Chapter 10 Comparisons Involving Means

The logical flow of this chapter:

- **Estimation** of the difference between the means of two populations (10.1)
- **Hypothesis tests** about the difference between the means of two populations (10.2~10.3)
- Testing for the difference among \( k \) \((k \geq 2)\) population means.

10.1. Estimation of the difference between the means:

**Motivating example:**

Objective: We want to determine the difference between the mean ages of the customers shopping at the inner-city and the suburban.

\( \mu_1 \): the mean age of all customers who shop at the inner-city store

\( \mu_2 \): the mean age of all customers who shop at the suburban store

\( \sigma_1^2 \): the variance of the ages of all customers who shop at the inner-city store

\( \sigma_2^2 \): the variance of the ages of all customers who shop at the suburban store

We want to estimate the difference \( \mu_1 - \mu_2 \).

Let \( x_{1,1}, x_{1,2}, \cdots, x_{1,36} \) be the ages of 36 customers shopping at the inner-city store;

\( x_{2,1}, x_{2,2}, \cdots, x_{2,49} \) be the ages of 49 customers shopping at the suburban store.

Then,

\[
\bar{x}_1 = \frac{\sum_{i=1}^{36} x_{1,i}}{36} = 40, \quad \bar{x}_2 = \frac{\sum_{i=1}^{49} x_{2,i}}{49} = 35,
\]

and

\[
s_1 = \sqrt{\frac{\sum_{i=1}^{36} (x_{1,i} - \bar{x}_1)^2}{36 - 1}} = 9, \quad s_2 = \sqrt{\frac{\sum_{i=1}^{49} (x_{2,i} - \bar{x}_2)^2}{49 - 1}} = 10.
\]

\((The \ point \ estimate \ of \ \mu_1 - \mu_2) = \bar{x}_1 - \bar{x}_2 = 40 - 35 = 5.\)

\((The \ point \ estimate \ of \ \sigma_1^2) = s_1^2 = 9^2 = 81.\)

\((The \ point \ estimate \ of \ \sigma_2^2) = s_2^2 = 10^2 = 100.\)

**General Case:**

\( \mu_1 \): The mean of population 1

\( \mu_2 \): The mean of population 2

\( \sigma_1^2 \): the variance of population 1

\( \sigma_2^2 \): the variance of population 2
Let \( x_{1,1}, x_{1,2}, \ldots, x_{1,n_1} \) be the random sample from population 1
\( x_{2,1}, x_{2,2}, \ldots, x_{2,n_2} \) be the random sample from population 2.

Then,
\[
\bar{x}_1 = \frac{\sum_{i=1}^{n_1} x_{1,i}}{n_1}
\]
and
\[
\bar{x}_2 = \frac{\sum_{i=1}^{n_2} x_{2,i}}{n_2}
\]
be the sample means of population 1 and population 2, respectively, and
\[
s_1 = \sqrt{\frac{\sum_{i=1}^{n_1} (x_{1,i} - \bar{x}_1)^2}{n_1 - 1}}
\]
and
\[
s_2 = \sqrt{\frac{\sum_{i=1}^{n_2} (x_{2,i} - \bar{x}_2)^2}{n_2 - 1}}
\]
be the standard deviations of population 1 and population 2, respectively.

(The point estimate of \( \mu_1 - \mu_2 \))

**Important properties of \( \bar{X}_1 - \bar{X}_2 \):**
- \( \bar{X}_1 \) the sample statistic with possible value \( \bar{x}_1 \)
- \( \bar{X}_2 \) the sample statistic with possible value \( \bar{x}_2 \)

1. \( \mu_{\bar{X}_1-\bar{X}_2} = E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2 \).
2. \( \sigma^2_{\bar{X}_1-\bar{X}_2} = Var(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2 \)) = E[(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)]^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}

I. Large sample case \((n_1 \geq 30, n_2 \geq 30)\):

3. Sampling distribution of \( \bar{X}_1 - \bar{X}_2 \) \((n_1 \geq 30, n_2 \geq 30)\):
\[
\bar{X}_1 - \bar{X}_2 \approx N(\mu_{\bar{X}_1-\bar{X}_2}, \sigma^2_{\bar{X}_1-\bar{X}_2}) = N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)
\]
Note:
\[
\frac{(\bar{X}_1 - \bar{X}_2) - \mu_{\bar{X}_1-\bar{X}_2}}{\sigma_{\bar{X}_1-\bar{X}_2}} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \approx Z.
\]

Derivation of \((1 - \alpha) \cdot 100\%\) C.I. of \(\mu_{\bar{X}_1-\bar{X}_2} = \mu_1 - \mu_2:\)

\(1 - \alpha\)
\[
= P \left( |Z| \leq z_{\alpha/2} \right) = P \left( \left| \frac{(\bar{X}_1 - \bar{X}_2) - \mu_{\bar{X}_1-\bar{X}_2}}{\sigma_{\bar{X}_1-\bar{X}_2}} \right| \leq z_{\alpha/2} \right)
= P \left( \frac{\mu_{\bar{X}_1-\bar{X}_2} - (\bar{X}_1 - \bar{X}_2)}{\sigma_{\bar{X}_1-\bar{X}_2}} \leq z_{\alpha/2} \right) = P \left( -z_{\alpha/2} \leq \frac{\mu_{\bar{X}_1-\bar{X}_2} - (\bar{X}_1 - \bar{X}_2)}{\sigma_{\bar{X}_1-\bar{X}_2}} \leq z_{\alpha/2} \right)
= P \left( -z_{\alpha/2} \sigma_{\bar{X}_1-\bar{X}_2} \leq \mu_{\bar{X}_1-\bar{X}_2} - (\bar{X}_1 - \bar{X}_2) \leq z_{\alpha/2} \sigma_{\bar{X}_1-\bar{X}_2} \right)
= P \left( (\bar{X}_1 - \bar{X}_2) - z_{\alpha/2} \sigma_{\bar{X}_1-\bar{X}_2} \leq \mu_{\bar{X}_1-\bar{X}_2} \leq (\bar{X}_1 - \bar{X}_2) + z_{\alpha/2} \sigma_{\bar{X}_1-\bar{X}_2} \right)
= P \left( \mu_{\bar{X}_1-\bar{X}_2} = \mu_1 - \mu_2 \right)
\]

Thus, \((1 - \alpha) \cdot 100\%\) confidence interval is
\[
(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sigma_{\bar{X}_1-\bar{X}_2} = (\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}
\]

\((1 - \alpha) \cdot 100\%\) C.I. of \(\mu_{\bar{X}_1-\bar{X}_2} = \mu_1 - \mu_2 \) \((n_1 \geq 30, n_2 \geq 30):\)

- As \(\sigma_1, \sigma_2\) are known,
\[
(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sigma_{\bar{X}_1-\bar{X}_2} = (\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}
\]

\[
= \left[ \bar{x}_1 - \bar{x}_2 - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}, \bar{x}_1 - \bar{x}_2 + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} } \right]
\]

is a \((1 - \alpha) \cdot 100\%\) confidence interval of the population difference \(\mu_1 - \mu_2\).
As $\sigma_1, \sigma_2$ are unknown,

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

$$= \left[ \bar{x}_1 - \bar{x}_2 - \frac{z_\alpha}{2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \bar{x}_1 - \bar{x}_2 + \frac{z_\alpha}{2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right]$$

is a $(1 - \alpha) \cdot 100\%$ confidence interval of the population difference $\mu_1 - \mu_2$.

Note:

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

is the standard error of $\bar{X}_1 - \bar{X}_2$, i.e., the estimate of $\sigma_{\bar{x}_1 - \bar{x}_2}$.

**Motivating Example (continue):**

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{9^2}{36} + \frac{10^2}{49}} = 2.07.$$

A 95\% ($\alpha = 0.05$) confidence interval for $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) \pm \frac{z_\alpha}{2} s_{\bar{x}_1 - \bar{x}_2} = (40 - 35) \pm 1.96 \cdot 2.07 = [0.94, 9.06].$$

II. Small sample case ($n_1 < 30, n_2 < 30$):

Two assumptions are made:

1. Both populations have normal distribution.
2. The variance of the populations are equal ($\sigma_1^2 = \sigma_2^2 = \sigma^2$)

**Pooled estimate of $\sigma^2$ and $\sigma^2_{\bar{x}_1 - \bar{x}_2}$**:

The pooled estimate of $\sigma^2$, denoted by $s_p^2$,

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

$$= \frac{\sum_{i=1}^{n_1}(x_{1,i} - \bar{x}_1)^2 + \sum_{i=1}^{n_2}(x_{2,i} - \bar{x}_2)^2}{n_1 + n_2 - 2}$$

is a weighted average of the two sample variance $s_1^2$ and $s_2^2$.

The estimate of

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} = \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)s_p^2}$$
is

\[ s_{\bar{x}_1-\bar{x}_2}^* = \sqrt{\frac{s_p^2}{n_1} + \frac{1}{n_2}}. \]

\((1 - \alpha) \cdot 100\% \text{ C.I. of } \mu_{\bar{x}_1-\bar{x}_2} = \mu_1 - \mu_2 \ (n_1 < 30, n_2 < 30):\)

\[
(\bar{x}_1 - \bar{x}_2) \pm t_{n_1+n_2-2, \alpha/2} s_{\bar{x}_1-\bar{x}_2}^* = (\bar{x}_1 - \bar{x}_2) \pm t_{n_1+n_2-2, \alpha/2} \sqrt{\frac{s_p^2}{n_1} + \frac{1}{n_2}}
\]

\[
= \left[ (\bar{x}_1 - \bar{x}_2) - t_{n_1+n_2-2, \alpha/2} \sqrt{\frac{s_p^2}{n_1} + \frac{1}{n_2}} \right],
\]

\[
(\bar{x}_1 - \bar{x}_2) + t_{n_1+n_2-2, \alpha/2} \sqrt{\frac{s_p^2}{n_1} + \frac{1}{n_2}}
\]

is a \((1 - \alpha) \cdot 100\%\) confidence interval of the population difference \(\mu_1 - \mu_2\).

Note:

\[
\frac{(\bar{X}_1 - \bar{X}_2) - \mu_{\bar{x}_1-\bar{x}_2}}{S_{\bar{x}_1-\bar{x}_2}^*} = \frac{(\bar{X}_1 - \bar{X}_2) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{s_p^2}{n_1} + \frac{1}{n_2}}} \sim T(n_1 + n_2 - 2),
\]

where

\[
S_p^2 = \frac{\sum_{i=1}^{n_1}(X_{1,i} - \bar{X}_1)^2 + \sum_{i=1}^{n_2}(X_{2,i} - \bar{X}_2)^2}{n_1 + n_2 - 2}
\]

Then, \((1 - \alpha) \cdot 100\%\) confidence interval in small sample case can be derived analogous to the one in large sample case.

Example 1:

Let

\(\mu_1: \text{The mean number of productions in factory A}\)

\(\mu_2: \text{The mean number of productions in factory B}\)

\([
\bar{x}_1 = 1000, n_1 = 12, s_1 = 150,
\bar{x}_2 = 920, n_2 = 10, s_2 = 120.
\]

Please find a 90% confidence interval for \(\mu_1 - \mu_2\).

[Solution:]

\[
s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{11 \cdot 150^2 + 9 \cdot 120^2}{12 + 10 - 2} = 18855.
\]
Thus,

\[
\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{s_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)} = \sqrt{18855 \left( \frac{1}{12} + \frac{1}{10} \right)} = 58.79.
\]

Then, a 90% confidence interval for \( \mu_1 - \mu_2 \) is

\[
(x_1 - x_2) \pm t_{n_1+n_2-2, \alpha/2} \cdot \sigma_{\bar{x}_1 - \bar{x}_2} = (1000 - 920) \pm 1.725 \cdot 58.79
\]

\[
= 80 \pm 101.41 = [-21.41, 181.41].
\]