STATISTICS & PROBABILITY FOR AI —COURSE VERSION

Course Tools: Python (>=3.8), NumPy, Pandas, Matplotlib, Seaborn

Ideal Learners: AI/ML students, data scientists, analysts, or developers transitioning into AI who want to understand how statistics and probability power machine learning.

Course Learning Outcomes

By the end of this course, learners will be able to:

- Explain and calculate descriptive statistics (mean, median, mode, variance, std).
- Understand and visualize probability distributions (normal, binomial, Poisson).
- Conduct hypothesis testing using t-test and chi-square test.
- Analyze and visualize relationships using correlation and covariance.
- Apply statistical reasoning to real-life data-driven AI problems.

DESCRIPTIVE STATISTICS

1.1 Concept Overview

Descriptive statistics summarize and describe features of a dataset.

Measure	**Meaning**	**Use-case**
Mean	Average of all data points	For symmetric distributions
Median	Middle value when sorted	When outliers are present
Mode	Most frequent value	For categorical or discrete data
Variance	Average squared distance from mean	Measures spread
Standard Deviation	Square root of variance	Spread in same unit as data

1.2 Coding Example

```
import numpy as np

import pandas as pd
```

```
# Example: Daily active users for a mobile app over 30 days
np.random.seed(0)
dau = np.random.poisson(lam=120, size=30)

df = pd.DataFrame({'DAU': dau})
print('Mean:', df['DAU'].mean())
print('Median:', df['DAU'].median())
print('Mode:', df['DAU'].mode()[0])
print('Variance:', df['DAU'].var())
print('Standard Deviation:', df['DAU'].std())
```

1.3 Visualization

```
import seaborn as sns
import matplotlib.pyplot as plt

sns.set()
plt.figure(figsize=(10,4))
plt.subplot(1,2,1)
sns.histplot(df['DAU'], kde=True, color='skyblue')
plt.title('Histogram of DAU')

plt.subplot(1,2,2)
sns.boxplot(x=df['DAU'], color='lightgreen')
plt.title('Boxplot of DAU')
plt.show()
```

Classwork: Discuss how outliers affect the mean but not the median.

Real-life Scenario

A product manager observes DAU spikes. You calculate the mean and median to understand user engagement trends and detect abnormal peaks (e.g., promotions, technical issues).

Exercises

1. Compute all descriptive statistics for monthly revenue of a store (12 values). Interpret your findings.
2. Add a few extreme values (outliers) and see which measure changes the most.

PROBABILITY DISTRIBUTIONS

2.1 Key Concepts

Distribution Type		**Example**
Normal	Continuous	Heights, exam scores
Binomial	Discrete	Coin flips, conversions
Poisson	Discrete	Number of calls per hour

2.2 Coding Practice

```
# Normal distribution
normal_data = np.random.normal(loc=50, scale=10, size=1000)

# Binomial distribution
binom_data = np.random.binomial(n=10, p=0.6, size=1000)

# Poisson distribution
poisson_data = np.random.poisson(lam=3, size=1000)

plt.figure(figsize=(12,4))
```

```
for i, (data, title) in enumerate([(normal_data, 'Normal'), (binom_data, 'Binomial'),
(poisson_data, 'Poisson')]):
    plt.subplot(1,3,i+1)
    sns.histplot(data, kde=True, color='coral')
    plt.title(title)
plt.show()
```

Classwork:
Explain that the binomial models successes in fixed trials, while the Poisson models event counts over time.

Exercises

1. Simulate coin flips (n=20, p=0.5) and compute how many times heads appears.
2. Model website hits per minute using a Poisson distribution with $\lambda = 4$.

HYPOTHESIS TESTING

3.1 Concept Overview

Hypothesis testing helps determine if a result is statistically significant.

Concept	Meaning
Null Hypothesis (H_o)	Assumes no effect or difference
Alternative (H_1)	Claims there is an effect
p-value	Probability of getting results as extreme as observed if H_o is true

3.2 t-Test Example

```
from scipy import stats

# Example: Comparing CTR for two versions of a webpage
```

```
np.random.seed(1)

A = np.random.binomial(1, 0.12, 200)

B = np.random.binomial(1, 0.15, 210)


# Two-sample t-test

t_stat, p_val = stats.ttest_ind(A, B)

print(f"t-statistic = {t_stat}, p-value = {p_val}")
```

3.3 Chi-square Test Example

```
import numpy as np

from scipy.stats import chi2_contingency


# Device type vs conversion

contingency = np.array([[50, 450], [80, 420]])

chi2, p, dof, expected = chi2_contingency(contingency)

print(f"Chi2 = {chi2}, p-value = {p}")
```

Classwork:
If $p < 0.05$, reject H_o. There's evidence of difference or association.

Exercises

1. Conduct a paired t-test for student performance before and after a training.
2. Use chi-square to test independence between gender and product preference.

CORRELATION & COVARIANCE

4.1 Core Concepts

- **Covariance:** Direction of relationship between variables.
- **Correlation:** Strength and direction ($-1 \leq r \leq 1$).

- **Spearman correlation:** Non-linear rank-based relationship.

4.2 Coding Example

```
x = np.linspace(0, 10, 100)

y = 2.5 * x + np.random.normal(scale=3.0, size=100)


df = pd.DataFrame({'x': x, 'y': y})

print(df.corr())  # Pearson correlation

sns.lmplot(x='x', y='y', data=df)

plt.title('Correlation Example')

plt.show()
```

Exercises

1. Load Seaborn's iris dataset and visualize pairwise correlations.
2. Create a nonlinear relationship (y = x**2 + noise) and compare Pearson vs Spearman.

QUIZZES

Quiz 1 (Conceptual)

1. Which measure is least affected by outliers? → **Median**
2. Which test compares proportions between two groups? → **Z-test or Chi-square**
3. p-value = 0.03 means? → **3% chance data this extreme if H_o is true**

Quiz 2 (Coding)

1. Generate 1000 samples from N(10, 2) and find the 95% confidence interval for the mean.
2. Write a function that computes and plots Pearson & Spearman correlations.

CAPSTONE PROJECTS

📈 Project 1 — A/B Test for Conversion Rate

Analyze an A/B experiment to see if new website design increases conversions.

- Load simulated dataset with group labels & conversions.
- Perform descriptive stats and hypothesis test.
- Visualize results and report findings.

🔢 Project 2 — Predicting Customer Calls (Poisson)

Model number of daily customer support calls.

- Fit Poisson regression using statsmodels.
- Evaluate model performance and visualize predictions.

💰 Project 3 — Correlation in Sales Data

Find relationships between marketing spend, website traffic, and revenue.

- Compute correlation and covariance.
- Use heatmaps to visualize dependencies.

📚 Suggested Readings

- *Practical Statistics for Data Scientists* (O'Reilly)
- *An Introduction to Statistical Learning* (ISL)
- SciPy and Statsmodels documentation for tests & GLMs

✅ OPTIONAL: Learners proceed to "Python for AI: Linear Algebra & Calculus Essentials."