# Kolokwium SAD 2022
# grupa 2

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

Kwiecień 2022

**Zadanie 1 [Autor: MM, gr 2]** (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 $(2.6, 1.7, 1.3, 1.4, 1.0, 1.3, 1.2, 1.1, 2.1, 5.0)$ 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 $\theta$ 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 $\theta$, oszacuj prawdopodobieństwo $P(X > \max_i x_i)$. Wskazówka: $\sum_{i=1}^{10} \ln(x_i) = 4.98$

Wartości estymatora parametru $\theta$ i szukanego prawdopodobieństwa wynoszą odpowiednio:

- 2.0 i $\frac{1}{125}$

**SOL** 2.0 i $\frac{1}{25}$

- 3.0 i $\frac{1}{125}$
- 3.0 i $\frac{1}{25}$

**Zadanie 2 [Autor: ŁR, punkty: 1, gr 2]** (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:

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

**SOL** $(TPR - ACC) \cdot (ACC - TNR) \geqslant 0$

**Zadanie 3 [Autor: ŁR, punkty: 2, gr 2]** (2 pkt) Na podstawie wagi $10^4$ noworodków płci żeńskiej oraz $10^4$ noworodków płci męskiej obliczono przedziały ufności na poziomie 95% dla wag (wyrażonych w kilogramach) noworodków: $[3.3, 3.32]$ dla chłopców oraz $[3.19, 3.21]$ 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:

**SOL** Estymator największej wiarogodności 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 jest równe 95%.
- Prawdopodobieństwo zdarzenia, że losowo wybrany z populacji noworodek płci męskiej jest cięższy od noworodka płci żeńskiej jest równe 90.25%.
- Żadne z powyższych zdań nie jest prawdziwe.

**Zadanie 4 [Autor: JS, punkty: 1, gr 2]** (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 1 tydzień. 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 wydłużał się do 2 tygodni. Wyznacz wartość estymatorów parametrów $\alpha$ i $\beta$ krzywej regresji liniowej, gdzie $\epsilon$ jest niezależną zmienną błędu:

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

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

**SOL** $\hat{\beta} = 0,25, \hat{\alpha} = 0,85,$

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

**Zadanie 5 [Autor: JS, punkty: 2, gr 2]** (2 pkt) Rozkład Poissona z nieznanym parametrem $\lambda$ można z powodzeniem zastosować do modelowania odwiedzin klientów w butiku, gdzie $\lambda$ odpowiada średniej liczbie klientów odwiedzających butik w ciągu godziny. Jednego dnia kierownik sklepu zauważył, że pomiędzy godziną 12 a 13 przybyło trzech 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 $\lambda$.

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

**SOL** $\hat{\lambda} = 2$,

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

**Zadanie 6 [Autor: PP, punkty: 2, gr 2 ]** (2 pkt) Na podstawie $n = 20$ obserwacji z rozkładu normalnego o nieznanych parametrach, testowano hipotezę $\sigma^2 = 8$ przeciw $\sigma^2 > 8$ za pomocą statystyki testowej $T = \frac{n\hat{\sigma}^2}{8}$, gdzie $\hat{\sigma}^2 = 12$ jest estymatorem największej wiarogodności $\sigma^2$. Następnie obliczono (i) p-wartość $T$ oraz (ii) moc testu postaci $\{T > c\}$ na poziomie istotności $\alpha = 0.1$ dla alternatywy $\sigma^2 = 20$. Wskaż prawidłowy wynik (p-wartość, moc):

**SOL** $(0.052, 0.928)$

- $(0.948, 0.928)$
- $(0.070, 0.936)$
- $(0.930, 0.936)$

**Zadanie 7 [Autor: DCK, punkty: 1, gr 2 ]** (2 pkt) Mamy oszacowany model regresji liniowej 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)

- W modelu proporcja zmienności zmiennych objaśniających, którą można wyjaśnić, używając zmienności zmiennej objaśnianej to 20%.
- Spadek wartości zmiennej $x_1$ o jedną jednostkę przekłada się na wzrost wartości zmiennej $y$ o 0.25 jednostki, ceteris paribus.

**SOL** Wraz ze wzrostem zmiennej $x_2$ o jedną jednostkę wartość zmiennej $y$ zmniejsza się o 0.73 jednostki, ceteris paribus.

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

**Zadanie 8 [Autor: DCK, punkty: 1, gr 2 ]** (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 niepopełnienia błędu pierwszego rodzaju.
b) Prawdopodobieństwo odrzucenia H0 gdy jest ona prawdziwa.
c) Prawdopodobieństwo przyjęcia H0 gdy jest ona fałszywa.
d) Całka po regionie krytycznym z gęstości rozkładu statystyki testowej przy prawdziwości hipotezy $H_0$.

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

**SOL** b) oraz d)

**Zadanie 9 [Autor: DCK, punkty: 1, gr 2 ]** (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. Którego z wymienionych testów statystycznych najlepiej użyć do sprawdzenia, czy rozkłady długości płatków w kwiatach, które otrzymywały odżywkę i kwiatach z grupy kontrolnej są takie same?

- Testu t-Studenta dla prób niesparowanych (niezależnych)

**SOL** Testu Kołmogorowa-Smirnova

- Testu Wilcoxona
- Testu $\rho$-Spearmana

**Zadanie 10 [Autor: ES, punkty: 1, gr 2, SPRAWDZONE – OK]** (2 pkt) Niech model $\hat{f}$ będzie klasyfikatorem dla kategorycznej zmiennej objaśnianej $Y$ i przy użyciu wektora zmiennych objaśniających $X = (X_1, \ldots, X_p)^T$. Niech $x_0$ będzie obserwacją testową (niewidzianą przez model w trakcie jego uczenia). Która z odpowiedzi jest **nieprawdziwą**:

**SOL** Model $\hat{f}$ o zerowym obciążeniu nie może być przeuczony.

- Błąd nieredukowalny nie zależy od elastyczności modelu.
- Dla $\hat{f}$ będącego klasyfikatorem Bayesowskim, suma wariancji modelu $\hat{f}$ oraz kwadratu jego obciążenia jest minimalna.
- Elastyczność modelu KNN wzrasta wraz z $\frac{1}{K}$, gdzie $K$ to parametr odpowiadający liczbie sąsiadów dla $x_0$ branych pod uwagę przy klasyfikacji.