

# Statystyka w analizie i planowaniu eksperymentu

Wykład 9

Przedziały ufności i błąd standardowy.

Przemysław Biecek

Dla 1 roku studentów Biotechnologii

Proszę na (niewielkiej) kartce napisać:

- 1 Imię, nazwisko,
- 2 Nr. indeksu,
- 3 Nazwisko osoby prowadzącej ćwiczenia

Wybierz dwie ostatnie różne cyfry swojego numeru indeksu.

Badamy płęć rodzeństwa w rodzinach dwudziętych.

pierwsze dziecko / drugie dziecko	M	K
M	20	10
K	XX	30

Na poziomie istotności  $\alpha = 0.05$  odpowiedz na następujące pytania

- Czy płęć jednego dziecka zależy od płci drugiego dziecka?
- Czy więcej jest rodzeństw dwóch chłopców czy rodzeństw dwóch dziewczynek?

- Wykonać analizę statystyczną danych dotyczących pacjentów oddziału Nefrologii.
- Użyć narzędzi statystycznych i zaprezentować wynik w czytelnej postaci.
- Interesuje nas przede wszystkim zależność zmiennej Kreatynina od innych zmiennych.
- Najlepsze opracowanie nagrodzone +1 do oceny, wszystkie poprawne i ciekawe opracowania +0.5.

Dane dotyczą

- Wiek i płeć pacjenta,
- Informacje o czasie pomiędzy operacją u dawcy do czasu operacji u biorcy (WIT i CIT).
- Poziomy Kreatyniny, Mocznika i GFR u pacjentów w 1, 3 i 7 dobie po zabiegu.

```
> dane = read.table("http://biecek.pl/statystyka/daneBioTech.csv", header=T,  
sep=";", dec=",")
```

```
> summary(dane)
```

Wiek	Płeć.K.O.M.1	WIT	CIT..h.	Kreatynina.1
Min. :24.00	K: 9	brak :16	Min. :13.50	Min. : 2.200
1st Qu.:42.00	M:15	obecny: 8	1st Qu.:19.00	1st Qu.: 4.725
Median :53.50			Median :21.50	Median : 6.900
Mean :50.54			Mean :21.42	Mean : 6.429
3rd Qu.:58.00			3rd Qu.:23.50	3rd Qu.: 8.000
Max. :70.00			Max. :31.00	Max. :10.400
Kreatynina.3	Kreatynina.7	Mocznik.1	Mocznik.3	
Min. : 1.100	Min. :0.600	Min. : 8.70	Min. : 7.30	
1st Qu.: 2.625	1st Qu.:1.650	1st Qu.:13.55	1st Qu.:12.65	
Median : 4.550	Median :2.550	Median :16.25	Median :18.55	
Mean : 5.088	Mean :3.421	Mean :17.46	Mean :18.30	
3rd Qu.: 7.600	3rd Qu.:4.625	3rd Qu.:21.77	3rd Qu.:22.45	
Max. :10.000	Max. :9.600	Max. :31.80	Max. :31.40	

Zadanie:

Zmierzono ekspresje genu BRCA1 u 10 pacjentek. Wyniki to

$$X = 4, 15, 9, 16, 6, 5, 16, 4, 11, 8, 35$$

Pytanie:

- Czy któraś z obserwacji nie jest obarczona błędem grubym?
- Ile obserwacji jest obarczonych błędem?

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  $t_{\alpha/N, N-2}$ .

Obserwacje

$$X = 4, 15, 9, 16, 6, 5, 16, 4, 11, 8, 35$$

Liczymy średnia i odchylenie standardowe

$$\bar{X} = 11.72, S_X = 8.99$$

Wartości  $|X_i - \bar{X}|/S_X$

$$1.17, 0.49, 0.41, 0.64, 0.86, 1.02, 0.64, 1.17, 0.11, 0.56, 3.53.$$

Ponieważ  $t_{1-0.05/22,9} = 3.75$  to

$$c_{0.05} = \frac{10}{\sqrt{11}} \sqrt{\frac{t_{1-0.05/22,9}^2}{9 + t_{1-0.05/22,9}^2}} = 2.35.$$

Jaka jest nasza decyzja?

Do testowania hipotezy

$H_0$  : brak wartości odstającej

wobec alternatywy z jedną wartością odstającą wykorzystać można test oparty na statystyce testowej

$$T_{3-7}(X) = \frac{x_2 - x_1}{x_N - x_1}.$$

$$T_{8-10}(X) = \frac{x_2 - x_1}{x_{N-1} - x_1}.$$

$$T_{11-13}(X) = \frac{x_3 - x_1}{x_{N-1} - x_1}.$$

$$T_{14-30}(X) = \frac{x_3 - x_1}{x_{N-2} - x_1}.$$

W indeksie dolnym  $T$  podane jest dla jakich liczebności należy stosować dany wariant statystyki testowej.

Wartości krytyczne dla testu Dixona należy odczytać z tablic.

Obserwacje

$$X = 4, 15, 9, 16, 6, 5, 16, 4, 11, 8, 35$$

Po uporządkowaniu

$$sX = 4, 4, 5, 6, 8, 9, 11, 15, 16, 16, 35$$

Podejrzana jest obserwacja ostatnia 11, liczymy

$$T_{11-13}(X) = \frac{x_3 - x_1}{x_{N-1} - x_1} = (16 - 35)/(4 - 35) = 0.612$$

Porównujemy z odpowiednim kwantylem z tablic  $q = 0.576$ .

Jaka jest nasza decyzja?

Zadanie:

Proszę w domu 1000 razy rzucić symetryczną monetą, i zapisać wyniki kolejnych rzutów w postaci

*ROOOORORRORORORROROROOOOORRORRRROOORRRRORRRRO....*

Pytanie:

- Czy prowadzący jest w stanie rozpoznać, czy student sumiennie rzucał monetą czy wyniki zmyślił?

Do testowania hipotezy

$H_0$  : kolejne obserwacje sa niezależne

można test serii oparty na statystyce testowej

$$T(X) = \text{liczba serii.}$$

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 więc odczytywać z tablic dla rozkładu normalnego.

Przykładowe wyniki rzutów

*ROOOORORRORORORROROOOOORRORRRROOORRRR*

Liczba serii = 17 (seria to blok takich samych wartości).

Liczebności  $N_O = 18$ ,  $N_R = 17$ , wyznaczamy średnią i wariancję  
 $\mu = 18.5$ ,  $\sigma^2 = 8.478$ ,  $\sigma = 2.912$ .

Odczytujemy wartość krytyczną z tablic

$$W = (q_{0.025}, q_{0.975}) = c(12.78, 24.19)$$

W R test serii zaimplementowany jest w funkcji  
`runs.test(lawstat)`.

Jeżeli obserwacje nie mają dychotomicznego charakteru (ale np. liczbowy) to aby użyć testu serii można zmienną liczbową zamienić na binarną, określając czy dana wartość jest większa/mniejsza od średniej lub mediany.

Tego typu zabieg jest często wykorzystywany, np. w teście znaków.

Zobaczmy jak używając testu znaków sprawdzić czy średnia danej cechy jest istotnie różna od określonej wartości (test na wartość średnią).

Zmierzyliśmy ekspresję BRCA1, policzmy wartość średnią i odchylenie standardowe

$$X = 4, 15, 9, 16, 6, 5, 16, 4, 11, 8$$

$$\bar{X} = 9.4$$

$$S_X = 4.88$$

Na ile jesteśmy pewni tych wyników? Na ile są one charakterystyczne dla populacji.

Przedział ufności dla parametru to przedział (wyznaczony na bazie obserwacji), w którym z określonym prawdopodobieństwem znajduje się prawdziwa wartość parametru.

Jeżeli obserwacje pochodzą z rozkładu normalnego, to

$$E(\bar{X}) = \mu,$$

$$\text{Var}(\bar{X}) = \sigma^2/N.$$

A więc dla naszych pomiarów

$$\bar{X} \sim N(9.4, 4.88/\sqrt{N})$$

W powyższym wzorze przyjmujemy, że próba jest duża, dla małych prób powinniśmy użyć kwantyli z rozkładu t-Studenta. Z prawdopodobieństwem 0.95% możemy stwierdzić, że

$$\mu \in (6.38, 12.42)$$

Przedział  $(6.38, 12.42)$  jest 95% przedziałem ufności dla parametru średniej w naszej populacji.

Podobne postępowanie możemy przeprowadzić dla parametru wariancji, musimy tylko wiedzieć jaki rozkład ma wariancja. Jeżeli obserwacje pochodzą z rozkładu normalnego, to wiadomo, że

$$E(S_X^2) = \frac{N-1}{N} \sigma^2,$$

$$\text{Var}(S_X^2) = \frac{2(N-1)}{N^2} \sigma^4.$$

Znając rozkład wariancji możemy określić przedział ufności dla otrzymanej wariancji w próbie.

- Błąd standardowy to odchylenie standardowe dla wartości średniej.
- Błąd standardowy NIE JEST równy odchyleniu standardowemu.
- Oznaczamy go symbolami  $\sigma_{\bar{X}}$  lub  $s_{\bar{X}}$ .

Policzmy błąd standardowy dla rozkładu normalnego i dwumianowego.

# Co trzeba zapamiętać?

- Jak działa i po co jest test serii?
- Jak działa i po co jest test Grubbsa?
- Jak działa i po co jest test Dixona?
- Po co jest przedział ufności?
- Po co jest błąd standardowy i jak ma się do odchylenia standardowego?