Testowanie średnich

Preliminaria

Rozpatrujemy dwie próbki: $X = \{x_1, \ldots, x_{n_x}\}$ oraz $Y = \left\{y_1, \ldots, y_{n_y}\right\}$. Zakładamy, że obserwacje są niezależne i podlegają rozkładom $\mathrm{N}(\mu_x, \sigma_x^2)$, $\mathrm{N}(\mu_y, \sigma_y^2)$. Przy tych założeniach mamy:

(i) $\bar{X} \sim \mathrm{N}\left(\mu_x, \dfrac{\sigma_x^2}{n_x}\right), \quad \bar{Y} \sim \mathrm{N}\left(\mu_y, \dfrac{\sigma_y^2}{n_y}\right),$

(ii) $\dfrac{n_x S_x^2}{\sigma_x^2} \sim \chi^2(n_x - 1), \quad \dfrac{n_y S_y^2}{\sigma_y^2} \sim \chi^2(n_y - 1),$

(iii) $\mathrm{V}\left(\dfrac{n_x S_x^2}{\sigma_x^2}\right) = 2\,(n_x - 1), \quad \mathrm{V}\left(\dfrac{n_y S_y^2}{\sigma_y^2}\right) = 2\,(n_y - 1),$

(iv) $V(\bar{X} - \bar{Y}) = \dfrac{\sigma_x^2}{n_x} + \dfrac{\sigma_y^2}{n_y},$

(v) $\bar{X}, S_x^2$ są niezależne; także $\bar{Y}, S_y^2$ są niezależne.

Przypomnijmy też, że zmienna $t(k)$ ma rozkład t-Studenta z k stopniami swobody wtedy gdy

$$t(k) = \frac{U}{\sqrt{V/k}}, \quad \text{gdzie } U \sim \mathrm{N}(0,1),\ \ V \sim \chi^2(k),\ \ U, V \text{ - niezależne.}$$

Hipoteza H_0

H_0: $\mu_x = \mu_y$.

1. Różne (znane) wariancje

Przy założeniu, że hipoteza H_0 jest prawdziwa i znane są wariancje σ_x^2, σ_y^2, mamy:

(c) $\bar{X} - \bar{Y} \sim \mathrm{N}\left(0, \dfrac{\sigma_x^2}{n_x} + \dfrac{\sigma_y^2}{n_y}\right)$

Stąd otrzymujemy

$$z = \frac{\bar{X} - \bar{Y}}{\sqrt{\left(\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}\right)}} \sim N(0,1). \tag{1}$$

2. Równe (nieznane) wariancje

Przy założeniu, że hipoteza H_0 jest prawdziwa i wariancje są równe ($\sigma_x^2 = \sigma_y^2 =: \sigma^2$) jest:

(a) $\bar{X} - \bar{Y} \sim \mathrm{N}\left(0, \sigma^2\left(\dfrac{1}{n_x} + \dfrac{1}{n_y}\right)\right)$

(b) $\dfrac{n_x S_x^2}{\sigma^2} + \dfrac{n_y S_y^2}{\sigma^2} \sim \chi^2(n_x + n_y - 2).$

Ze wzorów (a) oraz (b) wynika (wobec niezależności zmiennych)

$$\left[\frac{\bar{X} - \bar{Y}}{\sigma\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}\right] \Bigg/ \sqrt{\frac{n_x S_x^2 + n_y S_y^2}{\sigma^2}\frac{1}{n_x + n_y - 2}} \sim t(n_x - n_y - 2),$$

skąd otrzymujemy ostatecznie

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\left(n_x S_x^2 + n_y S_y^2\right)\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}} \sqrt{n_x + n_y - 2} \sim t(n_x + n_y - 2). \tag{2}$$

Zwróćmy uwagę na to iż nie jest istotne czy znamy wariancje σ_x^2, σ_y^2, korzystaliśmy wyłącznie z równości $\sigma_x^2 = \sigma_y^2$.

3. Różne (nieznane) wariancje – test Welcha

Rozważmy ponownie zmienną $\bar{X} - \bar{Y}$. Jak w poprzednim punkcie $\sigma^2 = \mathrm{Var}(\bar{X} - \bar{Y}) = \dfrac{\sigma_x^2}{n_x} + \dfrac{\sigma_y^2}{n_y}$. Jako przybliżoną wartość σ^2 przyjmijmy $S^2 = \dfrac{S_x^2}{n_x} + \dfrac{S_y^2}{n_y}$. Załóżmy też, dla pewnego r zmienna $\dfrac{rS^2}{\sigma^2}$ podlega rozkładowi $\chi^2(r)$. Z własności rozkładu χ^2 wynika, że $\mathrm{V}\left(\dfrac{rS^2}{\sigma^2}\right) = 2r$. Stąd mamy $\mathrm{V}\left(\dfrac{rS^2}{\sigma^2}\right) = \dfrac{r^2}{\sigma^4}\mathrm{V}(S^2) = 2r$ czyli $\dfrac{1}{\sigma^4}\mathrm{V}(S^2) = \dfrac{2}{r}$.
Wstawiając do ostatniego wzoru wyrażenie na S^2 otrzymujemy wzór

$$\frac{1}{\sigma^4}\left(\frac{\mathrm{V}(S_x^2)}{n_x^2} + \frac{\mathrm{V}(S_y^2)}{n_y^2}\right) = \frac{2}{r}. \tag{3}$$

Wariancję S_x^2 obliczamy z zależności $2 \cdot (n_x - 1) = \mathrm{V}\left(\dfrac{n_x S_x^2}{\sigma_x^2}\right) = \dfrac{n_x^2}{\sigma_x^4} \cdot \mathrm{V}(S_x^2)$ oraz z analogicznego wzoru dla S_y^2. Następnie przybliżamy wielkości $\sigma^2, \sigma_x^2, \sigma_y^2$ przez S^2, S_x^2, S_y^2. Równanie (3) przepisujemy w (przybliżonej) postaci:

$$\frac{2}{r} = \frac{1}{S^4}\left(\frac{2 \cdot (n_x - 1)S_x^4}{n_x^4} + \frac{2 \cdot (n_y - 1)S_y^4}{n_y^4}\right),$$

skąd ostatecznie wynika postać wzoru jak poniżej

$$r = \frac{\left(\dfrac{S_x^2}{n_x} + \dfrac{S_y^2}{n_y}\right)^2}{\left(\dfrac{S_x^2}{n_x}\right)^2 \cdot \frac{n_x - 1}{n_x^2} + \left(\dfrac{S_y^2}{n_y}\right)^2 \cdot \frac{n_y - 1}{n_y^2}}. \tag{4}$$

Statystyka testowa jest postaci:

$$t = \frac{\bar{X} - \bar{Y}}{S} \approx t(r). \tag{5}$$

Wzory (1), (2), (5) opisują statystyki testowe, $S^2 = \dfrac{S_x^2}{n_x} + \dfrac{S_y^2}{n_y}$.

Witold Karczewski