Testing Hypothesis

STUDENT'S t-DISTRIBUTION:

At the beginning of the 20th century, a statistician named William S. Gosset, an employee of Guinness Breweries in Ireland, was interested in making inferences about mean when σ was unknown. Because Guinness employees were not permitted to publish research work under their names, Gosset adopted the pseudonym "Student". The distribution that he developed has come to be known as Student's t- distribution.

If the random variable X is normally distributed, then the following statistic has a t distribution with n-1 degrees of freedom i-e

$$t = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$$

Notice that this expression has same form as the Z-Statistic, except that S is used to estimate σ, which is unknown in this case.

In appearance, the t-distribution is very similar to the standardized normal distribution. Both are bell-shaped and symmetrical. However, the t- distribution has more area in the tails and less in the centre than does the standardized normal distribution. Because the value of σ is unknown and S is used to estimate it, the values of t that are observed will be more variable than for Z. As the number of degrees freedom increases, the t-distribution gradually approaches the standardized normal distribution until the two are virtually identical. Since S becomes a better estimate of σ as the sample size gets large.

ASSUMPTIONS OF t-DISTRIBUTION:

I. The parent population from which the sample has been drawn is normal.

II. The sample observations are independent.

III. The population standard deviation σ is unknown.

APPLICATIONS:

The t-distribution has a number of applications in Statistics and other disciplines, of which are:

I. t- test for significance of single mean, population variance being unknown.
   t-test for the significance of the difference between two sample means, the population variances being equal but unknown.

II. t- test for significance of an observed sample correlation coefficient.

III. t- test for significance of an observed regression coefficient.

TEST FOR SINGLE MEAN:

Suppose we are interested to test:

a). If the given normal population has a specified value of the population mean say μ_0.

b). If the sample mean $\bar{x}$ differs significantly from specified value of population mean.

c). If a given random sample $x_1, x_2, \ldots\ldots\ldots x_n$ of size n has been drawn from a normal population with specified mean μ_0.

Basically, all the three problems are same. For all the cases we set up the null hypothesis as:

$H_0 = \mu = \mu_0$ i.e the population mean is μ.

Under H_0, the test statistic is

$$t = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$$

Where $S^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$. The above test statistic is computed and is compared with the tabulated value of t for (n-1) d.f at certain level of significance. If the calculated value of t is lea than the tabulated value of t, then H_0 is accepted, otherwise rejected.

Ex: The life expectancy of people in the year 1970 in Brazil is expected to be 50 years. A survey was conducted in eleven regions of Brazil and the data obtained are given below. Do the data confirm the expected view.

Life expectancy (Years)x: 54.2, 50.4, 44.2, 49.7, 55.4, 57.0, 58.2, 56.6, 61.9, 57.5, 53.4

Sol: Here we have to test, H_0=$\mu = 50$ against $H_1 \neq 50$

Under H_0, the test statistic t is

$$t = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$$

Here $\bar{x} = \frac{598.5}{11} = 54.41$ and $s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2 = \frac{1}{n-1}\left[\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right]$

$$= \frac{1}{10}\left\{32799.91 - {(598.5)^2}/{11}\right\}$$

$$= 23.607$$

Therefore s= 4.859

Therefore t= 14.626/40859= 3.01.

The tabulated value of t at α = 0.05 and 10 d.f is 2.228. Since calculated t is greater than the tabulated value of t. Therefore, H_0 is rejected and we conclude that the life expectancy is more than 5o years.

t- TEST FOR DIFFERENCE OF MEANS

Suppose we want to test if two independent samples have been drawn from two normal populations having the same means, the population variances being equal.

Let $x_1, x_2, \ldots\ldots\ldots x_{n1}$ and $y_1, y_2, \ldots\ldots\ldots y_{n2}$ be two independent random samples from the given normal populations. We set up the null hypothesis H_0: $\mu_x = \mu_y$ i.e the two samples have been drawn from the normal populations with the same means. In other words, the sample means $\bar{x}$ and $\bar{y}$ do not differ significantly. Under the assumption that ${\sigma_1}^2 = {\sigma_2}^2 = \sigma^2$ i.e population variances are equal but unknown. The test statistic under H_0 is

$$t = \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

Where $S^2 = \frac{1}{n_1+n_2-2}[\sum(x_i - \bar{x})^2 + \sum(y_i - \bar{y})^2]$

Proof: Here we know that

$$Z = \frac{(\bar{x} - \bar{y}) - E(\bar{x} - \bar{y})}{\sqrt{V((\bar{x} - \bar{y})}} \sim N(0,1)$$

But $E(\bar{x} - \bar{y}) = E(\bar{x}) - E(\bar{y}) = \mu_x - \mu_y = 0$ (By assumption)

and $V((\bar{x} - \bar{y}) = V(\bar{x}) + V(\bar{y}) = \frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_1} = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$ (By assumption)

Therefore

$$Z = \frac{(\bar{x} - \bar{y})}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

Let

$$\chi^2 = \frac{\sum(x_i - \bar{x})^2 + \sum(y_i - \bar{y})^2}{\sigma^2}$$

Then Fisher's t- statistic is given by

$$t = \frac{Z}{\sqrt{\frac{\chi^2}{n_1 + n_2 - 2}}}$$

Which gives

$$t = \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

Now by comparing the computed value of t with the tabulated value of t for $n_1 + n_2 - 2$ d.f at a certain level of significance, we may reject or accept the null hypothesis.

In case if the assumption ${\sigma_1}^2 = {\sigma_2}^2 = \sigma^2$ does not hold , then the t- statistic is given as:

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}}$$

PAIRED t- TEST FOR DIFFERENCE OF MEANS:

In the t-test for difference of means, the two samples were independent of each other. Let us now take a particular situation where

i) The sample sizes are equal i.e n_1 and n_2=n

ii) The sample observations $x_1, x_2, \ldots\ldots\ldots x_{n1}$ and $y_1, y_2, \ldots\ldots\ldots y_{n2}$ are nor completely independent but are dependent in pairs i.e $(x_1,y_1),(x_2,y_2)\ldots\ldots..(x_n,y_n)$ corresponding to ist, 2nd …….nth unit respectively.

Let $d_i=x_i-y_i$ (i= 1,2 ….n) denote the difference in the observations for ith unit.

Under the null hypothesis that the increments are just by chance i.e H_0: $\mu_x = \mu_y$, the test statistic is given by

$$t = \frac{\bar{d}}{S/\sqrt{n}} \sim t_{n-1}$$

Where d=x-y, $\bar{d}=\frac{1}{n}\sum d_i$ and $S^2=\frac{1}{n-1}\left[\sum d^2 - \frac{(\sum d)^2}{n}\right]$

Ex: The following table gives the monthly average of total solar radiation on a horizontal and an inclined surface at a particular place.

Month	Jan.	Feb.	March	April	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
Radiation on Horizontal surface(x)	363	404	518	521	613	587	365	412	469	468	371	330
Radiation on inclined surface (y)	536	474	556	549	479	422	315	414	505	552	492	507

Test whether the average daily radiation in a year on a horizontal and an inclined surface are equal

Sol: Here we set up the null hypothesis H_0: $\mu_x = \mu_y$ against H_1: $\mu_x \neq \mu_y$. Under H_0, the test statistic is given as :

$$t = \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

Here $\sum x = 5421$, $\bar{x} = 451.75$, $\sum x^2 = 2543583$, $\sum y = 5801$, $\bar{y} = 483.42$, $\sum y^2 = 2859497$

$$S^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum (x_i - \bar{x})^2 + \sum (y_i - \bar{y})^2\right]$$

$$= \frac{1}{n_1+n_2-2}\left\{\left[\sum {x_i}^2 - \frac{(\sum x_i)^2}{n_1}\right] + \left[\sum {y_i}^2 - \frac{(\sum y_i)^2}{n_2}\right]\right\}$$

= 6811.06

Therefore, t= -0.94, $|t| = 0.94$

The tabulated value of t at α=0.05 for 22 d.f is 2.074. Since calculated value of t is less than the tabulated value of t, therefore, we accept our null hypothesis and conclude that the average daily total radiations on the horizontal surface and the inclined surface are equal.

Ex: The following table gives the pulsality index (PI) of 11 patients:

Patient No.	1	2	3	4	5	6	7	8	9	10	11
PI value during seizure(x)	0.45	0.54	0.48	0.62	0.48	0.60	0.45	0.46	0.35	0.40	0.44
PI value after seizure(y)	0.60	0.65	0.63	0.78	0.63	0.80	0.69	0.62	0.68	0.50	0.57
Difference	0.15	0.11	0.15	0.16	0.15	0.20	0.24	0.16	0.33	0.10	0.13

Test whether there is a significant increase on the average in PI values after seizure as compared to during seizure.

Sol: We set up the null hypothesis that there is no increase in the average value of PI values after seizure and during seizure. i.e H_0= $\mu_x = \mu_y$ against H_1: $\mu_x > \mu_y$. Under H_0, the test statistic is:

$$t = \frac{\bar{d}}{s/\sqrt{n}} \sim t_{n-1}$$

Where d=x-y, $\bar{d}=\frac{1}{n}\sum d_i$ and $s^2=\frac{1}{n-1}\left[\sum d^2 - \frac{(\sum d)^2}{n}\right]$

$\sum d_i = 1.88, \sum d^2 = 0.3642, \bar{d}=\frac{1}{n}\sum d_i = 0.171$

Therefore, $s^2=\frac{1}{n-1}\left[\sum d^2 - \frac{(\sum d)^2}{n}\right]$ = 0.004289

Therefore, t = 8.72

The tabulated value of t at α= 0.05 for 10 d.f is 1.812. Since calculated value of t is greater than the tabulated value, thus, we reject our null hypothesis and conclude that there is significant increase in the PI value after seizure in comparison to during seizure.