Mathematics For Computer Science Engineers

UE24MA242A

Teaching Assistant: **KANTHI HEGDE**

Problem Set: Parkinsons Disease Progression Analysis

Parkinson’s Disease Progression Dataset Description

The Parkinson’s Disease Progression Dataset contains simulated voice-based biomarkers, demographic details, and neurological indicators associated with Parkinson’s disease patients and healthy individuals. Each row represents a voice recording session with key acoustic features known to correlate with motor deterioration and neurological impairment.
This dataset enables advanced data analysis tasks such as exploratory data analysis, statistical inference, and predictive modeling — helping students uncover patterns in disease biomarkers and voice degradation severity.

Background
Parkinson’s disease leads to progressive loss of motor control, affecting speech clarity, pitch stability, and breathing patterns. Modern clinical screening increasingly uses voice-based features because vocal tremors appear early even before motor impairments become visible.
This dataset supports investigation into questions like:

- Do Parkinson’s patients show lower harmonic-to-noise ratio (HNR)?
- Does jitter/shimmer rise with disease severity?
- Can we predict disease status using voice features?

Dataset Overview

Each row = one recorded voice sample
~650 observations, ~12 attributes

Feature	Description
Age	Patient age (years)
Sex	Male / Female
MDVP:Fo(Hz)	Average vocal fundamental frequency
MDVP:Fhi(Hz)	Maximum fundamental frequency
MDVP:Flo(Hz)	Minimum vocal frequency
Jitter(%)	Frequency variation measure
Shimmer	Amplitude variation measure
NHR	Noise-to-harmonic ratio
HNR	Harmonic-to-noise ratio
RPDE	Vocal signal complexity
DFA	Signal fractal scaling exponent

Feature	Description
Status	1 = Parkinson’s, 0 = Healthy

Unit 1: Exploratory Data Analysis & Preprocessing

1. Dataset Feature Identification

- List column names and data types
- Confirm appropriateness of data types
- Show first 5 and last 5 rows

(Hint: df.dtypes, df.head(), df.tail())

2. Data Quality & Cleaning

- Check missing values, duplicates
- Report missing values per column
- Write cleaning strategy (drop/impute/flag)
- Show shape before & after cleaning

3. Descriptive Statistics

Compute and interpret:

- mean, median, SD, range for
  Age, MDVP:Fo(Hz), Jitter(%), HNR

Discuss:

- Which measure represents each variable best and why?
  (e.g., jitter tends to skew → median > mean)

4. Visual Data Exploration

a) Histograms — Jitter(%) & HNR
b) Boxplot — HNR across Status groups
Explain:

- distribution shape
- spread
- presence of outliers

(**Bonus**) Remove outliers (IQR method) and re-plot

5. Outlier Detection & Fix

Apply IQR to:

- MDVP:Fo(Hz)
- Shimmer

Plot boxplots **before & after** removal
Describe impact on data spread

6. Normality Check (Q-Q Plot)

Q-Q plot for **Jitter(%)**

- Does it follow normal distribution?
- Comment on deviations & impact on t-test usage

7. Correlation Heatmap

- Plot correlation matrix for all numeric features
- Identify feature most strongly correlated with **Status**
- Interpret relationship in simple language

Example:
Higher jitter values strongly correlate with Parkinson’s presence.

Unit 2: Confidence Interval and MOE

8. Confidence Interval for Jitter(%)
Calculate **95% CI** for mean Jitter(%)
Interpretation example:
We are 95% confident the mean jitter for the population lies between X and Y.

9. Margin of Error vs Confidence Level
Calculate ME for HNR at:

- 90%
- 95%
- 99%

Explain how higher confidence = wider interval (less precision)

Unit 3: Hypothesis Testing & Inference

10. Gender-Based Vocal Differences
Test if **mean HNR differs by gender**

- H0: No difference in mean HNR between genders
- H1: There is a difference
- Apply Two-Sample t-test (or Mann-Whitney if non-normal)
- Interpret p-value

11. Jitter vs Disease Status Correlation

- Compute Pearson/Spearman between Jitter(%) and Status
- Scatter plot + fitted regression line

Interpretation:
Higher jitter increases likelihood of Parkinson's.

12. Linear Regression Model
Predict HNR using: Jitter(%)

Tasks:
- Fit a simple linear regression model
- Provide the regression equation
($HNR = \beta_0 + \beta_1 \times Jitter(\%)$)
- Interpret coefficients in plain English
- Report model accuracy metric