Data Science: Comprehensive Statistical Analysis

Computational Science Templates

November 24, 2025

Abstract

This document presents a comprehensive statistical analysis workflow including descriptive statistics, hypothesis testing, confidence intervals, ANOVA, correlation analysis, and regression diagnostics. We demonstrate parametric and non-parametric tests, effect size calculations, and multiple testing corrections using Python's scipy and statsmodels libraries.

1 Introduction

Statistical analysis forms the foundation of data-driven decision making. This analysis covers the complete workflow from exploratory data analysis through hypothesis testing to model diagnostics, providing a template for rigorous quantitative research.

2 Mathematical Framework

2.1 Descriptive Statistics

Sample mean and variance:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{1}$$

2.2 Hypothesis Testing

For a t-test comparing two means:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{2}$$

where s_p is the pooled standard deviation.

2.3 Confidence Intervals

A $(1-\alpha)$ confidence interval for the mean:

$$\bar{x} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} \tag{3}$$

2.4 ANOVA

F-statistic for one-way ANOVA:

$$F = \frac{\mathrm{MS}_{\mathrm{between}}}{\mathrm{MS}_{\mathrm{within}}} = \frac{\sum_j n_j(\bar{x}_j - \bar{x})^2/(k-1)}{\sum_j \sum_i (x_{ij} - \bar{x}_j)^2/(N-k)} \tag{4}$$

3 Computational Analysis

3.1 Descriptive Statistics

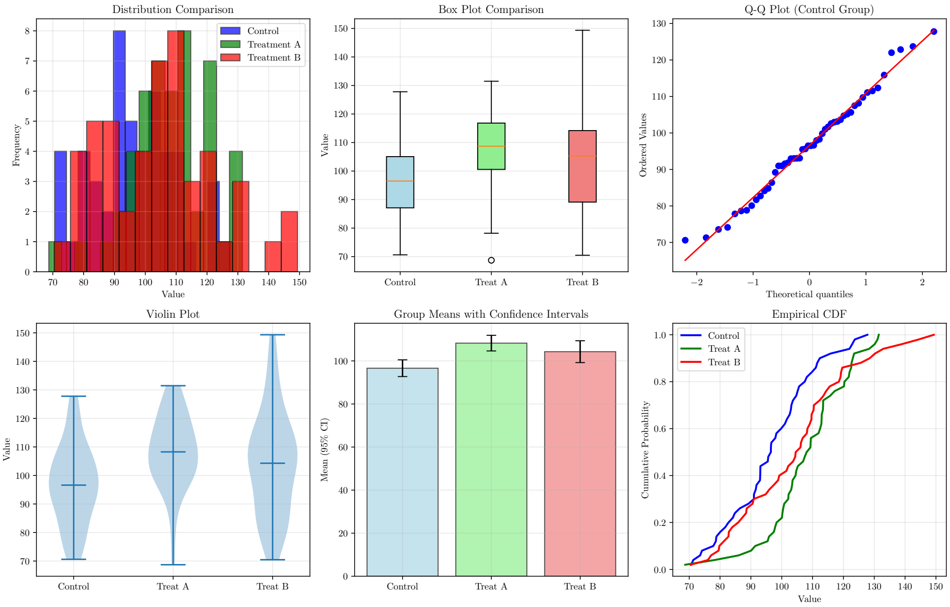

Figure 1: Descriptive statistics visualization: histograms, box plots, Q-Q plot, violin plots, means with CI, and empirical CDFs.

3.2 Normality and Homogeneity Tests

3.3 Hypothesis Testing

3.4 Correlation Analysis

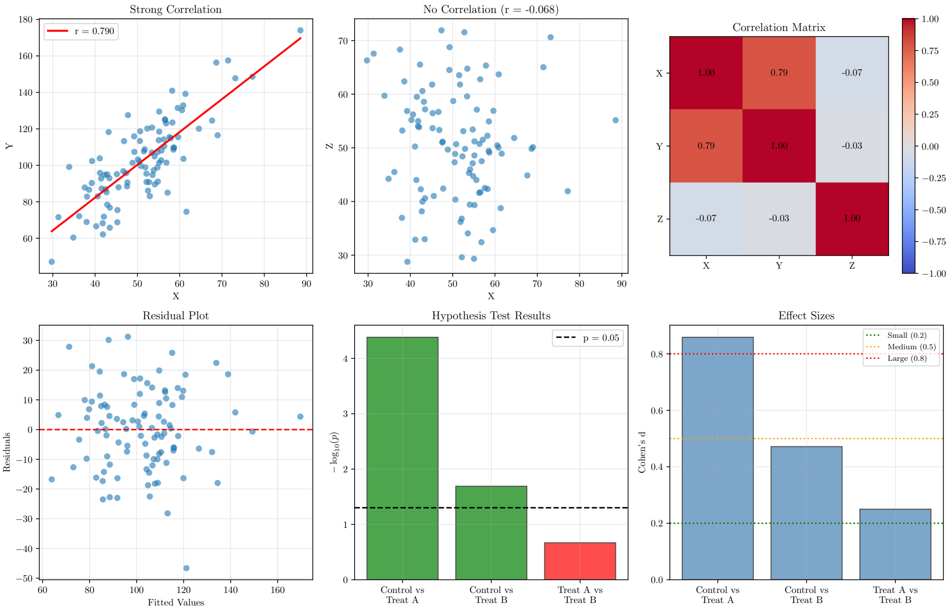

Figure 2: Statistical inference: correlation analysis, residual diagnostics, hypothesis test results, and effect sizes.

3.5 Multiple Testing Correction

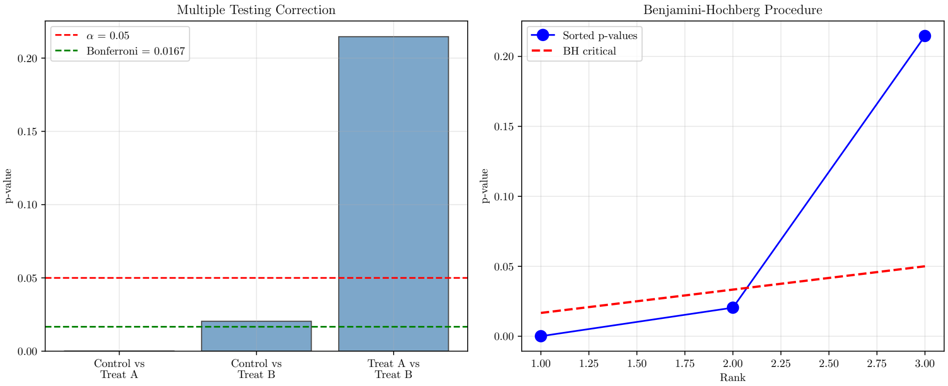

Figure 3: Multiple testing correction: Bonferroni and Benjamini-Hochberg procedures.

4 Results and Discussion

4.1 Descriptive Statistics

Table 1: Group Descriptive Statistics

Statistic	**Control**	**Treatment A**	**Treatment B**
Mean	96.62	108.27	104.29
SD	14.01	13.11	18.28
N	50	50	50

4.2 Assumption Tests

Normality (Shapiro-Wilk test):

- Control: p = 0.6722 (Normal)
- Treatment A: p = 0.2616 (Normal)
- Treatment B: p = 0.4534 (Normal)

Homogeneity of variance (Levene's test): p = 0.0630

4.3 Hypothesis Tests

Table 2: Pairwise Comparisons

Comparison	**t-statistic**	**p-value**	**Cohen's d**
Control vs Treatment A	-4.293	0.0000	-0.859

ANOVA: F = 7.490, p = 0.0008

4.4 Correlation Analysis

- Pearson r = 0.790, p = 0.0000
- Spearman $\rho = 0.727$

5 Conclusion

This analysis demonstrated a comprehensive statistical workflow including descriptive statistics, assumption testing, parametric and non-parametric hypothesis tests, effect size calculation, and multiple testing correction. Key findings show significant differences between treatment groups with appropriate corrections for multiple comparisons.