

Statystyka w analizie i planowaniu eksperymentu

Wykład 5

Testowanie wartości średnich

Przemysław Biecek

Dla 1 roku studentów Biotechnologii

- Za tydzień (16 kwiecień) będzie wejściówka z podstawowych pojęć dotyczących testowania,
- Za dwa tygodnie (23 kwiecień) będzie kolokwium z materiału poznanego do 17 kwietnia (podstawy rachunku prawdopodobieństwa, statystyki opisowe, podstawy testowania).

Będą nas interesowały testy dotyczące wartości średniej w dwóch lub więcej populacjach. Przyjmujemy założenie, że obserwowane wartości zgodne są z rozkładem normalnym.

- Analiza dla dwóch grup
 - test t-Studenta dla dwóch grup
 - test t-Studenta dla dwóch grup o różnej wariancji
 - test t-Studenta dla zmiennych sparowanych
- Analiza większej liczby grup (jednokierunkowa analiza wariancji)

Wykonaliśmy dwie serie pomiarów.

W pierwszej serii wykonano n_1 pomiarów, które będziemy oznaczać X_1, \dots, X_{n_1} .

W drugiej serii wykonano n_2 pomiarów, które będziemy oznaczać Y_1, \dots, Y_{n_2} .

Przyjmujemy, że wartości $X_i \sim \mathcal{N}(\mu_1, \sigma_1^2)$, oraz $Y_i \sim \mathcal{N}(\mu_2, \sigma_2^2)$.

Przyjmujemy również (o ile nie zaznaczymy, że jest inaczej), że zarówno zmienne X_i jak i Y_i są niezależne.

W wymienionych poniżej testach, interesującą nas hipotezą zerową będzie dotyczyła równości średnich w obu grupach

$$H_0 : \mu_X = \mu_Y.$$

Za alternatywę, podobnie jak dla jednej grupy, możemy wybrać jedną z trzech hipotez

- dwustronna

$$H_{A1} : \mu_x \neq \mu_y$$

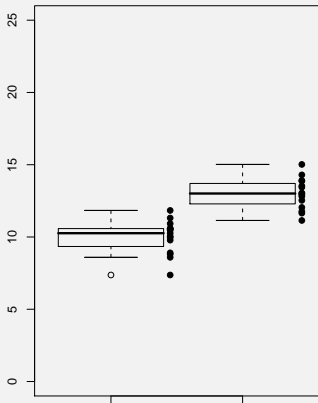
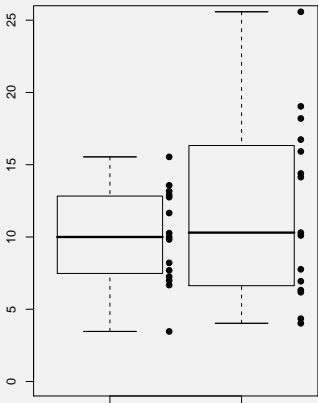
- jednostronna

$$H_{A2} : \mu_x > \mu_y$$

$$H_{A3} : \mu_x < \mu_y$$

Jak to ugryźć?

Mamy dwie próby, średnie to odpowiednio 10 i 12.
Czy to istotna statystycznie różnica?



Dwie próby o znanej wariancji

Jeżeli wariancje w obu grupach są znane, to za statystykę testową wybieramy

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Przy prawdziwej hipotezie zerowej, ta statystyka ma rozkład normalny $\mathcal{N}(0, 1)$.

Ten test, nazywany jest też testem **U**.

Przykład

Wykonaliśmy pomiary stężenia globulin w osoczu w dwóch grupach pacjentów. Przyjmujemy, że wariancja pomiaru w pierwszej grupie wynosi 20^2 a w drugiej grupie 30^2 .

Otrzymane pomiary to

$$X = (87, 88, 55, 122, 105, 63, 82, 95, 96, 97)$$

$$Y = (55, 97, 106, 95, 135, 67, 104, 130)$$

Wyznaczamy

$$\bar{X} = 89$$

$$\bar{Y} = 98.625$$

$$T = \frac{89 - 98.625}{\sqrt{20^2/10 + 30^2/8}} = -0.779$$

Dla dwustronnej alternatywy, odpowiadająca temu wynikowi p-wartość wynosi $p = 0.42$.

A obszar przyjęcia dla $\alpha = 0.05$ to

$$B^C = (-1.96, 1.96).$$

Jeżeli wariancje w obu grupach są równe ($\sigma_1^2 = \sigma_2^2$) ale nie są znane, to za statystykę testową wybieramy

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}.$$

Przy prawdziwej hipotezie zerowej, ta statystyka ma rozkład t-Studenta o $n_1 + n_2 - 2$ stopniach swobody.

Przykład

Wykonaliśmy pomiary absorpcji próbek zawierających dwie nieznanne substancje. Interesuje nas weryfikacja hipotezy, że absorpcja obu substancji jest sobie równa.

Otrzymane pomiary to

$$X = (0.48, 0.57, 0.46, 0.46, 0.55, 0.77, 0.64, 0.56, 0.55, 0.43)$$

$$Y = (0.63, 0.68, 0.60, 0.52, 0.71, 0.54, 0.63, 0.63, 0.84)$$

Wyznaczamy

$$\begin{aligned}\bar{X} &= 0.547 & \bar{Y} &= 0.642 \\ S_1^2 &= 0.01027 & S_2^2 &= 0.00909 \\ T &= \frac{-0.0952}{\sqrt{0.1651/17*(1/10+1/9)}} = -2.103\end{aligned}$$

Dla dwustronnej alternatywy, odpowiadająca temu wynikowi p-wartość wynosi $p = 0.0507$.

A obszar przyjęcia dla $\alpha = 0.05$ to

$$B^C = (-2.1098, 2.1098).$$

Dwie próby o nie znanej ale różnej wariancji

Jeżeli wariancje w obu grupach są różne i nie są znane ($\sigma_1^2 \neq \sigma_2^2$), to za statystykę testową wybieramy

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}.$$

Kwantyle rozkładu statystyki testowej przy prawdziwej hipotezie zerowej wyznacza się ze wzoru

$$q(x, n_1, n_2) = \frac{w_1 t_{n_1-1}(x) + w_2 t_{n_2-1}(x)}{w_1 + w_2},$$

gdzie $w_1 = \frac{S_1^2}{n_1}$, $w_2 = \frac{S_2^2}{n_2}$ a $t_k(x)$ to kwantyl rozkładu t-Studenta o k stopniach swobody w punkcie x .

Wykorzystajmy dane z poprzedniego przykładu, przyjmiemy teraz jednak, że wariancje niekoniecznie są równe

Wyznaczamy

$$\begin{aligned}\bar{X} &= 0.547 & \bar{Y} &= 0.642 \\ S_1^2 &= 0.01027 & S_2^2 &= 0.00909 \\ w_1 &= 0.0010267 & w_2 &= 0.0010105 \\ T &= -2.1097\end{aligned}$$

Wyznaczamy obszar przyjęcia dla $\alpha = 0.05$ to

$$\begin{aligned}q(0.025, 10, 9) &= \frac{-2.228*w_1 + -2.262*w_2}{w_1 + w_2} = -2.245 \\ q(0.975, 10, 9) &= 2.245\end{aligned}$$

Próby sparowane (zależne)

Jeżeli pomiary dotyczą tych samych obiektów ale w różnych warunkach i interesuje nas weryfikacja hipotezy, czy średnia wartość badanej cechy pozostała niezmienną, należy zastosować test dla danych sparowanych.

W tym przypadku, za statystykę testową wybieramy

$$T = \frac{\bar{Z}}{S_Z} \sqrt{n}$$

gdzie $Z_i = X_i - Y_i$ oznacza różnica elementów w parze.

Przy prawdziwej hipotezie zerowej, statystyka ta ma rozkład t-Studenta o $n - 1$ stopniach swobody (tutaj $n = n_1 = n_2$).

Przykład

W próbkach mamy próbki nieznaney mieszaniny, chcemy sprawdzić, czy zmieni się absorpcja jeżeli tą mieszaninę podgrzejemy.

Otrzymaliśmy następujące pomiary

$$X = (0.48, 0.57, 0.46, 0.46, 0.55, 0.77, 0.64, 0.56, 0.55)$$

$$Y = (0.63, 0.68, 0.60, 0.52, 0.71, 0.54, 0.63, 0.63, 0.84)$$

Wyznaczamy

$$Z = (-0.15, -0.11, -0.14, -0.06, -0.16, 0.23, 0.01, -0.07, -0.29)$$

$$\bar{Z} = -0.082$$

$$S_Z^2 = 0.0206 \quad S_Z = 0.1434$$

$$T = -1.720$$

Dla dwustronnej alternatywy, odpowiadająca temu wynikowi p-wartość wynosi $p = 0.119$.

A obszar przyjęcia dla $\alpha = 0.05$ to

$$B^C = (-2.262, 2.262).$$

Rozkład t-Studenta wraz z wzrostem liczby stopni swobody zbiega do rozkładu normalnego.

Z tego powodu, dla dużych liczebności próby ($n > 50$) można zamiast kwantyli rozkładu t, wykorzystywać kwantyle rozkładu normalnego $\mathcal{N}(0, 1)$.

Taki test, nazywany jest też testem z.

Jak to zrobić w pakiecie R?

W pakiecie R test na równość średnich można wykonać funkcją

```
t.test(x, y, alternative = c("two.sided", "less", "greater"),  
      paired = FALSE, var.equal = FALSE)
```

- argument **x** określa pierwszy wektor obserwacji,
- argument **y** określa drugi wektor obserwacji,
- argument **alternative** określa jaka hipoteza alternatywna jest testowana,
- argument **paired** określa czy obserwacje są sparowane, czy nie,
- argument **var.equal** określa czy wariancje są równe w obu grupach.

Jak to zrobić w pakiecie R?

```
> y = round(100*rnorm(10) + 320)
> x = round(100*rnorm(10) + 220)
> x
[1] 350 287 393 69 98 276 238 121 315 276
> y
[1] 334 253 339 313 364 292 302 409 351 476
>
> t.test(x, y)
Welch Two Sample t-test
data: x and y
t = -2.513, df = 14.334, p-value = 0.0245
alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:
-187.01365 -14.98635
sample estimates:
mean of x mean of y
242.3 343.3
```

Jak to zrobić w pakiecie R?

```
> t.test(x, y, alternative="less")
Welch Two Sample t-test
data: x and y
t = -2.513, df = 14.334, p-value = 0.01225
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
-Inf -30.32708
sample estimates:
mean of x mean of y
242.3 343.3
```

Jak to zrobić w pakiecie R?

```
> t.test(x, y, paired=TRUE)
Paired t-test
data: x and y
t = -2.3865, df = 9, p-value = 0.04079
alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:
-196.738525 -5.261475
sample estimates:
mean of the differences
-101
```

Jak to zrobić w pakiecie R?

```
> t.test(x, y, paired=TRUE, alternative="less")
Paired t-test
data: x and y
t = -2.3865, df = 9, p-value = 0.02040
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
-Inf -23.41939
sample estimates:
mean of the differences
-101
```

Przypuśćmy, że interesuje nas większa (niż dwie) liczba podpopulacji. Aby porównać średnie w kilku grupach, można przeprowadzić analizę wariancji.

Wykonaliśmy k serii pomiarów. W serii i wykonaliśmy n_i pomiarów. Pomiar w serii i oznaczamy przez $X_1^i, \dots, X_{n_i}^i$. Przyjmujemy, że wartości $X_j^i \sim \mathcal{N}(\mu_i, \sigma^2)$ (wariancje są równe!!!) oraz, że zmienne X_j^i są niezależne.

Interesująca nas hipoteza zerowa jest postaci

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k$$

a hipotezą alternatywną jest

$$H_A : \exists_{i,j} \mu_i \neq \mu_j.$$

Statystyką testową w analizie wariancji jest

$$F = \frac{SSA/(k-1)}{SSE/(n-k)}$$

gdzie $n = \sum_i n_i$,

$$SSA = n \sum_{i=1}^k (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2, \quad SSE = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})^2.$$

Dla prawdziwej hipotezy zerowej, ta statystyka testowa ma rozkład \mathcal{F} Snedecora z $k-1$ i $n-k$ stopniami swobody.

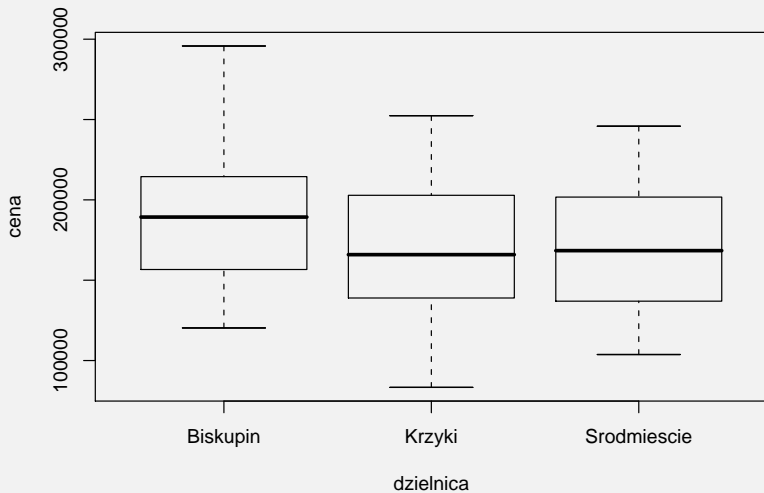
Uwaga

Jeżeli odrzucimy hipotezę zerową, a więc przyjmiemy, że przynajmniej dwie średnie się różnią, to powinniśmy wykonać kolejny krok, określający które zmienne się różnią. W tym celu wykonuje się testy post-hoc.

Aby przykład zapadał w pamięć będzie on dotyczył pieniędzy.

```
> summary(mieszkania)
cena pokoi powierzchnia dzielnica
Min. : 83280 Min. :1.00 Min. :17.00 Biskupin :65
Mean :175934 Mean :2.55 Mean :46.20
Max. :295762 Max. :4.00 Max. :87.70
```

Przykład



Interesuje nas weryfikacja hipotezy, czy średnie ceny mieszkań, w różnych dzielnicach, są równe.

```
> (a1 = anova(lm(cena dzielnica, data = mieszkania)))
```

```
Analysis of Variance Table
```

```
Response: cena
```

```
Df Sum Sq Mean Sq F value Pr(>F)
```

```
dzielnica 2 1.7995e+10 8.9977e+09 5.0456 0.007294 **
```

```
Residuals 197 3.5130e+11 1.7833e+09
```

```
---
```

```
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Co trzeba zapamiętać?

- Jakie założenia muszą być spełnione, by móc wykonywać testy omówione na tym wykładzie?
- Które testy można wykorzystywać gdy wariancje są znane?
- Które testy można wykorzystywać gdy wariancje są nieznane?
- Które testy można wykorzystywać gdy wariancje są równe?
- Na czym polega różnica pomiędzy grupami sparowanymi a niesparowanymi?