Chapter 5

Estimation, testing hypothesis and regression analysis

Professor Jung Jin Lee
Soongsil University, Korea
New Uzbekistan University, Uzbekistan

Chapter 5 Estimation, testing hypothesis and regression analysis

5.1 Sampling distribution and estimation

- 5.1.1 Sampling distribution of sample means
- 5.1.2 Estimation of a population mean

5.2 Testing hypothesis for a population mean

5.3 Testing hypothesis for two populations means

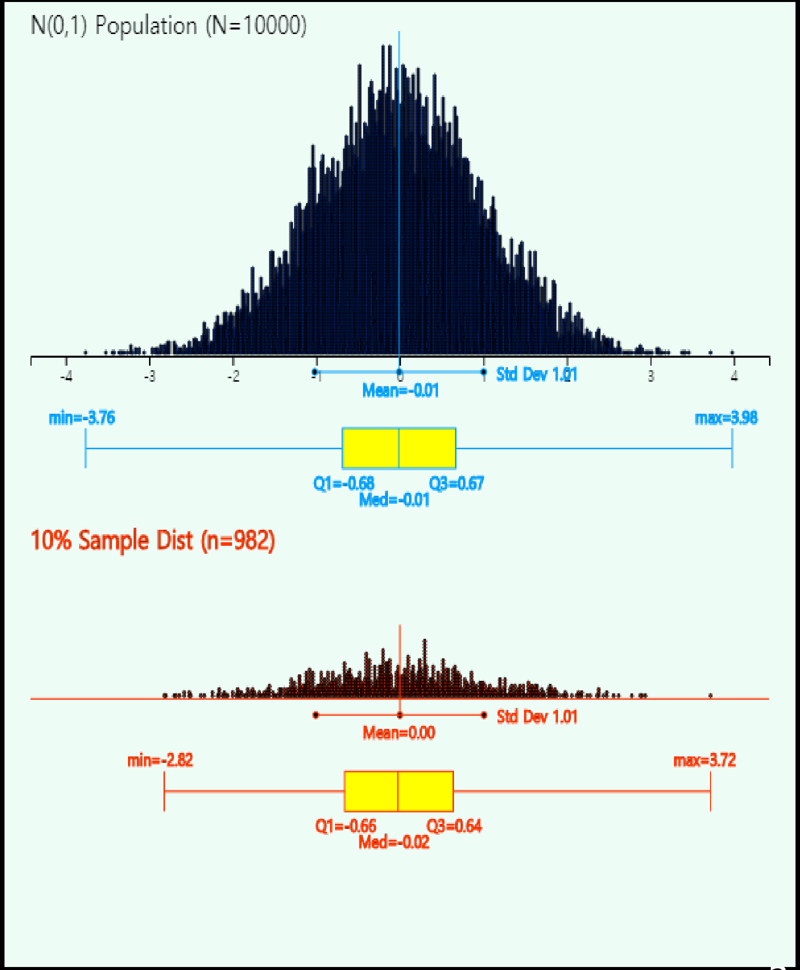
5.4 Testing hypothesis for several population means: Analysis of variance

5.5 Regression analysis

- 5.5.1 Correlation analysis
- 5.5.2 Simple linear regression
- 5.5.3 Multiple linear regression

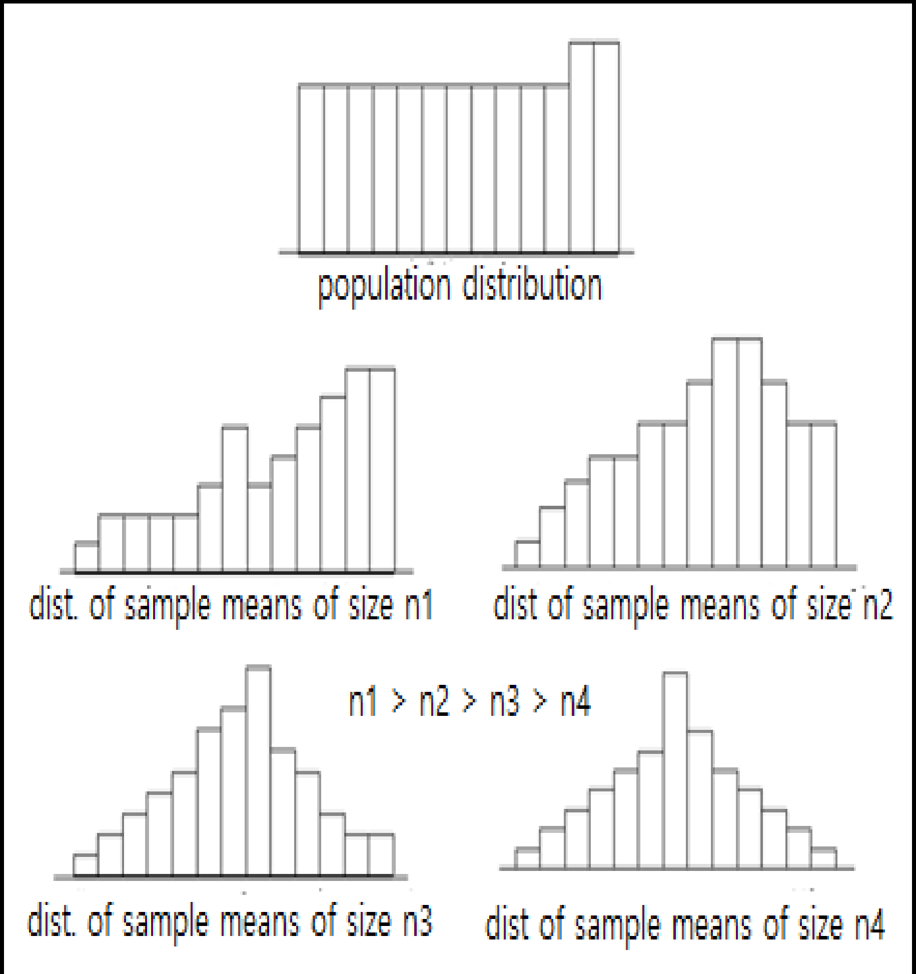
5.1 Sampling distribution and estimation

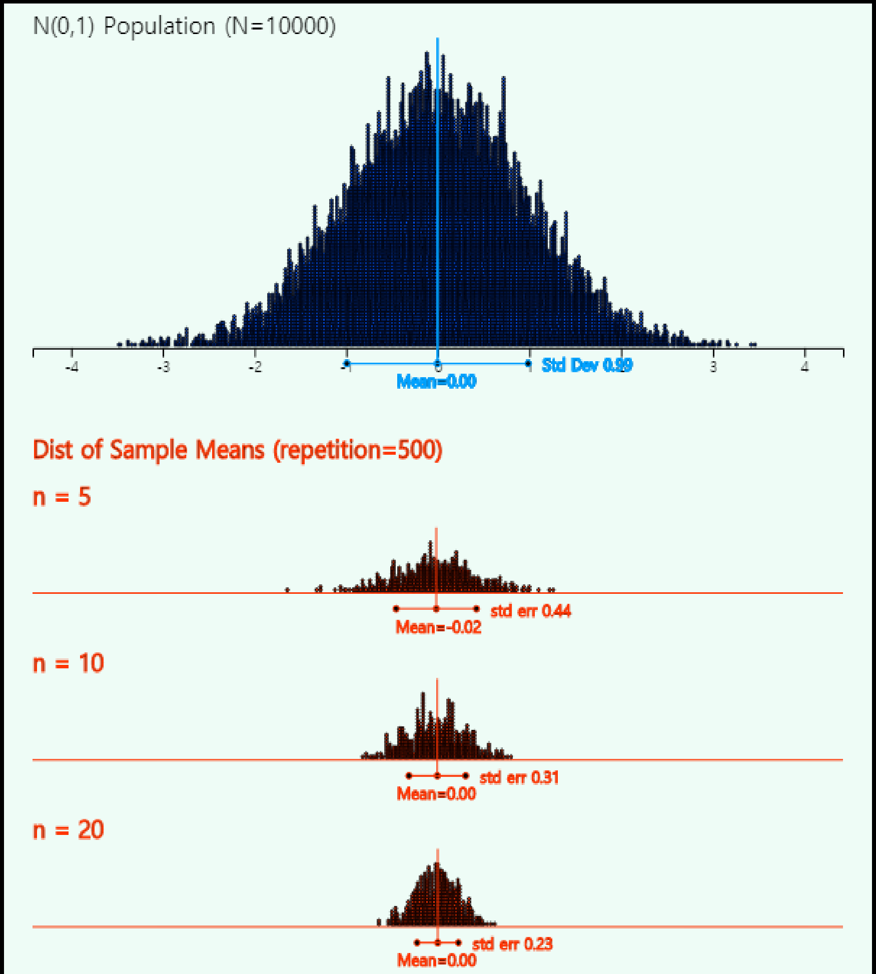
- Inferential statistics is used to find out characteristics of unknown populations.
- Characteristics of a population usually refers to the population mean, variance, etc.,
- Characteristic values of a population are called parameters.
- The parameters are estimated using sample characteristics, such as a sample mean and sample variance.



5.1 Sampling distribution and estimation

❖ Central Limit Theorem(CLT)





5.1 Sampling distribution and estimation

Central limit theorem

If a population has an infinite elements with a mean μ and variance σ^2, then, if the sample size is large enough, the distribution of all possible sample means is an approximately normal distribution $N(\mu, \frac{\sigma^2}{n})$. We can summarize specifically the central limit theorem as follows.

1) The average of all possible sample means, $\mu_{\overline{X}}$, is equal to the population mean μ.
   (i.e., $\mu_{\overline{X}} = \mu$)
2) The variance of all possible sample means, $\sigma^2_{\overline{X}}$, is the population variance divided by n.
   (i.e., $\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$)
3) The distribution of all possible sample means is approximately a normal distribution.

The above facts can be briefly written as $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$.

5.1 Sampling distribution and estimation

❖ Estimation

A. Point estimation of population mean

- An observed value of the sample mean is a point estimate of the population mean.

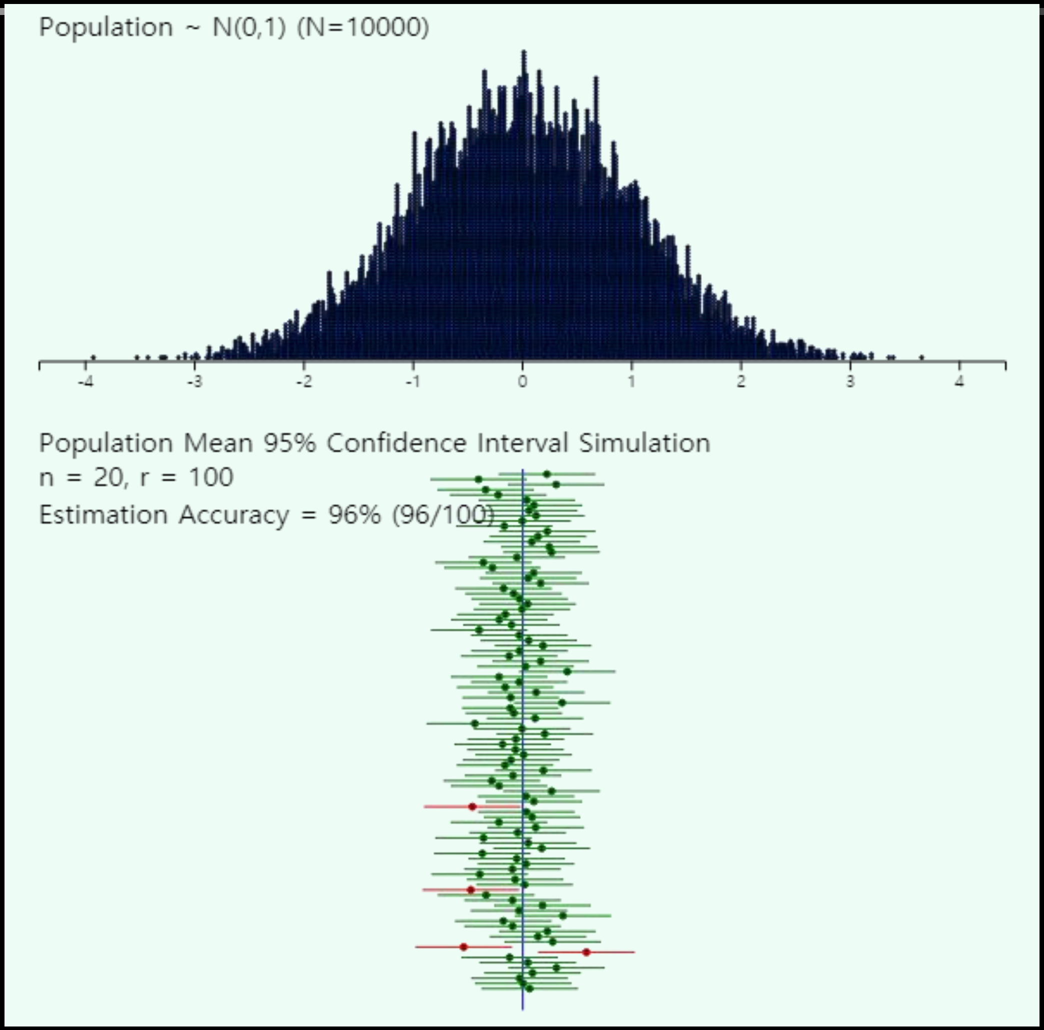
B. Interval estimation of population mean

☞ 100(1-α)% Confidence Interval for Population Mean μ

--- Population is normal and population variance σ^2 is known

$$\left[\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\ ,\ \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

5.1 Sampling distribution and estimation



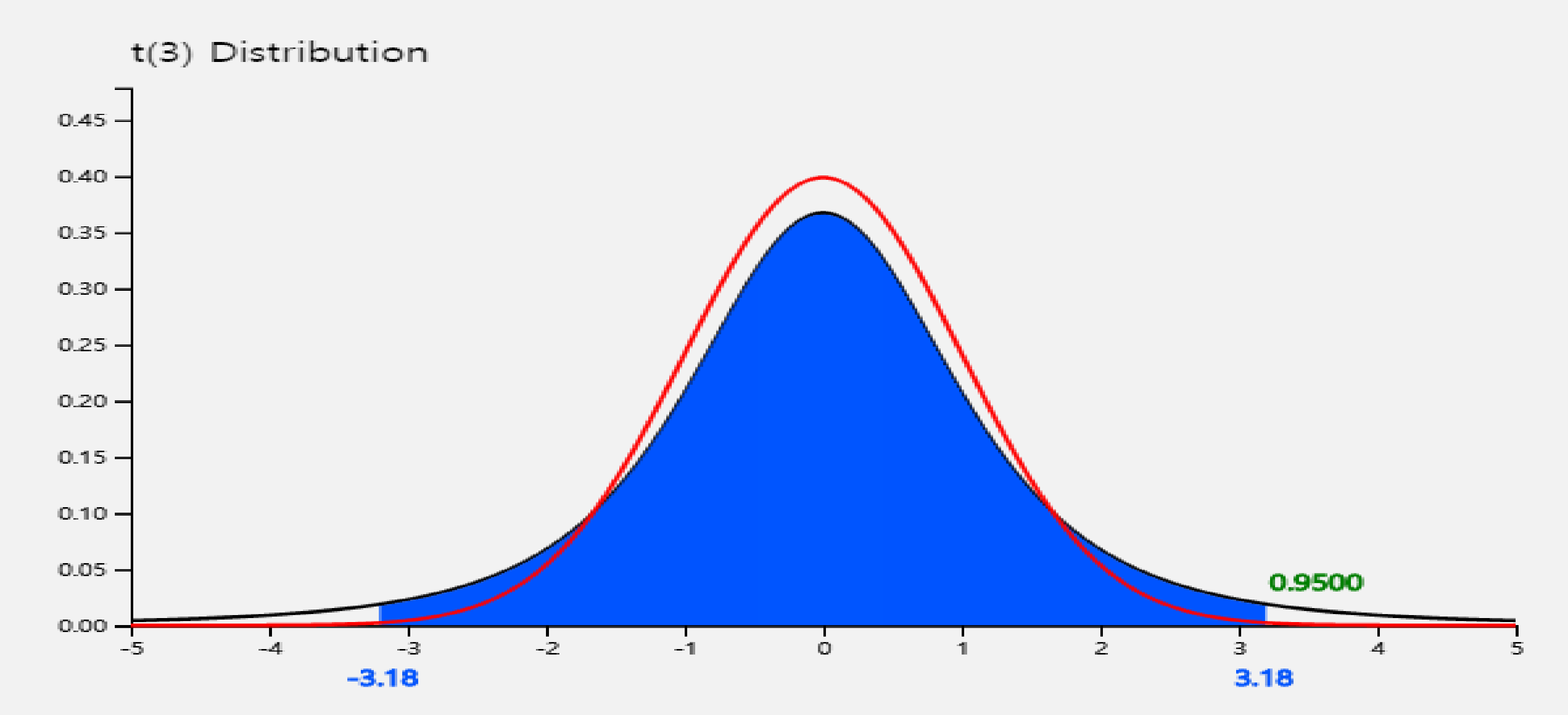
5.1 Sampling distribution and estimation

☞ 100(1-α)% Confidence Interval for Population Mean μ

--- Population is normal and population variance σ^2 is unknown

$$\left[\overline{X} - t_{n-1\,;\,\alpha/2}\cdot\frac{S}{\sqrt{n}}\ ,\ \overline{X} + t_{n-1\,;\,\alpha/2}\cdot\frac{S}{\sqrt{n}}\right]$$

n is the sample size and S is the sample standard deviation.



5.1 Sampling distribution and estimation

Example 5.1.2 Suppose we do not know the population variance in Example 4.4.2. If the sample size is 25 and the sample standard deviation is 5 (unit: 10,000 KRW), estimate the mean of the starting salary of college graduates at the 95% confidence level.

Answer

Since we do not know the population variance, we should use the t distribution for interval estimation of the population mean. Since $t_{n-1:\ \alpha/2} = t_{25-1:\ 0.05/2} = t_{25-1:\ 0.025} = 2.0639$, the 95% confidence interval of the population mean is as follows.

$$\left[\overline{X} - t_{n-1:\ \alpha/2}\frac{S}{\sqrt{n}}, \overline{X} + t_{n-1:\ \alpha/2}\frac{S}{\sqrt{n}}\right]$$

$$\Leftrightarrow [275 - 2.0639(5/5), 275 + 2.0639(5/5)]$$

$$\Leftrightarrow [272.9361, 277.0639]$$

Note that the smaller the sample size, the wider the interval width.

5.2 Testing hypothesis for a population mean

- **Examples of testing hypothesis for a population mean.**

  - Capacity of a cookie bag is indicated as 200g. Will there be enough cookies in the indicated capacity?
  - At a light bulb factory, a newly developed light bulb advertises a longer bulb life than the past. Is this propaganda reliable?
  - In this year's academic test, students said that there will be an average English score of 5 points higher than last year. How can you investigate if this is true?

5.2 Testing hypothesis for a population mean

[Example 5.2.1] At a light bulb factory, the average life expectancy of a light bulb made by a conventional method is known to be 1,500 hours, and the standard deviation is 200 hours. The company introduced a new production method with an average life expectancy of 1,600 hours. To confirm this argument, 30 samples were taken, and the sample mean was 1555 hours. Does the new type of light bulb have a life of 1600 hours?

<Answer>

- Make two assumptions about the different arguments for the population mean μ.

  $H_0 : \mu = 1500, \quad H_1 : \mu = 1600$
- H_0 is a null hypothesis and H_1 is an alternative hypothesis

5.2 Testing hypothesis for a population mean

- **Unless there is a significant reason, keep the null hypothesis**

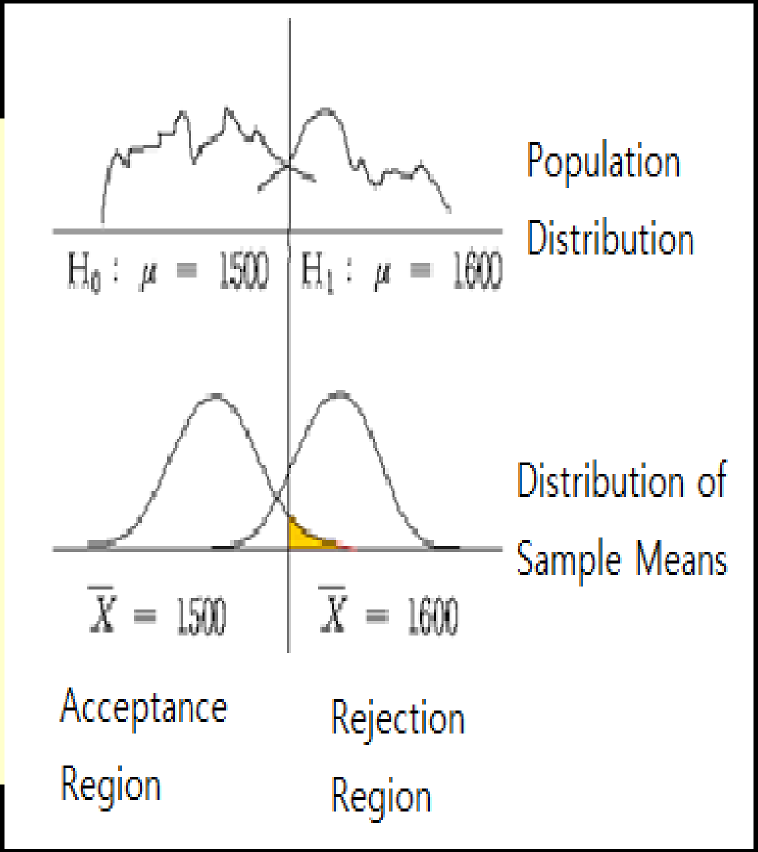
⇨ 'conservative decision making'

- **Testing hypothesis is based on sampling distribution of $\bar{X}$.**

⇨ Select a critical value C based on sampling distribution

⇨ Decision rule:

'If $\bar{X}$ < C, then accept H_0, else reject H_0'



5.2 Testing hypothesis for a population mean

Table 7.1.1 Two types of errors in testing hypothesis

		Actual	
		H_0 is true	H_1 is true
Decision:	H_0 is true	Correct	Type 2 Error
	H_1 is true	Type 1 Error	Correct

- If we set the significance level is 5%, C can be calculated by finding the percentile of $N(1500, \frac{200^2}{30})$

  C = 1500 + 1.645 $\sqrt{\frac{200^2}{30}}$ = 1560.06
- Decision rule: 'If $\overline{X}$ < 1560.06, then accept H_0, else reject H_0'
- Since $\overline{X}$ = 1555, it is less than 1560.06, accept H_0.

5.2 Testing hypothesis for a population mean

- p-value is the probability of a type 1 error when the observed sample mean value is considered as the critical value for decision
  ⇨ p-value indicates where the observed sample mean is located among all possible sample means

5.2 Testing hypothesis for a population mean

Table 5.2.2 Testing hypothesis for a population mean - unknown σ case	
Type of Hypothesis	Decision Rule
1) $H_0 : \mu = \mu_0$ $H_1 : \mu > \mu_0$	If $\frac{\overline{X}-\mu_0}{\frac{S}{\sqrt{n}}} > t_{n-1:\, \alpha}$, then reject H_0
2) $H_0 : \mu = \mu_0$ $H_1 : \mu < \mu_0$	If $\frac{\overline{X}-\mu_0}{\frac{S}{\sqrt{n}}} < -t_{n-1:\, \alpha}$, then reject H_0
3) $H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	If $\left\lvert \frac{\overline{X}-\mu_0}{\frac{S}{\sqrt{n}}} \right\rvert > t_{n-1;\, \alpha/2}$, then reject H_0

Note: Assume that the population is a normal distribution.
The H_0 of 1) can be written as $H_0 : \mu \leq \mu_0$, 2) as $H_0 : \mu \geq \mu_0$

5.2 Testing hypothesis for a population mean

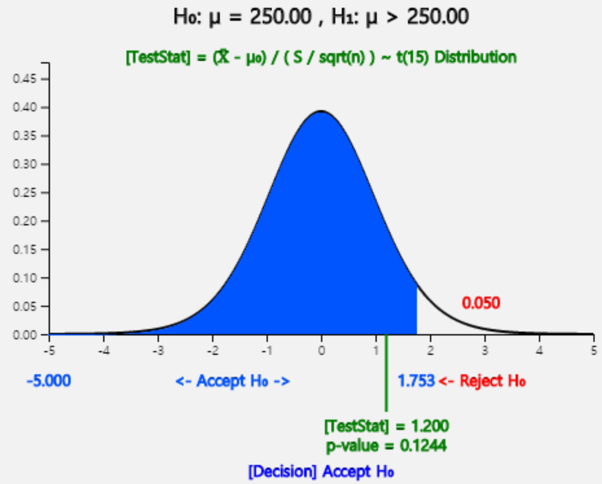
Example 5.2.2 The weight of a bag of cookies is supposed to be 250 grams. Suppose the weight of all bags of cookies follows a normal distribution. In the survey of 16 random samples of bags, the sample mean was 253 grams, and the sample standard deviation was 10 grams. Test the hypothesis whether the weight of the bag of cookies is 250g or larger using α = 1% and find the p-value. Use 「eStatU」 to test the hypothesis above.

Answer

Since the population standard deviation is unknown and the sample

$$\text{'If } \frac{\overline{X} - \mu_0}{\frac{S}{\sqrt{n}}} > t_{n-1:\,\alpha}, \text{ then reject } H_0 \text{ else accept } H_0'$$

$$\text{'If } \frac{253 - 250}{\frac{10}{\sqrt{16}}} > t_{16:\,0.01}, \text{ then reject } H_0 \text{ else accept } H_0'$$



Since the value of test statistic is $\frac{253-250}{\frac{10}{\sqrt{16}}} = 1.2$, and $t_{15:\,0.01} = 2.602$, we accept H_0. Note that the decision rule can be written as follows.

$$\text{'If } \overline{X} > 250 + 2.602\frac{10}{\sqrt{16}}, \text{ then reject } H_0 \text{ else accept } H_0'$$

5.3 Testing hypothesis for two population means

❖ Test statistic and sampling distribution

$$\frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} \quad \text{where } s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$\sim t_{n_1+n_2-2}$$

Table 5.3.1 Testing hypothesis of two populations means

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

5.3 Testing hypothesis for two population means

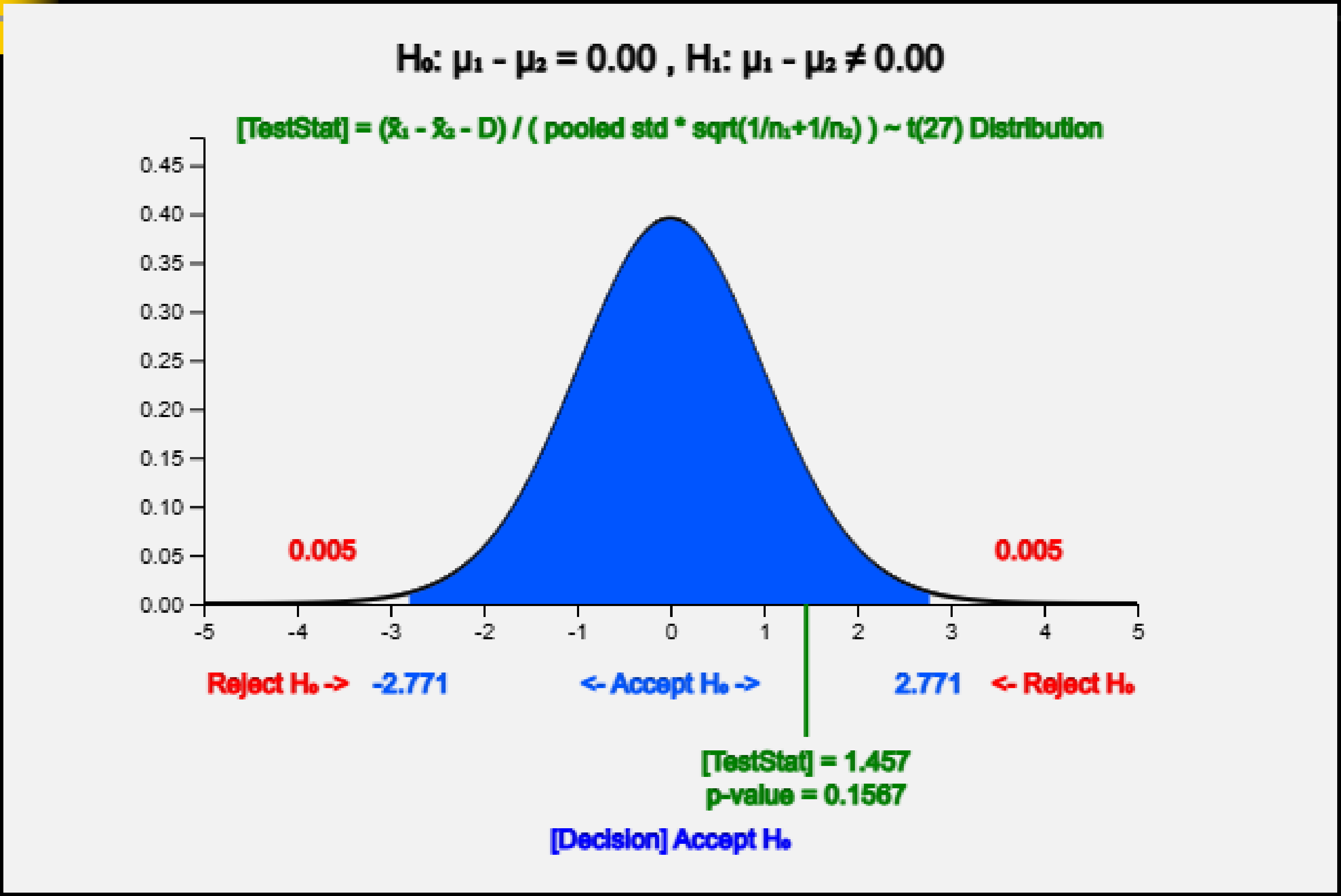
Example 5.3.2 (Monthly wages by male and female)
Random samples of 10 male and female college graduates this year showed their monthly wages as follows. (Unit 10,000 KRW)

Male 272 255 278 282 296 312 356 296 302 312
Female 276 280 369 285 303 317 290 250 313 307
Ex ⇨ DataScience ⇨ WageByGender.csv.

Using 「eStat」, answer the following questions.

1) If population variances are assumed to be the same, test the hypothesis at the 5% significance level of whether the average monthly wage for males and females is the same.
2) If population variances are assumed to be different, test the hypothesis at the 5% significance level of whether the average monthly wage for males and females is the same.

5.3 Testing hypothesis for two population means



5.3 Testing hypothesis for two population means

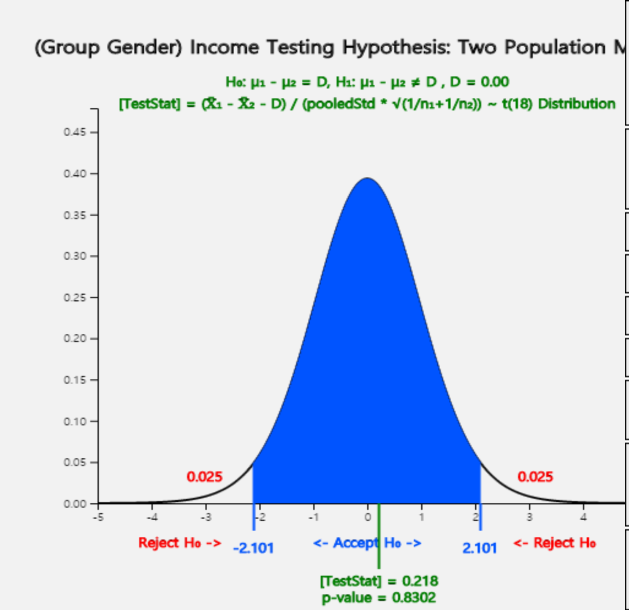
File: EX080103_WageByGender.csv

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		



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	$\mu_1-\mu_2$ 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)

5.4 Testing hypothesis for several population means: ANOVA

- **Examples to compare means of several populations.**
  - **Are average hours of library usage for each grade the same?**
  - **Are yields of three different rice seeds equal?**
  - **In a chemical reaction, are response rates the same at four different temperatures?**
  - **Are average monthly wages of college graduates the same in three different cities?**

- **A factor is a variable that distinguishes populations, such as grade or rice.**

5.4 Testing hypothesis for several population means: ANOVA

[Example 5.4.1] To compare the English proficiency of each grade at a university, samples were randomly selected from each grade to take the same English test.

Grade	English Proficiency Score	Average
1	81 75 69 90 72 83	$\bar{y}_{1\cdot}$ = 78.3
2	65 80 73 79 81 69	$\bar{y}_{2\cdot}$ = 74.5
3	72 67 62 76 80	$\bar{y}_{3\cdot}$ = 71.4
4	89 94 79 88	$\bar{y}_{4\cdot}$ = 87.5

1) Using『eStat』, draw a dot graph of exam scores for each grade and compare average.
2) We want to test a hypothesis whether the average scores of each grade are the same or not. Write a null hypothesis and an alternative hypothesis.
3) Apply the analysis of variances to test the hypothesis in question 2).
4) Use『eStat』to check the results of the ANOVA test.

5.4 Testing hypothesis for several population means: ANOVA

<Answer of Example 5.4.1>

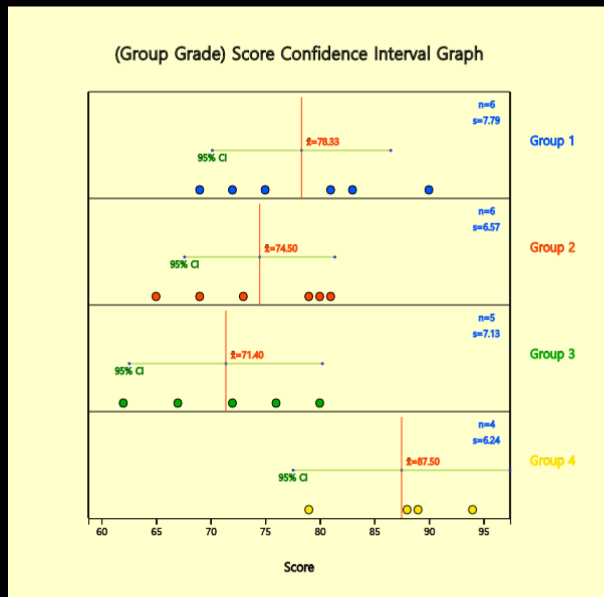
File: Ex911EnglishScoreByGrade.csv

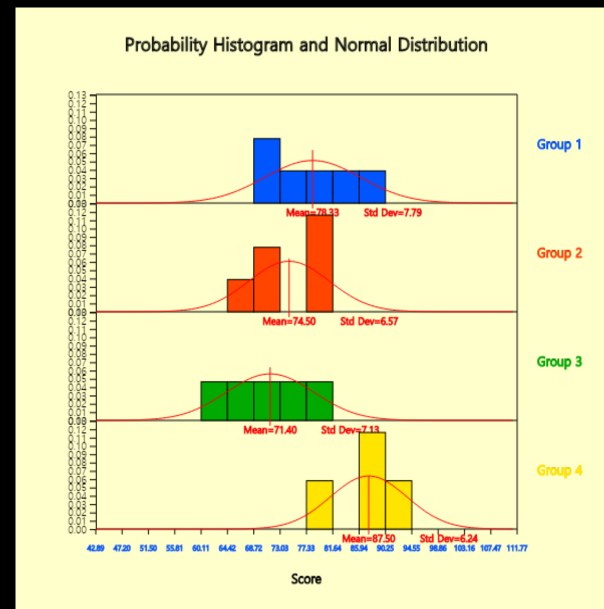
Analysis Var: 2: Score by Group: 1: Grade

(Selected data: Raw Data) (Select up to two groups)

SelectedVar: V2 by V1,

	Grade	Score	V3	V4
1	1	81		
2	1	75		
3	1	69		
4	1	90		
5	1	72		
6	1	83		
7	2	65		
8	2	80		
9	2	73		
10	2	79		
11	2	81		
12	2	69		
13	3	72		
14	3	67		
15	3	62		
16	3	76		
17	3	80		
18	4	89		
19	4	94		
20	4	79		
21	4	88		





Confidence Interval Graph | Histogram

$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ H_1 : At least one pair of means is different

Significance Level α = 5% 1% Confidence Level 95% 99%

ANOVA F test | Standardized Residual Plot | Kruskal-Wallis Test

5.4 Testing hypothesis for several population means: ANOVA

<Answer of Example 5.4.1>

2) Null hypothesis $H_o: \mu_1 = \mu_2 = \mu_3 = \mu_4$

Alternative hpothesis H_1: *at least one pair of* μ_i *is not the same*

3) Between sum of squares (SSB) or Treatment sum of squares (SSTr)

$$SSTr = 6(78.3 - \bar{y}_{..})^2 + 6(74.5 - \bar{y}_{..})^2 + 5(71.4 - \bar{y}_{..})^2 + 4(87.5 - \bar{y}_{..})^2 = 643.633$$

⇨ If SSTr is close to zero, all sample means for four grades are similar.

Within sum of squares (SSW) or Error sum of squares (SSE)

$$\begin{aligned} SSE = & (81 - \bar{y}_{1.})^2 + (75 - \bar{y}_{1.})^2 + \cdots + (83 - \bar{y}_{1.})^2 \\ & + (65 - \bar{y}_{2.})^2 + (80 - \bar{y}_{2.})^2 + \cdots + (69 - \bar{y}_{2.})^2 \\ & + (72 - \bar{y}_{3.})^2 + (67 - \bar{y}_{3.})^2 + \cdots + (80 - \bar{y}_{3.})^2 \\ & + (89 - \bar{y}_{4.})^2 + (94 - \bar{y}_{4.})^2 + \cdots + (88 - \bar{y}_{4.})^2 = 839.033 \end{aligned}$$

5.4 Testing hypothesis for several population means: ANOVA

<Answer of Example 9.1.1>

$$F_0 = \frac{\frac{SSTr}{(4-1)}}{\frac{SSE}{(21-4)}} = \frac{Treatment\ \ Mean\ Square\ (MSTr)}{Error\ Mean\ Square\ (MSE)} \quad \sim \quad F_{3,17}$$

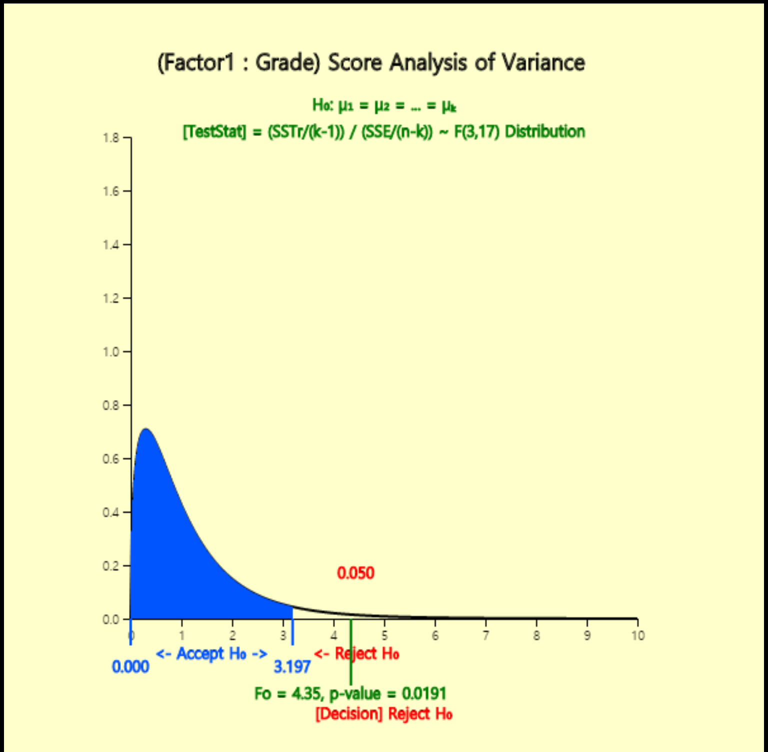
$$F_0 = \frac{\frac{643.633}{3}}{\frac{839.033}{17}} = 4.347 \qquad F_{3,17;0.05} = 3.20$$

- **Hence Reject** $H_o : \mu_1 = \mu_2 = \mu_3 = \mu_4$
- **ANOVA Table**

Factor	Sum of Squares	Degree of freedom	Mean Squares	F value
Treatment Error	SSTr= 643.633 SSE = 839.033	4-1 21-4	MSTr = 643.633/3 MSE = 839.033/17	Fo = 4.347
Total	SST =1482.666	20		

5.4 Testing hypothesis for several population means: ANOVA

<Answer of Example 5.4.1>



Statistics	Analysis Var	Score	Group Name	Grade		
Group Variable (Grade)	Observation	Mean	Std Dev	std err	Population Mean 95% Confidence Interval	Population Variance 95% Confidence Interval
1 (Group 1)	6	78.333	7.789	3.180	(70.159, 86.507)	(23.638, 364.929)
2 (Group 2)	6	74.500	6.565	2.680	(67.610, 81.390)	(16.793, 259.260)
3 (Group 3)	5	71.400	7.127	3.187	(62.550, 80.250)	(18.235, 419.472)
4 (Group 4)	4	87.500	6.245	3.122	(77.563, 97.437)	(12.516, 542.181)
Total	21	77.333	8.610	1.879	(73.414, 81.253)	(43.391, 154.593)
Missing Observations	0					

Analysis of Variance

Factor	Sum of Squares	deg of freedom	Mean Squares	F value	p value
Treatment	643.633	3	214.544	4.347	0.0191
Error	839.033	17	49.355		
Total	1482.667	20			

5.4 Testing hypothesis for several population means: ANOVA

<Answer of Example 5.4.1>

Testing Hypothesis ANOVA

Menu

[Hypothesis] $H_0: \mu_1 = \mu_2 = \dots = \mu_k$

H_1 : *At least one pair of means is different*

[Test Type] *F test* (ANOVA)

Significance Level $\alpha =$ (●) 5% ( ) 1%

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

Sample 1	81 75 69 90 72 83
Sample 2	65 80 73 79 81 69
Sample 3	72 67 62 76 80
Sample 4	89 94 79 88

[Sample Statistics]

$n_1 =$	6	$n_2 =$	6	$n_3 =$	5	$n_4 =$	4
$\bar{x}_1 =$	78.33	$\bar{x}_2 =$	74.50	$\bar{x}_3 =$	71.40	$\bar{x}_4 =$	87.50
$s_1^2 =$	60.67	$s_2^2 =$	43.10	$s_3^2 =$	50.80	$s_4^2 =$	39.00

Execute

5.4 Testing hypothesis for several population means: ANOVA

Table 9.1.2 Notation of one-way ANOVA

Factor	Observed values of sample				Average
Level 1	Y_{11}	Y_{12}	$\cdots$	Y_{1n_1}	$\overline{Y}_{1\cdot}$
Level 2	Y_{21}	Y_{22}	$\cdots$	Y_{2n_2}	$\overline{Y}_{2\cdot}$
Level k	Y_{k1}	Y_{k2}	$\cdots$	Y_{kn_k}	$\overline{Y}_{k\cdot}$

- **ANOVA Model**

$$\begin{aligned} Y_{ij} &= \mu_i + \varepsilon_{ij} \\ &= \mu + \alpha_i + \varepsilon_{ij}, \quad i = 1, 2, \ldots, k;\ j = 1, 2, \ldots, n_i \end{aligned}$$

- **Hypothesis**

$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$
H_1 : **At least one pair of α_i is not equal to 0**

5.4 Testing hypothesis for several population means: ANOVA

Table 9.1.3 Analysis of variance table of one-way ANOVA

Factor	Sum of Squares	Degree of freedom	Mean Squares	F value
Treatment	SSTr	$k-1$	MSTr = SSTr / $(k-1)$	F_0 = MSTr/MSE
Error	SSE	$n-k$	MSE = SSE / $(n-k)$	
Total	SST	$n-1$	$\left(n=\sum_{i=1}^{k} n_i\right)$	

- **Total Sum of Squares** **SST** $= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{..})^2$
- **Treatment Sum of Squares** **SSTr** $= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{Y}_{i.} - \bar{Y}_{..})^2$
- **Error Sum of Squares** **SSE** $= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i.})^2$
- **SST = SSTr + SSE**
- **If** $F_0 > F_{k-1,n-k\,;\,\alpha}$ **, then reject** H_0

5.5 Regression analysis

❖Correlation analysis

- Population with (μ_X, μ_Y) and (σ_X^2, σ_Y^2)
- Random Sample $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$
- Population Covariance $\sigma_{XY} = Cov(X,Y) = E(X_i - \mu_X)(Y_i - \mu_Y)$
- Sample Covariance $S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$
- Population Correlation $\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$
- Sample Correlation $r = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$

5.5 Regression analysis

❖ Correlation analysis

- **Characteristics of ρ**

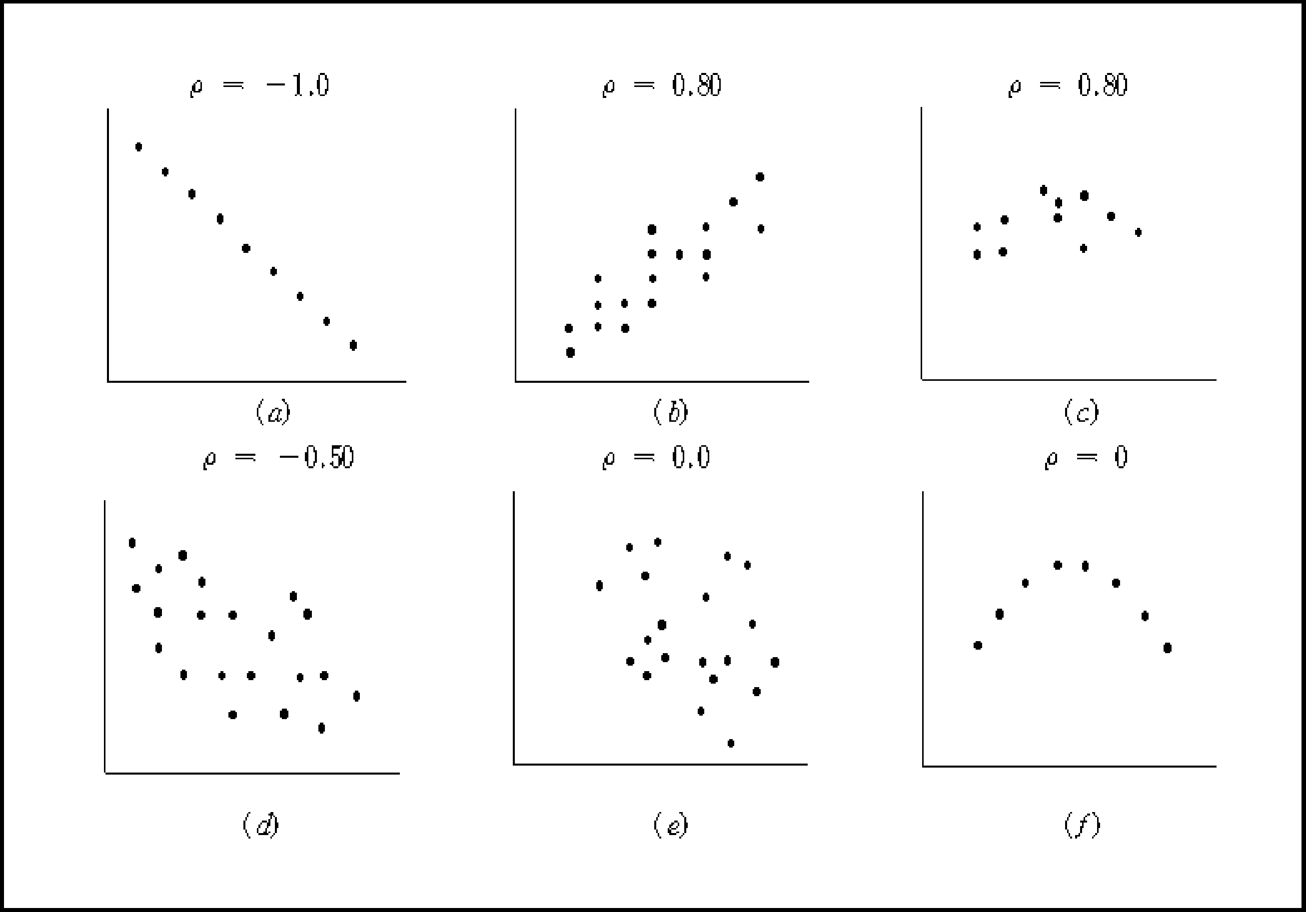
1) ρ has a value between -1 and +1.
   - closer to +1 ⇨ strong positive linear relation
   - closer to -1 ⇨ strong negative linear relation.
   - closer to 0 ⇨ weak linear relation

2) If all values of X and Y are located on a straight line, ρ is either +1 or -1.

3) ρ is only a measure of linear relationship between two variables.
   - if ρ =0, there is no linear relationship between the two variables, but there may be a different relationship

5.5 Regression analysis

❖Correlation analysis



5.5 Regression analysis

❖Correlation analysis

❑ **Testing population correlation coefficient ρ**

Null Hypothesis: $H_0 : \rho = 0$

Test Statistic: $t_0 = \sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}} \sim t_{n-2}$

Rejection Region of H_0 :

1) $H_1 : \rho < 0$ Reject H_0 if $t_0 < -t_{n-2;\,\alpha}$
2) $H_1 : \rho > 0$ Reject H_0 if $t_0 > t_{n-2;\,\alpha}$
3) $H_1 : \rho \neq 0$ Reject H_0 if $|t_0| > t_{n-2;\,\alpha/2}$

5.5 Regression analysis

❖Correlation analysis

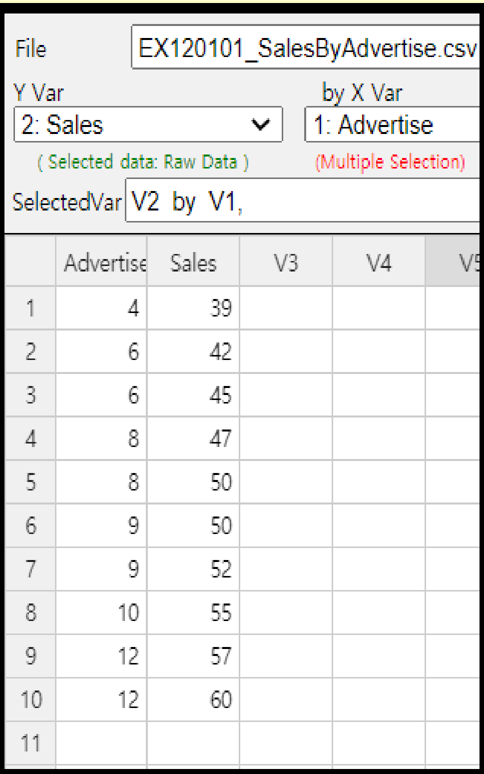
[Example 5.5.1] Based on the survey of advertising costs and sales for 10 companies that make the same product, we obtained the following data. Test the hypothesis that the population correlation coefficient is zero with the significance level 0.05.

Company	1	2	3	4	5	6	7	8	9	10
Advertise (X)	4	6	6	8	8	9	9	10	12	12
Sales (Y)	39	42	45	47	50	50	52	55	57	60

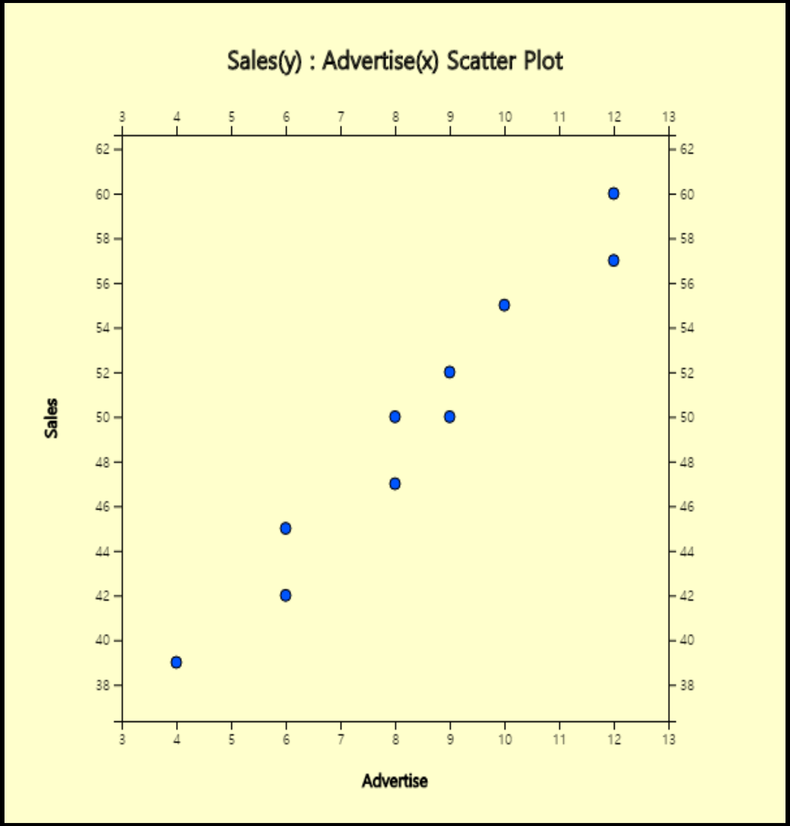
5.5 Regression analysis

❖Correlation analysis

<Answer of Example 5.5.1>


File: EX120101_SalesByAdvertise.csv
Y Var: 2: Sales
by X Var: 1: Advertise
(Selected data: Raw Data)
(Multiple Selection)
SelectedVar: V2 by V1,

	Advertise	Sales	V3	V4	V5
1	4	39			
2	6	42			
3	6	45			
4	8	47			
5	8	50			
6	9	50			
7	9	52			
8	10	55			
9	12	57			
10	12	60			
11					



5.5 Regression analysis

❖ Correlation analysis

<Answer of Example 5.5.1>

$$\mathbf{SXX} = \sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n} X_i^{\,2} - n\,\bar{X}^2 = 766 - 10\times 8.4^2 = 60.4$$

$$\mathbf{SYY} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n} Y_i^{\,2} - n\,\bar{Y}^2 = 25097 - 10\times 49.7^2 = 396.1$$

$$\mathbf{SXY} = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{n} X_iY_i - n\bar{X}\bar{Y} = 4326 - 10\times 8.4\times 49.7 = 151.2$$

id	X	Y	X^2	Y^2	XY
1	4	39	16	1521	156
2	6	42	36	1764	252
3	6	45	36	2025	270
4	8	47	64	2209	376
5	8	50	64	2500	400
6	9	50	81	2500	450
7	9	52	81	2704	468
8	10	55	100	3025	550
9	12	57	144	3249	684
10	12	60	144	3600	720
Sum	84	497	766	25097	4326
Mean	8.4	49.7			

$$S_{XY} = \frac{1}{n-1}\mathrm{SXY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \frac{151.2}{10-1}$$

$$r = \frac{S_{XY}}{S_X\,S_Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{SXY}{\sqrt{SXX\ \ SXY}} = \frac{151.2}{\sqrt{60.4\ \times 396.1}} = \mathbf{0.978}$$

5.5 Regression analysis

❖Correlation analysis

<Answer of Example 5.5.1>

$$t_0 = \sqrt{n-2}\ \frac{r}{\sqrt{1-r^2}} = \sqrt{10-2}\ \frac{0.978}{\sqrt{1-0.978^2}} = 13.26$$

$$t_{10-2;0.025} = 2.306$$

Hence $H_0 : \rho = 0$ is rejected

Regression Analysis				
Regression	y =	28.672 +	2.503 x	
Correlation Coefficient	r = 0.978	H_0: ρ = 0 H_1: ρ ≠ 0	t value = 13.117	p value < 0.0001
Coefficient of Determination	r^2 = 0.956			
Standard Error	s = 1.483			

5.5 Regression analysis

❖ Simple regression analysis

- **Regression analysis** is a statistical method
  - a mathematical model of relationships between variables,
  - estimates model using measured values of the variables,
  - uses estimated model to describe the relationship between variables, or to apply it to the analysis such as forecasting.
- Mathematical model ⇨ **regression equation**
- A variable affected by other variables is called a **dependent variable**. ⇨ **response variable**
- Variables that affect dependent variable are called **independent variables**. ⇨ **explanatory variable**

5.5 Regression analysis

❖Simple regression analysis

- **Population regression model** $Y_i = \alpha + \beta X_i + \epsilon_i \ , i = 1, 2, \dots, n$

  Estimated regression equation $\hat{Y}_i = a + b X_i$

  Residuals $e_i = Y_i - \hat{Y}_i$

- **Method of Least Squares**

  A method of estimating regression coefficients so that total sum of the squared errors occurring in each observation is minimized.

  Find α and β which minimize $\sum_{i=1}^{n} {\epsilon_i}^2 = \sum_{i=1}^{n} (Y_i - \alpha - \beta X_i)^2$

- **Least square estimator of α and β**

$$b = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$$

$$a = \overline{Y} - b\,\overline{X}$$

5.5 Regression analysis

❖Simple regression analysis

❑ **Goodness of Fit for Regression Line**

- **Residual standard error s is a measure of the extent to which observations are scattered around the estimated line.**

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$$

The residual standard error s is defined as the square root of s^2 .

- **SST = SSE + SSR**

$$SST = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 \qquad df \quad n-1$$

$$SSE = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 \qquad df \quad n-2$$

$$R^2 = \frac{SSR}{SST}$$

$$SSR = \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2 \qquad df \quad 1$$

5.5 Regression analysis

❖Simple regression analysis

❑ Inference for the parameter β

- Point estimate: $b = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \sim N(\beta, \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2})$
- Standard error of estimate b: $SE(b) = \frac{s}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$
- Confidence interval of β: $b \pm t_{n-2;\,\alpha/2} \times SE(b)$
- Testing hypothesis:

  Null hypothesis: $H_0 :\ \beta = \beta_0$

  Test statistic: $t = \frac{b - \beta_0}{SE(b)}$

  1) $H_1 : \beta < \beta_0$ Reject H_0 if t < - $t_{n-2;\ \alpha}$
  2) $H_1 : \beta > \beta_0$ Reject H_0 if t > $t_{n-2;\ \alpha}$
  3) $H_1 : \beta \neq \beta_0$ Reject H_0 if | t | > $t_{n-2;\ \alpha/2}$

5.5 Regression analysis

❖ Simple regression analysis

❑ Inference for the parameter α

- Point estimate: $a = \overline{Y} - b\,\overline{X} \quad \sim N\left(\alpha,\ \left(\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right)\sigma^2\right)$
- Standard error of estimate a: $SE(a) = s\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}}$
- Confidence interval of β: $a \pm t_{n-2;\,\alpha/2} \times SE(a)$
- Testing hypothesis:

  Null hypothesis: $H_0 : \alpha = \alpha_0$

  Test statistic: $t = \frac{a - \alpha_0}{SE(a)}$

  1) $H_1 : \alpha < \alpha_0$ Reject H_0 if t < - $t_{n-2;\ \alpha}$
  2) $H_1 : \alpha > \alpha_0$ Reject H_0 if t > $t_{n-2;\ \alpha}$
  3) $H_1 : \alpha \neq \alpha_0$ Reject H_0 if | t | > $t_{n-2;\ \alpha/2}$

5.5 Regression analysis

❖ Simple regression analysis

❑ **Inference for the average value** $\mu_{Y|x} = \alpha + \beta X_0$

- **Point estimate:** $\hat{Y}_0 = a + bX_0$
- **Standard error of estimate $\hat{Y}_0$:** $SE(\hat{Y}_0) = s\sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$
- **Confidence interval of $\mu_{Y|x}$:** $\hat{Y}_0 \pm t_{n-2;\,\alpha/2} \times SE(\hat{Y}_0)$

Table 5.5.2 Analysis of variance table for simple linear regression

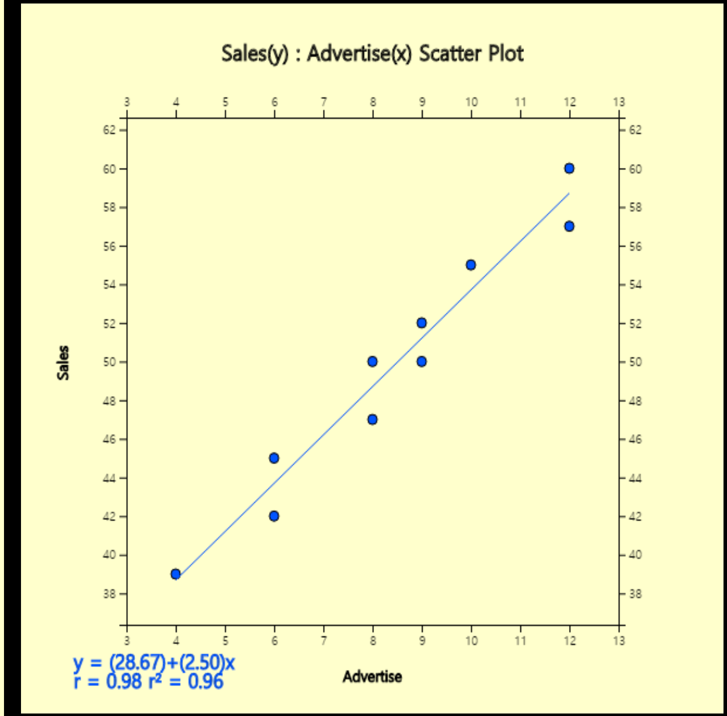
Source	Sum of squares	Degrees of freedom	Mean Squares	F value
Regression	SSR	1	MSR = $\frac{SSR}{1}$	$F_0 = \frac{MSR}{MSE}$
Error	SSE	$n-2$	MSE = $\frac{SSE}{n-2}$	
Total	SST	$n-1$		

5.5 Regression analysis

❖Simple regression analysis

[Example 5.5.2] In [Example 5.5.1], find the least squares estimate of the slope and intercept if the sales amount is a dependent variable and the advertising cost is an independent variable.

- **Predict amount of sales when you have spent on advertising by 10.**
- **Calculate the value of the residual standard error and the coefficient of determination in the data on advertising costs and sales**
- **Prepare an ANOVA table and test it using the 5% significance level**



5.5 Regression analysis

❖ Simple regression analysis

<Answer of Example 5.5.2>

$$b = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{151.2}{60.4} = 2.503$$

$$a = \bar{Y} - b\,\bar{X} = 49.7 - 2.503 \times 8.4 = 28.672$$

- **Forecasting $\hat{Y}_i = 28.672 + 2.503\ X_i$**

  $28.671 + 2.503 \times 10 = 53.705$

- $$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

  $$= \frac{17.622}{(10-2)} = 2.203$$

- $$R^2 = \frac{SSR}{SST} = \frac{378.429}{396.1} = 0.956$$

5.5 Regression analysis

❖ Simple regression analysis

	X_i	Y_i	$\hat{Y}_i$	SST $\sum(Y_i-\bar{Y})^2$	SSR $\sum(\hat{Y}_i-\bar{Y})^2$	SSE $\sum(Y_i-\hat{Y}_i)^2$
1	4	39	38.639	114.49	122.346	0.130
2	6	42	43.645	59.29	36.663	2.706
3	6	45	43.645	22.09	36.663	1.836
4	8	47	48.651	7.29	1.100	2.726
5	8	50	48.651	0.09	1.100	1.820
6	9	50	51.154	0.09	2.114	1.332
7	9	52	51.154	5.29	2.114	0.716
8	10	55	53.657	28.09	15.658	1.804
9	12	57	58.663	53.29	80.335	2.766
10	12	60	58.663	106.09	80.335	1.788
Sum	84	497	496.522	396.1	378.429	17.622
Average	8.4	49.7				

[ANOVA]

Factor	Sum of Squares	deg of freedom	Mean Squares	F value	p value
Regression	378.501	1	378.501	172.052	< 0.0001
Error	17.599	8	2.200		
Total	396.100	9			

5.5 Regression analysis

❖ Simple regression analysis

1) Inference for β

- $b = 2.50333$

$$SE(b) = \frac{s}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}} = \frac{1.484}{60.4} = 0.1908$$

- Confidence interval of β: $b \pm t_{n-2;\,\alpha/2} \times SE(b)$

$$2.5033 \pm 3.833 \times 0.1908 \quad \Leftrightarrow \quad (1.7720, 3.2346)$$

- Test statistic for $H_0 : \beta = 0 \quad H_1 : \beta \neq 0$

Reject H_0 if $|t| > t_{n-2;\,\alpha/2}$

$$t = \frac{b - \beta_0}{SE(b)} = \frac{2.5033 - 0}{0.1908} = 13.22$$

Since $t_{8;\,0.025} = 3.833$, H_0 is rejected.

5.5 Regression analysis

❖Simple regression analysis

2) Inference for α

- $a = 29.672$

$$SE(a) = s\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}} = 1.484\sqrt{\frac{1}{10} + \frac{8.4^2}{60.4}} = 1.670$$

- Test statistic for $H_0 : \alpha = 0 \quad H_1 : \alpha \neq 0$

Reject H_0 if $|t| > t_{n-2;\ \alpha/2}$

$$t = t = \frac{a - \alpha_0}{SE(a)} = \frac{29.672 - 0}{1.670} = 17.1657$$

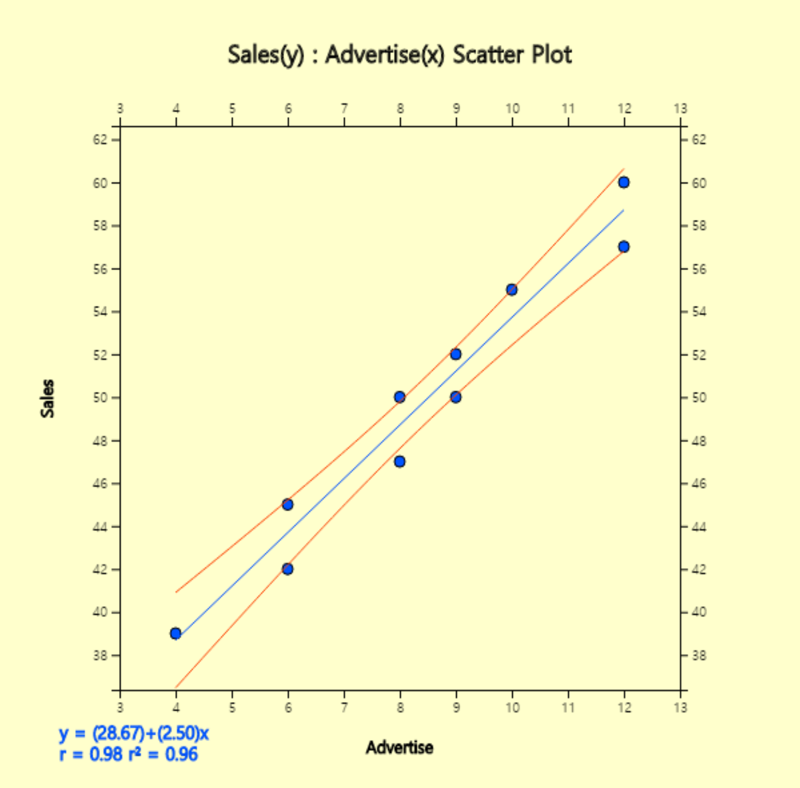
Since $t_{8;\ 0.025} = 3.833$, H_0 is rejected.

3) Confidence interval of $\mu_{Y|x}$: $\hat{Y}_0 \pm t_{n-2;\,\alpha/2} \times SE(\hat{Y}_0)$

if x= 8, $\hat{Y}_0 = 49.699$, ⇨ $49.699 \pm 3.833 \times 0.475$

5.5 Regression analysis

❖ Simple regression analysis

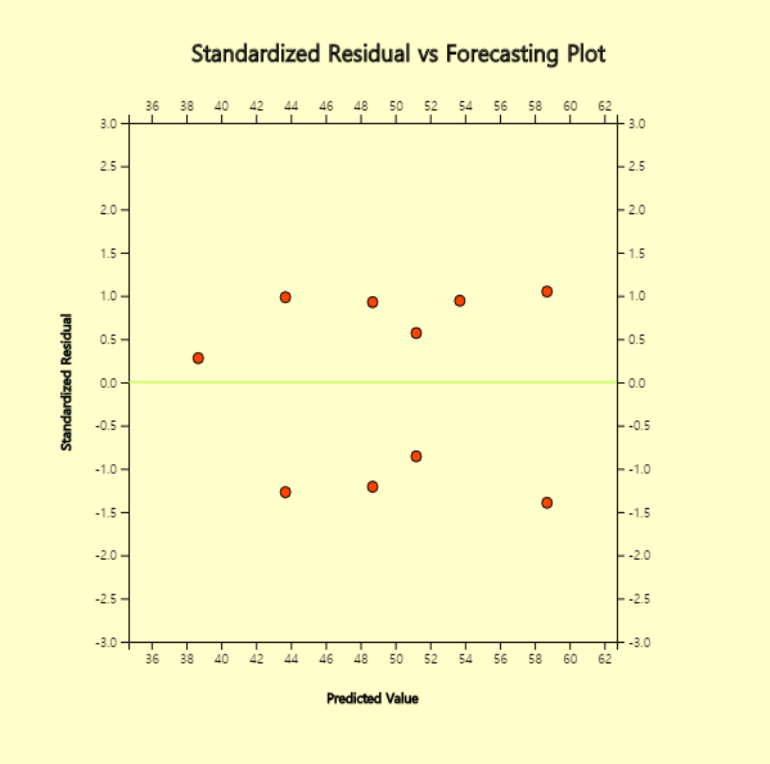


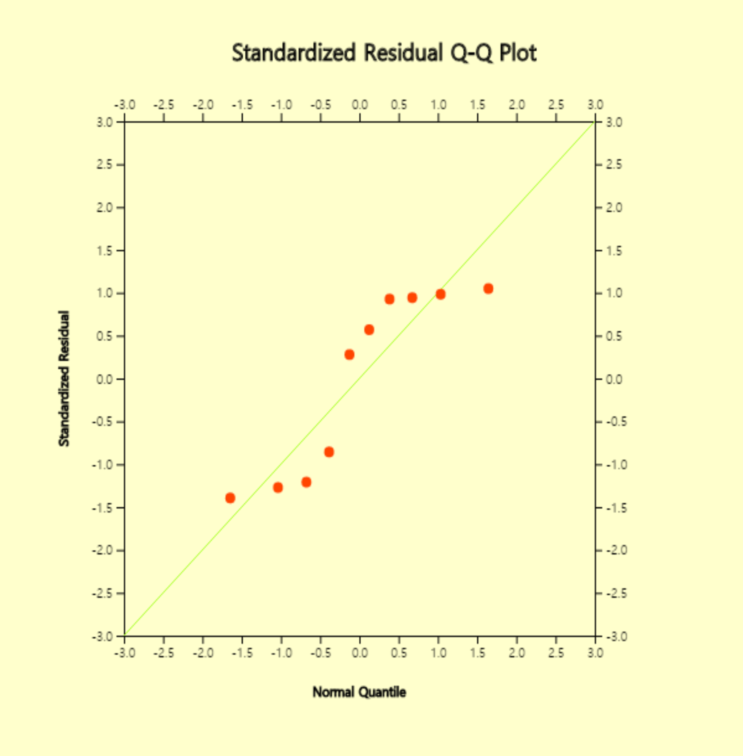
Parameter	Estimated Value	std err	t value	p value
Intercept	28.672	1.670	17.166	< 0.0001
Slope	2.503	0.191	13.117	< 0.0001

5.5 Regression analysis

❖Simple regression analysis

- **Residual analysis**





5.5 Regression analysis

❖ Multiple regression analysis

- **Population regression model**

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + +\beta_k X_k + \varepsilon_i \ , i = 1,2, \dots , n$$

$$\boldsymbol{Y} = \boldsymbol{X}\,\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \boldsymbol{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & X_{1k} \\ 1 & X_{21} & X_{22} & X_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & X_{nk} \end{bmatrix} \ \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \ \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

5.5 Regression analysis

❖ Multiple regression analysis

- **Least squares method**

  A method of estimating regression coefficients so that total sum of the squared errors occurring in each observation is minimized.

  Find α and β which minimize

  $$\sum_{i=1}^{n} \epsilon_i{}^2 = \varepsilon'\varepsilon = (Y - X\beta)'(Y - X\beta)$$

- **Least Square Estimator of α and β**

  $$b = (X'X)^{-1}(X'Y)$$

- **Residuals** $e_i = Y_i - \hat{Y}_i = Y_i - b_0 + b_1 X_{i1} + b_2 X_{i2} + \cdots + +b_k X_{ik}$

  Residual standard error s

  $$s = \sqrt{\frac{1}{n-k-1} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}$$

5.5 Regression analysis

❖ Multiple regression analysis

- **Analysis of Variance for Multiple Linear Regression**

Source	Sum of Squares	Degrees of Freedom	Mean Squares	F value
Regression Error	SSR SSE	k $n-k-1$	MSR=SSR / k MSE=SSE/ $(n-k-1)$	$F_0=\frac{MSR}{MSE}$
Total	SST	$n-1$		

- H_0 : $\beta_1 = \beta_2 = \cdots = \beta_k = 0$
  H_1 : At least one of k number of $\beta_i's$ is not equal to 0
- Reject H_0 if $F_0 > F_{k,n-k-1;\,\alpha}$

5.5 Regression analysis

❖ Multiple regression analysis

❑ Inference for the parameter β_i

- Point estimate: b_i
- Standard error of estimate b_i: $SE(b_i) = \sqrt{c_{ii}}\ s$
- Confidence interval of b_i: $b_i \pm t_{n-k-1;\,\alpha/2} \times SE(b_i)$
- Testing hypothesis:

  Null hypothesis: $H_0 : \beta_i = \beta_{i0}$

  Test statistic: $t = \frac{b_i - \beta_{i0}}{SE(b_i)}$

  1) $H_1 : \beta_i < \beta_{i0}$ Reject H_0 if $t <$ - $t_{n-k-1;\ \alpha}$
  2) $H_1 : \beta_i > \beta_{i0}$ Reject H_0 if $t >$ $t_{n-k-1;\ \alpha}$
  3) $H_1 : \beta_i \neq \beta_{i0}$ Reject H_0 if $|t| >$ $t_{n-k-1;\ \alpha/2}$

5.5 Regression analysis

❖ Multiple regression analysis

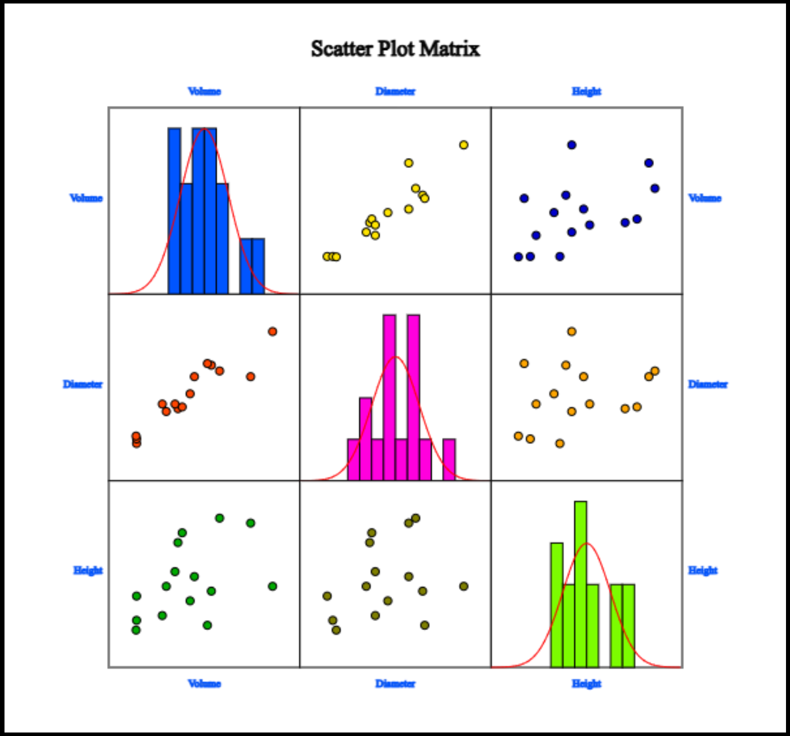
[Example 5.5.3] When logging trees in forest areas, it is necessary to investigate the amount of timber in those areas. Since it is difficult to measure the volume of a tree directly, we can think of ways to estimate the volume using the diameter and height of a tree that is relatively easy to measure. Draw a scatter plot matrix of this data and consider a regression model for this problem.

Diameter(cm)	Height(m)	Volume( )
21.0	21.33	0.291
21.8	19.81	0.291
22.3	19.20	0.288
26.6	21.94	0.464
27.1	24.68	0.532
27.4	25.29	0.557
27.9	20.11	0.441
27.9	22.86	0.515
29.7	21.03	0.603
32.7	22.55	0.628
32.7	25.90	0.956
33.7	26.21	0.775
34.7	21.64	0.727
35.0	19.50	0.704
40.6	21.94	1.084

5.5 Regression analysis

❖ Multiple regression analysis

<Answer of Example 5.5.3>



Regression Analysis					
Regression y =	(-1.024)	+ (0.037) X_1	+ (0.024) X_2		
Multiple Correlation Coeff	0.961	Coefficient of Determination	0.924	Standard Error	0.069

Parameter	Estimated Value	std err	t value	p value	95% Confidence Interval
β_0	-1.024	0.188	-5.458	0.0001	(-1.358 ,-0.689)
β_1 Diameter	0.037	0.003	10.590	< 0.0001	(0.031 ,0.043)
β_2 Height	0.024	0.008	2.844	0.0148	(0.009 ,0.038)

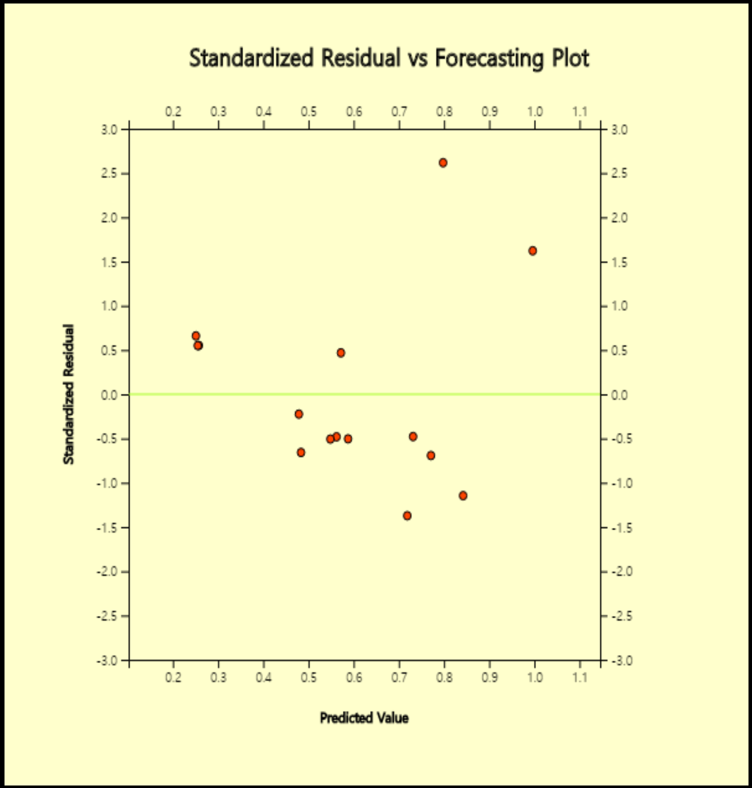
[ANOVA]

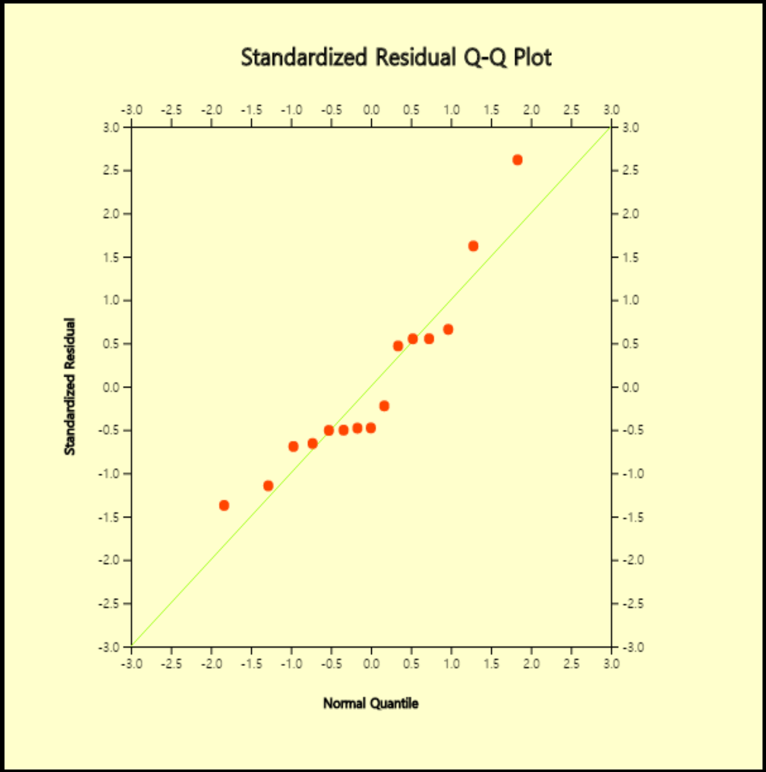
Factor	Sum of Squares	deg of freedom	Mean Squares	F value	p value
Regression	0.7058	2	0.3529	73.1191	< 0.0001
Error	0.0579	12	0.0048		
Total	0.7638	14			

5.5 Regression analysis

❖ Multiple regression analysis

[Example 5.5.4]





Summary

- Sampling distribution and estimation:
  - Central limit theorem
  - Point and interval estimation for a population mean
- Testing hypothesis for a population mean:
  - Type 1 and 2 error, significance level, p-value
  - Residual analysis
- Testing hypothesis for two population means
- Testing hypothesis for several population means
- Regression analysis:
  - Correlation analysis
  - Simple linear regression and multiple linear regression

Thank you !!!