Two-sample t-test (pooled)

Setup

Consider independent random variables

$$Y_{ij}, i = 1, 2; j = 1, 2, \dots, n_i,$$

such that

$$Y_{ij} \sim N(\mu_i, \underline{\sigma^2}).$$

Sample 1: $y_{11}, y_{12}, \dots, y_{1n_1}$ from $N(\mu_1, \sigma^2)$

Sample 2: $y_{21}, y_{22}, \dots, y_{2n_2}$ from $N(\mu_2, \sigma^2)$

We wish to make inference on $\mu_1 - \mu_2$.

Estimation of $\mu_1 - \mu_2$

Let

$$\bar{Y}_i = \frac{1}{n_i}\sum_{j=1}^{n} Y_{ij}\,, \text{for } i = 1, 2,$$

$\bar{Y}_1$ is the BLUE for μ_1.
$\bar{Y}_2$ is the BLUE for μ_2.
$\bar{Y}_1$ and $\bar{Y}_2$ are independent.

then $\bar{Y}_1 \sim N\left(\mu_1, \frac{\sigma^2}{n_1}\right)$, $\bar{Y}_2 \sim N\left(\mu_2, \frac{\sigma^2}{n_2}\right)$

$$\bar{Y}_1 - \bar{Y}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}\right).$$

$$Z = \frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}} = \frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0,1)$$

Estimation of $\mu_1 - \mu_2$

When σ^2 is known and we want to test

$$H_0 \colon \mu_1 - \mu_2 = 0,$$
$$H_1 : \mu_1 - \mu_2 \neq 0,$$

$H_0 : \mu_1 = \mu_2$

$H_a : \mu_1 \neq \mu_2$

the test statistic is

$$Z = \frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$\sim N(0,1)$ under H_0.

We reject H_0 iff

$$|Z| \geq z_{\alpha/2}.$$

Estimation of $\mu_1 - \mu_2$

The P-value in this case is

$$P(|Z| \geq |z|)$$

$Z \sim N(0,1)$ observed test statistic

The corresponding interval is $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$ is given by

$$\boxed{\bar{Y}_1 - \bar{Y}_2 \pm z_{\alpha/2}\, \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Estimation of σ^2

When σ^2 is unknown, we need to find an estimator of the common variance σ^2. $\sigma_1^2 = \sigma_2^2 = \sigma^2$

An unbiased estimator of the common variance σ^2 can be obtained by pooling the sample data to obtain the pooled estimator S_p^2.

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$$

$$= \frac{\sum_{i=1}^{n_1}(Y_{i1}-\bar{Y}_1)^2 + \sum_{j=1}^{n_2}(Y_{j2}-\bar{Y}_2)^2}{n_1+n_2-2}$$

If $n_1 = n_2$, then S_p^2 is the average of S_1^2 and S_2^2.

If $n_1 \neq n_2$, then S_p^2 is the weighted average of S_1^2 and S_2^2.

e.g. if $n_1 > n_2$, then S_1^2 will have a larger weighting than S_2^2.

Definition 2.9

The pooled estimator is given by

$$\boxed{S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}},$$

where

$$S_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n}\left(Y_{ij} - \bar{Y}_i\right)^2.$$

It is an unbiased estimator of the common variance σ^2, and

$$\boxed{\frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2}} \sim \chi^2_{n_1+n_2-2}.$$

We will prove this distributional result in the next Tutorial.
We will also prove that $E[S_p^2] = \sigma^2$.

Pooled two-sample t-test

As S_p^2 is independent of $\hat{Y}_i$, $i = 1{,}2$, it follows that

$$T = \boxed{\frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}} \sim t_{n_1+n_2-2}.$$

From this we can get the hypothesis test and confidence interval.

We had $Z = \dfrac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0,1)$

and $W = \dfrac{(n_1+n_2-2)\,S_p^2}{\sigma^2} \sim \chi^2_{n_1+n_2-2}$

Z and W are independent

$$T = \frac{Z}{\sqrt{\frac{W}{n_1+n_2-2}}} = \frac{\dfrac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\cancel{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}}{\sqrt{\dfrac{\frac{\cancel{(n_1+n_2-2)}\,S_p^2}{\cancel{\sigma^2}}}{\cancel{n_1+n_2-2}}}} = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

Confidence interval for $\mu_1 - \mu_2$

The interval

$$\boxed{\bar{Y}_1 - \bar{Y}_2 \pm t_{n_1+n_2-2,\frac{\alpha}{2}} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

is a $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$.

We will derive this in the next tutorial.

Hypothesis test

To test

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

the test statistic is

$$T = \frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \sim t_{n_1+n_2-2} \text{ under } H_0$$

We reject H_0 iff

$$\boxed{|T| \geq t_{n_1+n_2-2, \frac{\alpha}{2}}.} \quad \text{where } T \sim t_{n_1+n_2-2}$$

P-value is $P(|T| \geq |t|)$

observed test statistic

Example 2.12

A study was conducted to compare the 100m running times of men and women. Independent random samples of 9 men and 9 women were employed in the experiment. The results are shown in the table below. Do the data represent sufficient evidence to suggest a difference between the true mean 100m running times for men and women? Use $\alpha = 0.05$.

Men	Women
$n_1 = 9$	$n_2 = 9$
$\bar{y}_1 = 31.56$ seconds	$\bar{y}_2 = 35.22$ seconds
$s_1^2 = 20.0275$	$s_2^2 = 24.445$

Example 2.12 Solution

Let μ_1 = true mean 100m running time for men

μ_2 = true mean 100m running time for women

$H_0: \mu_1 = \mu_2$ against $H_a: \mu_1 \neq \mu_2$

$$T = \frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$= \frac{31.56 - 35.22}{\sqrt{22.236\left(\frac{1}{9} + \frac{1}{9}\right)}}$$

$$\approx -1.65$$

$$S_p^2 = \frac{(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2}{n_1 + n_2 - 2}$$

$$= \frac{8(20.0275) + 8(24.445)}{9 + 9 - 2}$$

$$\approx 22.236$$

The critical value is $t_{16, 0.025} \approx 2.120$. qt(0.975, 16)

We reject H_0 if $|T| > 2.120$.

As $|T| = 1.65 < 2.120$, we fail to reject H_0 in this case.

There is insufficient evidence to suggest that men and women have different 100m mean running time, at the $\alpha = 0.05$ significance level.