DSBDA LAB 3

RAVITEJ.R

160119733160

AIM:-To perform hypothesis tests on a given dataset (T-test, Z-test)

DESCRIPTION:-

Hypothesis Testing : In hypothesis testing, two mutually exclusive statements about a parameter are evaluated to decide which statement is best supported by sample data. The two mutually exclusive statements are Null Hypothesis (Ho) which claims that there is no significant difference between features and Alternate Hypothesis (Ha) which claims there is a significant difference between features.

Let’s discuss the types with an example of a dataset mentioning the students’ performances in math and reading sections.

Null Hypothesis (Ho): There is no difference in performance of students between math, reading skills.

Alternate Hypothesis (Ha): There is a difference in performance of students between math, reading skills

```
da = pd.read_csv(io.BytesIO(uploaded['ds_salaries.csv']),index_col='job_title')
first=da.loc['Data Science Manager']
second=da.loc['Machine Learning Scientist']
y=first[['salary_in_usd']]
z=second[['salary_in_usd']]
print("Data Science Manager-mean",y.mean())
print("Data Science Manager-median",y.median())
print("Data Science Manager-varience",np.var(y))
print("Machine Learning Scientist-mean",z.mean())
print("Machine Learning Scientist-median",z.median())
print("Machine Learning Scientist",np.var(z))
```

```
Data Science Manager-mean salary_in_usd    158328.5
dtype: float64
Data Science Manager-median salary_in_usd    155750.0
dtype: float64
Data Science Manager-varience salary_in_usd    2.535153e+09
dtype: float64
Machine Learning Scientist-mean salary_in_usd    158412.5
dtype: float64
Machine Learning Scientist-median salary_in_usd    156500.0
dtype: float64
Machine Learning Scientist salary_in_usd    5.532266e+09
dtype: float64
```

Mean ,median and variance of Data Science Manager and Machine Learning Scientist

Types of Hypothesis Testing:

-> A **t-test** is a statistical test that compares the means of two samples. It is used in hypothesis testing, with a null hypothesis that the difference in group means is zero and an alternate hypothesis that the difference in group means is different from zero.

```
da=pd.read_csv(io.BytesIO(uploaded['ds_salaries.csv']),index_col='job_title')
first=da.loc['Data Science Manager']
second=da.loc['Machine Learning Scientist']
y=first[['salary_in_usd']]
z=second[['salary_in_usd']]
a1,a2=stats.ttest_ind(y,z)
print("pval")
print(a2)
if a2<0.05:
  print("null hypothesis rejected(t test)")
else:
  print("hypothesis accepted")
```

OUTPUT:-

pval=0.99775175

```
pval
[0.99775175]
pval
[0.99772032]
hypothesis accepted
hypothesis accepted
```

-> A **z-test** is a statistical test to determine whether two population means are different when the variances are known and the sample size is large. A z-test is a hypothesis test in which the z-statistic follows a normal distribution.

```
a3,a4=ztest(y,z,value=0)
OUTPUT:-
0.99772032
```

```
pval
[0.99775175]
pval
[0.99772032]
hypothesis accepted
hypothesis accepted
```

DSBDA LAB 4

Ravitej.R
160119733160

AIM:- to perform chisquare,anova and pearson's correlation tests

DESCRIPTION:-

Null Hypothesis (Ho) which claims that there is no significant difference between features and Alternate Hypothesis (Ha) which claims there is a significant difference between features.Here, we are taking a dataset mentioning the students' performances in math and reading sections. We have taken a parameter 'verdict' which says 'acceptable' if the overall performance of the students is greater than 40 else 'not acceptable. Here, we are testing if there is any dependency between gender and verdict.

CHI square Test:- Mathematically, a Chi-Square test is done on two distributions two determine the level of similarity of their respective variances.

$$X^2 = \sum \frac{(\text{Observed value - Expected value})^2}{\text{Expected value}}$$

One-Way Anova - The one-way ANOVA compares the means between the groups you are interested in and determines whether any of those means are statistically significantly different from each other.

$$F = \frac{MS\ (between)}{MS\ (within)}$$

Ho - there is no significant difference in performance
Ha:- there is a significant difference in performance

PEARSON's correlation:-

Pearson's correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.

1)

CODE:-

```
from statsmodels.stats.weightstats import ztest as ztest
import scipy.stats as stats
da=pd.crosstab(df.company_size,df.work_year,margins=True)
print(da)
stat, p, dof, expected = chi2_contingency(da)
alpha = 0.05
print("p value is " + str(p))
if p <= alpha:
   print('Dependent (reject H0)')
else:
   print('Independent (H0 holds true)')
```

OUTPUT:

```
work_year      2020  2021  2022  All
company_size
L                33   119    46  198
M                14    53   259  326
S                25    45    13   83
All              72   217   318  607
p value is 7.414263119108927e-42
Dependent (reject H0)
```

AS we can see there is a significant difference in hiring of people of size L,S,M in the year 2020-2022

2)

CODE:-

```
import pandas as pd
from scipy.stats import pearsonr
list1 = df['YearsExperience']
list2 = df['Salary']
corr, _ = pearsonr(list1, list2)
print('Pearsons correlation: %.3f' % corr)

import pandas as pd
```

```
from scipy.stats import pearsonr
list1 = df['YearsExperience']
list2 = df['Salary']
corr, _ = pearsonr(list1, list2)
print('Pearsons correlation: %.3f' % corr)
```

OUTPUT:-

```
Pearsons correlation: 0.982
```

This indicates that years of exp and salary are highly correlated i.e as years of exp increases so does salary

3)

CODE:-

f_oneway(da[2020],da[2021],da[2022])#(annova/ftest) between years doing one way anova

OUTPUT:-

```
from scipy.stats import f_oneway

f_oneway(da[2020],da[2021],da[2022])
```

```
F_onewayResult(statistic=1.5244430658794075, pvalue=0.2690482936807047)
```

Since pval>0.05 reject null hypothesis i.e significant difference in hiring between years