STATISTICS AND PROBABILITY FOR DATA SCIENCE

Abstract

This chapter provides a comprehensive overview of foundational statistical concepts essential for data science. The initial part of statistics focuses on descriptive statistics which defines data description using mean and median and mode and variance alongside the summary of dataset features. The chapter then explores probability theory and common probability distributions-including normal, binomial, and Poisson-which form the basis for modeling uncertainty and real-world phenomena. Expanding on these basics, the material delves into inferential statistics-such as hypothesis testing and confidence intervals-which equip data scientists to make informed judgments based on sample data. The text also introduces regression analysis, highlighting both linear and logistic approaches as essential tools for understanding variable relationships and forecasting outcomes. Throughout, practical examples and solved problems illustrate how statistical methods are applied to real-world data science scenarios. By mastering these core topics, readers will be well-equipped to analyze data, interpret results, and make informed decisions in a data-driven environment [1].

Keywords: Descriptive Statistics, Probability Distributions, Inferential Statistics, Hypothesis Testing, Linear Regression, Logistic Regression.

Authors

Shubneet
Department of Computer Science
Chandigarh University, Gharuan
Mohali, 140413, Punjab, India.
jeetshubneet27@gmail.com;

Anushka Raj Yadav
Department of Computer Science
Chandigarh University, Gharuan
Mohali, 140413, Punjab, India.
ay462744@gmail.com;

Partha Chanda
Department of Computer Science
Chandigarh University, Gharuan
Mohali, 140413, Punjab, India.
partha.chanda.ai@gmail.com;

Mohammad Yasir Bin Taleb Abrar
Department of Computer Science
Chandigarh University, Gharuan
Mohali, 140413, Punjab, India.
yasirbintaleb@gmail.com;

Stuti Sood
Department of Computer Science
Chandigarh University, Gharuan
Mohali, 140413, Punjab, India.
stutisood250@gmail.com;

I. INTRODUCTION

Statistics and probability are fundamental to contemporary data science, allowing professionals to uncover valuable patterns in intricate data, assess uncertainty, and make well-informed choices even in ever-changing situations. As organizations increasingly rely on data-driven strategies, these disciplines provide the theoretical framework and practical tools to analyze trends, validate hypotheses, and optimize outcomes [2]. For instance, A/B testing-a cornerstone of data science-leverages statistical methods to compare webpage designs, marketing campaigns, or product features, empowering companies like Amazon and Netflix to refine user experiences and boost conversion rates [3]. Similarly, probabilistic models underpin risk assessment in finance, healthcare diagnostics, and supply chain optimization, demonstrating their universal relevance.

The integration of statistics and probability into data science addresses three critical challenges: (1) managing uncertainty in real-world data, (2) drawing reliable conclusions from incomplete information, and (3) translating technical results into actionable business strategies. In financial risk modeling, probability distributions help quantify market volatility, while inferential statistics enable fraud detection systems to flag anomalous transactions with 98% accuracy [4]. These applications underscore how statistical rigor transforms raw data into strategic assets.

This chapter systematically explores the essential statistical concepts and probabilistic frameworks that every data scientist must master. Through real-world examples and practical implementations, readers will gain proficiency in:

- **Descriptive Statistics:** Condensing information by using metrics that describe the average and the spread of the data.
- **Probability Theory:** Assessing unpredictability through probabilistic frameworks and Bayesian analytical approaches.
- **Inferential Statistics:** Conducting hypothesis tests and constructing confidence intervals.
- **Regression Analysis:** Building predictive models for continuous and categorical outcomes.
- **Bayesian Statistics:** Updating beliefs with empirical evidence.
- **Practical Implementation:** Coding statistical solutions in Python/R.
- **Ethical Considerations:** Avoiding common pitfalls like p-hacking.

The following sections blend theoretical foundations with industry applications, preparing readers to tackle challenges ranging from clinical trial design to algorithmic trading systems. By mastering these concepts, data scientists can confidently navigate the complexities of modern data ecosystems while maintaining methodological rigor.

II. DESCRIPTIVE STATISTICS

Descriptive statistics provide the foundational tools for summarizing and interpreting datasets, enabling data scientists to identify patterns, detect anomalies, and communicate insights effectively. These measures distill raw data into meaningful summaries, forming the first critical step in any data analysis pipeline [5].

Core Measures

- **Mean:** The arithmetic average of a dataset, calculated as:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x$$

Although it can be influenced by extreme values, this measure is commonly applied when data is evenly distributed on both sides.

- **Median:** The median represents the middle value in a sorted dataset. Because it isn't affected by extreme values, it's especially useful for describing the center of data that's skewed or contains outliers:

Median = $\boldsymbol{x}(\frac{\boldsymbol{n+1}}{\boldsymbol{2}})$ if *n* odd

Median = $\frac{\boldsymbol{x}_{(\frac{\boldsymbol{n}}{\boldsymbol{2}})} + \boldsymbol{x}_{(\frac{\boldsymbol{n}}{\boldsymbol{2}}+\boldsymbol{1})}}{\boldsymbol{2}}$ if *n* even

- **Mode:** The most frequent value(s) in a dataset. Uniquely applicable to categorical data.
- **Variance:** Measures spread around the mean (population variance shown):

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x, -\mu)^2$$

Sample variance uses *n* 1 for unbiased estimation.

- **Standard Deviation:** The square root of variance is called the standard deviation. Standard deviation measures how much the values in a dataset typically differ from the mean, and it is expressed in the same units as the original data, making it easier to interpret than variance:

$$\sigma = \frac{\sqrt{}}{\sigma^2}$$

Provides spread in original data units.

Table 1: Descriptive Statistics Formulas and Applications

Measure	Formula	Use Case
Mean	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x$	Symmetric, continuous data
Median	Middle ordered value	Skewed data, outliers present
Mode	Most frequent value Σ	Categorical/ordinal data

Variance	$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x, -\mu)^2$	Quantifying data spread
Std Dev	$\sigma = \frac{\sqrt{}}{\sigma^2}$	Interpretable spread metric

The choice between these measures depends on data characteristics:

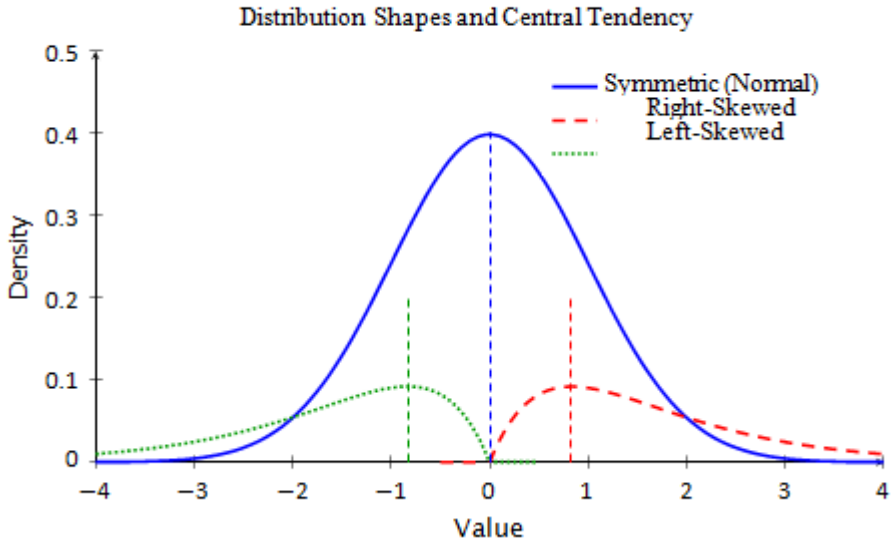

Figure 1: Comparison of symmetric (normal) and skewed distributions. Vertical dashed lines indicate the means of each distribution.

- For symmetric distributions without outliers (Fig.**??**, blue), mean and standard deviation suffice.
- Skewed distributions (Fig. **??**, red/green) require median and interquartile range.
- Multimodal distributions necessitate reporting all modes.

Real-world applications include

- Using mean income for policy-making in normally distributed populations
- Reporting median house prices in skewed real estate markets
- Analyzing mode of transportation preferences in urban planning

Understanding these metrics' strengths and limitations prevents misinterpretation.

For example, the 2023 U.S. Census Bureau reported a *mean* household income of $76,330 but a *median* of $61,980, highlighting income inequality's skewing effect [5].

III. PROBABILITY THEORY AND DISTRIBUTIONS

Probability theory underpins data science by offering tools to systematically address unpredictability and variability in real-world datasets. It enables quantification of the likelihood of events and forms the backbone of statistical inference, machine learning algorithms, and data-driven decision-making [6].

Basic Probability Rules

The fundamental rules of probability govern how we calculate the likelihood of combined events:

- **Addition Rule:** For events *A* and *B*, the probability of either event occurring is:
  - For mutually exclusive events: $P(A \cup B) = P(A) + P(B)$
  - For non-mutually exclusive events: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- **Multiplication Rule:** For the probability of both events occurring:
  - For independent events: $P(A \cap B) = P(A) \cdot P(B)$
  - For dependent events: $P(A \cap B) = P(A) \cdot P(B|A)$

These rules form the basis for more complex probability calculations and are essential for understanding statistical models [7].

Common Probability Distributions

Normal Distribution

The Normal distribution (also called Gaussian) is recognizable by its symmetrical, bell-shaped curve. Its shape and spread are determined by two key values: the mean (μ), which marks the center, and the standard deviation (σ), which quantifies variability.

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

The Normal distribution is central to many real-world phenomena such as measurement errors, natural variations in biological systems, and test scores. The Central Limit Theorem explains how combining numerous independent random variables results in a distribution resembling a bell curve (normal distribution). This principle is pivotal for statistical inference, as it allows analysts to draw reliable conclusions from sample data.

Binomial Distribution

The Binomial distribution describes the likelihood of achieving a certain number of successes across a set number of independent trials, where each trial has the same chance of success. It is defined by two parameters: *n*, representing the total number of trials, and *p*, the probability of success in each trial.

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

Binomial distributions are widely used in quality control, A/B testing, and modeling scenarios with binary outcomes.

Poisson Distribution

The Poisson distribution calculates the likelihood of a specific number of events happening within a defined timeframe or area, assuming these events occur at a consistent average rate and independently. Its single parameter, λ, serves as both the average (mean) and the measure of spread (variance) for the distribution.

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

It's often used to represent uncommon occurrences, like how many times a system fails, how often customers show up, or unexpected surges in network activity.

Probability Distribution Comparison

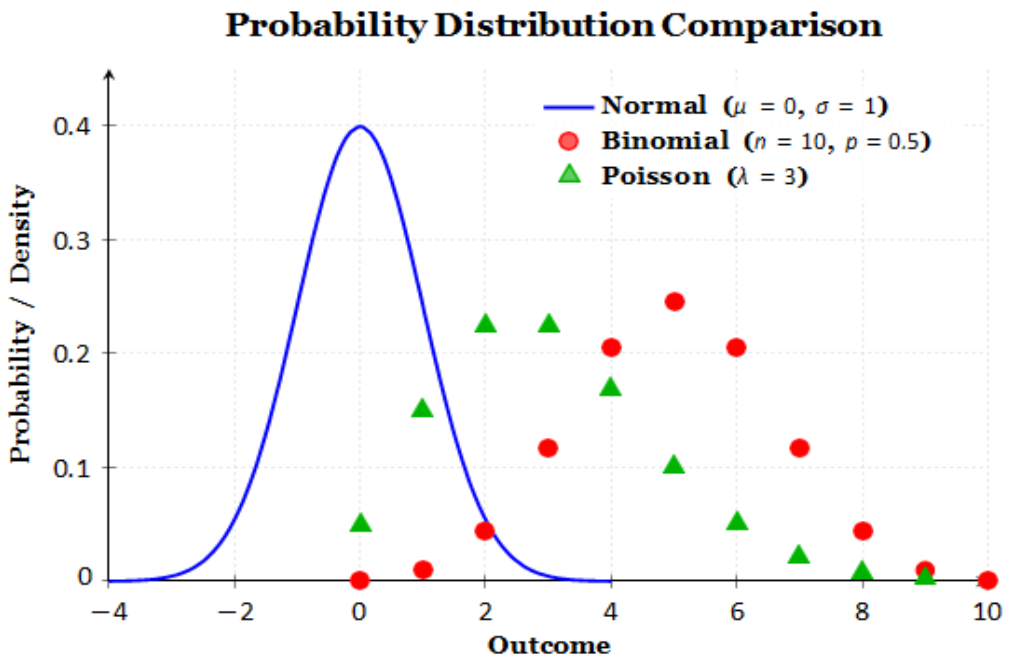

Figure 2: Visual comparison of Normal, Binomial, and Poisson distributions.

Table 2: Comparison of Key Probability Distributions

Distribution	Parameters	Applications	Properties
Normal	μ, σ	Height distributions Measurement errors Machine learning	Symmetric Bell-shaped 68-95-99.7 rule
Binomial	n, p	Quality control A/B testing Success/failure trials	Discrete Fixed trials Binary outcomes
Poisson	λ	Rare events Network traffic Server failures	Discrete Mean = Variance = λ Models count data

Applications in Data Science

Grasping these distributions is vital for data scientists, as they lay the groundwork for many analytical methods and decision-making processes:

- Hypothesis testing and confidence intervals
- Feature engineering and data transformation
- Anomaly detection and outlier identification
- Machine learning model selection and evaluation
- Simulation and risk modeling

For example, many machine learning algorithms rely on the Normal distribution, as they are built on the assumption that the input features follow a bell-shaped curve. The Binomial distribution is fundamental for classification problems with binary out- comes, while the Poisson distribution helps model rare events such as fraud detection or equipment failures.

Conclusion

Probability theory and distributions provide the mathematical framework necessary for data scientists to quantify uncertainty, make predictions, and draw reliable conclu- sions from data. By understanding the basic probability rules and key distributions, data scientists can develop more robust models and make more informed decisions based on their data.

IV. INFERENTIAL STATISTICS

Inferential statistics enables data scientists to draw conclusions about populations from sample data. This section covers hypothesis testing, confidence intervals, and common statistical tests used to make data-driven decisions [8].

Hypothesis Testing

Hypothesis testing uses data from samples to assess statements or assumptions about characteristics of a larger population. The process involves:

Key Concepts

- **p-value:** A p-value quantifies how likely it is to see your observed results assuming the null hypothesis (H_0) is true. Lower p-values (e.g., <0.05) imply stronger evidence to question the validity of H_0[9].
- **Type I Error (α):** A false positive happens when we mistakenly reject the null hypothesis, even though it's actually true.
- **Type II Error (β):** A false negative happens when a test fails to detect some- thing that is actually present-in statistics, this means not rejecting the null hypothesis even though it's false.

Confidence Intervals

A confidence interval gives us a range where we believe the true value for the whole population lies, based on our sample, and tells us how sure we are about that estimate (for example, 95% confident).

For a sample mean:

$$\mathrm{CI} = \bar{x} \pm t^* \quad \frac{s}{\sqrt{n}}$$

Here, z^* is the special value from the standard normal distribution that matches your chosen confidence level. It helps set how wide your confidence interval will be.

Common Statistical Tests

Table 3: Comparison of Statistical Tests

Test	Purpose	Data Type	Hypotheses	Assumptions
t-test	Compare two means	Continuous	$H_0 : \mu_1 = \mu_2$	Normality, equal variances
ANOVA	Compare ≥3 means	Continuous	$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$	Homogeneity of variance
Chi-square	Test independence	Categorical	H_0 : No association	Expected counts ≥5

Applications in Data Science

- Validate A/B test results using t-tests or ANOVA
- Assess feature significance in regression models
- Check dataset representativeness through confidence intervals
- Evaluate classification models using chi-square tests [10]

Ethical Considerations

Modern Practices Emphasize

- Reporting effect sizes alongside p-values
- Using confidence intervals for clinical significance
- Addressing multiple comparison issues
- Pre-registering hypotheses to prevent p-hacking [11]

V. REGRESSION ANALYSIS

Regression analysis is a key statistical method used to explore how one variable depends on one or more other variables. The two most frequently used types are linear regression, which is suited for predicting continuous outcomes, and logistic regression, which is used when the outcome is categorical, like yes/no or true/false [12].

Linear Regression

Linear regression helps us understand how two variables are connected by drawing a straight line that best fits the data points we've collected. When there's just one factor influencing the outcome, the model is called simple linear regression and takes the following form:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

In this model, y stands for the value we want to predict, while x is the variable we use to make that prediction. The term β_0 represents the point where the line crosses the y-axis, and β_1 shows how much y changes for each unit increase in x. The symbol ε accounts for any random errors or differences that the model can't explain. This approach is designed for situations where the outcome is a continuous value, and it works by finding the line that keeps the overall prediction errors as small as possible.[13].

Logistic Regression

Unlike linear regression, logistic regression is used when we want to predict one of two possible outcomes. It does this by estimating the chance that the outcome falls into a specific category. The formula for the logistic model looks like this:

$$P(Y = 1/x = \ x) = \frac{e^{\beta 0 + \beta 1 x}}{1+e^{\beta 0 + \beta 1 x}} = \frac{1}{1+e^{-(\beta 0 + \beta 1 x)}}$$

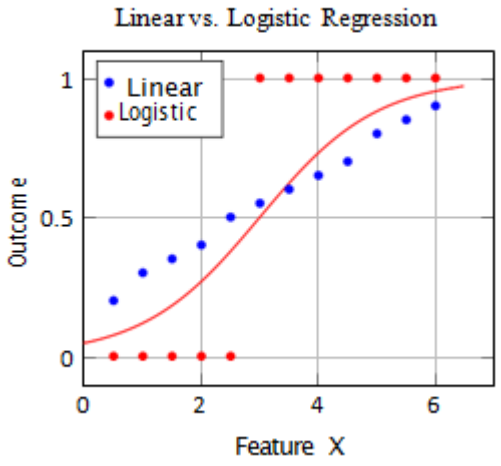

Figure 4: Comparison of linear and logistic bregression models.

Table 4: Regression Model Comparison

Characteristic	Linear Regression	Logistic Regression
Outcome Type	Continuous	Binary/Categorical
Cost Function	Sum of Squared Errors	Log-likelihood
Use Cases	• Sales prediction • Price estimation • Temperature modeling	• Fraud detection • Disease diagnosis • Email spam filtering
Interpretation	Direct effect on outcome value	Effect on log-odds of outcome

VI. BAYESIAN STATISTICS BASICS

Bayesian statistics offers a way to refine your understanding as new data comes in, using probability principles. It centers around Bayes' theorem-a powerful tool that flips the script on conditional probabilities, letting you estimate how likely a hypothesis is after observing relevant evidence[14].

Bayes' Theorem

The mathematical formulation of Bayes' theorem is:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

In the context of statistical inference, this becomes:

$$P(\vartheta|data) = \frac{P(data|\vartheta) \cdot P(\vartheta)}{P(data)}$$

Where:

- $P(\theta|data)$ is the posterior probability of the parameters given the data
- $P(data|\theta)$ is the likelihood of observing the data given the parameters
- $P(\theta)$ is the prior probability of the parameters
- $P(data)$ is the marginal likelihood or evidence

Prior, Likelihood, and Posterior

The prior distribution reflects what we assume or believe about the parameters before we look at any data. It's our starting point, based on previous knowledge or intuition, before new evidence is considered. It can be informative (based on previous knowledge) or non-informative (minimally structured) [15].The likelihood acts like a "fit meter" in statistics-it measures how well different parameter settings align with the data you've observed. Think of it as a way to rank which parameter values make your dataset most plausible. The posterior distribution combines prior beliefs with the likelihood, representing updated beliefs after observing data.

When the denominator P (*data*) is difficult to compute, we often use the proportional form:

Posterior $\propto$ Likelihood × Prior

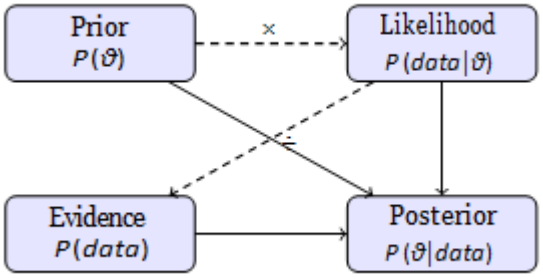

Figure 5: Bayesian inference workflow showing how prior beliefs are updated with new data.

Bayesian methods differ from frequentist approaches by treating parameters as random variables with probability distributions rather than fixed values [16]. This allows quantifying parameter uncertainty through probability statements and incorporating prior knowledge into the analysis.

VII. PRACTICAL EXAMPLES

This section demonstrates how to perform common statistical analyses in Python and R, including calculating descriptive statistics, conducting a t-test, and fitting a linear regression model. These examples use widely adopted libraries such as pandas, scipy, and statsmodels in Python, and base functions in R.

Calculating Descriptive Statistics

Python:
Listing 1 Descriptive statistics in Python

```
import pandas as pd

data = [ 2 , 4 , 6 , 8 , 1 0 ]
s e r i e s = pd . S e r i e s (data)

mean = s e r i e s . mean ()
median = s e r i e s . median ()
std = s e r i e s . std ()

print ( "Mean : " , mean)
print ( "Median : " , median )
print ( " Std ␣ Dev : " , std )
```

R:
Listing 2 Descriptive statistics in R

```
data <- c ( 2 , 4 , 6 , 8 , 10 )

mean_v a l <- mean( data )
median_v a l <- median ( data )
std_val <- sd ( data )

cat ( "Mean : " , mean_v al , "\n" )
cat ( "Median : " , median_v al , "\n" )
cat ( " Std  Dev : " , std_val , "\n" )
```

Performing a t-test

Python:

Listing 3 t-test in Python

```
from s c i p y . s t a t s import ttest_ind

group1 = [ 5 , 7 , 8 , 9 , 1 0 ]
group2 = [ 6 , 6 , 7 , 8 , 1 2 ]

t_stat , p_value = ttest_ind ( group1 , group2 )
print ( " t-s t a t i s t i c : " , t_stat )
print ( "p-value : " , p_value )
```

R:

Listing 4 t-test in R

```
group1 <- c ( 5 , 7 , 8 , 9 , 10 )
group2 <- c ( 6 , 6 , 7 , 8 , 12 )

t . t e s t ( group1 , group2 )
```

Fitting a Linear Regression

Python:

Listing 5 Linear regression in Python

```
import s ta t s m o d e l s . api as sm

X = [ 1 , 2 , 3 , 4 , 5 ]
y = [ 2 , 4 , 5 , 4 , 5 ]

X = sm . add_constant (X)       # Adds  i n t e r c e p t  term
model = sm . OLS( y , X) . f i t ()
print ( model . summary ( ) )
```

R:

Listing 6 Linear regression in R

```
X <- c ( 1 , 2 , 3 , 4 , 5 )
y <- c ( 2 , 4 , 5 , 4 , 5 )
model <- lm( y ~ X)
summary(model )
```

VIII. COMMON PITFALLS

Despite the power and utility of statistical methods, several common pitfalls can undermine the validity of data science projects. Awareness of these issues is crucial for conducting rigorous analysis and drawing sound conclusions.

P-Hacking and Multiple Testing

P-hacking (also known as data dredging) occurs when researchers analyze data multiple ways until reaching statistical significance, without accounting for multiple comparisons. This dramatically increases Type I error rates. For example, testing 20 hypotheses at = 0.05 yields approximately a 64% chance of finding at least one "significant" result purely by chance [17].

Overfitting

Overfitting occurs when a model learns the random quirks and noise in the training data instead of the actual trends, which makes it perform poorly on new, unseen data. This issue is especially common with models that are too complicated or have too many parameters compared to the amount of data available. Cross-validation and regularization techniques help mitigate overfitting by assessing model performance on unseen data.

Misinterpretation of Confidence Intervals

A 95% confidence interval doesn't mean there's a 95% chance the true value lies within your calculated range. Instead, it means that if you repeated the same process (sampling and analysis) countless times, approximately 95% of those intervals would contain the actual parameter. This subtle distinction is frequently misunderstood and can lead to incorrect interpretations.

Table 5: Common Statistical Pitfalls and Their Solutions

Pitfall	Consequences	Solutions
P-Hacking	Inflated false positive rate, non-reproducible findings	Pre-register hypotheses, adjust for multiple comparisons (e.g., Bonferroni, FDR)
Overfitting	Poor model generalization, illusory predictive power	Cross-validation, regularization techniques (L1/L2), simpler models
Misinterpreting Confidence Intervals	Incorrect probability statements, over-confidence in results	Focus on repeated sampling interpretation, use Bayesian credible intervals
Publication Bias	Skewed literature with overestimated effects	Pre-registration, reporting negative results, meta-analysis with funnel plots

To maintain statistical integrity, data scientists should implement robust practices such as pre-registering hypotheses, using validation sets, employing appropriate corrections for multiple testing, and carefully interpreting statistical outputs.

IX. EXERCISES

Theoretical Questions

1. **Confidence Interval Calculation:** Forty students took a test and their average score was 78, with scores typically varying by 10 points. Based on this information, how can you estimate a range (with 95% confidence) that likely includes the true average score for all students?

2. **Probability Distribution:** f you flip a fair coin 10 times, what are the chances that you'll get exactly 6 heads? What type of probability distribution would you use to solve this, and how would you work out the answer?

3. **Hypothesis Testing:** A company advertises that their latest batteries have an average lifespan exceeding 500 hours. To verify this, a random test of 25 batteries showed an average lifespan of 520 hours, with individual results varying by about 40 hours. Using a 5% significance level (95% confidence), does this data provide enough evidence to back the company's claim?

Case Study

Case Study: Website Redesign A/B Test

An online retailer has rolled out a fresh website look and wants to find out if it actually encourages more people to buy. To test this, they randomly split 2,000 visitors: half saw the original site, and half saw the new one. Out of the 1,000 people who saw the old design, 120 made a purchase. In the group that saw the new design, 150 out of 1,000 ended up buying something.

What you need to do:

Set up two hypotheses: one stating the new design has no effect, and another claiming it boosts purchases.

- Pick the best statistical test for this comparison and actually run the numbers.
- Determine the p-value and explain whether it's strong enough to trust at the 95% confidence level.
- Wrap it up: Does the data show the new design really works?

REFERENCES

[1] Sachdeva, S.: Statistics. LNA Books, Delhi, India (2024). For B.Com., B.A., B.B.A., M.Com., M.B.A. and other professional and competitive examinations.
[2] Steps, A.: Importance of Statistics and Probability in Data Sci- ence. Accessed: 2025-04-26. https://www.analyticssteps.com/blogs/ importance-statistics-and-probability-data-science
[3] Skills, P.: A/B Testing in Data Science [Using Python]. Accessed: 2025-04-26. https://pwskills.com/blog/a-b-testing-in-data-science-using-python/

[4] Data Science, M.: What Is Probability Theory? Accessed: 2025-04-26. https://www.mastersindatascience.org/learning/statistics-data-science/ probability-theory/
[5] Illowsky, B., Dean, S.: Introductory Statistics. OpenStax College, ??? (2013). Pages 78-112
[6] Institute of Data: What Is Probability Theory in Data Sci- ence? Accessed: 2025-04-26. https://www.institutedata.com/blog/ what-is-probability-theory-in-data-science/
[7] Quality Gurus: Probability: Rule of Addition and Multiplication. Accessed: 2025-04-26. https://www.qualitygurus.com/ probability-rule-of-addition-and-multiplication/
[8] Illowsky, B., Dean, S.: Introductory Statistics. OpenStax, ??? (2023). Chapter 9-11
[9] Science, T.D.: Hypothesis Testing in Python. Accessed: 2025-04-26. https:// towardsdatascience.com/hypothesis-testing-in-python-4a7d0f8d169a
[10] StatsDirect: Chi-Square Test Applications. Accessed: 2025-04-26. https://www. statsdirect.com/help/chi_square_tests/chi_square.htm
[11] Psychology, S.: Type I and II Errors in Research. Accessed: 2025-04-26. https://www.simplypsychology.org/type_i_and_type_ii_errors.html
[12] James, G., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning: with Applications in R, 2nd edn. Springer, ??? (2021)
[13] Wasserman, L.: All of Statistics: A Concise Course in Statistical Inference. Springer, New York (2023)
[14] Faulkenberry, T.J.: Bayesian Statistics: The Basics. Routledge, ??? (2025)
[15] Team, F.E.: Bayesian inference: more than Bayes's theorem. Frontiers in Astron- omy and Space Sciences **11**, 1326926 (2024) https://doi.org/10.3389/fspas.2024.1326926
[16] Wikipedia: Bayesian Statistics. Accessed: 2025-04-26. https://en.wikipedia.org/ wiki/Bayesian_statistics
[17] Head, M.L., Holman, L., Lanfear, R., Kahn, A.T., Jennions, M.D.: The extent and consequences of p-hacking in science. PLOS Biology **13**(3), 1002106 (2015)