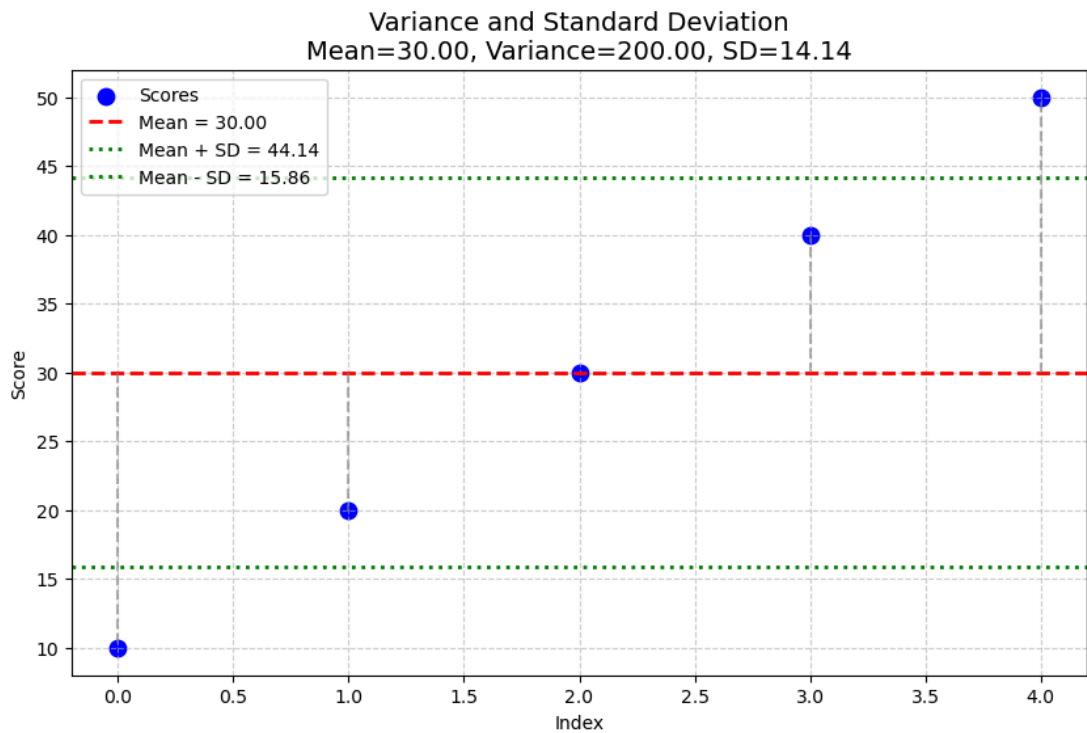
The middle value (30) sits on the mean.

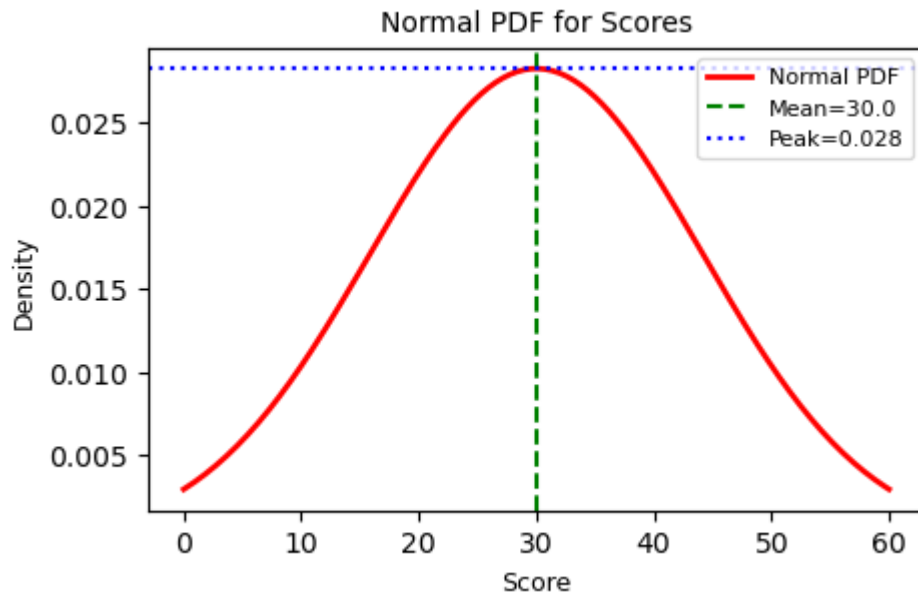
Values (10 and 50) are 20 away from the mean (biggest deviations).

Values (20 and 40) are 10 away from the mean.

these deviations give you the variance = 200 and SD  $\approx 14.14$ .

- **Mean** is the average score: 30
- **Variance** shows the average of the squared differences from the mean: 200
- **Standard Deviation** is the square root of variance:  $\approx 14.14$
- A **low standard deviation** means scores are tightly clustered around the mean.
- A **high standard deviation** means the scores are more spread out.





## Percentile

- A percentile is a number that tells us what percentage of the data lies below that value.
- It helps us understand where a particular data point stands in the dataset.

## Step-by-Step Quartile Calculation

Dataset:[12, 18, 25, 26, 30, 34, 40, 45, 50]

Number of elements, **n = 9**

---

$$position = (n - 1) * q$$

Where:

- **n** is the number of elements
  - **q** is the desired quantile (e.g., 0.25, 0.5, 0.75)
- 

## Step-by-Step Example

Q1 (25th percentile):

- $Position = (9 - 1) * 0.25 = 2.0 \rightarrow index2$
  - Value at index 2 = **25.0**
-

## Q2 (50th percentile / Median):

- $Position = (9 - 1) * 0.50 = 4.0 \rightarrow index4$
  - Value at index 4 = **30.0**
- 

## Q3 (75th percentile):

- $Position = (9 - 1) * 0.75 = 6.0 \rightarrow index6$
- Value at index 6 = **40.0**

## IQR (Interquartile Range)

$$IQR = Q3 - Q1 = 40.0 - 25.0 = 15.0$$

$$LowerBound = Q1 - 1.5 \times IQR = 25.0 - 1.5 \times 15.0 = 25.0 - 22.5 = ** 2.5 **$$

$$UpperBound = Q3 + 1.5 \times IQR = 40.0 + 1.5 \times 15.0 = 40.0 + 22.5 = ** 62.5 **$$

Metric	Value
Q1	25.0
Q2 (Median)	30.0
Q3	40.0
IQR	15.0
Lower Bound	2.5
Upper Bound	62.5

Any data point outside **[2.5, 62.5]** would be considered an **outlier**. The dataset is: [12, 18, 25, 26, 30, 34, 40, 45, 50] (N = 9).

Q1 = 25.0, Q2 (Median) = 30.0, Q3 = 40.0

This means that 25% of the values are less than or equal to 25.

50% of the values are less than or equal to 30 (so 30 is the middle of the dataset).

75% of the values are less than or equal to 40.

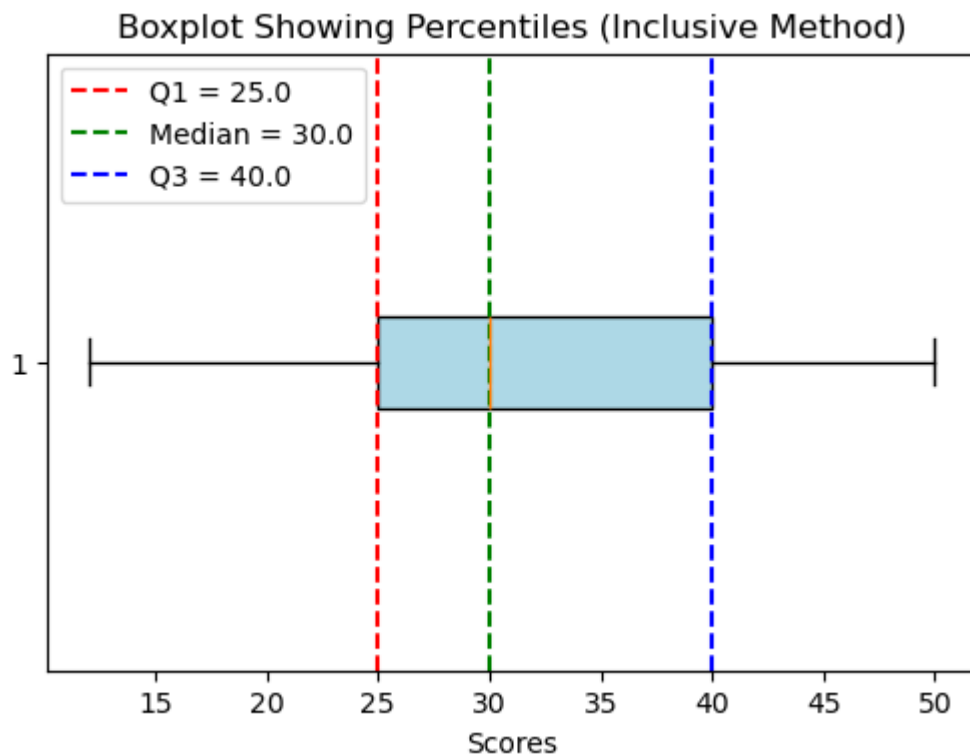
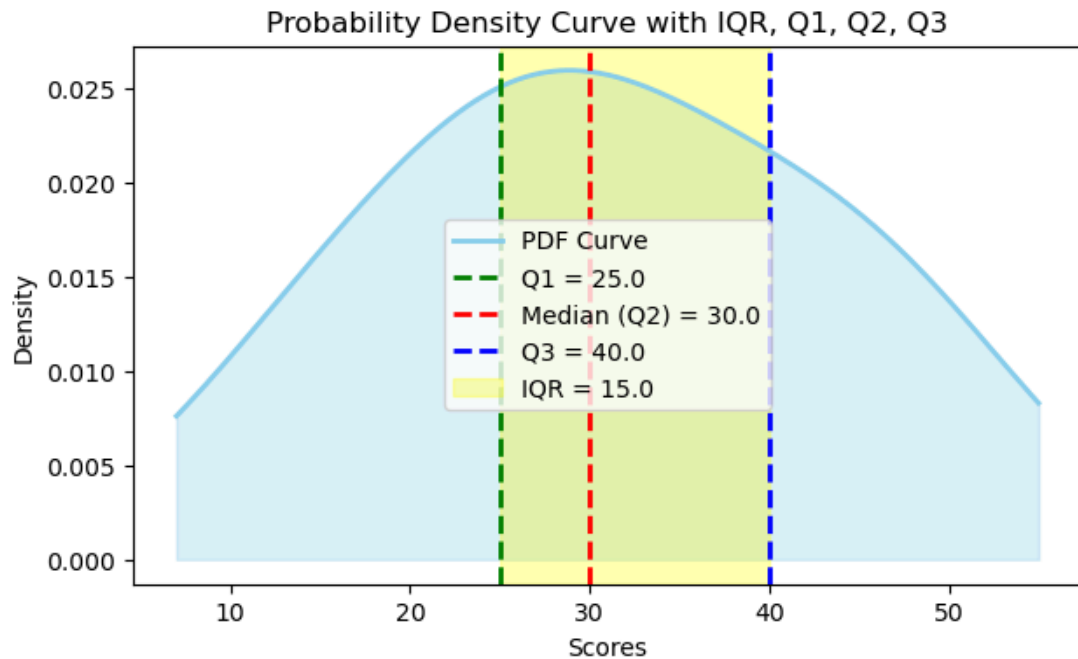
So the box in the boxplot (between Q1 = 25 and Q3 = 40) contains the middle 50% of the data values.

The box plot shows:

Middle 50% of values (IQR = Q3 - Q1 = 40 - 25 = 15) lies between 25 and 40.

Median (30) is exactly in the center of the dataset, showing balanced distribution.

Whiskers extend to min = 12 and max = 50, and no outliers are detected.



## Probability Density Curve (PDF) for Continuous Data – Normal Distribution

### What is a Probability Density Function (PDF)?

A **Probability Density Function (PDF)** describes the **likelihood** of a **continuous random variable** taking on a specific range of values.

For continuous data:

- The **probability at a single point is zero**
- Probability is calculated **over an interval**
- The **area under the curve** between two values gives the probability of falling in that range

The **total area under the curve** = 1, representing 100% probability.

---

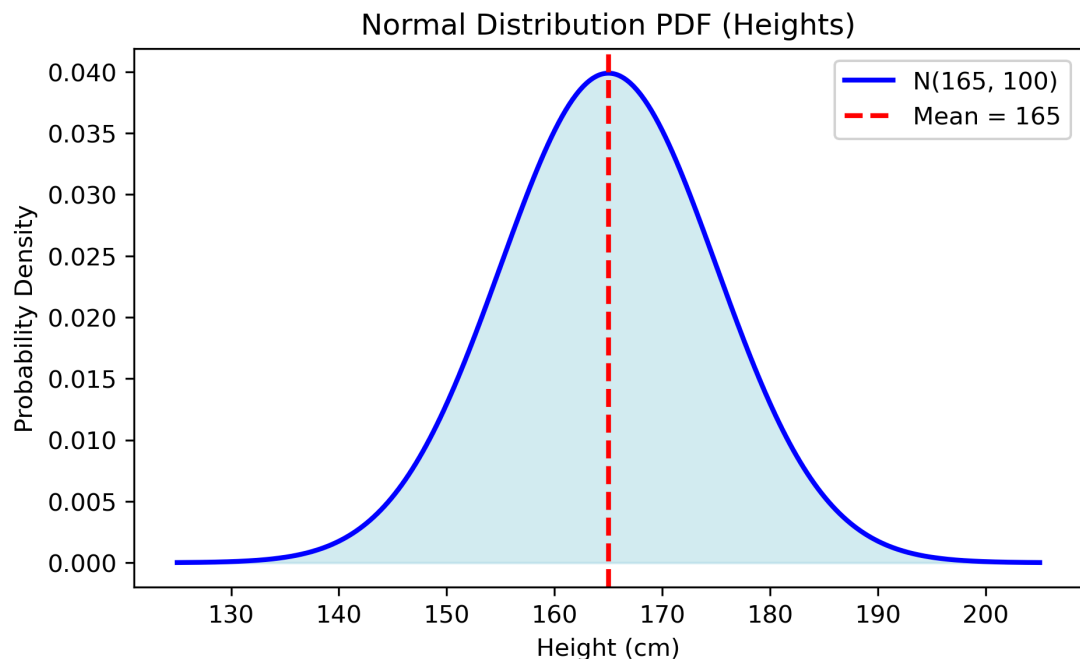
## Example: Heights of Students

Suppose we collected the **heights of 200 students**, and they follow a **normal distribution**.

- Mean height ( $\mu$ ) = 165 cm
- Standard deviation ( $\sigma$ ) = 10 cm

We want to understand:

- The shape of the distribution
  - What percentage of students are between 155 and 175 cm
- 



■ The curve represents the **probability density** of students' heights.

The shaded area between 155 cm and 175 cm gives the **probability that a randomly selected student falls in that range**.

---

### Curve Properties

- **Peak (Center):** At  $\mu = 165$  cm

- **Symmetry:** The curve is symmetrical around the mean
- **Spread:** Controlled by standard deviation ( $\sigma = 10$  cm)

Using the normal distribution rule:

- ~68% of students fall within **1 standard deviation** (155 cm to 175 cm)
- ~95% fall within **2 standard deviations** (145 cm to 185 cm)

So, **probability that height is between 155 and 175 cm  $\approx$  68%**

---

## Formula for Normal Distribution PDF

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

Where:

- $\mu$ : Mean
  - $\sigma$ : Standard deviation
  - $x$ : Value of the random variable
- 

Feature	Value
Variable	Height of students
Mean ( $\mu$ )	165 cm
Std. Deviation ( $\sigma$ )	10 cm
Distribution Type	Normal
Probability (155–175)	$\approx$ 68%

- PDF is used to understand **how values are distributed** in continuous data.
  - The **area under the curve** between two points gives the **actual probability**.
- 

## CLT Central Limit Theorem

### Central Limit Theorem (CLT)

The **Central Limit Theorem (CLT)** states that the distribution of sample means approximates a **normal distribution** as the sample size gets larger (assuming that all samples are identical in size), **regardless of the population distribution shape**.

**CLT in one sentence:**

*"Even if I'm not normal, the average is normal."*



### Rule of thumb:

When collecting means of samples from any distribution, the sample size should be  $n \geq 30$  for the normal approximation to be reliable.

## Why is the Central Limit Theorem (CLT) Useful?

- **Hypothesis Testing**

We can use z-tests and t-tests because CLT ensures sample means follow an approximately normal distribution.

- **Confidence Intervals**

Allows estimation of population parameters (like mean or proportion) with a known margin of error and reliability.

- **Machine Learning**

Forms the foundation for statistical inference, model evaluation, and many algorithms that rely on sampling distributions.

## Classroom Analogy for CLT

Imagine you want to know “**How tall is the average person in a classroom of 30?**”

- Take **one classroom ( $n = 30$ )**.
- Compute the **average height**.
- This gives you **one number** (e.g., 5.6 feet).

To study the **distribution of averages**, you’d need to:

- Sample **many classrooms**, each with 30 people.
- Compute the **average height** for each classroom.
- Collect all those averages.

Plotting those averages shows the **Central Limit Theorem in action**:

the distribution of averages approaches a **normal (bell-shaped) curve**, even if individual heights aren’t perfectly normal.



## Central Limit Theorem (Step-by-Step with Dice)

### Step 1: One sample ( $n = 30$ )

- You roll a die **30 times**.
- You calculate the **average** of those 30 rolls.
- This gives you **one number** (e.g., 3.47).

So **one sample of size 30** → **one average value**.

## Step 2: Many samples

To study CLT, we don't stop at one sample.

We **repeat the process** (say 10,000 times):

- Roll 30 dice.
- Take the average.
- Record that average.

Now we have **10,000 average values**.

## Step 3: Histogram of averages

- We then plot those **10,000 averages** in a **histogram**.
  - That histogram is called the **sampling distribution of the mean**.
  - For **n = 30**, the histogram looks **smooth and bell-shaped (close to normal)**.
- 

## Scaling

**Scaling** is the process of transforming features (variables) so they fit into a specific range or distribution.

This is important in machine learning and statistics because many algorithms (e.g., KNN, SVM, gradient descent) are sensitive to differences in scale.

## Standardization and Normalization

### 1. Standardization (Z-score Scaling)

- **Definition:** Rescales data so it has mean = 0 and standard deviation = 1.
- **Formula:**

$$Z_{score} = \frac{X - \bar{x}}{s}$$

- **Example:**  
Data: [10, 20, 30]
  - Mean ( $\bar{x}$ ) = 20, Std ( $s$ )  $\approx$  8.16
  - Transformed: [-1.22, 0, +1.22]

#### Z-Score Standardization

- **Formula:**  $(X - \mu) / \sigma$
  - **Range:** Not fixed (can be negative, >1, or < -1).
  - Typically: most values lie between **-3 and +3** (for normal data).
- 

### 2. Normalization (Min-Max Scaling)

- **Definition:** Rescales data into a fixed range, usually  $[0, 1]$ .

- **Example:**

Data:  $[10, 20, 30]$

- Min = 10, Max = 30
- Transformed:  $[0, 0.5, 1]$

## Importance of Z-Score

### 1. Standardization

Converts different datasets into a common scale (mean = 0, std = 1).

### 2. Outlier Detection

- If  $|Z| > 3 \rightarrow$  the point is usually considered an outlier.

### 3. Comparison Across Variables

- Useful when features have different units (e.g., height in cm vs. weight in kg).

### 4. Probability & Normal Distribution

- Z-scores are directly linked to probabilities in the **standard normal distribution**.
- Example: About 68% of data falls within  $Z = -1$  to  $+1$ .

## Z-Score and Probability Examples

We'll explore how to calculate:

- Z-scores
- Probabilities from Z-scores
- Number of values above or below a certain score in a dataset

---

### Dataset Used:

Student	Score
A	60
B	70
C	80
D	90
E	100

- Mean ( $\mu$ ) = 80
- Standard Deviation ( $\sigma$ )  $\approx 14.14$

- Total Data Points = 5
- 

## Example 1: How many values are **below 80**?

### Step 1: Z-score for $X = 80$

$$Z = \frac{X - \mu}{\sigma} = \frac{80 - 80}{14.14} = 0$$

### Step 2: Z-table Probability

$$P(Z \leq 0) = 0.5$$

👉 This means **50%** of values are less than 80.

### Step 3: Apply to Dataset

$$0.5 \times 5 = 2.5 \approx 2 \text{ or } 3 \text{ values}$$

### ✅ Manual Check:

Score	< 80?
60	✅
70	✅
80	❌ (equal)
90	❌
100	❌

✔ Matches: 2 values below 80.

---

## ✅ Example 2: How many values are **greater than 90**?

### Step 1: Z-score for $X = 90$

$$Z = \frac{90 - 80}{14.14} \approx 0.71$$

### Step 2: Z-table Probability

$$P(Z < 0.71) \approx 0.7611$$

$$P(Z > 0.71) = 1 - 0.7611 = 0.2389$$

So **23.89%** of values are greater than 90.

## Step 3: Apply to Dataset

$$0.2389 \times 5 = 1.1945 \approx 1 \text{ value}$$

### ✓ Manual Check:

Score	> 90?
60	✗
70	✗
80	✗
90	✗ (equal)
100	✓

✓ Matches: 1 value greater than 90.

## Summary Table

Query	Z-score	Probability	Approx. Count in Dataset
X < 80	0	0.5	2 or 3 values
X > 90	0.71	0.2389	~1 value

```
In [21]: import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Sample dataset (column)
#scores = [45, 52, 67, 70, 81, 90, 55, 60, 72, 85, 95, 100]
scores = [60,70,80,90,100]

# Convert to numpy array
data = np.array(scores)

# Plot histogram + probability density curve
plt.figure(figsize=(8,5))

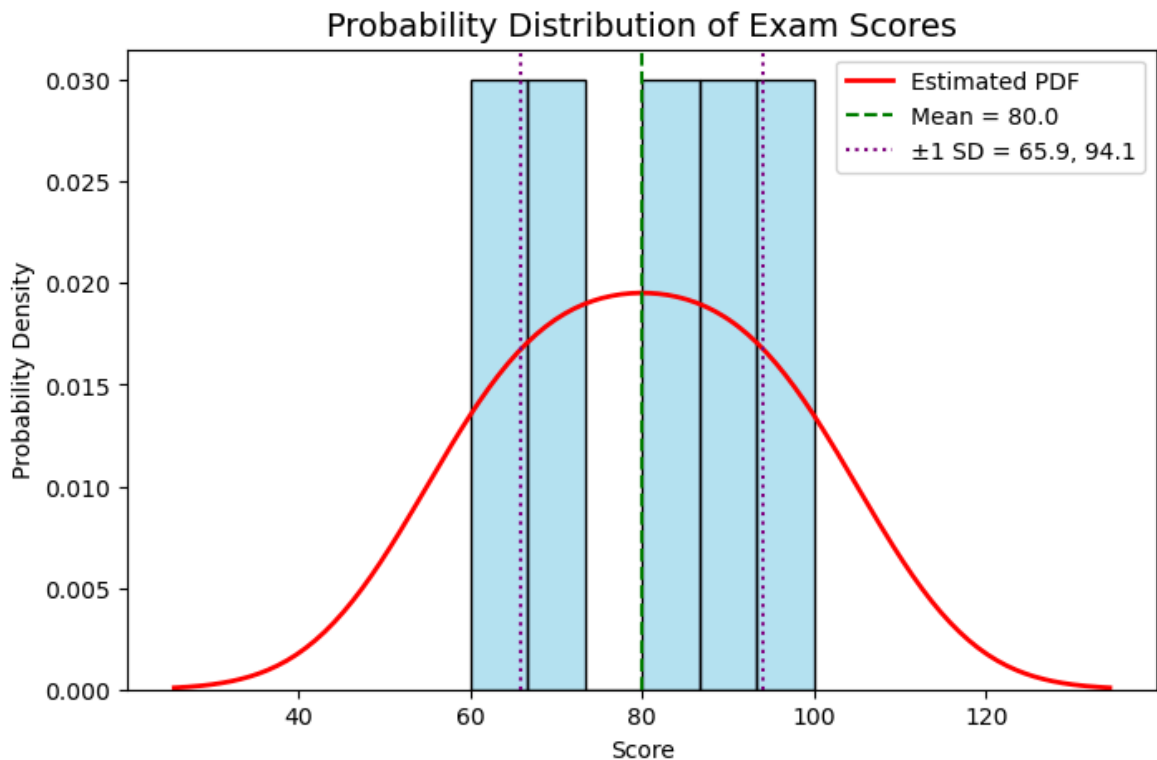
# Histogram (normalized so area = 1 → probability density)
sns.histplot(data, bins=6, kde=False, stat="density", color="skyblue", edgecolor="black")

# KDE = kernel density estimate (smooth probability distribution curve)
sns.kdeplot(data, color="red", linewidth=2, label="Estimated PDF")

# Mean and SD
mean = np.mean(data)
std_dev = np.std(data)

plt.axvline(mean, color="green", linestyle="--", label=f"Mean = {mean:.1f}")
plt.axvline(mean+std_dev, color="purple", linestyle=":", label=f"+1 SD = {mean+std_dev:.1f}")
plt.axvline(mean-std_dev, color="purple", linestyle=":", label=f"-1 SD = {mean-std_dev:.1f}")
```

```
plt.title("Probability Distribution of Exam Scores", fontsize=14)
plt.xlabel("Score")
plt.ylabel("Probability Density")
plt.legend()
plt.show()
```



## Hypothesis Testing



### What Is Hypothesis Testing?

Hypothesis testing is a way for scientists or researchers to **test ideas or claims** using data.

Think of it like a **trial**: you make a claim (the hypothesis), collect evidence (data), and then decide if your claim makes sense based on the evidence.

---

### Real-Life Example

Let's say school canteen says:

"Students eat **an average of 2 apples a day**."

You and your friends think, "That can't be right. We don't eat that many apples!"

So you decide to **test their claim** by collecting data.

---

# Steps in Hypothesis Testing

## 1. State the Hypotheses

- **Null Hypothesis ( $H_0$ ):** The canteen is correct. Students eat 2 apples a day.
- **Alternative Hypothesis ( $H_1$ ):** The canteen is wrong. Students eat **less** (or more) than 2 apples a day.

## 2. Collect Data

Ask 30 students how many apples they eat per day and find the average.

## 3. Analyze the Data

Use math (like calculating the mean and standard deviation) to check if the new average is **significantly different** from 2.

## 4. Make a Decision

- If your results show a big enough difference, you can **reject the null hypothesis**.
  - If not, you **fail to reject it** — meaning the canteen's claim might still be true.
- 

# Important Terms

- **Hypothesis:** A guess or claim you test.
  - **Null Hypothesis ( $H_0$ ):** The claim you're testing (usually that nothing has changed).
  - **Alternative Hypothesis ( $H_1$ ):** The opposite of the null (something **has** changed).
  - **Significance level ( $\alpha$ ):** A cutoff (like 5%) to decide if the result is rare or not.
- 

# Type I and Type II Errors

When we make a decision in hypothesis testing, there's always a chance we could be wrong.

These wrong decisions are called **Type I and Type II errors**.

---

# Continuing Our Apple Example

The school canteen claims:

▮ "Students eat 2 apples per day."

You test this claim using hypothesis testing.

- **Null Hypothesis ( $H_0$ ):** Students eat 2 apples per day.
  - **Alternative Hypothesis ( $H_1$ ):** Students do **not** eat 2 apples per day.
- 

## ✗ Type I Error (False Positive)

You **reject  $H_0$**  even though it's **actually true**.

Example: You say the canteen is wrong (students *don't* eat 2 apples), but in reality, they **do** eat 2 apples per day.

- ◆ It's like **punishing an innocent person**.
- ◆ The chance of making this error is called **alpha ( $\alpha$ )**, usually 5%.

## ✗ Type II Error (False Negative)

You **fail to reject  $H_0$**  even though it's **actually false**.

Example: You say the canteen is correct (students eat 2 apples), but in reality, they **don't** — maybe they eat only 1 apple per day.

- ◆ It's like **letting a guilty person go free**.
- ◆ The chance of this error is called **beta ( $\beta$ )**.

## Quick Summary

Decision	Reality ( $H_0$ True)	Reality ( $H_0$ False)
Reject $H_0$	✗ Type I Error	✓ Correct Decision
Fail to Reject $H_0$	✓ Correct Decision	✗ Type II Error

## What Is a Correct Decision?

In hypothesis testing, a **correct decision** happens when your conclusion **matches the actual truth** about the claim.

### Example: Apple-Eating at School

The school canteen says:

**"Students eat 2 apples per day."**

You test this claim using data from your classmates.

## Two Types of Correct Decisions

### 1. Fail to Reject $H_0$ — and $H_0$ is True

- You say:  
"The canteen's claim seems correct — students eat 2 apples."



- And in reality, this is **true**.

✓ You **accepted a true claim** → **Correct Decision!**

---

## 2. Reject $H_0$ — and $H_0$ is False

- You say:  
"The canteen's claim is wrong — students do NOT eat 2 apples."
- And in reality, they actually **don't** (maybe only 1 apple per day).

✓ You **rejected a false claim** → **Correct Decision!**

---

## Summary Table

Your Decision	Reality	Outcome
Reject $H_0$	$H_0$ is False	✓ Correct Decision
Fail to Reject $H_0$	$H_0$ is True	✓ Correct Decision

---

A correct decision is like giving the **right answer** in a multiple-choice question after checking the facts!

## Hypothesis Testing: Significance Value ( $\alpha$ ) and Confidence Interval (CI)

### 1. Significance Value ( $\alpha$ )

The **significance value** ( $\alpha$ ) is the **threshold probability** we set **before hypothesis testing**.

It determines **when to reject the null hypothesis ( $H_0$ )**.

---

#### 1.1 What is $\alpha$ ?

$\alpha = 0.05$  (5%),  $\alpha = 0.01$  (1%),  $\alpha = 0.10$  (10%)

**Formula:**

$$\alpha = 1 - \text{Confidence Level}$$

**Example:**

- Confidence level = **95%**

$$\alpha = 1 - 0.95 = 0.05$$

---

## Step 1 — Choose Confidence Level

- Typical values: 90%, 95%, 99%.
- **Z-values** table:

Confidence Level	$\alpha$	Z-value
90%	0.10	1.645
95%	0.05	1.96
99%	0.01	2.576

---

## Step 2 — Gather Data

- Sample size  $n$
- Sample mean  $\bar{x}$
- Standard deviation  $s$  or  $\sigma$

---

## Step 3 — Apply Formula

### Example (Engineering Context)

- $n = 30$
- $\bar{x} = 5.3$  mm
- $s = 0.4$  mm
- Confidence level = 95%

**Step 1: Z-value**  $Z = 1.96$

**Step 2: Standard Error**  $SE = \frac{s}{\sqrt{n}} = \frac{0.4}{\sqrt{30}} \approx 0.073$

**Step 3: CI Calculation**

$$CI = 5.3 \pm 1.96 \times 0.073$$

$$CI = [5.16, 5.44] \text{ mm}$$

## P-value

### P-Value (Beginner Definition)

- The **p-value** tells us **how likely our sample result (or something more extreme) is, if the null hypothesis ( $H_0$ ) were true.**
  - **Small p-value** → **evidence against  $H_0$ .**
  - Rule of thumb:
    - If  **$p < 0.05$**  → Reject  $H_0$  (significant).
    - If  **$p \geq 0.05$**  → Fail to reject  $H_0$  (not significant).
-

## Example

Claim: Students eat **1 apple/day** ( $H_0: \mu = 1$ ).

Sample: Average = **1.3 apples/day**, p-value = **0.02**.

Interpretation:

- If students really ate 1 apple/day, the chance of seeing a sample average of 1.3 or more is only **2%**.
- Since  $p = 0.02 < 0.05$ , we **reject  $H_0$**  → evidence that students eat more than 1 apple/day.

**one Tail and two tail**

## P-value: One-Tailed vs Two-Tailed Tests

### **1** One-Tailed Test

- Hypothesis checks only **one direction** (greater than or less than).
  - Example: Test if students eat **more than 1 apple/day** ( $H_1: \mu > 1$ ).
  - The p-value is the **area in one tail** (right side for  $\mu >$ , left side for  $\mu <$ ).
- 

### **2** Two-Tailed Test

- Hypothesis checks for **any difference** (not equal).
  - Example: Test if students eat **different from 1 apple/day** ( $H_1: \mu \neq 1$ ).
  - The p-value is the **area in both tails** (extreme low and extreme high).
- 

 Rule:

- **One-tailed** → Use when you care about only one direction.
- **Two-tailed** → Default choice when testing for “difference.”

## Hypothesis Testing

Hypothesis testing is a method used to **make decisions using data**. It helps us decide whether a claim about a **population** is likely to be true.

## Hypothesis Testing Process – With Simple Examples

# Step-by-Step Process

## 1. Set up Null and Alternate Hypotheses

- **H<sub>0</sub> (Null Hypothesis)**: The statement we test (no change/no effect)
- **H<sub>1</sub> (Alternative Hypothesis)**: What we want to prove (there is a change/effect)

## 2. Decide the Significance Level ( $\alpha$ )

- Common choices: 0.05 (5%) or 0.01 (1%)

## 3. Select the Right Test

- Use **Z-test** if population standard deviation is known
- Use **T-test** if population standard deviation is unknown and sample size  $< 30$

## 4. Calculate the Test Statistic (Z or T score)

- Formula for Z-score:

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

- Formula for T-score:

$$T = \frac{\bar{X} - \mu}{s / \sqrt{n}}$$

## 5. Find the p-value

## 6. Compare p-value and $\alpha$

- If **p <  $\alpha$** , reject the null hypothesis
- If **p >  $\alpha$** , accept the null hypothesis

---

# Hypothesis Testing with P-value and Confidence Intervals

---

## Example Scenario

A machine produces **bolts** with a target diameter of **5 mm**.

We want to check if the **average diameter** has **increased**.

### Given:

- Population standard deviation,  $\sigma = 0.4$  mm
  - Sample size,  $n = 25$
  - Sample mean,  $\bar{x} = 5.3$  mm
  - Significance level,  $\alpha = 0.05$
-

## Step 1 — Define Hypotheses

### Case A: Right-Tailed Test (Check if diameter has increased)

$$H_0 : \mu = 5 \quad (\text{machine is producing correct size})$$

$$H_1 : \mu > 5 \quad (\text{machine produces larger bolts})$$

---

## Step 2 — P-value Approach

### Formula:

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

### Calculation:

$$Z = \frac{5.3 - 5}{0.4 / \sqrt{25}} = \frac{0.3}{0.08} = 3.75$$

### Find P-value

Using **Z = 3.75** and right-tailed test:

$$p = P(Z > 3.75) \approx 0.00009$$

### Decision Rule:

- If  $p < \alpha \rightarrow$  **Reject**  $(H_0)$
- Here,  $p = 0.00009 < 0.05$  ✓

### Conclusion:

The bolts are **significantly larger** than 5 mm.

---

## Step 3 — Confidence Interval Approach

### Formula for CI:

#### (a) Right-Tailed CI

$$CI = \left[ \bar{x} - Z_{\alpha} \cdot \frac{\sigma}{\sqrt{n}}, +\infty \right)$$

- For  $\alpha = 0.05$ ,  $Z_{\alpha} = 1.645$
- Lower limit:

$$5.3 - 1.645 \cdot \frac{0.4}{5} = 5.3 - 0.1316 = 5.17$$

$$CI = [5.17, +\infty) \text{ mm}$$

Since  $\mu_0 = 5$  is **below the lower limit**, reject  $\setminus H_0$  ✓

## (b) Two-Tailed CI

If we test whether **diameter has changed** (larger or smaller):

$$H_1 : \mu \neq 5$$

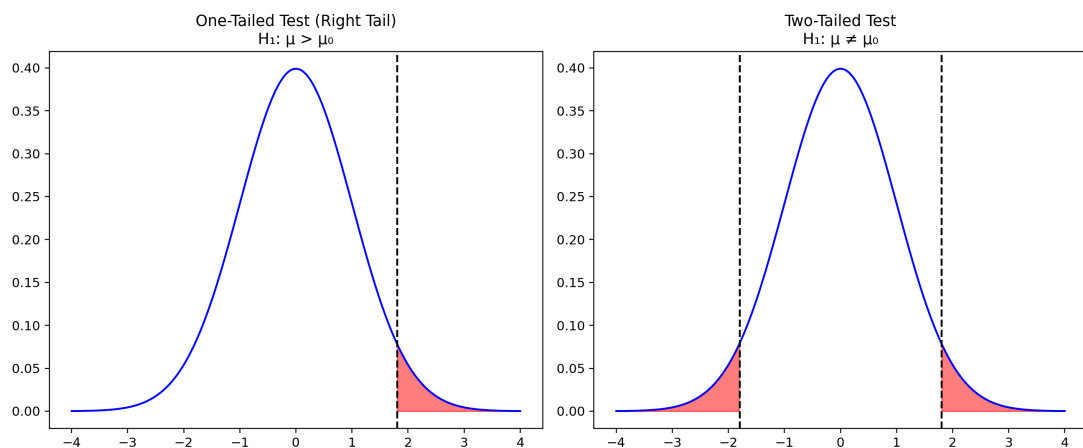
$$CI = \left[ \bar{x} - Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \right]$$

- For  $\alpha = 0.05$ ,  $Z_{\alpha/2} = 1.96$
- Lower limit:  $5.3 - 1.96 \cdot \frac{0.4}{5} = 5.3 - 0.1568 = 5.143$
- Upper limit:  $5.3 + 1.96 \cdot \frac{0.4}{5} = 5.3 + 0.1568 = 5.456$

$$CI = [5.143, 5.456] \text{ mm}$$

Since  $\mu_0 = 5$  lies **outside** this range → **Reject  $\setminus (H_0 \setminus)$**  ✓

## Step 4 — Visualization



## Step 5 — Summary Table

Aspect	Right-Tailed Test	Two-Tailed Test
Alternative (H <sub>1</sub> )	$\mu > \mu_0$	$\mu \neq \mu_0$
Rejection Region	Right tail only	Both tails
Z critical	$Z_{\alpha} = 1.645$	$Z_{\alpha/2} = 1.96$

Aspect	Right-Tailed Test	Two-Tailed Test
Confidence Interval	$[5.17, +\infty)$ mm	$[5.143, 5.456]$ mm
Decision	Reject $(H_0)$	Reject $(H_0)$

## Final Conclusion

- **P-value Approach:** ( $p = 0.00009 < 0.05$ ) → **Reject ( $H_0$ )**
- **Confidence Interval Approach:**
  - **One-tailed CI:** (5) lies **outside** → Reject ( $H_0$ )
  - **Two-tailed CI:** (5) lies **outside** → Reject ( $H_0$ )

✅ The machine is **producing bolts larger than 5 mm**.

## Independent Samples t-test

### What it does

- Compares the **means** of two *independent groups*.
- Helps to check if the difference in means is **significant** or just due to chance.

### When to use

- Groups are **different people**, not the same measured twice.
- Examples:
  - Test scores of **boys vs. girls**
  - Average weight of **smokers vs. non-smokers**
  - Customer satisfaction for **Brand A vs. Brand B**

## Assumptions

1. Groups are independent (no overlap of participants).
2. Data in each group is approximately normal.
3. Variances of the two groups are equal (or use a corrected test if not).

## Simple Example

- Group 1: Students who studied with music → mean score = 75
  - Group 2: Students who studied without music → mean score = 70
  - Independent t-test checks if the **5-point difference** is statistically significant.
- 

## Applications

- **Education:** Compare exam scores of students in two different teaching methods.
  - **Medicine:** Compare recovery time of patients given two different treatments.
- 

Perfect! Let's walk through **one simple example each** for:

1. **Independent Two-Sample t-Test**
  2. **Dependent (Paired) t-Test**
- 

## When to use

- Same subjects measured twice (before/after), or matched pairs.
- Examples: weight before vs after diet, BP with vs without drug.

## 1. Independent t-test Example (Unpaired Groups)

### Scenario:

You want to compare **marks of boys and girls** in a math test.

- Group 1 (Boys): [55, 60, 65, 70, 75]
  - Group 2 (Girls): [65, 70, 72, 68, 74]
- 

## Step-by-Step:

### 1 Hypotheses:

- **H<sub>0</sub>:**  $\mu_1 = \mu_2$  (No difference in scores)
- **H<sub>1</sub>:**  $\mu_1 \neq \mu_2$  (There is a difference)

### 2 Calculate sample means:

- Mean (Boys) = 65
- Mean (Girls) = 69.8

### 3 Use independent t-test formula or calculator:

- Test statistic **t**  $\approx$  **-2.02**
- Degrees of freedom  $\approx$  8



- Critical t-value ( $\alpha = 0.05$ , two-tailed)  $\approx \pm 2.306$

#### 4 Decision:

- $|t| = 2.02 < 2.306 \rightarrow \text{Fail to reject } H_0$

✓ **Conclusion:** No significant difference between boys' and girls' scores.

## 2. Dependent t-test Example (Paired Groups)

### Scenario:

You want to test if a **coaching class improved scores**. You take scores of **same students** before and after the class.

- **Before:** [50, 55, 52, 60, 58]
- **After:** [55, 60, 58, 65, 64]

### Step-by-Step:

#### 1 Hypotheses:

- $H_0: \mu_d = 0$  (No improvement)
- $H_1: \mu_d > 0$  (Improved after class)

#### 2 Find differences:

Student	After	Before	Difference (d)
1	55	50	5
2	60	55	5
3	58	52	6
4	65	60	5
5	64	58	6


- Mean of d = 5.4
- SD of d  $\approx 0.55$
- n = 5

#### 3 t-statistic:

$$t = \frac{\bar{d}}{s_d / \sqrt{n}} = \frac{5.4}{0.55 / \sqrt{5}} \approx 21.97$$

- Critical t ( $\alpha = 0.05$ , df = 4, one-tailed)  $\approx 2.132$

#### 4 Decision:

- $21.97 > 2.132 \rightarrow$   **Reject  $H_0$**

 **Conclusion:** The class **significantly improved** scores.

---



## Summary:

Type	Groups	Example	Test Used
<b>Independent</b>	Different (boys vs girls)	Math test comparison	Independent t-test
<b>Dependent</b>	Same students (before vs after)	Coaching class	Paired t-test

---

# ANOVA (Analysis of Variance)

## Full Form

**ANOVA = Analysis of Variance**

---

## Introduction

- A statistical method to compare the **means of 3 or more groups**.
  - Checks whether differences in sample means are **statistically significant** or just due to random variation.
  - Based on partitioning the total variation into:
    - **Between-group variation** (differences due to treatments/groups)
    - **Within-group variation** (random error, natural differences within groups)
- 

## Uses

- Compare effectiveness of **multiple teaching methods**.
- Compare **average yield** of different fertilizers.
- Compare **customer satisfaction** across different brands.
- In general: any case with **3 or more group means**.



## One-Way ANOVA Example (Manual Calculation)



### Objective:

Test if there is a significant difference in average scores among **three different classes** of students.

---

## Data Table:

Class A	Class B	Class C
40	42	55
45	41	60
50	44	65

- **k** = 3 groups
  - **n** = 3 observations per group
  - **N** = 9 total observations
- 

### Step 1: Calculate Group Means

- Mean A =  $(40 + 45 + 50) / 3 = 45$
  - Mean B =  $(42 + 41 + 44) / 3 \approx 42.33$
  - Mean C =  $(55 + 60 + 65) / 3 = 60$
  - **Grand Mean (GM)** = Total sum / 9 =  $442 / 9 \approx 49.11$
- 

### Step 2: Sum of Squares

#### ◆ Between Groups (SSB)

$$SSB = n \cdot \sum (\bar{X}_{group} - GM)^2$$
$$= 3 \cdot [(45 - 49.11)^2 + (42.33 - 49.11)^2 + (60 - 49.11)^2] = 544.56$$

---

#### ◆ Within Groups (SSW)

$$SSW = \sum (X_{ij} - \bar{X}_{group})^2 = 50(A) + 4.66(B) + 50(C) = 104.66$$

---

#### ◆ Total Sum of Squares (SST)

$$SST = SSB + SSW = 544.56 + 104.66 = 649.22$$

---

### Step 3: Degrees of Freedom

Source	Formula	Value
Between Groups	$(k - 1)$	2
Within Groups	$(N - k)$	6

	Source	Formula	Value
	Total	(N - 1)	8

## Step 4: Mean Squares

- $MSB = \frac{SSB}{df_{between}} = \frac{544.56}{2} = 272.28$
- $MSW = \frac{SSW}{df_{within}} = \frac{104.66}{6} \approx 17.44$

## Step 5: F-Ratio

$$F = \frac{MSB}{MSW} = \frac{272.28}{17.44} \approx 15.61$$

## Final ANOVA Table

Source	SS	df	MS	F
Between Groups	544.56	2	272.28	15.61
Within Groups	104.66	6	17.44	
Total	649.22	8		

## Conclusion

Compare the F-value (15.61) with the F-critical value from the F-table at  $\alpha = 0.05$ .

- ( $F_{\text{critical}} \approx 5.14$ ) for ( $df1 = 2, df2 = 6$ )
- Since **15.61** > **5.14**, we **reject the null hypothesis**.

 **At least one group has a significantly different average score.**

```
In [13]: import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import f_oneway

# Raw data
class_A = [40, 45, 50]
class_B = [42, 41, 44]
class_C = [55, 60, 65]

# Store in dictionary
data = {
    "Class A": class_A,
    "Class B": class_B,
```

```

    "Class C": class_C
}

# Calculate means and standard deviations
classes = list(data.keys())
means = [np.mean(scores) for scores in data.values()]
std_devs = [np.std(scores, ddof=1) for scores in data.values()] # sample SD

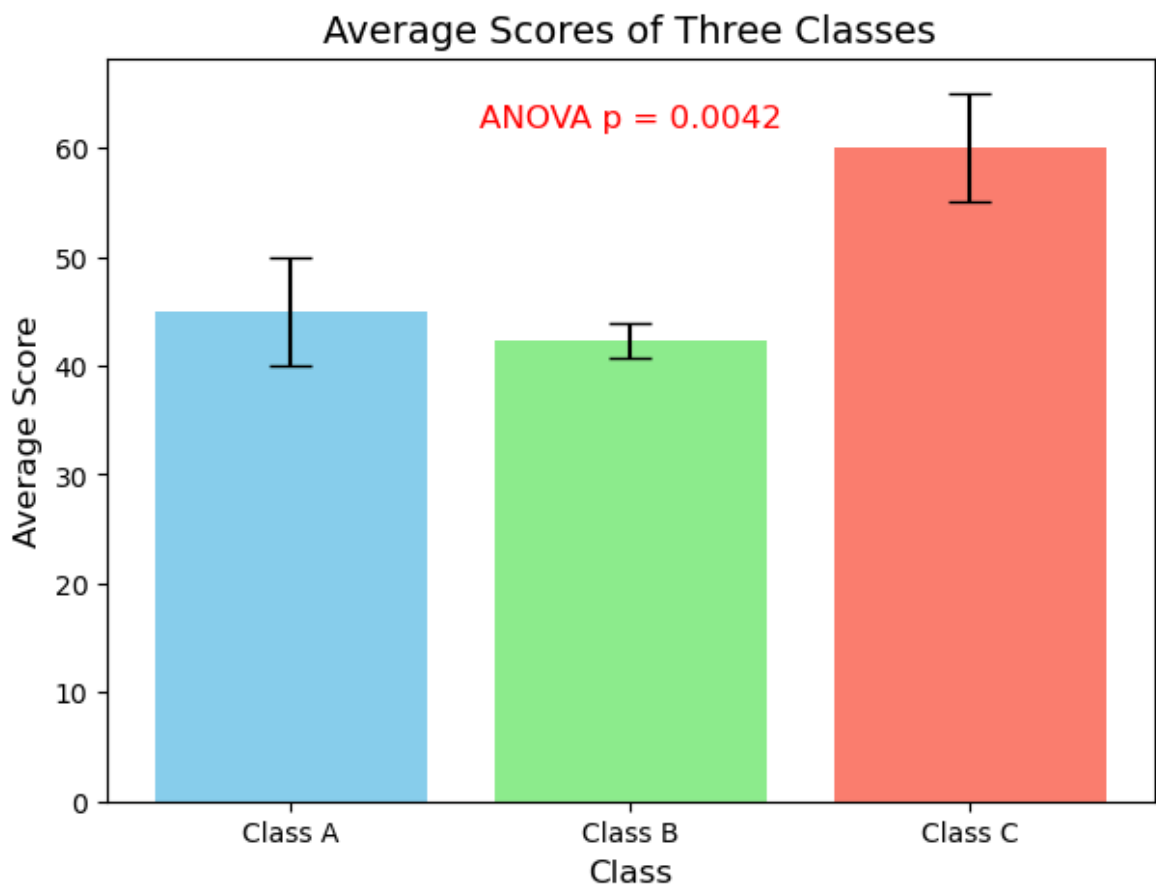
# Run ANOVA
F_stat, p_value = f_oneway(class_A, class_B, class_C)
print(f"F-statistic = {F_stat:.2f}, p-value = {p_value:.4f}")

# Plot bar chart with error bars
plt.figure(figsize=(7,5))
plt.bar(classes, means, yerr=std_devs, capsize=8, color=['skyblue', 'lightgreen',
# Add title + annotation with p-value
plt.title("Average Scores of Three Classes", fontsize=14)
plt.ylabel("Average Score", fontsize=12)
plt.xlabel("Class", fontsize=12)
plt.text(1, max(means)+2, f"ANOVA p = {p_value:.4f}", ha='center', fontsize=12,

# Save and show graph
plt.savefig("anova_with_pvalue.png", dpi=300)
plt.show()

```

F-statistic = 15.60, p-value = 0.0042



ince  $p\text{-value} = 0.0009 < 0.05$ , we reject  $H_0$ .

That means: Not all class averages are equal. At least one class differs significantly.

Looking at the bar chart, Class C (mean = 60) is much higher, so it's likely the group that differs.

# Chi-Square Tests

## Types of $\chi^2$ Tests

Feature	Goodness of Fit (Type 1)	Test of Independence
Purpose	Does one categorical variable follow a specified distribution?	Are two categorical variables related?
Data	One variable, k categories	Contingency table (r × c)
H <sub>0</sub>	Observed = Expected (per theory)	Variables are independent
df	k – 1 – (#params estimated)	(r – 1)(c – 1)
Example	Fair die? (1–6 equally likely)	Gender × Preference

## Common Formula

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

- O = observed frequency, E = expected frequency

## Example A — Dice

- Rolls = 60
- Observed: [8, 12, 9, 11, 10, 10]
- Expected (fair): [10, 10, 10, 10, 10, 10]
- Compute:  $\chi^2 = \sum (O - E)^2 / E = 1.0$
- df = 6 – 1 = 5
- Decision ( $\alpha=0.05$ ):  $1.0 < 11.07 \Rightarrow$  **Fail to reject H<sub>0</sub>** (die looks fair)

## Example B — Independence (2×2)

	Science	Arts	Total
Boys	30	10	40
Girls	10	30	40
<b>Total</b>	40	40	80

Expected: all cells =  $(40 \times 40) / 80 = 20$

$$\chi^2 = \sum (O - E)^2 / E = 4 \times (100 / 20) = 20$$

$$df = (2 - 1)(2 - 1) = 1$$

Decision ( $\alpha = 0.05$ ):  $20 > 3.84 \Rightarrow \text{Reject } H_0$  (not independent)

---



## Chi-Square Test ( $\chi^2$ Test)



### What is a Chi-Square Test?

The **Chi-Square ( $\chi^2$ ) Test** is a statistical method used to:

- Compare **observed** values with **expected** values.
- Check if there is a **significant association** between two categorical variables.

There are two main types:

1. **Chi-Square Goodness-of-Fit Test** – Tests how well observed data fits an expected distribution.
  2. **Chi-Square Test of Independence** – Tests if two variables are related in a contingency table.
- 



### Example: Chi-Square Test of Independence



#### Situation:

We want to know whether **gender** and **preference for a subject** are related.

### Step 1 — Hypotheses

- $H_0$ : Gender and subject choice are **independent** (no relation).
- $H_1$ : Gender and subject choice are **not independent** (there is a relation)



#### Data (Observed):

	Science	Arts	Total
Boys	30	10	40
Girls	10	30	40
Total	40	40	80

---



### Step 1: Find Expected Values

Use the formula:

$$E_{ij} = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$$

### Example:

Expected value for Boys–Science:

$$E = \frac{40 \times 40}{80} = 20$$

	Science (E)	Arts (E)
Boys	20	20
Girls	20	20



## Step 2: Apply Chi-Square Formula

Formula:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

### Calculations:

- For Boys–Science:  $(\frac{(30 - 20)^2}{20} = 5)$
- For Boys–Arts:  $(\frac{(10 - 20)^2}{20} = 5)$
- For Girls–Science:  $(\frac{(10 - 20)^2}{20} = 5)$
- For Girls–Arts:  $(\frac{(30 - 20)^2}{20} = 5)$

**Total  $\chi^2 = 5 + 5 + 5 + 5 = 20$**



## Step 3: Degrees of Freedom

$$df = (r - 1)(c - 1) = (2 - 1)(2 - 1) = 1$$



## Step 4: Compare with $\chi^2$ Table

- At  $\alpha = 0.05$  and **df = 1**, critical value  $\approx 3.84$
- Our  $\chi^2 = 20$ , which is **greater** than 3.84

**Conclusion:** Reject  $H_0$

There is a significant relationship between gender and subject preference.



## Summary



Step	Result
Observed Values	Given in table
Expected Values	Calculated using row $\times$ col / total
$\chi^2$ Value	20
Degrees of Freedom	1
Decision	Reject $H_0$ (significant relationship)

In [ ]: