



# Statistics & Data Analysis Concepts for Data Science and ML **10**

---

## *10* *Inference Procedures for Two Populations*

*Estimation & Hypothesis Testing Involving Two  
Population Parameters*



2

# **Hypothesis Testing**

## *Inference Procedures for Two Populations*

*Estimation & Hypothesis Testing Involving  
Two Population Parameters*

# Learning Objectives...What you will learn

---

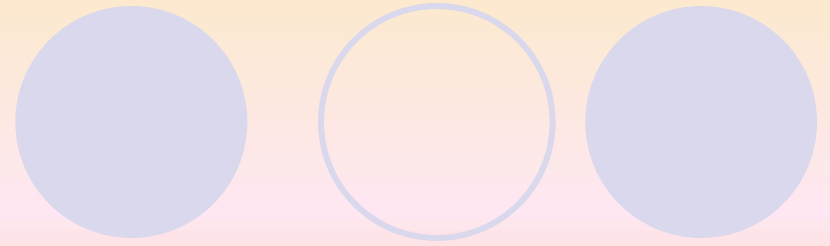
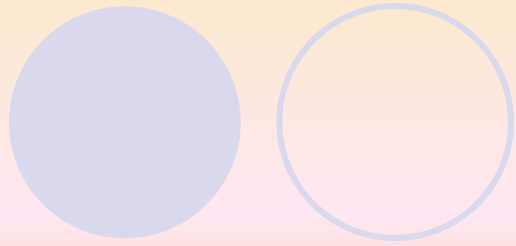
Develop the confidence interval estimates and hypothesis testing procedures involving two population parameters. After completing this chapter, you should be able :

- Develop the confidence interval estimates for the difference between two population means
- Formulate the null and alternate hypothesis involving two population parameters including means, proportions, and variances, and learn different methods of testing hypothesis involving two population parameters
- Develop both one-and two-tailed null and alternate hypotheses for testing the equality of two means or the difference between two means and apply correct distribution and test statistic to test these hypotheses

# Objectives...What you will learn

---

- Test two means when the populations are independent and the population variances are equal
- Test two means involving two independent populations but unequal variances
- Test the equality of two variances using the F-distribution
- Test two proportions involving large samples using the normal distribution
- Test appropriate hypothesis for related populations- the paired t-test



The point and interval estimates  
(Confidence Interval) for the difference  
between two population means and two  
proportions

# Point and Interval Estimates for the Difference between Two Population Means (Independent p Populations)

---

- Point Estimate for the Difference between Two Population Means

The point estimate of the difference between two population means  $(\mu_1 - \mu_2)$  is  $(\bar{x}_1 - \bar{x}_2)$

Interval Estimates for the Difference between Two Population Means depends on the sample size, whether the populations are independent/dependent, and whether population variances are known.

Specific cases are discussed in the following slides.

## Case 1

Assumptions: (1) the two populations are independent (2) the population variances are equal (3) the samples  $n_1$  and  $n_2$  are large and population variances  $\sigma_1^2$  and  $\sigma_2^2$  are **known** so that normal distribution applies.

The confidence interval for the difference between two population means is given by

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \quad (10.1)$$

or,

$$(\bar{x}_1 - \bar{x}_2) - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \quad (10.2)$$

Where,

$\bar{x}_1, \bar{x}_2$  = sample means from populations 1 and 2 respectively

$n_1, n_2$  = sample size of sample 1 and 2

$\sigma_1^2, \sigma_2^2$  = variances of 1<sup>st</sup> and 2<sup>nd</sup> populations respectively (known in this case)

## Case 2

Assumptions: (1) the two populations are independent (2) the population variances are equal (3) the samples  $n_1$  and  $n_2$  are large and population variances  $\sigma_1^2$  and  $\sigma_2^2$  are **unknown** so that normal distribution applies.

The confidence interval for the difference between two population means is given by

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad (10.3)$$

or,

$$(\bar{x}_1 - \bar{x}_2) - z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad (10.4)$$

where,  $s_1^2, s_2^2$  = sample variances of 1<sup>st</sup> and 2<sup>nd</sup> sample respectively

$\bar{x}_1, \bar{x}_2$  = sample means from populations 1 and 2 respectively

$n_1, n_2$  = sample size of sample 1 and 2

### Case 3

Assumptions: (1) the two populations are independent (2) the population variances are equal (3) the samples  $n_1$  and  $n_2$  are small ( $n < 30$ ) and population variances  $\sigma_1^2$  and  $\sigma_2^2$  are **unknown** so that the **t-distribution** applies. The confidence interval:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{n_1+n_2-2, \alpha/2} \sqrt{S_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)} \quad (10.5)$$

or,

$$(\bar{x}_1 - \bar{x}_2) - t_{n_1+n_2-2, \alpha/2} \sqrt{S_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)} \leq \mu_1 - \mu_2 \leq (\bar{x}_1 - \bar{x}_2) + t_{n_1+n_2-2, \alpha/2} \sqrt{S_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)} \quad (10.6)$$

Where,  $\bar{x}_1, \bar{x}_2$  = sample means from populations 1 and 2 respectively

$n_1, n_2$  = sample size of sample 1 and 2

$s_1^2, s_2^2$  = sample variances of 1<sup>st</sup> and 2<sup>nd</sup> sample respectively

$n_1 + n_2 - 2$  = degrees of freedom

$S_p^2$  is the "pooled" or combined variance, given by

$$S_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

## Example 1

A production manager wants to determine whether there is a significant difference in the average number of production units produced by the night shift and the day shift. Random samples from the day and night shifts are collected and the mean and standard deviation of the samples are calculated. The results are shown below

### Day Shift

$$n_1 = 50$$

$$\bar{x}_1 = 13.6$$

$$S_1 = 2.2$$

### Night Shift

$$n_2 = 35$$

$$\bar{x}_2 = 11.6$$

$$S_2 = 3.0$$

In this problem, we assume that the population variances are equal, (that is,  $\sigma_1^2 = \sigma_2^2$ ).

(a) What is the point estimate of the difference between the two population means?

The point estimate for the difference of population means is given by

$$\bar{x}_1 - \bar{x}_2 = 13.6 - 11.6 = 2$$

### ***Example 1...cont.***

(b) Find a 90% confidence interval for the difference in the mean number of units for the two shifts.

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

$$(13.6 - 11.6) \pm (1.645) \sqrt{\frac{(2.2)^2}{50} + \frac{(3.0)^2}{35}}$$

$$(2.0) \pm (1.645)(0.595)$$

$$2.0 \pm 0.98$$

$$1.02 \text{ to } 2.98$$

$$\text{or, } 1.02 \leq \mu_1 - \mu_2 \leq 2.98$$

- The above is a 90% confidence interval for the difference between the population means. Note that we are estimating the difference between the population means ( $\mu_1 - \mu_2$ ) using the sample means ( $x_1 - x_2$ )
- The interval does not include a zero indicating that there is a difference between the items produced during the day and the night shift.

### Example 1...cont.

(c) Find a 95% confidence interval for the above case.

Using the confidence interval formula in part (b), formula, a 95% confidence interval can be calculated as:

$$\begin{aligned} & (13.6 - 11.6) \pm (1.96) (0.595) \\ & (2.0) \pm 1.17 \\ & 0.83 \text{ to } 3.17 \end{aligned}$$

**From the normal (Z) table - see the next slide**

$$\text{or, } 0.83 \leq \mu_1 - \mu_2 \leq 3.17$$

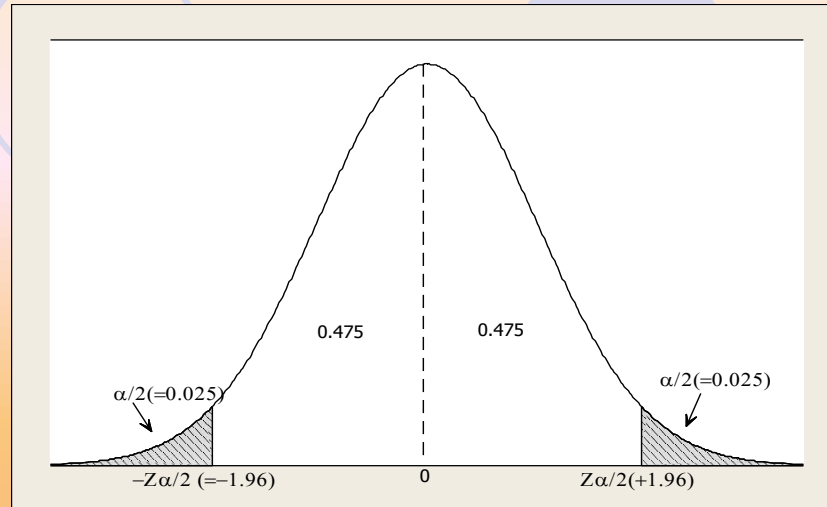
Note that the interval gets wider, as we increase the confidence interval.

(d) What is the margin of error at 95% confidence?

The margin of error is given by

$$\begin{aligned} e &= z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \\ e &= (1.96) \sqrt{\frac{(2.2)^2}{50} + \frac{(3.0)^2}{35}} \\ e &= 1.17 \end{aligned}$$

*Z-value for a 95%  
C.I. using the Normal  
Distribution* ➔



The table shows the Z values for commonly used confidence levels. You should verify these values using the normal table on the next page.

Confidence Level	$\alpha$	$\alpha/2$	Corresponding Z values ( $z_{\alpha/2}$ )
90%	0.10	0.05	$z_{\alpha/2} = 1.645$
95%	0.05	0.025	$z_{\alpha/2} = 1.96$
98%	0.02	0.01	$z_{\alpha/2} = 2.33$
99%	0.01	0.005	$z_{\alpha/2} = 2.58$



## Standard Normal Distribution Table

Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.00	0.004	0.008	0.012	0.016	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.091	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.148	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.17	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.195	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.219	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.258	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.291	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.334	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.377	0.379	0.381	0.383
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.398	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.437	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.475	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.483	0.4834	0.4838	0.4842	0.4846	0.485	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.489
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.492	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.494	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.496	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.497	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.498	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.499	0.499



## Example 2

The average hourly wage of union and nonunion workers are of interest. Assume that (i) the populations are independent, and (ii) the variances of the populations are equal ( $\sigma_1^2 = \sigma_2^2$ ). Sample from both groups were taken with the following results:

### Union Workers

$$\begin{aligned}n_1 &= 15 \\ \bar{x}_1 &= 15.54 \\ s_1 &= 2.24\end{aligned}$$

### Non-union Workers

$$\begin{aligned}n_2 &= 20 \\ \bar{x}_2 &= 15.36 \\ s_2 &= 1.99\end{aligned}$$

(a) Find the “pooled” estimate of the variance.

$$\begin{aligned}S_p^2 &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \\ &= \frac{(14)(2.24)^2 + (19)(1.99)^2}{33} \\ &= 4.41\end{aligned}$$

(b) Find a 95% confidence interval for the difference in the mean hourly wage.

$$\left(\bar{x}_1 - \bar{x}_2\right) \pm t_{n_1+n_2-2, \alpha/2} \sqrt{Sp^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

↓

$$t_{33, 0.025} = 2.035$$
$$(14.54 - 15.36) \pm (2.035)(0.72)$$
$$-0.82 \pm 1.47$$
$$-2.29 \text{ to } 0.65$$

or,  $-2.29 \leq \mu_1 - \mu_2 \leq 0.65$

**From the t-table. See the next slide**

Note that this interval contains zero (-2.29 to 0.65). This means that the difference is zero at some point in the interval indicating there is no difference in the average wage of union and nonunion workers.

*Area in Upper Tail*

Degrees of Freedom(Df)	0.100	0.050	0.025	0.010	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
:					
33			2.035		

# Point and Interval Estimates for the Difference Between Two Population Means for Paired or Matched Samples (dependent or related populations)

---

**In some situations, there may be a relationship between the data values between the samples of two populations. Because of this relationship, the data values or the sample values from one population will not be independent of the sample values from the other population.**

Suppose that  $X_{11}, X_{12}, X_{13}, \dots, X_{1n}$  and  $X_{21}, X_{22}, X_{23}, \dots, X_{2n}$  are samples of the same size from two different normal populations with means  $\mu_1$  and  $\mu_2$

The pairs of data values  $X_{1n}$  and  $X_{2n}$  are related, and are not independent. In this case, we say that the values in one sample are **paired** or **matched** with the values in other sample. In other words, the samples are dependent or related. The data structure is shown in the next slide.



Observation	Sample 1	Sample 2	Difference, $d_i$
1	$X_{11}$	$X_{21}$	$d_1 = X_{11} - X_{21}$
2	$X_{12}$	$X_{22}$	$d_2 = X_{12} - X_{22}$
3	$X_{13}$	$X_{23}$	$d_3 = X_{13} - X_{23}$
4	$X_{14}$	$X_{24}$	$d_4 = X_{14} - X_{24}$
:	:	:	
:	:	:	
n	$X_{1n}$	$X_{2n}$	$d_n = X_{1n} - X_{2n}$

If the pairs of data values  $X_{1n}$  and  $X_{2n}$  are related, then they are not independent. The average of the differences ( $\bar{d}$ ) follows a t-distribution and the point and interval estimates for the population mean paired difference,  $\mu_d$  is given by

Point estimate of  $\mu_d$

$$\bar{d} = \frac{\sum_{i=1}^n d_i}{n} \quad (10.7)$$



Confidence Interval for  $\mu_d$

$$\bar{d} \pm t_{n-1, \frac{\alpha}{2}} \frac{s_d}{\sqrt{n}} \quad (10.8)$$

where,  $\bar{d}$  = average of the differences

$s_d$  = standard deviation of the differences;  $s_d = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n-1}} = \sqrt{\frac{\sum d_i^2 - \frac{(\sum d_i)^2}{n}}{n-1}}$

$n$  = number of observations (sample size)

$t_{n-1, \frac{\alpha}{2}}$  = critical t-value from the t-table for  $(n-1)$  degrees of freedom and the appropriate  $\alpha$



### Example 3

An oil company is proposing a new type of additive for the gasoline that is expected to increase the average mileage. Ten cars of the same make and model are selected and tested with and without the additive. The mileage (miles per gallon or mpg) with and without the additive were recorded for these cars. The data are shown below. Based on these data, can we say that there is a difference in mileage due to the additive? Construct a 95% confidence interval to answer the question.

Car	Mpg (without additive)	Mpg (with additive)	Difference (di)
1	25.0	26.2	-1.2
2	23.5	25.0	-1.5
3	24.0	27.5	-3.5
4	23.5	30.5	-7.0
5	26.5	27.0	-0.5
6	25.0	25.0	0.0
7	24.8	27.8	-3.0
8	26.0	28.3	-2.3
9	25.9	29.0	-3.1
10	29.0	25.0	4.0



The mean and standard deviation of the difference are;

$$\bar{d} = -1.81$$

$$s_d = 2.84$$

A 95% confidence interval can be calculated using

$$\bar{d} \pm t_{n-1, \frac{\alpha}{2}} \frac{s_d}{\sqrt{n}}$$

$$-1.81 \pm t_{9, 0.025} \frac{2.84}{\sqrt{10}}$$

$$-1.81 \pm (2.262)(0.9)$$

$$-1.81 \pm 2.04$$

$$-3.85 \text{ to } 0.23$$

**The interval of -3.85 to 0.23 includes zero. This shows that there is no difference in mileage when the gasoline is used with or without the additive.**



# Point and Interval Estimates for the Difference Between Two Population Proportions

---

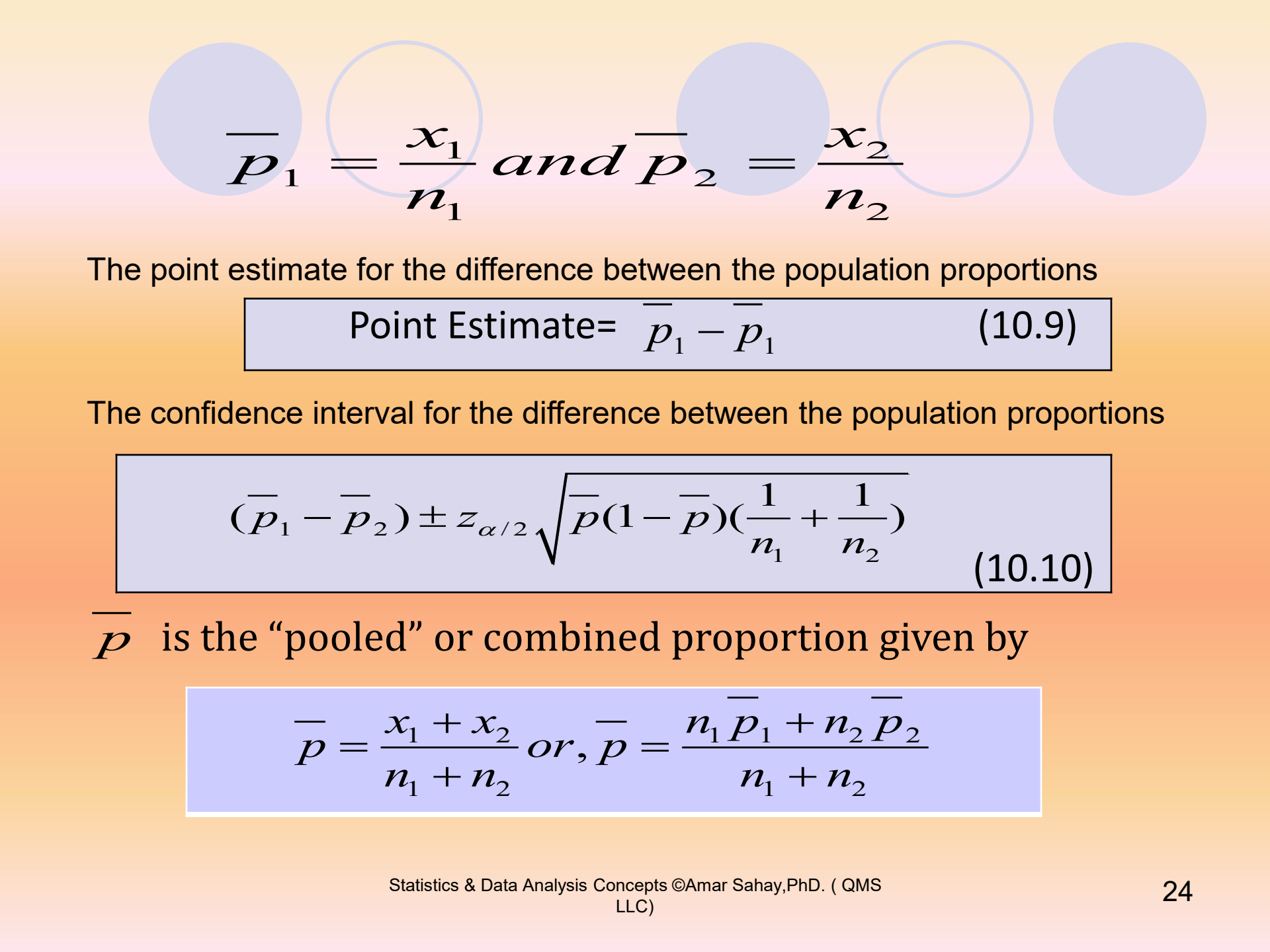
Consider two large populations. Let  $p_1$  and  $p_2$  denote the proportions of these two populations that have certain characteristic of interest.

Suppose we want to find the difference between the two population proportions. To do this, we will take two random samples of size  $n_1$  and  $n_2$  from the two populations.

$x_1$  is the number having the characteristic of interest (for example, number of defects) in the first population,

$x_2$  the number having the characteristic of interest in the second population, then

$\bar{p}_1$  and  $\bar{p}_2$  are the sample proportions of the two samples that have the characteristic of interest then



$$\bar{p}_1 = \frac{x_1}{n_1} \text{ and } \bar{p}_2 = \frac{x_2}{n_2}$$

The point estimate for the difference between the population proportions

$$\text{Point Estimate} = \bar{p}_1 - \bar{p}_2 \quad (10.9)$$

The confidence interval for the difference between the population proportions

$$(\bar{p}_1 - \bar{p}_2) \pm z_{\alpha/2} \sqrt{\bar{p}(1 - \bar{p}) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)} \quad (10.10)$$

$\bar{p}$  is the “pooled” or combined proportion given by

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2} \text{ or, } \bar{p} = \frac{n_1 \bar{p}_1 + n_2 \bar{p}_2}{n_1 + n_2}$$

### Example 4

A manufacturer of computer chips has two plants in two different cities. An improved method of producing the chips is suggested by the research and development department which is expected to reduce the proportion of defective chips. After implementing the new method in one of the plants, the manufacturer wanted to test if the improved method actually reduced the proportion of defects. To do this, they collected a sample of 400 chips from the improved method of production, and a sample of 450 chips from the old production method. The numbers of defective chips were found to be 80 and 108 respectively. At a 95% confidence, can we conclude that there is a difference in proportion of defective chips?

Proportion of defective using the improved method:  $\bar{p}_1 = \frac{x_1}{n_1} = \frac{80}{400} = 0.20$

Proportion of defective using the old method:  $\bar{p}_2 = \frac{x_2}{n_2} = \frac{108}{450} = 0.24$

Combined or the “pooled” proportion:  $\bar{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{80 + 108}{400 + 450} = 0.221$



The confidence interval:

Thus, a 95% confidence interval is

$$(\bar{p}_1 - \bar{p}_2) \pm z_{\alpha/2} \sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$(0.20 - 0.24) \pm (1.96) \sqrt{(0.221)(1-0.221)\left(\frac{1}{400} + \frac{1}{450}\right)}$$

$$-0.04 \pm 0.06$$

$$-0.1 \leq p_1 - p_2 \leq 0.02$$

**This interval contains zero which indicates that there is no difference between the improved method and the old method of production.**



# Hypothesis Tests involving two means, two variances, and two proportions



- This section extends the concepts of hypothesis testing to two populations.
- This part involves testing two population means. For example, testing the average salaries of male and female employees for the same job, or testing the difference between the average starting salaries of business and engineering majors.

# Hypothesis testing for the equality of two means or the differences between two population means

---

The hypothesis for testing the two means can be a two-sided test or a one-sided test. The tests in this sections are based on the following two assumptions:

- (1) The populations are independent
- (2) The population variances are equal

The hypothesis test involving two means can be written in one of the following three ways:

- (a) Test if the two population means are equal or there is no difference between the two means: ***a two-sided test***

$$\begin{array}{ll}
 H_0 : \mu_1 = \mu_2 \text{ or,} & H_0 : \mu_1 - \mu_2 = 0 \\
 H_0 : \mu_1 \neq \mu_2 \text{ or,} & H_0 : \mu_1 - \mu_2 \neq 0 \quad (10.11)
 \end{array}$$

(b) Test if one population mean is greater than the other: ***a right-sided test***

$$\begin{array}{ll}
 H_0 : \mu_1 \leq \mu_2 \text{ or,} & H_0 : \mu_1 - \mu_2 \leq 0 \\
 H_0 : \mu_1 > \mu_2 \text{ or,} & H_0 : \mu_1 - \mu_2 > 0 \quad (10.12)
 \end{array}$$

(c) Test if one population mean is smaller than the other: ***a left-sided test***

$$\begin{array}{ll}
 H_0 : \mu_1 \geq \mu_2 \text{ or,} & H_0 : \mu_1 - \mu_2 \geq 0 \\
 H_0 : \mu_1 < \mu_2 \text{ or,} & H_0 : \mu_1 - \mu_2 < 0 \quad (10.13)
 \end{array}$$

To test two means, the test statistics are selected based on the following cases:

**Case 1:** Sample sizes  $n_1$  and  $n_2$  are large ( $\geq 30$ ) and the population variances  $\sigma_1^2$  and  $\sigma_2^2$  are known

If the sample sizes  $n_1$  and  $n_2$  are large ( $\geq 30$ ) and the population variances  $\sigma_1^2$  and  $\sigma_2^2$  are known, then the sampling distribution of the difference between the sample means follows a normal distribution and the test statistic is given by

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad (10.15)$$

$\bar{x}_1, \bar{x}_2$  = sample means from populations 1 and 2 respectively

$n_1, n_2$  = sample size of sample 1 and 2

$\sigma_1^2, \sigma_2^2$  = variances of 1<sup>st</sup> and 2<sup>nd</sup> populations respectively (known in this case)



The confidence interval for this case can be written as:

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \quad (10.16)$$

or,

$$(\bar{x}_1 - \bar{x}_2) - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

To test two means, the test statistics are selected based on the following cases:

**Case 2:** Sample sizes  $n_1$  and  $n_2$  are large ( $\geq 30$ ) and the population variances  $\sigma_1^2$  and  $\sigma_2^2$  are **unknown**

If the sample sizes  $n_1$  and  $n_2$  are large ( $\geq 30$ ) and the population variances  $\sigma_1^2$  and  $\sigma_2^2$  are **unknown**, then the sampling distribution of the difference between the sample means follows a normal distribution and the test statistic is given by

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad (10.17)$$

$\bar{x}_1, \bar{x}_2$  = sample means from populations 1 and 2 respectively

$n_1, n_2$  = sample size of sample 1 and 2

$s_1^2, s_2^2$  = variances of 1<sup>st</sup> and 2<sup>nd</sup> populations respectively

The confidence interval for this case can be written as:

or,

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad (10.18)$$

$$(\bar{x}_1 - \bar{x}_2) - z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad (10.19)$$

To test two means, the test statistics are selected based on the following cases:

**Case 3:** Sample sizes  $n_1$  and  $n_2$  are small ( $n < 30$ ) and the population variances  $\sigma_1^2$  and  $\sigma_2^2$  are **unknown**

If the sample sizes  $n_1$  and  $n_2$  are small ( $< 30$ ) and the population variances  $\sigma_1^2$  and  $\sigma_2^2$  are **unknown**, then the sampling distribution of the difference between the sample means follows a **t-distribution** and the test statistic is given by

$$t_{n_1+n_2-2} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \quad (10.20)$$

$\bar{x}_1, \bar{x}_2$  = sample means from populations 1 and 2 respectively

$n_1, n_2$  = sample size of sample 1 and 2

$s_1^2, s_2^2$  = variances of 1<sup>st</sup> and 2<sup>nd</sup> populations respectively

$S_p^2$  is the 'pooled' or combined variance, given by

$$S_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \quad (10.21)$$

**Confidence interval for the difference between the two population means  $(\mu_1 - \mu_2)$**

$$(\bar{x}_1 - \bar{x}_2) \pm t_{n_1+n_2-2, \alpha/2} \sqrt{S_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)} \quad (10.22)$$

In expanded form:

$$(\bar{x}_1 - \bar{x}_2) - t_{n_1+n_2-2, \alpha/2} \sqrt{S_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)} \leq \mu_1 - \mu_2 \leq (\bar{x}_1 - \bar{x}_2) + t_{n_1+n_2-2, \alpha/2} \sqrt{S_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)} \quad (10.23)$$

## Example 5

The mean assembly time of two products produced in two different plants is of interest. The operations manager believes that plant 1 has higher time on the average compared to plant 2. The product samples from the two plants for the past two weeks are used to test the hypothesis to determine if the mean assembly time for plant 1 is higher than that of plant 2. Assume that the variances of the two populations are equal. Note that we want to test the following hypothesis:

$$H_0: \mu_{\text{plant 1}} \leq \mu_{\text{plant 2}}$$

$$H_1: \mu_{\text{plant 1}} > \mu_{\text{plant 2}}$$

The data from the two plants are shown below

The data from the two plants are shown below

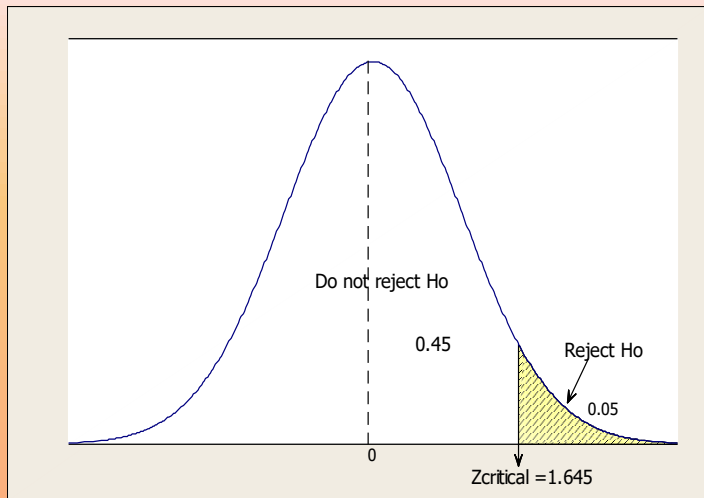
<u>Plant 1</u>	<u>Plant 2</u>
– $n_1 = 50$ $x_1 = 6.72$ min.	– $n_2 = 50$ $x_1 = 6.34$ min.
$s_1 = 2.374$	$s_2 = 2.163$

$$\alpha = 0.05$$

Test the hypothesis using four different methods and state your conclusion clearly.

## Method (1) z-value approach

Critical point: The critical point test for this shown below



**Decision Rule:** Reject  $H_0$  if  
 $z = 1.645$



Test Statistic Value:

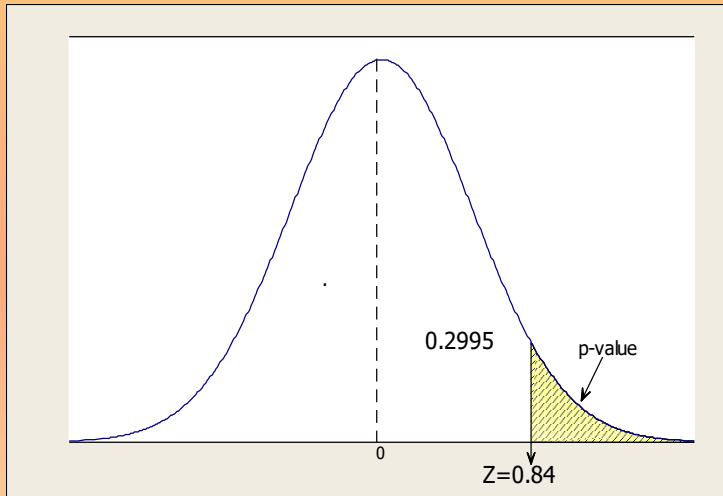
$$\begin{aligned} z &= \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \\ &= \frac{(6.72 - 6.34) - 0}{\sqrt{\frac{(2.374)^2}{50} + \frac{(2.163)^2}{50}}} \\ &= 0.84 \end{aligned}$$

Since,  $z=0.84 < z = 1.645$ ; do not reject  $H_0$

## Method (2) p-value approach

Decision Rule: if  $p \geq \alpha$ ; Do not reject  $H_0$   
if  $p < \alpha$ ; reject  $H_0$

To calculate p, first find z (test statistic value) using the test statistics formula in method (1) – note that this  $z=0.84$ . Convert this z value to a p-value (see the figure below).



From the z-table, the area corresponding to  $z = 0.84$  is 0.2995 (see the figure ).

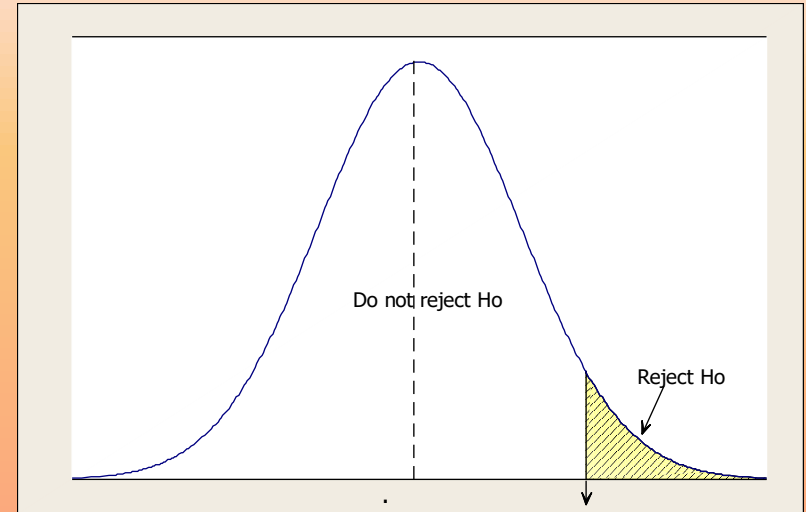
Probability of  $z > .84 = 0.5 - 0.2995 = 0.2005$ .  
(this is the p-value).

**Since,  $p = 0.2005 > \alpha = 0.05$ ; do not reject  $H_0$ .**

### Method (3) critical-value approach

Find the critical value of  $(\bar{x}_1 - \bar{x}_2)_c$  and compare to  $(\bar{x}_1 - \bar{x}_2)$

$$\begin{aligned}(\bar{x}_1 - \bar{x}_2)_c &= +z \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \\ &= +(1.645) \sqrt{\frac{(2.374)^2}{50} + \frac{(2.163)^2}{50}} \\ &= +(1.645)(0.454) \\ &= +0.745\end{aligned}$$



$$(\bar{x}_1 - \bar{x}_2)_c = 0.745$$

Decision rule: Reject  $H_0$  if  $(\bar{x}_1 - \bar{x}_2) > 0.75$

Since,  $(\bar{x}_1 - \bar{x}_2) = 6.72 - 6.34 = 0.38 < (\bar{x}_1 - \bar{x}_2)_c = 0.745$  ; do not reject  $H_0$ .

## ***Method (4): Confidence Interval approach***

Note: Since this is a one-sided test, double the value of  $\alpha$  and find a 90% confidence interval.

Using the Confidence Interval Formula for this case:

$$\begin{aligned} & (\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \\ & (6.72 - 6.34) \pm (1.645) \sqrt{\frac{(2.374)^2}{50} + \frac{(2.163)^2}{50}} \\ & 0.38 \pm (1.645)(0.454) \\ & 0.38 \pm 0.75 \\ & -0.37 \text{ to } 1.130 \end{aligned}$$

$$\text{or, } -0.37 \leq \mu_1 - \mu_2 \leq 1.130$$

***The above interval includes “zero” therefore, do not reject  $H_0$ .***

### Example 5

Two production lines are compared to determine if there is difference in the assembly times for the same job on two lines. The following hypothesis is of interest (assuming that the variances of the two lines are equal).

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{or,} \quad H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 - \mu_2 \neq 0 \quad H_1 : \mu_1 \neq \mu_2$$

The data from the two plants are shown below

The data from the two plants are shown below

$$\begin{array}{l} \text{Plant 1} \\ \bar{x}_1 = 1.4 \text{ min.} \\ s_1 = 0.4 \end{array}$$

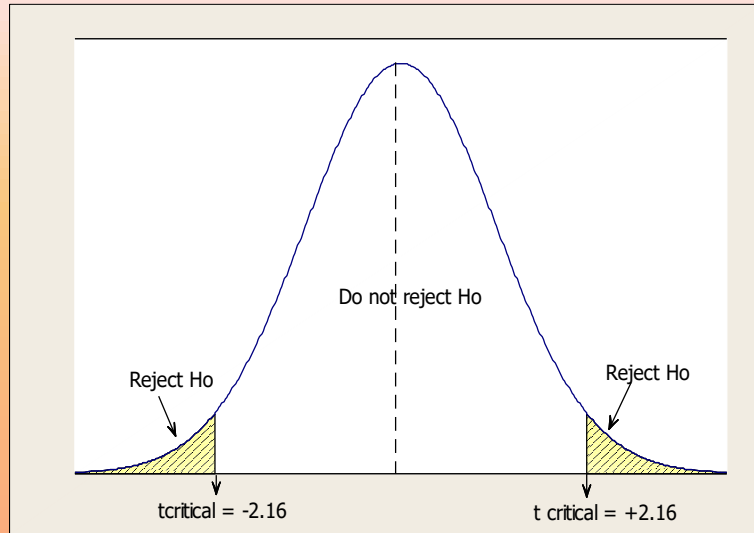
$$\begin{array}{l} \text{Plant 2} \\ \bar{x}_2 = 1.0 \text{ min.} \\ s_2 = 0.6 \end{array}$$

$$\alpha = 0.05$$

Test the hypothesis using four different methods and state your conclusion clearly.

## Method (1) t-value approach

Critical point: The critical points for this test are shown below:



$$t_{n_1+n_2-2, \alpha/2} = t_{13, 0.025} = 2.16$$

**Decision Rule:** Reject  $H_0$  if

$$t_{n_1+n_2-2} > 2.16$$

$$\text{or, } t_{n_1+n_2-2} < -2.16$$

Before calculating the test statistic value, calculate the value of  $S_p^2$

$$S_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$S_p^2 = \frac{(7)(0.4)^2 + (6)(0.6)^2}{8 + 7 - 2} = 0.2523$$

Test Statistic Value:

$$\begin{aligned} t_{n_1+n_2-2} &= \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \\ &= \frac{(1.4 - 1.0) - 0}{\sqrt{0.2523 \left( \frac{1}{8} + \frac{1}{7} \right)}} \\ &= 1.54 \end{aligned}$$

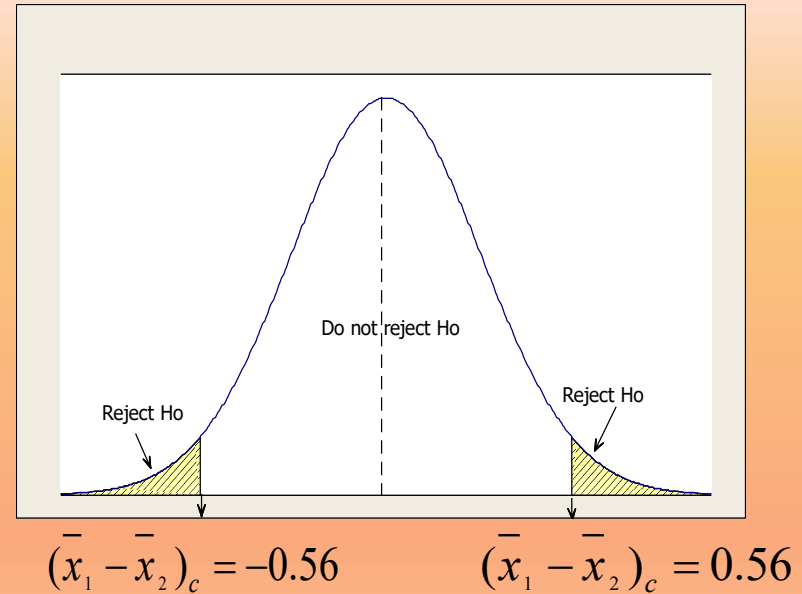
Since,  $t_{n_1+n_2-2} = 1.54 < t_{critical} = 2.16$

Do not reject  $H_0$

## Method (2) critical-value approach

Find the critical value of  $(\bar{x}_1 - \bar{x}_2)_c$  and compare to  $(\bar{x}_1 - \bar{x}_2)$

$$\begin{aligned}(\bar{x}_1 - \bar{x}_2)_c &= \pm t_{n_1+n_2-2, \alpha/2} \sqrt{Sp^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)} \\ &= \pm t_{13, 0.025} \sqrt{Sp^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)} \\ &= \pm (2.16) \sqrt{0.2523 \left( \frac{1}{8} + \frac{1}{7} \right)} \\ &= \pm (2.16)(0.26) \\ &= \pm 0.56\end{aligned}$$



Decision rule: Reject  $H_0$  if  $(\bar{x}_1 - \bar{x}_2) > 0.56$   
or, if  $(\bar{x}_1 - \bar{x}_2) < -0.56$

Since,  $(\bar{x}_1 - \bar{x}_2) = (1.4 - 1.0) = 0.4 < (\bar{x}_1 - \bar{x}_2)_c = 0.56$  ; do not reject  $H_0$ .

### **Method (3): Confidence Interval approach**

**Decision Rule:** Reject  $H_0$  if “zero” is not included in the confidence interval.  
For this case, the confidence interval can be calculated as

Using the Confidence Interval Formula for this case:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{n_1+n_2-2, \alpha/2} \sqrt{Sp^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}$$

↓

$$t_{13, 0.025} = 2.16$$

$$(1.4 - 1.0) \pm (2.16) \sqrt{0.2523 \left( \frac{1}{8} + \frac{1}{7} \right)}$$

$$0.4 \pm (2.16)(0.26)$$

$$0.4 \pm 0.56$$

$$-0.16 \text{ to } 0.96$$

or,  $-0.16 \leq \mu_1 - \mu_2 \leq 0.96$

**The above interval includes “zero” therefore, do not reject  $H_0$ .**

# Testing Two Means When the Variances are Not Equal (Ignore pages 46 to 78)

---

- The example here demonstrates *how to test two means when variances are not equal* or,  $\sigma_1^2 \neq \sigma_2^2$
- In practice, whenever we cannot safely assume that the population variances are equal, we should first test the variances rather than assuming that the variances are equal and then test the means. Thus, testing the equality of two means involves two steps:
  - (1) **Test to determine if the two population variances are equal.**
  - (2) **Based on whether the variances are equal, test the equality of means.**

# Testing the Equality of Two Variances

The test statistics for testing the equality of two variances is written as:

$$\begin{aligned} H_0 : \sigma_1^2 &= \sigma_2^2 \\ H_1 : \sigma_1^2 &\neq \sigma_2^2 \end{aligned} \quad (10.24)$$

To test if the variances of two populations are equal, we test the ratio of the two sample variances  $s_1^2 / s_2^2$  (not the difference), and use the F-distribution. The test statistic is given by

$$F_{(n_1-1),(n_2-1)} = \frac{s_1^2}{s_2^2} \quad (10.25)$$



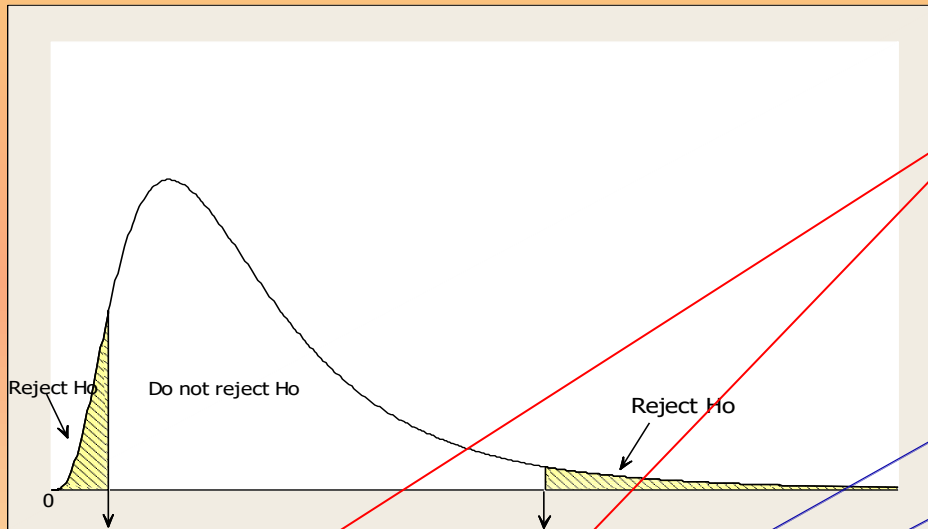
$n_1, n_2$  = sample size of sample 1 and 2  
 $s_1^2, s_2^2$  sample variances of 1<sup>st</sup> and 2<sup>nd</sup> sample respectively

$F_{(n_1-1),(n_2-1)}$  F- value from the F-table for  $(n_1 - 1), (n_2 - 1)$  degrees of freedom and appropriate level of significance  $\alpha$

# Critical Values of F-distribution

(Note: the ratio of the two sample variances  $s_1^2 / s_2^2$  follows an F-distribution.

Figure below shows the critical values for the F-distribution for testing two variances. The critical values are read from the F-table. **The F-distribution has two degrees of freedom (df):  $(n_1 - 1), (n_2 - 1)$  -- numerator and the denominator degrees of freedom (note: the first value is the numerator df and the second value is the denominator df).**



Critical Values using F-distribution for Testing Two Variances

**Numerator degrees of freedom (df1)**

**Denominator degrees of freedom (df2)**

$$F_{(n_2-1), (n_1-1), \alpha/2}$$

$$F_{(n_1-1), (n_2-1), \alpha/2}$$

# Testing Two Means or the Equality of Two Means When the Variances Are Not Equal

When the population variances are not equal; that is,  $\sigma_1^2 \neq \sigma_2^2$  the test of two means has a different test statistic than when the population variances are equal.

The test statistic for testing two means when the **population variances are not equal** is given by

$$t' = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad (10.26)$$

*Critical point:* For a two-sided test  $t_{df, \alpha/2}$  For a one-sided test  $t_{df, \alpha}$

*df* is the degrees of freedom,

The degrees of freedom  $df$ , determined by the following formula:

$$df = \frac{(w_1 + w_2)^2}{\frac{w_1^2}{n_1 - 1} + \frac{w_2^2}{n_2 - 1}} \text{ where, } w_1 = \frac{s_1^2}{n_1} \text{ and } w_2 = \frac{s_2^2}{n_2} \quad (10.27)$$

***Confidence Interval for this case:***

$$(\bar{x}_1 - \bar{x}_2) \pm t_{df, \alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad (10.28)$$

## Example 6

### ***Testing Two Means when the Variances are not Equal***

A job is processed in two different machines: Machine 1 and Machine 2. The average processing times of the jobs and the standard deviations are given below. The production manager would like to determine if there is a difference in the average processing time of two machines. Test the appropriate hypothesis. ***It is not known whether the population variances are equal.*** Use a 5% level of significance ( $\alpha = 0.05$ ).

Sample 1

$$n_1 = 16$$

$$\bar{x}_1 = 3.33hrs.$$

$$s_1 = 0.68$$

Sample 2

$$n_2 = 21$$

$$\bar{x}_2 = 3.98hrs.$$

$$s_2 = 0.38$$

$$\alpha = 0.05$$

**Step 1:** Since it is not known if the population variances are equal; test the hypothesis to determine if the population variances are equal

---

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

**Test Statistic:** (use F-distribution to test the two variances)

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

**Critical Points: Upper critical point:**

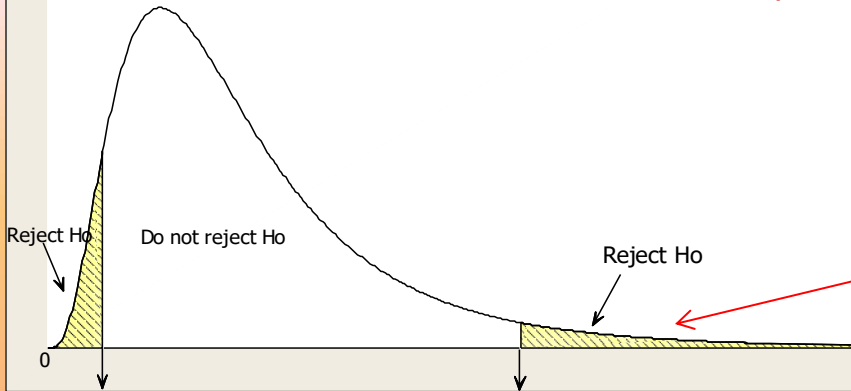
$$F_{(n_1 - 1), (n_2 - 1), \alpha/2} = F_{15, 20, 0.025} = 2.57$$

**This value is obtained from the F-table. Since  $\alpha = 0.05$  and it is a two-sided test; we look into the F-table of 0.025 (see the next two slides)**

**Lower critical point**

$$\frac{1}{F_{(n_2 - 1), (n_1 - 1), \alpha/2}} = \frac{1}{F_{20, 15, 0.025}} = \frac{1}{2.76} = 0.362$$

(See the next slide for F-table values)



0.362

2.57

$$\frac{1}{F_{(n_2-1), (n_1-1), \alpha/2}} = \frac{1}{2.76} = 0.362$$

$$F_{(n_1-1), (n_2-1), \alpha/2} = F_{15, 20, 0.025} = 2.57$$

**Decision Rule:** Reject  $H_0$  if  $F_{(n_1-1), (n_2-1)} > 2.57$   
or,  $F_{(n_1-1), (n_2-1)} < 0.362$

Test Statistic Value:

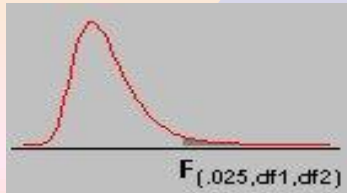
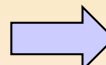
$$F_{(n_1-1), (n_2-1)} = \frac{S_1^2}{S_2^2} = \frac{(0.68)^2}{(0.38)^2} = 3.20$$

$$F_{(n_1-1), (n_2-1)} = 3.20 > F_{critical} = 2.57; \text{reject } H_0$$

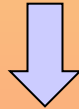
Thus, we conclude that the variances are not equal or,

$$\sigma_1^2 \neq \sigma_2^2.$$

# Numerator degrees of freedom



Denominator  
degrees of  
freedom



df2/df1	1	2	3	6	7	.....	12	15	20	24
1	647.789	799.500	864.163	937.111	948.217		976.708	984.867	993.103	997.249
2	38.506	39.000	39.166	39.332	39.355		39.415	39.431	39.448	39.456
3	17.443	16.044	15.439	14.735	14.624		14.337	14.253	14.167	14.124
4	12.218	10.649	9.979	9.197	9.074		8.751	8.657	8.560	8.511
5	10.007	8.434	7.764	6.978	6.853		6.525	6.428	6.329	6.278
6	8.813	7.260	6.599	5.820	5.696		5.366	5.269	5.168	5.117
7	8.073	6.542	5.890	5.119	4.995		4.666	4.568	4.467	4.415
8	7.571	6.060	5.416	4.652	4.529		4.200	4.101	4.000	3.947
9	7.209	5.715	5.078	4.320	4.197		3.868	3.769	3.667	3.614
10	6.937	5.456	4.826	4.072	3.950		3.621	3.522	3.419	3.365
11	6.724	5.256	4.630	3.881	3.759		3.430	3.330	3.226	3.173
12	6.554	5.096	4.474	3.728	3.607		3.277	3.177	3.073	3.019
13	6.414	4.965	4.347	3.604	3.483		3.153	3.053	2.948	2.893
14	6.298	4.857	4.242	3.501	3.380		3.050	2.949	2.844	2.789
15	6.200	4.765	4.153	3.415	3.293		2.963	2.862	2.756	2.701
16	6.1151	4.6867	4.0768	3.3406	3.2194		2.889	2.7875	2.6808	2.6252
17	6.042	4.6189	4.0112	3.2767	3.1556		2.8249	2.723	2.6158	2.5598
18	5.9781	4.5597	3.9539	3.2209	3.0999		2.7689	2.6667	2.559	2.5027
19	5.9216	4.5075	3.9034	3.1718	3.0509		2.7196	2.6171	2.5089	2.4523
20	<del>5.8745</del>	<del>4.4613</del>	<del>3.8587</del>	<del>3.1283</del>	<del>3.0074</del>		<del>2.6758</del>	<del>2.5731</del>	2.4645	2.4076
21	5.8266	4.4199	3.8188	3.0895	2.9686		2.6368	2.5338	2.4247	2.3675
22	5.7863	4.3828	3.7829	3.0546	2.9338		2.6017	2.4984	2.389	2.3315
23	5.7498	4.3492	3.7505	3.0232	2.9023		2.5699	2.4665	2.3567	2.2989
24	5.7166	4.3187	3.7211	2.9946	2.8738		2.5411	2.4374	2.3273	2.2693
25	5.6864	4.2909	3.6943	2.9685	2.8478		2.5149	2.411	2.3005	2.2422
26	5.6586	4.2655	3.6697	2.9447	2.824		2.4908	2.3867	2.2759	2.2174
:										
:										
120	5.1523	3.8046	3.2269	2.5154	2.3948		2.0548	1.945	1.8249	1.7597
inf	5.0239	3.6889	3.1161	2.4082	2.2875		1.9447	1.8326	1.7085	1.6402

**Step 2: Test for the equality of means, or test if the average processing times of the two machines do not differ.**

---

The hypothesis to be tested:  $H_0 : \mu_1 - \mu_2 = 0$  or,  $H_0 : \mu_1 = \mu_2$   
 $H_1 : \mu_1 - \mu_2 \neq 0$  or,  $H_1 : \mu_1 \neq \mu_2$

Test Statistic:

$$t' = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Critical Points:  $t_{df}, \frac{\alpha}{2}$

Where, df = degrees of freedom, given by

$$df = \frac{(w_1 + w_2)^2}{\frac{w_1^2}{n_1 - 1} + \frac{w_2^2}{n_2 - 1}} \quad \text{where,} \quad w_1 = \frac{s_1^2}{n_1}; w_2 = \frac{s_2^2}{n_2}$$



The value of  $df$  (degrees of freedom) can be calculated as shown below.

$$w_1 = \frac{s_1^2}{n_1} = \frac{(0.68)^2}{16} = 0.0289$$

$$w_2 = \frac{s_2^2}{n_2} = \frac{(0.38)^2}{21} = 0.006876$$

$$\frac{w_1^2}{n_1 - 1} = \frac{(0.0289)^2}{15} = 0.0000556$$

$$\frac{w_2^2}{n_2 - 1} = \frac{(0.006876)^2}{20} = 0.0000023$$

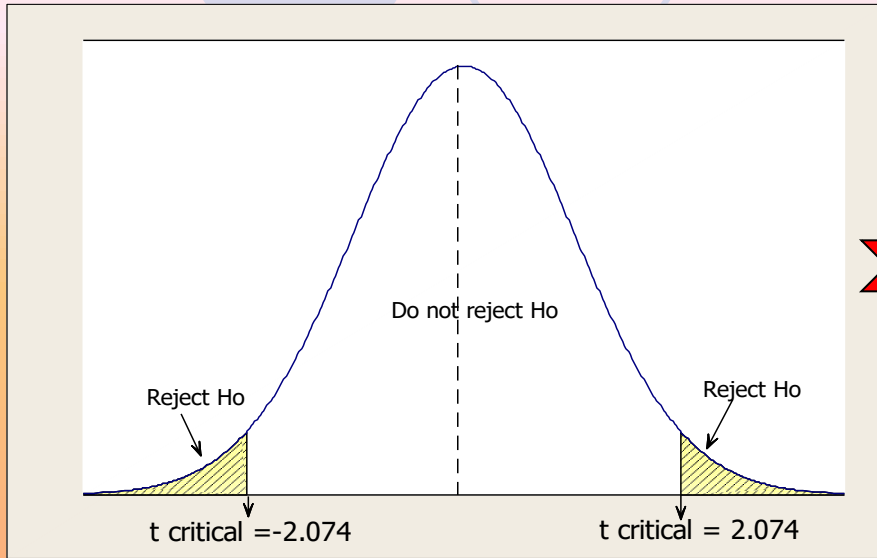
$$df = \frac{(w_1 + w_2)^2}{\frac{w_1^2}{n_1 - 1} + \frac{w_2^2}{n_2 - 1}} = \frac{(0.0289 + 0.00687)^2}{0.0000556 + 0.0000023} = 22.09$$

***Therefore, the degrees of freedom  $df = 22$ .***

The critical points using the t-table

$$t_{df, \alpha/2} = t_{22, 0.025} = 2.074$$

Figure shows the critical values and the areas of rejection and non-rejection.



Decision rule:

Reject  $H_0$  if  $t' > 2.074$   
or, if  $t' < -2.074$

Test Statistic value:

$$t' = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$
$$= \frac{(3.33 - 3.98) - 0}{\sqrt{(0.0289 + 0.006876)}} = -3.44$$

**$t' = -3.44 < t_{\text{critical}} = -2.074$ ; therefore; reject  $H_0$  and conclude that the average processing times differ.**

# Testing the Equality of Two Variances using MINITAB

- The test of variances or the homogeneity of variance is used to ensure that variances are equal before testing the means ( $\mu$ ).
- As indicated earlier, this test is important because unequal variances affect the test statistic formula used to test the equality of two means.
- It is recommended to test the normality first because if the data are not normally distributed, it may affect the variance test and the conclusion from the equality of variance test.

**Follow the steps below for the test of variances:**

- 1. State the null and alternate hypothesis**

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

- 2. Test the data for normality**
- 3. (a) If the data are normal, use the F-test for testing two variances**  
**(b) If the data are not normal, use the Levine's test**

**Note: If the data are summarized, normality test cannot be performed. Here we test variances when the data are summarized and when they are not.**

# Testing Two Variances using MINITAB (Data not summarized)

## **Example 7**

Test if the variances of the two processes (process 1 and 2) are equal.

### **Process 1**

5.20 5.70 5.75 6.00 5.73 5.51 5.58 5.45 5.41 5.45  
5.47 5.73 6.15 5.61 5.50 5.53 5.45 5.47 5.23 5.70  
5.95 5.59 5.43 5.60 5.37

### **Process 2**

5.47 5.55 5.59 5.53 5.54 5.50 5.55 5.52 5.63 5.59  
5.60 5.35 5.60 5.30 5.61 5.53 5.55 5.53 5.55 5.45  
5.41 5.58

## 1. State the null and alternate hypothesis

$$H_0 : \sigma_1^2 = \sigma_2^2$$

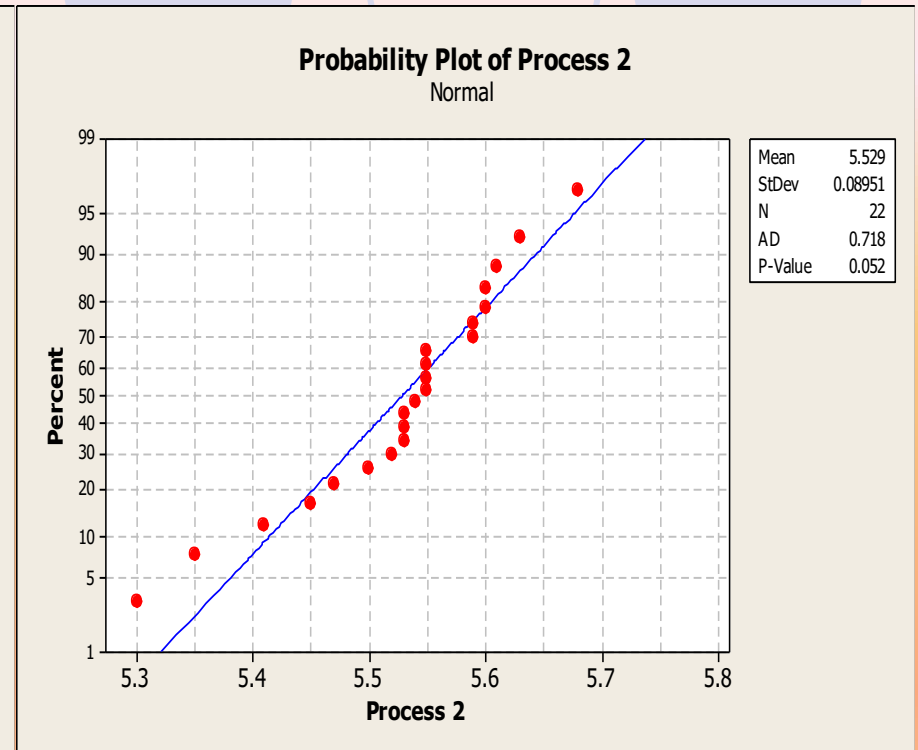
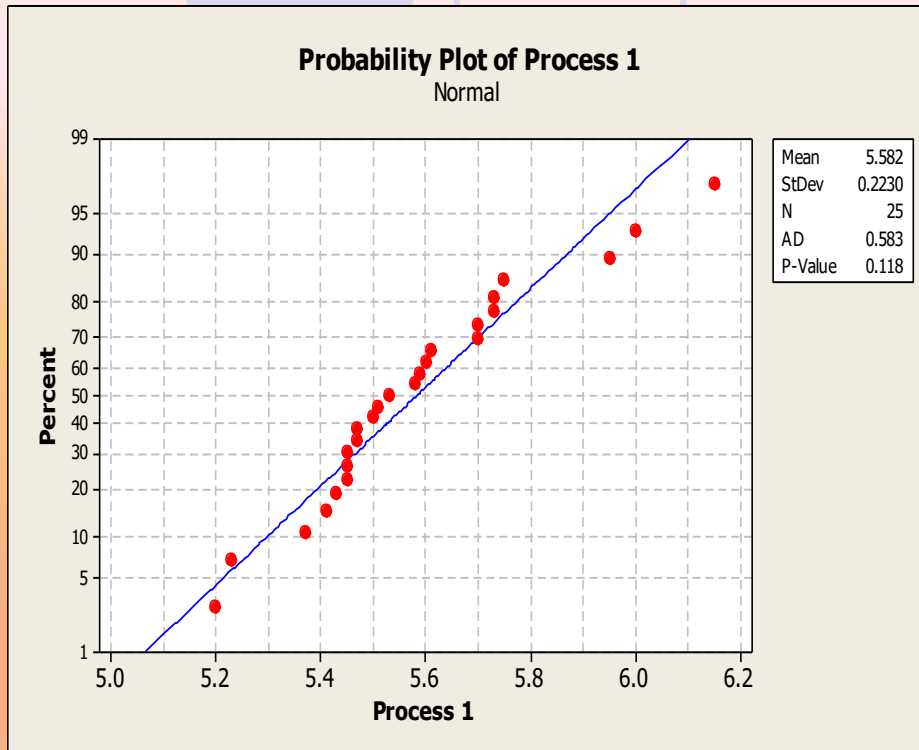
$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

## 2. Test the data for normality – perform normality test (see next slide)

Since the F-test for testing the two variances is based on the assumption of normality, the data should be checked for normality. To check for normality, open the worksheet **TwoVarianceTest.MTW** and follow the instructions:

## **Stat & Basic Statistics & Normality Test**

For variable, select Process 1 and then Process 2 - one at a time and click OK. The normal probability plots are shown in the next slide..

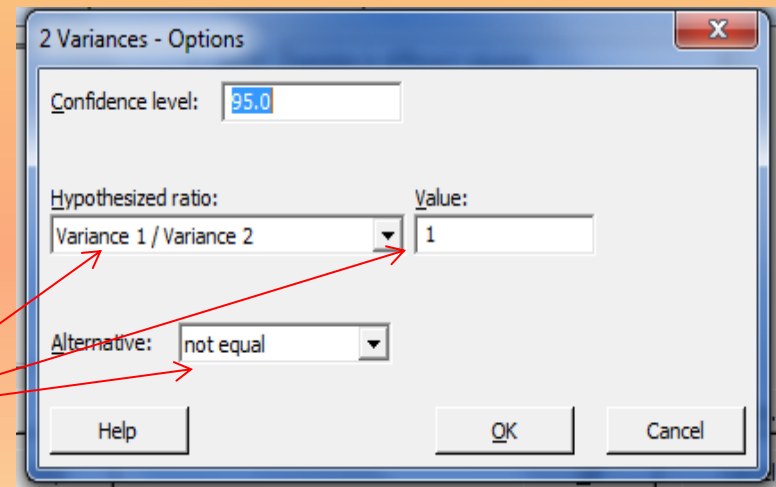
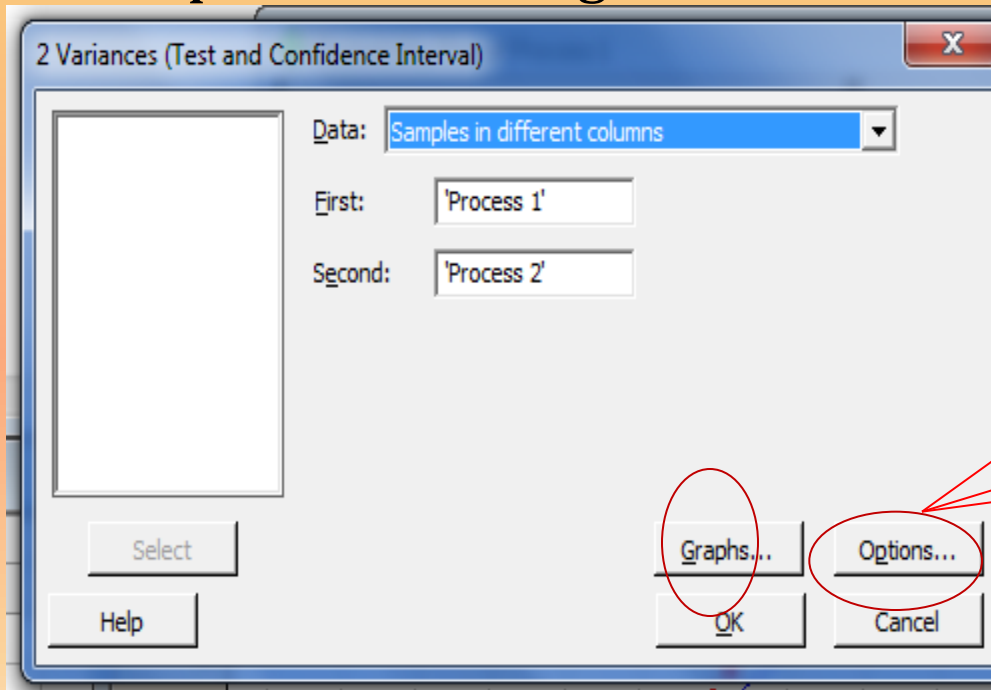


The above probability plots show that both process 1 and 2 data are normal as the p-value in both the cases are more than 0.05. At a significance level,  $\alpha=0.05$ , we cannot reject the null hypothesis that the data follow a normal distribution.

3. Since the data are normal, use the F-test for testing two variances . Open the worksheet `TwoVarianceTest.MTW` and execute the following commands in MINITAB:

## Stat & Basic Statistics & 2-Variances

Complete the dialog box as shown.



Click the graph option to do the box plots and interval plots.

Results 

## Test and CI for Two Variances: Process 1, Process 2

### Method

Null hypothesis                      Variance(Process 1) / Variance(Process 2) = 1  
 Alternative hypothesis          Variance(Process 1) / Variance(Process 2) not = 1  
 Significance level                Alpha = 0.05

### Statistics

Variable	N	StDev	Variance
Process 1	25	0.223	0.050
Process 2	22	0.090	0.008

Ratio of standard deviations = 2.491  
 Ratio of variances = 6.205

### 95% Confidence Intervals

Distribution of Data	CI for StDev Ratio	CI for Variance Ratio
Normal	(1.619, 3.787)	(2.621, 14.340)
Continuous	(1.439, 5.034)	(2.072, 25.346)

### Tests

Method	DF1	DF2	Test Statistic	P-Value
F Test (normal)	24	21	6.21	0.000
Levene's Test (any continuous)	1	45	8.56	0.005

**Since the data are normal, use F-Test to test the hypothesis otherwise; use the Levene's test**

Test the variances

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

p=0.000 < α=0.05;  
 Reject Ho.

# Testing the Equality of Two Means when Variances are not Equal using MINITAB

---

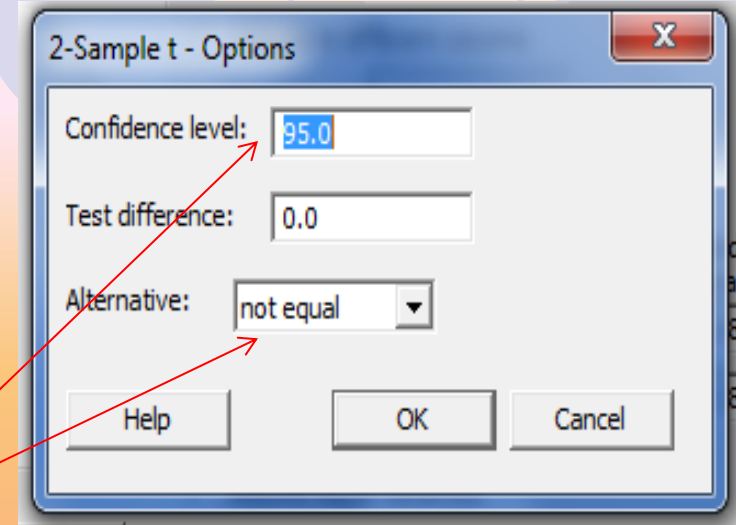
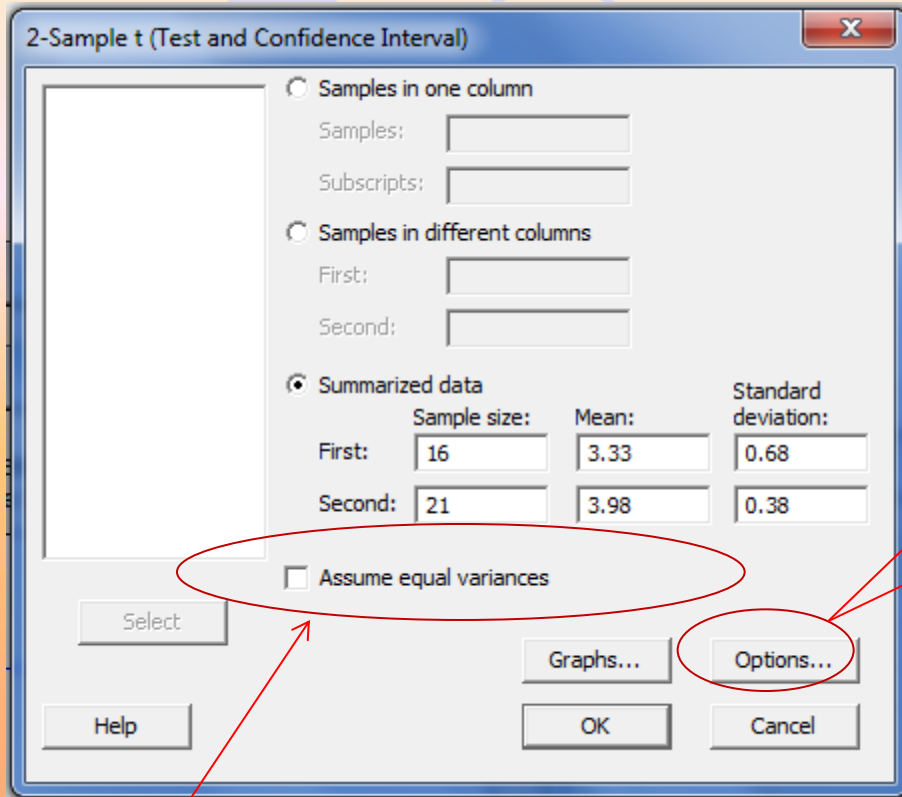
In *Example 6*, we showed the procedure for testing two means when the variances of the two populations were not equal. We performed the test manually.

In this section we will use MINITAB to test the equality of two mean when the variances are not equal. Use the data of *Example 6* and Follow the instructions to perform this test.

Using the MINITAB main menu, perform the following command sequence

## **Stat & Basic Statistics & 2-Sample t**

**Complete the dialog box using the data in Example 6 as shown**



Do not check this box

Results 

## Two-Sample T-Test and CI

Sample	N	Mean	StDev	SE Mean
1	16	3.330	0.680	0.17
2	21	3.980	0.380	0.083

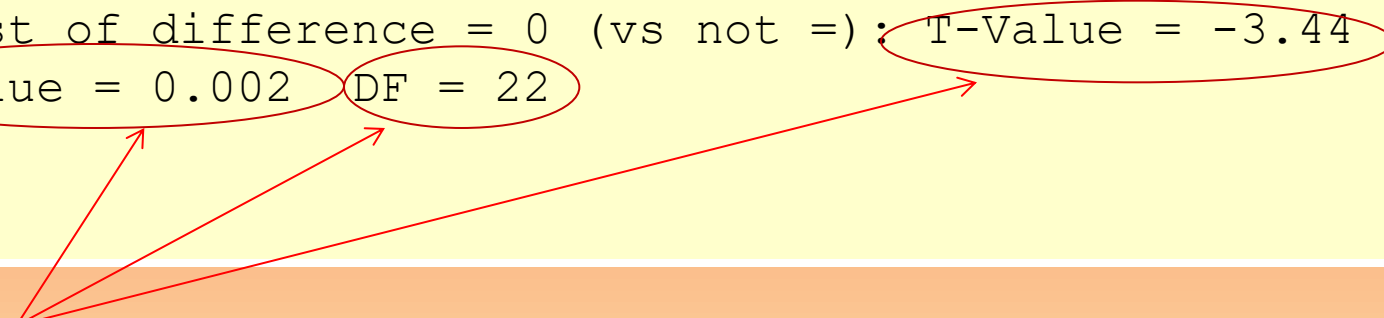
Difference =  $\mu$  (1) -  $\mu$  (2)

Estimate for difference: -0.650

95% CI for difference: (-1.042, -0.258)

T-Test of difference = 0 (vs not =): T-Value = -3.44

P-Value = 0.002 DF = 22



**We can use these values to test the hypothesis. The values are the same we obtained in Example 6 – Step 2 – testing the equality of two means when the variances are not equal.**

# Paired t-test - Testing two means for dependent (related) populations, or hypothesis testing for paired samples

---

- *In some situations, there may be a relationship between the data values of the samples of two populations. The data values from one population may not be independent of the sample values from the other population. The two populations may be considered dependent in such cases.*
- *In cases where the populations are considered related, the observations are paired to prevent other factors from inflating the estimate of the variance. This method is used to improve the precision of comparisons between means.*

## Paired t-test...cont.

---

*The method of testing the difference between the two means when the populations are related is also known as matched sample test or the paired-test. In such cases, the two methods or processes are tested under similar conditions (e.g., we compare the two production methods with the same workers). This leads to a smaller sampling error compared to independent sample test.*

## Paired t-test...cont.

---

- Suppose that  $X_{11}, X_{12}, X_{13}, \dots, X_{1n}$  and  $X_{21}, X_{22}, X_{23}, \dots, X_{2n}$  are samples of the same size from two different normal populations with means  $\mu_1$  and  $\mu_2$ .
- The pairs of data values  $X_{1n}$  and  $X_{2n}$  are related, and are not independent. In this case, we say that the values in one sample are ***paired*** or ***matched*** with the values in other sample; the samples are dependent or related.
- The test is commonly known as ***paired t-test***

## Paired t-test...cont.

The data structure for testing paired samples is shown below.

Observation	Sample 1	Sample 2	Difference $d_i$
1	$X_{11}$	$X_{21}$	$d_1 = X_{11} - X_{21}$
2	$X_{12}$	$X_{22}$	$d_2 = X_{12} - X_{22}$
3	$X_{13}$	$X_{23}$	$d_3 = X_{13} - X_{23}$
4	$X_{14}$	$X_{24}$	$d_4 = X_{14} - X_{24}$
:	:	:	
:	:	:	
n	$X_{1n}$	$X_{2n}$	$d_n = X_{1n} - X_{2n}$

The hypothesis for the matched or paired samples can take one of forms shown below.

$H_0 : \mu_d = 0$	$H_0 : \mu_d \leq 0$	$H_0 : \mu_d \geq 0$
$H_1 : \mu_d \neq 0$	$H_1 : \mu_d > 0$	$H_1 : \mu_d < 0$
<i>Two-tailed or Two-sided test</i>	<i>Right tailed or Right-sided test</i>	<i>Left-tailed or Left-sided test</i>

**Note: the difference,  $d$  can be taken in any order; (sample 1-sample 2) or (sample 2-sample1).**

## Paired t-test...cont.

**Test Statistic:** If the pairs of data values  $X_{1n}$  and  $X_{2n}$  are related, and are not independent; the average of the differences ( $\bar{d}$ ) follows a t-distribution and the test statistics is given by

$$t_{n-1} = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}} \quad (10.29)$$

where,

$$\bar{d} = \text{average of the differences} = \frac{\sum d_i}{n}$$

$$S_d = \text{standard deviation of the differences; } s_d = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n-1}} = \sqrt{\frac{\sum d_i^2 - \frac{(\sum d_i)^2}{n}}{n-1}}$$

n= number of observations (sample size)

$t_{n-1, \frac{\alpha}{2}}$  = critical t-value from the t-table for (n-1) degrees of freedom and appropriate  $\alpha$

The confidence interval given below can also be used to test the hypothesis

$$\bar{d} \pm t_{n-1, \frac{\alpha}{2}} \frac{s_d}{\sqrt{n}} \quad (10.30)$$

### ***Example 8***

A department store wants to determine whether advertising increased its sales of video games. To do this, the management selected 8 of their stores during a two-week period. The number of games sold for the two weeks were recorded. Note that there were no advertisement placed in the first week and the advertising began at the beginning of the second week. The data below shows the number of video games sold after and before the advertisement was placed. Can we conclude from this data that advertising increases the mean number of sales? Use a 5% level of significance.

<i>Store</i>	<i>After</i>	<i>Before</i>
1	60	54
2	65	60
3	73	75
4	67	65
5	75	70
6	80	74
7	85	80
8	70	73

### **Example 8 ... cont.**

The hypothesis we want to test - whether the advertising increases the mean number of sales can be written as:

$$H_0 : \mu_d \leq 0$$

$$H_1 : \mu_d > 0$$

Where,  $d = \text{Difference} = (\text{sales after} - \text{sales before})$ . We can also write the hypothesis as:

$$H_0: \mu_{\text{after}} \leq \mu_{\text{before}} \quad \text{or,} \quad \mu_{\text{after}} - \mu_{\text{before}} \leq 0$$

$$H_1: \mu_{\text{after}} > \mu_{\text{before}} \quad \text{or,} \quad \mu_{\text{after}} - \mu_{\text{before}} > 0$$

The populations are dependent; therefore the test statistic is given by:

$$t_{n-1} = \frac{\bar{d} - \mu_d}{\frac{S_d}{\sqrt{n}}} \quad \text{where,}$$

$$\bar{d} = \frac{\sum d_i}{n} \quad \text{and} \quad S_d = \sqrt{\frac{\sum d_i^2 - \frac{(\sum d_i)^2}{n}}{n-1}}$$

### Example 8... cont.

First calculate,  $\bar{d}$  = average of the differences, and

$S_d$  = standard deviation of the differences.

Store	After	Before	Difference $d_i = \text{after} - \text{before}$	$d_i^2$
1	60	54	+6	36
2	65	60	+5	25
3	73	75	-2	4
4	67	65	+2	4
5	75	70	+5	25
6	80	74	+6	36
7	85	80	+5	25
8	70	73	-3	9
			$\sum d_i = 24$	$\sum d_i^2 = 164$

$$\bar{d} = \frac{\sum di}{n} = \frac{24}{8} = 3.0$$

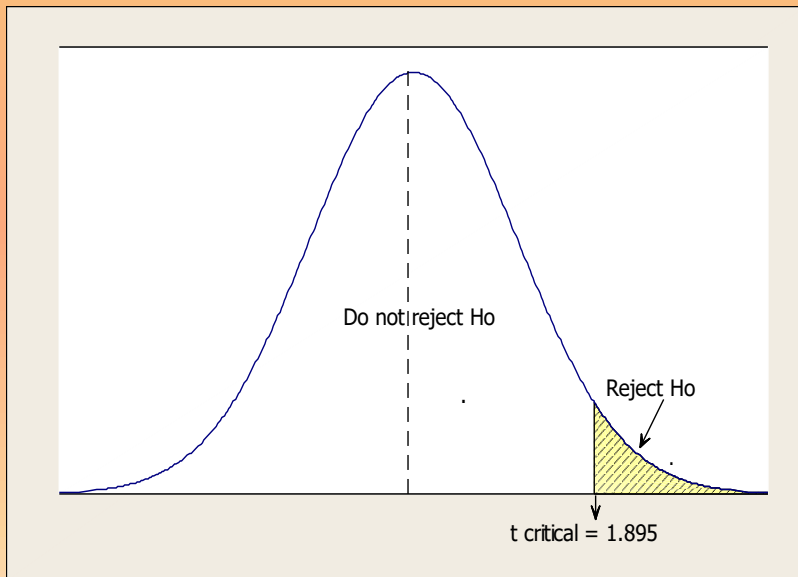
$$S_d = \sqrt{\frac{\sum d_i^2 - \frac{(\sum di)^2}{n}}{n-1}} = \sqrt{\frac{164 - \frac{(24)^2}{8}}{7}} = 3.63$$

## Example 8... cont.

**Test Statistic value**

$$t_{n-1} = \frac{\bar{d} - \mu_d}{Sd / \sqrt{n}}$$
$$= \frac{3 - 0}{3.63 / \sqrt{8}} = 2.34$$

**Critical Point:**  $t_{n-1,0.05} = t_{7,0.05} = 1.895$



Decision rule: Reject  $H_0$  if  $t_{n-1} > 1.895$

$t_{n-1} = 2.34 > t_{\text{crit}} = 1.895$ ; reject  $H_0$   
which means that the average sales after the advertisement was greater than before the advertisement.

# Paired t-test using MINITAB

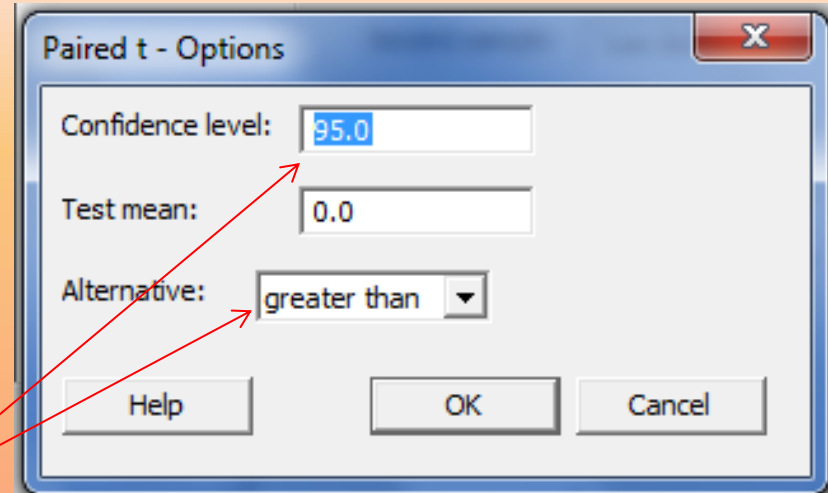
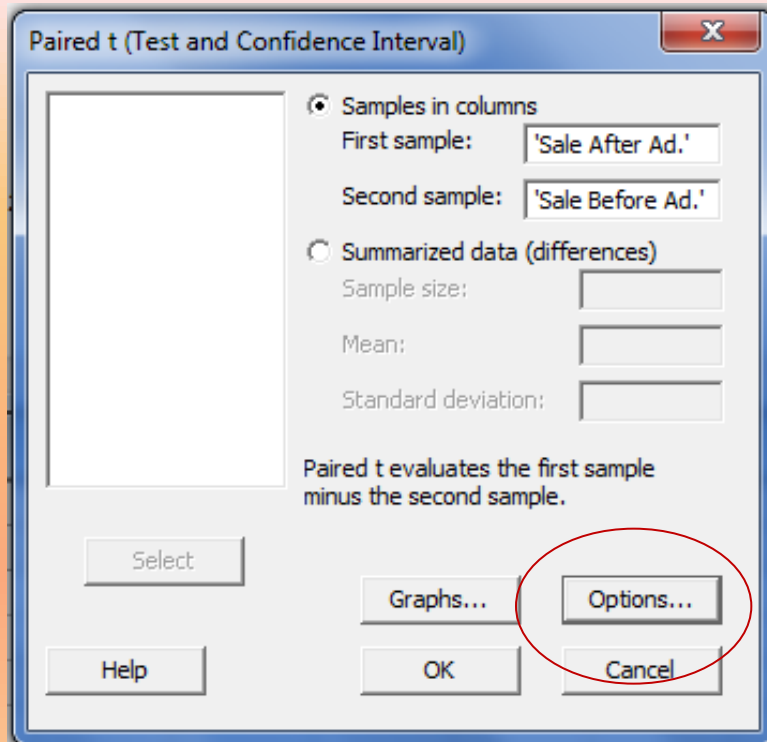
We will demonstrate the procedure of paired test in MINITAB using the data of Example 8. Label columns C1 and C2 of MINITAB worksheet as shown or open the file **Paired t-test 1.MTW** then follow the steps:

	C1	C2
	Sale	Sale
	After	Before
Row	Ad.	Ad.
1	60	54
2	65	60
3	73	75
4	67	65
5	75	70
6	80	74
7	85	80
8	70	73

**Stat & Basic Statistics & Paired t**

**Complete the dialog box shown on the next slide**

# Paired t-test using MINITAB...cont.



**Click OK in all dialog boxes. The results are shown in the next slide.**

## Paired t-test using MINITAB...cont.

### Paired T-Test and CI: Sale After Ad., Sale Before Ad.

Paired T for Sale After Ad. - Sale Before Ad.

	N	Mean	StDev	SE Mean
Sale After Ad.	8	71.88	8.15	2.88
Sale Before Ad.	8	68.88	8.63	3.05
Difference	8	3.00	3.63	1.28

95% lower bound for mean difference: 0.57

T-Test of mean difference = 0 (vs > 0): T-Value = 2.34

P-Value = 0.026

**The values calculated above match the manual calculations in Example 8. The hypothesis can be tested using the T-value or the p-value above (see Example 8 for the test).**

# Testing Two Proportions - Attribute Data

The hypothesis test involving two proportions can be a one-sided or a two-sided test. The hypotheses are written as

$$H_0 : p_1 = p_2 \quad \text{or,} \quad H_0 : p_1 - p_2 = 0$$

$$H_0 : p_1 \neq p_2 \quad H_0 : p_1 - p_2 \neq 0$$

**Two-sided test**

$$H_0 : p_1 \geq p_2 \quad \text{or,} \quad H_0 : p_1 = p_2$$

$$H_0 : p_1 < p_2 \quad H_0 : p_1 = p_2$$

**Left-sided test**

$$H_0 : p_1 \leq p_2 \quad \text{or,} \quad H_0 : p_1 - p_2 \leq 0$$

$$H_0 : p_1 > p_2 \quad H_0 : p_1 - p_2 > 0$$

**Right-sided test**

**Test Statistic:** The difference between two sample proportions  $(\bar{p}_1 - \bar{p}_2)$  can be approximated by a normal distribution for large samples. Therefore, the test statistic using a normal approximation is given by

$$z = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\sqrt{\bar{p}(1 - \bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$x_1$  is the number having the characteristic of interest in the first population and,  $x_2$  is the number having the characteristic in the second population;

$p_1$  and  $p_2$  are the sample proportions of the two samples that have the characteristic of interest where,

$$\bar{p}_1 = \frac{x_1}{n_1} \text{ and}$$

$$\bar{p}_2 = \frac{x_2}{n_2}$$

$\bar{p}$  is the “pooled” or combined proportion given by

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2} \text{ or, } \bar{p} = \frac{n_1 \bar{p}_1 + n_2 \bar{p}_2}{n_1 + n_2}$$

### Example 9

Two numerical controlled lathe machines (machine A and machine B) are used in a production process to produce aluminum rods with specification  $5 \pm 0.01$  cm. If the diameters of the rods do not meet specifications, they are considered to be nonconforming. The production supervisor suspects that the percent of nonconforming rods produced by machine A is more than that produced by machine B because of some error in the part program used in machine A. However, the machine operator believes that there is no difference in the proportion of nonconforming rods produced by two machines. The data below shows the number of nonconforming parts produced by two machines.

Machine A	Machine B
$n_1 = 200$	$n_2 = 200$
Number of nonconforming, = 43	Number of nonconforming, = 37

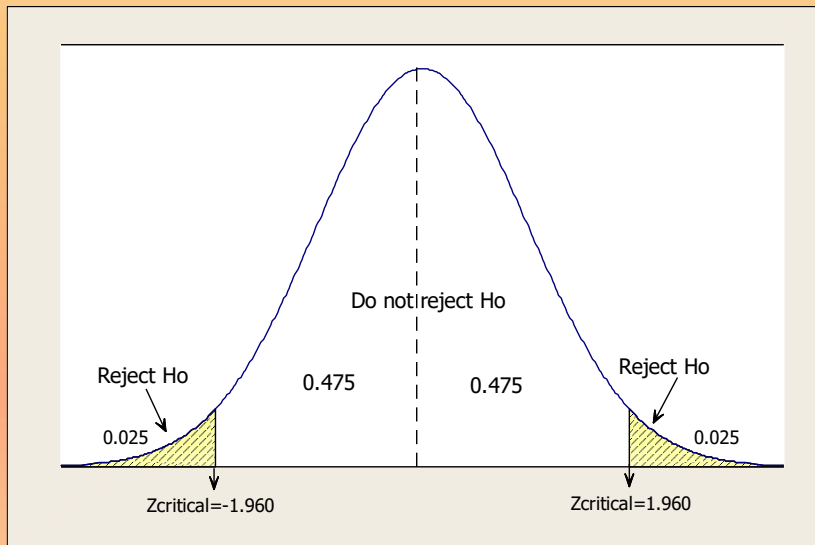
Test the hypothesis that the proportion of nonconforming circles produced by two machines do not differ. Use a 5% level of significance.

## The hypothesis to be tested:

$$H_0 : p_1 = p_2 \Rightarrow p_1 - p_2 = 0$$

$$H_1 : p_1 \neq p_2 \Rightarrow p_1 - p_2 \neq 0$$

## Critical points:



## Decision rule:

Reject  $H_0$  if  $z > 1.96$

or, if  $z < -1.96$

## Test statistic

Before calculating the test statistic value, calculate the following proportions:

$$\bar{p}_1 = \frac{x_1}{n_1} = \frac{43}{200} = 0.215$$

$$\bar{p}_2 = \frac{x_2}{n_2} = \frac{37}{200} = 0.185$$

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{43 + 37}{200 + 200} = 0.20$$

## Test statistic value:

$$z = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{(0.215 - 0.185) - 0}{\sqrt{(0.20)(1-0.20)\left(\frac{1}{200} + \frac{1}{200}\right)}} = 0.75$$

$z = 0.75 < z_{critical} = 1.96$  Do not reject  $H_0$ .

## (b) Use the p-value approach to test the hypothesis in part (a).

The decision rule for the p-value approach

If  $p \geq \alpha$  do not reject  $H_0$

if  $p < \alpha$  Reject  $H_0$