SECTION A: 5 MARKS

1. What does P-value signify about the statistical data?

When you perform a hypothesis test in statistics, a p-value helps you determine the significance of your results. Hypothesis tests are used to test the validity of a claim that is made about a population. The alternative hypothesis is the one you would believe if the null hypothesis is concluded to be untrue.Hypothesis tests use a p-value to know the strength of the . The p-value is a number between 0 and 1 and interpreted in the following way:

- A small p-value (<0.05) indicates you have strong evidence against the null hypothesis, so you reject the null hypothesis.
- A large p-value (> 0.05) indicates you have weak evidence against the null hypothesis, so you fail to reject the null hypothesis.
- p-values which are equal to 0.05 are considered to be marginal (could go either way).

p-value is used to come to a conclusion For example, suppose Zomato claims their delivery times are 30 minutes or less on average

But you think it’s more than that and want to prove it.

1.You conduct a hypothesis test because you believe the null hypothesis(H0):that delivery time is 30 minutes max (as claimed by Zomato) alternative hypothesis (Ha): is that the delivery time is greater than 30 minutes.

Take random sample of delivery times and run the data through the hypothesis test, and if the p-value 0.01,that is it's less than 0.05 this senario indicated that Sample is against the Null Hypothesis, soreject null

That is your claim of delivery time greater than 30 minutes is true.

Note: a. Under the assumption that taken sample data represents the total data b. The conclusion may be wrong if the sample data taken has mostly high delivery time compared to the actual data.

SECTION B: 10 MARKS

2.The mass of N1=20 acorns from oak trees up wind from a coal power plant and N2=30 acorns from oak trees downwind from the same coal power plant are measured. Are the acorns from trees downwind less massive then the ones from up wind? The sample sizes are not equal but we will assume that the population variance of sample 1 and sample 2 are equal. alpha = 0.05, t-critical for specified alpha level = -1.677

Samples are given below. State the Null and Alternate hypothesis. Which test would be used here?

Do you reject or retain the H0 based on your test? Find the t-statistic, Confidence Interval and RSquare value. (10 MARKS)

SAMPLE UP WIND: x1 = [10.8, 10.0, 8.2, 9.9, 11.6, 10.1, 11.3, 10.3, 10.7, 9.7, 7.8, 9.6, 9.7, 11.6, 10.3, 9.8, 12.3, 11.0, 10.4, 10.4]

SAMPLE DOWNWIND: x2 = [7.8, 7.5, 9.5, 11.7, 8.1, 8.8, 8.8, 7.7, 9.7, 7.0, 9.0, 9.7, 11.3, 8.7, 8.8, 10.9, 10.3, 9.6, 8.4, 6.6, 7.2, 7.6, 11.5, 6.6, 8.6, 10.5, 8.4, 8.5, 10.2, 9.2]

In [3]:

```
SAMPLE_UP_WIND_x1 = [10.8, 10.0, 8.2, 9.9, 11.6, 10.1, 11.3, 10.3, 10.7,
                     9.7, 7.8, 9.6, 9.7, 11.6, 10.3, 9.8, 12.3, 11.0, 10.4, 10.4]

SAMPLE_DOWNWIND_x2 = [7.8, 7.5, 9.5, 11.7, 8.1, 8.8, 8.8, 7.7, 9.7,
                      7.0, 9.0, 9.7, 11.3, 8.7, 8.8, 10.9, 10.3, 9.6,
                      8.4, 6.6, 7.2, 7.6, 11.5, 6.6, 8.6, 10.5, 8.4, 8.5, 10.2, 9.2]
```

Null Hypothesis(H0): acorns from trees downwind are not less massive than the ones from up wind

Alternate Hypothesis(H1): acorns from trees downwind are less massive than the ones from up wind

We can use T-test to find whether the Null Hypothesis is true or false

In [11]:

```
import seaborn as sns
import matplotlib.pyplot as plt
sns.distplot(SAMPLE_UP_WIND_x1)
sns.distplot(SAMPLE_DOWNWIND_x2)
plt.show()
```

In [107]:

```
import scipy.stats as stats
stats.ttest_ind(SAMPLE_UP_WIND_x1,SAMPLE_DOWNWIND_x2)
```

Out[107]:

```
Ttest_indResult(statistic=3.5981947686898033, pvalue=0.0007560337478801464)
```

In [109]:

```
t=3.5981947686898033
stats.norm.isf(0.95,loc=50,scale=se)
```

Out[109]:

```
45.53025606842054
```

p-value (0.0007) is much less than 0.05 so reject Null Hypothesis(H0)

In [110]:

```
#CONFIDENCE INTERVAL
import numpy as np
n1=len(SAMPLE_UP_WIND_x1)
n2=len(SAMPLE_DOWNWIND_x2)
n=n1+n2
xbar1=np.mean(SAMPLE_UP_WIND_x1)
xbar2=np.mean(SAMPLE_DOWNWIND_x2)
xbar=xbar1+xbar2
se1=xbar1/np.sqrt(len(SAMPLE_UP_WIND_x1))

se2=xbar2/np.sqrt(len(SAMPLE_DOWNWIND_x2))
se=xbar/np.sqrt(n)
```

In [111]:

```
stats.norm.isf(0.05,loc=xbar,scale=50)
```

Out[111]:

```
101.45768134757364
```

In [112]:

```
import numpy as np, statsmodels.stats.api as sms
cm = sms.CompareMeans(sms.DescrStatsW(SAMPLE_UP_WIND_x1), sms.DescrStatsW(SAMPLE_DOWNWIND_x
print (cm.tconfint_diff(usevar='unequal'))
```

```
(0.6286487921600237, 2.041351207839978)
```

In [113]:

```
#r2 value
df = n1 + n2 - 2
```

In [114]:

```
r2=t*t/t*t+df
r2
```

Out[114]:

```
60.94700559342667
```

SECTION B: 15 MARKS

3.A machine is supposed to run for 300 minutes at a go, as told by a company on one unit of regular gas. A random sample of 50 machines is tested. The machine run for an average of 295 minutes, with a standard deviation of 20 minutes. Check the hypothesis if the mean run-time of a machine is 300 minutes or not. (15 MARKS)

a. Use a 0.05 level of significance. What is the region of acceptance? (5 marks)

b. What hypothesis test would you choose to do for this problem and why? (5 marks)

c. Would you reject or fail to reject the null hypothesis? (5 marks)

In [56]:

```
n=50
xbar=300
sigma=20
alpha=0.05
se=sigma/np.sqrt(n)
```

Null(h0): mean run_time machine is 300 minutes

alternate(h1): mean run_time of machine is not 300 minutes

the region of acceptance, gives us whether the minimal difference from actual value to the observed value.

Two-tailed test

To find this hypothesis test we can do two-tailed test

In [84]:

```
print('critical value')
print(stats.norm.isf(0.25,loc=xbar,scale=se))
print(stats.norm.isf(0.975,loc=xbar,scale=se))
```

```
critical value
301.9077451048179
294.45638470260127
```

The observed sample mean does not lie in the Critical region,fail to reject null

In [68]:

```
print('p-value')
print(2*stats.norm.cdf(295,loc=xbar,scale=se))
print('p-value is not less than alpha 5%')
```

```
p-value
0.0770998717435417
p-value is not less than alpha 5%
```

p-value 0.07 is > 0.05 , so fail to reject the null hypothesis.

In [ ]: