

Statystyczne modelowanie danych biologicznych

Egzamin

Zadanie 1. Znaleźć rozkład macierzy $\Sigma = \begin{bmatrix} 4 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 4 \end{bmatrix}$ na iloczyn macierzy dolno- i górno-trójkątnej przy pomocy metody Choleskiego.

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix} = Q Q^T, \text{ gdzie } Q = \begin{bmatrix} q_{11} & 0 & 0 \\ q_{21} & q_{22} & 0 \\ q_{31} & q_{32} & q_{33} \end{bmatrix} \text{ jest macierzą dolno-trójkątną.}$$

$$q_{11} = \sqrt{\sigma_{11}} = \sqrt{4} = 2, q_{21} = \frac{\sigma_{12}}{q_{11}} = \frac{1}{2} = 0.5, q_{22} = \sqrt{\sigma_{22} - q_{21}^2} = \sqrt{2 - 0.25} = \sqrt{1.75} \approx 1.323$$

$$q_{31} = \frac{0}{q_{11}} = 0, q_{32} = \frac{\sigma_{23} - q_{31} \cdot q_{21}}{q_{22}} \approx \frac{1 - 0 \cdot 0.5}{1.323} \approx 0.75586,$$

$$q_{33} = \sqrt{\sigma_{33} - (q_{31}^2 + q_{32}^2)} \approx \sqrt{4 - 0 - 0.5713212} \approx \sqrt{3.4286788} \approx 1.851669$$

Za każdą policzoną wartość były przyznawany był 1 punkt. łącznie 6 punktów.

Zadanie 2. Wypełnij brakujące dane w tabeli, a następnie odpowiedz na poniższe pytania.

Czas przeżycia (miesiące)	Ilość pacjentów (n_t)	Obserwowana ilość zgonów (d_t)	$p_t = 1 - \frac{d_t}{n_t}$	Funkcja przeżycia $S(t)$
0	30	0	1.00	1.000
3	30	3	0.90	1.00x0.90=0.900
6	25	3	0.88	0.90x0.88=0.792
9	22	5	0.773	0.792x0.773=0.612
12	15	4	0.733	0.612x0.733=0.449
18	10	1	0.90	0.449x0.90=0.404
24	8	0	1.00	0.404x1.00=0.404

- Ilu było pacjentów cenzorowanych? Było 6 pacjentów cenzorowanych (30-16-8=6)
- Ile wynosi mediana czasu przeżycia? Mediana wynosi 12 miesięcy. Wystarczy narysować funkcję przeżycia na podstawie tabeli i odczytać dla jakiej wartości na osi X funkcja $S(t)$ przyjmuje wartość 0.5.
- Obliczyć funkcję hazardu dla pierwszego i drugiego roku.

W przedziale czasu $[0,12)$ mamy 11 zgonów. Osoby które zmarły przeżyły $f_1=3 \times 3 + 3 \times 6 + 5 \times 9 = 72$ miesiące. 4 osoby cenzorowane przeżyły $F_1=2 \times 3 + 2 \times 9 = 24$ miesiące. Pozostałe przy życiu 15 osób przeżyły $G_1=15 \times 12 = 180$ miesięcy. Wszystkie osoby w pierwszym roku przeżyły $f_1 + F_1 + G_1 = 72 + 24 + 180 = 276$ miesięcy. Funkcja hazardu dla pierwszego roku wynosi $H = 11/276 = 0.039855$ na miesiąc. Co w skali rocznej daje $0.039855 \times 12 = 0.4782609$.

Dla drugiego roku sytuację mamy analogiczną. W przedziale czasu $[12,24)$ mamy 5 zgonów. Osoby które zmarły przeżyły $f_2=4 \times 1 + 1 \times 7 = 11$ miesięcy. 2 osoby cenzorowane przeżyły $F_2=1 \times 1 + 1 \times 7 = 8$ miesięcy. Pozostałe przy życiu 8 osób przeżyły $G_2=8 \times 12 = 96$ miesięcy.

miesiący. Wszystkie osoby w drugim roku przeżyły $f_2 + F_2 + G_2 = 11 + 8 + 96 = 115$ miesięcy. Funkcja hazardu dla drugiego roku wynosi $H = 5/115 = 0.04347826$ na miesiąc. Co w skali rocznej daje $0.04347826 \times 12 = 0.5217391$.

d) Wyznaczyć 95% przedział ufności dla $S(12)$ przy pomocy metody Greenwooda.

Wiemy, że $S(12)$ wynosi 0.4488. Ze wzoru na błąd standardowy dla funkcji przeżycia wyznaczony metodą Greenwooda mamy

$$SE_{GR} = 0.4488 \sqrt{\left[\frac{3}{30(30-3)} + \frac{3}{25(25-3)} + \frac{5}{22(22-5)} \right]} = 0.067361.$$

Na tej podstawie wyznaczamy 95% przedział ufności dla $S(12)$

$$(0.4488 - 1.96 \times 0.067361; 0.4488 + 1.96 \times 0.067361) = (0.316773; 0.580827).$$

Za wypełnienie każdej kolumny w tabeli przyznawane były po 2 punkty. Za podpunkt a i b przyznawane było po 1 punkcie. Za podpunkt c i d po 2 punkty. Łącznie 10 punktów.

Zadanie 3. Uzupełnij poniższą tabelę, następnie na podstawie testu log-rank odpowiedz na pytanie, czy płeć pacjenta wpływa na śmiertelność ludzi chorych na ostrą niewydolność serca.

Płeć	Ilość pacjentów	Obserwowana ilość zgonów	Oczekiwana ilość zgonów	HR	$\chi^2_{log-rank}$	p-wartość
kobieta	30	22	20	1.375	0.6	0.438578
mężczyzna	20	8	10			
Suma	50	30	30			

$$HR = \frac{\frac{22}{20}}{\frac{8}{10}} = \frac{22 \cdot 10}{20 \cdot 8} = 1.375$$

$HR > 1$. Ryzyko śmierci u kobiet jest wyższe niż ryzyko śmierci u mężczyzn.

$$\chi^2_{log-rank} = \frac{(22 - 20)^2}{20} + \frac{(8 - 10)^2}{10} = 0.2 + 0.4 = 0.6$$

`pchisq(0.6,1,lower.tail=F)` #p-wartość

P wartość dla testu log-rank wynosi 0.438578, więc na poziomie istotności $\alpha = 0.05$ nie mamy podstaw do odrzucenia hipotezy zerowej, która mówi, że płeć nie ma wpływu na śmiertelność u ludzi o ostrej niewydolności serca.

Za wpisanie poprawnych wartości, HR, $\chi^2_{log-rank}$ i p-wartości przyznawane były po 2 punkty. Dodatkowe 2 punkty można było uzyskać podając interpretację p-wartości i współczynnika HR. Łącznie 8 punktów.

Zadanie 4. Wykazać, że dla funkcji przeżycia $S(t) = \exp(-\lambda t)$ (funkcja przeżycia dla rozkładu wykładniczego z parametrem λ) prawdopodobieństwo śmierci w dowolnym przedziale czasu zależy tylko i wyłącznie od szerokości tego przedziału.

$$S(t) = \exp(-\lambda t)$$

$$\log(S(t)) = -\lambda t$$

$$\text{dla dowolnego } t_2 > t_1 \text{ mamy } \log(S(t_1)) - \log(S(t_2)) = -\lambda t_1 + \lambda t_2 = \lambda(t_2 - t_1)$$

łącznie 4 punkty.

Zadanie 5. W zbiorze danych opisującym ludzi chorych na ostrą białaczkę szpikową było dwustu pacjentów z pośród których stu zmarło. Naukowiec postanowił oszacować funkcję hazardu dla powyższych danych przy pomocy rozkładu wykładniczego i uzyskał $\hat{\lambda} = 0.03$. Dla funkcji hazardu określ 95% przedział ufności i zinterpretuj uzyskany wynik.

$$\hat{\lambda} = 0.03 \quad \text{i} \quad SE(\hat{\lambda}) = \sqrt{\frac{\hat{\lambda}}{d-1}} = \sqrt{\frac{0.03}{99}} = 0.017, \quad CI(95\%) = (0.03 - 1.96 * 0.017; 0.03 + 1.96 * 0.017) = (-0.004; 0.064)$$

Zadanie 6. Opisać proces losowania liczb pseudolosowych z rozkładu Weibulla z parametrem $\lambda = 3$ i $k = 2$ przy pomocy dystrybuanty odwrotnej. Dystrybuanta dla rozkładu Weibulla ma postać: $F(x) = 1 - \exp\left(-\left(\frac{x}{\lambda}\right)^k\right)$. Wyznaczyć dwie zmienne pseudolosowe dla tego rozkładu odpowiadające następującym dwóm zmiennym z rozkładu jednostajnego $u_1 = 0.35$ i $u_2 = 0.65$. Przedstaw niezbędne obliczenia!!!

Zadanie 7. Dla dyskretnej zmiennej losowej, mającej rozkład zadany tabelą

X	1	2	3
P(x)	0.2	0.5	0.3

wyznaczyć średnią i wariancję. Przedstaw niezbędne obliczenia!