12.1 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***.

μ_1 : **the mean age of all customers who shop at the *inner-city* store**

μ_2 : **the mean age of all customers who shop at the *suburban* store**

σ_1^2 : **the variance of the ages of all customers who shop at the *inner-city* store**

σ_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

$x_{1,1}, x_{1,2}, \ldots, x_{1,36}$: the ages of 36 customers shopping at the ***inner-city*** store.

$x_{2,1}, x_{2,2}, \ldots, x_{2,49}$: 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

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

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:

μ_1 : **the mean of population 1**

μ_2 : **the mean of population 2**

σ_1^2 : **the variance of population 1**

σ_2^2 : **the variance of population 2**

Let

$x_{1,1}, x_{1,2}, \ldots, x_{1,n_1}$: the random sample from population 1

$x_{2,1}, x_{2,2}, \ldots, x_{2,n_2}$: the random sample from population 2.

Then,

$$\bar{x}_1 = \frac{\sum_{i=1}^{n_1} x_{1,i}}{n_1} :$$ the sample mean of population 1.

$$\bar{x}_2 = \frac{\sum_{i=1}^{n_2} x_{2,i}}{n_2} :$$ the sample mean of population 2

and

$$s_1 = \sqrt{\frac{\sum_{i=1}^{n_1} \left(x_{1,i} - \bar{x}_1\right)^2}{n_1 - 1}}$$: the standard deviation of population 1.

$s_2 = \sqrt{\dfrac{\sum_{i=1}^{n_2}\left(x_{2,i} - \bar{x}_2\right)^2}{n_2 - 1}}$: the standard deviation of population 2.

The point estimate of $u_1 - u_2 = \bar{x}_1 - \bar{x}_2$

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 $\bar{x}_2$

Then,

$$\mu_{\bar{X}_1 - \bar{X}_2} = E\left(\bar{X}_1 - \bar{X}_2\right) = \mu_1 - \mu_2$$

and

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

I. Large Sample Case ($n_1 \geq 30, n_2 \geq 30$):

Sampling Distribution of $\bar{X}_1 - \bar{X}_2$ ($n_1 \geq 30, n_2 \geq 30$):

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

$$\Leftrightarrow \frac{\left(\bar{X}_1 - \bar{X}_2\right) - \mu_{\bar{X}_1 - \bar{X}_2}}{\sigma_{\bar{X}_1 - \bar{X}_2}} = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \approx N(0,1)$$

Derivation of $(1-\alpha)\times 100\%$ confidence interval:

$$
\begin{aligned}
1-\alpha &= P\left(|Z| \le z_{\alpha/2}\right) \\
&\approx P\left(\left|\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sigma_{\bar{X}_1-\bar{X}_2}}\right| \le z_{\alpha/2}\right) \\
&= P\left(\left|\frac{(\mu_1-\mu_2)-(\bar{X}_1-\bar{X}_2)}{\sigma_{\bar{X}_1-\bar{X}_2}}\right| \le z_{\alpha/2}\right) \\
&= P\left(-z_{\alpha/2} \le \frac{(\mu_1-\mu_2)-(\bar{X}_1-\bar{X}_2)}{\sigma_{\bar{X}_1-\bar{X}_2}} \le z_{\alpha/2}\right) \\
&= P\left(-z_{\alpha/2}\sigma_{\bar{X}_1-\bar{X}_2} \le (\mu_1-\mu_2)-(\bar{X}_1-\bar{X}_2) \le z_{\alpha/2}\sigma_{\bar{X}_1-\bar{X}_2}\right) \\
&= P\left((\bar{X}_1-\bar{X}_2)-z_{\alpha/2}\sigma_{\bar{X}_1-\bar{X}_2} \le \mu_1-\mu_2 \le (\bar{X}_1-\bar{X}_2)+z_{\alpha/2}\sigma_{\bar{X}_1-\bar{X}_2}\right)
\end{aligned}
$$

Thus, $(1-\alpha)\times 100\%$ confidence interval is

$$
\begin{aligned}
(\bar{x}_1-\bar{x}_2)\pm z_{\alpha/2}\sigma_{\bar{X}_1-\bar{X}_2} &\equiv (\bar{x}_1-\bar{x}_2)\pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \\
&\equiv \left[\bar{x}_1-\bar{x}_2-z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}, \bar{x}_1-\bar{x}_2+z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right]
\end{aligned}
$$

$(1-\alpha)\times 100\,\%$ **confidence interval (** $n_1 \ge 30, n_2 \ge 30$ **):**

- **As** σ_1, σ_2 **are known,**

$$
\begin{aligned}
(\bar{x}_1-\bar{x}_2)\pm z_{\alpha/2}\sigma_{\bar{X}_1-\bar{X}_2} &\equiv (\bar{x}_1-\bar{x}_2)\pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \\
&\equiv \left[\bar{x}_1-\bar{x}_2-z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}, \bar{x}_1-\bar{x}_2+z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right]
\end{aligned}
$$

is a $(1-\alpha)\times 100\%$ **confidence interval estimate of the population**

difference $\mu_1 - \mu_2$.

- **As** σ_1, σ_2 **are unknown,**

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

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

is a $(1-\alpha)\times 100\%$ **confidence interval estimate of the population**

difference $\mu_1 - \mu_2$.

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% confidence interval for $\mu_1 - \mu_2$ is

$$\left(\bar{x}_1 - \bar{x}_2\right) \pm z_{\alpha/2} s_{\bar{X}_1 - \bar{X}_2} = (40 - 35) \pm z_{0.025} \cdot 2.07 = 5 \pm 1.96 \cdot 2.07 = 5 \pm 4.06 = [0.94, 9.06]$$

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 σ^2 **and** $\sigma^2_{\bar{X}_1 - \bar{X}_2}$ **:**

The pooled estimate of σ^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

$$\sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} = \sqrt{\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

is

$$s^*_{\bar{X}_1-\bar{X}_2} = \sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}.$$

$(1-\alpha)\times 100\%$ **confidence interval (** $n_1 < 30, n_2 < 30$ **):**

$$(\bar{x}_1-\bar{x}_2) \pm t_{n_1+n_2-2,\alpha/2}\, s^*_{\bar{X}_1-\bar{X}_2} \equiv (\bar{x}_1-\bar{x}_2) \pm t_{n_1+n_2-2,\alpha/2}\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

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

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

Note:

$$\frac{\bar{X}_1-\bar{X}_2-(u_1-u_2)}{s^*_{\bar{X}_1-\bar{X}_2}} = \frac{\bar{X}_1-\bar{X}_2-(\mu_1-\mu_2)}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \sim T(n_1+n_2-2)$$

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

Example:

Let

μ_1 : **the mean balance of checking account in *Chekry Grove* bank**

μ_2 : **the mean balance of checking account in *Beechmont* bank.**

$$\bar{x}_1 = 1000\,,\ n_1 = 12\,,\ s_1 = 150$$
$$\bar{x}_2 = 920\,,\ n_2 = 10\,,\ s_2 = 120$$

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,

$$s^*_{\bar{X}_1-\bar{X}_2} = \sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = \sqrt{18855\left(\frac{1}{12}+\frac{1}{10}\right)} = 58.79\,.$$

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

$$(\bar{x}_1-\bar{x}_2)\pm t_{n_1+n_2-2,\alpha/2}\, s^*_{\bar{X}_1-\bar{X}_2} = (1000-920)\pm t_{20,0.05}\cdot 58.79 = 80\pm 1.725\cdot 58.79$$
$$= 80\pm 101.41$$
$$= [-21.41, 181.41]$$

Online Exercise:

Exercise 12.1.1
Exercise 12.1.2