DS3030: Data Analytics Lab
Assignment 13

Date: Nov 10, 2025

Timing: 2:00 to 4:30 PM Max marks: 12

Instructions

- Submit .ipynb file named as
[student name]_assignment[number]_part[number].ipynb
containing all questions (text blocks) and solutions (code blocks)
- Write **justifications/comments** as required.

Part I: F-test & t-test [4 Marks]

DATASET: "ds_salaries.csv"

Question 1 [2 Marks]: A company wants to check whether the salary variability between Data Scientists and Data Engineers is the same. Use the dataset available at ds_salaries.csv, which contains the columns job_title and salary_in_usd. Perform an F-test for equality of two variances at a 5% significance level ($\alpha = 0.05$).
Tasks:

- State the Null and Alternative Hypotheses. [0.5]
- Perform the F-Test on the salaries of 'Data Scientists' and 'Data Engineers' using stats.levene. [0.5]
- Report the F-statistic obtained from the test. [0.5]
- Write your interpretation based on the result — clearly mention whether you reject or fail to reject H_0, and what it means about the salary variability between the two roles. [0.5]

Question 2 [2 Marks]: It is claimed that the average annual salary of all employees listed in the dataset is $100,000. Using the dataset 'ds_salaries.csv', test this claim using a one-sample t-test with the 'salary_in_usd' column. Use a significance level of $\alpha = 0.05$.
Tasks:

- State the null and alternative hypotheses. [0.5]
- Perform the one-sample t-test using scipy.stats.ttest_1samp. [0.5]

- Report the t-statistic. [0.5]
- Write your interpretation of the result, indicating whether the sample supports or contradicts the claim. [0.5]

Part II: Comprehensive Z-Test Analysis [4 Marks]

DATASET: Employee Performance & Productivity Dataset

You are a data analyst working for a large corporation. The HR department has asked you to conduct statistical tests on employee data to verify certain claims and compare different employee groups. Use the provided Employee Performance & Productivity Dataset to answer the following:

Part A: Single Sample Z-Test (2.0 marks)
The company's executive team claims that:

- The average *Performance_Score* of employees is 3.0.
- The average *Employee_Satisfaction_Score* is 3.2.

Test both claims at a 5% significance level. Assume:

- Population standard deviation for *Performance_Score* = 1.41
- Population standard deviation for *Employee_Satisfaction_Score* = 1.15

Tasks:

- For each claim, state the null and alternative hypotheses
- Calculate the sample mean, z-statistic.
- Make a decision (Reject/Fail to Reject H_0) for each test.
- Provide a summary table with results for both tests

Part B: Two-Sample Z-Test (2.0 marks)
The HR department wants to investigate:

1. Whether there is a significant difference in average *Work_Hours_Per_Week* between employees who have Resigned (True) vs those who have Not Resigned (False). Use $\alpha = 0.05$.

2. Whether there is a significant difference in the average *Monthly_Salary* between Male and Female employees. Use $\alpha = 0.05$.

Assume population standard deviations:

- Work_Hours_Per_Week: 8.94 hours for both groups
- Monthly_Salary: $1,370 for both groups

Tasks:

- For each comparison, state the null and alternative hypotheses.
- Calculate sample means for both groups, z-statistic.
- Make a decision (Reject/Fail to Reject H_0) for each test.
- Provide a summary table with results for both tests.

Part III: Chi-Squared Test [4 Marks]

OBJECTIVE: Chi-Square Test of Independence for analyzing how certain key features of the dataset are related to social media usage

DATASET: "Students Social Media Addiction.csv"

DATASET DESCRIPTION: Students' Distraction & Social Media Addiction Dataset; As quoted from Kaggle, "A dataset exploring how social media affects students' daily life, study patterns; Includes data on how much students use social media daily, and whether it affects their academic performance". More details can be found here

PRELIMINARY:
Form two new categorical features as follows:

1. "*Usage*" and fill the values as "*Overuse*" or "*Controlled*" based on the corresponding feature "*Avg_Daily_Usage_Hours*" $\geq$ or $<$ *median* of the values, respectively.
2. "*Sleep*" and fill the values as "*Sufficient*" or "*Insufficient*" based on the corresponding variable *Sleep_Hours_Per_Night* ≥ 7 or <7, respectively.

TASK:
Generate the contingency table and apply the Chi-Square Test of Independence on each of the following pairs formed by the respective features and "*Usage*".

- (Academic_Level, Usage)
- (Gender, Usage)
- (Sleep, Usage)

Provide a summary table of the results, and interpret how the variables in each pair are related.

WORKFLOW (4 marks):
Generate the following in order

1. Two new categorical variables, "*Usage*", "*Sleep*", based on the conditions provided above. **(1 mark)**
2. Three contingency tables for each of the 3 pairs. The output for each table should be similar to the following. **(1 mark)**

Usage	Controlled	Overuse
Academic Level		
Graduate	181	144
High School	2	25
Undergraduate	194	159

3. The summary table and result analysis. The output should be similar to the following table **(2 marks)**

	Chi-Square Statistic	p-value	Is Related to Usage
Academic_Level	23.9853	0.0000	Yes
Gender	0.6327	0.4264	No
Sleep	405.9778	0.0000	Yes