Chapter 5. Estimation, testing hypothesis, and regression analysis

[presentation] (./pdf/ppt5.pdf) **[book] (./pdf/book5.pdf)**

CHAPTER OBJECTIVES

We introduce the following in this chapter.

- What is the sampling distribution of sample means is, and how can we we estimate a population mean using the sampling distribution in section 5.1?
- What a testing hypothesis is and the testing hypothesis for a single population mean in section 5.2.
- Testing hypothesis for comparing two population means in section 5.3.
- Testing hypothesis for comparing several populations means using analysis of variance in section 5.4.
- Correlation and regression analysis to analyze the relation between several continuous variables in section 5.5.

5.1 Sampling distribution and estimation

A population is usually very massive, and it is difficult and costly to investigate the entire population. Therefore, characteristic values of the population, such as a population mean and variance called **population parameters**, are usually estimated using a set of samples. Characteristic values of samples, such as a sample mean and sample variance called **sample statistic**. The distribution of all possible values of the sample statistic is called a **sampling distribution**. The sampling distribution identifies a relationship between the sample statistic and population parameter, making it possible to estimate and to test a population parameter. Section 5.1.1 discusses the sampling distribution of all possible sample means, and section 5.1.2 discusses how to estimate the population mean using the sampling distribution.

5.1.1 Sampling distribution of sample means

A population mean μ is called a **parameter** of a population, one of the characteristic values of the population. We collect samples of size n and calculate a sample mean to estimate the population mean. We hope this sample mean can estimate the population mean correctly, but there are many ways to collect samples of size n, and therefore, so many possible sample means. We hope this sample mean can estimate the population mean correctly, but there are many ways to collect samples of size n, and therefore, so many possible sample means. The distribution of all possible sample means is called a **sampling distribution of all possible sample means**. Since the sample mean is a random variable that can have many different values, it is usually denoted with a capital letter such as $\overline{X}$ and called an **estimator** of the population parameter μ. An observed sample mean, marked $\overline{x}$ with a lowercase letter, is called an **estimate** of μ.

If a population is a normal distribution $N(\mu, \sigma^2)$, the distribution of all possible sample means is exactly a normal distribution $N(\mu, \frac{\sigma^2}{n})$. If a population is not a normal distribution but the sample size is large enough, the distribution of all possible sample means is approximately a normal distribution such as $N(\mu, \frac{\sigma^2}{n})$. We call this the **central limit theorem**, which is a key theory underlying modern statistics. Theoretical proof of this theorem is beyond the scope of this book; please refer to any book on mathematical statistics.

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.

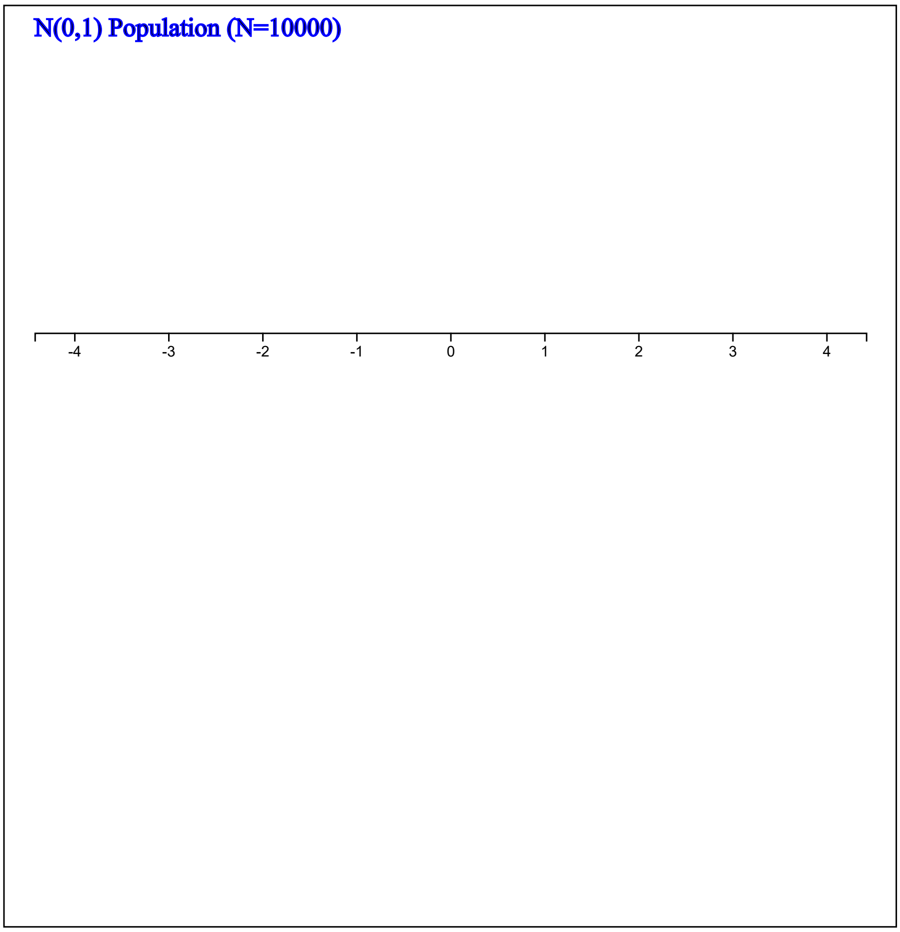
1) The average of all possible sample means, $\mu_{\overline{X}}$, is equal to the population mean μ.
(i.e., $\mu_{\overline{X}} = \mu$)
2) The variance of all possible sample means, $\sigma^2_{\overline{X}}$, is the population variance divided by n.
(i.e., $\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$)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

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

The following simulation using 『eStatU』 shows that when a population is a normal distribution, the distribution of all possible sample means is approximately normal, but variances become smaller as the sample size increases.

[Central Limit Theorem]

Dist of Sample Means

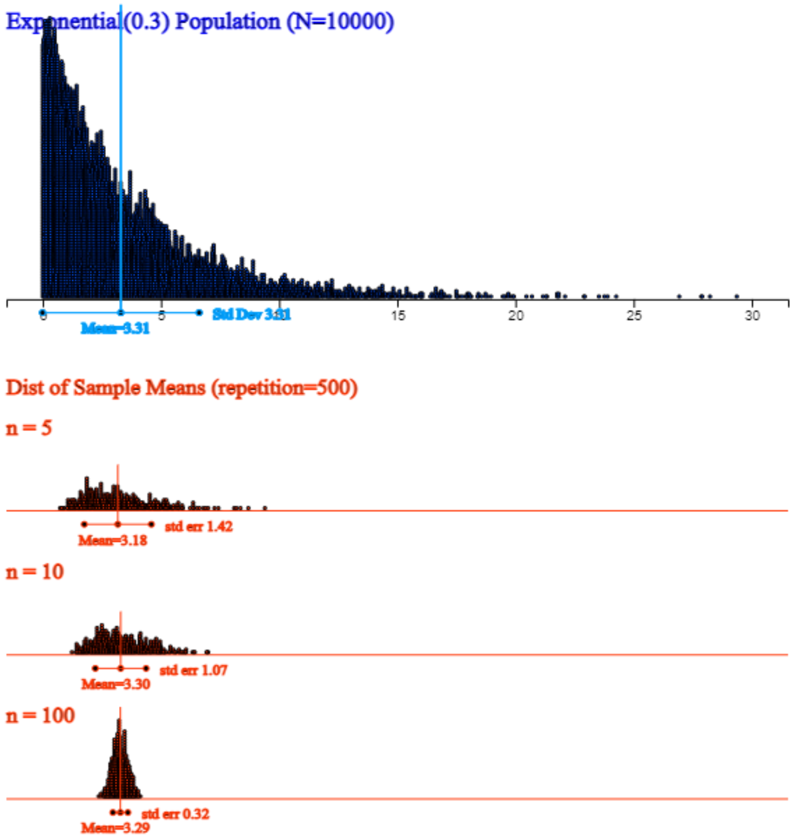


Population : ● N(0,1) ○ Exponential(0.3) ○ Uniform(0,1)

Execute Sample Size n1 = 5 n2 = 10 n3 = 20

Graph Save

<Figure 5.1.1> shows a simulation using 『eStatU』 in case a population is skewed from its mean. The distribution of all possible sample means is closer to normal as the sample size increases.



<Figure 5.1.1> 『eStatU』 Simulation of the central limit theorem

5.1.2 Estimation of a population mean

When a sample survey is conducted, only one set of samples, usually smaller than the population size, is selected from a population to estimate a characteristic value of the population, such as the population mean. We typically consider the sample mean of the selected samples to estimate the population mean. Can this sample mean can estimate the population mean well, even if the sample mean is only calculated from one set of small samples? This question is fundamental in estimating the population parameter that everyone can think about at least once. The sampling distribution of all possible sample means answers this question. Whatever the population distribution is, if the sample size is large enough, all possible sample means are distributed around the population mean in the form of a normal distribution by the central limit theorem. Therefore, the sample mean obtained from one set of samples is usually close to the population mean. Even in the worst case, the difference between the population mean and sample mean, known as an error, is not so significant, and it is possible to estimate the population mean using the sample mean. The larger the sample size, the more sample means are concentrated around the population mean based on the central limit theorem and hence, we can reduce the error of the estimation.

The value of an observed sample mean is called a **point estimate** of the population mean. In general, the sample statistic used to estimate a population parameter must have good characteristics to be accurate. The sample mean has all the good characteristics to estimate the population mean, and the sample variance also has all the good characteristics to estimate the population variance.

In contrast to the point estimate for a population mean, estimating the population mean using an interval is called an **interval estimation**. If a population follows a normal distribution with the mean μ and variance σ^2, the distribution of all possible sample means follows a normal distribution with the mean μ and variance $\frac{\sigma^2}{n}$, so the probability that one sample mean will be included in the interval $[\, \mu - z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}},\ \mu + z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}\,]$ is $1 - \alpha$ as follows.

$$P(\mu - z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} < \overline{X} < \mu + z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

We can rewrite this formula as follows.

$$P(\overline{X} - z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

Assuming σ is known, the meaning of the above formula is that 95% of intervals obtained by applying the formula $[\overline{X} - z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}]$ for all possible sample means include the population mean. The formula of this interval is referred to as the 100(1- α)% confidence interval of the population mean.

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

100(1-α)% here is called a **confidence level**, which refers to the probability of intervals that will include the population mean among all possible intervals calculated by the confidence interval formula. Usually, we use 0.01 or 0.05 for α. z_α is the upper α percentile of the standard normal distribution. In other words, if Z is the random variable that follows the standard normal distribution, the probability that Z is greater than z_α is α, i.e.,

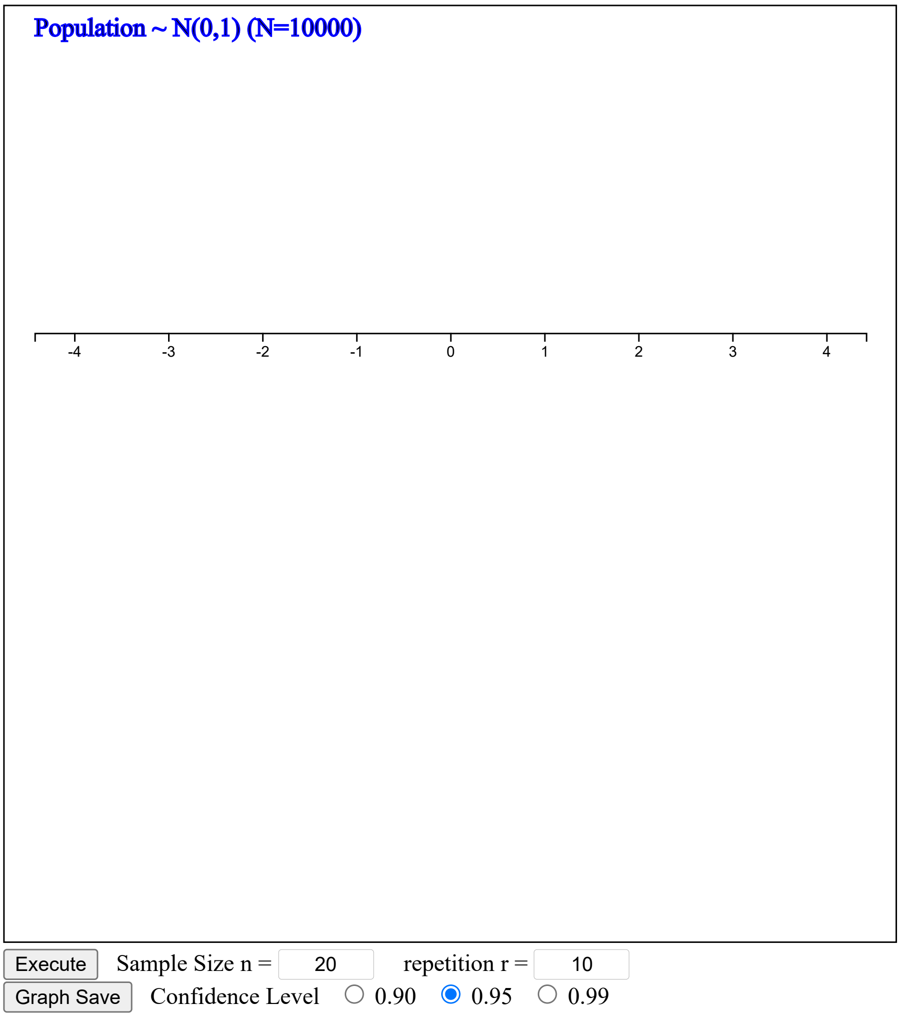
$$P(Z > z_\alpha) = \alpha$$

For example, $z_{0.025;}$ = 1.96, $z_{0.05}$ = 1.645, $z_{0.01}$ = 2.326, and $z_{0.005}$ = 2.575.

The following simulation shows the 95% confidence intervals for the population mean by extracting 100 sets of samples with the sample size n = 20 from a population of 10,000 numbers which follow the standard normal distribution N(0,1). In this case, 96 of the 100 confidence intervals contain the population mean 0. This result might be different on your computer because the program uses a random number generator, which depends on the computer. Whenever we repeat these experiments, the result may also vary slightly.

[Confidence Interval Simulation]

Confidence Interval Simulation



Sample Size n = 20 repetition r = 10

Confidence Level ○ 0.90 ● 0.95 ○ 0.99

Example 5.1.1 The average monthly starting salary of college graduates was 275 (unit: 10,000 KRW) after a simple random sampling of 100 this year. Assume that the starting salary for all college graduates follows a normal distribution with a standard deviation of 5.

1) What is the point estimate of the average monthly starting salary of all college graduates?
2) Estimate a 95% confidence interval of the average monthly starting salary of college graduates.
3) Estimate a 99% confidence interval of the average monthly starting salary of college graduates. Compare the width of this interval to the 95% confidence interval.
4) If the sample size is increased to 400 and its average is the same, estimate a 95% confidence interval of the average monthly starting salary for all college graduates. Compare the width of the interval to question 2).

Answer

1) Point estimation of the average monthly starting salary is the sample mean which is 275 (unit: 10,000 KRW).

2) Since the 95% confidence interval implies α = 0.05, z value is as follows.

   $$z_{\alpha/2} = z_{0.05/2} = 1.96$$

   Therefore, the 95% confidence interval is as follows.

   $$\left[\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$
   $$\Leftrightarrow [275\ \text{-}\ 1.96\tfrac{5}{10},\ 275 + 1.96\tfrac{5}{10}]$$
   $$\Leftrightarrow [274.02,\ 275.98]$$

3) Since the 99% confidence interval implies α = 0.01, z value is as follows.

   $$z_{\alpha/2} = z_{0.01/2} = 2.575$$

   Hence, the 99% confidence interval is as follows.

   $$[275\ \text{-}\ 2.575\tfrac{5}{10},\ 275 + 2.575\tfrac{5}{10}]$$
   $$\Leftrightarrow [273.71,\ 276.29]$$

   Therefore, if the confidence level is increasing, the width of the confidence interval becomes wider.

4) If the sample size is 400, the 95% confidence interval is as follows.

   $$\Leftrightarrow [275\ \text{-}\ 1.96\tfrac{5}{20},\ 275 + 1.96\tfrac{5}{20}]$$
   $$\Leftrightarrow [274.51,\ 275.49]$$

   Therefore, as the sample size increases, the width of the confidence interval becomes narrower, which is more accurate.

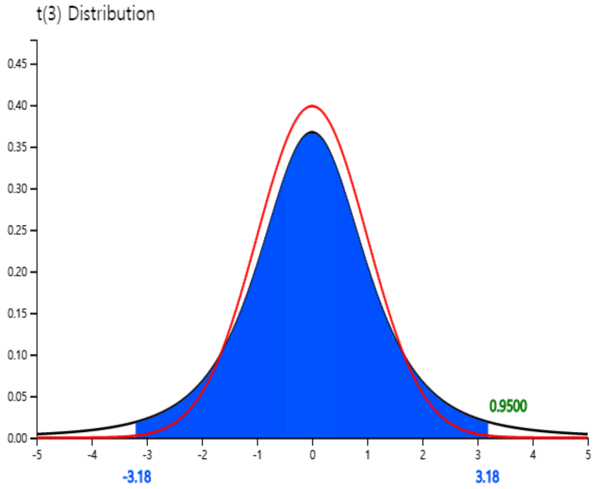
Practice 5.1.1 A large manufacturer's quality manager wants to know raw materials' average weight. Twenty-five samples were collected by simple random sampling, and their sample mean was 60 kg. Assume the population standard deviation is 5 kg. Use 『eStatU』 to answer the following.

1) What is a point estimation of the population mean weight of raw materials?
2) Estimate a 95% confidence interval of the population mean weight of raw materials.
3) Estimate a 99% confidence interval of the population mean weight of raw materials. Compare the width of this interval to the 95% confidence interval.
4) If the sample size is increased to 100 and its average is the same, estimate a 95% confidence interval of the population mean weight of raw materials. Compare the width of the interval to question 2).

Interval estimation of a population mean – Unknown population variance

One problem in estimating the unknown population mean using the confidence interval formula in the previous section is that the population variance may be unknown. If the sample size is large enough, a confidence interval of the population mean can be obtained approximately using the sample variance instead of the population variance in the confidence interval formula. However, if the sample size is small and the sample variance is used, we should use a confidence interval based on the t distribution. The t distribution was studied by a statistician W. S. Gosset, who worked for a brewer in Ireland and published his study result in 1907 under the alias Student. So t distribution is often referred to as Student's t distribution. The t distribution is not just a single distribution, but it is a family of distributions with a parameter called a degree of freedom, 1,2, ... , 30, ... and denoted as $t_1, t_2, \ldots, t_{30}, \ldots$

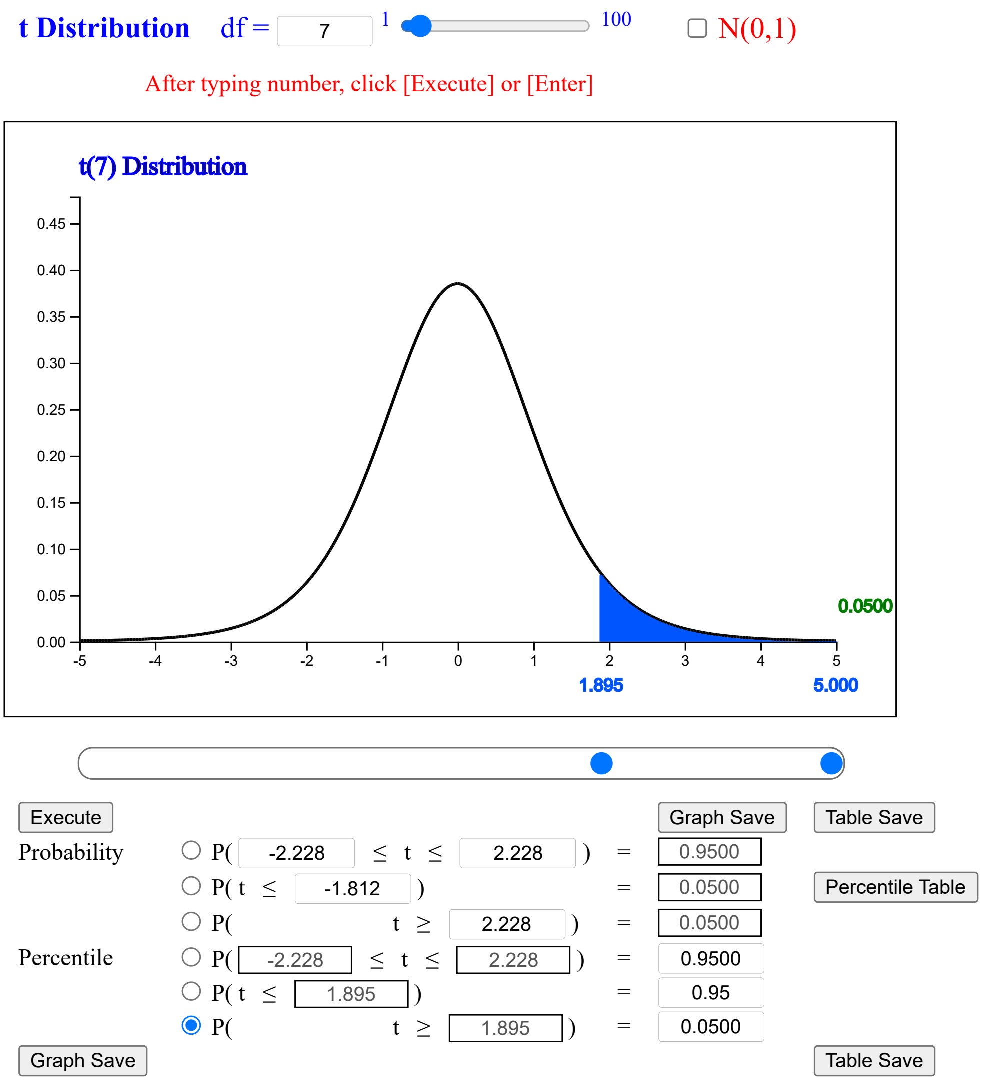
The shape of the t distribution is symmetrical about zero (y axis), similar to the standard normal distribution, but it has a tail that is flat and longer than the standard normal distribution. <Figure 5.1.2> shows the standard normal distribution N(0,1), and t distribution with 3 degrees of freedom simultaneously using the t distribution module of 『eStatU』.



<Figure 5.1.2> Comparison of t_3 and N(0,1)

The t distribution is closer to the standard normal distribution as degrees of freedom increase above 100, which is why a confidence interval can be obtained approximately using the standard normal distribution if the sample size is greater than 100. Denote $t_{n:\,\alpha}$ as the 100×α% percentile from the right tail of the t distribution with n degrees of freedom. For example, $t_{7:\,0.05}$ is the 5% percentile of the t distribution from the right tail and its value is 1.895 as <Figure 5.1.3>. In the standard normal distribution, this value was 1.645. Since the t distribution is symmetrical, $t_{n:\,1-\alpha} = -t_{n:\,\alpha}$. To find a percentile value from the right tail of the t_7 distribution using 『eStatU』, click on 't distribution' in the main menu of 『eStatU』 and then set the degree of freedom (df) to 7, and set the probability value in the sixth option below the t distribution graph to 0.05, then $t_{7:\,0.05}$ = 1.895 will appear.

[t Distribution]



df = 7

Probability

○ P(-2.228 ≤ t ≤ 2.228) = 0.9500

○ P(t ≤ -1.812) = 0.0500

○ P(t ≥ 2.228) = 0.0500

Percentile

○ P(-2.228 ≤ t ≤ 2.228) = 0.9500

○ P(t ≤ 1.895) = 0.95

● P(t ≥ 1.895) = 0.0500

<Figure 5.1.3> The 5% percentile of the t distribution from the right tail

Assume that a population follows a normal distribution, and consider an interval estimation of the population mean in case of the unknown population variance. If $X_1, X_2, \ldots, X_n$ is a random sample of size n from the population, then it can be shown that the distribution of $\frac{\overline{X}-\mu}{S/\sqrt{n}}$, where σ is replaced with S, is the t distribution with $n-1$ degrees of freedom.

$$\frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$$

Hence, the probability of the following interval is (1 - α).

$$P\left(-t_{n-1;\ \alpha/2} < \frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}} < t_{n-1:\ \alpha/2}\right) = 1-\alpha$$

The above formula can be summarized as the confidence interval for the population mean when the population variance is unknown.

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

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

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.

Example 5.1.3 The following data shows a simple random sampling of 10 new male students' heights in a college this year. Use 『eStatU』 to make a 95% confidence interval of the height of the first-year college students.

171 172 185 169 175 177 174 179 168 173

Answer

Click [Estimation : μ Confidence Interval] on the menu of 『eStatU』 and enter data at the [Sample Data] box. Then the confidence intervals [170.68, 177.92] are calculated using the t_9 distribution. In this 『eStatU』 module, confidence intervals can also be obtained by entering the sample sizes, sample mean, and sample variance without entering data.

[Estimation : μ Confidence Interval]

Estimation : μ Confidence Interval

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

171,172,185,169,175,177,174,179,168,173

[Sample Statistics]

Sample Size	n	=	10	*(>1)*
Sample Mean	$\bar{x}$	=	174.30	
Sample Variance	s^2	=	25.57	

[Confidence Level]

$1 - \alpha$ ● 95% ○ 99%

[Sampling Distribution] ● *t Distribution* ○ *Normal Distribution* σ^2 =

Execute Erase Data

[Confidence Interval]

$t_{n-1\,;\,\alpha/2}$ = $s / \sqrt{n}$ =

$\bar{x}$ ± $t_{n-1\,;\,\alpha/2}$ $(s / \sqrt{n})$ ⇔ [,]

n = 10 1 200 1-α = 0.95 0.60 0.99

Graph Save

In this module of 『eStatU』, a simulation experiment to investigate the size of the confidence interval can be done by changing the sample size n and the confidence level 1 - α. If you increase n, the interval size becomes narrower. If you increase 1 - α, the interval size becomes wider.

Practice 5.1.2 In [Practice 5.1.1], suppose you do not know the population standard deviation, and the sample standard deviation is 5 kg. Answer the same questions in [Practice 5.1.1] using 『eStatU』.

5.2 Testing hypothesis for a population mean

Examples of testing hypotheses for a population mean are as follows.

- The weight of a cookie bag is indicated as 200g. Would there be enough cookies to meet the indicated weight?
- At a light bulb factory, a newly developed light bulb advertises a longer bulb life than the past one. Is this propaganda reliable?
- Immediately after completing this year's academic test, students said there would be a 5-point increase in the average English score, which is higher than last year. How can you investigate if this is true?

The testing hypothesis is an answer to the above questions (hypothesis). The testing hypothesis is a statistical decision-making method using samples, which is used to compare two hypotheses about the population parameter. This section discusses the test of the population mean, which is frequently used in applications. The following example explains the theory of the testing hypothesis about a single population mean.

Example 5.2.1 At a light bulb factory, the average life expectancy of a light bulb made by a conventional production method is known to be 1500 hours, and the standard deviation is 200 hours. Recently, the company has been trying to introduce a new production method, with an average life expectancy of 1600 hours for light bulbs. Thirty samples were taken by simple random sampling from the new type of light bulbs to confirm this argument, and the sample mean was $\overline{x}$ = 1555 hours. Can you tell me that the new light bulb has an average life of 1600 hours?

Answer

A statistical approach to the question of this issue is first to make two assumptions about the different arguments for the population mean μ . Namely,

$$H_0 : \mu = 1500$$
$$H_1 : \mu = 1600$$

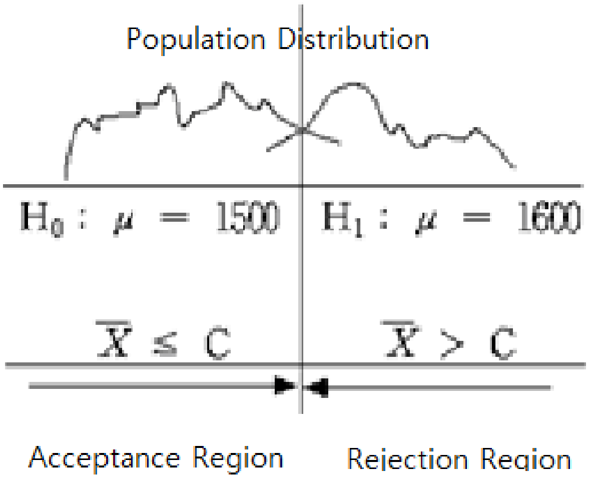
H_0 is called a null hypothesis and H_1 is an alternative hypothesis. In most cases, the null hypothesis is defined as an ‘existing known fact’ and the alternative hypothesis is defined as ‘new facts or changes in current beliefs’. So when choosing between two hypotheses, the basic idea of testing a hypothesis is 'unless there is a significant reason, we accept the null hypothesis (current fact) without choosing the alternative hypothesis (the fact of the matter). This idea of testing a hypothesis is referred to as ‘conservative decision making’.

A common sense for choosing between two hypotheses is 'which population mean of two hypotheses is closer in the distance to the sample mean'. Based on this common sense, which uses the concept of distance, the sample mean of 1555 is closer to $H_1 : \mu = 1600$, so we choose the alternative hypothesis.

However, a statistical testing hypothesis makes a decision using the sampling distribution of $\overline{X}$ to select a critical value C and to make a decision rule as follows.

‘If $\overline{X}$ is smaller than C, then the null hypothesis H_0 will be chosen, else reject H_0’

The area of $\{\overline{X} \le C\}$ is called an **acceptance region** of H_0 and the area $\{\overline{X} > C\}$ is called a **rejection region** of H_0 (<Figure 5.2.1>).



<Figure 5.2.1> Acceptance and rejection region of H_0

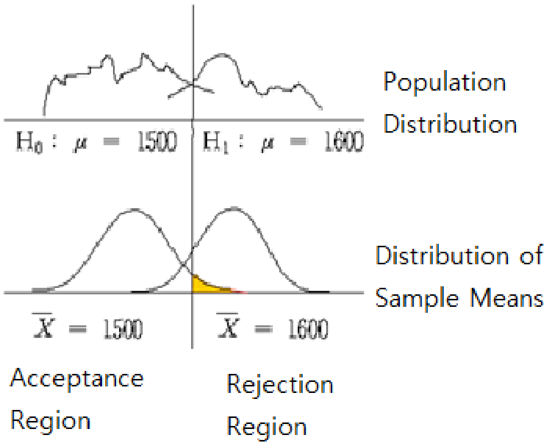
If this decision rule chooses a hypothesis, there are always two possible errors. One is a **Type 1 Error** which accepts H_1 when H_0 is true, the other is a **Type 2 Error** which accept H_0 when H_1 is true. We can summarize these errors as in Table 5.2.1.

Table 5.2.1 Two types of errors in testing hypothesis

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

If you try to reduce one type of error when the sample size is fixed, the other type of error will increase. That is why we came up with a conservative decision-making method that defines the null hypothesis H_0 as 'past or present facts' and 'accept the null hypothesis unless there is significant evidence for the alternative hypothesis.' In this conservative way, we try to reduce the type 1 error as much as possible that selects H_1 when H_0 is true, which would be more risky than the type 2 error. The testing hypothesis determines the tolerance for the probability of the type 1 error, usually 5% or 1% for rigorous tests, and uses the selection criteria that satisfy this limitation. The tolerance for the probability that this type 1 error will occur is called the **significance level** and denoted as α. The probability of the type 2 error is denoted as ß.

If the significance level is established, the decision rule for the two hypotheses can be tested using the sampling distribution of all possible sample means. <Figure 5.2.2> shows two population distributions of two hypotheses and their sampling distributions of all possible sample means in each hypothesis.



<Figure 5.2.2> Testing Hypothesis

The sampling distribution of all possible sample means, which corresponds to the population of the null hypothesis H_0 : μ = 1500, is approximately normal $N(1500, 200^2)$ by the central limit theorem. The sampling distribution of all possible sample means, which corresponds to the population of the alternative hypothesis H_1 : μ = 1600, is approximately normal $N(1600, 200^2)$. The population standard deviation for each population is assumed to be 200 from historical data. Then, the decision rule becomes as follows.

‘If $\overline{X} \le C$, then accept H_0, else accept H_1 (i.e. reject H_0)’

In Figure 5.2.2, the shaded area represents the probability of the type 1 error. If we set the significance level, which is the tolerance level of the type 1 error, is 5%, i.e., $P(\overline{X} \le C) = 0.95$, C can be calculated by finding the percentile of the normal distribution $N(1500, \frac{200^2}{30})$ as follows.

$$1500 + 1.645 \frac{200}{\sqrt{30}} = 1560.06$$

Therefore, the decision rule can be written as follows.

‘If $\overline{X} \le 1560.06$, then accept H_0, else reject H_0 (accept H_1).’

In this problem, the observed sample mean of the random variable $\overline{X}$ is $\overline{x}$= 1555 and H_0 is accepted. In other words, the hypothesis of H_0 : μ = 1500 is judged to be correct, which contradicts the result of common sense criteria that $\overline{x}$ = 1555 is closer to H_1 : μ = 1600 than H_0 : μ = 1500. We can interpret that the sample mean of 1555 is insufficient evidence to reject the null hypothesis using a conservative decision-making method.

The above decision rule is often written as follows, emphasizing that it results from a conservative decision-making method.

‘If $\overline{X} \le 1560.06$, then do not reject H_0, else reject H_0.’

In addition, this decision rule can be written for calculation purposes as follows.

‘If $\frac{\overline{X}-1500}{\frac{200}{\sqrt{30}}} \le 1.645$, then accept H_0 , else reject H_0.’

In this case, since $\overline{x}$ = 1555, $\frac{1555-1500}{\frac{200}{\sqrt{30}}}$ = 1.506, and it is less than 1.645. Therefore, we accept H_0.

Since the testing hypothesis by the conservative decision-making is only based on the probability of the type 1 error as seen in [Example 5.2.1], even if the alternative hypotheses is $H_1 : \mu > 1500$, we will have the same decision rule. Generally, there are three types of alternative hypotheses in the testing hypothesis for the population mean as follows.

1) $H_1 : \mu > \mu_0$	2) $H_1 : \mu < \mu_0$	3) $H_1 : \mu \neq \mu_0$

Since 1) has the rejection region on the right side of the sampling distribution of all possible sample means under the null hypothesis, it is called a **right-sided test**. Since 2) has the rejection region on the left side of the sampling distribution, it is called a **left-sided test**. Since 3) has rejection regions on both sides of the sampling distribution, it is called a **two-sided test**.

In [Example 5.2.1], if the sample mean is either 1555 or 1540, we cannot reject the null hypothesis, but the degrees of evidence that the null hypothesis is not rejected are different. The degree of evidence that the null hypothesis is not rejected is measured by calculating the probability of the type 1 error when the observed sample mean value is considered as the critical value for decision, which is called the **p-value**. That is, the p-value indicates where the observed sample mean is located among all possible sample means by considering the location of the alternative hypothesis. In [Example 5.2.1], the p-value for $\overline{X} = 1540$ is the probability of sample means which is greater than $\overline{X} = 1540$ using $N(1500, \frac{200^2}{30})$ as follows.

$$p\text{-value} = P(\overline{X} > 1540) = P(\frac{\overline{X} - 1500}{\frac{200}{\sqrt{30}}}) = 0.0660$$

The higher the p-value, the stronger the reason for not being rejected. If H_0 is rejected, the smaller the p-value, the stronger the grounds for rejection. Therefore, if the p-value is less than the significance level the analyst decided, then H_0 is rejected because it means that the sample mean is in the rejection region. Statistical packages provide this p-value. The **decision rule using p-value** is as follows.

'If p-value < α, then H_0 is rejected, else H_0 is accepted.'

If the population standard deviation, σ, is unknown and the population follows a normal distribution, the test statistic

$$\frac{\overline{X} - \mu_0}{\frac{S}{\sqrt{n}}}$$

is a t distribution with $(n-1)$ degrees of freedom. If the population standard deviation is unknown, the decision rule for each type of three alternative hypothesis are summarized in Table 5.2.2 where α is the significance level.

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 $H_0 : \mu \geq \mu_0$

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 size is small, the decision rule is as follows.

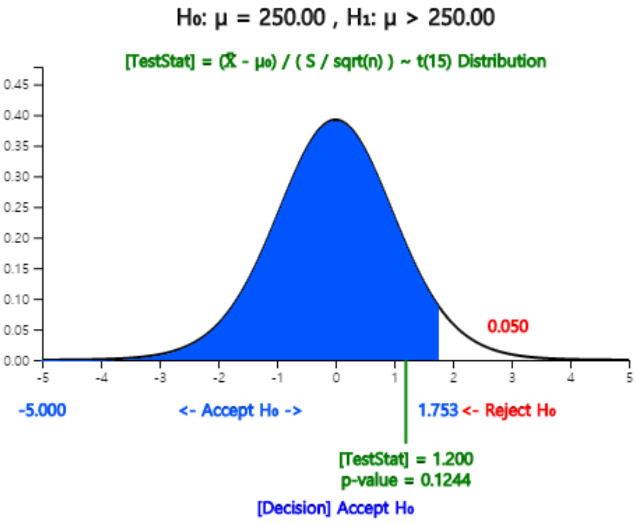
$$\text{'If } \frac{\overline{X} - \mu_0}{\frac{S}{\sqrt{n}}} > t_{n-1:\ \alpha}, \text{ then reject } H_0 \text{ else accept } H_0'$$

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

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.

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

In『eStatU』menu, select [Testing Hypothesis μ], enter 250 at the box on [Hypothesis] and select the alternative hypotheses as the right test. Check [Test Type] as t test and enter &alpha = 0.01. At the [Sample Statistics], enter sample size 16, sample mean 253, and sample variance $10^2 = 100$. If you click the [Execute] button, the confidence Interval for μ is calculated, and the testing result will appear as in <Figure 5.2.3>.



<Figure 5.2.3> Testing hypothesis for μ with t distribution using『eStatU』

Since the p-value is the probability that t_{15} is greater than the test statistics 1.200, the p-value is 0.124 using the module of t distribution in『eStatU』.

[Testing Hypothesis μ]

Testing Hypothesis μ

[Hypothesis] $H_o : \mu = \mu_o$ 250

○ $H_1 : \mu \neq \mu_o$ ● $H_1 : \mu > \mu_o$ ○ $H_1 : \mu < \mu_o$

[Test Type] ● *t test* ○ *Z test* $\sigma^2 =$

Significance Level $\alpha =$ 0.01 *(0 < α < 1)*

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

[Sample Statistics]

Sample Size	n	=	16	*(>1)*
Sample Mean	$\bar{x}$	=	253	
Sample Variance	s^2	=	100	

[Confidence Interval] $t_{n-1;\ \alpha/2} =$

$\bar{x} \ \pm \ t_{n-1\ ;\ \alpha/2} \ (s\ /\sqrt{\ n}\ ) \ \Leftrightarrow \ [\quad , \quad]$

Execute $\alpha =$ 0.05 0 ● 1 Erase Data

Graph Save

Practice 5.2.1 The following data are weights of the 7 employees randomly selected who are working in the shipping department of a wholesale food company.

> 154, 186, 159, 174, 183, 163, 181 (unit pound)
> Ex ⇨ DataScience ⇨ Weight.csv.

Based on this data, is the average weight of employees working in the shipping department 160 or greater than 160? Use the significance level of 5%.

5.3 Testing hypothesis for two populations means

When samples are selected independently from two populations, an estimator for the difference of two population means, $\mu_1 - \mu_2$, is the difference of two sample means, $\overline{x}_1 - \overline{x}_2$. The sampling distribution of all possible sample means differences is approximately a normal distribution with the mean $\mu_1 - \mu_2$ and variance $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$ if both sample sizes are large enough. Since the population variances σ_1^2 and σ_2^2 are usually unknown, sample variances, S_1^2 and S_2^2, are used. If the two populations follow normal distributions and their variances can be assumed to be the same, we can show that the following sample statistic for the sample means difference follows t-distribution with $n_1 + n_2 - 2$ degrees of freedom.

$$\frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}} \qquad \text{where } S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

s_p^2 is an estimator of the population variance called as a **pooled variance** which is an weighted average of two sample variances s_1^2 and s_2^2 using the sample sizes as weights when population variances are assumed to be the same.

Assume that two populations follow normal distributions as $N(\mu_1, \sigma_1^2)$, and $N(\mu_1, \sigma_1^2)$. Consider the interval estimation of the population mean difference when you do not know the population variances, but they can be assumed to be the same. Using the sampling distribution of the sample mean differences described above, the 100(1 - α)% confidence interval for the population mean difference when the population variances are unknown can be shown as follows.

$$\left[(\overline{X}_1 - \overline{X}_2) - t_{n_1+n_2-2:\ \alpha/2}\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}},\ (\overline{X}_1 - \overline{X}_2) + t_{n_1+n_2-2:\ \alpha/2}\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}} \right]$$

where n_1 and n_2 are the sample size, $\overline{X}_1$ and $\overline{X}_2$ are sample means of each population. s_p^2 is an estimator of the population variance, called the **pooled variance**.

A comparison of two populations means, μ_1 and μ_2. is possible by testing the hypothesis that the difference in the population means is equal to zero or not. There are many examples comparing the means of two populations as follows.

- Is there a difference between the starting salary of male and female graduates in this year's college graduates?
- Is there a difference in the weight of the products produced in the two production lines?

Generally, testing hypothesis for two populations means can be divided into three types, depending on the type of alternative hypothesis.

1) $H_0 : \mu_1 - \mu_2 = D_0 \qquad H_1 : \mu_1 - \mu_2 > D_0$
2) $H_0 : \mu_1 - \mu_2 = D_0 \qquad H_1 : \mu_1 - \mu_2 < D_0$
3) $H_0 : \mu_1 - \mu_2 = D_0 \qquad H_1 : \mu_1 - \mu_2 \neq D_0$

Here D_0 is the value for the difference in population means to be tested. When samples are selected independently from two populations, the estimator of the difference of two population means, $\mu_1 - \mu_2$, is the difference of sample means, $\overline{x}_1 - \overline{x}_2$. If two populations follow normal distributions and their variances can be assumed to be the same, the testing hypothesis for the difference between the two populations means uses the following statistic.

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

The test statistic follows a t-distribution with $n_1 + n_2 - 2$ degrees of freedom. The decision rule for testing the difference between the two populations' means is as follows.

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 $\frac{(\overline{x}_1 - \overline{x}_2) - D_0}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} > t_{n_1+n_2-2;\,\alpha}$, 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 $\frac{(\overline{x}_1 - \overline{x}_2) - D_0}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} < -t_{n_1+n_2-2;\,\alpha}$, 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 $\left| \frac{(\overline{x}_1 - \overline{x}_2) - D_0}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} \right| > t_{n_1+n_2-2;\,\alpha/2}$, then reject H_0, else accept H_0

Note: Assume independent samples, normal populations, population variances are equal.
If sample sizes are large enough ($n_1 > 30, n_2 > 30$), t-distribution is approximately close to the standard normal distribution and the decision rule may use the standard normal distribution.

Example 5.3.1 Two machines produce cookies at a factory, and a cookie bag's average weight should be 270g. We sampled cookie bags from each of the two machines to examine the weight of the cookie bags. The average weight of 15 cookie bags extracted from machine 1 was 275g, and their standard deviation was 12g. The average weight of 14 cookie bags extracted from machine 2 was 269g, and the standard deviation was 10g.

1) Find a 99% confidence interval for the difference between two population means.
2) Test whether the two machines' cookie bag weights are different. Use α = 0.01.
3) Check the test result using『eStatU』.

Answer

1) We can summarize the sample information in this example as follows.

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

Therefore, the pooled variance of two samples is as follows.

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{(15-1)12^2 + (14-1)10^2}{15+14-2} = 122.815$$

Since the t-value for 99% confidence interval is $t_{15+14-2;\ 0.01/2} = t_{27:\ 0.005} = 2.7707$, the 99% confidence interval is as follows.

$$\left[(\overline{X}_1 - \overline{X}_2) - t_{n_1+n_2-2:\ \alpha/2}\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}},\ (\overline{X}_1 - \overline{X}_2) + t_{n_1+n_2-2:\ \alpha/2}\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}} \right]$$

$$\left[(275-269) - 2.7707\sqrt{\frac{122.815}{15} + \frac{122.815}{14}},\ (275-269) + 2.7707\sqrt{\frac{122.815}{15} + \frac{122.815}{14}} \right]$$

$$[\,-5.410,\ 17.410\,]$$

2) 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.

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

$D_0 = 0$ in this exaample. The calculation of the test statistic is as follows.

$$\left| \frac{275-269}{\sqrt{\frac{122.815}{15} + \frac{122.815}{14}}} \right| = 1.457$$

Since 1.457 < 2.7707, H_0 can not be rejected.

3) In 『eStatU』 menu, select [Testing Hypothesis μ_1, μ_2]. At the window shown in <Figure 5.3.1>, check the alternative hypotheses of not equal case at [Hypothesis], check the variance assumption of [Test Type] as the equal case, check the significance level of 1%, check the independent sample, and enter sample sizes n_1, n_2, sample means $\overline{x}_1, \overline{x}_2$, and sample variances as the following window. Click [Execute] button to see the confidence interval and result of the testing hypothesis.

[Testing Hypothesis μ₁, μ₂]

Testing Hypothesis μ_1, μ_2

[Hypothesis] $H_o : \mu_1 - \mu_2 = D$ 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 Significance Level $\alpha =$ 0.01 *$(0 < \alpha < 1)$*

[Sampling Type] ● independent sample ○ paired sample

[Variance Assumption] ● $\sigma_1^2 = \sigma_2^2$ ○ $\sigma_1^2 \neq \sigma_2^2$

[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	n_d	=	
Sample Mean	$\bar{x}_1$	=	275	$\bar{x}_2$	=	269	$\bar{d}$	=	
Sample Variance	s_1^2	=	144	s_2^2	=	100	s_d^2	=	

[Confidence Interval : μ_1 - μ_2] $t_{n_1 + n_2 - 2 \,;\, \alpha/2} =$

$$(\bar{x}_1 - \bar{x}_2) \;\pm\; [D + t_{n_1 + n_2 - 2 \,;\, \alpha/2} \sqrt{(s_p^2 / n_1 + s_p^2 / n_2)}\,] \;\Leftrightarrow\; [\quad , \quad]$$

$$s_p^2 = [(n_1-1)\, s_1^2 + (n_2-1)\, s_2^2] / (n_1 + n_2 - 2) \;=$$

Execute $\alpha =$ 0.01 0 ● 1 Erase Data

Graph Save

If variances of two populations are different, the test statistic

$$\frac{(\overline{x}_1 - \overline{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

does not follow a t-distribution even if populations are normally distributed. The testing hypothesis for two populations means when their population variances are different is called a Behrens-Fisher problem, and several methods to solve this problem have been studied. The Satterthwaite method approximates the degrees of freedom of the t-distribution in the decision rule in Table 5.3.1 with ϕ as follows.

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

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)

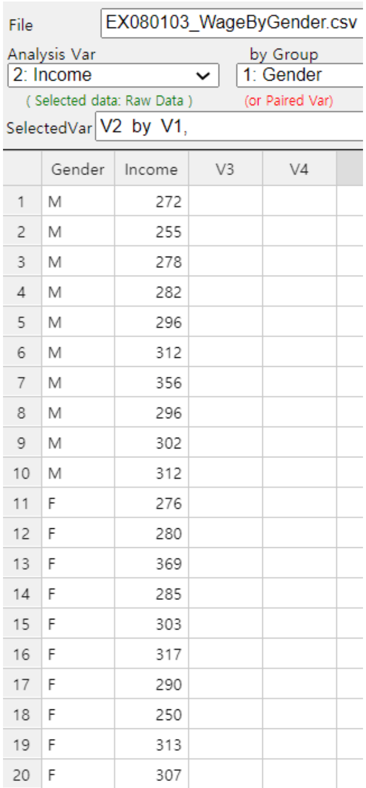
Male 272 255 278 282 296 312 356 296 302 312
Female 276 280 369 285 303 317 290 250 313 307
Ex ⇨ DataScience ⇨ WageByGender.csv.

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.

Answer

1) In『eStat』, enter raw data of gender (M or F) and income as shown in <Figure 5.3.1> on the sheet. This type of data input is similar to all statistical packages. After entering the data, click the icon for testing two populations' means and select 'Analysis Var' as V2 and 'By Group' variable as V1. A 95% confidence interval graph that compares the sample means of two populations will be displayed as <Figure 5.3.2>.



File EX080103_WageByGender.csv

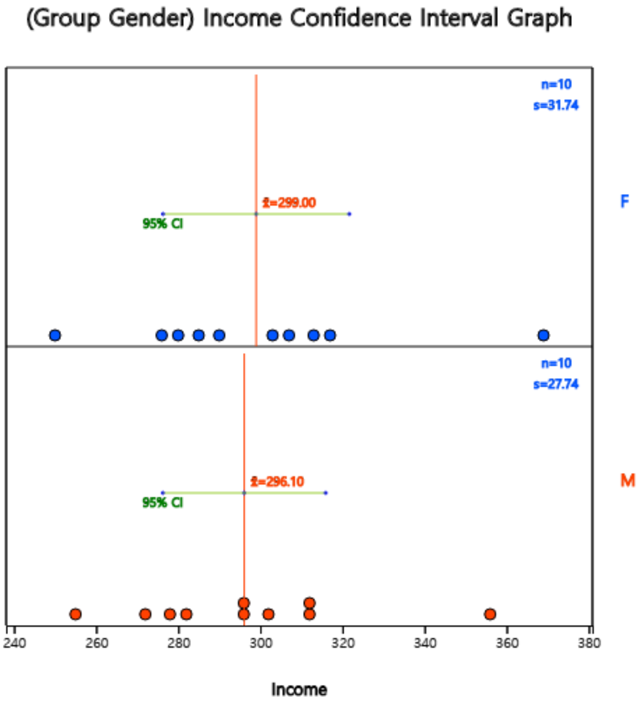
Analysis Var 2: Income by Group 1: Gender

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

SelectedVar V2 by V1,

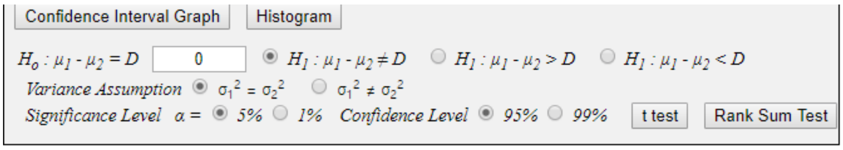
	Gender	Income	V3	V4
1	M	272		
2	M	255		
3	M	278		
4	M	282		
5	M	296		
6	M	312		
7	M	356		
8	M	296		
9	M	302		
10	M	312		
11	F	276		
12	F	280		
13	F	369		
14	F	285		
15	F	303		
16	F	317		
17	F	290		
18	F	250		
19	F	313		
20	F	307		

<Figure 5.3.1> Data input for testing two populations means



<Figure 5.3.2> Dot graph and confidence Intervals by gender for testing two populations means

In the options window, as in <Figure 5.3.3> located below the Graph Area, enter the average difference $D = 0$ for the desired test, select the variance assumption $\sigma_1^2 = \sigma_2^2$, the 5% significance level and click the [t-test] button. Then, the graphical result of the testing hypothesis for two populations' means will be shown as in <Figure 5.3.4> and the test result as in <Figure 5.3.5>.

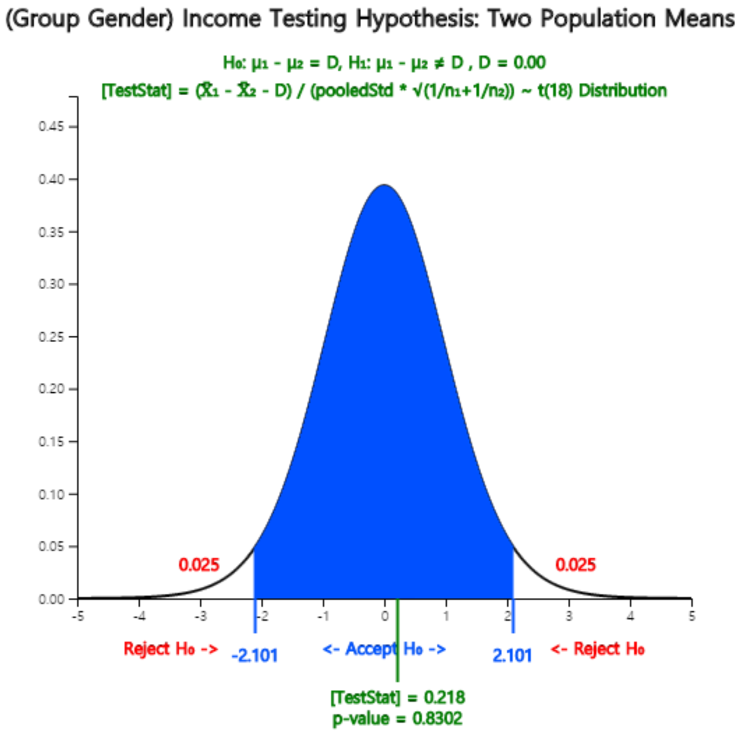


$H_0 : \mu_1 - \mu_2 = D$ 0 ● $H_1 : \mu_1 - \mu_2 \neq D$ ○ $H_1 : \mu_1 - \mu_2 > D$ ○ $H_1 : \mu_1 - \mu_2 < D$

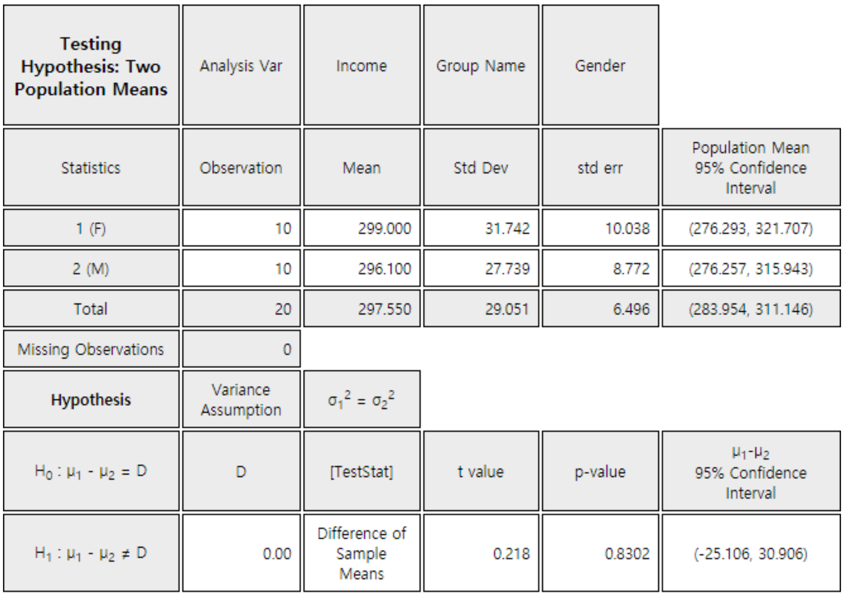
Variance Assumption ● $\sigma_1^2 = \sigma_2^2$ ○ $\sigma_1^2 \neq \sigma_2^2$

Significance Level $\alpha =$ ● 5% ○ 1% Confidence Level ● 95% ○ 99%

<Figure 5.3.3> Options to test for two populations means



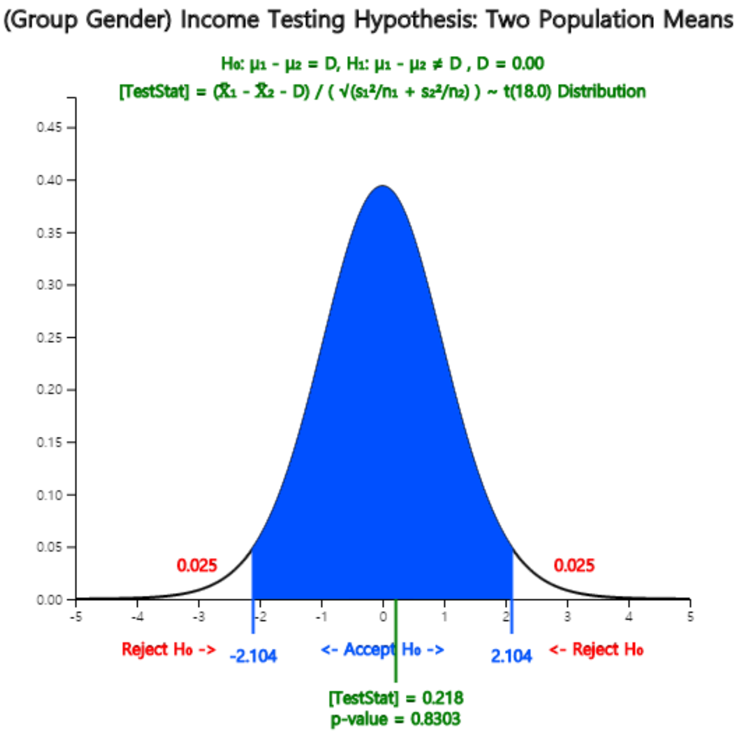
<Figure 5.3.4> Testing hypothesis for and - case of the same population variances



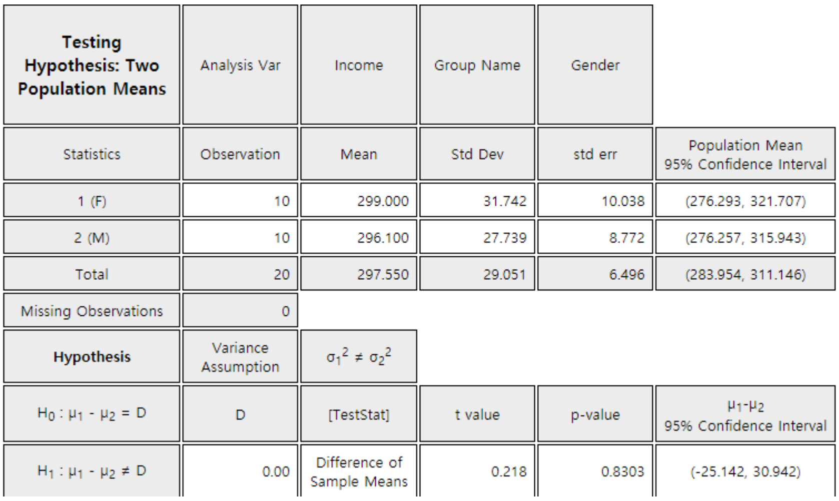
Testing Hypothesis: Two Population Means	Analysis Var	Income	Group Name	Gender	
Statistics	Observation	Mean	Std Dev	std err	Population Mean 95% Confidence Interval
1 (F)	10	299.000	31.742	10.038	(276.293, 321.707)
2 (M)	10	296.100	27.739	8.772	(276.257, 315.943)
Total	20	297.550	29.051	6.496	(283.954, 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)

<Figure 5.3.5> The result of testing hypothesis for two populations means if population variances are the same

2) Select the variance assumption $\sigma_1^2 \neq \sigma_2^2$ at the option window and click [t-test] button under the graph to display the graph of the hypothesis test and the test result table as in <Figure 5.3.6> and <Figure 5.3.7>.



<Figure 5.3.6> Testing hypothesis for and – case of the different population variances



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)

<Figure 5.3.7> result of testing hypothesis for two populations means if population variances are different

Practice 5.3.1 (Oral Cleanliness by Brushing Methods)
Oral cleanliness scores were examined for eight samples using the basic brushing method (coded 1) and seven samples using the rotation method (coded 2). The data are saved at the following location of 『eStat』.

Ex ⇨ DataScience ⇨ ToothCleanByBrushMethod.csv

1) If population variances are the same, test the hypothesis at the 5% significance level to determine whether scores for both brushing methods are the same using 『eStat』.
2) If population variances are different, test the hypothesis at the 5% significance level to determine whether scores for both brushing methods are the same using 『eStat』.

5.4 Testing hypothesis for several population means: Analysis of variances

Section 5.3 discussed comparing the means of two populations using the testing hypothesis. This section discusses comparing the means of several populations. There are many examples of comparing means of several populations as follows.

- 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 the average monthly wages of college graduates the same in three different cities?

The group variable used to distinguish population groups, such as the grade or the rice, is called a **factor**. This section describes the one-way analysis of variance (ANOVA), which compares population means when there is a single factor. Let us take a look at the following example.

Example 5.4.1 We collected samples randomly from each grade to compare the English proficiency scores of each grade at a university, and the data are in Table 5.4.1. The last column is the average $\overline{y}_{1\cdot}$, $\overline{y}_{2\cdot}$, $\overline{y}_{3\cdot}$, $\overline{y}_{4\cdot}$ for each grade.

Table 5.4.1 English Proficiency Score by Grade

Socre	Student 1	Student 2	Student 3	Student 4	Student 5	Student 6	Student Average
Grade 1	81	75	69	90	72	83	$\overline{y}_{1\cdot}$=78.3
Grade 2	65	80	73	79	81	69	$\overline{y}_{2\cdot}$=74.5
Grade 3	72	67	62	76	80		$\overline{y}_{3\cdot}$=71.4
Grade 4	89	94	79	88			$\overline{y}_{4\cdot}$=87.5

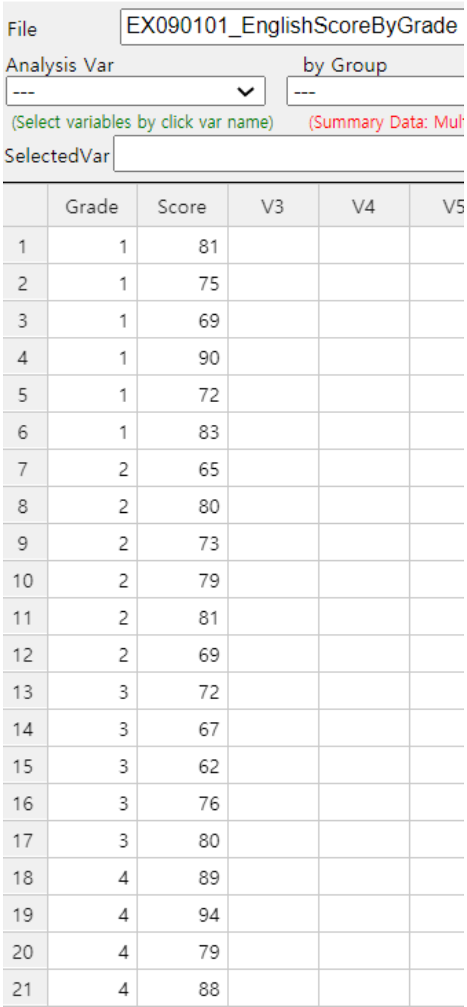
[Ex] ⇨ DataScience ⇨ EnglishScoreByGrade.csv.

1) Draw a dot graph of test scores for each grade and compare their averages using 『eStat』.
2) Set up a null hypothesis and an alternative hypothesis. Test a hypothesis whether the average scores of each grade are the same or not.
3) Apply the one-way analysis of variances to test the hypothesis in question 2).
4) Check the result of the ANOVA test using 『eStat』.

Answer

1) Enter data on the sheet to draw a dot graph with data shown in Table 5.4.1 using 『eStat』, and set variable names to 'Grade' and 'Score' as shown in <Figure 5.4.1>. In the variable selection box appeared by clicking the ANOVA icon on the main menu of 『eStat』, select 'Analysis Var' as ‘Score’ and 'By Group' as ‘Grade’. The dot graph of English scores by each grade and the 95% confidence interval are displayed in <Figure 5.4.2>. Clicking the 'Confidence Interval Graph' button, we can see a more detailed comparison of the population mean on each dot graph. <Figure 5.4.2> shows sample means as $\overline{y}_{1\cdot}$= 78.3, $\overline{y}_{2\cdot}$ = 74.5, $\overline{y}_{3\cdot}$ = 71.4, $\overline{y}_{4\cdot}$ = 87.5. The sample mean of the 4th grade is relatively larger than the other grades and $\overline{y}_{2\cdot}$ and $\overline{y}_{3\cdot}$ are similar. Therefore, we can expect that the population

mean μ_2 and μ_3 would be the same and μ_4 will differ from three other population means. However, we need to test whether these differences of sample means are statistically significant.



File EX090101_EnglishScoreByGrade

Analysis Var ---

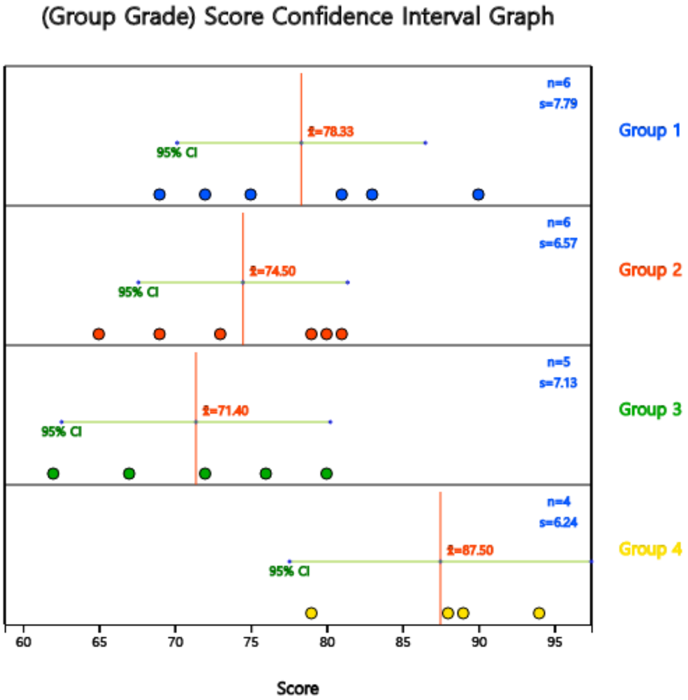
by Group ---

(Select variables by click var name) (Summary Data: Mul

SelectedVar

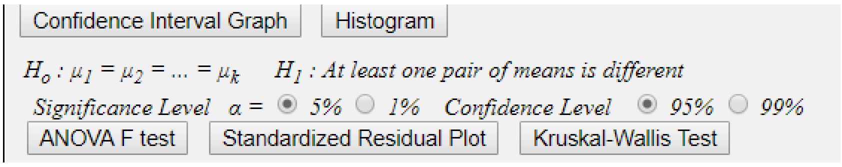
	Grade	Score	V3	V4	V5
1	1	81			
2	1	75			
3	1	69			
4	1	90			
5	1	72			
6	1	83			
7	2	65			
8	2	80			
9	2	73			
10	2	79			
11	2	81			
12	2	69			
13	3	72			
14	3	67			
15	3	62			
16	3	76			
17	3	80			
18	4	89			
19	4	94			
20	4	79			
21	4	88			

<Figure 5.4.1> 『eStat』 data input for ANOVA



<Figure 5.4.2> 95% Confidence Interval by grade

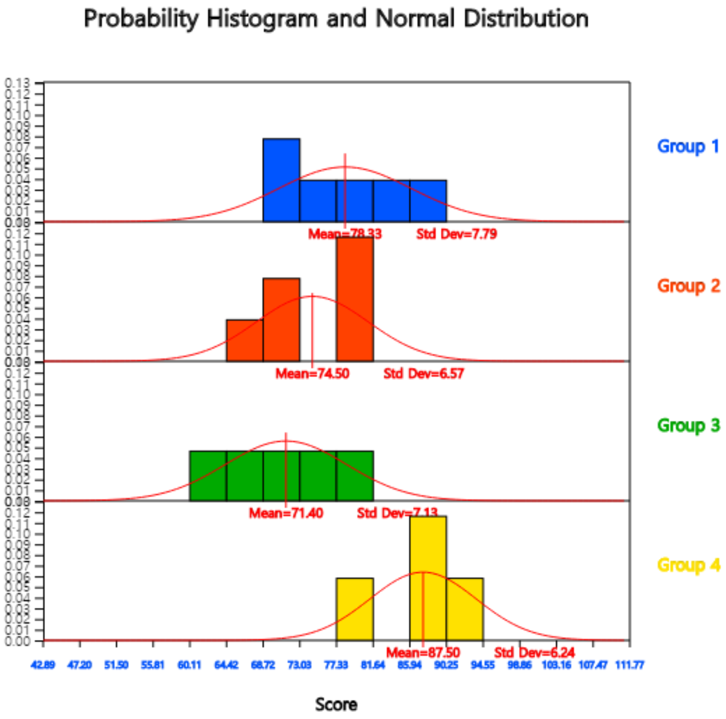
Clicking the [Histogram] button under this graph, as in <Figure 5.4.3>, to check the normality of the data will draw histograms and normal distributions simultaneously, as shown in Figure 5.4.4>



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

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

<Figure 5.4.3> Options of ANOVA



<Figure 5.4.4> Histogram of English score by grade

2) In this example, the null hypothesis to test is that the population means of English scores of the four grades are all the same, and the alternative hypothesis is that the population means of the English scores are not the same. In other words, if μ_1, μ_2, μ_3, and μ_4 are the population means of English scores for each grade, the hypothesis to test can be written as follows,

Null hypothesis H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$
Alternative hypothesis H_1: at least one pair of μ_i is not the same

3) A measure that can be considered first as a basis for testing differences in multiple sample means would be the distance from each mean to the overall mean. In other words, if the overall sample mean for all 21 students is expressed as $\overline{y}_{\cdot\cdot}$, the squared distance from each sample mean to the overall mean is as follows when the number of samples in each grade is weighted. This squared distance is called the **between sum of squares (SSB)** or the **treatment sum of squares (SSTr)**.

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

If the squared distance $SSTr$ is close to zero, all sample means of English scores for four grades are similar. However, this treatment sum of squares can be larger if the number of populations increases. Modifications are required to become a test statistic to determine whether several population means are equal. The squared distance from each observation to its sample mean of the grade is called the **within sum of squares (SSW)** or the **error sum of squares (SSE)** as defined below.

$$\begin{aligned} SSE = & (81 - \overline{y}_{1\cdot})^2 + (75 - \overline{y}_{1\cdot})^2 + \cdots + (83 - \overline{y}_{1\cdot})^2 \\ & + (65 - \overline{y}_{2\cdot})^2 + (80 - \overline{y}_{2\cdot})^2 + \cdots + (69 - \overline{y}_{2\cdot})^2 \\ & + (72 - \overline{y}_{3\cdot})^2 + (67 - \overline{y}_{3\cdot})^2 + \cdots + (80 - \overline{y}_{3\cdot})^2 \end{aligned}$$

$$+(89-\overline{y}_{4\cdot})^2+(94-\overline{y}_{4\cdot})^2+\cdots+(88-\overline{y}_{4\cdot})^2$$
$$=839.033$$

If population distributions of English scores in each grade follow normal distributions and their variances are the same, the following test statistic has the $F_{3,17}$ distribution.

$$F_0=\frac{\frac{SSTr}{(4-1)}}{\frac{SSE}{(21-4)}}$$

This statistic can be used to test whether or not the population's English scores in four grades are the same. In the test statistic, the numerator $\frac{SSTr}{4-1}$ is called the **treatment mean square (MSTr)**, which implies a variance between grade means. The denominator $\frac{SSE}{21-4}$ is called the **error mean square (MSE)**, which implies a variance within each grade. The MSE is a pooled variance of four sample variances. Thus, the above test statistics are based on the ratio of two variances, which is why the test of multiple population means is called an **analysis of variance (ANOVA)**.

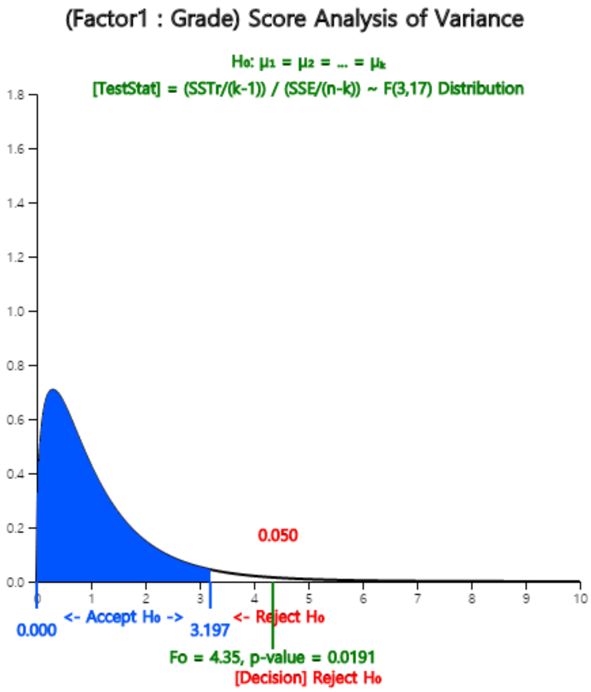
The calculated test statistic, which is the observed F value F_0, using data of English scores for each grade is as follows.

$$F_0=\frac{\frac{SSTr}{(4-1)}}{\frac{839.033}{(21-4)}}=\frac{\frac{643.633}{(4-1)}}{\frac{SSE}{(21-4)}}=4.347$$

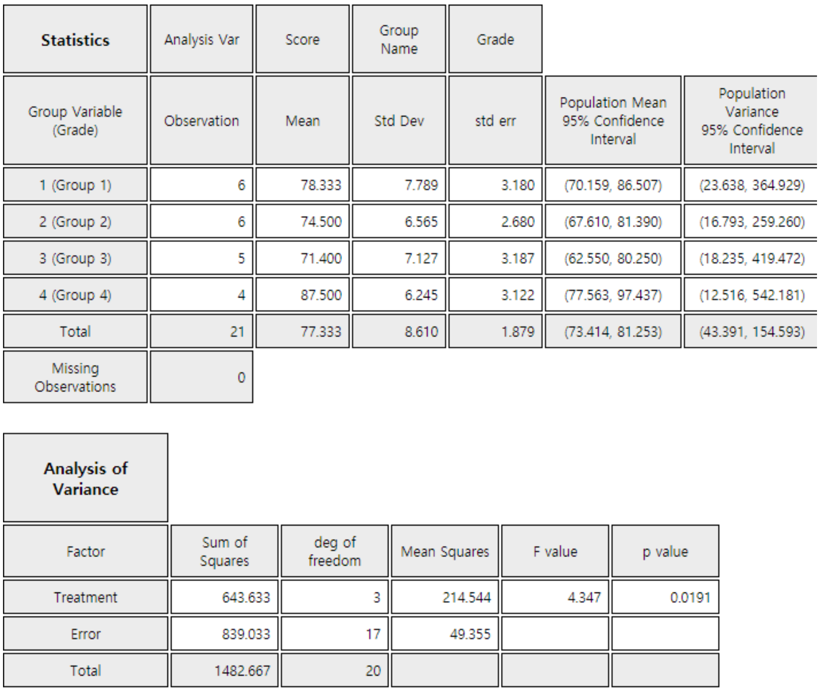
Since $F_{3,17;\ 0.05}$ = 3.20, the null hypothesis that population means of English scores of each grade are the same, $H_0:\mu_1=\mu_2=\mu_3=\mu_4$, is rejected at the 5% significance level. In other words, there is a difference in population means of English scores of each grade. The following ANOVA table provides a single view of the above calculation.

Factor	Sum of Squares	Degree of freedom	Mean Squares	F ratio
Treatment	SSTr = 643.633	4-1	MSTr = $\frac{643.633}{3}$	$F_0=4.347$
Error	SSE = 839.033	21-4	MSE = $\frac{839.033}{17}$	
Total	SST = 1482.666	20		

4) In <Figure 5.4.3>, if you select the significance level of 5%, the confidence level of 95%, and click [ANOVA F test] button, a graph showing the location of the test statistic in the F distribution is appeared as shown in <Figure 5.4.5>. Also, in the Log Area, the mean and confidence interval tables and test results for each grade appear in <Figure 5.4.6>.



<Figure 5.4.5> 『eStat』 ANOVA F test



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			

<Figure 5.4.6> 『eStat』 Basic Statistics and ANOVA table

The analysis of variance is also possible using 『eStatU』 as below. Entering the data as below, and clicking the [Execute] button will have the same result as in <Figure 5.4.5> and <Figure 5.4.6>.

[Single Factor ANOVA]

Single Factor ANOVA

[Hypothesis] $H_o: \ \mu_1 = \mu_2 = \ldots = \mu_k$

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

[Test Type] *F test* (ANOVA)

Significance Level $\alpha =$ 0.05 *(0 < α < 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 $\alpha =$ 0.05 0 1 Erase Data

Multiple Comparison ● LSD ○ 5%HSD ○ 1%HSD Graph Save Table Save

The above example refers to two variables: the English score and grade. The variable, such as the English score, is called an **analysis variable** or a **response variable**. The response variable is mostly a continuous variable. The variable used to distinguish populations, such as the grade, is called a **group variable** or a **factor variable**, which is mostly a categorical variable. Each value of a factor variable is called a **level** of the factor, and the number of these levels is the number of populations to be compared. In the above example, the factor has four levels, 1st, 2nd, 3rd and 4th grade. The term 'response' or 'factor' is originated to analyze data in engineering, agriculture, medicine, and pharmacy experiments. The analysis of variance method that examines the effect of a single factor on the response variable is called the **one-way ANOVA**. Table 5.4.2 shows the typical data structure of the one-way ANOVA when the number of levels of a factor is k, and the numbers of observations at each level are $n_1, n_2, \ldots, n_k$.

Table 5.4.2 Notation of the 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	$Y_{21}\ Y_{22}\ \cdots\ Y_{2n_2}$	$\overline{Y}_{2\cdot}$
$\cdots$	$\cdots$	$\cdots$
Level k	$Y_{k1}\ Y_{k2}\ \cdots\ Y_{kn_k}$	$\overline{Y}_{k\cdot}$
	Total	$\overline{Y}_{\cdot\cdot}$

Statistical model for the one-way analysis of variance is given as follows.

$$\begin{aligned} Y_{ij} &= \mu_i + \epsilon_{ij} \\ &= \mu + \alpha_i + \epsilon_{ij},\ i = 1, 2, \ldots, k;\ j = 1, 2, \ldots, n_i \\ &\text{where}\ \ \epsilon_{ij} \backsim N(0, \sigma^2) \end{aligned}$$

Y_{ij} represents the j^{th} observed value of the response variable for the i^{th} level of factor. The population mean of the i^{th} level, μ_i, is represented as $\mu + \alpha_i$ where μ is the mean of entire population and α_i is the effect of i^{th} level for the response variable. ϵ_{ij} denotes an error term of the j^{th} observation for the i^{th} level, and the all error terms are assumed independent of each other and follow the same normal distribution with the mean 0 and variance σ^2. The error term ϵ_{ij} is a random variable in the response variable due to reasons other than levels of the factor. For example, in the English score example, differences in English performance for each grade can be caused by other variables besides the variables of grade, such as individual study hours, gender and IQ. However, by assuming that these variations are relatively small compared to variations due to differences in grade, the error term can be interpreted as the sum of these various reasons. The hypothesis to test can be represented using α_i instead of μ_i as follows.

Null hypothesis H_0: $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$
Alternative hypothesis H_1: at least one α_i is not equal to 0

The analysis of variance table as Table 5.4.3 is used to test the hypothesis.

Table 5.4.3 Analysis of variance table of the one-way ANOVA

Factor	Sum of Squares	Degree of freedom	Mean Squares	F ratio
Treatment	SSTr	$k-1$	MSTr=$\frac{SSTr}{k-1}$	$F_0 = \frac{MSTr}{MSE}$
Error	SSE	$n-k$	MSE=$\frac{SSE}{n-k}$	
Total	SST	$n-1$		
		where $n = \sum_{i=1}^{n} n_i$		

The three sum of squares for the analysis of variances can be described as follows.

SST = $\sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{\cdot\cdot})^2$:
The sum of squared distances between observed values of the response variable and the mean of total observations is called the **total sum of squares** (SST).

SSTr = $\sum_{i=1}^{k} \sum_{j=1}^{n_i} (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2$:
The sum of squared distances between the mean of each level and the mean of total observations is called the **treatment sum of squares** (SSTr). It represents the variation between level means.

SSE = $\sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i\cdot})^2$:
The sum of squared distances between observations of the i^{th} level and the mean of the i^{th} level is referred to as 'within variation', and is called the **error sum of squares** (SSE).

The following logic determines the degree of freedom of each sum of squares. The SST consists of n number of squares, $(Y_{ij} - \overline{Y}_{\cdot\cdot})^2$, but $\overline{Y}_{\cdot\cdot}$ should be calculated first, before SST is calculated, and hence the degree of freedom of SST is $n-1$. The SSE consists of n number of squares, $(Y_{ij} - \overline{Y}_{i\cdot})^2$, but the number of values, $\overline{Y}_{1\cdot}, \overline{Y}_{2\cdot}, \ldots, \overline{Y}_{k\cdot}$ should be calculated first before SSE is calculated, and hence, the degree of freedom of SSE is $n-k$. The degree of freedom of SSTr is calculated as the degree of freedom of SST minus the degree of freedom of SSE, which is $k-1$. In the one-way analysis of variance, the following partition of the sum of squares and degree of freedom are always established;

Sum of squares: SST = SSTr + SSE
Degrees of freedom: $(n-1) = (k-1) + (n-k)$

The sum of squares divided by the corresponding degrees of freedom is referred to as the mean squares, and Table 5.4.3 defines the treatment mean squares (MSTr) and error mean squares (MSE). The treatment mean square implies the average variation between each level of the factor, and the error mean square implies the average variation within observations in each level. Therefore, if MSTr is relatively much larger than MSE, we can conclude that the population means of each level, μ_i, are not the same. So by what criteria can you say it is relatively much larger?

The calculated F value, F_0, in the last column of the ANOVA table represents the relative size of MSTr and MSE. If the assumptions of ϵ_{ij} are satisfied, and if the null hypothesis $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$ is true, then the test statistic follows a F distribution with degrees of freedoms $k-1$ and $n-k$.

$$F_0 = \frac{\frac{SSTr}{(k-1)}}{\frac{SSE}{(n-k)}}$$

Therefore, when the significance level is α for a test, if the calculated value F_0 is greater than the value of $F_{k-1,n-k;\alpha}$, then the null hypothesis is rejected. That is, it is determined that the population means of each factor level are different. (Note: 『eStat』 calculates this test's p-value. Hence, if the p-value is smaller than the significance level α, then reject the null hypothesis.)

Practice 5.4.1 (Plant Growth by Condition)
Results from an experiment to compare yields (as measured by the dried weight of plants) obtained under a control (leveled ‘ctrl’) and two treatment conditions (leveled ‘trt1’ and ‘trt2’). The weight data with 30 observations on control and two treatments (‘crtl’, ‘trt1’, ‘trt2’), are saved at the following location of 『eStat』. Answer the following using 『eStat』 ,

[Ex] ⇨ DataScience ⇨ PlantGrowth.csv

1) Draw a dot graph of weights for each control and treatment.
2) Test a hypothesis whether the weights are the same or not. Use the 5% significance level.

5.5 Regression analysis

5.5.1 Correlation analysis

Sample correlation coefficient r can be used for testing the hypothesis of a population correlation coefficient ρ. We test usually $H_0 : \rho = 0$ which tests the existence of linear correlation. This test can be done using t distribution as follows.

Testing a population correlation coefficient

Null hypothesis: $H_0 : \rho = 0$

Test statistic: $t_0 = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}}$ follows t distribution with $n-2$ degrees of freedom

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

We can also test a hypothesis $H_0 : \rho = \rho_0$ when $\rho_0 \neq 0$, but please refer other statistics book.

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 as in Table 5.5.1. Draw a scatter plot for this data using 『eStat』, and find the sample correlation coefficient of the two variables. Test the hypothesis that the population correlation coefficient is zero with the significance level 0.05.

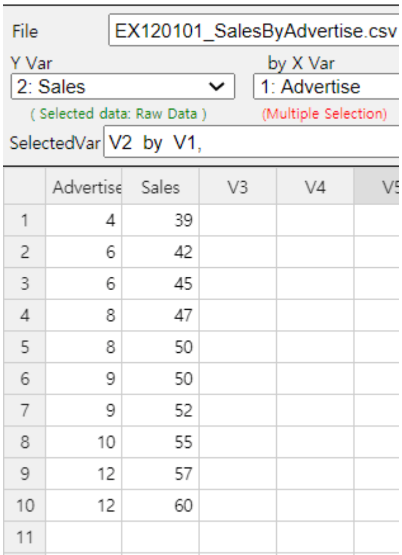
Table 5.5.1 Advertising costs and sales (unit: 1 million USD)

Company	Advertise (X)	Sales (Y)
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

[Ex] ⇨ DataScience ⇨ SalesByAdvertise.csv.

Answer

Using『eStat』, enter data as shown in <Figure 5.5.1>. If you select the Sales as 'Y Var' and the Advertise 'by X Var' in the variable selection box that appears when you click the scatter plot icon on the main menu, the scatter plot will appear as shown in <Figure 5.5.2>. As we can expect, the scatter plot show that the more investments in advertising, the more sales increase, and not only that, the form of increase is linear.



File: EX120101_SalesByAdvertise.csv

Y Var: 2: Sales

by X Var: 1: Advertise

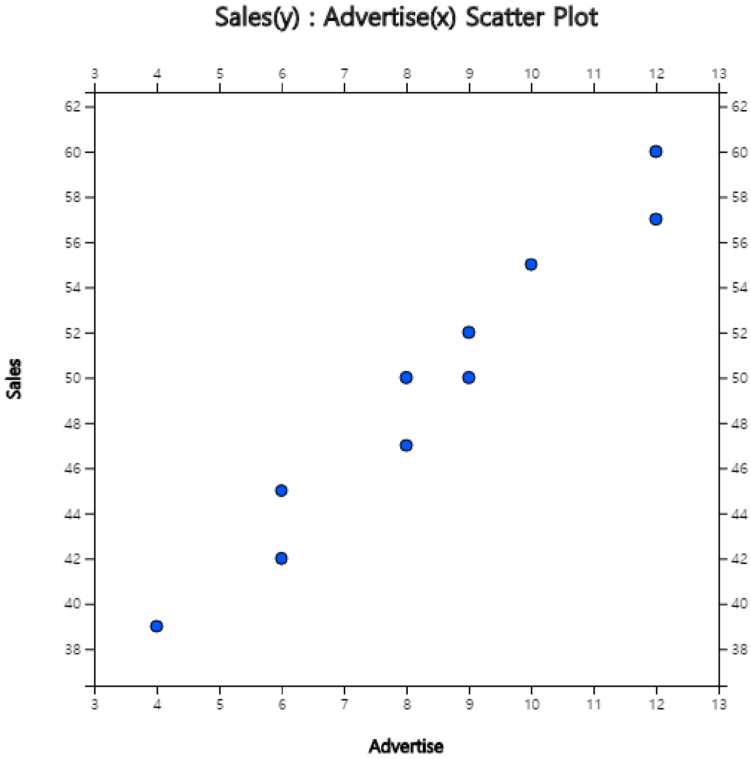
(Selected data: Raw Data)

(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					

<Figure 5.5.1> Data input in『eStat』



<Figure 5.5.2> Scatter plot of sales by advertise

To calculate the sample covariance and correlation coefficient, it is convenient to make the following table. This table can also be used for calculations in regression analysis.

Table 5.5.1 A table for calculating the covariance and correlation coefficient

Number	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	**64**	**497**	**766**	**25097**	**4326**
Mean	**8.4**	**49.7**			

Terms which are necessary to calculate the covariance and correlation coefficient are as follows:

$$SXX = \sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2 = 766 - 10 \times 8.4^2 = 60.4$$
$$SYY = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2 = 25097 - 10 \times 49.7^2 = 396.1$$
$$SXY = \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \sum_{i=1}^{n} X_i Y_i - n\overline{XY} = 4326 - 10 \times 8.4 \times 49.7 = 151.2$$

SXX, SYY, SXYrepresent the sum of squares of X, the sum of squares of Y, the sum of squares of XY . Hence, the covariance and correlation coefficient are as follows:

$$S_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y}) = \frac{151.2}{10-1} = 16.8$$

$$r = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2 \sum_{i=1}^{n} (Y_i - \overline{Y})^2}} = \frac{151.2}{\sqrt{60.4 \times 396.1}} = 0.978$$

This sample correlation coefficient is consistent with the scatter plot which shows a strong positive correlation of the two variables. The value of the test statistic t_0 is as follows.

$$t_0 = \sqrt{10-2} \frac{0.978}{\sqrt{1-0.978^2}}$$ = 13.117

Since it is greater than $t_{8;\ 0.025}$ = 2.306, $H_0 : \rho = 0$ should be rejected.

The correlation analysis can be done using『eStatU』by following data input and clicking [Execute] button..

[Correlation Analysis]

Correlation Analysis

[Hypothesis] $H_o : \rho = 0$ ● $H_1 : \rho \neq 0$ ○ $H_1 : \rho > 0$ ○ $H_1 : \rho < 0$

[TestStat] $t_0 = \sqrt{(n-2)}\ r / \sqrt{(1-r^2)} =$ p-value =

[Sample Data] (Sample size of each cell should be the same.)

X Data Input	4,6,8,8,9,9,10,12,12
Y Data Input	39,42,45,47,50,50,52,55,57,60
Main Title	
y title	
x title	

Number of Data	n_x		n_y			
Mean	$\bar{X}$		$\bar{Y}$			
Sample Variance(n-1)	$S_x^{\ 2}$		$S_y^{\ 2}$		Sample Covariance	S_{xy}
Sample Std Deviation	S_x		S_y		Sample Correlation Coefficient	r

Execute Erase Data

□ Regression Line

Graph Save

Practice 5.5.1 A professor of statistics argues that a student's final test score can be predicted from his midterm score. Ten students were randomly selected, and their mid-term and final exam scores are as follows.

id	Mid-term X	Final Y
1	92	87
2	65	71
3	75	75
4	83	84
5	95	93
6	87	82
7	96	98
8	53	42
9	77	82
10	68	60

[Ex] ⇨ DataScience ⇨ MidtermFinal.csv.

1) Draw a scatter plot of this data with the X-axis mid-term and Y-axis final scores. What do you think is the relationship between mid-term and final scores?
2) Find the sample correlation coefficient and test the hypothesis that the population correlation coefficient is zero with a significance level 0.05.

5.5.2 Simple linear regression

Data are concentrated around a straight line when two variables show a strong correlation. In this case, **linear regression analysis** is a statistical model to estimate the straight line which describes the data's relationship suitably. The estimated model can be applied to the forecasting analysis. For example, a mathematical model of the relationship between sales (Y) and advertising costs (X) would not only explain the relationship between sales and advertising costs but would also be able to predict the sales for a given investment for advertisement. As such, the regression analysis is intended to investigate and predict the degree of relation between variables and the shape of the relation.

In regression analysis, a mathematical model of the relation between variables is called a **regression equation**, and the variable affected by other related variables is called a **dependent variable**. The dependent variable is the variable we would like to describe, which is usually observed in response to other variables, so it is also called a **response variable**. In addition, variables that affect the dependent variable are called **independent variables**. The independent variable is also referred to as the **explanatory variable** because it is used to describe the dependent variable. In the previous example, if the objective is to analyze the change in sales amounts resulting from increases and decreases in advertising costs, the sales is a dependent variable, and the advertising cost is an independent variable. If the number of independent variables included in the regression equation is one, it is called a **simple linear regression**. If the number of independent variables is two or more, it is called a **multiple linear regression**, explained in section 5.5.3.

Simple linear regression analysis has only one independent variable, and the regression equation is as follows.

$$Y = f(X, \alpha, \beta) = \alpha + \beta X$$

In other words, the regression equation is represented by a linear equation of the independent variable, and α and β are unknown parameters that represent the intercept and slope, respectively. The α and β are called the **regression coefficients**. The above equation represents an unknown linear relationship between Y and X in population and is referred to as the population regression equation.

To estimate the regression coefficients α and β, observations of the dependent and independent variables are required, i.e., samples. In general, all of these observations are not located in a line. It is because, even if the Y and X have an exact linear relation, there may be a measurement error in the observations, or there may not be an exact linear relationship between Y and X. Therefore, we can write the regression formula by considering these errors as follows.

$$Y_i = \alpha + \beta X_i + \epsilon_i, \quad i = 1, 2, \dots, n$$

Where i is the subscript representing the i^{th} observation, and ϵ_i is the random variable indicating an error with a mean of zero and a variance σ^2 which is independent of each other. The error ϵ_i indicates that the observation Y_i is how far away from the population regression equation. The above equation includes unknown population parameters α, β, and σ^2 and is referred to as a population regression model.

If a and b are the estimated regression coefficients using samples, the fitted regression equation can be written as follows. It is referred to as the sample regression equation.

$$\hat{Y}_i = a + bX_i$$

In this expression, $\hat{Y}_i$ represents the estimated value of Y at $X = X_i$ as predicted by the appropriate regression equation. These predicted values can not match the actual observed values of Y, and differences between these values are called residuals and denoted as e_i.

$$\text{residuals} \qquad e_i = Y_i - \hat{Y}_i, \quad i = 1, 2, \ldots, n$$

The regression analysis makes assumptions about the unobservable error ϵ_i. Since the residuals e_i calculated using the sample values have similar characteristics as ϵ_i, they are used to investigate the validity of these assumptions.

When sample data, $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, are given, a straight line representing it can be drawn in many ways. Since one of the main objectives of a regression analysis is prediction, we would like to use the estimated regression line that would make the residuals smallest that the error occurs when predicting the value of Y. However, it is impossible to minimize the residuals' value at all points, and it should be chosen to make the residuals 'totally' smaller. The most widely used of these methods is a method that minimizes the total sum of squared residuals, called a **method of least squares**.

Method of least squares

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

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

To obtain the values of α and β by the least squares method, the sum of squares above should be differentiated partially with respect to α and β, and equate them zero respectively. If the solution of α and β of these equations is a and b, the equations can be written as follows.

$$a \cdot n + b \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i$$
$$a \sum_{i=1}^{n} X_i + b \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i$$

The above expression is called a **normal equation**. The solution a and b of this normal equation is called a **least squares estimator** of α and β, and is given as follows.

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

After estimating the regression line, how valid it is should be investigated. Since a regression analysis aims to describe a dependent variable as a function of an independent variable, it is necessary to find out how much the explanation is. A residual standard error and a coefficient of determination are used for such validation studies. Residual standard error s measures the extent to which observations are scattered around the estimated line. First, you can define the sample variance of residuals as follows.

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

The residual standard error s is the square root of s^2. The s^2 is an estimate of σ^2 which is the extent that the observations Y are spread around the population regression line. A small value of s or s^2 indicates that the observations are close to the estimated regression line, which in turn implies that the regression line represents well the relationship between the two variables.

However, it is not clear how small the residual standard error s is, although the smaller the value is, the better. In addition, the size of the value of s depends on the unit of Y. A relative measure called the coefficient of determination is defined to eliminate this shortcoming. The **coefficient of determination** is the ratio of the variation described by the regression line over the total variation of observation Y_i, so that it is a relative measure that can be used regardless of the type and unit of a variable. As in the analysis of variance in the previous section, the following partitions of the sum of squares and degrees of freedom are established in the regression analysis:

$$\begin{array}{rc} \text{Sum of squares:} & SST \ = \ SSE \ + \ SSR \\ \text{Degrees of freedom:} & (n-1) = (n-2) + 1 \end{array}$$

Description of the above three sums of squares is as follows.

Total sum of squares : $SST = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$
The total sum of squares indicating the total variation in observed values of Y is called the total sum of squares (SST). This SST has the degree of freedom, $n-1$, and if SST is divided by the degree of freedom, it becomes the sample variance of Y_i.

Error sum of squares : $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$
The error sum of squares (SSE) of the residuals represents the unexplained variation of the total variation of the Y. Since the calculation of this sum of squares requires the estimation of two parameters α and β, SSE has the degree of freedom $n-2$. This is the reason why, in the calculation of the sample variance of residuals s^2, it was divided by $n-2$.

Regression sum of squares : $SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$
The regression sum of squares (SSR) indicates the variation explained by the regression line among the total variation of Y. This sum of squares has the degree of freedom of 1.

If the estimated regression equation fully explains the variation in all samples (i.e., if all observations are on the sample regression line), the unexplained variation SSE will be zero. Thus, if the portion of SSE is small among the total sum of squares SST, or if the portion of SSR is large, the estimated regression model is more suitable. Therefore, the ratio of SSR to the total variation SST, called the **coefficient of determination**, is defined as a measure of the suitability of the regression line as follows.

$$R^2 = \frac{Explained\ Variation}{Total\ Variation} = \frac{SSR}{SST}$$

The value of the coefficient of determination is always between 0 and 1, and the closer the value is to 1, the more concentrated the samples are around the regression line, which means that the estimated regression line explains the observations well.

If we divide three sums of squares obtained in the above example by their degrees of freedom, each becomes a variance. For example, if you divide the SST by $n-1$ degrees of freedom, then it becomes the sample variance of the observed values $Y_1, Y_2, \ldots, Y_n$. If you divide the SSE by $n-2$ degrees of freedom, it becomes s^2 which is an estimate of the variance of error σ^2. For this reason, addressing the problems associated with the regression using the partition of the sum of squares is called the ANOVA of regression. Information required for ANOVA, such as a calculated sum of squares and degrees of freedom, can be compiled in the ANOVA table, as shown in Table 5.5.2.

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**	$n-1$		

The sum of squares divided by its degrees of freedom is referred to as mean squares, and Table 5.5.2 defines the regression mean squares (MSR) and error mean squares (MSE) respectively. As the expression indicates, MSE is the same statistic as s^2 which is the estimate of σ^2. The F value given in the last column is used for testing the hypothesis $H_0: \beta = 0,\ H_1: \beta \neq 0$. If β is not 0, the F value can be expected to be large because the assumed regression line is valid and the variation of Y is explained in large part by the regression line. Therefore, we can reversely decide that β is not zero if the calculated F ratio is large enough. If the assumptions about the error terms mentioned in the population regression model are valid and if the error terms follow a normal distribution, the distribution of F value, when the null hypothesis is true, follows F distribution with 1 and $n-2$ degrees of freedom. Therefore, if $F_0 > F_{1,n-2;\,\alpha}$, then we can reject $H_0: \beta = 0$. (In 『eStat』, the p-value for this test is calculated, and the decision can be made using this p-value. That is, if the p -value is less than the significance level, the null hypothesis H_0 is rejected.)

One assumption of the error term ϵ in the population regression model is that it follows a normal distribution with a mean of zero and variance of σ^2. Under this assumption, the regression coefficients and other parameters can be estimated and tested. Note that, under the assumption above, the regression model $Y = \alpha + \beta X + \epsilon$ follows a normal distribution with the mean $\alpha + \beta X$ and variance σ^2.

1) Inference on the parameter β

The parameter β, the slope of the regression line, indicates the existence and extent of a linear relationship between the dependent and the independent variables. The inference for β can be summarized as follows. The test for hypotheses $H_0: \beta = 0$ is used to determine the independent variable describes the dependent variable significantly or not. The F test for the hypothesis $H_0: \beta = 0$ described in the ANOVA of regression is theoretically the same as in the test below.

Point estimate: $b = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}, \quad b \sim N(\beta, \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2})$

Standard error of estimate b: $SE(b) = \frac{s}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}}$

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

Testing hypothesis:

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

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

rejection region:

$H_1 : \beta < \beta_0$: $t < -t_{n-2;\alpha}$

$H_1 : \beta > \beta_0$: $t > t_{n-2;\alpha}$

$H_1 : \beta \neq \beta_0$: $|t| > t_{n-2;\alpha/2}$

2) Inference on the parameter α

The inference for the parameter α, which is the intercept of the regression line, can be summarized below. The parameter α is not so interesting in most of the analysis because it represents the average value of the response variable when an independent variable is 0.

Point estimate: $a = \overline{Y} - b\overline{X}, \quad a \sim N(\alpha, (\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}) \cdot \sigma^2)$

Standard error of estimate a: $SE(a) = s \cdot \sqrt{\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2})}$

Confidence interval of α: $a \pm t_{n-2;\alpha/2} \cdot SE(a)$

Testing hypothesis:

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

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

rejection region:

$H_1 : \alpha < \alpha_0$: $t < -t_{n-2;\alpha}$

$H_1 : \alpha > \alpha_0$: $t > t_{n-2;\alpha}$

$H_1 : \alpha \neq \alpha_0$: $|t| > t_{n-2;\alpha/2}$

3) Inference on the average value $\mu_{Y|x} = \alpha + \beta X_0$

At any point in $X = X_0$, the dependent variable Y has an average value $\mu_{Y|x} = \alpha + \beta X_0$. Estimation of $\mu_{Y|x}$ is also considered an important parameter because it means predicting the mean value of Y .

Point estimate: $\hat{Y}_0 = a + bX_0$

Standard error of estimate $\hat{Y}_0$: $SE(\hat{Y}_0) = s \cdot \sqrt{\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}}$

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

The confidence interval formula of the mean value $\mu_{Y|x}$ depends on the value of the X given the standard error of the estimate, so the width of the confidence interval depends on the value of the given X. As the formula for the standard error shows, this width is the narrowest at a time $X = \overline{X}$, and if X is the farther away from $\overline{X}$, the wider it becomes. If we calculate the confidence interval for the mean value of Y at each point of X, and then if we connect the upper and lower limits, we have a **confidence band** of the regression line on the above and below of the sample regression line.

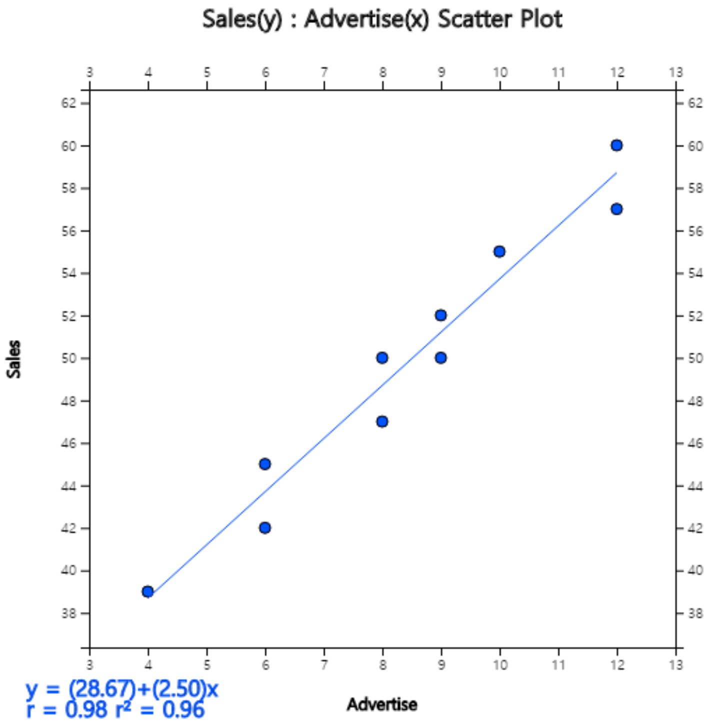
Example 5.5.2 In Example 5.5.1, find the least squares estimate of the slope and intercept if the sales amount is a dependent variable and the advertising cost is an independent variable.
1) Predict the amount of sales when you have spent on advertising by 10.
2) Calculate the value of the residual standard error and the coefficient of determination.
3) Prepare an ANOVA table and test it using the 5% significance level.

Answer

1) In Example 5.5.1, the calculation required to obtain the intercept and slope has already been made. The intercept and slope using this are as follows.

$$b = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{151.2}{60.4} = 2.503$$
$$a = \overline{Y} - b\overline{X} = 49.7 - 2.503 \times 8.4 = 28.672$$

Therefore, the fitted regression line is $\hat{Y}_i = 28.672 + 2.503X_i$. <Figure 5.5.3> shows the fitted regression line on the original data. The slope value, 2.5033, means that if advertising cost increases by one (i.e., one million), sales increase by about 2.5 million.



<Figure 5.5.3> Simple linear regression using『eStat』

Prediction of the sales amount of a company with an advertising cost of 10 can be obtained using the fitted sample regression line as follows.

$$28.672 + (2.503)(10) = 53.702$$

In other words, sales of 53.705 million are expected. That is not to say that all companies with advertising costs of 10 million USD have sales of 53.705 million USD, but that the average amount of their sales is about that. Therefore, there may be some differences in individual companies.

2) To obtain the residual standard error and the coefficient of determination, it is convenient to make the following Table 12.2.1. Here, the estimated value $\hat{Y}_i$ of the sales from each value of X_i uses the fitted regression line.

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

Table 5.5.3 Useful calculations for the residual standard error and coefficient of determination

Number	X_i	Y_i	$\hat{Y}_i$	SST $(Y_i - \overline{Y}_i)^2$	SSR $(\hat{Y}_i - \overline{Y}_i)^2$	SSE $(Y_i - \hat{Y}_i)^2$
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	**64**	**497**	**496.522**	**396.1**	**378.429**	**17.622**
Average	**8.4**	**49.7**				

In Table 12.2.1, SST = 396.1, SSR = 378.429, SSE = 17.622. Here, the relationship of $SST = SSE + SSR$ does not exactly match because number of digits calculation error. The sample variance of residuals is as follows.

$$s^2 = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \frac{17.622}{(10-2)} = 2.203$$

Hence, the residual standard error is s = 1.484. The coefficient of determination is as follows.

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

It means that 95.6% of the total variation in the observed 10 sales amounts can be explained by the simple linear regression model using a variable of advertising costs, so this regression line is quite useful.

3) The ANOVA table using the calculated sum of squares is as follows.

Source	Sum of squares	Degrees of freedom	Mean Squares	F value
Regression	378.42	1	MSR = $\frac{378.42}{1}$ = 378.42	$F_0 = \frac{378.42}{2.20} = 172.0$
Error	17.62	10-2	MSE = $\frac{17.62}{8} = 2.20$	
Total	**396.04**	**10-1**		

Since the calculated F value of 172.0 is much greater than $F_{1,8;\ 0.05} = 5.32$, we reject the null hypothesis $H_0 : \beta = 0$ with the significance level α = 0.05. Inferences about each parameter with the result of a regression analysis are as follows.

(a) Inference for β

The point estimate of β is b = 2.5033, and the standard error of b is as follows.

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

Hence, the 95% confidence interval of β using $t_{8;\ 0.025}$ = 2.056 is as follows.

$$2.5033 \pm (2.056)(0.1908)$$
$$2.5033 \pm 0.3922$$

i.e. the interval (2.1110, 2.8956).

The test statistic for the hypothesis $H_0 : \beta = 0$, is as follows.

$$t = \frac{2.5033 - 0}{0.1908} = 13.12$$

Since $t_{8;0.025}$ = 2.056, the null hypothesis $H_0 : \beta = 0$ is rejected with the significance level of α = 0.05. This result of the two-sided test can be obtained from the confidence interval. Since the 95% confidence interval (1.7720, 3.2346) does not include 0, the null hypothesis $H_0 : \beta = 0$ can be rejected.

(b) Inference for α

The point estimate of α is a = 29.672, and its standard error is as follows.

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

Since the value of t statistic is $\frac{29.672}{1.67}$ = 17.1657 and $t_{8;0.025}$ = 2.056, the null hypothesis $H_0 : \alpha = 0$ is also rejected with the significance level α = 0.05.

(c) Inference for the average value of Y

In 『eStat』 , the standard error of $\hat{Y}$, which is the estimate of $\mu_{Y|x}$, is calculated at each point of X. For example, the point estimate of Y at X = 8 is $\hat{Y}$ = 28.672 + 2.503 × 8 = 48.696 and its standard error is 0.475.

$$SE(\hat{Y}_0) = s \cdot \sqrt{\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}}$$
$$= 1.484 \cdot \sqrt{\frac{1}{10} + \frac{(8-8.4)^2}{60.4}} = 0.475$$

Hence, the 95% confidence interval of $\mu_{Y|x}$ is as follows.

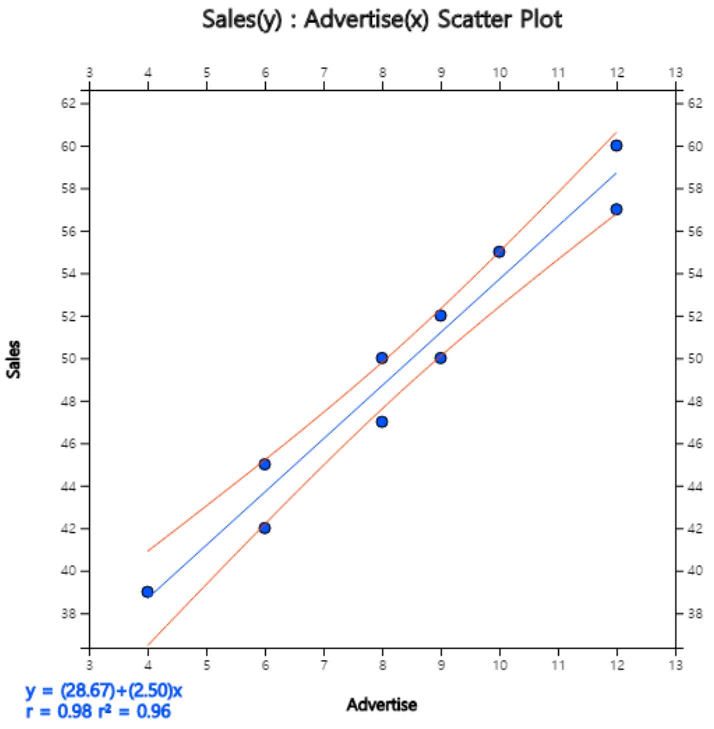
$48.696 \pm (2.056) \times (0.475)$
48.696 ± 0.978
i.e., the inteval is (47.718, 49.674).

We can calculate the confidence interval for other values of X similarly as follows.

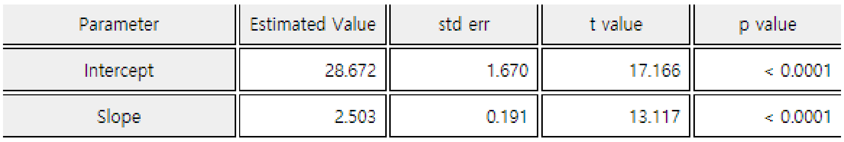
At $X = 4, \quad 38.684 \pm (2.056) \times (0.962) \Rightarrow (36.705, 40.663)$
At $X = 6, \quad 47.690 \pm (2.056) \times (0.656) \Rightarrow (42.341, 45.039)$
At $X = 9, \quad 51.199 \pm (2.056) \times (0.483) \Rightarrow (50.206, 52.192)$
At $X = 12, \quad 58.708 \pm (2.056) \times (0.832) \Rightarrow (56.997, 60.419)$

As we discussed, the confidence interval becomes wider as X is far from $\overline{X}$.

If you select the [Confidence Band] button from the options below, the regression graph of <Figure 5.5.3>, you can see the confidence band graph on the scatter plot together with the regression line as <Figure 5.5.4>. If you click the [Correlation and Regression] button, the inference result of each parameter will appear in the Log Area, as shown in <Figure 5.5.3>.



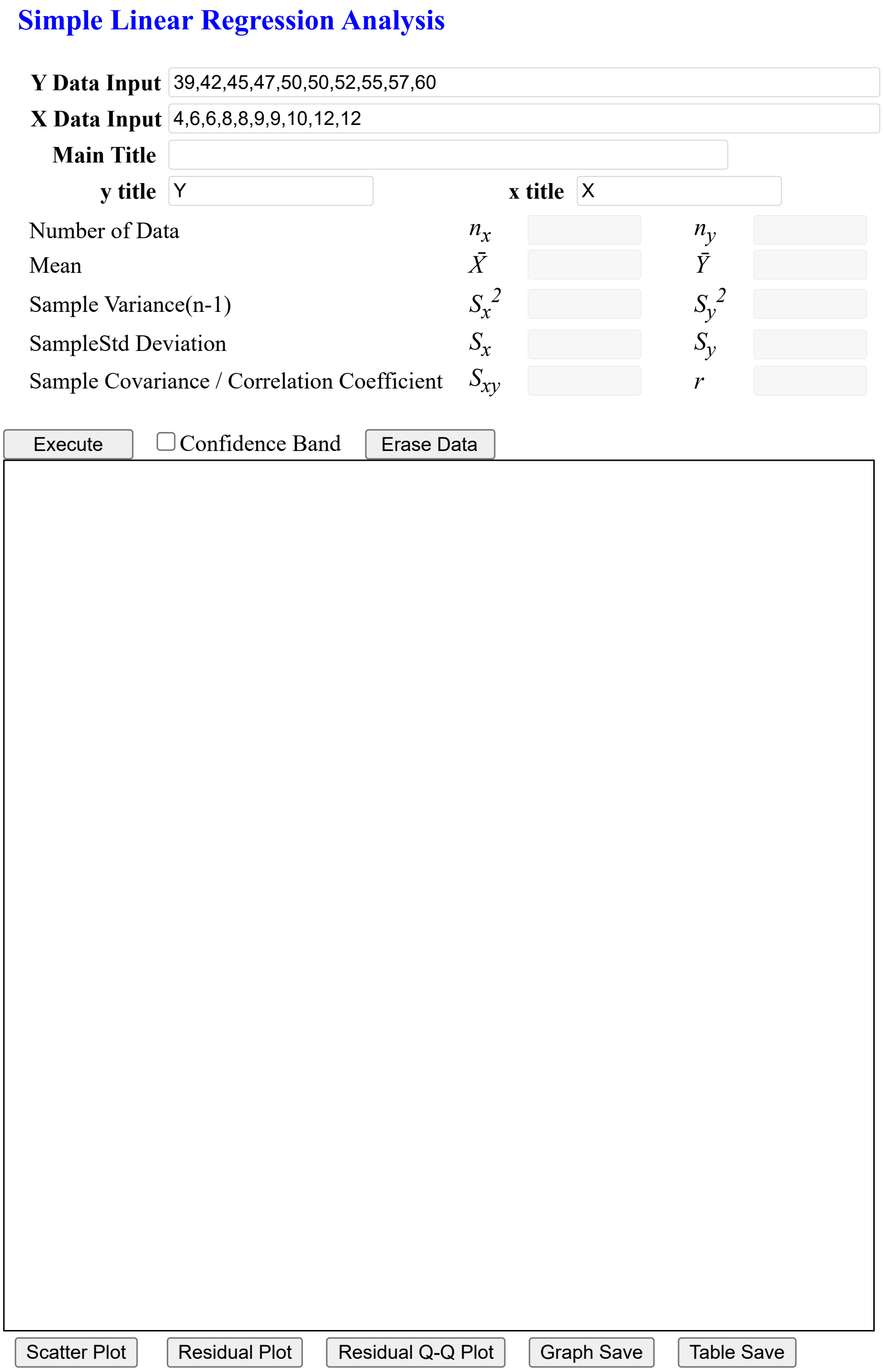
<Figure 5.5.4> Confidence band using『eStat』



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

<Figure 5.5.5> Testing hypothesis of regression coefficients

[Simple Linear Regression Analysis]



Simple Linear Regression Analysis

Y Data Input	39,42,45,47,50,50,52,55,57,60		
X Data Input	4,6,6,8,8,9,9,10,12,12		
Main Title			
y title	Y	**x title**	X

Number of Data	n_x		n_y	
Mean	$\bar{X}$		$\bar{Y}$	
Sample Variance(n-1)	$S_x{}^2$		$S_y{}^2$	
SampleStd Deviation	S_x		S_y	
Sample Covariance / Correlation Coefficient	S_{xy}		r	

Practice 5.5.2 Using the data of [Practice 5.5.1] for the mid-term and final exam scores, find the following:

1) Least squares estimates for the slope and intercept if the final exam score is a dependent variable and the mid-term scores is an independent variable.
2) Predict the final exam score when you have a mid-term score of 80.
3) Residual standard error and coefficient of determination.
4) Prepare an ANOVA table and test it using the 5% significance level.
5) Make inferences about each parameter using 『eStat』 and draw the confidence band.

5.5.3 Multiple linear regression

For actual applications of the regression analysis, the multiple regression models with two or more independent variables are more frequently used than the simple linear regression with one independent variable. It is rare for a dependent variable to be sufficiently explained by a single independent variable; in most cases, a dependent variable has a relationship with several independent variables. For example, we can expect that sales will be significantly affected by advertising costs, examples of simple linear regression, and product quality ratings, and the number and size of stores sold. The statistical model used to identify the relationship between one dependent variable and several independent variables is called a **multiple linear regression analysis**. However, the simple linear regression and multiple linear regression analysis differ only in the number of independent variables involved, and there is no difference in the analysis method.

In the multiple linear regression model, it is assumed that the dependent variable Y and k number of independent variables have the following relational formulas:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \epsilon_i$$

It means that the dependent variable is represented by the linear function of the independent variables and a random variable that represents the error term as in the simple linear regression model. The assumption of the error terms is the same as the assumption in the simple linear regression. In the above equation, β_0 is the intercept of Y axis and β_i is the slope of the Y axis and X_i which indicates the effect of X_i to Y when other independent variables are fixed.

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 estimate the volume using the diameter and height of a tree, which is relatively easy to measure. The data in Table 5.5.4 showes the values for measuring diameter, height, and volume after sampling 15 trees in a region. (The diameter was measured 1.5 meters above the ground.) Draw a scatter plot matrix of this data and consider a regression model for this problem.

Table 5.5.4 Diameter, height and volume of tree

Diameter(cm)	Height(m)	Volume(m^3)
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

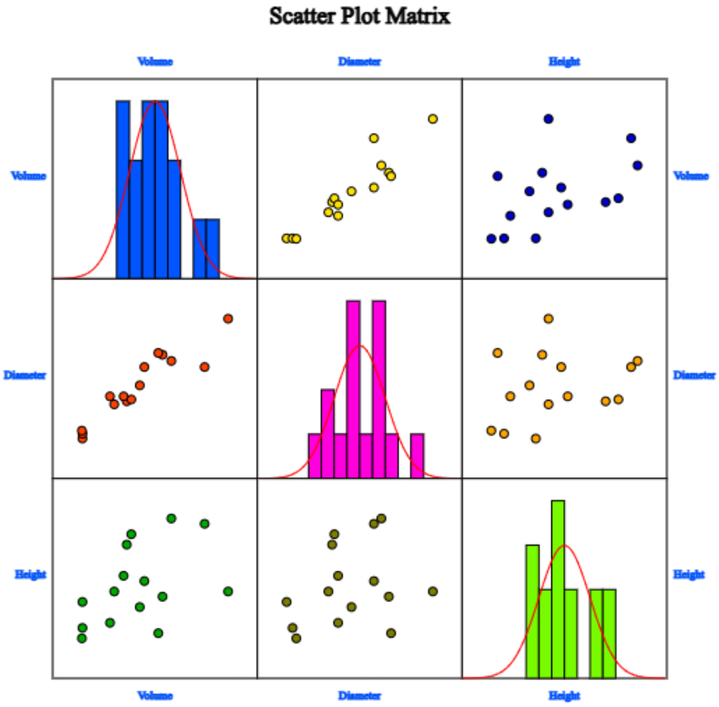
[Ex] ⇨ DataScience ⇨ TreeVolume.csv.

Answer

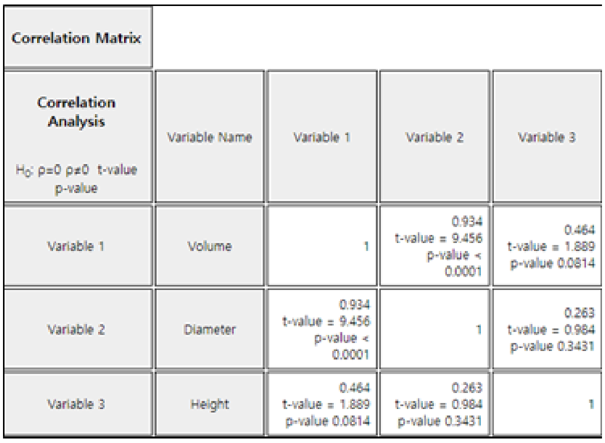
Load the data saved at the following location of 『eStat』.

[Ex] ⇨ DataScience ⇨ TreeVolume.csv

In the variable selection box, which appears by selecting the regression icon, select 'Y variable' by volume and select ‘by X variable’ as the diameter and height to display a scatter plot matrix, as shown in <Figure 5.5.6>. It can be observed that there is a high correlation between volume and diameter, and that volume and height, as well as diameter and height, are also somewhat related.



<Figure 5.5.6> Scatterplot matrix



Correlation Matrix				
Correlation Analysis H$_0$: ρ=0 ρ≠0 t-value p-value	Variable Name	Variable 1	Variable 2	Variable 3
Variable 1	Volume	1	0.934 t-value = 9.456 p-value < 0.0001	0.464 t-value = 1.889 p-value 0.0814
Variable 2	Diameter	0.934 t-value = 9.456 p-value < 0.0001	1	0.263 t-value = 0.984 p-value 0.3431
Variable 3	Height	0.464 t-value = 1.889 p-value 0.0814	0.263 t-value = 0.984 p-value 0.3431	1

<Figure 5.5.7> Correlation matrix

Since the volume is to be estimated using the diameter and height of the tree, the volume is the dependent variable Y, and the diameter and height are independent variables X_1, X_2 respectively, and we can consider the following regression model.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i, \quad i = 1, 2, \ldots, 15$$

The same analysis of multiple linear regression can be done using『eStatU』by following data input and clicking [Execute] button..

[Multiple Linear Regression Analysis]

Multiple Linear Regression Analysis

Menu

Variable Name		Data Input
Y		0.291,0.291,0.288,0.464,0.532,0.557,0.441,0.515,0.603,0.628,0.956,0.775,0.727,
X_1		21.0,21.8,22.3,26.6,27.1,27.4,27.9,27.9,29.7,32.7,32.7,33.7,34.7,35.0,40.6
X_2		21.33,19.81,19.20,21.94,24.68,25.29,20.11,22.86,21.03,22.55,25.90,26.21,21.64,
X_3		
X_4		
X_5		

Execute Erase Data

Scatter Plot Matrix Residual Plot Residual Q-Q Plot Graph Save Table Save

Practice 5.5.3 A health scientist randomly selected 20 people to determine the effect of smoking and obesity on their physical strength and examined the average daily smoking rate (x_1, number/day), the ratio of weight by height (x_2, kg/m), and the time to continue to exercise with a certain intensity (y, in hours). Draw a scatter plot matrix of this data and consider a regression model for this problem.

smoking rate x_1	ratio of weight by height x_2	time to continue to exercise y
24	53	11
0	47	22
25	50	7
0	52	26
5	40	22
18	44	15
20	46	9
0	45	23
15	56	15
6	40	24
0	45	27
15	47	14
18	41	13
5	38	21
10	51	20
0	43	24
12	38	15
0	36	24
15	43	12
12	45	16

[Ex] ⇨ DataScience ⇨ SmokingObesityExercis.csv.

In general, matrices and vectors are used to facilitate the expression of formulas and the calculation of expressions. For example, if there are k number of independent variables, the population multiple regression model at the observation point $i = 1, 2, \ldots, n$ is presented as follows.

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

Here $\mathbf{Y}, \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\epsilon}$ are defined as follows.

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ Y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ & & \cdots & & \\ & & \cdots & & \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \cdot \\ \beta_k \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \cdot \\ \cdot \\ \epsilon_n \end{bmatrix}$$

In a multiple regression analysis, it is necessary to estimate the $k+1$ number of regression coefficients $\beta_0, \beta_1, \ldots, \beta_k$ using samples. In this case, the least squares method, which minimizes the sum of the squared errors is also used. We find $\boldsymbol{\beta}$, which minimizes the following sum of the error squares.

$$S = \sum_{i=1}^{n} \epsilon_i^2 = \boldsymbol{\epsilon}'\boldsymbol{\epsilon} = (\mathbf{Y} - \mathbf{X}'\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}'\boldsymbol{\beta})$$

As in the simple linear regression, the above error sum of squares is differentiated with respect to $\boldsymbol{\beta}$ and then equate to zero, called a normal equation. The solution of the equation denoted as $\mathbf{b}$ which is called the least squares estimate of $\boldsymbol{\beta}$, should satisfy the following normal equation.

$$(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{y}$$

Therefore, if there exists an inverse matrix of $\mathbf{X}'\mathbf{X}$, the least squares estimator of $\boldsymbol{\beta}$, $\mathbf{b}$, is as follows.

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

(Note: Statistical packages uses a different formula, because the above formula causes large amount of computing error)

If the estimated regression coefficients are $\mathbf{b} = (b_0, b_1, \ldots, b_k)$, the estimate of the response variable Y is as follows.

$$\hat{Y}_i = b_0 + b_1 X_{i1} + \cdots + b_k X_{ik}$$

The residuals are as follows.

$$\begin{aligned} e_i &= Y_i - \hat{Y}_i \\ &= Y_i - (b_0 + b_1 X_{i1} + \cdots + b_k X_{ik}) \end{aligned}$$

using a vector notation, the residual vector $\mathbf{e}$ can be defined as follows.

$$\mathbf{e} = \mathbf{Y} - \mathbf{X}\mathbf{b}$$

The standardized residual error and coefficient of determination are also used to investigate the validity of the estimated regression line in the multiple regression analysis. In the simple linear regression analysis, the computational formula for these measures was given as a function of the residuals, i.e., the observed value of Y and its predicted value have nothing to do with the number of independent variables. Therefore, the same formula can be used in the multiple linear regression, and there is only a difference in the value of the degrees of freedom that each sum of squares has. In the multiple linear regression analysis, the standard error of residuals is defined as follows.

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

As in simple linear regression, s^2 is a statistic such as the residual mean squares (MSE).

The coefficient of determination is given in $R^2 = \frac{SSR}{SST}$ and its interpretation is as shown in the simple linear regression. The same formula defines the sum of squares as in the simple linear regression, and it can be divided with corresponding degrees of freedom as follows. The table of the analysis of variance is shown in Table 5.5.5.

$$\begin{array}{rl} \text{Sum of squares} & SST = SSE + SSR \\ \text{Degrees of freedom} & (n-1) = (n-k-1) + k \end{array}$$

Table 5.5.5 Analysis of variance table for multiple linear regression analysis

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

The F value in the above ANOVA table is used to test the significance of the regression equation, the null hypothesis is that all independent variables are not linearly related to the dependent variables.

$$\begin{array}{l} H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ H_1 : \text{At least one of } k \text{ number of } \beta_i\text{s is not equal to 0} \end{array}$$

Since F_0 follows F distribution with k and $(n-k-1)$ degrees of freedom under the null hypothesis, we can reject H_0 at the significance level α if $F_0 > F_{k,n-k-1;\alpha}$. Each β_i can also be tested, which is described in the following sections. (Also, 『eStat』 calculates the p-value for this test, so use this p-value to test. That is, if the p-value is less than the significance level, the null hypothesis is rejected.)

Parameters of interest in multiple linear regression, as in the simple linear regression, are the expected value of Y and each regression coefficient $\beta_0, \beta_1, \cdots, \beta_k$. The inference of these parameters $\beta_0, \beta_1, \cdots, \beta_k$ is made possible by obtaining a probability distribution of the point estimates b_i. Under the assumption that the error terms ϵ_i are independent and all have a distribution of $N(0, \sigma^2)$, it can be shown that the distribution of b_i is as follows.

$$b_i \sim N(\beta_i, c_{ii} \cdot \sigma^2), \quad i = 0, 1, 2, \ldots, k$$

The above c_{ii} is the i^{th} diagonal element of the $(k+1) \times (k+1)$ matrix $\mathbf{(X'X)^{-1}}$. In addition, using an estimate s^2 instead of a parameter σ^2, you can make inferences about each regression coefficient using the t distribution.

Inference on regression coefficient β_i

Point estimate: b_i

Standard error of estimate b: $SE(b_i) = \sqrt{c_{ii}} \cdot s$

Confidence interval of β_i: $b_i \pm t_{n-k-1;\alpha/2} \cdot SE(b_i)$

Testing hypothesis:

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

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

Rejection region:

$H_1 : \beta_i < \beta_{i0}$: $t < -t_{n-k-1;\ \alpha}$

$H_1 : \beta_i > \beta_{i0}$: $t > t_{n-k-1;\ \alpha}$

$H_1 : \beta_i \neq \beta_{i0}$: $|t| > t_{n-k-1;\ \alpha/2}$

Residual analysis of the multiple linear regression is the same as in the simple linear regression.

Example 5.5.4 For the tree data of [Example 5.5.3], obtain the least squares estimate of each coefficient of the proposed regression equation using 『eStat』 and apply the analysis of variance, test for goodness of fit and test for regression coefficients.

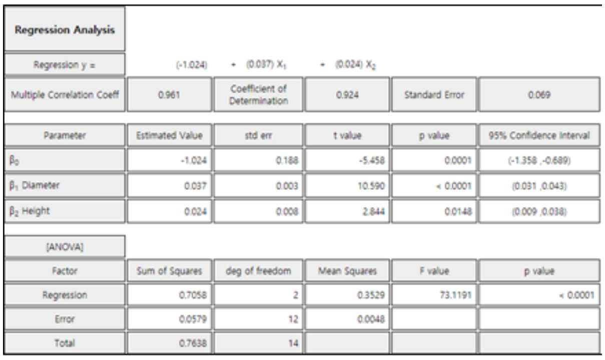
Answer

In the options window below the scatter plot matrix in <Figure 5.5.6>, click [Regression Analysis] button. Then, you can find the estimated regression line, ANOVA table, as shown in <Figure 5.5.8> in the Log Area. The estimated regression equation is as follows.

$$\hat{Y}_i = -1.024 + 0.037X_1 + 0.024X_2$$

In the above equation, 0.037 represents the increase of the volume of the tree when the diameter (X_1) increases 1(cm).

The p-value calculated from the ANOVA table in <Figure 5.5.8> at F value of 73.12 is less than 0.0001, so you can reject the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$ at the significance level α = 0.05. The coefficient of determination, R^2 = 0.924, implies that 92.4% of the total variances of the dependent variable are explained by the regression line. Based on the above two results, we can conclude that the diameter and height of the tree are quite useful in estimating the volume.



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			

<Figure 5.5.8> Result of Multiple Linear Regression

Since $SE(b_1) = 0.003,\ SE(b_2) = 0.008$ and $t_{12;0.025}$ = 2.179 from the result in <Figure 5.5.8>, the 95% confidence intervals for each regression coefficients can be calculated as follows. The difference between this result and the <Figure 5.5.8> due to the error in the calculation below the decimal point.

95% confidence interval for β_1 : 0.037 $\pm$ (2.179)(0.003) $\Rightarrow$ (0.029, ~0.045)
95% confidence interval for β_2 : 0.024 $\pm$ (2.179)(0.008) $\Rightarrow$ (0.006,~ 0.042)

In the hypothesis test of $H_0 : \beta_i = 0,\ H_1 : \beta_i \neq 0$, each p-value is less than the significance level of 0.05, so you can reject each null hypothesis.

Practice 5.5.4 Apply a multiple regression model using『eStat』on the regression model of [Practice 5.5.3]. Obtain the least squares estimate of each coefficient of the proposed regression equation and apply the analysis of variance, test for goodness of fit, and test for regression coefficients.

5.6 R practice

R practice

Let us practice testing hypothesess in this chapter using R commands. Since there is no package for the testing hypothesis in R, we have to calculate test statistic and p-value one by one. The distribution functions of the normal, t, and F distribution in R will be used to calculate the p-value.

Normal Distribution	
pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)	
q	quantile
mean	mean of normal distribution, default = 0.
sd	standard deviation of normal distribution, default = 1.
lower.tail	logical; if TRUE (default), probabilities are P[X≤x], otherwise P[X>x].
log.p	logical; if TRUE, probabilities p are given as log(p).

Student t Distribution	
pt(q, df, ncp, lower.tail = TRUE, log.p = FALSE)	
q	quantile
df	degree of freedom(>0, maybe non-integer). df = Inf is allowed.
ncp	non-centrality parameter δ; currently except for rt(), accurate only for abs(ncp) <= 37.62. If omitted, use the central t distribution.
lower.tail	logical; if TRUE (default), probabilities are P[X≤x], otherwise P[X>x].
log.p	logical; if TRUE, probabilities p are given as log(p).

F Distribution	
pf(q, df1, df2, ncp, lower.tail = TRUE, log.p = FALSE)	
q	quantile
df1, df2	degree of freedom. Inf is allowed.
ncp	non-centrality parameter. If omitted, the central F distribution.
lower.tail	logical; if TRUE (default), probabilities are P[X≤x], otherwise P[X>x].
log.p	logical; if TRUE, probabilities p are given as log(p).

```
# Example 5.2.1 Testing hypothesis for a population mean with known population standard deviation.
# Enter mu_0, n, xbar, and population standard deviation, then calculate test statistic and p-value.
```

```
> mu0 <- 1500
> n <- 30
> xbar <- 1555
> sigma <- 200
> teststat <- (xbar - mu0) / (sigma / sqrt(n))
> teststat
[1] 1.506237
> pvalue <- pnorm(teststat,0,1,lower.tail = FALSE)
> pvalue
[1] 0.06600317
```

```
# Since p-value is greater than the significance level 5%, we cannot reject the null hypothesis.
```

```
# Example 5.2.2 Testing hypothesis for a population mean with unknown population standard deviation.
# Enter mu_0, n, xbar, and sample standard deviation, then calculate test statistic and p-value.
```

```
> mu0 <- 250
> n <- 16
> xbar <- 253
> s <- 10
> teststat <- (xbar - mu0) / (s / sqrt(n))
> teststat
[1] 1.2
> pvalue <- pt(teststat, n-1, lower.tail = FALSE)
> pvalue
[1] 0.1243749
```

```
# Since p-value is greater than the significance level 5%, we cannot reject the null hypothesis.
```

```
# Example 5.3.1 Testing hypothesis for two populations means when population variances are equal.
# Enter n1, n2, xbar1, xbar2, s1, and s2, then calculate the pooled variance, test statistic and p-value.
```

```
> n1 <- 15
> n2 <- 14
> xbar1 <- 275
> xbar2 <- 269
> s1 <- 12
> s2 <- 10
> pooledvar <- ((n1-1)*s1^2 + (n2-1)*s2^2) / (n1+n2-2)
> pooledvar
[1] 122.8148
> teststat <- (xbar1 - xbar2) / sqrt(pooledvar/n1 + pooledvar/n2)
> teststat
[1] 1.456923
> pvalue <- 2 * pt(teststat, n1+n2-2, lower.tail = FALSE)
> pvalue
[1] 0.1566703
```

```
# Since p-value is greater than the significance level 5%, we cannot reject the null hypothesis.
```

We can use aov() function in R for the analysis of variance.

Fit an Analysis of Variance Model using aov()	
aov(formula, data = NULL, projections = FALSE, qr = TRUE, contrasts = NULL, ...)	
formula	A formula specifying the model.
data	A data frame in which the variables specified in the formula will be found. If missing, the variables are searched for in the standard way.

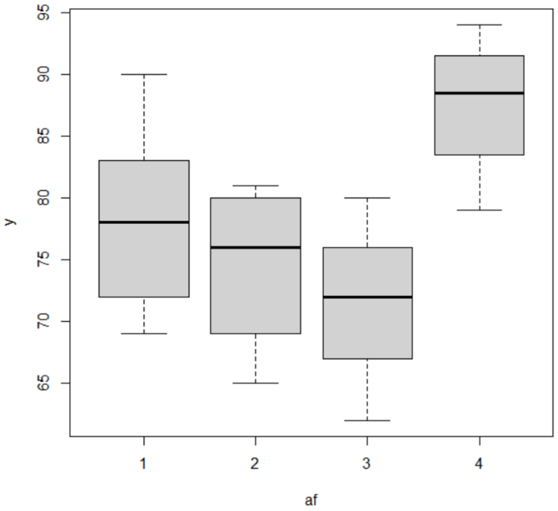
projections	Logical flag: should the projections be returned?
qr	Logical flag: should the QR decomposition be returned?
contrasts	A list of contrasts to be used for some of the factors in the formula. These are not used for any Error term, and supplying contrasts for factors only in the Error term will give a warning.

# Example 5.4.1 Testing hypothesis for several populations means; one-way ANOVA.
# Enter data and change the level as a factor (as.factor()).

```
> y <- c(81,75,69,90,72,83, 65,80,73,79,81,69, 72,67,62,76,80, 89,94,79,88)
> f <- c(rep(1,6), rep(2,6), rep(3,5), rep(4,4))
> af <- as.factor(f)
> ymean <- tapply(y, af, mean); ymean
```

```
       1        2        3        4 
78.33333 74.50000 71.40000 87.50000 
```

```
> boxplot(y ~ af)
```



<Figure 5.6.1> Box plot for each grade

```
> an1 <- aov(y ~ af)
> summary(an1)
```

```
Call:
   aov(formula = y ~ af)

Terms:
                      af Residuals
Sum of Squares  643.6333  839.0333
Deg. of Freedom        3        17

Residual standard error: 7.025304
Estimated effects may be unbalanced
> summary(an1)
            Df Sum Sq Mean Sq F value Pr(>F)  
af           3  643.6  214.54   4.347 0.0191 *
Residuals   17  839.0   49.35                 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

# Since p-value (Pr(>F) = 0.0191 is less than the significance level 5%, we reject the null hypothesis.

We can use lm() function in R for the regression analysis.

<table>
<tr><th colspan="2"><b>Fitting Linear Models using lm()
lm is used to fit linear models, including multivariate ones.</b></th></tr>
<tr><th colspan="2"><b>lm(formula, data, subset, weights, na.action, method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE, singular.ok = TRUE, contrasts = NULL, offset, ...)</b></th></tr>
<tr><td>formula</td><td>an object of class "formula": a symbolic description of the model to be fitted.</td></tr>
<tr><td>data</td><td>an optional data frame, list or environment containing the variables in the model.</td></tr>
<tr><td>subset</td><td>an optional vector specifying a subset of observations to be used in the fitting process.</td></tr>
<tr><td>weights</td><td>an optional vector of weights to be used in the fitting process. Should be NULL or a numeric vector. If non-NULL, weighted least squares is used with weights weights (that is, minimizing sum(w*e^2)); otherwise ordinary least squares is used.</td></tr>
<tr><td>na.action</td><td>a function which indicates what should happen when the data contain NAs. The default is set by the na.action setting of options, and is na.fail if that is unset. The ‘factory-fresh’ default is na.omit. Another possible value is NULL, no action. Value na.exclude can be useful.</td></tr>
<tr><td>method</td><td>the method to be used; for fitting, currently only method = "qr" is supported; method = "model.frame" returns the model frame (the same as with model = TRUE, see below).</td></tr>
<tr><td>model, x, y, qr</td><td>logicals. If TRUE the corresponding components of the fit (the model frame, the model matrix, the response, the QR decomposition) are returned.</td></tr>
<tr><td>singular.ok</td><td>logical. If FALSE (the default in S but not in R) a singular fit is an error.</td></tr>
<tr><td>contrasts</td><td>an optional list. See the contrasts.arg of model.matrix.default.</td></tr>
<tr><td>offsets</td><td>this can be used to specify an a priori known component to be included in the linear predictor during fitting. This should be NULL or a numeric vector or matrix of extents matching those of the response. One or more offset terms can be included in the formula instead or as well, and if more than one are specified their sum is used. See model.offset.</td></tr>
</table>

```
# Example 5.5.1 Regression analysis.
# Enter data and change the level as a factor (as.factor()).
```

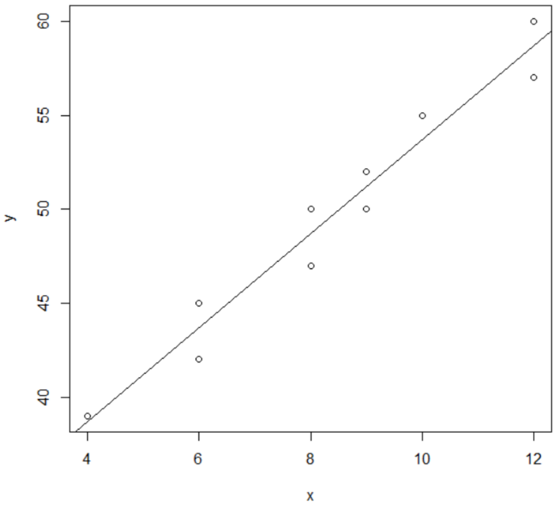
```
> x <- c(4,6,6,8,8,9,9,10,12,12)
> y <- c(39,42,45,47,50,50,52,55,57,60)
> rg <- lm(y ~ x); rg
```

```

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
     28.672        2.503
```

```
> plot(x, y); abline(rg)
```



<Figure 5.6.2> Simple linear regression

```
> summary(rg)
```

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.7119 -1.5695  0.5563  1.2931  1.3079 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  28.6722     1.6703   17.17 1.35e-07 ***
x             2.5033     0.1908   13.12 1.09e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.483 on 8 degrees of freedom
Multiple R-squared:  0.9556,    Adjusted R-squared:   0.95 
F-statistic: 172.1 on 1 and 8 DF,  p-value: 1.085e-06
```

```
# Example 5.5.3 Multiple regression analysis.
# Enter data and change the level as a factor (as.factor()).
```

```
> y <- c(0.291,0.291,0.288,0.464,0.532,0.557,0.441,0.515,0.603,0.628,0.956,0.775,0.727,0.704,1.084)
> x1 <- c(21.0,21.8,22.3,26.6,27.1,27.4,27.9,27.9,29.7,32.7,32.7,33.7,34.7,35.0,40.6)
> x2 <- c(21.33,19.81,19.20,21.94,24.68,25.29,20.11,22.86,21.03,22.55,25.90,26.21,21.64,19.50,21.94)
> rg2 <- lm(y ~ x1+x2); rg2
```

```

Call:
lm(formula = y ~ x1 + x2)

Coefficients:
(Intercept)           x1           x2  
   -1.02357      0.03697      0.02366  
```

```
> summary(rg2)
```

```
Call:
lm(formula = y ~ x1 + x2)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.09087 -0.03822 -0.02772  0.03320  0.15786 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.023572   0.187535  -5.458 0.000146 ***
x1           0.036968   0.003491  10.590 1.92e-07 ***
x2           0.023663   0.008321   2.844 0.014792 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06947 on 12 degrees of freedom
Multiple R-squared:  0.9242,	Adjusted R-squared:  0.9115 
F-statistic: 73.12 on 2 and 12 DF,  p-value: 1.902e-07
```

5.7 Exercise

5.1 A psychologist is working on physically disabled workers. Based on experience, the psychologist believed that the average social (relationship) score of these disabled workers was greater than 80. Twenty employees were sampled from the score population to obtain the following result:

99, 69, 91, 97, 70, 99, 72, 74, 74, 76, 96, 97, 68, 71, 99, 78, 76, 78, 83, 66

The psychologist wants to know if the average social score of the population is correct. Assume that the population follows a normal distribution and its standard deviation is 10. Test with a significance level of 0.05.

5.2 The following is the weights of the 10 employees randomly selected who are working in the shipping department of a wholesale food company.

154, 154, 186, 243, 159, 174, 183, 163, 192, 181 (unit pound)

Based on this data, can you say that the average weight of employees working in the shipping department is greater than 160 pound? Use the significance level of 5%.

5.3 In a large manufacturer, the company manager claims that the average adaptation score of all unskilled workers is greater than 60. Forty unskilled workers were selected randomly to check this claim, and their test scores of adaptation scores were as follows.

73 57 96 78 74 42 55 44 91 91 50 65 46 63 82 60 97 79 85 79
92 50 42 46 86 81 81 83 64 76 40 57 78 66 84 96 94 70 70 81

Test the hypothesis at the significance level of 0.05 whether the manager's argument is correct. What is the p-value?

5.4 An analyst studies two types of advertising methods (A and B) retailers tried. The variable is the sum spent on advertising over the past year. The following are the sample statistics extracted independently from retailers of each type. (Unit million USD)

Type A: $n_1 = 60,\ \overline{x}_1 = 14.8,\ s_1^2 = 0.180$
Type B: $n_1 = 70,\ \overline{x}_2 = 14.5,\ s_2^2 = 0.133$

From these data, can you conclude that type A retailers have invested more in advertising than type B retailers? (Significance level = 0.05)

5.5 Below are the entrance exam results for selecting new employees at a particular company. Test whether the male population mean is equal to the female population mean using the significance level 5%.

Male	Female

49 86 40 45 48 93 97 58 58 98
58 82 52 56 50 85 80 60 62 80
62 72 65 60 64 70 78 67 69 88
	60 72 66 65 75 78 62 64 74 58
	68 72 67 61 62 72 79 71 74 73

5.6 An industrial psychologist thinks that the significant factor that workers change jobs is self-esteem to workers' work. The scholar believes that workers who change jobs frequently (group A) have lower self-esteem than those who do not (group B). The score data on self-esteem were collected independently by sampling from each group.

Group A: 60 45 42 62 68 54 52 55 44 41
Group B: 70 72 74 74 76 91 71 78 78 83 50 52 66 65 53 52

Can this data support the psychologist's idea? Assume that the population scores are normally distributed and that the population variance is unknown but the same. (Significance level = 0.01)

5.7 A company used four exhibition methods to test customers' responses to new products (A, B, C, and D). Each exhibition method was used in nine stores by selecting 36 stores that met the company's criteria. The total sales in USD for the weekend are shown in the following table.

Method A	Method B	Method C	Method D
5	2	2	6
6	2	2	6
7	2	3	7
7	3	3	8
8	3	2	8
6	2	2	8
7	3	2	6
7	3	3	6
6	2	3	6

1) Draw a scatter plot of sales (y-axis) and exhibition method (x-axis). Mark the average sales of each exhibition method and connect them with a line.
2) Test that the sales by each exhibition method are different in the amount of sales with the 5% significance level. Can you conclude that one of the exhibition methods significantly affects on sales?

5.8 The following table shows mileages in km per liter of gasoline obtained from experiments to compare three brands of gasoline. In this experiment, seven cars of the same type were used in a similar situation to reduce the variation of the car.

Gasoline A	Gasoline B	Gasoline C
14	20	20

19	21	26
19	18	23
16	20	24
15	19	23
17	19	25
20	18	23

1) Calculate the average mileage of each gasoline brand. Draw a scatter plot of gas mileage (y-axis) and gasoline brand (x-axis) to compare.
2) From this data, test whether there are differences between gasoline brands for gas mileage with a 5% significance level.

5.9 The result of a survey on job satisfaction of three companies (A, B, and C) is as follows: Test whether the averages of job satisfaction of the three companies are different with a 5% significance level.

Company A	Company B	Company C
69	56	71
67	63	72
65	55	70
59	59	68
68	52	74
61	57	
66		

5.10 The following data shows studying time for a week (X) and the grade (Y) of six students.

Studying time (X)	Grade (Y)
15	2.0
28	2.7
13	1.3
20	1.9
4	0.9
10	1.7

1) Find a regression line.
2) Calculate a 95% confidence interval in the average score of a student who studies an average of 12 hours a week.
3) Test for hypothesis $H_0 : \beta = 0.10, H_1 : \beta < 0.10,$ (significance level α = 0.01).

5.11 An economist argues that there is a clear relationship between coffee and sugar prices. 'When people buy coffee, they will also buy sugar. Isn't it natural that the higher the demand, the higher the price?' We collected the following sample data to test his theory.

Year	Coffee price	Sugar Price
1985	0.68	0.245
1986	1.21	0.126

1987	1.92	0.092
1988	1.81	0.086
1989	1.55	0.101
1990	1.87	0.223
1991	1.56	0.212

1) Prepare a scatter plot with the coffee price on X axis and the sugar price on Y axis. Is this data true to this economist's theory?
2) Test this economist's theory by using a regression analysis.

5.12 A health scientist randomly selected 20 people to determine the effects of smoking and obesity on their physical strength and examined the average daily smoking rate (x_1, number/day), the ratio of weight by height (x_2, kg/m), and the time to continue to exercise with a certain intensity (y, in hours). Test whether smoking and obesity can affect your exercising time with a certain intensity. Apply a multiple regression model by using 『eStat』.

smoking rate x_1	ratio of weight by height x_2	Atime to continue to exercise y
24	53	11
0	47	22
25	50	7
0	52	26
5	40	22
18	44	15
20	46	9
0	45	23
15	56	15
6	40	24
0	45	27
15	47	14
18	41	13
5	38	21
10	51	20
0	43	24
12	38	15
0	36	24
15	43	12
12	45	16