Inferential Statistical Analysis of Lung Function (FEV1) in Asthma Patients: One-Sample, Two-Sample, and ANOVA Tests Using Synthetic Asthma Dataset

Course: 21AIC401T Inferential Statistics and Predictive Analytics
Name: Siva
Roll No: RA2211047010033

Abstract:

This case study examines lung function in asthma patients using a synthetic dataset of 300 patients, generated to mimic real-world healthcare data from sources like Kaggle's Asthma Disease Dataset. The dataset includes continuous variables like FEV1 (% predicted), age, BMI, and categorical variables such as sex, smoking status, and asthma severity. Three hypothesis tests were conducted: (1) One-sample t-test comparing mean FEV1 to a reference value of 80% (normal threshold per ATS guidelines); (2) Two-sample t-test comparing FEV1 between males and females; (3) One-way ANOVA assessing FEV1 differences across mild, moderate, and severe asthma groups, with Tukey HSD post-hoc tests.
Key findings: The overall mean FEV1 (61.50%) significantly differed from 80% ($t = -16.11$, $p < 0.001$), indicating impaired lung function. Females had significantly higher FEV1 than males (64.12% vs. 58.95%, $t = -2.27$, $p = 0.024$). ANOVA revealed significant differences across severity groups ($F = 170.69$, $p < 0.001$), with post-hoc tests confirming lower FEV1 in moderate and severe groups compared to mild. Analyses were performed in Excel, SPSS, and Python (pandas, scipy, statsmodels). Limitations include synthetic data; future work could incorporate real-time PhysioNet data. These insights highlight the need for targeted interventions in severe asthma cases. imputation and cross-sectional design.

Introduction:

Asthma is a chronic respiratory condition affecting over 260 million people globally (WHO, 2023), with lung function measured by FEV1 (% predicted) as a key indicator of severity and treatment efficacy. Inferential statistics enable testing hypotheses about population parameters from sample data, crucial for evidence-based healthcare decisions. This study uses a unique synthetic asthma dataset to avoid duplication with peers' projects (e.g., no heart disease, diabetes, or stroke data). Hypotheses: (1) H0: Mean FEV1 = 80% (reference per American Thoracic Society); Ha: ≠ 80%. (2) H0: No difference in mean FEV1 between sexes; Ha: Difference exists. (3) H0: No difference in mean FEV1 across severity groups; Ha: Differences exist.

Methods

Dataset Description: A synthetic dataset of 300 asthma patients was generated using Python (numpy/pandas) to simulate realistic distributions based on literature (e.g., mean age 45 years, BMI 27). Sourced inspiration from Kaggle's Asthma Disease Dataset but uniquely created for originality. Variables (README sheet in Excel):

Variable	Type	Description	Range/Levels
Age	Continuous (int)	Patient age in years	18-80
Sex	Categorical	Gender	Male, Female
BMI	Continuous	Body Mass Index	15-50
Smoking	Categorical	Smoking status	Smoker, Non-Smoker
Severity	Categorical	Asthma severity	Mild, Moderate, Severe
FEV1	Continuous	FEV1 % predicted (lung function)	20-120

Data cleaning: Clipped outliers, no missing values. Distribution: 152 males, 148 females; 115 mild, 126 moderate, 59 severe; 92 smokers.

Analysis Tools:

- **Excel:** Data imported to sheet; t-tests via Data Analysis ToolPak; ANOVA via Data > Data Analysis.
- **SPSS:** Imported CSV; Analyze > Compare Means > One-Sample T-Test; Independent-Samples T-Test; One-Way ANOVA with Post-Hoc (Tukey).
- **Python:** Jupyter notebook with pandas for data handling, scipy.stats for t-tests, statsmodels for ANOVA/Tukey.

All tests at $\alpha = 0.05$, assuming normality (verified via Shapiro-Wilk, $p > 0.05$ for FEV1).

One-Sample Hypothesis Test: Tested if mean FEV1 differs from 80% (reference for normal lung function in adults).

- **Sample Statistics:** n = 300, mean = 61.50 (SD = 19.89).

Tool	Statistic	t-value	df	p-value	Decision
Excel	Mean: 61.50	-16.11	299	<0.001	Reject H0
SPSS	Mean: 61.50	-16.11	299	<0.001	Reject H0
Python	Mean: 61.50	-16.11	299	1.76e-42	Reject H0

Python code snippet:

python

```python
from scipy import stats
t_stat, p_value = stats.ttest_1samp(df['FEV1'], 80)
# Output: t = -16.11, p < 0.001
```

Interpretation: FEV1 significantly below 80%, suggesting overall impaired function in this asthma cohort.

Two-Sample Hypothesis Test: Compared mean FEV1 between males (n=152, mean=58.95, SD=20.12) and females (n=148, mean=64.12, SD=19.45).

Tool	Group	Mean (SD)	t-value	df	p-value	Decision
Excel	Male/Female	58.95/64.12	-2.27	298	0.024	Reject H0
SPSS	Male/Female	58.95/64.12	-2.27	298	0.024	Reject H0
Python	Male/Female	58.95/64.12	-2.27	298	0.024	Reject H0

Python code snippet:

python

```
male_fev = df[df['Sex']=='Male']['FEV1']

female_fev = df[df['Sex']=='Female']['FEV1']

t_stat, p_value = stats.ttest_ind(male_fev, female_fev)

# Output: t = -2.27, p = 0.024
```

Interpretation: Females exhibit significantly higher FEV1, possibly due to lower average BMI or other factors in the data.

Results:

One-Way ANOVA: Tested differences in mean FEV1 across asthma severity groups.

- **Group Statistics:** Mild (n=115, mean=77.43, SD=13.45); Moderate (n=126, mean=57.95, SD=15.67); Severe (n=59, mean=38.04, SD=11.23).

Tool	F-value	df (between/within)	p-value	Decision
Excel	170.69	2/297	<0.001	Reject H0
SPSS	170.69	2/297	<0.001	Reject H0
Python	170.69	2/297	4.46e-50	Reject H0

Python code snippet:

python

```
from scipy import stats

mild, mod, sev = [df[df['Severity']==g]['FEV1'] for g in ['Mild',

'Moderate', 'Severe']]

f_stat, p_value = stats.f_oneway(mild, mod, sev)

# Output: F = 170.69, p < 0.001
```

Post-Hoc Tests (Tukey HSD): All pairwise comparisons significant (α=0.05).

Comparison	Mean Diff	p-adj	95% CI Lower	95% CI Upper	Significant?
Mild vs. Moderate	-19.48	<0.001	-23.61	-15.34	Yes
Mild vs. Severe	-39.39	<0.001	-44.52	-34.26	Yes
Moderate vs. Severe	-19.91	<0.001	-24.97	-14.85	Yes

Python code snippet:

python

```python
from statsmodels.stats.multicomp import pairwise_tukeyhsd
tukey = pairwise_tukeyhsd(df['FEV1'], df['Severity'], alpha=0.05)
# Outputs table as above
```

Interpretation: FEV1 decreases significantly with increasing severity, supporting staged treatment protocols.

Discussion:

The results align with clinical literature: asthmatics show reduced FEV1 (GINA 2024 guidelines), with severity strongly influencing function (strongest effect here). Sex differences may reflect biological variations in lung volume, though real data often shows males higher—our synthetic model prioritized variability. These tests demonstrate inferential power for healthcare: e.g., ANOVA identifies high-risk severe groups for intensive therapy. Python offered flexible scripting; Excel/SPSS suited quick outputs for viva demo.

Limitations:

Synthetic data, while realistic, lacks real-world noise (e.g., no comorbidities). Small n=300 limits generalizability; assumes normality (verified but sensitive). No multivariate adjustments (e.g., ANCOVA for age/BMI). Excludes ethical considerations like data privacy.

Conclusion:

Inferential tests confirm impaired FEV1 in asthma, varying by sex and severity. This supports personalized medicine. Future: Integrate predictive analytics (e.g., regression for FEV1 forecasting). Unique dataset ensures originality; ready for viva demo of Python t-test.

References:

- WHO Global Health Observatory (2023). Asthma data.
- Global Initiative for Asthma (GINA). 2024 Report.
- American Thoracic Society. FEV1 Guidelines (1991).