Introduction to Statistics and Data Science using *eStat*

Chapter 8 Testing Hypothesis for Two Populations

8.1 Testing hypothesis for two population means - Independent sample -

Jung Jin Lee
Professor of Soongsil University, Korea
Visiting Professor of ADA University, Azerbaijan

8.1 Testing hypothesis for two population means

- Examples comparing the mean of the two populations.
  - Is there a difference between starting salary of male and female for this year's college graduates?
  - Is there a difference in the weight of the products produced in two production lines?
  - Did special training given to the typist to increase the speed of typing really bring about an increase in the speed of typing?

- Comparison of two population means is possible by testing hypothesis.
  - independent sample
  - paired sample

8.1 Testing hypothesis for two population means

8.1.1 Two Independent Samples

- Testing hypothesis for two population means:

  1) $H_0 : \mu_1 - \mu_2 = D_0$
     $H_1 : \mu_1 - \mu_2 > D_0$

  2) $H_0 : \mu_1 - \mu_2 = D_0$
     $H_1 : \mu_1 - \mu_2 < D_0$

  3) $H_0 : \mu_1 - \mu_2 = D_0$
     $H_1 : \mu_1 - \mu_2 \neq D_0$

  * D_0 is value for difference in population means

- Estimator of difference in population means $\mu_1 - \mu_2$

  ⇨ difference in sample means $\bar{X}_1 - \bar{X}_2$

8.1 Testing hypothesis for two population means

8.1.1 Two Independent Samples

- Sampling distribution of $\bar{X}_1$ - $\bar{X}_2$ if σ_1^2 and σ_2^2 are known, large samples

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

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \approx N(0,1)$$

- Sampling distribution of $\bar{X}_1$ - $\bar{X}_2$ if σ_1^2 and σ_2^2 are unknown

  - If two populations follow normal distributions and variances are same

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{n_1}+\frac{s_p^2}{n_2}}} \approx t_{n_1+n_2-2}$$

where $s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{(n_1-1)+(n_2-1)}$ is a pooled variance

8.1 Testing hypothesis for two population means

8.1.1 Two Independent Samples

- **Sampling distribution of $\bar{X}_1$ - $\bar{X}_2$ if σ_1^2 and σ_2^2 are unknown**

- If two populations follow normal distributions and variances are different

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \approx t_\varphi$$

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

8.1 Testing hypothesis for two population means

Table 8.1.1 Testing hypothesis of two population means
- independent samples, populations are normal distributions, case of two population variances are equal

Type of Hypothesis	Decision Rule
1) $H_0: \mu_1 - \mu_2 = D_0$ $H_1: \mu_1 - \mu_2 > D_0$	If $\dfrac{(\overline{X}_1 - \overline{X}_2) - D_0}{\sqrt{\dfrac{s_p^2}{n_1} + \dfrac{s_p^2}{n_2}}} > t_{n_1+n_2-2;\alpha}$, then reject H_0, else accept H_0
2) $H_0: \mu_1 - \mu_2 = D_0$ $H_1: \mu_1 - \mu_2 < D_0$	If $\dfrac{(\overline{X}_1 - \overline{X}_2) - D_0}{\sqrt{\dfrac{s_p^2}{n_1} + \dfrac{s_p^2}{n_2}}} < -t_{n_1+n_2-2;\alpha}$, then reject H_0, else accept H_0
3) $H_0: \mu_1 - \mu_2 = D_0$ $H_1: \mu_1 - \mu_2 \neq D_0$	If $\left\| \dfrac{(\overline{X}_1 - \overline{X}_2) - D_0}{\sqrt{\dfrac{s_p^2}{n_1} + \dfrac{s_p^2}{n_2}}} \right\| > t_{n_1+n_2-2;\alpha/2}$, then reject H_0, else accept H_0

8.1 Testing hypothesis for two population means

Table 8.1.2 Testing hypothesis of two population means
- independent samples, populations are normal distributions,
two population variances are assumed to be different

Type of Hypothesis	Decision Rule
1) $H_0: \mu_1 - \mu_2 = D_0$ $H_1: \mu_1 - \mu_2 > D_0$	If $\dfrac{(\overline{X}_1 - \overline{X}_2) - D_0}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} > t_{\phi;\alpha}$, then reject H_0, else accept H_0
2) $H_0: \mu_1 - \mu_2 = D_0$ $H_1: \mu_1 - \mu_2 < D_0$	If $\dfrac{(\overline{X}_1 - \overline{X}_2) - D_0}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} < -t_{\phi;\alpha}$, then reject H_0, else accept H_0
3) $H_0: \mu_1 - \mu_2 = D_0$ $H_1: \mu_1 - \mu_2 \neq D_0$	If $\left\lvert \dfrac{(\overline{X}_1 - \overline{X}_2) - D_0}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} \right\rvert > t_{\phi;\alpha/2}$, then reject H_0, else accept H_0

8.1 Testing hypothesis for two population means

[Ex 8.1.1] Two machines produce a cookie at a factory and a cookie package has a static capacity of 270 grams. Assume two population variance are equal.

- Samples were extracted from each of packages by two machines to examine weight of package.
- Average weight of 15 packages extracted from machine 1 was 275g, standard deviation was 12g, and average weight of 14 packages extracted from machine 2 was 269g and standard deviation was 10g.
- Test whether weights of cookie bags produced by two machines are different at the 1% significance level.
- Check the test result using 『eStatU』.

8.1 Testing hypothesis for two population means

<Answer of Ex 8.1.1>

- The hypothesis of this problem is $H_0 : \mu_1 = \mu_2$, $H_1 : \mu_1 \neq \mu_2$. Hence the decision rule is as follows.

If $\left| \dfrac{(\overline{X}_1 - \overline{X}_2) - D_0}{\sqrt{\dfrac{s_p^2}{n_1} + \dfrac{s_p^2}{n_2}}} \right| > t_{n_1+n_2-2\,;\,\alpha/2}$, then reject H_0

The information in the example can be summarized as follows.

$$n_1 = 15,\ \overline{X}_1 = 275,\ s_1 = 12,$$
$$n_2 = 14,\ \overline{X}_2 = 269,\ s_2 = 10$$

Therefore,

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$
$$= \frac{(15-1)12^2 + (14-1)10^2}{15 + 14 - 2} = 122.815$$
$$\left| \frac{275 - 269}{\sqrt{\dfrac{122.815}{15} + \dfrac{122.815}{14}}} \right| = 1.457$$
$$t_{15+14-2\,;\,0.01/2} = t_{27\,;\,0.005} = 2.7707$$

Since 1.457 < 2.7707, H_0 cannot be rejected.

8.1 Testing hypothesis for two population means

<Answer of Ex 8.1.1>

Testing Hypothesis μ_1, μ_2 Menu

[Hypothesis] $H_0: \mu_1 - \mu_2 = D$ [0]

◉ $H_1: \mu_1 - \mu_2 \neq D$ ○ $H_1: \mu_1 - \mu_2 > D$ ○ $H_1: \mu_1 - \mu_2 < D$

[Test Type] t test , Variance Assumption ◉ $\sigma_1^2 = \sigma_2^2$ ○ $\sigma_1^2 \neq \sigma_2^2$

Significance Level $\alpha =$ ○ 5% ◉ 1%

Sampling Type ◉ independent sample ○ paired sample

[Sample Data] *Input either sample data using BSV or sample statistics at the next boxes*

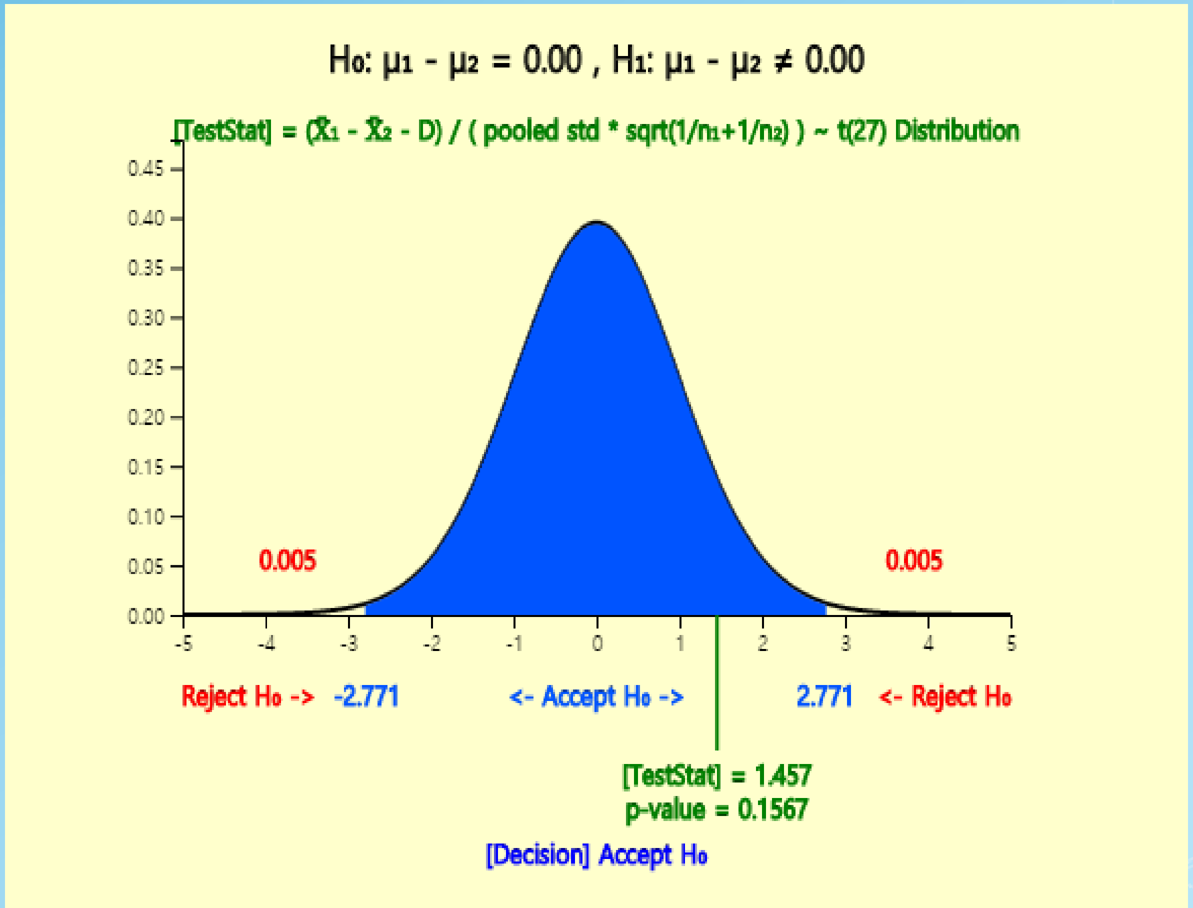
Sample 1 [ ]

Sample 2 [ ]

[Sample Statistics]

Sample Size	n_1	=	15	n_2	=	14		
Sample Mean	$\bar{x}_1$	=	275	$\bar{x}_2$	=	269	$\bar{x}_d$	=
Sample Variance	s_1^2	=	144	s_2^2	=	100	s_d^2	=

Execute



8.1 Testing hypothesis for two population means

[Example 8.1.2] If two population variances are assumed to be different in [Example 8.1.1], test whether weights of cookie bags produced from two machines are equal or not at a 1% significance level. Check the test result using 『eStatU』.

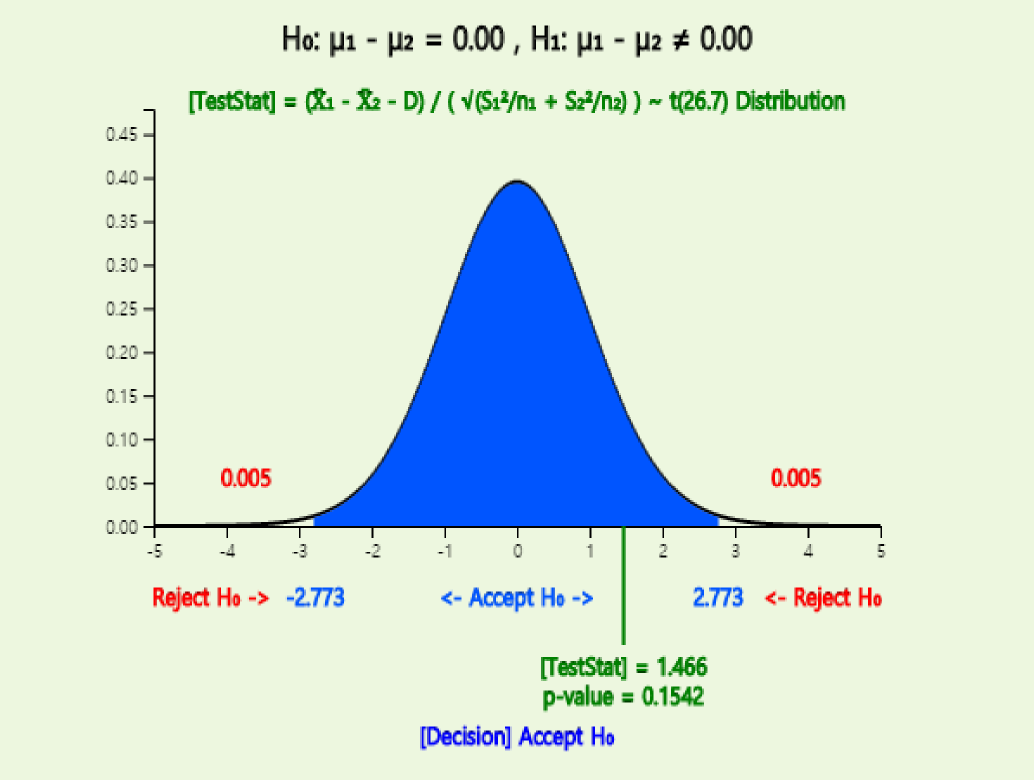
<Answer>

$$t_{obs} = \left| \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \right| = \left| \frac{275 - 269}{\sqrt{\frac{144}{15} + \frac{100}{14}}} \right| = 1.466$$

$$\varphi = \frac{\{\frac{144}{15} + \frac{100}{14}\}^2}{\frac{\{\frac{144}{15}\}^2}{(15-1)} + \frac{\{\frac{100}{14}\}^2}{(14-1)}} = 26.67$$

$$t_{26.7;0.005} = 2.773$$

Since 1.466 < 2.773, H_0 cannot be rejected



8.1 Testing hypothesis for two population means

[Example 8.1.3]
A sample of 10 men and women in the male and female populations of college graduates this year was taken and the monthly average wage was examined as follows. (Unit 10,000 KRW)

Men	272 255 278 282 296 312 356 296 302 312
Women	276 280 369 285 303 317 290 250 313 307

1) If the population variances are the same, test the hypothesis at a significant level of 5% whether the average monthly wage for male and female is the same.
2) If you assume that the population variances are different, test the hypothesis at a significant level of 5% whether the average monthly wage for male and female is the same.

8.1 Testing hypothesis for two population means

<Answer of Ex 8.1.3>

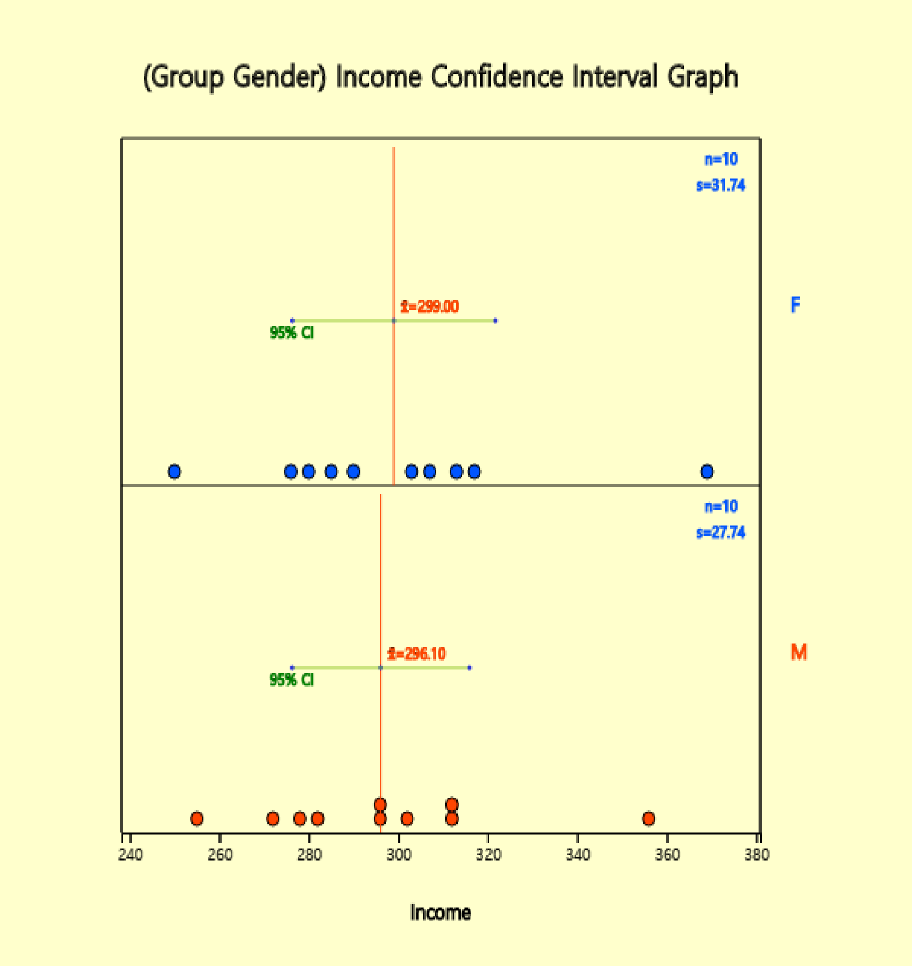
File: Ex813IncomByGender.csv

Analysis Var: 2: Income — by Group: 1: Gender

(Selected data: Raw Data) (or Paired Var)

SelectedVar: V2 by V1,

	Gender	Income	V3	V4
1	M	272		
2	M	255		
3	M	278		
4	M	282		
5	M	296		
6	M	312		
7	M	356		
8	M	296		
9	M	302		
10	M	312		
11	F	276		
12	F	280		
13	F	369		
14	F	285		
15	F	303		
16	F	317		
17	F	290		
18	F	250		
19	F	313		
20	F	307		



Confidence Interval Graph | Histogram

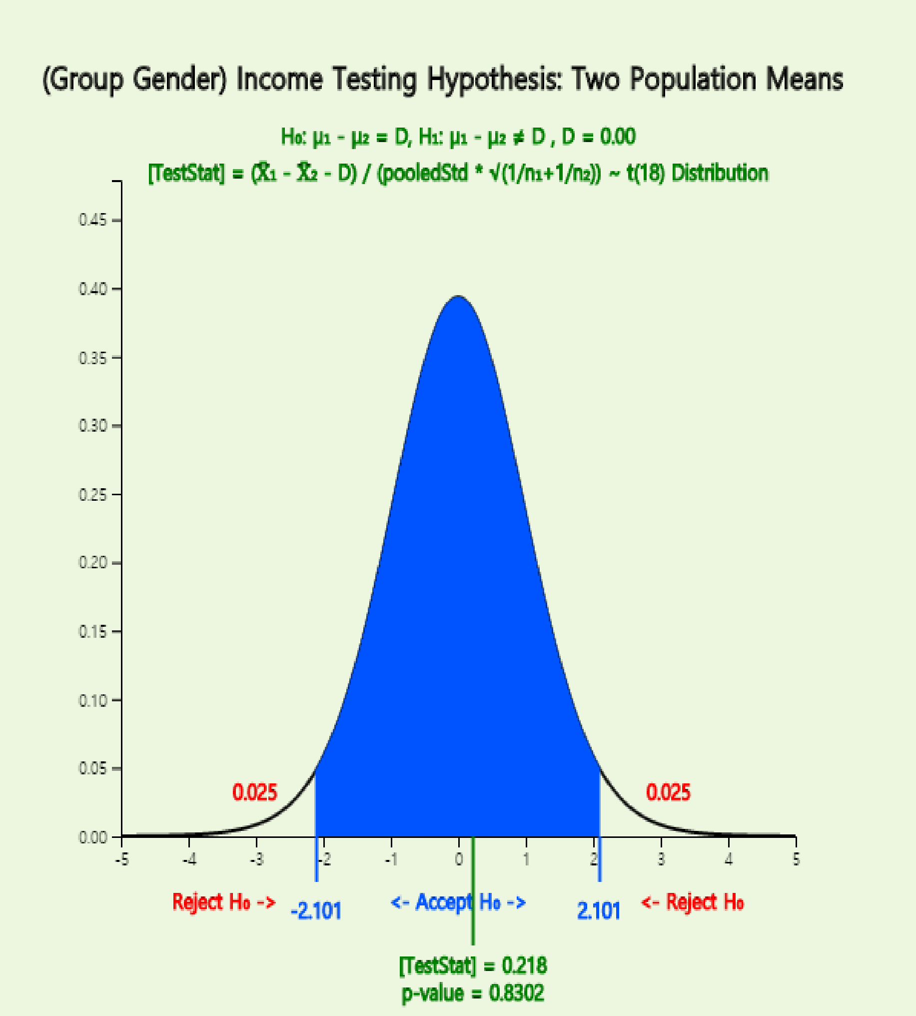
$H_0 : \mu_1 - \mu_2 = D$ [0] ◉ $H_1 : \mu_1 - \mu_2 \neq D$ ○ $H_1 : \mu_1 - \mu_2 > D$ ○ $H_1 : \mu_1 - \mu_2 < D$

Variance Assumption ◉ $\sigma_1^2 = \sigma_2^2$ ○ $\sigma_1^2 \neq \sigma_2^2$

Significance Level α = ◉ 5% ○ 1% *Confidence Level* ◉ 95% ○ 99% | t test | Rank Sum Test

8.1 Testing hypothesis for two population means

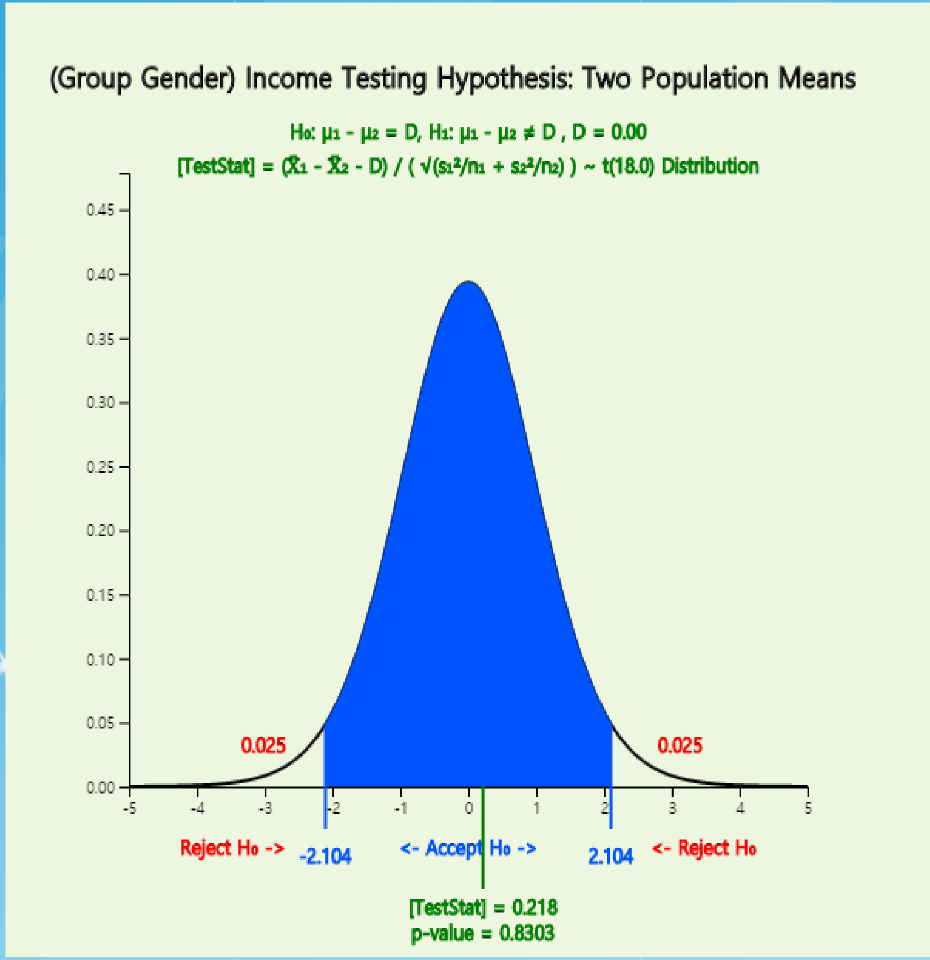
<Answer of Ex 8.1.3>



Testing Hypothesis: Two Population Means	Analysis Var	Income	Group Name	Gender	
Statistics	Observation	Mean	Std Dev	std err	Population Mean 95% Confidence Interval
1 (F)	10	299.000	31.742	10.038	(276.293, 321.707)
2 (M)	10	296.100	27.739	8.772	(276.257, 315.943)
Total	20	297.550	29.051	6.496	(283.954, 311.146)
Missing Observations	0				
Hypothesis	Variance Assumption	$\sigma_1^2 = \sigma_2^2$			
$H_0 : \mu_1 - \mu_2 = D$	D	[TestStat]	t value	p-value	μ_1-μ_2 95% Confidence Interval
$H_1 : \mu_1 - \mu_2 \neq D$	0.00	Difference of Sample Means	0.218	0.8302	(-25.106, 30.906)

8.1 Testing hypothesis for two population means

<Answer of Ex 8.1.3>



Testing Hypothesis: Two Population Means	Analysis Var	Income	Group Name	Gender	
Statistics	Observation	Mean	Std Dev	std err	Population Mean 95% Confidence Interval
1 (F)	10	299.000	31.742	10.038	(276.293, 321.707)
2 (M)	10	296.100	27.739	8.772	(276.257, 315.943)
Total	20	297.550	29.051	6.496	(283.954, 311.146)
Missing Observations	0				
Hypothesis	Variance Assumption	$\sigma_1^2 \neq \sigma_2^2$			
$H_0 : \mu_1 - \mu_2 = D$	D	[TestStat]	t value	p-value	μ_1-μ_2 95% Confidence Interval
$H_1 : \mu_1 - \mu_2 \neq D$	0.00	Difference of Sample Means	0.218	0.8303	(-25.142, 30.942)

Thank you