ADP – Projekt 3

(2022Z)

Zad. 1. W tabeli zestawiono wyniki pomiarów stężenia glukozy we krwi u osób zdrowych (L0=bez podawania żadnego leku) i leczonych lekami L1 i L2.

lek	stężenie glukozy [mmol/l]						
L0	**4,5**	**4,8**	**6,1**	**5,7**	**5,2**	**4,6**	**6,6**
L1	**5,2**	**5,3**	**6,7**	**7,2**	**6,9**	**6,8**	**5,4**
L2	**5,1**	**4,4**	**4,9**	**5,1**	**5,6**	**5,7**	**6,1**

Zweryfikuj, na poziomie istotności α=0,01; α=0,05 oraz α=0,1 hipotezę H_0 o braku wpływu leków na stężenie glukozy we krwi.

Czynimy założenie, że badane zmienne losowe podlegają rozkładowi normalnemu i mają równe nieznane wariancje, a pomiarów dokonano jedną metodą o stałym błędzie przypadkowym.

Wzory wykorzystane do wyliczenia średnich:

Średnia globalna:

$$\bar{x} = \frac{1}{N}\sum_{i=0}^{r-1}\sum_{j=1}^{N_i} x_{ij}$$

Średnia grupowa:

$$\bar{x}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} x_{ij}$$

gdzie:
N – całkowita liczba pomiarów – N = 21,
N_i– liczba elementów w i-tej grupie – $N_i = N_i = N_i = 7$,
r – liczba badanych grup – r = 3,
x_{ij} – element j-ty w i-tej grupie.

Wyliczone wartości średnie:

$$\bar{x}_0 = \frac{4{,}5 + 4{,}8 + 6{,}1 + 5{,}7 + 5{,}2 + 4{,}6 + 6{,}6}{7} = \frac{37{,}5}{7} = 5{,}3571$$

$$\bar{x}_1 = \frac{5{,}2 + 5{,}3 + 6{,}7 + 7{,}2 + 6{,}9 + 6{,}8 + 5{,}4}{7} = \frac{43{,}5}{7} = 6{,}2143$$

$$\bar{x}_2 = \frac{5{,}1 + 4{,}4 + 4{,}9 + 5{,}1 + 5{,}6 + 5{,}7 + 6{,}1}{7} = \frac{36{,}9}{7} = 5{,}2714$$

$$\bar{x} = \frac{5{,}3571 + 6{,}2143 + 5{,}2714}{3} = \frac{16{,}8428}{3} = 5{,}6143$$

Ze względu na poczynione założenia możemy wykorzystać analizę wariancji metodą Fishera.

Wariancje:

a) Między grupami:

$$s_1^2 = \frac{1}{r-1}\sum_{j=1}^{N_i} N_i \cdot (\bar{x}_i - \bar{x})^2 = \frac{7 \cdot (0{,}0662 + 0{,}36 + 0{,}1176)}{2} = 1{,}9033$$

b) Wewnątrz grup:

$$s_2^2 = \frac{1}{N-r}\sum_{i=0}^{r-1}\sum_{j=1}^{N_i}\left(x_{ij} - \bar{x}_i\right)^2 = 0{,}5744$$

$$F = \frac{s_1^2}{s_2^2} = 3{,}3135$$

Hipoteza H_0: Wartości średnie we wszystkich grupach są równe (lek nie wpływa na poziom glukozy we krwi).
Stopnie swobody: n = r – 1= 2, m = N $-$ r = 18. Wartości zmiennej F dla obecnych w zdaniu stopni swobody dla poziomów istotności 0,01, 0,05 oraz 0,1:

$$F_{0{,}01} = 6{,}013$$
$$F_{0{,}05} = 3{,}555$$
$$F_{0{,}1} = 2{,}624$$

Stosujemy jednostronny test Fishera (P (F ≥ F_α)) ze względu na to, że interesuje nas przypadek, w którym iloraz s_1/s_2 przyjmuje duże wartości, tzn. między grupami występują różnice.

Sprawdzenie hipotezy dla różnych poziomów istotności:

a) dla α = 0,01
   Nie mamy podstaw do odrzucenia hipotezy H_0, ponieważ F < $F_{0{,}01}$.
b) dla α = 0,05
   Nie mamy podstaw do odrzucenia hipotezy H_0, ponieważ F < $F_{0{,}05}$.
c) dla α = 0,1
   Odrzucamy hipotezę H_0, ponieważ F > $F_{0{,}1}$.

Zad. 2. W klinice opracowano test do przewidywania zachorowania niemowląt na żółtaczkę. Ustalono, że wskaźnikami prognozującymi o zachorowalności są: x_0 – wiek matki, x_1 – waga noworodka [g], x_2 – czas trwania ciąży [d], x_3 – stopień zażółcenia skóry. W/w wskaźniki ustalono dla dwóch grup dzieci: grupy badanej – dzieci, które zapadły na żółtaczkę oraz grupy kontrolnej – dzieci, które nie zapadły na żółtaczkę i zebrano w tabelach.

Korzystając z tych danych, dokonać metodą SFS (Sequential Forward Selection), selekcji 3 cech spośród 4 – za kryterium przyjąć miarę dyskryminacyjną T^2.

grupa badana (noworodki, które zachorowały na żółtaczkę)

x_0	x_1	x_2	x_3
28	2430	254	2.0
21	1900	228	4.0
24	1950	224	2.0
21	2350	263	3.0
19	2250	271	2.5
22	2230	255	1.5
22	2300	242	1.5
28	1800	210	4.0
32	2190	235	2.5
22	2050	253	3.0
20	2400	242	3.0

grupa kontrolna (noworodki, które nie zachorowały na żółtaczkę)

x_0	x_1	x_2	x_3
25	2300	247	1.0
17	1800	262	2.5
25	2500	273	0.0
21	2400	263	1.0
24	2090	291	2.0
18	2300	272	2.0
18	2000	272	2.5
32	2230	273	2.0
35	2100	278	2.0
33	2300	273	2.0
22	2230	270	1.5
32	1440	206	2.5
30	2150	261	2.0
28	2100	260	2.0
27	2400	273	1.5
24	2100	275	1.5
19	2450	277	1.0
25	1850	246	2.5
28	2480	246	2.0
34	2450	266	2.0

Dane mamy obserwacje o liczbie cech p = 4 oraz J = 2 klasy zdarzeń (zachorowanie na żółtaczkę j = 1, oraz brak zachorowania j = 2). Początkowo przedstawimy sposób wyznaczania miary dyskryminacyjnej T^2 – dla przykładu dla wszystkich cech. W dalszej części sprawozdania przedstawione zostaną tylko wyniki dla poszczególnych zestawów cech. Sposób obliczenia pozostanie taki sam, z tym tylko, że do obliczenia nie będą użyte wszystkie kolumny cech, a jedynie wybrane.

T^2 dla wszystkich cech

Wartość średnia z próby:

$$\bar{x}_j = \frac{1}{N_j}\sum_{k=1}^{N_j} x_{kj}$$

$$\bar{x}_1 = \frac{1}{11}\sum_{k=1}^{11} x_{k,1} = \begin{bmatrix} 23.5455 \\ 2168.1818 \\ 243.3636 \\ 2.6364 \end{bmatrix} \qquad \bar{x}_2 = \frac{1}{20}\sum_{k=1}^{20} x_{k,2} = \begin{bmatrix} 25.85 \\ 2183.50 \\ 264.20 \\ 1.775 \end{bmatrix}$$

Średnia ogólna z próby:

$$\bar{\boldsymbol{x}} = \frac{1}{N_1 + N_2}\sum_{j=1}^{2} N_j \bar{\boldsymbol{x}}_j = \begin{bmatrix} 25.0324 \\ 2178.0645 \\ 256.8065 \\ 2.0806 \end{bmatrix}$$

Estymaty S_j macierzy kowariancji w poszczególnych klasach

$$\hat{\boldsymbol{\Sigma}}_j = \boldsymbol{S}_j = \frac{1}{N_j - 1}\sum_{k=1}^{N_j} (\boldsymbol{x}_{kj} - \bar{\boldsymbol{x}}_j)(\boldsymbol{x}_{kj} - \bar{\boldsymbol{x}}_j)^T$$

$$\boldsymbol{S_1} = \begin{bmatrix} 16.4727 & -147.9090 & -34.3182 & -0.0318 \\ -147.9090 & 45276.3636 & 2814.7273 & -102.2273 \\ -34.3182 & 2814.7272 & 328.8545 & -6.5045 \\ -0.0318 & -102.2272 & -6.5045 & 0.7545 \end{bmatrix}$$

$$\boldsymbol{S_2} = \begin{bmatrix} 31.6079 & -19.4474 & -21.3368 & 0.5697 \\ -19.4474 & 70760.7895 & 2536.1053 & -112.8553 \\ -21.3368 & 2536.1053 & 317.7474 & -3.1632 \\ 0.5697 & -112.8553 & -3.1632 & 0.4072 \end{bmatrix}$$

Estymator uśrednionej macierzy kowariancji

$$\boldsymbol{S} = \frac{1}{N_1 + N_2 - 2}\sum_{j=1}^{2} (N_j - 1)\boldsymbol{S}_j = \begin{bmatrix} 26.3889 & -63.7445 & -25.8132 & 0.3623 \\ -63.7445 & 61973.0564 & 2632.1818 & -109.1904 \\ -25.8132 & 2632.1818 & 321.5774 & -4.3154 \\ 0.3623 & -109.1904 & -4.3154 & 0.5270 \end{bmatrix}$$

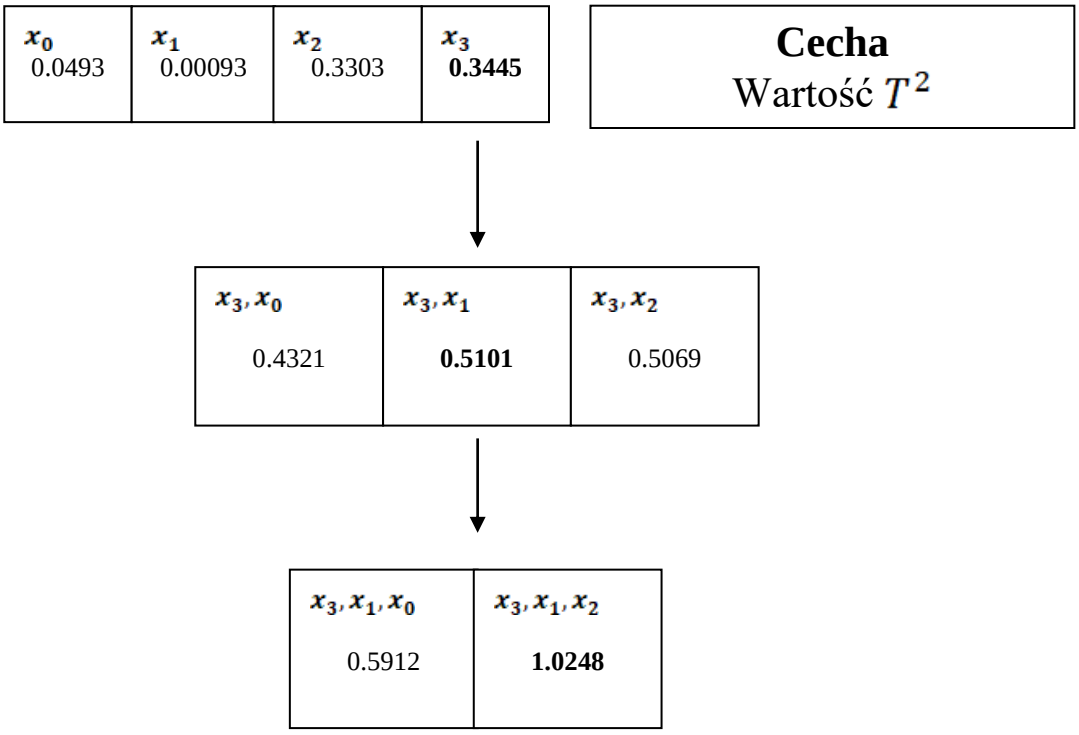
T^2 jest w przypadku dwóch klas równe

$$T^2 = \frac{1}{N_1 + N_2 - 2}\frac{N_1 N_2}{N_1 + N_2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^T \boldsymbol{S}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$$

$$T^2 = 1.3150$$

Algorytm SFS w celu selekcji cech rozpoczyna od pustego zbioru cech. Pierwszą wyselekcjonowaną cechą zostanie ta, dla której wartość T^2 będzie największa. Następnie, spośród pozostałych cech wybierana jest ta, która w połączeniu z pierwszą da największą wartość T^2. I ostatecznie, spośród pozostałych dwóch cech wybierzemy jedną, która zapewni największą wartość T^2.

Cechy:
x0 – wiek matki,
x1 – waga noworodka [g],
x2 – czas trwania ciąży [d],
x3 – stopień zażółcenia skóry.



Z czego wynika że wyselekcjonowanymi cechami są:
x1 – waga noworodka[g],
x2 – czas trwania ciąży [d],
x3 – stopień zażółcenia skóry.