Inference on Difference in Means of Two Independent Normal Populations

Case – I When population variances σ_1^2 *and* σ_2^2 are known-

#.1 **Two machines are used for filling plastic bottles with a net volume of 16.0 ounce. The filling process can be assumed normal with standard deviations of $\sigma_1 = 0.015$ and $\sigma_2 = 0.018$. Quality engineering suspects that both machines fill to the same net volume, whether or not this volume is 16.0 ounce. Random samples of 10 units from each machine are observed to fill to the following volumes-**

Machine 1		Machine 2	
16.03	16.01	16.02	16.03
16.04	15.96	15.97	16.04
16.05	15.98	15.96	16.02
16.05	16.02	16.01	16.01
16.02	15.99	15.99	16.00

Do you think that quality engineering is correct? Test at $\alpha = 0.05$. Also construct 95% confidence interval for difference in mean volumes of two machines.

Solution-

Let μ_1 *and* μ_2 denote actual (or true) means of fill volumes of Machine1 and Machine2, respectively.

$$H_0: \mu_1 = \mu_2 \ \ (or\ the\ suspect\ of\ quality\ enginerring\ is\ not\ true)$$

$$H_1: \mu_1 \neq \mu_2 \ \ (or\ the\ suspect\ of\ quality\ enginerring\ is\ true)$$

Here, population standard deviations are given, so the test statistic is

$$Z_0 = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Here,

$$\overline{X_1} = \frac{1}{10}(16.03 + 16.01 + \cdots \dots \dots \dots \dots \dots . +15.99) = 16.015$$

$$\overline{X_2} = \frac{1}{10} \times (16.02 + 16.03 + \cdots \dots \dots \dots . 16.00) = 16.005$$

Also, $\sigma_1 = 0.015$ *and* $\sigma_2 = 0.018$.

Hence,

$$Z_0 = \frac{16.015 - 16.005}{\sqrt{\frac{0.015^2}{10} + \frac{0.018^2}{10}}} = \frac{0.01}{0.0074} = 1.35$$

It is two-tailed test, so critical values are $Z_{\frac{0.05}{2}} = 1.96 \ \ and - Z_{\frac{0.05}{2}} = -1.96$

Since, it is not true that $Z_0 \geq 1.96$ and it is not also true that $Z_0 \leq -1.96$, so the null hypothesis is not rejected and it is concluded that the suspect of quality engineering is not correct.

Next, 95% confidence interval for difference in mean volumes of two machines is given by

$$P\left((\overline{X_1} - \overline{X_2}) - Z_{\frac{\alpha}{2}}. \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq (\overline{X_1} - \overline{X_2}) + Z_{\frac{\alpha}{2}}. \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right) = 1 - \alpha$$

$$P\left((16.015 - 16.005) - Z_{\frac{0.05}{2}}. \sqrt{\frac{0.015^2}{10} + \frac{0.018^2}{10}} \leq \mu_1 - \mu_2 \right.$$

$$\left. \leq (16.015 - 16.005) + Z_{\frac{0.05}{2}}. \sqrt{\frac{0.015^2}{10} + \frac{0.018^2}{10}} \right) = 1 - 0.05$$

$$P(0.01 - 1.96 \times 0.0074 \leq \mu_1 - \mu_2 \leq 0.01 + 1.96 \times 0.0074) = 0.95$$

$$P(0.01 - 0.0145 \leq \mu_1 - \mu_2 \leq 0.01 + 0.0145) = 0.95$$

$$P(-0.0045 \leq \mu_1 - \mu_2 \leq 0.0245) = 0.95$$

2. The plant manager of an orange juice canning facility is interested in comparing the performance of two different production lines in her plant. As the line number 1 is relatively new, she suspects that its output in number of cases per day is greater than the number of cases produced by the older line 2. Ten days of data are selected at random for each line, for which it is found that $\overline{x_1}$ = 824.9 cases per day and $\overline{x_2}$ = 818.6 cases per day. From experience with this type of equipment it is known that $\sigma_1^2 = 40$ and $\sigma_2^2 = 50$.
Test the null hypothesis $H_0 : \mu_1 = \mu_2$ against alternative hypothesis $H_1 : \mu_1 > \mu_2$.

[Hints: $|Z_0|$=2.10, H_0 Rejected]

3. Two different formulations of a lead-free gasoline are being tested to study their road octane numbers. The variance of road octane number for formulation 1 is $\sigma_1^2 = 1.5$ and for formulation 2 it is σ_2^2=1.2. Two random sample of size $n_1 = 15$ and $n_2 = 20$ are tested, and the mean road octane number observed are $\bar{x}_1 = 89.6$ and $\bar{x}_2$ = 92.5. If formulation 2 produces a higher road octane number than formulation 1, the manufacturer would like to detect this. Formulate and test an appropriate hypothesis, using α = 0.05.
[Hints: $|Z_0|$ = 6.32, H_0 rejected against H_1: $\mu_1<\mu_2$]

Case – II When population variances σ_1^2 *and* σ_2^2 are unknown-

Sub-case (a) When samples are large, i.e., when $n_1 > 30$ *or* $n_2 > 30$

1. **To test the claim that the resistance of electric wire can be reduced by more than 0.050 ohm by alloying, 32 values obtained for standard wire yielded $\overline{X_1} = 0.136$ ohm and s_1 = 0.004 ohm, and 32 values obtained for alloyed wire yielded $\overline{X_2} = 0.083$ ohm and s_2 = 0.005 ohm. At the 0.05 level of significance, does this support the claim? Also, obtain the 95% confidence interval for reduction in resistance of electric wire due to alloying.**

Solution-

Let μ_1 *and* μ_2 denote actual means of resistance of standard wire and of alloyed wire respectively.

Here, the null and alternative hypotheses:

$$H_0: \mu_1 - \mu_2 = 0.050 \ \ (Or, claim\ is\ not\ true)$$

$$H_1: \mu_1 - \mu_2 > 0.050 \ \ (Or, claim\ is\ true)$$

Since population variances are not known and samples are of large size, so the test statistic is

$$Z_0 = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Here, we are given

$$\overline{X_1} = 0.136 \qquad s_1 = 0.004 \qquad n_1 = 32$$

$$\overline{X_2} = 0.083 \qquad s_1 = 0.005 \qquad n_1 = 32$$

Hence,

$$Z_0 = \frac{(0.136 - 0.083) - 0.050}{\sqrt{\frac{0.004^2}{32} + \frac{0.005^2}{32}}} = \frac{0.053}{0.0011} = 2.56$$

It is right-tailed test, so critical value is $Z_{0.05} = 1.64$.

Since $Z_0 > 1.64$, so the null hypothesis is rejected and it is concluded that claim is true.

Next, confidence interval is given by

$$P\left((\overline{X_1} - \overline{X_2}) - Z_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \le \mu_1 - \mu_2 \le (\overline{X_1} - \overline{X_2}) + Z_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right) = 1 - \alpha$$

$$P\left((0.136-0.083)-Z_{\frac{0.05}{2}}\sqrt{\frac{0.004^2}{32}+\frac{0.005^2}{32}}\leq\mu_1-\mu_2\right.$$

$$\left.\leq(0.136-0.083)+Z_{\frac{0.05}{2}}\sqrt{\frac{0.004^2}{32}+\frac{0.005^2}{32}}\right)=1-0.05$$

$$P(0.053-1.96\times0.0011\leq\mu_1-\mu_2\leq0.053+1.96\times0.0011)=0.95$$

$$P(0.0508\leq\mu_1-\mu_2\leq0.0552)=0.95$$

Subcase (b) When samples are small, i.e., $n_1 \leq 30\ and\ n_2 \leq 30$

#.1 **Following are the number of sales which a sample of 9 salespersons of industrial chemicals in California and a sample of 6 salespersons of industrial chemicals in Oregon made over a certain fixed period of time**

California	**59**	**68**	**44**	**71**	**63**	**46**	**69**	**54**	**48**
Oregon	**50**	**36**	**62**	**52**	**70**	**41**			

Assuming that populations sampled are normally distributed with unknown means μ_1 and μ_2 but having common variance, construct 95% confidence interval for difference of means of sales in two cities. Also test whether the mean sales in California is significantly greater than in Oregon at 5% level of significance.

Solution – Let $\mu_1 and\ \mu_2$ denote actual means of sales in California and in Oregon.

Her, the null and alternative hypotheses are

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 > \mu_2$$

Here, population variances are unknown, samples are small and population variances are given to be equal, so the test statistic is

$$t_0 = \frac{\overline{X_1}-\overline{X_2}}{\sqrt{\frac{S_p^2}{n_1}+\frac{S_p^2}{n_2}}}$$

Working Table

X_1	X_2	X_1^2	X_2^2
59	50	3481	2500
68	36	4624	1296

44	62	1936	3844
71	52	5041	2704
63	70	3969	4900
46	41	2116	1681
69		4761	
54		2916	
48		2304	
522	311	31148	16925

Calculation of sample means

$$\overline{X_1} = \frac{1}{n_1}\sum X_1 = \frac{1}{9} \times 522 = 58$$

$$\overline{X_2} = \frac{1}{n_2}\sum X_2 = \frac{1}{6} \times 311 = 51.83$$

Calculation of sample variances

$$S_1^2 = \frac{n_1}{n_1 - 1}\left(\frac{1}{n_1}\sum X_1^2 - \overline{X_1}^2\right) = \frac{9}{9-1}\left(\frac{1}{9} \times 31148 - 58^2\right)$$

$$= \frac{9}{8} \times 96.89 = 109$$

$$S_2^2 = \frac{n_2}{n_2 - 1}\left(\frac{1}{n_2}\sum X_2^2 - \overline{X_2}^2\right) = \frac{6}{6-1}\left(\frac{1}{6} \times 16925 - 51.83^2\right)$$

$$= \frac{6}{5} \times 134.139 = 160.97$$

Hence,

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = \frac{(9-1) \times 109 + (6-1) \times 160.97}{9 + 6 - 2} = 128.99$$

Hence, test statistic is

$$t_0 = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}} = \frac{58 - 51.83}{\sqrt{\frac{128.99}{9} + \frac{128.99}{6}}} = \frac{6.17}{5.9859} = 1.0307$$

It is right-tailed test, so critical value is $t_{\alpha, n_1 + n_2 - 2} = t_{0.05, 9+6-2} = 1.77$

Since it is not true that $t_0 \geq 1.77$, so the null hypothesis is not rejected.

Next, confidence interval is given by

$$P\left((\overline{X_1}-\overline{X_2})-t_{\frac{\alpha}{2},n_1+n_2-2}\cdot\sqrt{\frac{S_p^2}{n_1}+\frac{S_p^2}{n_2}}\le\mu_1-\mu_2\le(\overline{X_1}-\overline{X_2})+t_{\frac{\alpha}{2},n_1+n_2-2}\cdot\sqrt{\frac{S_p^2}{n_1}+\frac{S_p^2}{n_2}}\right)=1-\alpha$$

$$P\left((58-51.83)-t_{\frac{0.05}{2},9+6-2}\sqrt{\frac{128.99}{9}+\frac{128.99}{6}}\le(\mu_1-\mu_2)\right.$$
$$\left.\le(58-51.83)+t_{\frac{0.05}{2},9+6-2}\sqrt{\frac{128.99}{9}+\frac{128.99}{6}}\right)=1-0.05$$

$$P(6.17-2.16\times5.9859\le(\mu_1-\mu_2)\le6.17-2.16\times5.9859)=0.95$$

....

....

#.2 Administrators at Rans Hospital (RH) wished to know whether the average stay at their hospital significantly differed from the average stay at nearby Gen Hospital (GH). A random sample of 25 records from the RH for the year 2006 revealed a mean hospital stay of 6.6 days, with a standard deviation of 2.5 days. A random sample of 25 records taken from the files of the GH revealed a mean hospital stay of 5.7 days, with a standard deviation of 2.0 days. By using the most appropriate statistical procedure to analyze the data obtained in this study, how do you help the administrators to formulate most appropriate conclusion?

[Hints: t_0 = 0.2583, not rejected H_0 that average stay at two hospitals are not significantly different.]

#.3 As part of an industrial program, some trainees are instructed by method A, which is straight computer-based instruction, and some are instructed method B, which also involves the personal attention of an instructor. If a random sample of size 10 are taken from large group of trainees instructed by each of these two methods, and the scores which they obtained in an appropriate achievement test are:

Method A	71	75	65	69	83	66	68	71	74	68
Method B	72	77	84	78	69	70	77	73	65	75

Test whether method B is better than method A.

[Hints: mean (A) = 70, mean (B) = 74, S_A^2 = 11.33, S_B^2 = 29.11, t = -1.9889, d.f. = 16, reject H_0]

Inference on Difference in Means between Two Dependent Samples (Paired t-test) (Most Imp.)

Following are the average weekly losses of worker-hours due to accidents in 10 industrial plants before and after certain safety training program was put into operation-

Plant No.	Before	After
1	45	36
2	73	60
3	46	44
4	124	119
5	33	35
6	57	51
7	83	77
8	34	29
9	26	24
10	17	11

Use the 0.01 level of significance to test whether the safety training program is effective in reducing losses of work hours. Use paired t-test.

Also, obtain 98% confidence interval for reduction in weekly work hour loss.

Solution-

Let μ_1 and μ_2 denote weekly work hour losses before and after safety training program.

Here, the null and alternative hypotheses are:

$$H_0: training\ program\ is\ not\ effective. (\mu_1 = \mu_2)$$

$$H_1: training\ program\ is\ effective. (\mu_1 > \mu_2)$$

Working Table:

Before	After	Difference (d)	d^2
45	36	9	81
73	60	13	169
46	44	2	4
124	119	5	25
33	35	-2	4
57	51	6	36
83	77	6	36
34	29	5	25
26	24	2	4
17	11	6	36
Total		52	420

Here,

$$\bar{d} = \frac{1}{n}\sum d_i = \frac{1}{10} \times 52 = 5.2$$

Next,

$$S_{\bar{d}}^2 = \frac{1}{n-1}\sum\left(d - \bar{d}\right)^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum d^2 - \bar{d}^2\right)$$

$$= \frac{10}{10-1}\left(\frac{1}{10} \times 420 - 5.2^2\right) = 16.62$$

Now the test statistic is

$$t_0 = \frac{\bar{d}}{\sqrt{\frac{S_{\bar{d}}^2}{n}}} = \frac{5.2}{\sqrt{\frac{16.62}{10}}} = 4.03$$

Here, the critical value is $t_{\alpha,n-1} = t_{0.01,10-1} = 2.82$.

Since $t_0 > 2.82$, so the null hypothesis is rejected and it is concluded that training program is effective in reducing the number work hour losses in plants.

Next confidence interval is given by

$$P\left(\bar{d} - t_{\frac{\alpha}{2},n-1}\sqrt{\frac{S_{\bar{d}}^2}{n}} \le \mu_1 - \mu_2 \le \bar{d} + t_{\frac{\alpha}{2},n-1}\sqrt{\frac{S_{\bar{d}}^2}{n}}\right) = 1 - \alpha$$

$$P\left(5.2 - t_{\frac{0.02}{2},10n-1}\sqrt{\frac{16.62}{10}} \le \mu_1 - \mu_2 \le 5.2 + t_{\frac{0.02}{2},10n-1}\sqrt{\frac{16.62}{10}}\right) = 1 - 0.02$$

$$P\big(5.2 - t_{0.01,9} \times 1.29 \le \mu_1 - \mu_2 \le 5.2 + t_{0.01,9} \times 1.29\big) = 0.98$$

$$P(5.2 - 2.82 \times 1.29 \le \mu_1 - \mu_2 \le 5.2 + 2.82 \times 1.29) = 0.98$$

……….

……..

#.2 The sales data of an item in six shops before and after a special promotional campaign are as follows:

Shops	A	B	C	D	E	F
Before campaign	53	28	31	48	50	42
After campaign	58	29	30	55	56	45

Can the campaign be judged to be a success? test at 5% level of significance.

[Hints: calc. t = - 2.78, H_0 rejected]

#.3 Memory capacity of 10 students was tested before and after training, state whether the training program was effective or not from the following scores

Roll No.	1	2	3	4	5	6	7	8	9	10
Before training	12	14	11	8	7	10	6	0	5	6
After training	15	16	10	7	5	12	10	2	3	8

[Hints: calc. t = -1.3, not rejected]