T-Test

- A T-test is a statistical test used in hypothesis testing to compare the mean of two samples or to compare a sample mean to a known population mean.
- It is used when the population standard deviation is unknown and the sample size is small.
- There are three main type of t-test ——

(A) one sample t-test

the one sample t-test is used to compare the mean of a single sample to a known population mean. Here our Null hypothesis ~~be~~ states that there is no significant ~~different~~ difference between the sample mean and the population mean.

- **Assumption :**

(i) Normality (ii) Independence (the observation of sample must be independent (iii) Random sampling (iv) Unknown population std.

(Q) Suppose a manufacturer claims that the average weight of their new chocolate bars is 50 grams, We highly doubt that and want to check this so we drew out a sample of 25 chocolate bars and measured their weight, the sample mean came out to be 49.7 gm. and the sample SD was 1.2 gm. Consider the significance level to be 0.05.

① Null Hypothesis (Ho): ~~There is no significance difference between~~ Average weight of chocolate bar is 50 gm.

$$\mu = 50$$

Alternative Hypothesis (Ha): $\mu \neq 50$

We have population mean (μ) = 50, sample mean ($\bar{x}$) = 49.7

sample S.D (s) = 1.2 N = 25

Now, We need to find the T-statistic.

Now,

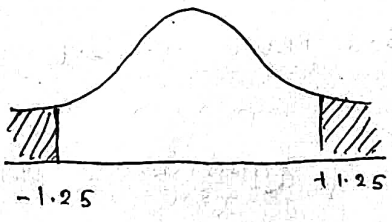
$$\text{T-statistic} = \frac{\text{Sample Mean} - \text{population mean}}{\frac{\text{sample SD}}{\sqrt{\text{Sample size}}}}$$

Here

$$\text{T-statistic} = \frac{49.7 - 50}{1.2/\sqrt{25}} = -1.25.$$

(Here degree of freedom (dF) = (N-1) = (25-1) = 24)

② Now, We are going to find out the p-value.



So, our p-value = (0.11167 X 2) = 0.22334.

As our p-value > significance level (0.05)

then We can't reject our Null hypothesis.

(8) so, the average weight of chocolate bar is 50 gm. (Proved)

(B) **Independent two-sample t-test**

The independent two sample t-test is used to compare the means of two independent samples. The null hypothesis states that there is no significance difference between the means of two samples.

It is also known as unpaired t-test.

(1) **Assumption**

① Independence of observations (two sample must be independent)
② Normality
③ Equal variance (Homoseedasticity)
④ Random sampling

(Q) Suppose a website owner claims that there is no difference in the average time spend on their website between desktop and mobile users. To test this claim, we collect data for 30 desktop users and 30 mobile users regarding the time spend on the website in minutes. The sample statistics are as following —

desktop users = [12, 15, 18, 16, 20, 17, 14, 22, 19, 21, 23, 18, 25, 17, 16, 24, 20, 19, 22, 18, 15, 14, 23, 16, 12, 21, 19, 17, 20, 14].

Mobile users = [10, 12, 14, 13, 16, 15, 11, 17, 14, 16, 18, 14, 20, 15, 14, 19, 16, 15, 17, 14, 12, 11, 18, 15, 10, 16, 15, 13, 16, 11].

and desktop SD = 3.5 and Mobile SD = 2.7.

(1) First let construct our null hypothesis as well as alternative hypothesis —

Null hypothesis (H_0): ~~There is~~ Desktop User Mean = Mobile user Mean

Alternative hypothesis (H_a): Desktop user mean ≠ Mobile user Mean

Now, we need to find out the t-statistic.

The formula of T-statistic be —

$$t = \frac{\text{First sample mean} - \text{second sample mean}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where s_1 = First sample SD, s_2 = Second sample SD, n_1 = First sample size, n_2 = Second sample size

Here t-statistic $= \frac{18.5 - 14.3}{\sqrt{\frac{3.5^2}{30} + \frac{2.7^2}{30}}} = \frac{4.2}{0.81} = 5.20$

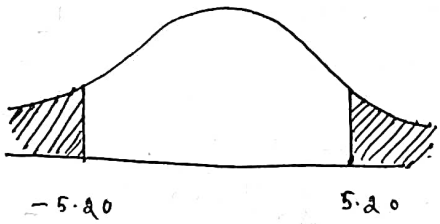
Now, degree of freedom $= (n_1 - 1) + (n_2 - 1)$

$= 29 + 29$

$= 58$

Now, corrosponding this t-statistic and df (58) we have p-value = [redacted]

$= 2.7 \times 10^{-06}$



so, clearly p-value < level of significance (0.05).

so, Null hypothesis canbe regected.

so, the average time spend on desktop and average time spend on mobile are different. (Proved)

⑥ Pair-2 sample t test

A paired two sample t-test, also known as a dependent or paired samples t-test, is a statistical test used to compare the means of two related or dependent groups.

Assumptions

① paired observations ② Normality ③ Independence of pairs (Each pair of observations should be independent).

(Q) Let's assume that a fitness center is evaluating the effectiveness of a new 8-week weight loss program. They enroll 15 participants in the program and measure their weights before and after the program.

The goal is to test whether the new weight loss program leads to a significant reduction in the participant weights.

Before the program —

[80, 92, 75, 68, 85, 78, 73, 90, 70, 88, 76, 84, 82, 77, 91]

After the program —

[78, 93, 81, 67, 88, 76, 74, 91, 69, 88, 77, 81, 80, 79, 88]

Null hypothesis : (H_0) : $\text{Mean}|_{\text{before}} = \text{Mean}|_{\text{After}}$

Alternative hypothesis (H_a): $\text{Mean}|_{\text{before}} > \text{Mean}|_{\text{after}}$

basically it is —

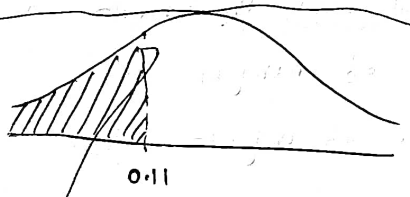
$$t\text{-statistic} = \frac{\bar{X}_{diff}}{S.D_{diff}/\sqrt{n}}$$

so, here t-statistic should be —

$$= \frac{-0.0667}{2.379/\sqrt{15}}$$

$$= -0.108$$

$$\approx -0.11$$



So, our p-value will be —

$$= 0.457$$

so, we see that —

p-value > level of significance (0.05)

so, we cannot reject our null-hypothesis.

so, we have not enough evidence to say that mean of before is less than mean of after program.

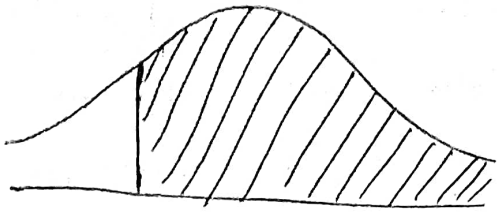
($\bar{X}_{diff}$ = Mean of the difference columns of before and after)

Means first we calculate the difference of before columns and After column then we calculate the mean of that new column.

and $S.D_{diff}$ = S.D of that new column

Here —

Before	After	Difference column
80	78	2
92	93	-1
75	81	-6
68	67	1
85	88	-3
78	76	2
73	74	-1
90	91	-1
70	69	1
88	88	0
76	77	-1
84	81	3
82	80	2
77	79	-2
91	88	3



so, with respect to the t-statistic (-0.11) and degree of freedom (14) we got p-value = 0.5424

so, p-value > level of significance (0.05)

so, we cant reject our null hypothesis, so that mean we haven't enough evedance to prove that the mean average of weight before program is greater than the mean average of weight after program. (Proved)

(11) Common Senario where paired t-test is used

(1) Match or correlated group (2) Before and After studies problems

One simple or single simple t test :

Let we have the titanic dataset and we are assuming that average age of the passengers would be 35,whereas some of us claim that it would be less than 35 . So we are going to do the t-test for this problem statement.

Answer :

so our null hypothesis (Ho): Average age of the passenger is 35

alternative hypothesis(Ha):Average age is less than 35

In [1]:

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

In [2]:

```
df=sns.load_dataset("titanic")
```

In [37]:

```
df["age"].isnull().sum()
```

Out[37]:

```
177
```

In [38]:

```
pop=df["age"].dropna()
```

In [39]:

```
pop.isnull().sum()
```

Out[39]:

```
0
```

let suppose we have taken 25 sample from the population.

In [40]:

```
sample=pop.sample(25).values
```

In [41]:

```
sample
```

Out[41]:

```
array([40.  , 42.  , 19.  , 37.  , 30.  , 29.  ,  2.  , 22.  , 26.  ,
       27.  , 22.  , 15.  , 10.  , 51.  , 31.  , 43.  , 19.  ,  4.  ,
       39.  , 32.  ,  0.42, 21.  , 38.  , 35.  , 31.  ])
```

now we have to test the assumptions

In [42]:

```
#checking for the normality
from scipy.stats import shapiro
```

In [43]:

```
result=shapiro(sample)
p_value=result[1]
if p_value>0.05:
    print("our sample is normally distributed.")
else:
    print("our sample isn't normally distributed.")
```

```
our sample is normally distributed.
```

In [44]:

```
pop_mean=35
```

In [45]:

```
from scipy.stats import ttest_1samp
```

In [46]:

```
result=ttest_1samp(sample,pop_mean)
p_value=result[1]/2 # as it is left-tailed test
if p_value>0.05:
    print("Our null hypothesis is true.Average age of the passenger is 35 yr")
else:
    print("we reject our null hypothesis and Average age is less than 35 yr")
```

```
we reject our null hypothesis and Average age is less than 35 yr
```

In [47]:

```
result
```

Out[47]:

```
Ttest_1sampResult(statistic=-3.157360962130363, pvalue=0.00425752320269086
2)
```

In [49]:

```
# checking our test with the population dataset
df["age"].mean()
```

Out[49]:

```
29.69911764705882
```

Independent 2 sample t-test :

now we are claiming that the average age of male and average age for female would be similar.for this we are going to do the t-test.

Answer

null hypothesis(Ho): average age of male is similar to the average age of female

alternative hypothesis(Ha):average age of male is not similar to average age of female.there is significant differrence between them.

In [50]:

```
male_age=df[df["sex"]=='male']["age"]
female_age=df[df["sex"]=='female']["age"]
```

In [51]:

```
pop_male_age=male_age.dropna()
pop_female_age=female_age.dropna()
```

In [52]:

```
sample_male=pop_male_age.sample(25).values
sample_female=pop_female_age.sample(25).values
```

In [53]:

```
sample_male
```

Out[53]:

```
array([11.  , 74.  , 18.  ,  1.  , 65.  , 45.  , 19.  , 38.  , 22.  ,
       31.  ,  7.  ,  4.  , 44.  , 28.  , 40.5 , 21.  , 36.  , 34.  ,
        0.42, 33.  , 50.  , 70.5 , 70.  , 32.  , 18.  ])
```

In [54]:

```
sample_female
```

Out[54]:

```
array([14., 30.,  7., 18., 16., 40., 49., 38., 43., 38., 45., 29., 45.,
       36.,  4., 35., 41., 29., 51., 22., 19., 17., 30., 54., 21.])
```

checking for assumptions

In [55]:

```
#normality check:
```

In [56]:

```
l={"sample_male":sample_male,"sample_female":sample_female}
for i in l:
    result=shapiro(l[i])
    p_value=result[1]
    if p_value>0.05:
        print(f"{i} is normally distributed")
    else:
        print(f"{i} isn't normally distributed")
```

```
sample_male is normally distributed
sample_female is normally distributed
```

In [57]:

```
# variance test
```

In [58]:

```
result=stat.levene(sample_male,sample_female)
if result[1]>0.05:
    print("these two sample columns have equal variance")
else:
    print("variance of these two columns are not equal")
```

```
these two sample columns have equal variance
```

testing the t-test:

In [62]:

```
result=stat.ttest_ind(sample_male,sample_female)
if result[1]>0.05:
    print("we can't reject our null hypothesis so average age of male is similar to aver
else:
    print("average age of male is not similar to average age of female.there is signific
```

```
we can't reject our null hypothesis so average age of male is similar to a
verage age of female
```

pair two sample t test :

Let's assume that a fitness center is evaluating the effectiveness of a new 8-week weight loss program. They enroll 15 participants in the program and measure their weights before and after the program. The goal is to test whether the new weight loss program leads to a significant reduction in the participants' weight. Before the program: [80, 92, 75, 68, 85, 78, 73, 90, 70, 88, 76, 84, 82, 77, 91] After the program: [78, 93, 81, 67, 88, 76, 74, 91, 69, 88, 77, 81, 80, 79, 88] Significance level (α) = 0.05

answer :

so we have data of weight Before the program: [80, 92, 75, 68, 85, 78, 73, 90, 70, 88, 76, 84, 82, 77, 91] and After the program: [78, 93, 81, 67, 88, 76, 74, 91, 69, 88, 77, 81, 80, 79, 88]

now our null hypothesis(Ho):mean(before)=mean(after)

alternative hypothesis(Ha):mean(before)>mean(after)

now we need to test our assumptions :

In [64]:

```
# checking the normality of the different columns of these two columns
column_before=np.array([80, 92, 75, 68, 85, 78, 73, 90, 70, 88, 76, 84, 82, 77, 91])
column_after=np.array([78, 93, 81, 67, 88, 76, 74, 91, 69, 88, 77, 81, 80, 79, 88])
```

In [65]:

```
diff_column=column_before-column_after
```

In [66]:

```
diff_column
```

Out[66]:

```
array([ 2, -1, -6,  1, -3,  2, -1, -1,  1,  0, -1,  3,  2, -2,  3])
```

In [71]:

```
result=shapiro(diff_column)
p_value=result[1]
if p_value>0.05:
    print("normally distributed")
else:
    print("not normally distreibuted")
```

```
normally distributed
```

now we need to find out the t-statistic.

In [72]:

```
t_statistic=diff_column.mean()/(diff_column.std()/np.sqrt(15))
```

In [73]:

```
t_statistic
```

Out[73]:

```
-0.10850778933039285
```

this the right-tailed test. now we need to find out the p_value with respect to t_statistic

In [76]:

```
p_value=stat.t.sf(t_statistic,14)
```

In [77]:

```
if p_value>0.05:
    print("we can't reject our null hypothesis and mean(before)=mean(after)")
else:
    print("we reject our null hypothesis and (Ha):mean(before)>mean(after)")
```

```
we can't reject our null hypothesis and mean(before)=mean(after)
```

In [ ]:

```

```