Chapter IV

Estimation and Hypothesis Testing: Two Populations

D. 4. 1. (***Independent Samples***)
Two samples from two populations are said to be *independent* if the selection of one sample from one population does not affect the selection of the second sample from the second population. Otherwise, the samples are *dependent*. Such samples are called *paired* or *matched samples.*

Ex. 4. 1.
Suppose we want to estimate the difference between the mean salaries of all male and female executives. To do so, we draw two samples, one from the population of male executives and another from the population of female executives.
These two samples are *independent* because they are drawn from two different populations, and the samples have no effect on each other.

Ex. 4. 2.
Suppose we want to estimate the difference between the mean weights of all participants before and after a weight loss programme. To accomplish this, suppose we take a sample of 40 participants and measure their weights before and after the completion of this programme. Note that these two samples include the same 40 participants.
This is an example of two *dependent* samples.

Th. 4. 1.
Let

$\mu_i,\ i=1,2$: the mean of population i
$\sigma_i,\ i=1,2$: the standard deviation of population i
$n_i,\ i=1,2$: the size of population i
$\bar{x}_i,\ i=1,2$: the mean of the sample drawn from population i .

If the following conditions are fulfilled,

1. The two samples are independent
2. The standard deviations $\sigma_i,\ i=1,2,$ are known
3. At least one of the following two conditions are satisfied:
   i. Both samples are large ($n_i \geq 30,\ i=1,2$)
   ii. If at least one of the samples is small, then both populations from which the samples are drawn are normally distributed.

then the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is (approximately) normally distributed with its mean and standard deviations, respectively

(4. 1.) $$\mu_{\bar{x}_1-\bar{x}_2} = \mu_1 - \mu_2$$

and

(4. 2.) $$\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\,.$$

Th. 4. 2.
When using the normal distribution, the $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$ is

(4. 3.) $$(\bar{x}_1 - \bar{x}_2) \pm z \cdot \sigma_{\bar{x}_1-\bar{x}_2}\,.$$

The value of z is obtained from the normal distribution table for the given confidence level.

R. 4. 1
Note that in the real world, σ_1 and σ_2 are never known.

Ex. 4. 3.
A survey of low- and middle-income households showed that consumers aged 65 years and older had an average credit card debt of \$10235 and consumers in the 50- to 64-year age group had an average credit card debt of \$9342 at the time of the survey.
Suppose that these averages were based on random samples of 1200 and 1400 people for the two groups, respectively. Further assume that the population standard deviations for the two groups were \$2800 and \$5200, respectively.

1. What is the point estimate of the difference between the two population means?
2. Construct a 97% confidence interval for this difference.

Solution:
Let us refer to consumers aged 65 years and older as population 1 and those in the 50- to 64-year age group as population 2. The respective samples are samples 1 and 2.
Then we have:

For 65 and older group: $n_1 = 1200,\quad \bar{x}_1 = \$10235,\quad \sigma_1 = \$2800$

For 50 -65 age group: $n_2 = 1400,\quad \bar{x}_2 = \$9342,\quad \sigma_2 = \$2500$

1.

Point estimate of $\mu_1 - \mu_2 = 10235 - 9342 = 893$.

2.

$$\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{2800^2}{1200} + \frac{2500^2}{1400}} = 104.8695335$$

$$\sigma_{\bar{x}_1-\bar{x}_2} \in \left[(10235-9342)-2.17\cdot 104.8695335,\ (10235-9342)-2.17\cdot 104.8695335\right]$$

$$= \left[893-227.57,\ 893-227.57\ \right] = \left[665.43,\ 1120.57\right].$$

R. 4. 2. (***Hypothesis Testing About*** $\mu_1-\mu_2$ **)**
We distinguish two approaches:

1. The critical value approach
2. The $p-$value approach.

R. 4. 3. (***The Critical Value Approach***)

Step 1:
State the null and alternative hypothesis.

Step 2:
Select the distribution to use.
If the following conditions are fulfilled,

1. The two samples are independent
2. The standard deviations $\sigma_i,\ i=1,2,$ are known
3. At least one of the following two conditions are satisfied:

   i. Both samples are large ($n_i \geq 30,\ i=1,2$)
   ii. If at least one of the samples is small, then both populations from which the samples are drawn are normally distributed.

then the sampling distribution of $\bar{x}_1-\bar{x}_2$ is (approximately) normally distributed.

Step 3:
Find the value of the *critical statistic* $z_{critical}$ and determine the *rejection* and *nonrejection* regions.

Step 4:

Calculate the value of the *test statistic* for $\bar{x}_1-\bar{x}_2$:

$$z_{statistic} = \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sigma_{\bar{x}_1-\bar{x}_2}}.$$

Step 5:
Make a decision by comparing the critical statistic with the test statistic.

Ex. 4. 4.
Refer to Ex. 4. 3. about the average credit card debt for consumers of two age groups.
Test at the 1% significance level whether the population mean for the credit card debts for the two groups are different.

Solution:
Step 1:

$$H_0:\ \mu_1 - \mu_2 = 0, \qquad H_1 = \mu_1 \neq \mu_2 = 0\text{`}.$$

Step 2:
Here, the population standard deviations, σ_1 and σ_2, are known, and both samples are large.

Therefore, the distribution of $\bar{x}_1 - \bar{x}_2$ is approximately normal, and we use the normal distribution to perform the hypothesis analysis.

Step 3:

$$z_{critical} = \pm 2.576\ .$$

Step 4:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{2800^2}{1200} + \frac{2500^2}{1400}} = 104.8695335$$

$$z_{stat} = \frac{(10235 - 9342) - 0}{104.8695335} = 8.52$$

Step 5:
The value of the test statistic $z_{statistic} = 8.52$ falls in the rejection region, we reject the null hypothesis H_0.
Therefore, we conclude that the mean credit card debts for the two age groups are different.

<u>R. 4. 4.</u> (*The $p-$Value Approach*)

Step 1:
State the null and alternative hypothesis.

Step 2:
Select the distribution to use.
If the following conditions are fulfilled,

1. The two samples are independent
2. The standard deviations $\sigma_i,\ i = 1, 2,$ are known
3. At least one of the following two conditions are satisfied:

    i. Both samples are large ($n_i \geq 30,\ i = 1, 2$)
    ii. If at least one of the samples is small, then both populations from which the samples are drawn are normally distributed.

Step 3:

Calculate the value of the *test statistic* for $\bar{x}_1 - \bar{x}_2$:

$$z_{statistic} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}}$$

and find the $p - value$ for this z from the normal distribution table.

Step 4:
Reject the null hypothesis if $p - value < \alpha$; do not reject it otherwise.

Ex. 4. 4. (*revisited*)
Step 1:

$$H_0 : \ \mu_1 - \mu_2 = 0, \qquad H_1 = \mu_1 \neq \mu_2 = 0\grave{}.$$

Step 2:
Here, the population standard deviations, σ_1 and σ_2 , are known, and both samples are large.

Therefore, the distribution of $\bar{x}_1 - \bar{x}_2$ is approximately normal, and we use the normal distribution to perform the hypothesis analysis.

Step 3:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{2800^2}{1200} + \frac{2500^2}{1400}} = 104.8695335$$

$$z_{stat} = \frac{(10235 - 9342) - 0}{104.8695335} = 8.52, \qquad p - value = 0.$$

Step 4:

$$p - value = 0 < 0.01 = \alpha .$$

Therefore, we reject the null hypothesis and conclude that the mean credit card debts for the two age groups are different.

Th. 4. 3.
Let

$\mu_i, \ i = 1,2$: the mean of population i
$s_i, \ i = 1,2$: the standard deviation of the sample drawn from population i
$n_i, \ i = 1,2$: the size of population i
$\bar{x}_i, \ i = 1,2$: the mean of the sample drawn from population i .

If the following conditions are fulfilled,

1. The two samples are independent
2. The standard deviations $\sigma_i,\ i = 1, 2,$ are unknown, but they can be assumed to be equal
3. At least one of the following two conditions are fulfilled:

   i. Both samples are large
   ii. If the size of at least one sample is small, then both populations from which the samples are drawn are normally distributed

then the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is t- distributed with

(4. 5.) $$s_{\bar{x}_1 - \bar{x}_2} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

(4. 6.) $$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

(s_p is called *the pooled standard deviation for the two samples*).

Th. 4. 4.
When using the t-distribution, the $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$ is

(4. 7.) $$(\bar{x}_1 - \bar{x}_2) \pm t \cdot \sigma_{\bar{x}_1 - \bar{x}_2}.$$

The value of t is obtained from the t distribution table for the given confidence level.

Ex. 4. 5.
A consumer agency wanted to estimate the difference in the mean amounts of caffeine in two brands of coffee. The agency took a sample of 15 one-pound jars of Brand I coffee that showed the mean amount of caffeine in these jars to be 80 milligrams per jar with a standard deviation of 5 milligrams. Another sample of 12 one-pound jars of Brand II coffee gave a mean amount of caffeine equal to 77 milligrams per jar with a standard deviation of 6 milligrams.
Construct a 95% confidence interval for the difference between the mean amounts of caffeine in on-pound jars of these two brands of coffee. Assume that the two populations are normally distributed and that that the standard deviations of the two populations are equal.

Solution:

$$n_1 = 15, \quad \bar{x}_1 = 80 \text{ milligrams}, \quad s_1 = 5 \text{ milligrams}$$

$$n_2 = 12, \quad \bar{x}_2 = 77 \text{ milligrams}, \quad s_2 = 6 \text{ milligrams}$$

$$s_p = \sqrt{\frac{(15-1)\cdot 5^2 + (12-1)\cdot 6^2}{15+12-2}} = 5.46260011$$

$$s_{\bar{x}_1 - \bar{x}_2} = 5.46260011\sqrt{\frac{1}{15}+\frac{1}{12}} = 2.11565593$$

$$\sigma_{\bar{x}_1 - \bar{x}_2} \in \left[(80-77) - 2.060\cdot 2.11565593,\ (80-77) - 2.060\cdot 2.11565593\right]$$

$$= \left[3 - 4.36,\ 3 + 4.36\ \right] = \left[-1.36,\ 7.36\right].$$

Because the lower limit of the interval is negative, it is possible that the mean amount of caffeine is greater in the second brand than in the first brand of coffee.

R. 4. 5. (***The Critical Value Approach***)

Step 1:
State the null and alternative hypothesis.

Step 2:
Select the distribution to use.
If the following conditions are fulfilled,

1. The two samples are independent
2. The standard deviations $\sigma_i,\ i = 1,2,$ are unknown known
3. At least one of the following two conditions are satisfied:

    i. Both samples are large ($n_i \geq 30,\ i = 1,2$)
    ii. If at least one of the samples is small, then both populations from which the samples are drawn are normally distributed.

Consequently, we use the t distribution.

Step 3:
Find the value of the *critical statistic* $t_{critical}$ and determine the *rejection* and *nonrejection* regions.

Step 4:

Calculate the value of the *test statistic* for $\bar{x}_1 - \bar{x}_2$:

$$t_{statistic} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}}.$$

Step 5:
Make a decision by comparing the critical statistic with the test statistic.

Ex. 4. 6.
A sample of 14 cans of Brand I diet soda gave the mean number of calories of 23 per can with a standard deviation of 3 calories. Another sample of 16 cans of Brand II diet soda gave the mean number of calories of 25 per can with a standard deviation of 4 calories.
At the 1% significance level, can you conclude that the mean numbers of calories per can are different for these two brands of diet soda? Assume that the calories per can of diet soda are normally distributed for each of the two brands and that the standard deviations of the two populations are equal.

Solution:
Step 1:

$$H_0: \mu_1 - \mu_2 = 0, \qquad H_1 = \mu_1 \neq \mu_2 = 0`.$$

Step 2:
We use the t distribution, since all conditions stated in R. 5. 4. are fulfilled.

Step 3:

$$t_{critical} = \pm 2.763.$$

Step 4:

$$s_p = \sqrt{\frac{(14-1)\cdot 3^2 + (16-1)\cdot 4^2}{14+16-2}} = 3.57071421$$

$$s_{\bar{x}_1 - \bar{x}_2} = 3.57071421 \cdot \sqrt{\frac{1}{14} + \frac{1}{16}} = 1.30674760$$

$$t_{statistic} = \frac{(23-25)-0}{1.30674760} = -1.531$$

Step 5:
Because the value of the test statistic $t_{statistic} = -1.531$ falls in the nonrejection region, we fail to reject the null hypothesis H_0.

R. 4. 6. (*The $p-$Value Approach*)

Step 1:
State the null and alternative hypothesis.

Step 2:
Select the distribution to use.
If the following conditions are fulfilled,

1. The two samples are independent

2. The standard deviations $\sigma_i,\ i=1,2,$ are known
3. At least one of the following two conditions are satisfied:

   i. Both samples are large ($n_i \geq 30,\ i=1,2$)
   ii. If at least one of the samples is small, then both populations from which the samples are drawn are normally distributed.

Step 3:

Calculate the value of the *test statistic* for $\bar{x}_1 - \bar{x}_2$:

$$t_{statistic} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}}$$

and find the $p-value$ for this z from the t distribution table.

Step 4:
Reject the null hypothesis if $p-value < \alpha$; do not reject it otherwise.

Ex. 4. 6. (*revisited*)

Solution:
Step 1:

$$H_0:\ \mu_1 - \mu_2 = 0, \qquad H_1 = \mu_1 \neq \mu_2 = 0\text{`}.$$

Step 2:
We use the t distribution, since all conditions stated in R. 5. 4. are fulfilled.

Step 3:

$$s_p = \sqrt{\frac{(14-1)\cdot 3^2 + (16-1)\cdot 4^2}{14+16-2}} = 3.57071421$$

$$s_{\bar{x}_1 - \bar{x}_2} = 3.57071421 \cdot \sqrt{\frac{1}{14} + \frac{1}{16}} = 1.30674760$$

$$t_{statistic} = \frac{(23-25)-0}{1.30674760} = -1.531, \quad p-value = 0.1370.$$

Step 4:

$$p-value = 0.1370 \geq 0.01 = \alpha .$$

Therefore, we do not reject the null hypothesis.

Th. 4. 5.
If

1. The two samples are independent
2. The standard deviations of the two populations are unknown and unequal
3. At least one of the following two conditions is fulfilled:

    i. Both samples are large
    ii. If either one or both samples are small, then both populations from which the samples are drawn are normally distributed

then the t distribution is used to make inferences about $\mu_1 - \mu_2$, and the *degrees of freedom* for t distribution are given by

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

The number given by this formula is always rounded down for *df*.
The value of $s_{\bar{x}_1 - \bar{x}_2}$ is calculated as

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

Th. 5. 6.
When using the t distribution, the $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) \pm t \cdot \sigma_{\bar{x}_1 - \bar{x}_2}.$$

The value of t is obtained from the t distribution table for the given confidence level.

(Last updated: 10.12.19)