problem

A Car manufacturing company claims mileage of all their cars is 25 kmpl. The population standard deviation is known from the past experience to be 1.5 kmpl. A car rental company wants to test the null hypothesis that the average milage is at least 25 kmpl. car representative drove 30 cars (samples) of the same company and recorded each car milage. the average milage of the sample turned out to be 24.2 kmpl.

Test the null hypothesis

(HO) Null Hypothesis>=25

(H1) Alternative Hypothesis <25

In [1]:

```
import numpy as np
mean =25
sample_mean=24.2
std=1.5
n=30
```

In [2]:

```
z=(sample_mean-mean)/(std/np.sqrt(30))
z
```

Out[2]:

```
-2.921186973360889
```

In [3]:

```
# using statistics tools
```

In [4]:

```
import scipy.stats as st
import numpy as np
p_value=st.norm.cdf(sample_mean,loc=mean,scale=std/np.sqrt(30))
p_value
```

Out[4]:

```
0.0017435024460706768
```

In [6]:

```
#statistical tools

#1. ttest or student's ttest
                        #. One sample ttest
                        #.  Two sample ttest
#2. ANOVA
```

In [10]:

```
# conduct ttest when standard deviation is not known.
```

In [14]:

```
import pandas as pd
hyp=pd.read_csv('Hypothesis ttest2.csv')
hyp.head()
```

Out[14]:

	ID	Gender	Race	SEB	School	Prog	Read	Write	Math1	Math2	SST
0	70	0	4	1	1	1	57	52	41	47	57
1	121	1	4	2	1	3	68	59	53	63	61
2	86	0	4	3	1	1	44	33	54	58	31
3	141	0	4	3	1	3	63	44	47	53	56
4	172	0	4	2	1	2	47	52	57	53	61

A teacher want to teach students for a competative exam. she collected sample of students marks from different schools and based on her past experience she believes that, these type of students will score more than 60 marks in the exams. now she want to conduct hypothesis test on the marks

In [ ]:

```
#Ho=mu>=60
#Ha=mu<=60
```

In [15]:

```
hyp['Math1'].mean()
```

Out[15]:

```
52.645
```

In [16]:

```
alpha=.05
xbar=52.645
mu=60
```

1 Sample t test

In [17]:

```
import scipy.stats as st
```

In [18]:

```
# since we are conducting test for one sample we will go for 1sampal ttest
st.ttest_1samp(hyp.Math1,mu)
```

Out[18]:

```
Ttest_1sampResult(statistic=-11.102736525713398, pvalue=1.3339158110547646e-
22)
```

In [19]:

```
# when you run ttest , we get 2 values 1. t statistics score and 2. pvalue for two tail
#pvalue/2
```

In [20]:

```
#let's write this way
s,p=st.ttest_1samp(hyp.Math1,mu)
```

In [21]:

```
s
```

Out[21]:

```
-11.102736525713398
```

In [22]:

```
p
```

Out[22]:

```
1.3339158110547646e-22
```

In [23]:

```
# for our problem we need to fond left tail p value
p/2
```

Out[23]:

```
6.669579055273823e-23
```

2 Sample t test

Independent sample

paired sample

In [24]:

```
hyp.head()
```

Out[24]:

	ID	Gender	Race	SEB	School	Prog	Read	Write	Math1	Math2	SST
0	70	0	4	1	1	1	57	52	41	47	57
1	121	1	4	2	1	3	68	59	53	63	61
2	86	0	4	3	1	1	44	33	54	58	31
3	141	0	4	3	1	3	63	44	47	53	56
4	172	0	4	2	1	2	47	52	57	53	61

In [26]:

```
hyp['Gender'].value_counts()
```

Out[26]:

```
1    109
0     91
Name: Gender, dtype: int64
```

In [28]:

```
#how to subset the data on gender
hyp[hyp['Gender']==1].Math1.head()
```

Out[28]:

```
1     53
92    65
93    52
94    46
95    61
Name: Math1, dtype: int64
```

In [29]:

```
# Independent sample
t,p=st.ttest_ind(hyp[hyp['Gender']==1].Math1,hyp[hyp['Gender']==0].Math1)
```

In [30]:

```
p
```

Out[30]:

```
0.6800544974232143
```

In [31]:

```
#paired sample
t,p=st.ttest_rel(hyp.Math1,hyp.Math2)
```

In [32]:

```
p
```

Out[32]:

```
0.17676525605507354
```

There is an another test which is more than 2 sample t test .But for each t test we allow 5% error(alpha),so if we conduct more than 2 sample t test we are addng 5% more error for each extra sample t test. In order to resolve this problem we goo for ANOVA test i.e ANalysis Of VAriance`