Statystyka i Analiza Danych

W6: Porównanie dwóch populacji

dr hab. inż. Katarzyna Filipiak, prof. PP

Instytut Matematyki
Politechnika Poznańska

2023/2024

Populacje

Próby z dwóch populacji: X i Y

Próba	Parametry	Statystyki
$X_1, X_2, \ldots, X_{n_1}$	μ_1, σ_1^2, p_1	$\overline{X},\ s_1^2,\ \widehat{p}_1$
$Y_1, Y_2, \ldots, Y_{n_2}$	μ_2, σ_2^2, p_2	$\overline{Y},\ s_2^2,\ \widehat{p}_2$

Założenie: $X_1, X_2, \ldots, X_{n_1}$, $Y_1, Y_2, \ldots, Y_{n_2}$ – próby niezależne

Różnica średnich populacyjnych

Parametr: $\mu_1 - \mu_2$

Estymator punktowy: $\overline{X} - \overline{Y}$

Własności: $E(\overline{X} - \overline{Y}) = \mu_1 - \mu_2$

$$\mathrm{Var}(\overline{X} - \overline{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Jeżeli $\sigma_1^2 = \sigma_2^2 = \sigma^2$ to

$$\mathrm{Var}(\overline{X} - \overline{Y}) = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

Przedział ufności dla $\mu_1 - \mu_2$ (1)

Założenia: $X \sim N(\mu_1, \sigma_1)$, $Y \sim N(\mu_2, \sigma_2)$, $\sigma_1^2 = \sigma_2^2$

$t_{\nu,\beta}$ – kwantyl rozkładu t_ν; $\quad S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$

Z ufnością $100(1-\alpha)\%$ przedział

$$\left(\overline{X} - \overline{Y} - t_{n_1+n_2-2,1-\alpha/2} \cdot \sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \; ; \; \overline{X} - \overline{Y} + t_{n_1+n_2-2,1-\alpha/2} \cdot \sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \right)$$

pokrywa nieznaną prawdziwą różnicę średnich populacyjnych $\mu_1 - \mu_2$.

```
t.test(dane1, dane2, var.equal = TRUE, conf.level = 1 − α)
```

Założenia: $X \sim N(\mu_1, \sigma_1)$, $Y \sim N(\mu_2, \sigma_2)$, $\sigma_1^2 = \sigma_2^2$

Statystyka testowa: $$t = \frac{\overline{X} - \overline{Y} - \mu_0}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \underset{H_0}{\sim} t_{n_1+n_2-2}$$

H_0	H_1	Obszar krytyczny R zależny od α
$\mu_1 - \mu_2 = \mu_0$	$\mu_1 - \mu_2 \neq \mu_0$	$(-\infty, -t_{n_1+n_2-2,1-\alpha/2}) \cup (t_{n_1+n_2-2,1-\alpha/2}, \infty)$
$\mu_1 - \mu_2 \leqslant \mu_0$	$\mu_1 - \mu_2 > \mu_0$	$(t_{n_1+n_2-2,1-\alpha}, \infty)$
$\mu_1 - \mu_2 \geqslant \mu_0$	$\mu_1 - \mu_2 < \mu_0$	$(-\infty, -t_{n_1+n_2-2,1-\alpha})$

```
t.test(dane1, dane2, var.equal = TRUE, mu = μ0, alternative = "two.sided")
```

Przedział ufności dla $\mu_1 - \mu_2$ (2)

Założenia: $X \sim N(\mu_1, \sigma_1)$, $Y \sim N(\mu_2, \sigma_2)$, $\sigma_1^2 \neq \sigma_2^2$

$t_{\nu,\beta}$ – kwantyl rozkładu t_ν, $\nu = \left\lfloor \frac{(S_1^2/n_1+S_2^2/n_2)^2}{(S_1^2/n_1)^2/(n_1-1)+(S_2^2/n_2)^2/(n_2-1)} \right\rfloor$

Z ufnością $100(1-\alpha)\%$ przedział

$$\left(\overline{X} - \overline{Y} - t_{\nu,1-\alpha/2} \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\ ;\ \overline{X} - \overline{Y} + t_{\nu,1-\alpha/2} \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right)$$

pokrywa nieznaną prawdziwą różnicę średnich populacyjnych $\mu_1 - \mu_2$.

`t.test(dane1, dane2, var.equal = FALSE, conf.level = 1 − α)`

Hipoteza o $\mu_1 - \mu_2$ (2)

Założenia: $X \sim N(\mu_1, \sigma_1)$, $Y \sim N(\mu_2, \sigma_2)$, $\sigma_1^2 \neq \sigma_2^2$

Statystyka testowa: $$\tilde{t} = \frac{\overline{X} - \overline{Y} - \mu_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \underset{H_0}{\sim} t_\nu$$

H_0	H_1	Obszar krytyczny R zależny od α
$\mu_1 - \mu_2 = \mu_0$	$\mu_1 - \mu_2 \neq \mu_0$	$(-\infty, -t_{\nu,1-\alpha/2}) \cup (t_{\nu,1-\alpha/2}, \infty)$
$\mu_1 - \mu_2 \leqslant \mu_0$	$\mu_1 - \mu_2 > \mu_0$	$(t_{\nu,1-\alpha}, \infty)$
$\mu_1 - \mu_2 \geqslant \mu_0$	$\mu_1 - \mu_2 < \mu_0$	$(-\infty, -t_{\nu,1-\alpha})$

t.test(dane1, dane2, var.equal = FALSE, mu=μ_0, alternative="two.sided")

Przedział ufności dla $\mu_1 - \mu_2$ (3)

Założenia: duże próby $(n_1, n_2 > 30)$

z_β – kwantyl rozkładu $N(0,1)$

Z ufnością $100(1-\alpha)\%$ przybliżony przedział

$$\left(\overline{X} - \overline{Y} - z_{1-\alpha/2} \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\ ;\ \overline{X} - \overline{Y} + z_{1-\alpha/2} \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right)$$

pokrywa nieznaną prawdziwą różnicę średnich populacyjnych $\mu_1 - \mu_2$.

```
zsum.test (mean(dane1), sd(dane1), length(dane1),
           mean(dane2), sd(dane2), length(dane2),
           conf.level = 1 - α)
```

UWAGA! Wymagany pakiet BSDA

Założenia: duże próby $(n_1, n_2 > 30)$

Statystyka testowa: $$\tilde{Z} = \frac{\overline{X} - \overline{Y} - \mu_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \underset{H_0}{\sim} N(0,1)$$

H_0	H_1	Obszar krytyczny R zależny od α
$\mu_1 - \mu_2 = \mu_0$	$\mu_1 - \mu_2 \neq \mu_0$	$(-\infty, -z_{1-\alpha/2}) \cup (z_{1-\alpha/2}, \infty)$
$\mu_1 - \mu_2 \leqslant \mu_0$	$\mu_1 - \mu_2 > \mu_0$	$(z_{1-\alpha}, \infty)$
$\mu_1 - \mu_2 \geqslant \mu_0$	$\mu_1 - \mu_2 < \mu_0$	$(-\infty, -z_{1-\alpha})$

```
zsum.test (mean(dane1), sd(dane1), length(dane1),
           mean(dane2), sd(dane2), length(dane2),
           mu = μ0, alternative = "two.sided")
```

UWAGA! Wymagany pakiet `BSDA`

Przykład 1

Dokonano pomiaru żywotności dwóch typów żarówek energooszczędnych typu LED (w godzinach). Uzyskano 5 obserwacji dla świetlówek I-go typu: 2830, 2840, 2800, 2880, 2820 oraz 5 obserwacji dla świetlówek II-go typu: 2790, 2720, 2770, 2780, 2760. Wiedząc, że czas świecenia żarówek I-go rodzaju podlega rozkładowi $N(\mu_1, \sigma)$, a żarówek II-go rodzaju rozkładowi $N(\mu_2, \sigma)$, wykonaj polecenia:

(a) oceń z ufnością 99% różnicę prawdziwych średnich żywotności żarówek dwóch typów;

(b) czy uzyskane dane potwierdzają przypuszczenie, że żywotność żarówek I-go typu jest większa od żywotności żarówek II-go typu? Przyjmij poziom istotności 0,01.

Przykład 1 (a)

Przykład 1 (a) – rozwiązanie w R

[1] 9.789727 130.210273
attr(,"conf.level")
[1] 0.99

Przykład 1 (b)

Przykład 1 (b) – rozwiązanie w R

```
        Two Sample t-test

data: czas1 and czas2
t = 3.9009, df = 8, p-value = 0.002269
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
  36.6316  Inf
sample estimates:
mean of x  mean of y
  2834       2764
```

Wariancje populacyjne

$X_1, X_2, \ldots, X_{n_1}$ – próba z rozkładu $N(\mu_1, \sigma_1)$

$Y_1, Y_2, \ldots, Y_{n_2}$ – próba z rozkładu $N(\mu_2, \sigma_2)$

Założenie: $X_1, X_2, \ldots, X_{n_1}$, $Y_1, Y_2, \ldots, Y_{n_2}$ – niezależne

Statystyka:

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1, n_2-1}$$

F_{n_1-1, n_2-1} – rozkład F-Snedecora z $(n_1 - 1, n_2 - 1)$ stopniami swobody

Przedział ufności dla $\sigma_1^2 \,/\, \sigma_2^2$

$F_{\nu_1,\nu_2;\beta}$ – kwantyl rozkładu F_{ν_1,ν_2}

Z ufnością $100(1-\alpha)\%$ przedział

$$\left(\frac{S_1^2}{S_2^2} \cdot \frac{1}{F_{n_1-1,n_2-1;1-\alpha/2}} \;;\; \frac{S_1^2}{S_2^2} \cdot \frac{1}{F_{n_1-1,n_2-1;\alpha/2}} \right)$$

pokrywa nieznany prawdziwy iloraz dwóch wariancji populacyjnych $\frac{\sigma_1^2}{\sigma_2^2}$.

`var.test (dane1, dane2, conf.level` $= 1-\alpha$`)`

UWAGA! Wymagany pakiet `PairedData`

Hipoteza o $\sigma_1^2 = \sigma_2^2$

Statystyka testowa: $F = \frac{S_1^2}{S_2^2} \underset{H_0}{\sim} F_{n_1-1,n_2-1}$

H_0	H_1	Obszar krytyczny R zależny od α
$\sigma_1^2 = \sigma_2^2$	$\sigma_1^2 \neq \sigma_2^2$	$(0, F_{n_1-1,n_2-1;\alpha/2}) \cup (F_{n_1-1,n_2-1;1-\alpha/2}, \infty)$
$\sigma_1^2 \leqslant \sigma_2^2$	$\sigma_1^2 > \sigma_2^2$	$(F_{n_1-1,n_2-1;1-\alpha}, \infty)$
$\sigma_1^2 \geqslant \sigma_2^2$	$\sigma_1^2 < \sigma_2^2$	$(0, F_{n_1-1,n_2-1;\alpha})$

```
var.test(dane1, dane2, alternative = "two.sided")
```

UWAGA! Wymagany pakiet `PairedData`

Przykład 1 - c.d.

Dokonano pomiaru żywotności dwóch typów żarówek energooszczędnych typu LED (w godzinach). Uzyskano 5 obserwacji dla świetlówek I-go typu: 2830, 2840, 2800, 2880, 2820 oraz 5 obserwacji dla świetlówek II-go typu: 2790, 2720, 2770, 2780, 2760. Wiedząc, że czas świecenia żarówek obu typów podlega rozkładom normalnym, na poziomie istotności 0,01 zweryfikuj założenie o jednorodności wariancji.

Przykład 1 - c.d.

Przykład 1 - rozwiązanie w R

```
        F test to compare two variances

data: czas1 and czas2
F = 1.2055, num df = 4, denom df = 4, p-value = 0.8607
alternative hypothesis: true ratio of variances is not equal to 1
99 percent confidence interval:
   0.05206242  27.91227571
sample estimates:
ratio of variances
   1.205479
```

Różnica proporcji populacyjnych $p_1 - p_2$

Różnica proporcji populacyjnych: $p_1 - p_2$

Proporcje próbkowe: $\widehat{p}_1 = \dfrac{T_1}{n_1}, \quad \widehat{p}_2 = \dfrac{T_2}{n_2}, \quad \widehat{q}_1 = 1 - \widehat{p}_1, \quad \widehat{q}_2 = 1 - \widehat{p}_2$

Założenie: n_1, n_2 – duże $(\geqslant 100)$

z_β – kwantyl rozkładu $N(0,1)$

Z ufnością $100(1-\alpha)\%$ przybliżony przedział

$$\left(\widehat{p}_1 - \widehat{p}_2 - z_{1-\alpha/2} \cdot \sqrt{\frac{\widehat{p}_1\widehat{q}_1}{n_1} + \frac{\widehat{p}_2\widehat{q}_2}{n_2}}\ ;\ \widehat{p}_1 - \widehat{p}_2 + z_{1-\alpha/2} \cdot \sqrt{\frac{\widehat{p}_1\widehat{q}_1}{n_1} + \frac{\widehat{p}_2\widehat{q}_2}{n_2}}\right)$$

pokrywa nieznaną prawdziwą różnicę proporcji populacyjnych $p_1 - p_2$.

$$\texttt{prop.test}(\texttt{c}(T_1, T_2), \texttt{c}(n_1, n_2), \texttt{conf.level} = 1 - \alpha)$$

Hipoteza o $p_1 - p_2$

Założenie: n_1, n_2 – duże $(\geqslant 100)$

Estymator wspólnej proporcji: $\widehat{p} = \dfrac{T_1 + T_2}{n_1 + n_2}$

Statystyka testowa: $Z = \dfrac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \underset{H_0}{\sim} N(0,1)$

H_0	H_1	Obszar krytyczny R zależny od α
$p_1 = p_2$	$p_1 \neq p_2$	$\left(-\infty, -z_{1-\alpha/2}\right) \cup \left(z_{1-\alpha/2}, \infty\right)$
$p_1 \leqslant p_2$	$p_1 > p_2$	$\left(z_{1-\alpha}, \infty\right)$
$p_1 \geqslant p_2$	$p_1 < p_2$	$\left(-\infty, -z_{1-\alpha}\right)$

`prop.test(c(T1, T2), c(n1, n2), alternative = "two.sided")`

Przykład 2

Rzecznik innowacyjnych metod nauczania chce porównać efektywność nauczania języka angielskiego dwoma metodami: tradycyjną oraz przez zastosowanie pomocy audio-wizualnych. Z grupy 250 uczniów wybrano losowo 150 i poddano ich metodzie uczenia z pomocą technik audio-wizualnych (AW). Pozostałych uczniów uczono sposobem tradycyjnym (T). Na koniec semestru przeporwadzono test, którego wyniki są następujące:

	AW	T
zdało:	107	63
nie zdało:	43	37

(a) Wyznacz 95% przedział ufności dla różnicy proporcji testów dla dwóch metod nauczania.

(b) Czy dane potwierdzają przypuszczenie, że proporcja zdanych testów jest lepsza w nowym sposobie nauczania? Przyjmij poziom istotności 0,05.

Przykład 2

Przykład 2

```
[1]   -0.0357941  0.2024608
attr(,"conf.level")
[1]   0.95
```

```
[1]   -0.04412743  0.21079410
attr(,"conf.level")
[1]   0.95
```

Przykład 2

Przykład 2

```
prop.test(c(107, 63), c(150, 100), alternative = "greater",
          correct = FALSE)
```

2-sample test for equality of proportions without continuity correction

data: c(107, 63) out of c(150, 100)
X-squared = 1.9148, df = 1, p-value = 0.08321
alternative hypothesis: greater
95 percent confidence interval:
-0.01664156 1.00000000
sample estimates:

prop 1	prop 2
0.7133333	0.6300000

Przykład 2

```
prop.test(c(107, 63), c(150, 100), alternative = "greater",
          correct = TRUE)
```

2-sample test for equality of proportions with continuity correction

data: c(107, 63) out of c(150, 100)
X-squared = 1.551, df = 1, p-value = 0.1065
alternative hypothesis: greater
95 percent confidence interval:
-0.02497489 1.00000000
sample estimates:

prop 1	prop 2
0.7133333	0.6300000