TITANIC PASSENGER SURVIVAL ANALYSIS USING HYPOTHESIS TESTING

By: Lokendra Chawla
Tools Used: Python, Pandas, Seaborn, SciPy

---

1. Introduction

The Titanic disaster is one of the most famous ship accidents in history.
The Titanic dataset contains details about passengers such as:

- Whether they survived
- Their gender
- Their age
- Their passenger class
- Their ticket fare
- Their family size
- Their port of embarkation

The goal of this project is:

To find which factors (gender, class, age, fare, family size, etc.) affected a passenger's chances of survival.

To do this, I used **Hypothesis Testing**, a statistical method that checks whether relationships between variables are real or random.

---

2. Dataset Overview

I used the **Seaborn Titanic dataset**, which includes important columns:

Column	Meaning
survived	0 = did not survive, 1 = survived
sex	male / female
pclass	passenger class (1, 2, 3)

Column	Meaning
age	age in years
fare	price of the ticket
embarked	port where passenger boarded
sibsp	siblings/spouses aboard
parch	parents/children aboard

I cleaned missing values by:

- Filling missing ages → median age
- Filling missing embarked → most common value
- Dropping rows missing essential values

I also created new columns:

- **family_size = sibsp + parch + 1**
- **is_alone = 1 if family_size == 1 else 0**
- **age_group = child / teen / adult / senior**

---

3. Hypothesis Testing Methods

I used 3 statistical tests:

✓ Chi-Square Test

Used when **both variables are categorical**.
Example: Gender vs Survival.

✓ t-test

Used when comparing **average numbers** between two groups.
Example: Age of survived vs not survived.

✓ ANOVA (F-test)

Used when comparing averages of **more than two groups**.
Example: Fare across 1st, 2nd, 3rd class.

✓ Pearson Correlation

Used for checking relationship between **two numeric variables**.

Significance is based on:

If p-value < 0.05 → Relationship is statistically significant.

---

4. Hypothesis Tests & Results

Below are all **10 hypothesis tests** with simple-English conclusions.

1️⃣ Gender vs Survival (Chi-square)

Result:

- Females survived **much more** than males.
- p-value < 0.05 → **Significant**

Conclusion:

Gender strongly affected survival. Women had much higher chances of survival.

2️⃣ Passenger Class vs Survival (Chi-square)

Result:

- 1st class had the highest survival
- 3rd class had the lowest
- p-value < 0.05 → **Significant**

Conclusion:

Passenger class affected survival. Richer passengers in 1st class survived more.

3️⃣ Embarked Port vs Survival (Chi-square)

Result:

Different ports showed different survival patterns.

Conclusion:

Where the passenger boarded the ship also affected survival.

4️⃣ Age Difference (t-test)

Result:

- Survived passengers were slightly younger
- p-value < 0.05 → Significant

Conclusion:

Age has some impact on survival, with younger passengers surviving more.

5️⃣ Fare Difference (t-test)

Result:

- Survived passengers paid higher fares
- p-value < 0.05 → Significant

Conclusion:

Passengers who paid more had better survival chances (likely because they were in higher classes).

6️⃣ Family Size vs Survival (t-test)

Result:

- Survival changed based on family size
- p-value < 0.05 → Significant

Conclusion:

Family size affected survival. Very large families and people traveling alone had lower chances.

7️⃣ Is Alone vs Survival (Chi-square)

Result:

- People traveling alone had lower survival
- p-value < 0.05 → Significant

Conclusion:

Being alone reduced survival chances. Families helped each other survive.

8️⃣ Class vs Fare (ANOVA)

Result:

- Clear differences in ticket prices between the 3 classes
- p-value < 0.05 → Significant

Conclusion:

Ticket price clearly depends on passenger class.

9️⃣ Age Group vs Survival (Chi-square)

Result:

Children had the highest survival.
Adults had lower survival.

Conclusion:

Age group affects survival. Children were protected first.

🔟 Age vs Fare Correlation (Pearson)

Result:

Very weak relationship between age and ticket price.

Conclusion:

Age does NOT influence how much fare a passenger paid.

---

5. Final Conclusion (Easy English)

From the statistical tests, the most important factors that influenced survival were:

✓ Gender — Women survived more

✓ **Passenger Class — 1st class had best survival**

✓ **Fare — Higher-paying passengers survived more**

✓ **Family — Not being alone helped survival**

✓ **Age Group — Children were more likely to survive**

Some factors had **little or no effect**, such as:

❌ Age vs Fare
❌ Exact numeric age differences were small

---

6. What I Learned

Through this project, I learned:

- How to clean data
- How to perform Chi-square tests
- How to perform t-tests
- How to use ANOVA
- How to interpret p-values
- How to convert real problems into statistical hypotheses
- How to create a full data analysis report

This project improved my understanding of:

✓ Hypothesis testing
✓ Python for data analysis
✓ Titanic dataset patterns
✓ Data storytelling
