

2つの群の平均値の検定（等分散を前提）

2つの群のデータを

第1群 $\{x_i\}_{i=1}^{N_1}$ 、第2群 $\{y_i\}_{i=1}^{N_2}$

とした時、2つの群の平均値が等しいかどうかを検定する。

但し 以下を前提とする。

- (1) 2つの群のデータは正規分布に従う。
- (2) 2つの群のデータの分散は等しい。(等分散性)
- (3) 2つの群のデータには対応がない。

μ_x を第1群の母平均、 μ_y を第2群の母平均、
 \bar{X} を第1群の標本平均、 \bar{Y} を第2群の標本平均、
 U_x^2 を第1群の不偏分散、 U_y^2 を第2群の不偏分散とする。

$$\text{つまり } \bar{X} = \frac{\sum_{i=1}^{N_1} x_i}{N_1}, \quad \bar{Y} = \frac{\sum_{i=1}^{N_2} y_i}{N_2}, \quad U_x^2 = \frac{\sum_{i=1}^{N_1} (x_i - \bar{X})^2}{N_1 - 1}, \quad U_y^2 = \frac{\sum_{i=1}^{N_2} (y_i - \bar{Y})^2}{N_2 - 1}$$

である。この時

$$\text{統計量 } T = \frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{\left(\frac{1}{N_1} + \frac{1}{N_2}\right) U_{xy}}}, \quad U_{xy}^2 = \frac{(N_1 - 1)U_x^2 + (N_2 - 1)U_y^2}{N_1 + N_2 - 2}$$

は自由度 $N_1 + N_2 - 2$ の t 分布に従う。

検定についての考え方には3通りあり、帰無仮説はいずれも

帰無仮説 : 2群間の平均値には差がない。つまり $\mu_x = \mu_y$

であるが、対立仮説は

- 対立仮説(1) : 2群間の平均値には差がないとはいえない(両側検定)
- 対立仮説(2) : $\mu_x < \mu_y$ (左片側検定)
- 対立仮説(3) : $\mu_x > \mu_y$ (右片側検定)

の3通りとなる。ここでは $\mu_x = \mu_y$ を仮定するので、実際には 以下を計算し

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\left(\frac{1}{N_1} + \frac{1}{N_2}\right) U_{xy}}}$$

T が棄却域に入る場合 対立仮説を採択し、そうでない場合 帰無仮説を採択する。

なお、サンプルサイズ N_1 , N_2 が大きいと、統計量 T の式の一部の $\frac{\sqrt{N_1 N_2}}{\sqrt{N_1 + N_2}}$ が大きくなり、つまり T の絶対値が大きくなり、p 値が小さくなり、帰無仮説が棄却されることが多くなる。

その為、サンプルサイズの大小に影響されにくい 効果量 d を同時に表示する。

$$\text{効果量 } d = \frac{\bar{X} - \bar{Y}}{U_{xy}} = T \times \sqrt{\left(\frac{1}{N_1} + \frac{1}{N_2}\right)} \quad \text{ただし } U_{xy}^2 = \frac{(N_1 - 1)U_x^2 + (N_2 - 1)U_y^2}{N_1 + N_2 - 2}$$