Submission Format and Evaluation Rubric

Group Information

- Group Number: 15
- Group Members: Siwoong Yoon, Minsun Jee, Seungwoo Kim, Jeonghyun Kim, Yewon Jeung
- Date of Discussion: 2025. 09. 21
- Any issues arising from collaborative work

Part A: The 4-level framework (Building from Week 2)

Original Hypothesis from Week 2

- H_0: Middle aged people use online shopping less than young generations.
- H_1: Middle aged people use online shopping and services as same as young generations or even more.

Any Refinements After Further Consideration

- No change

Four-Level Framework Construction

1. Level 1 - Population
   - ○ Define population parameters assumed (μ, σ², p, λ), stating what is unknown, known and etc.

   μ: average purchase amount
   σ²: dispersion within a group
   p: Distribution ratio of purchase frequency
   λ: not needed
   known: sample data(location, purchase amount, age etc.)
   unknown: The actual distribution of the population as a whole (especially differences by race and income)
2. Level 2 - Sample
   - ○ Define the primary Variables

Variable Name	Type	Role	How Measured
purchase amount	Continuous	Dependent	Kaggle data
frequency of purchase	Categorical	Dependent	Kaggle data(weekly, monthly, annually)
age group	categorical	Independent	Kaggle data(only age)

○ Confounding Variables Identified:

- difference of location, gender, seasonality(holiday vs non-holiday), and economic factors
  - Plan to address: apply statistical control and treat them as

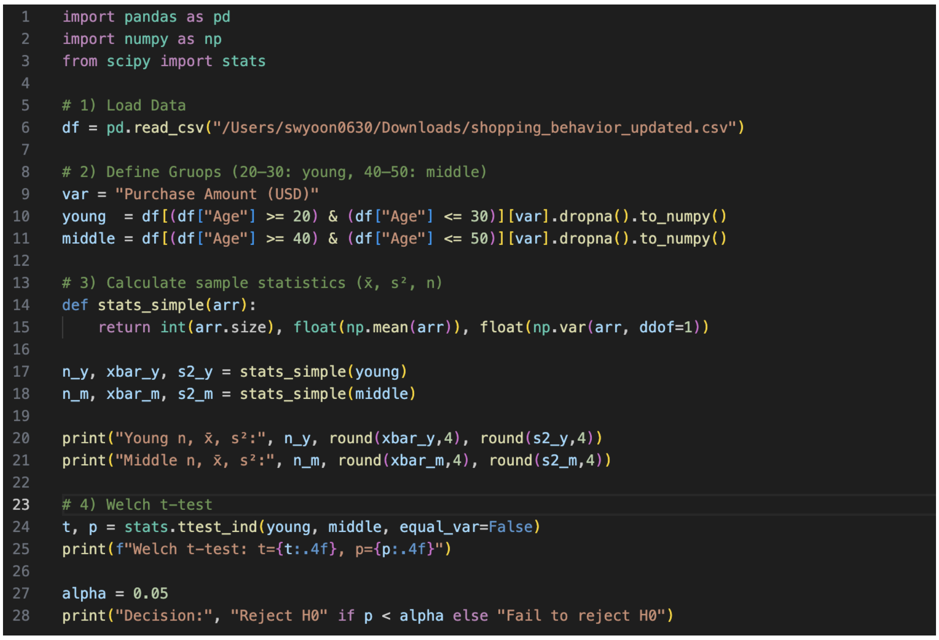
supplementary variables in the analysis

○ Calculate your ‘test’ sample statistics ($\bar{x}$,$\bar{x}s^2$, $\hat{p}$),

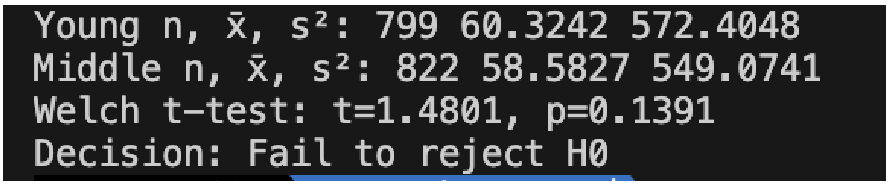
■ In case you have the data from the “Tackling the conventional wisdom”

■ otherwise, you can find relevant statistics from other resources (e.g., newspaper, research articles, and etc)

■ Show all computational work (i.e., Python codes)



```python
import pandas as pd
import numpy as np
from scipy import stats

# 1) Load Data
df = pd.read_csv("/Users/swyoon0630/Downloads/shopping_behavior_updated.csv")

# 2) Define Gruops (20–30: young, 40–50: middle)
var = "Purchase Amount (USD)"
young  = df[(df["Age"] >= 20) & (df["Age"] <= 30)][var].dropna().to_numpy()
middle = df[(df["Age"] >= 40) & (df["Age"] <= 50)][var].dropna().to_numpy()

# 3) Calculate sample statistics (x̄, s², n)
def stats_simple(arr):
    return int(arr.size), float(np.mean(arr)), float(np.var(arr, ddof=1))

n_y, xbar_y, s2_y = stats_simple(young)
n_m, xbar_m, s2_m = stats_simple(middle)

print("Young n, x̄, s²:", n_y, round(xbar_y,4), round(s2_y,4))
print("Middle n, x̄, s²:", n_m, round(xbar_m,4), round(s2_m,4))

# 4) Welch t-test
t, p = stats.ttest_ind(young, middle, equal_var=False)
print(f"Welch t-test: t={t:.4f}, p={p:.4f}")

alpha = 0.05
print("Decision:", "Reject H0" if p < alpha else "Fail to reject H0")
```



```
Young n, x̄, s²: 799 60.3242 572.4048
Middle n, x̄, s²: 822 58.5827 549.0741
Welch t-test: t=1.4801, p=0.1391
Decision: Fail to reject H0
```

○ Define Sampling Method

- Method chosen: Random
- Justification: Because we already have enough data in a given dataset, sampling randomly without specific standard and bias is the most fair and representative method. Also, we can process statistically other variables(gender, location, category) in the analysis step. So using random sampling is proper in the initial sampling step.
- Sample size: n = 200
- CLT consideration: If the sample size is over 30, by CLT, sampling distribution of the sample mean follows normal distribution approximately. So if n >= 30, it is enough sample size to use t-test and χ^2-test.

3. Level 3 - Sampling Distribution

○ Identify appropriate distribution (t, χ^2, F) based on your research questions

- When we compare mean of a continuous variable such as purchase amount, and frequency of purchase by age or gender -> t-test
- When we compare difference in proportions of a categorical variable such as category -> χ^2-test

○ Verify CLT conditions

Sample size n >= 30 -> Sampling distribution of the sample mean follows normal distribution approximately by CLT. The samples are independent of each other. The population distribution is not much distorted.

○ Find critical values from your sample data

t-test -> After calculating the t-statistic by using sampling distribution of the sample mean, standard deviation, and sample size, identify the t-critical value by df(n-1).

χ^2-test -> Calculate the χ^2-statistic by comparing observed frequency and expected frequency. And identify the χ^2-critical value by df{(rows − 1)(columns − 1)}.

4. Level 4 - Inference (Testing Plan)
    ○ Primary Test Selection
        ■ Test chosen: t-test/Chi-square
        ■ Why appropriate: From t-test, we can analyze the relationship between purchase amount and age(continuous variable).
        Additionally, from Chi-square test, we can analyze the relationship between categorical variables such as category and item purchased and age.
        ■ Parameter of interest: In t-test, we are interested in the difference of μ. And we want to find p value for Chi-square test.
    ○ Confidence Interval Plan
        ■ CI level: 95%
        ■ What it will tell us: 95% is the most common value for confidence. It will tell us the range of the possibility that there is a difference between middle-aged people and young people.

Part B: Group Process Reflection

Building on Week 2: (100 words)

- How did Week 2's Socratic questioning inform our data strategy?
- What creative elements from Week 2 carried forward?
- What new challenges emerged in operationalizing our hypothesis?

In week2's Socratic questioning, the question which is "Does online shopping usage simply mean the number of purchases or the amount of money?" played a decisive role in setting the direction for data analysis. Thanks to this question, using t-test to compare the average purchase amount or usage frequency between two generations became a natural choice and it also allowed us to plan alternative analyses in advance.

The creative element from week2 is the perspective that extends to considering economic value. This allows us to view things from a different perspective by recognizing that it is important to measure the actual purchasing power each generation contributes to the market. Additionally, how companies can encourage middle-aged consumers to online shopping was also a creative element. This will demonstrate the practical value of our group's analysis.

Since the data used is a given data, it may be difficult to generalize. Furthermore, variables such as income and race are absent from the data, making it challenging to control for these factors.

Slack Guidelines (Enhanced from Week 2)

DO:

- Reference specific Week 3 concepts (CLT, distributions)
- Use Week 2's Socratic question types when stuck
- Create a shared document for real-time collaboration (and also share it in the Slack’s group channel)
- Use polls for group decisions
- Pin important decisions for easy reference

DON'T:

- Abandon your creative Week 2 insights for "easier" hypotheses
- Choose tests without considering variable types
- Ignore practical constraints

Submission Requirements

- Format: PDF document
- Length: 3-4 pages maximum
- Submission: Upload to LXP by [21st Sep, 23:59pm]
- File naming: Group[#]_Assignment1_W3.pdf

Grading Criteria

- Consistency with Week 2 hypothesis (20%)
- Variable identification and classification (20%)
- Appropriate test selection with justification (20%)
- Sampling strategy and CLT application (20%)
- Feasibility and ethical considerations (10%)
- Group collaboration and progression (10%)

Common Pitfalls to Avoid

- Using the z-distribution when σ is unknown
- Forgetting to check CLT conditions
- Confusing statistical and practical significance
- Making causal claims from observational data
- Ignoring assumption violations
- Over-interpreting small samples
- Missing degrees of freedom calculation