Kolokwium SAD 2022
grupa 1

Dorota Celińska-Kopczyńska, Magda Markowska, Piotr Pokarowski,
Łukasz Rajkowski, Jacek Sroka, Ewa Szczurek

Kwiecień 2022

Zadanie 1 [Autor: MM gr 1] (2 pkt) Rozkład Pareto, (od nazwiska włoskiego ekonomisty Vilfreda Pareto) został po raz pierwszy użyty do opisania rozkładu bogactwa w społeczeństwie, którego 80% miało się znajdować w posiadaniu 20% obywateli. Niech $(1.1, 1.2, 1.3, 1.7, 1.0, 1.4, 1.2, 1.1, 3.0, 1.7)$ będzie próbą prostą z rozkładu Pareto o parametrach $a > 0$ (dodatnie minimum) oraz $\theta > 0$ (tzw. Pareto index) i gęstości:

$$f_{a,\theta}(x) = \begin{cases} \frac{\theta a^\theta}{x^{\theta+1}} & \text{jeżeli } x > a \\ 0 & \text{w p.p.} \end{cases} \tag{1}$$

Wyznacz estymator największej wiarogodności parametru θ i oblicz jego wartość dla zadanej próby (z dokładnością do jednego miejsca po przecinku), wiedząc, że estymator największej wiarogodności parametru a ma postać $\hat{a} = \min_i x_i$. Używając otrzymanych wartości estymatorów parametrów a i θ, oszacuj prawdopodobieństwo $P(X > \max_i x_i)$. Wskazówka: $\sum_{i=1}^{10} \ln(x_i) = 3.31$

Wartości estymatora parametru θ i szukanego prawdopodobieństwa wynoszą odpowiednio:

- 3.0 i $\frac{1}{9}$
- 2.0 i $\frac{1}{27}$
- 2.0 i $\frac{1}{9}$

SOL 3.0 i $\frac{1}{27}$

Zadanie 2 [Autor: ŁR, gr 1] (2 pkt) Rozważmy ocenę modelu klasyfikacji w przypadku, gdy istnieją tylko dwie klasy. Przyjmijmy standardowe oznaczenia: TPR (czułość), TNR (swoistość), PPV (precyzja), FDR (*false discovery rate*) i ACC (dokładność). Wskaż nierówność, która jest zawsze prawdziwa:

SOL $(ACC - TPR) \cdot (ACC - TNR) \leqslant 0$

- $(PPV - TPR) \cdot (PPV - TNR) \leqslant 0$
- $(FDR - TPR) \cdot (FDR - TNR) \leqslant 0$
- $(FDR + PPV) \cdot (TPR - TNR) \leqslant 0$

Zadanie 3 [Autor: ŁR, gr 1,] (2 pkt) Na podstawie wagi 100 noworodków płci żeńskiej oraz 100 noworodków płci męskiej obliczono przedziały ufności na poziomie 95% dla wag (wyrażonych w kilogramach) noworodków: $[3.28, 3.52]$ dla chłopców oraz $[3.1, 3.3]$ dla dziewczynek. Obliczenia zostały wykonane przy założeniu, że zaobserwowane wagi chłopców są próbą prostą z rozkładu normalnego o nieznanych parametrach wartości oczekiwanej oraz wariancji i takie samo założenie poczyniono w stosunku do wag dziewczynek; zastosowano standardowy wzór na symetryczny przedział ufności. Wskaż zdanie prawdziwe:

- Nieobciążony estymator wariancji wagi chłopców jest mniejszy od nieobciążonego estymatora wariancji wagi dziewczynek.
- Prawdopodobieństwo zdarzenia, że losowo wybrany z populacji noworodek płci męskiej jest cięższy od noworodka płci żeńskiej przekracza 50%.

SOL Nieobciążony estymator wariancji wagi chłopców jest większy od estymatora największej wiarogodności wariancji wagi dziewczynek.

- Żadne z powyższych zdań nie jest prawdziwe.

Zadanie 4 [Autor: JS, gr 1] (2 pkt) W pewnej klinice badano, jak liczba godzin spędzonych na słońcu wpływa na czas choroby w przypadku zakażenia wirusem chi. Na podstawie obserwacji ustalono, że w przypadku przebywania codziennie po 1 godzinie na słońcu choroba trwała 2 tygodnie. W przypadku przebywania codziennie po 3 godziny na słońcu choroba trwała 1.8 tygodnia. W przypadku przebywania na słońcu codziennie po 5 godzin czas choroby skracał się do 1 tygodnia. Wyznacz wartość estymatorów parametrów α i β krzywej regresji liniowej, gdzie ϵ jest niezależną zmienną błędu:

$$tygodnie\_choroby = \alpha + \beta * godziny\_na\_słońcu + \epsilon$$

SOL $\hat{\beta} = -0,25, \hat{\alpha} = 2,35,$

- $\hat{\beta} = 0,25, \hat{\alpha} = 0,85,$
- $\hat{\beta} = 1,2, \hat{\alpha} = -1,5,$
- $\hat{\beta} = -1,2, \hat{\alpha} = 3,5.$

Zadanie 5 [Autor: JS, gr 1] (2 pkt) Rozkład Poissona z nieznanym parametrem λ można z powodzeniem zastosować do modelowania odwiedzin klientów w butiku, gdzie λ odpowiada średniej liczbie klientów odwiedzających butik w ciągu godziny. Jednego dnia kierownik sklepu zauważył, że pomiędzy godziną 12 a 13 przybyło dwóch klientów, a jego pomocnik, że tego samego dnia między 13:15, a 14:15 przybył tylko jeden klient. Użyj tych obserwacji, żeby oszacować wartość estymatora największej wiarogodności parametru λ.

- $\hat{\lambda} = e^3/2$,

SOL $\hat{\lambda} = 3/2$,

- $\hat{\lambda} = e^{3/2}$,
- $\hat{\lambda} = 3 * e/2$.

Zadanie 6 [Autor: PP, gr 1] (2 pkt) Na podstawie $n = 25$ obserwacji z rozkładu normalnego o nieznanych parametrach, testowano hipotezę $\sigma^2 = 10$ przeciw $\sigma^2 > 10$ za pomocą statystyki testowej $T = \frac{n\hat{\sigma}^2}{10}$, gdzie $\hat{\sigma}^2 = 15$ jest estymatorem największej wiarogodności σ^2. Następnie obliczono (i) p-wartość T oraz (ii) moc testu postaci $\{T > c\}$ na poziomie istotności $\alpha = 0.05$ dla alternatywy $\sigma^2 = 20$. Wskaż prawidłowy wynik (p-wartość, moc):

- $(0.961, 0.793)$
- $(0.052, 0.805)$

SOL $(0.039, 0.793)$

- $(0.948, 0.805)$

Zadanie 7 [Autor: DCK, gr 1] (2 pkt) Dla pewnego modelu regresji liniowej, tłumaczącego zmienną y z wykorzystaniem zmiennych x_1, x_2, x_3 oraz stałej otrzymano oszacowanie postaci:

$$y = 34 + 0.25x_1 + 0.73x_2 - 0.45x_3 + e$$

gdzie e to składnik resztowy. R^2 dla tego modelu wyniosło 0.2. Wszystkie oszacowania parametrów okazały się indywidualnie istotnie różne od zera. Wskaż zdanie prawdziwe:

(Podpowiedź: ceteris paribus – przy pozostałych wartościach niezmienionych)

- Wraz ze wzrostem zmiennej x_2 o jedną jednostkę wartość zmiennej y zmniejsza się o 0.73 jednostki, ceteris paribus.
- Wzrost wartości zmiennej x_1 o jedną jednostkę przekłada się na spadek wartości zmiennej y o 0.25 jednostki, ceteris paribus.

SOL W modelu proporcja zmienności zmiennej objaśnianej, którą można wyjaśnić, używając zmienności zmiennych objaśniających to 20%.

- Na podstawie R^2 wnioskujemy, że w 20% przypadków model prawidłowo przewiduje wartość zmiennej y.

Zadanie 8 [Autor: DCK, gr 1] (2 pkt) Dana jest próba losowa z rozkładu ciągłego oraz proste hipotezy parametryczne (zerowa i alternatywna). Które dwa spośród następujących zdań są równoważnymi definicjami poziomu istotności testu w tym przypadku?

a) Prawdopodobieństwo popełnienia błędu pierwszego rodzaju.
b) Prawdopodobieństwo przyjęcia H0 gdy jest ona fałszywa.
c) Prawdopodobieństwo odrzucenia H0 gdy jest ona prawdziwa.
d) Całka po regionie krytycznym z gęstości rozkładu statystyki testowej przy prawdziwości hipotezy H_1.

- a) oraz b)
- c) oraz d)

SOL a) oraz c)

- b) oraz d)

Zadanie 9 [Autor: DCK, gr 1] (2 pkt) Ogrodnik-statystyk amator zauważył, że wśród jego klientów występuje wyższy popyt na irysy z dłuższymi płatkami. Natknął się na reklamę odżywki do kwiatów mającej zapewniać taki efekt. W szklarni założył dwie hodowle iris versicolor – jedna grupa wysianych donic była nawożona odżywką, druga grupa wysianych donic nie otrzymywała odżywki – była grupą kontrolną. Poza kwestią podawania odżywki obie grupy miały zapewnione te same warunki. Gdy kwiaty wyrosły, ogrodnik pomierzył w nich długość płatków. Wiadomo, że niezależnie od ożywiania, długości płatków pochodzą z rozkładu normalnego o pewnej nieznanej wariancji. Którego z wymienionych testów statystycznych najlepiej użyć do sprawdzenia, czy średnia długość płatka wśród kwiatów, którym podawano odżywkę a kwiatów z grupy kontrolnej jest taka sama?

SOL Testu t-Studenta dla prób niesparowanych (niezależnych)

- Testu Kołmogorowa-Smirnova
- Testu t-Studenta dla prób sparowanych (zależnych)
- Testu niezależności χ^2

Zadanie 10 [Autor: ES, punkty: 1, gr 1] (2 pkt) Rozważmy problem regresji dla zmiennej objaśnianej Y oraz wektora zmiennych objaśniających $X = (X_1, \ldots, X_p)^T$ postaci

$$Y = f(X) + \epsilon$$

gdzie ϵ jest błędem losowym, niezależnym od X, o wartości oczekiwanej $\mathbb{E}[\epsilon] = 0$, a f nieznaną funkcją, którą chcemy estymować modelem $\hat{f}$. Niech $\mathcal{M}$ będzie klasą nieobciążonych estymatorów f. Wybierz **nieprawdziwą** odpowiedź:

- Dla $\hat{f}$ będącego modelem regresji liniowej,

  $$\hat{f} = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

  elastyczność $\hat{f}$ rośnie z p.

SOL $\hat{f}$ nieobciążony ($\hat{f} \in \mathcal{M}$), o najmniejszej wariancji spośród modeli w $\mathcal{M}$ zawsze da minimalny błąd testowy wynoszący $Var[\epsilon]$.

- Model

  $$\hat{f} = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \beta_{1,2} X_1 X_2 + \ldots + \beta_{p-1,p} X_{p-1} X_p$$

  jest modelem liniowym

- Błąd testowy dla modelu $\hat{f}$ zależy od obciążenia, wariancji i błędu nieredukowalnego tego modelu.