1 Zadanie 10

W obliczu nowego, nieznanego dotąd wirusa, próbowano stworzyć szybszy niż dotychczasowy test wykrywający jego nosicielstwo. Wyniki badań nowym testem uzyskane dla reprezentatywnej próby 1000 osób, wskazały, że wśród osób faktycznie zdrowych nowy test wykazał 680 osób zdrowych i 170 zarażonych; wśród osób faktycznie zarażonych test wskazał 120 osób zarażonych i 30 osób zdrowych.

a) Oblicz wartości czułości i swoistości dla nowego testu.

b) Jaka jest dokładność oraz false discovery rate nowego testu?

c) U losowo wybranego obywatela nowy test wykazał obecność wirusa. Jakie jest prawdopodobieństwo, że ta osoba faktycznie jest zarażona?

Rozwiązanie:

Dane:

Stosując notację z wyjkładu, dostajemy:

TP = 120
FP = 170
FN = 30
TN = 680

N = FP + TN = 850
P = TP + FN = 150

a)
czułość: TPR = $\frac{TP}{P} = 0,8$
swoistość: TNR = $\frac{TN}{N} = 0,8$

b)
dokładność: ACC = $\frac{TP+TN}{P+N} = 0,8$
false discovery rate: FDR = $\frac{FP}{FP+TP} = \frac{17}{29}$

c)
Takie prawdopodobieństwo jest równoznaczne z poziomem precyzji testu.
precyzja: PPV = 1 - FDR = $\frac{12}{29}$

2 Zadanie 7

Pewnej grupie 12 pacjentów leczonych na nadciśnienie podano lek. Wyniki pomiaru ciśnienia krwi w tej grupie były następujące:

	1	2	3	4	5	6	7	8	9	10	11	12
przed lekiem	220	180	270	290	200	300	250	190	220	230	260	270
po leku	190	170	220	260	220	200	260	150	160	170	210	190

• Zakładając, że rozkład ciśnienia krwi jest normalny, zweryfikować hipotezę o nieskuteczności podanego leku. Przyjmij poziom istotności $\alpha = 0,01$.

• Prawdopodobieństwo tego, że lek zadziała jest równe 0,36. Wyznacz takie k, żeby liczba osób, u których lek zadziała w losowej próbie 400 osób należała do przedziału $(144 - k, 144 + k)$ z prawdopodobieństwem 0,9.

Rozwiązanie:

a)

H_0 = lek jest nieskuteczny tzn. średnia w grupie przed podaniem leku i po podaniu nie różnią się

H_1 = średnia w grupie po leku < średnia przed lekiem

Statystyką testową będzie znormalizowana różnica średnich.

Dane pochodzą z rozkładu normalnego o nieznanej wariancji, zatem użyję testu t Studenta na różnicę średnich w dwóch grupach.

Średnie odpowiednio w grupie przed lekiem i po leku wynoszą:

$\bar{x_1}$ = 240, $\bar{x_2}$ = 200, a liczebności grup N_1 = 12, N_2 = 12.

Wariancje w grupach zgodnie ze wzorem

$$s^2_{x_i} = \frac{1}{(N_i - 1)} \sum_{j=1}^{N_i} (x_j - \overline{x}_i)^2 \tag{1}$$

wynoszą odpowiednio

$s^2_{x_1}$ = 1545,45, $s^2_{x_2}$ = 1290,91.

Wariacja różnicy wartości średnich to suma wariancji średnich

$$s^2_\Delta = s^2_{\overline{x}_1} + s^2_{\overline{x}_2} = \frac{s^2_{x_1}}{N_1} + \frac{s^2_{x_2}}{N_2}, \tag{2}$$

tutaj $s^2_\Delta = 236,36$

Statystyka testowa t_1 wynosi

$$t_1 = \frac{\overline{x}_1 - \overline{x}_2}{s_\Delta} = 2,60 \tag{3}$$

Wartość rozkładu t Studenta dla liczby stopni swobody $f = N_1 + N_2 - 2 = 22$ oraz dla poziomu istotności $\alpha = 0,01$ wynosi

$$t[22, 0, 01] = 2,50832 \tag{4}$$

$t_1 > t[22, 0, 01]$ zatem na poziomie istotności 0,01 odrzucam hipotezę zerową o braku różnic pomiędzy średnimi w dwóch grupach, co oznacza, że terapia pomagała obniżyć ciśnienie.

b)

Lek pomoże średnio

400 · 36% = 144

osobom.

$\bar{x}$ = 144

Estymatorem wariancji średniej będzie wariancja średniej w grupie po leku z podpunktu a) $\frac{s^2_{x_2}}{N_2}$

Przedział ufności dla tej średniej to

$$\left[\bar{x} - t_{\frac{\alpha}{2}} \frac{s_{x_2}}{\sqrt{N_2}}, \quad \bar{x} + t_{\frac{\alpha}{2}} \frac{s_{x_2}}{\sqrt{N_2}}\right], \tag{5}$$

gdzie $\alpha = 0,1$
Zatem $k = t_{\frac{\alpha}{2}} \frac{s_{x_2}}{\sqrt{N_2}}$
$t[N_2 - 1, \alpha/2] = t[11, 0, 05] = 1,79588$
$k = 1,79588 \cdot \frac{\sqrt{1290{,}91}}{\sqrt{12}} = 18,63$

3 Zadanie 12

Dane są trzy próby:
(a) 6,6 ; 0,6 ; 3,4 ; 1,8 ; -0,6; 4,2 ; 2,6 ; 5,4
(b) 2,62; 0,92; 1,75; 4,84; 0,26; 1,27
(c) -1,5; 0,3; 0,8; 2,0; -2,0; -0,8; 0,6; -0,6; 1,5; -0,3
Wykorzystując test Kołmogorowa zweryfikuj hipotezy:

• Próbę (a) pobrano z populacji o rozkładzie $N(3, 2^2)$, poziom istotności $\alpha = 0,05$.

• Próbę (b) pobrano z populacji o rozkładzie $f(x) = \frac{1}{2}e^{-\frac{x}{2}}$, x > 0, poziom istotności α = 0,1.

• Próbę (c) pobrano z populacji o rozkładzie $N(0, 1)$, poziom istotności α = 0,1.

Rozwiązanie:

(a)

H_0 = próba (a) została wylosowana z rozkładu normalnego o wartości oczekiwanej 3 i odchyleniu standardowym 2

H_1 = próba (a) została wylosowania z innego rozkładu niż rozkład normalny o wartości oczekiwanej 3 i odchyleniu standardowym 2

Wartości z próby (a) sortuję rosnąco oraz normalizuję, odejmując średnią i dzieląc przez odchylenie standardowe, aby przeskalować wartości (a) na wartości z rozkładu $N[0, 1]$

$$a\_norm_i = \frac{a_i - 3}{2} \tag{6}$$

i otrzymuję

$$a\_norm = [-1,8; -1,2; -0,6; -0,2; 0,2; 0,6; 1,2; 1,8] \tag{7}$$

Następnie będę porównywać dystrybuantę empiryczną próby (a) z dystrybuantą $N[3, 2^2]$. Wartości dystrybuant wyznaczę dla wartości a_norm. Dystrybuanta empiryczna wynosi

$$cdf\_emp = [0,125; 0,25; 0,375; 0,5; 0,625; 0,75; 0,875; 1], \tag{8}$$

a dystrybuanta teoretyczna

$$cdf\_theory = [0,0359; 0,1151; 0,2743; 0,4207; 0,5793; 0,7257; 0,8849; 0,9641] \quad (9)$$

Obliczam wartość bezwzględną różnicy pomiędzy odpowiadającymi sobie wartościami dystrybuant i szukam maksymalnej wartości, która wynosi 0,1349.

Wartość statystyki testowej w teście Kołmogorowa wynosi zatem

$$D_1 = \sqrt{8} \cdot 0,1349 = 0,3816 \quad (10)$$

W tablicach sprawdzam wartości funkcji odwrotnej do dystrybuanty dla rozkładu Kołmogorowa (liczonego z wartością bezwzględną dla małych n) dla liczby badanych punktów $n = 8$ i dla $\alpha =$ 0,05, która wynosi $D\_kryt = 0,45427$.

$D_1 < D\_kryt$, co oznacza, że nie mogę odrzucić hipotezy zerowej o tym, że dane z próbki (a) pochodzą z rozkładu $N[3, 2^2]$.

b)

Posortowana próba: $b = [0.26; 0.92; 1.27; 1.75; 2.62; 4.84]$

$cdf\_emp = [0.1667; 0.3333; 0.5000; 0.6667; 0.8333; 1.0000]$

$cdf\_theory = [0.1219; 0.3687; 0.4701; 0.5831; 0.7302; 0.9111]$

$D = 0.2527$

$D\_kryt[6, \alpha = 0.1] = 0.46799$

$D < D\_kryt$ zatem na poziomie istotności 0.1 nie mogę odrzucić H_0

c)

H_0 = próba (c) została wylosowana z $N[0, 1]$

H_1 = próba (c) została wylosowania z innego rozkładu niż $N[0, 1]$

Analogicznie do podpunktu a), tylko nie ma konieczności normowania.

Posortowana próba: $c = [-2; -1,5; -0,8; -0,6; -0,3; 0,3; 0,6; 0,8; 1,5; 2]$

$cdf\_emp = [0.1; 0.2; 0.3; 0.4; 0.5; 0.6; 0.7; 0.8; 0.9; 1]$

$cdf\_theory = [0.0228; 0.0668; 0.2119; 0.2743; 0.3821; 0.6179; 0.7257; 0.7881; 0.9332; 0.9772]$

$D = 0,4212$

$D\_kryt[10, \alpha = 0,1] = 0,36866$

$D > D\_kryt$ zatem na poziomie istotności 0,1 odrzucam H_0