DATE 18.03.22

F.P/S

Further Probability and Statistics

Content.

Inference using normal and t-distribution

Notes and Revision

Suresh Goel
(Former Director)
Alliance World School,
Noida. Delhi-NCR.
INDIA.
(+91 9810444804)

FP/S

Inference using normal and t-distribution: (Notes)

§ "Hypothesis tests for population mean with unknown variance"

§ In this unit we will use test statistics T using t-distribution. The test sample has a small size 'n'. The test static T is given by:

$$T = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$$s^2 = \frac{\Sigma(x-\bar{x})^2}{(n-1)} \quad \text{or} \quad s^2 = \frac{1}{(n-1)}\left(\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right)$$

This value is compared to the critical t-values for the the t-distribution. The critical t-value is found in the table intersecting the row for the degree of freedom $(n-1)$ 'ν' with the column for the prob. 'p'.

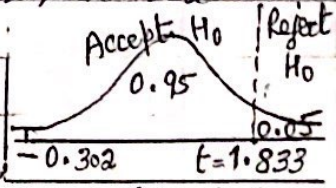
Example 1: Gemma thinks that the average length of leaves of her garden is greater than 10cm. A random sample of 10 leaves is selected and the length, X cm, of each leaf is measured: $\Sigma x = 96$ and $\Sigma x^2 = 1080$

(a) Write down a suitable null hypothesis.

(b) Write down a suitable alternate hypothesis.

(c) Describe the test statistic Gemma should use.

(d) Test at 5% level of significance whether the average length of leaves in Gemma's garden is greater than 10cm. State any assumption made.



Solution (a) H_0: $\mu = 10$cm. The average length of leaves is 10cm.

(b) H_1: $\mu > 10$cm. The average length is greater than 10cm.

(c) Gemma should use the test statistic 'T', because the population standard deviation is unknown and the sample size is small.

(d) Now $\bar{x} = \frac{\Sigma x}{n} = \frac{96}{10} = 9.6$ ✓ and $s^2 = \frac{1}{(n-1)}\left(\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right)$

$s^2 = \frac{1}{9}\left(1080 - \frac{96^2}{10}\right) = 17.6$ ✓

The test statistic $T = \frac{\bar{x} - \mu}{s/\sqrt{n}}$

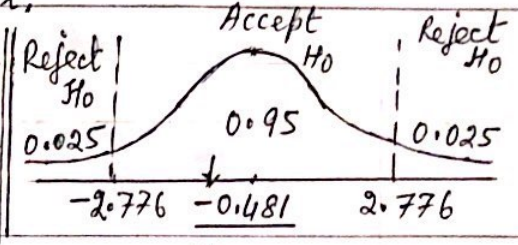
$$T = \frac{9.6 - 10}{\sqrt{17.6/10}} = -0.302$$

Now it is one-tail test to the right of $p = 0.95$ and $\nu = 10 - 1 = 9$.

The critical value of $t = 1.833$.

The test statistic $-0.302 < 1.833$ lies within the acceptance region. So H_0 is is accepted. There is no evidence to suggest that the average length of leaves is greater than 10cm.

Example 2: A restaurant chef dessolves 50g sugar into 100mL milk to make a dessert sauce. He repeats the process five times. The volumes of the sauce, in millilitres, are recorded as follows:
136, 141, 127, 132, 146.
Assume the volume of sauce has a normal distribution. Test at the 5% significance level whether these results show a difference from the expected volume of 138ml.



Solution: $H_0: \mu = 138ml$; $H_1: \mu \neq 138ml$
$\Sigma x = 682$, $n = 5$, $\bar{x} = \frac{682}{5} = 136.4$ ✓
$\Sigma x^2 = 93246$,

$$s^2 = \frac{1}{(n-1)}\left(\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right) = \frac{1}{4}\left(93246 - \frac{(682)^2}{5}\right) = 55.3$$

The test statistic $T = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{n}}} = \frac{136.4 - 138}{\sqrt{55.3/5}} = -0.481$ ✓

Two tailed test at 5% significance level, with $p = 0.975 \left(1 - \frac{0.05}{2}\right)$ and $\nu = 5 - 1 = 4$, the critical values of $t = \pm 2.776$ ✓
As $-2.776 < -0.481 < 2.776$, the test statistic T lies inside the acceptance region, so we accept H_0. There is no evidence to suggest that the mean volume is not as expected.

Example 3: A large number of children are competing in a throwing competition. The distances, in metres, thrown by a random sample of 8 children are as: 19.8, 22.1, 24.4, 21.5, 20.8, 26.3, 23.7, 25.0
Assuming that distances are normally distributed, test, at 5% significance level, whether the population mean distance thrown is more than 22.0 m.

S-20|41|Q5(a)] --[7]

Solution: $n = 8$, $\Sigma x = 183.6 \Rightarrow \bar{x} = 183.6/8 = 22.95$ ✓; $\Sigma x^2 = 4249.08$
$H_0: \mu = 22$; $H_1: \mu > 22$; $s^2 = \frac{1}{7}\left(4249.08 - \frac{183.6^2}{8}\right) = 5.066$ ✓
The test statistic $T = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{n}}} = \frac{22.95 - 22}{\sqrt{5.066/8}} = 1.194$
One tailed test at 5% significance level, with $p = 0.95$ and $\nu = 8 - 1 = 7$, $t = 1.895$
Here $1.194 < 1.895$ (critical value)... the statistic T lies inside the acceptance region, we accept H_0/ Mean distance thrown is not more than 22.0 m

Example 4: The heights of the members of a large sports club are normally distributed. A random sample of 11 members of the club is chosen and their heights, x cm, are measured. The results are summarised as follows, where $\bar{x}$ denotes the sample mean of x. $\bar{x} = 176.2$, $\Sigma(x-\bar{x})^2 = 313.1$
Test, at the 5% significance level, the null hypothesis that the population mean height for members of this club is equal to 172.5 cm against the alternative hypothesis that the mean differs from 172.5 cm. [W-20/42/Q1] ---[5]

Solution: $H_0: \mu = 172.5$; $H_1: \mu \neq 172.5$, $n = 11$,
$\bar{x} = 176.2$; $(x-\bar{x})^2 = 313.1$

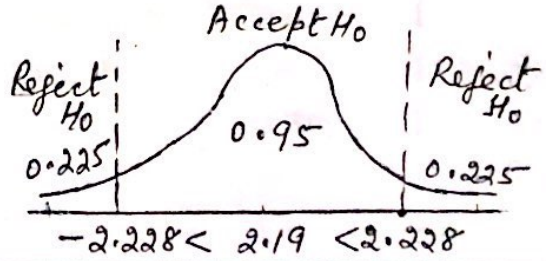
$$s^2 = \frac{(x-\bar{x})^2}{n-1} \Rightarrow s^2 = \frac{313.1}{10} = 31.3 \checkmark$$

Test statistic $T = \dfrac{\bar{x} - \mu}{\sqrt{s^2/n}}$

$$T = \frac{176.2 - 172.5}{\sqrt{31.3/11}} = 2.19$$

Two tailed test at 5% significance level with $p = 0.975 \left(1 - \frac{0.05}{2}\right)$ and $v = 11 - 1 = 10$, the critical value $t = 2.228$

$-2.228 < 2.19 < 2.228$ (critical value). Hence the test statistic T lies inside the acceptance region. We accept H_0.
Hence evidence to accept mean height is 172.5.



Example 5: A random sample of 7 observations of a variable X are as:
8.26, 7.78, 7.92, 8.04, 8.27, 7.95, 8.34
The population mean of X is μ.

(a) Test, at the 10% significance level, the null hypothesis $\mu = 8.22$ against the alternative hypothesis $\mu < 8.22$ ---[6]

(b) State an assumption necessary for the test in part (a) to be valid. --[1]

S-21/41/Q1

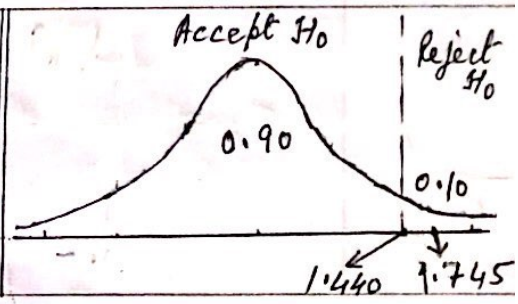
Solution: $H_0: \mu = 8.22$ and $H_1: \mu < 8.22$, $n = 7$

(a) $\bar{x} = \frac{\Sigma x}{n} = \frac{56.56}{7} = 8.08$ ✓, $\Sigma x^2 = 457.275$ ✓

$s^2 = \frac{1}{n-1}\left(\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right) \Rightarrow s^2 = \frac{1}{6}\left(457.275 - \frac{56.56^2}{7}\right) = 0.045033$

Test statistic:

$$T = \frac{\bar{x} - \mu}{\sqrt{s^2/n}} = \frac{8.08 - 8.22}{\sqrt{\frac{0.045033}{7}}} = -1.745$$



Now for one-tailed test at 10% significance level with $p = 0.90$ and $\nu = 7 - 1 = 6$, the critical value $t = 1.440$;

$1.745 > 1.440$; so reject H_0 [accept H_1 ✓]

hence there is sufficient evidence to support the hypothesis that the mean is less than 8.22 (or $\mu < 8.22$).

(b) Underlying distribution is normal / population is normal.

Example 6: Farmer A grows apples of a certain variety. Each tree produces 14.8 Kg of apples, on average, per year. Farmer B grows apples of the same variety and claims that his apple trees produce a higher mass of apples per year than Farmer A's trees. The masses of apples from Farmer B's trees may be assumed to be normally distributed. A random sample of 10 trees from Farmer B is chosen. The masses, x Kg, of apples produced in a year are as: $\Sigma x = 152.0$; $\Sigma x^2 = 2313.0$; Test at the 5% significance level, whether Farmer B's claim is justified. --[6]

S-21/43/Q1

Solution: $H_0: \mu = 14.8$; $H_1: \mu > 14.8$, $\bar{x} = \frac{152}{10} = 15.2$; $n = 10$,

$\left(s^2 = \frac{1}{9}\left[2313 - \frac{152^2}{10}\right] = 0.28889\right.$.

Test statistic $T = \frac{15.2 - 14.8}{\sqrt{0.28889/10}} = 2.35$

at 5% significance level, $\nu = 10 - 1 = 9$; Critical value $t = 1.833$.

$2.35 > 1.833$; hence Reject H_0; Sufficient evidence to accept Farmer B's claim.

Hypothesis tests for the difference between two population means.

§ We need to carry out a hypothesis test to compare two populations and determine whether there is a difference between the two means.

Here we can use the test statistic 'z' or the test statistic 'T' depending on the sample size and available data.

Case I § Testing two sample means using the normal distribution: $(n>30)$

The following assumptions are required:

- Two samples are independent,
- Each sample is randomly selected from a population that is normally distributed:
- The population variance is known or can be calculated by the Central Limit Theorem.

$X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$,

Hence $\bar{X}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right)$ and $\bar{X}_2 \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$

$(\bar{X}_1 - \bar{X}_2)$ is also normally distributed: $\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$

and the test statistic Z is given by: $Z = \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

under the null hypothesis when $\mu_1 - \mu_2 = 0$;

the test statistic is $Z = \dfrac{(\bar{X}_1 - \bar{X}_2) - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

Example 7: A company has two different machines, X and Y, each of which fills empty cups with coffee. The manager is investigating the volumes of coffee, x and y, measured in appropriate units, in the cups filled by machines X and Y respectively. She chooses a random sample of 50 cups filled by machine X and a random sample of 40 cups filled by machine Y. The volumes are summarised as:

$\Sigma x = 15.2$, $\Sigma x^2 = 5.1$, $\Sigma y = 13.4$, $\Sigma y^2 = 4.8$

The manager claims that there is no difference between the mean volume of coffee in cups filled by machine X and the mean volume of coffee in cups filled by machine Y.

Test the manager's claim at the 10% significance level. ---[9]

S-20|41|04

Solution: $H_0: \mu_x = \mu_y$; $H_1: \mu_x \neq \mu_y$, $n_x = 50$, $n_y = 40$ (>30)

$$\bar{x} = \frac{\Sigma x}{n_x} = \frac{15.2}{50} = 0.304 \; ; \quad \bar{y} = \frac{\Sigma y}{n_y} = \frac{13.4}{40} = 0.334$$

$$s_x^2 = \frac{1}{n_x - 1}\left(\Sigma x^2 - \frac{(\Sigma x)^2}{n_x}\right) = \frac{1}{49}\left(5.1 - \frac{15.2^2}{50}\right) = 0.0097796$$

$$s_y = \frac{1}{(n_y - 1)}\left(\Sigma y^2 - \frac{(\Sigma y)^2}{n_y}\right) = \frac{1}{39}\left(4.8 - \frac{13.4^2}{40}\right) = 0.007974$$

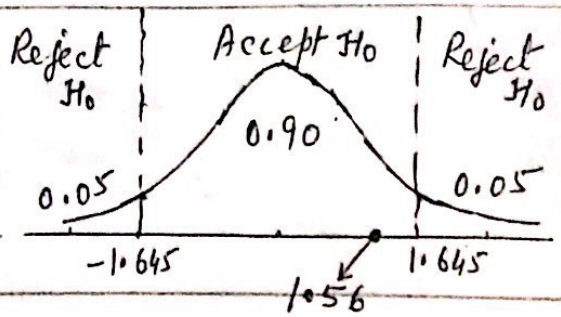
$$s^2 = \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y} = \frac{0.0097796}{50} + \frac{0.007974}{40} = 0.0003949 \checkmark$$

Test statistic $z = \dfrac{(\mu_x - \mu_y) - 0}{\sqrt{s^2}} = \dfrac{0.304 - 0.335}{\sqrt{0.0003949}} = (-)1.56$ (1.05)

Now for two-tailed test at 10% significant level with $p = 0.95$

critical value $= 1.645$.

$1.56 < 1.645 \Rightarrow$ Accept H_0. $\Rightarrow$ Insufficient evidence to reject manager's claim.



Case II

§ Testing two sample means Using t-distribution:

The following assumption are made to use t-distributions:

- The two samples with relatively small sample size are independent.
- Each sample is randomly selected from a population that is normally distributed.
- The population variance of both samples is the same – $\sigma_1^2 = \sigma_2^2$

Here use (Pooled estimate) $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$ $\left[s_p^2 = \frac{\Sigma(x_1-\bar{x}_1)^2 + \Sigma(x_2-\bar{x}_2)^2}{n_1+n_2-2} \right.$

The test statistic: $T = \frac{(\bar{x}_1 - \bar{x}_2) - d}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$

Example 8: Students at two colleges, A and B, are completing in a computer games challenge.

(a) The time taken for a randomly chosen student from college A to complete the challenge has a normal distribution with mean μ min. The time taken, x minutes, are recorded for a random sample of 10 students chosen from college A. The results are as follows: $\Sigma x = 828$, $\Sigma x^2 = 68622$; A test is carried out on the data at 5% significance level and the result support the claim that $\mu > k$. Find the greatest possible value of k. --[4]

(b) A random sample of 8 students is chosen from college B. Their times to complete the same challenge give a sample mean of 79.8 minutes and an unbiased variance estimate of 9.966 minutes^2. Use a two sample test at 5% significance level to test whether the mean time for students at college B to complete the challenge is same as the mean time for students at college A to compete the challenge. You should assume that the two distributions are normal and have the same population variance. [S-20/43/Q5] --[7]

(Solution on next page)

(Continued example 8)

Solution: $\Sigma x = 828$, $n = 10$, $\bar{x} = \frac{\Sigma x}{n} = \frac{828}{10} = 82.8$ ✓; $\Sigma x^2 = 68622$

(a) $H_0: \mu = 82.8$ and $H_1: \mu > k$; $s^2 = \frac{1}{n-1}\left(\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right) = \frac{1}{9}\left(68622 - \frac{828^2}{10}\right)$ ✓

$s^2 = 7.0667$ ✓

Test Statistic, $T = \frac{\bar{x} - \mu}{\sqrt{s^2/n}} = \frac{82.8 - k}{\sqrt{7.0667/10}} = \frac{82.8 - k}{0.8406}$ ---(i)

for one tailed test at 5% significance level, $p = 0.95$ and $\nu = 10 - 1 = 9$

critical value $t = 1.833$ ---(ii)

for $\mu > k$, $\frac{82.8 - k}{0.8406} \geq 1.833 \Rightarrow 82.8 - k \geq 1.54 \Rightarrow k \leq 81.259$

from (i) & (ii) $\Rightarrow$

$\therefore$ Greatest value of $k = 81.3$ (3 sf) ✓

(b) $H_0: \mu_A = \mu_B$ and $H_1: \mu_A \neq \mu_B$, $n_1 = 10$, $n_2 = 8$

$\bar{x}_A = 82.8$, $\bar{x}_B = 79.8$, $s_A^2 = 7.0667$ and $s_B^2 = 9.966$

Pooled variance $s_p^2 = \frac{(n_1 - 1)s_A^2 + (n_2 - 1)s_B^2}{n_1 + n_2 - 2}$

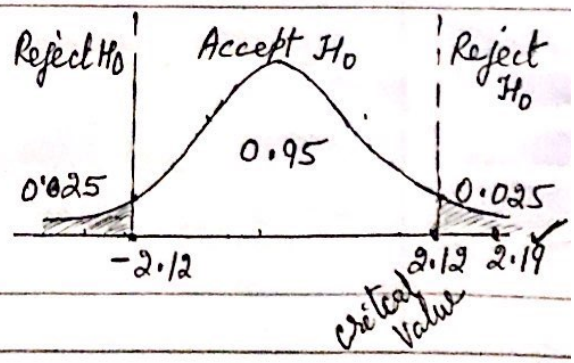
$s_p^2 = \frac{9 \times 7.0667 + 7 \times 9.966}{10 + 8 - 2} = 8.335$

Test statistic $T = \frac{(\bar{x}_A - \bar{x}_B) - 0}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{82.8 - 79.8}{\sqrt{8.335\left(\frac{1}{10} + \frac{1}{8}\right)}} = 2.19$ ✓ ---(iii)

Now for two tailed test, 5% significance level $p = 0.975$ $\left(1 - \frac{0.05}{2}\right)$ and $\nu = 10 + 8 - 2 = 16$, Critical value $t = 2.12$ --(iv)

$2.19 > 2.12 \Rightarrow$ Reject H_0.

Hence the population means are not same.



Example 9: A scientist is investigating the masses of a particular type of fish found in lakes A and B. He chooses a random sample of 10 fish of this type from lake A and records their masses, x kg, as: 2·1, 1·8, 0·9, 3·0, 2·4, 2·6, 1·8, 2·2, 1·9, 2·5

The scientist also chooses a random sample of 12 fish of this type from lake B, but he only has a summary of their masses, y kg, as:

$$\Sigma y = 24.48 \quad ; \quad \Sigma y^2 = 53.75$$

Test at the 10% significance level whether the mean mass of fish of this type in lake A is greater than the mean mass of fish of this type in lake B. You should state any assumptions that you need to make for the test to be valid. [W-21/42/Q6] --[10]

Solution: $H_0: \mu_A = \mu_B$ and $H_1: \mu_A > \mu_B$; $\Sigma x = 21.2$, $n_x = 10$, $\bar{x} = 2.12$ ✓

$\Sigma x^2 = 47.92$; $\bar{y} = \frac{24.48}{12} = 2.04$ ✓

$$s_x^2 = \frac{1}{n_x - 1}\left(\Sigma x^2 - \frac{(\Sigma x)^2}{n_x}\right) = \frac{1}{9}\left(47.92 - \frac{(21.2)^2}{10}\right) = 0.33067 \checkmark$$

$$s_y = \frac{1}{n_y - 1}\left(\Sigma y^2 - \frac{(\Sigma y)^2}{n_y}\right) = \frac{1}{11}\left(53.75 - \frac{24.48^2}{12}\right) = 0.34644 \checkmark$$

Pooled Variance $s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2} = \frac{9 \times 0.33067 + 11 \times 0.34644}{10 + 12 - 2}$

$$s_p^2 = 0.3393 \checkmark$$

Test statistic $T = \frac{(\bar{x} - \bar{y}) - 0}{\sqrt{s_p^2\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}} = \frac{2.12 - 2.04}{\sqrt{0.3393\left(\frac{1}{10} + \frac{1}{12}\right)}} = 0.321 \checkmark$

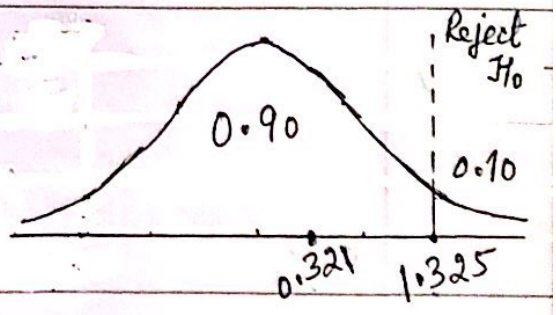
Now for one tailed test at 10% significance level $p = 0.90$, $v = 10 + 12 - 2 = 20$ ✓

Critical value $t = 1.325$

$0.321 < 1.325 \Rightarrow$ accept H_0.

Insufficient evidence that mean of A is greater than mean of B.

or (H_1 is rejected)



Assumptions: Distributions are normal and equal variances.

or Distributions are Normal.

Example 10: A year group of students took a mock science exam at the end of term. Their teach thinks that on average boys performed better than girls by at least 2 marks. A random sample of 60 students, comprising 30 boys and 30 girls, was selected and each mock exam score, X, was recorded. The results are as:

	Σx	Σx^2
Boys	2474	206044
Girls	2380	191094

Test, at the 5% significance level, whether these results support the teacher's claim.

Solution: $H_0 : \mu_B - \mu_G = 2$; $H_1 : \mu_B - \mu_G < 2$, $n_B = 30 = n_G$

$\bar{x}_B = \frac{2474}{30} = 82.47$, $\bar{x}_y = \frac{2380}{30} = 79.33$

$S_B = \frac{1}{n-1}\left(\Sigma x_B^2 - \frac{(\Sigma x_B)^2}{n_B}\right) = \frac{1}{29}\left(206044 - \frac{2474^2}{30}\right) = 69.71$ ✓

$S_G = \frac{1}{29}\left(191094 - \frac{2380^2}{30}\right) = 78.64$ ✓

The test statistic $Z = \frac{(\bar{x}_B - \bar{x}_G) - d}{\sqrt{\frac{S_B^2}{n_1} + \frac{S_G^2}{n_2}}} = \frac{(82.47 - 79.33)}{\sqrt{\frac{69.71}{30} + \frac{78.64}{30}}} = 0.5\,10$

Now one tailed test at 5% significance level with $p = 0.05$.
The critical value from the normal distribution table is -1.645.

The test statistic $0.510 > -1.645$, lies within the acceptance region. So H_0 accepted,
The evidence supports the claim that boys scored at least 2 marks more than girls in the mock science exam.

Paired sample t-test.

§ A paired sample t-test is applied to compare two means that come from the same sample, or population measured at two different points in time - that is before and after an experiment.

The following assumptions should hold.

- The sample is independent and randomly selected from the population.
- The difference between the paired values is normally distributed.
- There are no outliers in the difference between the two selected groups.

The difference between the paired data is normally distributed:

$$\bar{X}_d \sim N(\mu_d, \sigma_d^2)$$

where μ_d is the mean of the differences and σ_d^2 is the variance of the differences.

Under the null hypothesis $\mu_d = \mu_1 - \mu_2$

The statistic is: $T = \dfrac{\bar{X}_d - \mu_d}{\sqrt{\dfrac{S_d^2}{n}}}$ $\left\{ S_d^2 = \dfrac{1}{(n-1)}\left(\Sigma d^2 - \dfrac{(\Sigma d)^2}{n}\right) \right.$

and the critical value of the t-distribution with $\nu = (n-1)$ degrees of freedom.

Example 11: Members of the sprints athletics club have been taking part in an intense training scheme, aimed at reducing their times taken to run 400m. For a random sample of 9 athletes from the club, the times taken, in seconds, before and after the training scheme are given as:

Athlete	A	B	C	D	E	F	G	H	I
Time before	48.8	48.2	50.3	49.6	49.4	48.9	47.6	50.3	48.4
Time after	47.9	47.8	49.6	49.1	49.6	48.9	47.7	49.1	48.1

The organiser of the training scheme claims that on average an athlete's time will be reduced by at least 0.3 seconds

Test at 10% significance level whether the organiser's claim is justified, stating any assumption that you make. ...[8]

[W-20|41|Q4]

Solution: Assume (population) differences are normally distributed.

$H_0: \mu_x - \mu_y = 0.3$; $H_1: \mu_x - \mu_y > 0.3$, $n = 9$

Differences: 0.9, 0.4, 0.7, 0.5, 0.0, −0.1, 1.2, 0.3

$\Sigma d = 3.7$, $\Sigma d^2 = 3.29$; $\bar{d} = \frac{\Sigma d}{n} = \frac{3.7}{9} = 0.411$,

$s_d^2 = \frac{1}{(n-1)}\left(\Sigma d^2 - \frac{\Sigma d^2}{n}\right) = \frac{1}{9}\left(3.29 - \frac{3.7^2}{9}\right) = 0.2211$

The statistic $T = \frac{\bar{d} - (\mu_x - \mu_y)}{\sqrt{\frac{s_d^2}{n}}} = \frac{0.411 - 0.3}{\sqrt{\frac{0.2211}{9}}} = 0.708$ ✓

Now it is one tailed test at 10% significance level with $p = 0.90$ and $\nu = 9 - 1 = 8$; critical value is $t = 1.397$ ✓

$0.708 < 1.397 \Rightarrow$ Accept H_0.

Insufficient evidence to accept the claim.

Example 12: Sunflower seeds of the same type were planted in two different types of soils, X and Y, at the same time. Nine sunflowers from each site were selected at the end of the season and the diameter of each sunflower was measured in cm, as:

Soil X	6.3	7.2	5.9	6.8	7.6	8.3	5.1	7.4	8.1
Soil Y	7.6	5.1	5.8	6.9	7.2	6.4	6.7	5.5	7.8

(a) Find a pooled estimate of the population variance for the two samples.

(b) Test at 5% significance level whether the sunflowers planted in soil X are larger than those planted in soil Y. State any assumption made.

Solution: $\Sigma x = 62.7$, $\Sigma x^2 = 445.61$, $\Sigma y = 59$, $\Sigma y^2 = 393.8$, $n = 9$

(a) $\bar{x} = \frac{\Sigma x}{n} = 6.967$, $\bar{y} = \frac{\Sigma y}{n} = 6.556$,

$$s_x^2 = \frac{1}{n-1}\left[\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right] = \frac{1}{8}\left[445.61 - \frac{62.7^2}{9}\right] = 1.1$$

$$s_y^2 = \frac{1}{n-1}\left[\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right] = \frac{1}{8}\left[393.8 - \frac{59^2}{9}\right] = 0.8778$$

Pooled variance $s_p^2 = \frac{(n_1-1)s_x^2 + (n_2-1)s_y^2}{n_1+n_2-2} = \frac{(9-1)\times 1.1 + (9-1)\times 0.8778}{9+9-2} = 0.989$ ✓

(b) Assume that both the samples are independent and randomly selected. They have the same population variance.

$H_0: \mu_x - \mu_y = 0$ and $H_1: \mu_x - \mu_y > 0$

The test statistic $T = \frac{(\bar{x}-\bar{y}) - 0}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{(6.967-6.556)-0}{\sqrt{0.989\left(\frac{1}{9}+\frac{1}{9}\right)}} = 0.877$

Now one tailed test at 5% significance level $p = 0.95$, $V = 9+9-2 = 16$

the critical value of $t = 1.746$

As $0.877 < 1.746$, the test statistic T lies within the acceptance region. So we accept H_0.

There is no evidence to suggest that the sunflowers planted in soil X are larger than those planted in soil Y.

§ We will consider t-distribution to determine confidence intervals for population mean, μ, when a small normally normally distributed random sample ($n < 30$) with unknown population variance is used.

Confidence interval: $\bar{x} - t \cdot \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t \cdot \frac{s}{\sqrt{n}}$

where $\bar{x}$ = Sample mean, s^2 = unbiased estimate of population variance $= \frac{1}{(n-1)}\left(\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right)$

n = number of random sample.

t = critical value for the t-distribution, with degree of freedom $\nu = n-1$, at the given significance level.

Example 13: A large number of children are competing in a throwing competion. The distances, in metres, thrown by a random sample of 8 children are as follows:

19.8, 22.1, 24.4, 21.5, 20.8, 26.3, 23.7, 25.0

Find a 95% confidence interval for the population mean distance thrown. S-20/41/Q5(b) --[3]

Solution: $\Sigma x = 183.6$, $\Sigma x^2 = 4249.08$, $n = 8$, $\bar{x} = \frac{\Sigma x}{n} = 22.95$ ✓

unbiased estimate of population variance: $s^2 = \frac{1}{n-1}\left(\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right)$.

$\Rightarrow s^2 = \frac{1}{7}\left(4249.08 - \frac{183.6^2}{8}\right) = 5.066$ ✓

(Assuming sample is normally distributed)

Now at 95% significance level $p = 0.975$, $\nu = 8-1 = 7$

The critical value $t = 2.365$

$\therefore$ Confidence interval is: $\bar{x} \pm t \cdot \sqrt{\frac{s^2}{n}}$

$= 22.95 \pm 2.365 \times \sqrt{\frac{5.066}{8}} = 22.95 \pm 1.88$

$21.07 \leq \mu \leq 24.83$

or $21.1 \leq \mu \leq 24.8$ (3 sf)

Example 14: The time, t minutes, taken by fifteen students in a class to complete a puzzle was recorded. The results are as:

$$\Sigma t = 79.5 \text{ and } \Sigma t^2 = 451.6$$

(a) Stating any assumptions, calculate a 95% confidence interval for the population mean time for students to complete the puzzle.

(b) The manufacturers of the puzzle indicate a mean completion time of eight minutes. What conclusion might be drawn about the students in the class?

Solution: Assuming sample is normally distributed.

$$\bar{x}_t = \frac{\Sigma t}{n} = \frac{79.5}{15} = 5.3 \quad , n = 15$$

Unbiased estimate of population variance $s^2 = \frac{1}{n-1}\left(\Sigma t^2 - \frac{(\Sigma t)^2}{n}\right)$

$$s^2 = \frac{1}{14} \cdot \left(451.6 - \frac{79.5^2}{15}\right) = 2.161 \checkmark$$

Now at 95% confidence interval, $p = 0.975$, $v = 15 - 1 = 14$

Critical value $t = 2.145$ ✓

$\therefore$ The confidence interval is $\bar{x} \pm t\sqrt{\frac{s^2}{n}}$

$$= 5.3 \pm 2.145\sqrt{\frac{2.161}{15}}$$

$$= 5.3 \pm 0.81$$

$$\underline{4.49 \leq \mu \leq 6.11} \checkmark \qquad [4.49, 6.11]$$

(b) You are 95% confident that the mean time of completing the puzzle is between 4.49 and 6.11 minutes. That means the students can complete the puzzle more quickly than the time that the manufacturer suggested (8 minutes).

Confidence interval for a difference of population means. P-16

§ Let us have the assumptions:

- both random samples are normally distributed. ($n > 30$)
- the population variances are known or unbiased estimates of population variance can be used.

$(\bar{X}_1 - \bar{X}_2)$ is then also normally distributed with:

Case I

$$(\bar{X}_1 - \bar{X}_2) \sim \left(\mu_1 - \mu_2,\ \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

where μ_1, μ_2 are the means of the two populations.

$\bar{X}_1, \bar{X}_2$ are the means of the two samples.

σ_1^2, σ_2^2 are the population variances; n_1, n_2 are the sample sizes.

Then the confidence interval for the difference of two population means is: $(\bar{X}_1 - \bar{X}_2) \pm Z \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

where Z is the critical value, depending on the level of significance.

Case II

If two small samples ($n < 30$) with unknown population variance are taken, and the population variances can be assumed to be equal, the t-distribution with $(n_1 + n_2 - 2)$ degrees of freedom is considered for the determination of the confidence interval. and the pooled estimate variance S_p^2 is used.

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)}$$

$$\begin{cases} S_1^2 = \frac{1}{n_1 - 1}\left[\Sigma x_1^2 - \frac{(\Sigma x_1)^2}{n_1}\right] \\ \text{and} \\ S_2^2 = \frac{1}{n_2 - 1}\left[\Sigma x_2^2 - \frac{(\Sigma x_2)^2}{n_2}\right] \end{cases}$$

The confidence interval for the differences of two population means is given as:

$$(\bar{X}_1 - \bar{X}_2) \pm t \cdot \sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Example 15: The number, x, of a certain type of sea shell was counted at 60 randomly chosen sites, each one metre square, along the coastline in country A. The number, y, of the same type of sea shell was counted at 50 randomly chosen sites, each one metre square, along the coastline in country B. The results are as follows, where $\bar{x}$ and $\bar{y}$ denote the sample means of x and y respectively.

$\bar{x} = 29.2$, $\Sigma(x-\bar{x})^2 = 4341.6$, $\bar{y} = 24.4$, $(y-\bar{y})^2 = 3732.0$

Find a 95% confidence interval for the difference between the mean number of sea shells, per square metre, on the coastlines in country A and in country B. --[7]

SP-20|04|Q4

Solution: Sample variance $s_x^2 = \frac{\Sigma(x-\bar{x})^2}{n_1 - 1} = \frac{4341.6}{59} = 73.59$ ✓

and $s_y = \frac{\Sigma(y-\bar{y})^2}{n_2 - 1} = \frac{3732.0}{49} = 76.16$

Variance for difference of means $s^2 = \frac{s_x^2}{n_1} + \frac{s_y^2}{n_2} = \frac{73.59}{60} + \frac{76.16}{50}$

$\Rightarrow s^2 = 1.2265 + 1.5232 =$

$s^2 = 2.7497$

$\Rightarrow s = 1.658$ ✓

Now at 95% confidence interval, $p = 0.975$, hence the critical value $z = 1.96$

∴ The confidence interval is $(\bar{x} - \bar{y}) \pm z \cdot s$

$= (29.2 - 24.4) \pm 1.96 \times 1.658$

$= 4.8 \pm 3.25$ $[1.55, 8.05]$

$1.55 \leq \mu_d \leq 8.05$ ✓

Example 16: The heights, x m, of a random sample of 50 adult males from country A were recorded. The heights, y m, of a random sample of 40 adult males from country B were also recorded. The results are:

$\Sigma x = 89.0$, $\Sigma x^2 = 159.4$, $\Sigma y = 67.2$ $\Sigma y^2 = 113.1$

Find a 95% confidence interval for the difference between the mean heights of adult males from country A and adult males from country B.

S-21 | 43 | Q3 ---[8]

Solution: $\Sigma x = 89.0$, $\Sigma x^2 = 159.4$, $n_A = 50$; $\Sigma y = 67.2$, $\Sigma y^2 = 113.1$, $n_B = 40$

$\bar{x} = \frac{89}{50} = 1.78$; $\bar{y} = \frac{67.2}{40} = 1.68$

∴ Unbiased estimate of population variances;

$\sigma_x^2 = \frac{1}{n_A - 1}\left[\Sigma x^2 - \frac{(\Sigma x)^2}{n_A}\right] = \frac{1}{49}\left[159.4 - \frac{89^2}{5}\right] = 0.02$

$\sigma_y^2 = \frac{1}{n_B - 1}\left[\Sigma y^2 - \frac{(\Sigma y)^2}{n_B}\right] = \frac{1}{39}\left[113.1 - \frac{67.2^2}{40}\right] = 0.0052308$

Variance for difference of means $\sigma^2 = \frac{\sigma_x^2}{n_A} + \frac{\sigma_y^2}{n_B} = \frac{0.02}{50} + \frac{0.0052308}{40}$

$\Rightarrow \sigma^2 = 0.0005308$

Now at 95% confidence interval, $p = 0.975$

Hence critical value $z = 1.96$

∴ The confidence interval $= (\bar{x} - \bar{y}) \pm z \cdot \sqrt{\sigma^2}$

$= (1.78 - 1.68) \pm 1.96 \times \sqrt{0.0005308}$

$= 0.1 \pm 0.0452$

or $[0.0548, 0.145]$

or $0.0548 \leq \mu_d \leq 0.145$

Example 17: A school teacher is comparing the numbers of hours spent watching TV each week by boys and by girls. He randomly picked seven boys and nine girls. The hours spent watching TV each week by each pupil are recorded as follows:

Boys	7.5	5	6.5	8.6	7.5	12	6.5	-	-
Girls	7	8.5	6	6.5	9	7.5	7	8.5	8

(a) State any assumptions made.

(b) Calculate a 95% confidence interval for the difference in the mean hours spent watching TV each week by boys and girls.

(c) Comment on the results.

Solution (a) Assume population variances of the times spent watching TV by boys and girls are equal.

(b) $\bar{x}_B = \frac{\Sigma x}{n_1} = \frac{53.6}{7} = 7.657$; $\bar{y}_G = \frac{\Sigma y}{n_2} = \frac{68}{9} = 7.556$, $n_1 = 7$, $n_2 = 9$

Unbiased estimate of population variance. $s_B^2 = \frac{1}{n_1 - 1}\left(\Sigma x^2 - \frac{(\Sigma x)^2}{n_1}\right)$

$= \frac{1}{6}\left(439.96 - \frac{53.6^2}{7}\right) = 2.219$ ✓

and $s_G^2 = \frac{1}{n_2 - 1}\left[\Sigma y^2 - \frac{(\Sigma y)^2}{n_2}\right] = \frac{1}{8}\left[522 - \frac{68^2}{9}\right] = 1.014$ ✓

The pooled estimate of variance $s_p^2 = \frac{(n_1 - 1)s_B^2 + (n_2 - 1)s_G^2}{(n_1 + n_2 - 2)}$

$s_p^2 = \frac{6 \times 2.219^2 + 8 \times 1.014^2}{7 + 9 - 2} = 2.698$

Now at 95% confidence interval, $p = 0.975$ and $\nu = 7 + 9 - 2 = 14$

critical value $t = 2.145$,

$\therefore$ Confidence interval: $(\bar{X}_B - \bar{Y}_G) \pm t^* \sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$

$= (7.657 - 7.556) \pm 2.145\sqrt{2.698\left(\frac{1}{7} + \frac{1}{9}\right)}$

$-1.67 \leq \mu_d \leq 1.88$ ✓ or $[-1.67, 1.88]$

(c) We are 95% confident that, on average, the difference between the times spent by boys and girls watching TV each week lies in this interval. As the interval contains the value zero, there is no significant difference between boys and girls.

Example 18: A random sample of first year students is selected from two secondary schools, comprising 12 students from school A and 15 students from school B. A test is given to these students. Students from school A have an average score of 63 with standard deviation of 5.9. Students from school B have an average score of 57 with a standard deviation of 4.1. Assuming that the test scores came from normal distributions in both schools.

(a) Calculate a pooled estimate of the population variance for the two samples.

(b) Calculate a 90% confidence interval for the difference in the test scores at the two schools.

Solution: $n_1 = 12$, $\bar{x} = 63$, $\sigma_x = 5.9$ & $n_2 = 15$, $\bar{y} = 57$, $s_y = 4.1$

(a) The pooled estimate of variance $S_p^2 = \dfrac{(n_1-1)\sigma_x^2 + (n_2-1)\sigma_y^2}{n_1+n_2-2}$

$$S_p^2 = \frac{11 \times 5.9^2 + 14 \times 4.1^2}{12+15-2} = \underline{24.73} \checkmark$$

(b) Now at 90% confidence interval, $p = 0.95$, $\nu = 12+15-2 = 25$

Critical value $t = 1.708$.

$\therefore$ The confidence interval is: $(\bar{x}-\bar{y}) \pm t \cdot \sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$

$$= (63-57) \pm 1.708 \times \sqrt{24.73\left(\frac{1}{12}+\frac{1}{15}\right)}$$

$$\Rightarrow \quad \underline{2.71 \leq \mu_d \leq 9.29} \checkmark$$