DATA ANALYSIS & VISUALIZATION

U18AIE0219L

EXPERIMENT -2

Pre-lab Questions:

1. **Define inferential statistics and provide an example.**

Inferential statistics is a branch of statistics that involves making predictions, inferences, or generalizations about a population based on a sample of data. This method uses probability theory to estimate population parameters, test hypotheses, and draw conclusions that extend beyond the immediate data set.

(eg) A researcher uses sample data (e.g., height of 100 women) to estimate the average height of all adult women in a city.

2. **What is the difference between a t-test and a z-test?**

- **Z-Test:**
  - Used when the sample size is large (typically n > 30) or when the population variance is known**.**
  - Assumes the sample data follows a normal distribution or the sample size is large enough to apply the Central Limit Theorem (CLT).
  - The population variance (or standard deviation) is known.
- **T-Test:**
  - Used when the sample size is small (typically n < 30) or when the population variance is unknown.
  - Uses the t-distribution, which is wider and more spread out compared to the normal distribution, especially with small sample sizes. The t-distribution becomes more like the normal distribution as the sample size increases.
  - The population variance is unknown and is estimated from the sample data.

3. **Why are p-values important in inferential statistics?**

The p value reflects the degree of data compatibility with the null hypothesis. Some recommend abandoning p value, others lowering the significance threshold to 0.005. A 0.005 threshold could increase sample sizes and costs as well as depress spontaneous research.

4. Explain the concept of confidence intervals.

A confidence interval (CI) is a range of values that likely contains the true value of something you're measuring (e.g., average height, population proportion).

Confidence Level: It shows how confident you are that the true value is in the range. For example, a 95% CI means you're 95% confident the true value is inside the range.

Formula: The CI is calculated using the sample mean (or another statistic) plus or minus a margin of error**:**

Formula:

CI=Sample Value±Margin of Error

Example: If you calculate a 95% CI for average height and get (170 cm, 180 cm), you're 95% sure the true average height is between 170 and 180 cm.

Key Point: A confidence interval is a range, not a guarantee. The true value might still be outside the range**.**

5. How can outliers impact hypothesis testing?

Outliers are data points that deviate significantly from the rest of the distribution. They can have a large impact on the results of hypothesis testing, which is a method of evaluating the validity of a technical analysis claim based on statistical evidence.

Descriptive Statistical Analysis

Aim:

To apply inferential statistics to analyze and interpret data using hypothesis testing and confidence intervals

Resources used:

Programming Language – Python

Platform-Google Colab

Dataset- feedback_data.csv

Columns chosen for comparison- cohesion, vocabulary

Hypothesis:

- Null Hypothesis (H0): The mean cohesion score is equal to the mean vocabulary score.
- Alternative Hypothesis (H1): The mean cohesion score is not equal to the mean vocabulary score.

Confidence Intervals:

95% confidence intervals for the mean of each column.

Code:

```
import pandas as pd
from scipy.stats import ttest_ind
import numpy as np
# Load the data
data = pd.read_csv('feedback_data.csv')
print(data.head())
# Extract the columns
cohesion_scores = data['cohesion']
```

```
vocabulary_scores = data['vocabulary']

# Perform t-test

t_stat, p_value = ttest_ind(cohesion_scores, vocabulary_scores)

# Calculate confidence intervals

cohesion_mean = np.mean(cohesion_scores)

cohesion_std = np.std(cohesion_scores, ddof=1)

cohesion_ci = cohesion_mean + np.array([-1.96, 1.96]) * cohesion_std /
np.sqrt(len(cohesion_scores))


vocabulary_mean = np.mean(vocabulary_scores)

vocabulary_std = np.std(vocabulary_scores, ddof=1)

vocabulary_ci = vocabulary_mean + np.array([-1.96, 1.96]) *
vocabulary_std / np.sqrt(len(vocabulary_scores))

# Print results

print("T-statistic:", t_stat)

print("P-value:", p_value)

print("Cohesion Mean:", cohesion_mean, "Confidence Interval:",
cohesion_ci)

print("Vocabulary Mean:", vocabulary_mean, "Confidence Interval:",
vocabulary_ci)
```

Google Colab Link:

https://colab.research.google.com/drive/1J-GC9dSU_qjfxiIDtfb_MaoCfXmwJCA1#scrollTo=tuwaOSUeE_xw

Output:

```
T-statistic: -7.699623677887229
P-value: 1.5318250077458225e-14
Cohesion Mean: 3.127077473791869 Confidence Interval: [3.10631277
3.14784218]
Vocabulary Mean: 3.2357453336742523 Confidence Interval: [3.21746891
3.25402176]
```

Inference:

The output indicates that there is a statistically significant difference between the mean cohesion score and the mean vocabulary score.

- **T-statistic:** The value of -7.699623677887229 suggests a large difference between the means.
- **P-value:** The extremely small p-value (1.5318250077458225e-14) is much less than the typical significance level of 0.05. This indicates that the observed difference is unlikely to have occurred by chance.
- **Confidence Intervals:** The 95% confidence intervals for cohesion and vocabulary do not overlap. This further supports the conclusion that the means are different.

Conclusion:

Based on the statistical analysis, we can confidently conclude that the average cohesion score is significantly different from the average vocabulary score in the dataset.

Post-lab Questions:

1. **What does the p-value indicate in hypothesis testing?**

- The p-value measures the probability of obtaining results as extreme as observed, assuming the null hypothesis is true. Indicates the strength of evidence against the null hypothesis.
- The p-value does not indicate the probability that the null or alternative hypothesis is true.
- Low p-value (< 0.05): Strong evidence against the null hypothesis; reject the null hypothesis.
- High p-value (> 0.05): Insufficient evidence to reject the null hypothesis; fail to reject the null hypothesis.

2. **How can the t-statistics help in comparing two groups?**

- The t-statistic quantifies the difference between the sample means of two groups relative to the variability in the data.

$$t = \frac{\text{Difference in Means}}{\text{Standard Error of the Difference}}$$

- It considers sample size and variability (standard deviation) in the data, providing a normalized measure of the difference.
- Null Hypothesis: Assumes no difference between the group means ($\mu1=\mu2$).
- Alternative Hypothesis: Assumes a significant difference between the means ($\mu1\neq\mu2$).
- The calculated t-statistic is compared to a critical value (from the t-distribution table) based on the significance level (α\alphaα) and degrees of freedom.
- If |t|>critical value: Reject H0H_0H0 (significant difference).
- If |t|≤critical value: Fail to reject H0H_0H0 (no significant difference).

3. Explain why confidence intervals are preferred over point estimates.

Confidence intervals (CIs) are preferred over point estimates because they provide more comprehensive and informative insights about the parameter being estimated.

- Point Estimate: Provides a single value as an estimate for the parameter (e.g., sample mean).
- Confidence Interval: Provides a range of plausible values within which the true parameter is likely to lie, offering more context.
- Confidence intervals explicitly account for sampling variability, reflecting the uncertainty associated with the point estimate.
- CIs allow for better decision-making by showing the margin of error and helping assess the precision and reliability of the estimate.
- CIs enable comparisons across studies or datasets by showing the range of possible values, making results more interpretable and less prone to misinterpretation.
- CIs are more intuitive in graphical representations, making it easier to understand the variability and reliability of the estimate.

4. If the p-value is 0.03, then what does it mean in the context of hypothesis testing?

- There is a 3% probability of obtaining the observed data (or something more extreme) if the null hypothesis (H0H_0H0) is true.
- p (0.03) <α (0.05): The p-value is smaller than the significance level, so the null hypothesis is rejected.

- This suggests that the result is statistically significant, providing evidence in favor of the alternative hypothesis (H1H_1H1).
- The result is unlikely to occur by random chance if the null hypothesis is true.

5. Suggest improvements to handle unequal variances between two groups.

- Weighted Analysis: Assign weights to groups inversely proportional to their variances. This approach minimizes the impact of groups with high variability.
- Bootstrap Resampling: Use bootstrap methods to estimate the sampling distribution of the test statistic without assuming equal variances. This approach is computationally intensive but highly flexible.
- Non-Parametric Tests: Non-parametric tests (e.g., Mann-Whitney U test) do not assume equal variances or normality. These tests compare medians rather than means.
- Welch's t-test: Unlike the standard t-test, Welch's t-test does not assume equal variances between groups. It adjusts the degrees of freedom based on the variances and sample sizes, providing more reliable results in the presence of unequal variances.

Done by:
Deepika Ganesan- 22BCS019

Durga S- 22BCS026

Nandhini Priya- 22BCS073

Nithiyaa T- 22BCS082

Priya Dharshini- 22BCS094

Rithika- 22BCS103

Sandhiya G- 22BCS109

Santhiya P 22BCS111