2026-01-07

Introduction to Statistics: Essentials for Data Science

Ramatoulaye Diallo
DATACAMP COURS

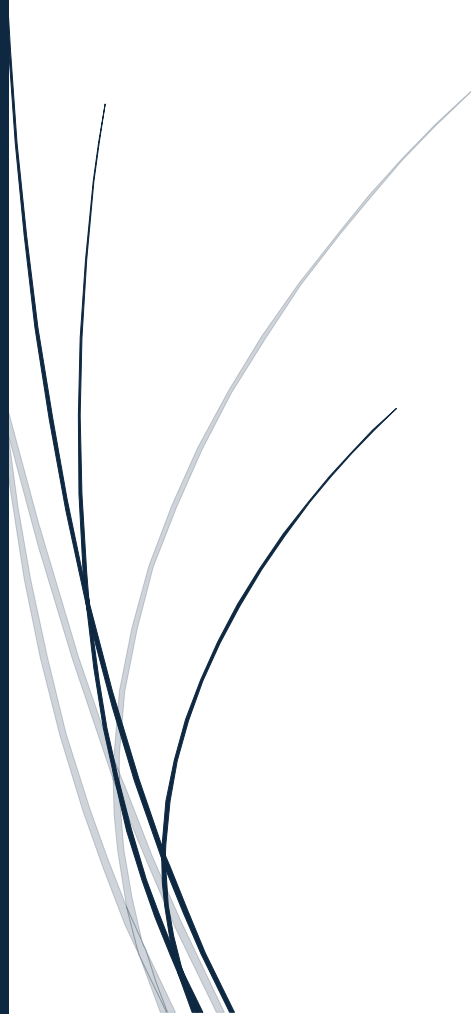

Table of Contents

Introduction to Statistics

Chapter 1: Basics of Statistics

What is Statistics?

- Definition: The practice of collecting, analyzing, and interpreting data.
- Branches:
    - Descriptive Statistics: Summarizes data (mean, median, mode).
    - Inferential Statistics: Uses sample data to make conclusions about a population.

Types of Data

- Numeric:
    - Continuous (e.g., stock prices).
    - Interval/count (e.g., cups of coffee per day).
- Categorical:
    - Nominal (e.g., eye color).
    - Ordinal (e.g., ranking preferences).

Measures of Center

- Mean:

$$\text{Mean} = \frac{\text{Sum of all values}}{\text{Number of values}}$$

- Median: Middle value when data is ordered.
- Mode: Most frequent value.

Measures of Spread

- Range:

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

- Variance:

$$\text{Variance} = \frac{\sum(x_i - \bar{x})^2}{n}$$

- Standard Deviation:

$$\text{SD} = \sqrt{\text{Variance}}$$

- Interquartile Range (IQR):

$$\text{IQR} = Q_3 - Q_1$$

Chapter 2: Probability & Distributions

Basic Probability

- Formula:

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

Conditional Probability

- Formula:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Discrete Distributions

- Expected Value:

$$E(X) = \sum x_i \cdot P(x_i)$$

Example for a fair die:

$$E = \frac{1+2+3+4+5+6}{6} = 3.5$$

Continuous Distributions

- Probability = Area under curve.
- For uniform distribution:

$$P(a \le X \le b) = \frac{b-a}{\text{Range}}$$

Chapter 3: Key Statistical Distributions

Binomial Distribution

- Definition: Probability of $k$successes in $n$independent trials.
- Formula:

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

- Expected Value:

$$E(X) = n \cdot p$$

Normal Distribution

- Symmetrical bell curve.
- Defined by mean (μ) and standard deviation (σ).
- Empirical Rule:
    - 68% within 1σ

- 95% within 2σ
- 99.7% within 3σ

Central Limit Theorem

- As sample size increases, sampling distribution of the mean approaches normal.

Poisson Distribution

- For events over time/space.
- Formula:

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

where λ= average rate.

Chapter 4: Hypothesis Testing, Experiments & Correlation

Why Hypothesis Testing?

- Used to compare populations and validate assumptions.
- Examples:
  - Does a price change increase revenue?
  - Is a medication effective?

Key Concepts

- Null Hypothesis (H_0): Assume no difference exists.
- Alternative Hypothesis (H_a): A difference exists.
- Workflow:
  1. Define populations.
  2. State $H_0$and H_a.
  3. Collect sample data.
  4. Perform statistical tests.
  5. Draw conclusions.

Variables

- Independent Variable: Unaffected by other data (e.g., treatment).
- Dependent Variable: Outcome affected by independent variable.

Experiments

- Treatment vs Control Groups:

    - Treatment group receives intervention.
    - Control group does not.
- Randomization: Assign participants randomly.
- Blinding:
    - Single-blind: Participants don't know group.
    - Double-blind: Neither participants nor administrators know.
- A/B Testing: Common in marketing (two groups only).

Correlation

- Pearson Correlation Coefficient (r): $-1 \leq r \leq 1$
    - Magnitude = strength.
    - Sign = direction.
- Important: Correlation ≠ Causation.
- Confounding Variables: Hidden factors affecting results.

Interpreting Results

- p-value: Probability of observing data assuming $H_0$is true.
- Significance Level (α): Commonly 0.05.
    - If $p \leq \alpha$, reject H_0.
- Errors:
    - Type I: Reject $H_0$when true (false positive).
    - Type II: Fail to reject $H_0$when false (false negative).

Global Conclusion for All Four Chapters

These chapters provide a foundation in statistics:

1. Chapter 1: Basics of data, measures of center (mean, median, mode) and spread (range, variance, SD).
2. Chapter 2: Probability concepts, conditional probability, discrete & continuous distributions.
3. Chapter 3: Key distributions (Binomial, Normal, Poisson), Central Limit Theorem.
4. Chapter 4: Hypothesis testing, experimental design, correlation, interpreting results.

Key Formulas and Concepts to Remember

- Mean:
$$\bar{x} = \frac{\sum x_i}{n}$$
- Variance:
$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n}$$
- Standard Deviation:
$$\sigma = \sqrt{\sigma^2}$$
- Probability:
$$P(A) = \frac{\text{favorable outcomes}}{\text{total outcomes}}$$
- Conditional Probability:
$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$
- Binomial:
$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$
- Poisson:
$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$
- Normal Distribution:
Defined by $\mu$and σ; 68%-95%-99.7% rule.
- Central Limit Theorem: Sampling distribution → normal as $n\uparrow$.
- Hypothesis Testing:
    - Null vs Alternative.
    - p-value and α.
    - Type I & II errors.
- Correlation: $r \in [-1,1]$