Lab: Evaluating Big O Notation for Basic Statistical Methods

1 Introduction

In this lab, you will explore and analyze the computational complexity (Big O notation) of several basic statistical methods. Understanding the time complexity of algorithms is crucial for efficient data analysis, especially with large datasets commonly encountered in statistical work.

2 Background

Big O notation describes the upper bound of an algorithm's running time, providing a measure of its efficiency as the input size grows. It abstracts away constants and lower-order terms to focus on the dominant factor affecting the growth rate.

3 Tasks

For each statistical method, you will:

1. **Implement** the method in a programming language of your choice (e.g., Python, R).
2. **Guess** the theoretical time complexity (Big O notation).
3. **Generate** datasets of varying sizes. (perhaps $n = 10^4, 10^5, 10^6, 10^7$.)
4. **Measure** the execution time for each dataset size.

5. **Plot** execution time against dataset size.

6. **Compute** time complexity using empirical resultes

3.1 General Instructions

- Use timing functions appropriate for your programming language to measure execution times accurately.
- Run each computation multiple times and take the average to mitigate variability.
- Ensure that your code is optimized to reflect the true complexity of the algorithms, avoiding unnecessary overhead.

4 Statistical Methods

4.1 Mean Calculation

Compute the mean of a dataset $\{x_1, x_2, \ldots, x_n\}$:

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

4.2 Variance Calculation

Compute the variance:

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

4.3 T-Test Statistic for Difference of Means

For two independent samples $\{x_1\}$ and $\{x_2\}$:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

4.4 Correlation Coefficient

Compute Pearson's correlation coefficient between two variables $\{x_i\}$ and $\{y_i\}$:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

4.5 Linear Regression (Matrix Perspective)

Compute the regression coefficients using:

$$\hat{\beta} = (X'X)^{-1}X'y$$

Where:

- X is an $n \times p$ matrix of predictors.
- y is an $n \times 1$ vector of responses.

Hint: Computing $(X'X)^{-1}$ takes inverting a $p \times p$ matrix takes $O(p^3)$, and computing $(X'X)^{-1}X'y$ takes $O(p^2n)$. If p is small compared to n, the dominant term is $O(np^2)$.

4.6 Principal Component Analysis (PCA)

Compute the principal components of a dataset:

1. **Standardize** the dataset if necessary.
2. **Compute** the covariance matrix C:

   $$C = \frac{1}{n-1}X'X$$

   where X is the data matrix of size $n \times p$.
3. **Perform** eigenvalue decomposition on C to find eigenvalues and eigenvectors.
4. **Project** Use the first k eigenvalues as predictors in a linear regression model

Hint: There are multiple steps here. Big O notation focuses on the most complex term.