

# Statystyka w analizie i planowaniu eksperymentu

Wykład 10

## Podsumowanie przerobionego materiału

Przemysław Biecek

Dla 1 roku studentów Biotechnologii

Proszę na kartce napisać:

- 1 Imię, nazwisko,
- 2 Nr. indeksu.

- 6 VI oddanie wejściówki.
- 11 VI drugie kolokwium.
- 11 VI termin oddawania raportów z badań własnych własnych.
- 18 VI wpis dla osób o jasnej sytuacji.
- 18 VI prezentacja najciekawszych raportów dotyczących badań własnych.

- Estymacja
  - statystyki podstawowe,
  - przedziały ufności i błąd standardowy,
  - współczynniki korelacji,
  - model regresji.
- Testowanie
  - testy zgodności: test K-S,  $\chi^2$ ,
  - test dla wartości odstających: test Grubbsa, Dixona.
  - testy dla parametrów skali,
  - testy dla parametrów położenia: t-studenta, Wilcoxon, test proporcji,
  - testy niezależności: test dla współczynnika korelacji, test  $\chi^2$ ,
  - inne testy.

Średnia w próbie

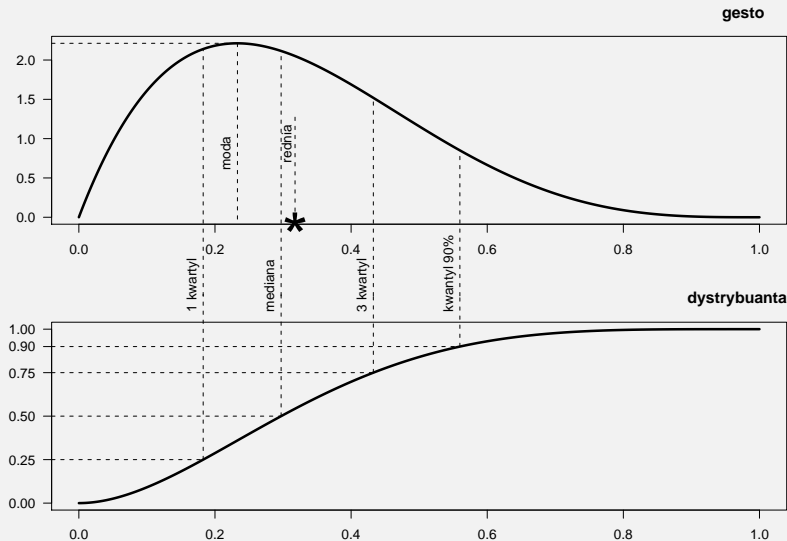
$$\bar{X} = \hat{\mu} = \frac{1}{N} \sum_{i=1}^N X_i.$$

Wariancja w próbie

$$S_X^2 = \hat{\sigma}^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2$$

Odchylenie w próbie

$$S_X = \hat{\sigma} = \sqrt{S_X^2}$$



Przedział ufności to przedział, w którym z określonym prawdopodobieństwem znajduje się prawdziwa wartość parametru z próby.

Jeżeli obserwacje pochodzą z rozkładu normalnego  $X \sim \mathcal{N}(\mu, \sigma^2)$ , to wiadomo, że

$$\bar{X} \sim \mathcal{N}(\mu, \sigma^2/N).$$

Przedział ufności dla średniej można wyznaczyć ze wzoru

$$\mu \in_{95\%} \left( \bar{X} - q_{0.025} \frac{S_X}{\sqrt{N}}, \bar{X} + q_{0.975} \frac{S_X}{\sqrt{N}} \right).$$

Błąd standardowy dla średniej wyznaczamy jako  $S_X/\sqrt{N}$ .

Kowariancje pomiędzy dwiema zmiennymi wyznaczyć można ze wzoru

$$\text{Cov}(X, Y) = \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y}).$$

Korelacje Pearsona pomiędzy dwiema zmiennymi wyznaczyć można ze wzoru

$$\text{Cor}(X, Y) = \frac{\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^N (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^N (Y_i - \bar{Y})^2}}.$$

# Współczynnik korelacji Spearmana

Współczynnik korelacji Spearmana można wyznaczyć zamieniając wartości na rangi.

$$\text{Cor}_{\text{spearmana}}(X, Y) = \text{Cor}(r(X), r(Y)),$$

Gdzie  $r(X_i)$  odpowiada randze obserwacji  $X_i$  w uporządkowanej próbie, czyli

$$r(x_i) = \sum_{j=1}^N \mathbb{1}_{x_i \geq x_j}.$$

Model regresji prostej, jest postaci:

$$y = \beta_0 + \beta_1 x + \varepsilon,$$

gdzie  $y$  to zmienna objaśniana,  $x$  zmienna objaśniająca a  $\varepsilon$  to zakłócenie losowe.

Postać modelu jest liniowa, a zakłócenia  $\varepsilon$  są niezależne, mają rozkład normalny, średnią 0 i stałą wariancję.

Oceny tych współczynników możemy wyznaczyć ze wzorów

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{\text{cov}(x, y)}{\text{var}(x)},$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Do oceny dopasowania wykorzystywany jest współczynnik  $R^2$ , nazywany współczynnikiem determinacji.

Przedstawia on procent wariancji wyjaśnionej przez model

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y})^2}{\sum_i (y_i - \bar{y})^2}.$$

Wysoka wartość tego współczynnika (bliska 1) oznacza, że użyty model dobrze i wyczerpująco wyjaśnia zmienność w danych.

Niska wartość tego współczynnika (bliska 0) oznacza, że użyty model wyjaśnia niewielki fragment całej zmienności.

Testowanie to bardzo szeroka dziedzina, testy które poznaliśmy to jedynie pakiet podstawowy. Większość hipotez dotyczy równości pewnych parametrów.

$$H_0 : \theta_X = \theta_Y.$$

Za alternatywę, najczęściej wybiera się jedną z trzech hipotez

- alternatywa dwustronna

$$H_{A1} : \theta_x \neq \theta_y,$$

- alternatywa jednostronna

$$H_{A2} : \theta_x > \theta_y,$$

$$H_{A3} : \theta_x < \theta_y.$$

Dla danych obserwacji przeprowadzić test można bazując na wartości statystyki testowej, lub p-wartości.

P-wartość (ang. p-value) jest równa najmniejszemu poziomowi istotności, na którym dla wyniku  $X$  odrzuca się hipotezę  $H_0$ .

Do testowania hipotezy

$$H_0 : \sigma_1^2 = \sigma_2^2$$

gdzie  $\sigma_i^2$  to wariancja w grupie  $i$ , wykorzystuje się test oparty o statystykę testową

$$T(X) = \frac{S_1^2}{S_2^2}$$

(większą wariancję zawsze wpisujemy do licznika).

Przy prawdziwej hipotezie zerowej statystyka ta ma rozkład normalny  $\mathcal{F}(n_1 - 1, n_2 - 1)$ . Obszary krytyczne wyznacza się ze wzorów

- dla dwustronnej hipotezy alternatywnej !!!

$$W_\alpha = [f_{1-\alpha/2}^{n_1-1, n_2-1}, \infty)$$

- dla jednostronnej hipotezy alternatywnej

$$W_\alpha = [f_{1-\alpha}^{n_1-1, n_2-1}, \infty).$$

Jest wiele testów do testowania średnich. Aby wybrać właściwy należy odpowiedzieć sobie na pytania:

- Czy zmienne mają rozkład normalny czy nie?
- Czy porównywana jest średnia z zadaną stałą, czy porównywane są dwie średnie?
- Czy dane są sparowane (związane) czy nie?
- Czy wariancja w grupach jest znana czy nie?
- Czy wariancje są takie same czy są różne?

# Test t-Studenta, gdy wariancja jest znana

Do testowania wartości średniej w podpopulacji, w sytuacji gdy wariancja jest znana wykorzystuje się test oparty na statystyce testowej

$$T(X) = \frac{\bar{X} - \mu_0}{\sigma} \sqrt{n}.$$

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

# Test t-Studenta, gdy wariancja jest nie znana

Do testowania wartości średniej w podpopulacji, w sytuacji gdy wariancja jest nieznana wykorzystuje się test t-Studenta oparty na statystyce testowej

$$T(X) = \frac{\bar{X} - \mu_0}{S} \sqrt{n}.$$

Przy prawdziwej hipotezie zerowej statystyka ta ma rozkład t-Studenta o  $n - 1$  stopniach swobody.

# Test t-Studenta, 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 testem **U**.

# Test t-Studenta, dwie próby o nie znanej ale równej wariancji

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.

# Test t-Studenta, 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  $x$ .

# Test t-Studenta, próby sparowane (zależne)

Jeżeli dwie serie pomiarowe dotyczą tych samych obiektów, a więc wartości pomiarów są zależne, należy zastosować test dla danych sparowanych.

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.

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 testem z.

# Test dla proporcji - duże próby

W dużych próbach rozkład częstości przybliżyć można rozkładem normalnym. Do testowania hipotezy

$$H_0 : p = p_0$$

gdzie  $p_0$  zadana wartość, wykorzystać można test oparty na statystyce testowej

$$T(X) = n \frac{p - p_0}{\sqrt{p_0(1 - p_0)n}}.$$

Przy prawdziwej hipotezie zerowej statystyka ta ma rozkład normalny  $\mathcal{N}(0, 1)$ . Obszary krytyczne wyznacza się ze wzorów

- dla dwustronnej hipotezy alternatywnej
$$W_\alpha = (-\infty, q_{\alpha/2}] \cup [q_{1-\alpha/2}, \infty),$$
- dla lewostronnej hipotezy alternatywnej
$$W_\alpha = (-\infty, q_\alpha],$$
- dla prawostronnej hipotezy alternatywnej
$$W_\alpha = [q_{1-\alpha}, \infty).$$

W dużych próbach rozkład częstości przybliżyć można rozkładem normalnym. Do testowania hipotezy

$$H_0 : p_1 = p_2,$$

wykorzystać można test oparty na statystyce testowej

$$T_1(X) = \frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

lub

$$T_2(X) = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

Przy prawdziwej hipotezie zerowej statystyka ta ma rozkład normalny  $\mathcal{N}(0, 1)$ . Obszary krytyczne wyznacza się jak dla testu dla jednej próby.

Nieparametryczny odpowiednik testu t Studenta.

W wersji sparowanej hipoteza zerowa ma postać

$$H_0 : med_{Y-X} = 0$$

gdzie  $med_{Y-X}$  to mediana różnic  $d_i = Y_i - X_i$ . Do testowania wykorzystuje się statystykę testową

$$S = \min(W^+, W^-)$$

gdzie

$$W^+ = \sum_{d_i > 0} r(d_i), \quad W^- = \sum_{d_i < 0} r(d_i)$$

a  $r(d_i)$  to ranga wartości  $d_i$  wyznaczona wektorze wartości bezwzględnych  $|d_i|$ .

Dla dużych prób ( $n > 20$ ) statystykę  $S$  można przybliżyć rozkładem normalnym o średniej  $\frac{n(n+1)}{4}$  i wariancji  $\frac{n(n+1)(2n+1)}{24}$ .

Nieparametryczny odpowiednik testu t Studenta.  
Hipoteza zerowa ma postać

$$H_0 : \theta_X = \theta_Y$$

gdzie  $\theta_X$  to mediana dla populacji  $X$  a  $\theta_Y$  dla  $Y$ .  
Do testowania wykorzystuje się statystykę testową

$$U = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} 1_{X_i < Y_j}$$

Dla dużych prób ( $n > 20$ ) statystykę  $U$  można przybliżyć rozkładem normalnym o średniej  $\frac{n_1 n_2}{2}$  i wariancji  $\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$ .

Nieparametryczny odpowiednik testu t Studenta.

Hipoteza zerowa ma postać

$$H_0 : med_X = \theta$$

gdzie  $med_X$  to mediana dla populacji  $X$  a  $\theta$  t pewna liczba.

Do testowania wykorzystuje się statystykę testową

$$B = \sum_{i=1}^N x_i > \theta,$$

czyli liczbę przypadków większych od  $\theta$ . Dla prawdziwej hipotezy zerowej, ta statystyka ma rozkład dwumianowy  $\mathcal{B}(N, 0.5)$ .

Dla dużych prób ( $n > 20$ ) statystykę  $B$  można przybliżyć rozkładem normalnym o średniej  $N/2$  i wariancji  $N/4$ .

Do testowania hipotezy

$$H_0 : X \sim F$$

wykorzystuje się test oparty o statystykę testową

$$T = \sum \frac{(O - E)^2}{E} = \sum_{i=1}^k \frac{(n_i - E_i)^2}{E_i}$$

gdzie

$$E_i = p_i \sum_{j=1}^k n_{ij}.$$

Przy prawdziwej hipotezie zerowej statystyka ta ma rozkład  $\chi^2_{(k-1)}$  ze  $(k - 1)$  stopniami swobody.

Obszary krytyczne wyznacza się ze wzoru

$$W_\alpha = [\chi^2_{1-\alpha, (k-1)}, \infty).$$

# Test zgodności, test Kołomogorova-Smirnova

Do testowania hipotezy

$$H_0 : X \sim F$$

wykorzystuje się test oparty o statystykę testową

$$D_n = \sup_x |F_n(x) - F(x)|$$

gdzie  $F_n(x)$  to dystrybuanta empiryczna zadana wzorem

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n I_{X_i \leq x}.$$

$$\sqrt{n}D_n \xrightarrow{n \rightarrow \infty} \sup_t |B(F(t))|$$

Kwantyli rozkładu tej statystyki testowej najlepiej szukać w tablicach.

Do testowania hipotezy

$$H_0 : \text{brak obserwacji odstających}$$

przy dwustronnej alternatywie wykorzystać można test oparty na statystyce testowej

$$T(X) = \frac{\max |X_i - \bar{X}|}{S_X}.$$

Wartość krytyczną dla tego testu wyznacza się ze wzoru

$$c_\alpha = \frac{N-1}{\sqrt{N}} \sqrt{\frac{t_{\alpha/(2N), N-2}^2}{N-2 + t_{\alpha/(2N), N-2}^2}}$$

gdzie  $t_{\alpha/(2N), N-2}$  to kwantyl rzędu  $1 - \alpha/(2N)$  rozkładu t-Studenta o  $N-2$  stopniach swobody.

Dla jednostronnej alternatywy, wykorzystuje się kwantyl rzędu  $t_{\alpha/N, N-2}$ .

# Testy niezależności, test $\chi^2$

Do testowania hipotezy

$$H_0 : X \text{ niezależne od } Y$$

wykorzystuje się test oparty o statystykę testową

$$T = \sum \frac{(O - E)^2}{E} = \sum_{i=1}^k \sum_{j=1}^p \frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

gdzie

$$E_{ij} = \frac{\sum_{i=1}^k n_{ij} \sum_{j=1}^p n_{ij}}{\sum_{i=1}^k \sum_{j=1}^p n_{ij}}.$$

Przy prawdziwej hipotezie zerowej statystyka ta ma rozkład

$\chi^2_{(k-1)(p-1)}$  ze  $(k-1)(p-1)$  stopniami swobody.

Obszary krytyczne wyznacza się ze wzoru

$$W_\alpha = [\chi^2_{1-\alpha}, \infty).$$

# Testy niezależności oparte na współczynniku korelacji Pearsona

Do testowania hipotezy

$$H_0 : X \text{ niezależne od } Y, \rho_{X,Y} = 0$$

wykorzystuje się test oparty o transformacje Fishera

$$f(\rho) = \frac{1}{2} \ln \left( \frac{1 + \rho}{1 - \rho} \right).$$

Przyjmuje się, że zmienna  $f(\rho)$  ma w przybliżeniu rozkład normalny o wariancji  $1/(N - 3)$ .

Do testowania wartości korelacji za statystykę testową przyjmuje się

$$z = \frac{f(\hat{\rho}) - f(\rho_0)}{\sqrt{1/(N - 3)}},$$

ta statystyka testowa ma asymptotyczny rozkład normalny.

Do testowania hipotezy

$$H_0 : b \text{ występuje równie często jak } c$$

wykorzystuje się test oparty o statystykę testową

$$T = \frac{(b - c)^2}{b + c}.$$

Przy prawdziwej hipotezie zerowej statystyka ta ma rozkład  $\chi_1^2$  z 1 stopniem swobody.

Obszary krytyczne wyznacza się ze wzoru

$$W_\alpha = [\chi_{1-\alpha}^{2,1}, \infty).$$

W modelu regresji liniowej możemy weryfikować, czy dany współczynnik jest istotnie różny od zera.

$$H_0 : \beta_1 = 0,$$

$$H_A : \beta_1 \neq 0.$$

Za statystykę testową wybiera się

$$T = \frac{\hat{\beta}_1}{\hat{\sigma}} \sqrt{\sum_i (x_i - \bar{x})^2}.$$

Ta statystyka testowa ma rozkład t-Studenta z  $n - 2$  stopniami swobody (nie będziemy z niej korzystać).

Do testowania hipotezy

$H_0$  : kolejne obserwacje są niezależne

można test serii oparty na statystyce testowej

$T(X)$  = liczba serii.

Przy prawdziwej hipotezie zerowej, liczba serii ma rozkład normalny o średniej

$$\mu = 1 + \frac{2N_R N_O}{N}$$

i wariancji

$$\sigma^2 = \frac{(\mu - 1)(\mu - 2)}{N - 1}.$$

Wartości krytyczne możemy odczytywać z tablic dla rozkładu normalnego.