SHRI VILEPARLE KELAVANI MANDAL'S
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC ACCREDITED with "A" GRADE (CGPA : 3.18)
DEPARTMENT OF INFORMATION TECHNOLOGY

COURSE CODE: DJS22ITL5013   **DATE:** 9/10/24

COURSE NAME: Statistical Analysis Lab   **Class: I1**

Name: Anish Sharma   **Roll No: I011**   **Sap: 60003220045**

LAB EXPERIMENT NO. 6

CO/LO: Perform Test of Hypothesis for independence and appropriateness of distribution using various statistical techniques.

AIM / OBJECTIVE:

To estimate confidence interval for mean, population proportion and variance (Single Population)

DESCRIPTION OF EXPERIMENT:

The aim of this experiment is to estimate the confidence intervals for three population parameters: the mean, the population proportion, and the population variance, using a single population. Confidence intervals provide a range within which we expect the true population parameter to lie, based on a sample from that population. Confidence intervals are calculated under the assumption that the sample is representative of the population, and the confidence level (typically 90%, 95%, or 99%) indicates the degree of certainty in our estimation. For constructing confidence intervals and testing hypotheses about the differences in two population means and two population proportions and, in addition, testing hypotheses about two population variances following tests are suggested:

1. For constructing confidence intervals or testing hypotheses about the difference in two population means and the population standard deviations or variances are known, the z test for μ will be used.

2. For constructing confidence intervals for the population where standard deviations or variances are unknown, then the appropriate technique is the t test for μ.

3. For constructing a confidence interval for population proportions, the z test for p will be used.

4. For constructing a confidence interval for population variances, the chi-square test will be used.

INPUT DATA / DATASET:

1. Select appropriate dataset. (different for each test)

2. Perform the following tests using Excel and Python tools for two population:

```python
import seaborn as sns
import pandas as pd
import numpy as np
import scipy.stats as stats


tips = sns.load_dataset('tips')
```

```python
print(tips.head())
```

```
       total_bill    tip    sex  smoker  day    time  size
   0        16.99   1.01  Female     No   Sun  Dinner     2
   1        10.34   1.66    Male     No   Sun  Dinner     3
   2        21.01   3.50    Male     No   Sun  Dinner     3
   3        23.68   3.31    Male     No   Sun  Dinner     2
   4        24.59   3.61  Female     No   Sun  Dinner     4
```

a. Estimate confidence interval about the population mean with a known population
standard deviation using the z statistic.

```python
population_std_dev = 10
confidence_level = 0.95
sample_size = len(tips)

sample_mean = tips['total_bill'].mean()

z_score = stats.norm.ppf(1 - (1 - confidence_level) / 2)

margin_of_error = z_score * (population_std_dev / np.sqrt(sample_size))

confidence_interval = (sample_mean - margin_of_error, sample_mean +
margin_of_error)
print("Confidence interval for mean (known σ):", confidence_interval)
```

```
Confidence interval for mean (known σ): (18.531204256144044, 21.040680989757597)
```

b. Estimate confidence interval about the population mean with an unknown population
standard deviation using the t statistic.

```python
# Sample standard deviation
sample_std_dev = tips['total_bill'].std()

# T-score for 95% confidence
t_score = stats.t.ppf(1 - (1 - confidence_level) / 2, df=sample_size - 1)

# Margin of error
margin_of_error_t = t_score * (sample_std_dev / np.sqrt(sample_size))

# Confidence interval
```

```python
confidence_interval_t = (sample_mean - margin_of_error_t, sample_mean +
margin_of_error_t)
print("Confidence interval for mean (unknown σ):", confidence_interval_t)
```

Confidence interval for mean (unknown σ): (18.663331704358473, 20.908553541543167)

c. Estimate confidence interval about the population proportion using the z statistic.

```python
p_female = tips['sex'].value_counts(normalize=True)['Female']
n = len(tips)

z_score_prop = stats.norm.ppf(1 - (1 - confidence_level) / 2)

margin_of_error_prop = z_score_prop * np.sqrt((p_female * (1 - p_female))
/ n)

confidence_interval_prop = (p_female - margin_of_error_prop, p_female +
margin_of_error_prop)
print("Confidence interval for population proportion:",
confidence_interval_prop)
```

Confidence interval for population proportion: (0.29645760787230985, 0.4166571462260508)

d. Estimate confidence interval about the population variance using the chi-square statistic

```python
sample_variance = tips['total_bill'].var()

chi_square_low = stats.chi2.ppf(0.025, df=sample_size - 1)
chi_square_high = stats.chi2.ppf(0.975, df=sample_size - 1)

confidence_interval_variance = (
    (sample_size - 1) * sample_variance / chi_square_high,
    (sample_size - 1) * sample_variance / chi_square_low
)
print("Confidence interval for population variance:",
confidence_interval_variance)
```

Confidence interval for population variance: (66.85336288187914, 95.47310092851704)

Observations / Discussion of Results

1. The confidence intervals for the mean total bill, population proportion of female diners, and population variance highlight the variability and distribution within the dataset. The results indicate a typical range for the total bill and suggest a balanced gender distribution among diners.
2. The chosen tests effectively assessed the population parameters in the dataset. The z-test and t-test provided insights into the mean's reliability, while the proportion test evaluated gender representation. These methods are appropriate given the sample's characteristics and the underlying assumptions of normality.

CONCLUSION:

This experiment demonstrated the application of statistical techniques to estimate population parameters. The findings reveal important insights into dining behavior, emphasizing the significance of understanding sample data to infer broader trends, which can guide restaurant management and marketing strategies.

REFERENCES:

(List the references as per format given below and citations to be included the document)

[1] Ponniah P., "Data Warehousing: Fundamentals for IT Professionals", 2nd Edition, Wiley India, 2013.

[2] Ageed, Z. S., Zeebaree, S. R., Sadeeq, M. M., Kak, S. F., Yahia, H. S., Mahmood, M. R., & Ibrahim, I. M. (2021), "Comprehensive survey of big data mining approaches in cloud systems", Qubahan Academic Journal, 1(2), 29-38.

Website References:

Author's Last Name, First Initial. Middle Initial. (Date of Publication or Update). Title of work. Site name. Retrieved Month Day, Year, from URL from Homepage

[3] U.S. Census Bureau. U.S. and world population clock. U.S. Department of Commerce. Retrieved July 3, 2019, from https://www.census.gov/popclock.