Tests to Compare Two Means

- Two-Sample Problems and a Comparison to Matched Pairs
- Comparing Two Means when $\sigma$ Known
- Comparing Two Means when $\sigma$ is Unknown
- Pooled Two-Sample $t$ Procedures

Two-sample problems

- The goal of two-sample inference is to compare the responses in two groups.
- Each group is considered to be a sample from a distinct population.
- The responses in each group are independent of those in the other group (in addition to being independent of each other).

For example, suppose we have a SRS of size $n_1$ drawn from a $N(\mu_1, \sigma_1)$ population and an independent SRS of size $n_2$ drawn from a $N(\mu_2, \sigma_2)$ population.

The first sample might be first year UC Berkeley male graduates and the second sample might be first year UC Berkeley female graduates. We are interested in comparing the salaries of the two groups, and we might test $H_0 : \mu_1 = \mu_2$ against $H_a : \mu_1 \neq \mu_2$.

How is this different from the matched pairs design? In a matched pairs design, subjects are matched in pairs and the outcomes are compared within each matched pair, e.g., we compute the difference for each pair.

For the two-sample problem:

1. There is no matching of the units in the two samples.
2. The two samples may be of different size.

Comparing Two Means when $\sigma$ is Known

Suppose we have a SRS of size $n_1$ drawn from a $N(\mu_1, \sigma_1)$ population (with sample mean $\bar{x}_1$) and an independent SRS of size $n_2$ drawn from a $N(\mu_2, \sigma_2)$ population (with sample mean $\bar{x}_2$). Suppose $\sigma_1$ and $\sigma_2$ are known.

What is the distribution of $\bar{X}_1 - \bar{X}_2$?

What is the mean of $\bar{X}_1 - \bar{X}_2$?

What is the variance of $\bar{X}_1 - \bar{X}_2$?
The two-sample z-statistic is
\[ Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1) \]

- A \((1 - \alpha)\) CI for \(\mu_1 - \mu_2\) is given by
\[ (\bar{x}_1 - \bar{x}_2) \pm z^* \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \]
where \(z^*\) is the \(\alpha/2\) critical value of the standard normal distribution.

- To test the hypothesis \(H_0: \mu_1 = \mu_2\) (or equivalently \(\mu_1 - \mu_2 = 0\)), we use
\[ Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1) \quad \text{under } H_0 \]
The p-value is calculated as before:
\[
\begin{align*}
H_a: \mu &> \mu_0 \quad P(Z \geq z) \\
H_a: \mu &< \mu_0 \quad P(Z \leq z) \\
H_a: \mu &\neq \mu_0 \quad 2P(Z \geq |z|)
\end{align*}
\]

### Comparing Two Means with \(\sigma\) Unknown

The two-sample t-statistic is
\[ T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim \text{t}_k \]
This statistic does not have an exact \(t\) distribution, it has an approximate \(t_k\) distribution with \(k = \min(n_1 - 1, n_2 - 1)\)

- A \((1 - \alpha)\) CI for \(\mu_1 - \mu_2\) is given by
\[ (\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \]
where \(t^*\) is the \(\alpha/2\) critical value of the \(t_k\) distribution.

- To test the hypothesis \(H_0: \mu_1 = \mu_2\), we use
\[ T = \frac{X_1 - X_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim \text{t}_k \quad \text{under } H_0 \]

### Pooled two-sample t procedures

In the previous procedure, we assumed that \(\sigma_1 \neq \sigma_2\). What if we have reason to believe \(\sigma_1 = \sigma_2 = \sigma\) (even though we don’t know either value)?

We know that the variance of \(X_1 - X_2\) is
\[
\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)
\]
and the SD is \(\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\)

To estimate \(\sigma^2\), we can gain information (i.e. power) by pooling the two samples together for estimating the variance:
\[
S_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}
\]
The \(t\) statistic is then
\[ T = \frac{(X_1 - X_2) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \text{t}_{(n_1 + n_2 - 2)} \]
Example

Weight gains (in kg) of babies from birth to age one year are measured. All babies weighed approximately the same at birth. Group A are babies that were breast fed and Group B are babies that were given formula.

<table>
<thead>
<tr>
<th>Group A</th>
<th>5</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>6</th>
<th>7</th>
<th>10</th>
<th>8</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group B</td>
<td>9</td>
<td>10</td>
<td>8</td>
<td>6</td>
<td>8</td>
<td>7</td>
<td>9</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Assume that the samples are randomly selected from independent normal populations. Is there any difference between the true means of the two groups?

i) Assume \( \sigma_1 = \sigma_2 = 1.5 \) is known. What type of test is used?

ii) Assume \( \sigma_1 \) and \( \sigma_2 \) are unknown and unequal. What type of test is used?

iii) Assume \( \sigma_1 \) and \( \sigma_2 \) are unknown but equal. What type of test is used.

State the hypothesis:

\[ H_0 : \mu_1 = \mu_2 \quad H_a : \mu_1 \neq \mu_2 \]

where \( \mu_1 \) is the true population mean of the Group A and \( \mu_2 \) is the true population mean of Group B.

\[
\begin{align*}
\bar{x}_1 &= 7.33 \\
s_1 &= 1.58 \\
n_1 &= 9
\end{align*}
\[
\begin{align*}
\bar{x}_2 &= 8.14 \\
s_2 &= 1.35 \\
n_2 &= 7
\end{align*}
\]

i) Assume \( \sigma_1 = \sigma_2 = 1.5 \) is known. Then, the two-sample z statistic is

\[
\begin{align*}
z &= \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \\
&= \frac{7.33 - 8.14}{\sqrt{\frac{1.5^2}{9} + \frac{1.35^2}{7}}} \\
&= \frac{-0.81}{1.5 \times \sqrt{\frac{1}{9} + \frac{1}{7}}} \\
&= -1.07
\end{align*}
\]

The two-sided p-value is

\[
2P(Z \geq |z|) = 2P(Z \geq 1.07) = 0.28
\]

where \( Z \sim N(0, 1) \).

So there is no difference between the true population mean of these two group at the significance level 0.1.

A 90% confidence interval for \( \mu_1 - \mu_2 \) is:

\[
(\bar{x}_1 - \bar{x}_2) \pm z^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}
\]

\[
= (7.33 - 8.14) \pm 1.645 \times 1.5 \times \sqrt{\frac{1}{9} + \frac{1}{7}}
\]

\[
= (-2.05, 0.43)
\]

As expected, the 90% confidence interval covers 0.
ii) Assume \( \sigma_1 \) and \( \sigma_2 \) are unknown and unequal. Then, the two-sample \( t \) statistic is

\[
t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} = \frac{7.33 - 8.14}{\sqrt{1.58^2/9 + 1.35^2/7}} = -1.10
\]

The two-sided \( p \)-value is

\[
2P(T \geq |z|) = 2P(T \geq 1.10) = 0.31
\]

where \( T \sim t_6 \).

A 90% confidence interval for \( \mu_1 - \mu_2 \) is given by

\[
(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{s_1^2/n_1 + s_2^2/n_2}
\]

\[
= (7.33 - 8.14) \pm 1.94 \times \sqrt{1.58^2/9 + 1.35^2/7}
\]

\[
= (-2.23, 0.61)
\]

where \( P(|T| < t^*) = 0.90 \). That is, \( P(T > t^*) = 0.05 \)

iii) Assume \( \sigma_1 \) and \( \sigma_2 \) are unknown but equal.

The pooled two-sample estimator of \( \sigma \) is

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

\[
= \sqrt{\frac{(9 - 1) \times 1.58^2 + (7 - 1) \times 1.35^2}{9 + 7 - 2}}
\]

\[
= 1.54
\]

Thus, the pooled two-sample \( t \) statistic is

\[
t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{7.33 - 8.14}{1.54 \sqrt{\frac{1}{9} + \frac{1}{7}}}
\]

\[
= -1.04
\]

The two-sided \( p \)-value is given by

\[
2P(T \geq |t|) = 2P(T \geq 1.04) = 0.32 \quad \text{where} \quad T \sim t_{13}
\]

A 90% confidence interval for \( \mu_1 - \mu_2 \) is

\[
(\bar{x}_1 - \bar{x}_2) \pm t^* s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}
\]

\[
= (7.33 - 8.14) \pm 1.77 \times 1.54 \times \sqrt{\frac{1}{9} + \frac{1}{7}}
\]

\[
= (-2.18, 0.56)
\]

Where \( P(|T| < t^*) = 0.90 \). That is, \( P(T > t^*) = 0.05 \).