EXPERIMENT 4

AIM: To perform hypothesis testing

SOFTWARE USED: Jupyter Notebook

THEORY:

- Null Hypothesis (H0): The initial assumption or claim about a population parameter that is assumed to be true unless proven otherwise.
- Alternative Hypothesis (H1 or Ha): The hypothesis that contradicts the null hypothesis and is accepted if there is enough evidence against the null hypothesis.
- Test Statistic: A value calculated from sample data used to assess the strength of evidence against the null hypothesis.
- Level of Significance (α): The probability of rejecting the null hypothesis when it is true, typically set at 0.05 or 0.01.
- Critical Region (Rejection Region): The set of values of the test statistic for which the null hypothesis is rejected.
- P-value: The probability of observing a test statistic as extreme as, or more extreme than, the one obtained from the sample data, assuming the null hypothesis is true.

Steps in Hypothesis Testing:

- State the hypotheses: Formulate the null and alternative hypotheses based on the research question.
- Choose the significance level: Decide on the level of significance, denoted by α.
- Select the appropriate test statistic: Choose a test statistic based on the type of data and the hypothesis being tested.
- Determine the critical region: Determine the critical values or critical region of the test statistic corresponding to the chosen significance level.
- Collect and analyze data: Collect sample data and calculate the test statistic.
- Make a decision: Compare the test statistic to the critical values or use the p-value to determine whether to reject the null hypothesis.
- Draw conclusions: Based on the decision made in the previous step, draw conclusions about the null hypothesis and interpret the results in the context of the research question.

Types of Hypothesis Tests:

- Parametric tests: Tests that make assumptions about the distribution of the population, such as t-tests, z-tests, and F-tests.
- Non-parametric tests: Tests that do not rely on distributional assumptions, such as chi-square tests and Wilcoxon signed-rank tests.
- One-tailed tests: Tests that examine whether a parameter is significantly greater than or less than a specific value.

- Two-tailed tests: Tests that examine whether a parameter is significantly different from a specific value.

Errors in Hypothesis Testing:

- Type I error (False Positive): Rejecting the null hypothesis when it is actually true.
- Type II error (False Negative): Failing to reject the null hypothesis when it is actually false.

OUTPUT:

In [1]:
```
#to perform hypothesis testing
#experiment 4
#subject: dsa
```

In [2]:
```
#call centre example
#anjali : day 6
```

In [3]:
```
#s = 3 mins
#u = 4 mins
#n = 50 days
# x_bar = 4.1 Anjali gothawal
```

In [17]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

In [5]:
```
call_data = pd.read_excel('Lab 4_DSA.xlsx')
```

In [6]:
```
call_data.head()
```

Out[6]:

	Day	Mean Call Duration	t-value	p-value
0	1	3.7	-0.447214	0.659777
1	2	4.1	0.149071	1.116932
2	3	3.5	-0.745356	0.465178
3	4	4.2	0.298142	1.231170
4	5	3.9	-0.149071	0.883068

In [7]:
```
ttest = (4.1- 4)/(3/np.sqrt(20))
```

In [8]:
```
ttest
```

Out[8]:
```
0.14907119849998546
```

In [9]:
```
from scipy import stats
```

In [10]:
```
p = 2*stats.t.cdf(0.14907,df=19)
print(p)
```

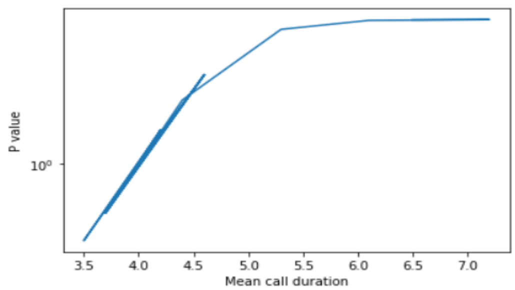
```
1.116930724940456
```

In [11]:
```
if p > 0.05:
    print('Null Hypothesis')
else:
    print('Alternate hypothesis or Reject null hypothesis')
```

```
Null Hypothesis
```

In [18]:
```
plt.plot(call_data['Mean Call Duration'],call_data['p-value'])
plt.xlabel('Mean call duration')
plt.ylabel('P value')
plt.yscale('symlog')
```

```
In [18]: plt.plot(call_data['Mean Call Duration'],call_data['p-value'])
         plt.xlabel('Mean call duration')
         plt.ylabel('P value')
         plt.yscale('symlog')
```




```
In [19]: #p value for all days and find minimum and maximum values of p value
```

```
In [20]: #an outbreak of salmonella typhi related illness was attributed to ice cream produced at a certain factory.
         #scientists measured the level of salmonella in 9 randomly sampled batches of ice cream.
         #level: 0.593, 0.142, 0.329, 0.691, 0.231, 0.793, 0.519, 0.392, 0.418.(mpn/g)
         #is there evidence that the mean level of salmonella in the ice cream is greater than 0.3?
         #h0<0.3
         #one sample one tail test
```

```
In [21]: data = pd.Series([0.593, 0.142, 0.329, 0.691, 0.231, 0.793, 0.519, 0.392, 0.418])
```


```
In [21]: data = pd.Series([0.593, 0.142, 0.329, 0.691, 0.231, 0.793, 0.519, 0.392, 0.418])
```


```
In [23]: import scipy
         p  = scipy.stats.ttest_1samp(data,0.3)[1]
         p_value = p/2
         p/2

Out[23]: 0.029265164842448822
```

```
In [ ]: #6 subjects were given a drug(treatment group) and additional 6 subjects a placebo.
        #their reaction time to a stimulus was measured(ms).
        #we want to perform a 2 sample ttest for comparing the means of treatment and control groups
        Control = pd.Series([91,87,99,77,88,91])
        Treat = pd.Series([101,110,103,93,99,104])
        p  = scipy.stats.ttest_2samp(Control, Treat)[1]
        #h0 = u1=u2
        #u1=control
        #u2=treat
```

CONCLUSION:

In conclusion, hypothesis testing through methods like the Z-test and t-test provides a structured approach to inferential statistics, allowing researchers to make informed decisions about population parameters based on sample data. By formulating hypotheses, selecting appropriate significance levels, collecting and analyzing sample data, and interpreting results within the context of critical regions, researchers can draw meaningful conclusions while controlling for Type I errors. These statistical methods are essential tools for data analysis across various fields, enabling evidence-based decision-making and advancing our understanding of real-world phenomena.