Kolokwium SAD 2023

Dorota Celińska-Kopczyńska, Krzysztof Gogolewski, Błażej Miasojedow, Szymon Nowakowski,
Kazimierz Oksza-Orzechowski, Piotr Pokarowski, Ewa Szczurek

Kwiecień 2023

Zadanie 1 [Autor: ES, gr 1] (2 pkt) Rozważmy deterministyczny zbiór danych $\mathbf{X}$ o $n+m$ wierszach odpowiadających obserwacjom, pierwszej kolumnie wypełnionej jedynkami i p pozostałych kolumnach odpowiadających predyktorom. Dla tego zbioru danych wygenerowano wektor $\mathbf{y}$ zmiennej objaśnianej z modelu liniowego o wektorze niezerowych współczynników β

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{X}_1 + \ldots + \beta_p \mathbf{X}_p + \varepsilon,$$

gdzie $\mathbb{E}[\varepsilon] = \mathbf{0}$ i $Var[\varepsilon] = \sigma^2 I_{n+m}$. Następnie do macierzy $\mathbf{X}$ dołożono p' deterministycznych wektorów, ortogonalnych wzajemnie i do kolumn macierzy $\mathbf{X}$, jako p' dodatkowych kolumn, otrzymując $(n+m) \times (p+p'+1)$ macierz $\mathbf{X}'$ taką, że $rank(\mathbf{X}') = p+p'+1$. Obie macierze podzielono na dane treningowe i testowe, biorąc pierwsze n wierszy do danych treningowych $\mathbf{X}_{Train}$ i $\mathbf{X}'_{Train}$ odpowiednio z macierzy $\mathbf{X}$ i $\mathbf{X}'$, a ostatnie m wierszy do danych testowych $\mathbf{X}_{Test}$ i $\mathbf{X}'_{Test}$. Podobnie pierwsze n elementów wektora $\mathbf{y}$ przydzielono do danych treningowych, a ostatnie m do danych testowych. Na danych $\mathbf{X}_{Train}$, $\mathbf{y}_{Train}$ metodą najmniejszych kwadratów wyestymowano współczynniki regresji liniowej $\hat{\beta}$. Tak otrzymany model oznaczono $\mathcal{M}$. Na danych $\mathbf{X}'_{Train}$, $\mathbf{y}_{Train}$ metodą najmniejszych kwadratów otrzymano $\hat{\beta}'$ i otrzymany model oznaczono $\mathcal{M}'$. Wiadomo, że

- $||\mathbf{y}_{Train} - \mathbf{X}_{Train}\hat{\beta}||^2 \geqslant ||\mathbf{y}_{Train} - \mathbf{X}'_{Train}\hat{\beta}'||^2$.
- $VIF(\hat{\beta}'_{p+1}) > \frac{1}{1-R^2_{\mathcal{M}}}$, gdzie $R^2_{\mathcal{M}}$ to R^2 dla modelu $\mathcal{M}$.
- Nieobciążony estymator wariancji σ^2 musiał być wiekszy dla $\mathcal{M}'$ i danych $\mathbf{X}'$ niż dla $\mathcal{M}$ i danych $\mathbf{X}$.
- $||\mathbf{y}_{Test} - \mathbf{X}_{Test}\hat{\beta}||^2 \geqslant ||\mathbf{y}_{Test} - \mathbf{X}'_{Test}\hat{\beta}'||^2$.

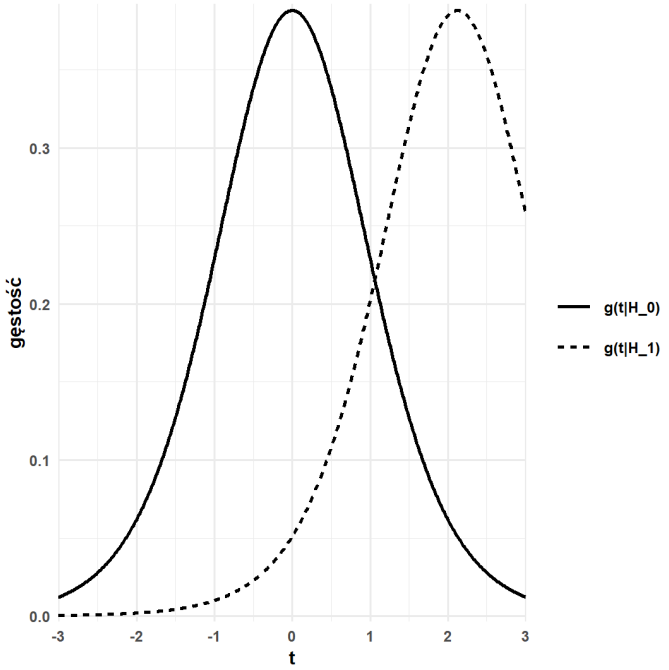
Rozwiązanie: T, F, F, F

Zadanie 1 [Autor: ES, gr 2] (2 pkt) treść j.w.

- $\mathbb{E}[\hat{\beta}'] = [\beta, 0, \ldots, 0]^T$, wektor wymiaru $p+p'+1$, z pierwszymi $p+1$ elementami równymi elementom wektora β, a pozostalymi p' elementami równymi 0.
- p-wartosci dla wszystkich $p+1$ wspolczynnikow z modelu $\mathcal{M}$ dla modelu $\mathcal{M}'$ byly takie same.
- W teście F dla wszystkich predyktorów (ang. overall F test) statystyka F dla modelu $\mathcal{M}'$ była bliska 1.
- Dla $i = 0, \ldots, p$, $Var[\hat{\beta}_i] = Var[\hat{\beta}'_i]$.

Rozwiązanie: T, F, F, T

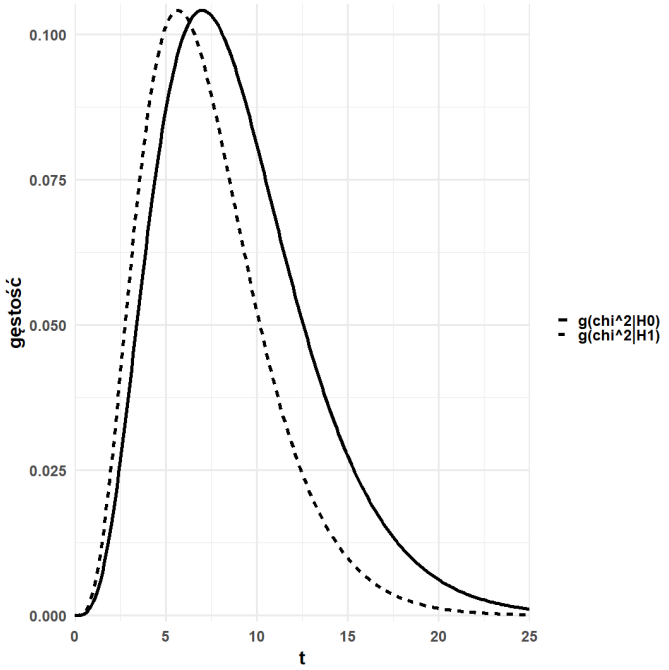
Zadanie 2 [Autor: KO, gr 1] (2 pkt) Mamy w klasie 10 uczniów, których wzrost w cm zmierzyliśmy. Niech X_i wzrost i-tego ucznia. Z naszych pomiarów dostaliśmy $\bar{X} = \frac{1}{10}\sum_i^{10} X_i = 175$, $\sum_{i=1}^{10}(X_i - \bar{X})^2 = 500$. Zakładamy, że wzrost uczniów jest niezależny i pochodzi z rozkładu normalnego $N(\mu, \sigma^2)$ o nieznanych parametrach μ, σ^2. Średni wzrost w Polsce to 170 cm, więc stawiamy hipotezy: $H_0 : \mu = 170$, $H_1 : \mu = 175$. Hipotezy będziemy testować na podstawie t: statystyki t-Studenta. Dany jest też wykres $g(t|H_0)$ i $g(t|H_1)$, czyli gęstości prawdopodobieństwa t pod warunkiem odpowiednio hipotezy zerowej i alternatywnej:



- Wartość nieobciążonego estymatora σ^2 otrzymana z naszych pomiarów to 50.
- Z danego wykresu wynika, że przy dowolnym $c > 0$ obszar krytyczny postaci $t > c$ da nam test o większej mocy niż gdyby był postaci $t < -c$.
- Początkujący statystyk wybrał obszar krytyczny postaci $|t| > c$. Czy to prawda, że przy poziomie istotności $\alpha = 0.1$ odrzuci on H_0 na podstawie danych z zadania? Przypomnienie: kwantyl rzędu 0.05 rozkładu t-Studenta o 9 stopniach swobody to $q(0.05, 9) \approx -1.83$.
- Czy optymalna (dająca węższe przedziały ufności) statystyka testowa zmieni się, jeśli poznamy prawdziwą wartość σ^2?

Rozwiązanie: F, T, T, T

Zadanie 2 [Autor: KO, gr 2] (2 pkt) Przez ostatnie 10 miesięcy notowaliśmy miesięczne zyski (w tys. \$) pewnej firmy. Niech X_i będzie zyskiem z i-tego miesiąca. Z naszych pomiarów dostaliśmy: $\bar{X} = \frac{1}{10}\sum_{i=1}^{10} X_i = 100$, $\sum_{i=1}^{10}(X_i - \bar{X})^2 = 16000$. Zakładamy, że zyski firmy są od siebie niezależne i pochodzą z rozkładu normalnego $N(\mu, \sigma^2)$ o nieznanych parametrach μ, σ^2. Według gazet typowa wariancja zysków to $\sigma^2 = 2500$, natomiast my będziemy zainteresowani inwestycją w tę firmę, jeśli jej zyski są bardziej stabilne $\sigma^2 = 2025$. Dlatego stawiamy hipotezy $H_0 : \sigma^2 = 2500$, $H_1 : \sigma^2 = 2025$, które będziemy testować za pomocą χ^2, statystyki chi-kwadrat. Do tego mamy dany wykres $g(\chi^2|H_0)$ i $g(\chi^2|H_1)$, czyli gęstości prawdopodobieństwa χ^2 pod warunkiem odpowiednio hipotezy zerowej i alternatywnej:



- Wartość nieobciążonego estymatora σ^2 otrzymana z naszych pomiarów to $\frac{16000}{9}$.
- Przy poziomie istotności $\alpha = 0.05$ i obszarze krytycznym postaci $\chi^2 < c, (c > 0)$ powinniśmy na podstawie naszych danych zainwestować w tę spółkę (tzn. odrzucamy H_0). Przypomnienie: kwantyl rzędu 0.05 rozkładu χ^2 o 9 stopniach swobody to $\chi^2(0.05, 9) \approx 3.33$.
- Odpowiedź w poprzednim podpunkcie zależy od parametrów naszej hipotezy alternatywnej H_1.
- Z danego wykresu wynika, że dla obszaru krytycznego postaci $\chi^2 < 5$ otrzymamy test o mocy większej niż poziom istotności.

Rozwiązanie: T, F, F, T

Zadanie 3 [Autor: KG, gr 1] (4 pkt) Pewna firma wprowadza do użycia nową taśmę produkcyjną M_2 (do tej pory korzystała tylko z taśmy M_1). Przeprowadzono kontrolę jakości wytwarzanych produktów, powstających na taśmach M_1 oraz M_2. Każdy z ocenionych produktów otrzymał kategorię jakości od najwyższej klasy (A), do najniższej (C). Wyniki zebrano w poniższej tabeli 1

	A	B	C
M_1	100	100	100
M_2	120	100	80

Tabela 1: Tabela zliczeń do zadania 3

Niech X będzie zmienną losową opisującą zaobserwowaną kateogorię jakości, zaś Y taśmę produkcyjną, z której pochodzi produkt. Ponadto, niech $F_{X=M}(Y)$ oznacza dystrybuantę empiryczną zmiennej Y dla taśmy M. Przy użyciu testów χ^2 postawiono zbadać dwie hipotezy: H_0^A: $F_{X=M_2}(Y)$ jest zgodna z $F_{X=M_1}(Y)$ oraz H_0^B: X oraz Y są niezależne.

	0.1	0.05	0.025	0.02	0.01	0.005
1	2.7055	3.8415	5.0239	5.4119	6.6349	7.8794
2	4.6052	5.9915	7.3778	7.8241	9.2103	10.5966
3	6.2514	7.8147	9.3484	9.8374	11.3449	12.8382

Tabela 2: Tablica wartości krytycznych rozkładu χ^2 dla zadanej liczby stopni swobody n i poziomu istotności

- na poziomie istotności $\alpha = 0.025$ istnieje podstawa do odrzucenia H_0^A

- na poziomie istotności $\alpha = 0.025$ istnieje podstawa do odrzucenia H_0^B
- jeśli przy ustalonym poziomie istotności $\alpha > 0$ nie mamy podstaw do odrzucenia hipotez H_0^A oraz H_0^B, należy przyjąć ich prawdziwość
- przeprowadzono test zgodności χ^2 przy założeniu pewnej H_0 i przyjęciu poziomu istotności $\alpha = 0.05$. Statystyka testowa ma wartość $\chi^2 = 4$ przy jednym stopniu swobody. Istnieje podstawa do odrzucenia H_0.

Rozwiązanie: T, F, F, T

Zadanie 3 [Autor: KG, gr 2] (2 pkt) Pewna firma wprowadza do użycia nową taśmę produkcyjną M_2 (do tej pory korzystała tylko z taśmy M_1). Przeprowadzono kontrolę jakości wytwarzanych produktów, powstających na taśmach M_1 oraz M_2. Każdy z ocenionych produktów otrzymał kategorię jakości od najwyższej klasy (A), do najniższej (C). Wyniki zebrano w poniższej tabeli 3.

	A	B	C
M_1	100	100	100
M_2	120	100	80

Tabela 3: Tabela zliczeń do zadania 3

Niech X będzie zmienną losową opisującą zaobserwowaną kateogorię jakości, zaś Y taśmę produkcyjną, z której pochodzi produkt. Ponadto, niech $F_{X=M}(Y)$ oznacza dystrybuantę empiryczną zmiennej Y dla taśmy M. Przy użyciu testów χ^2 postawiono zbadać dwie hipotezy: H_0^A: $F_{X=M_2}(Y)$ jest zgodna z $F_{X=M_1}(Y)$ oraz H_0^B: X oraz Y są niezależne.

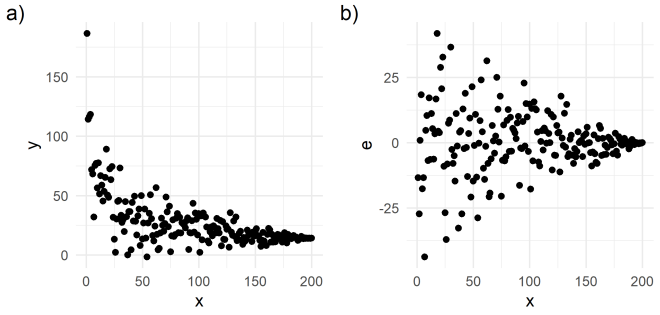
	0.1	0.05	0.025	0.02	0.01	0.005
1	2.7055	3.8415	5.0239	5.4119	6.6349	7.8794
2	4.6052	5.9915	7.3778	7.8241	9.2103	10.5966
3	6.2514	7.8147	9.3484	9.8374	11.3449	12.8382

Tabela 4: Tablica wartości krytycznych rozkładu χ^2 dla zadanej liczby stopni swobody n i poziomu istotności

- na poziomie istotności $\alpha = 0.05$ istnieje podstawa do odrzucenia H_0^B
- na poziomie istotności $\alpha = 0.05$ istnieje podstawa do odrzucenia H_0^A
- jeśli przy ustalonym poziomie istotności $\alpha > 0$ nie mamy podstaw do odrzucenia hipotez H_0^A oraz H_0^B, należy przyjąć ich prawdziwość
- przeprowadzono test zgodności χ^2 przy założeniu pewnej H_0 i przyjęciu poziomu istotności $\alpha = 0.05$. Statystyka testowa ma wartość $\chi^2 = 8$ przy trzech stopniach swobody. Istnieje podstawa do odrzucenia H_0.

Rozwiązanie: F, T, F, T

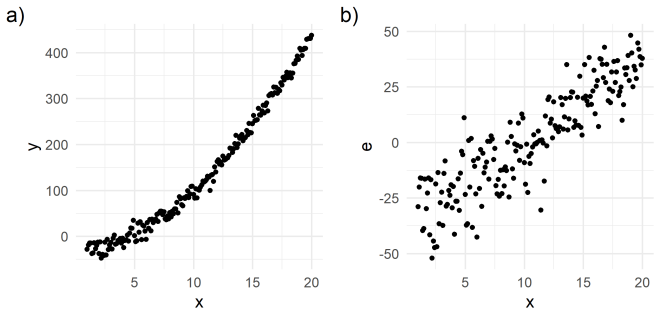
Zadanie 4 [Autor: KO, gr 1] (2 pkt) W trakcie rocznego badania w każdej z 200 firm zanotowano wzrost jej zadłużenia Y oraz jej przychodu netto X. Wyniki badania przedstawiono na wykresie **a)**, gdzie punkt (x_i, y_i) odpowiada obserwacjom dla i-tej firmy. Zakładamy, że x_i mogą być traktowane jako obserwacje deterministyczne. Odnotowaliśmy też, że żadna firma nie zarejestrowała takiego samego przychodu netto ($i \neq j \rightarrow x_i \neq x_j$).



- W modelu $\hat{Y} = \beta_1 \cdot Z + \beta_0$, gdzie $Z = \frac{1}{\sqrt{X}}$, y zależy liniowo od parametrów β oraz transformacji zmiennej X, czyli jest to model liniowy.
- Różnice $e = Y - \hat{Y}$ zostały przedstawione na wykresie b). Są one homoskedastyczne.
- Gdyby w modelu uwzględniono dodatkowy predyktor $W = X^2$, to resztowa suma kwadratów (RSS) w tym modelu byłaby większa w porównaniu do RSS w modelu bez tego predyktora.
- Całkowita suma kwadratów w tym modelu wyniosła $\sum_{i=1}^{n}(y_i - \bar{y})^2 = 107013.6$, a resztowa suma kwadratów w modelu $\hat{Y} = \beta_1 \cdot Z + \beta_0$ wyniosła $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 27986.26$. Wynika stąd, że model ten tłumaczy ponad 70% zmienności Y.

Rozwiązanie: T, F, F, T

Zadanie 4 [Autor: KO, gr 2] (2 pkt) Testujemy nową konstrukcję mostu. Dla stali o różnej grubości X użytej do jego budowy porównujemy, o ile większe obciążenie może on utrzymać w porównaniu do starszej konstrukcji, i notujemy tę różnicę obciążeń jako Y. Wyniki naszych testów przedstawia wykres **a)**, gdzie punkt (x_i, y_i) odpowiada różnicy obciążeń zanotowanej dla konstrukcji ze stalą o grubości x_i. Nie testowano dwa razy tej samej grubości stali $(i \neq j \rightarrow x_i \neq x_j)$.



- Model $\hat{Y} = \beta_1 \cdot Z + \beta_0$, gdzie $Z = X^2$, nie jest modelem liniowym, bo predyktor Z zależy nieliniowo od zmiennej X.
- Różnice $e = Y - \hat{Y}$ zostały przedstawione na wykresie b). Są one nieskorelowane z wielkością predyktora X.
- Gdyby w modelu uwzględniono dodatkowo predyktor $W = X$, to resztowa suma kwadratów (RSS) w tym modelu nie wzrosłaby w porównaniu do RSS w modelu bez tego predyktora.
- Całkowita suma kwadratów w tym modelu wyniosła $\sum_{i=1}^{n}(y_i - \bar{y})^2 = 3728485$, a resztowa suma kwadratów w modelu $\hat{Y} = \beta_1 \cdot Z + \beta_0$ wyniosła $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 102915.9$. Wynika stąd, że model ten tłumaczy ponad 95% zmienności Y.

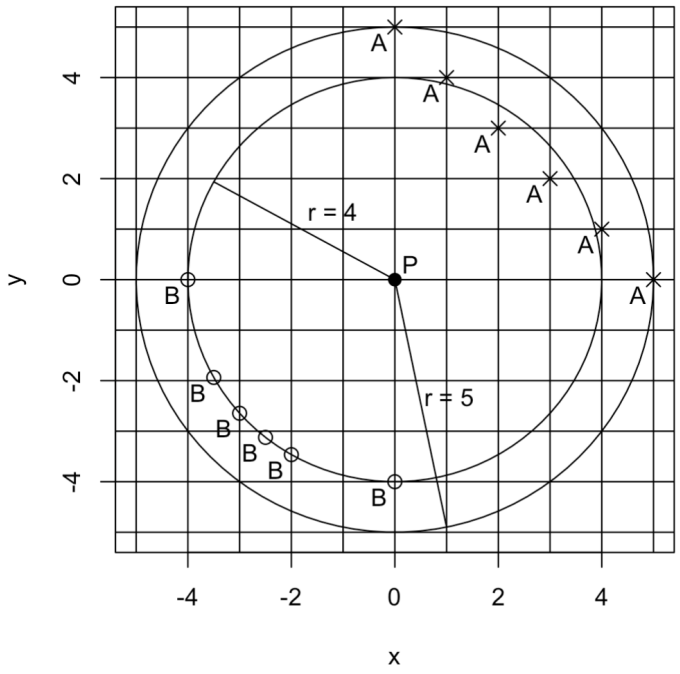
Rozwiązanie: F, F, T, T

Zadanie 5 [Autor: KG, gr 1] (2 pkt) Dany jest zbiór punktów $(x, y) \in \mathbb{R}^2$. Każdy z punktów jest przyporządkowany do jednej z klas: A (punkty oznaczone jako $\times$, mające współrzędne postaci $(i, 5-i)$ dla $i = 0, ..., 5$ lub B (punkty oznaczone jako $\circ$, leżące na okręgu o promieniu 4).

Dla dowolnych dwóch punktów $p = (x_1, y_1)$ oraz $q = (x_2, y_2)$ rozważamy dwie metryki:

- miejską: $d_M(p, q) = |x_1 - x_2| + |y_1 - y_2|$
- euklidesową: $d_E(p, q) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$

Chcemy ustalić przyporządkowanie klasy dla nowego punktu $P = (0, 0)$ za pomocą algorytmu kNN dla różnych wartości k i przyjętej metryki.



- dla $k = 5$ i metryki d_M punkt P otrzyma klasę B
- dla $k = 3$ i metryki d_M punkt P otrzyma klasę B
- dla $k = 5$ i metryki d_E punkt P otrzyma klasę B
- dla $k = 3$ i metryki d_E punkt P otrzyma klasę B

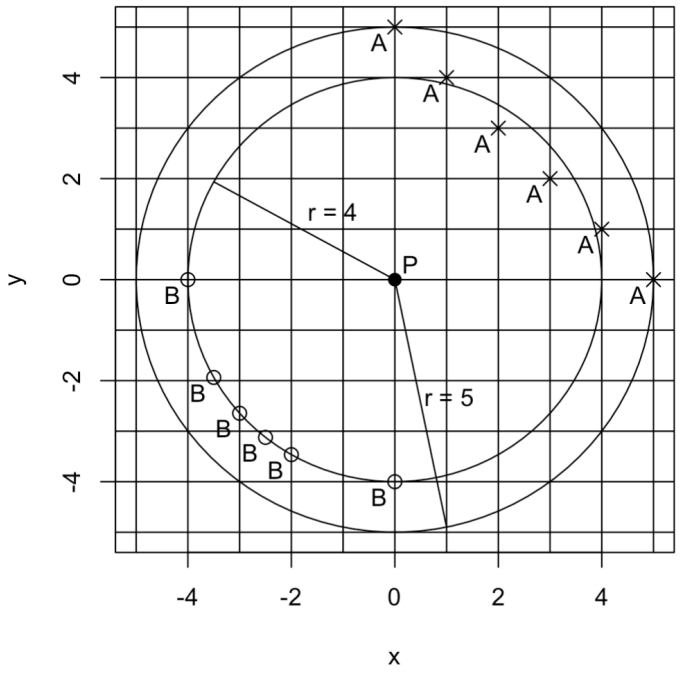
Rozwiązanie: F, T, T, F

Zadanie 5 [Autor: KG, gr 2] (2 pkt) Dany jest zbiór punktów $(x, y) \in \mathbb{R}^2$. Każdy z punktów jest przyporządkowany do jednej z klas: A (punkty oznaczone jako $\times$, mające współrzędne postaci $(i, 5-i)$ dla $i = 0, ..., 5$ lub B (punkty oznaczone jako $\circ$, leżące na okręgu o promieniu 4).

Dla dowolnych dwóch punktów $p = (x_1, y_1)$ oraz $q = (x_2, y_2)$ rozważamy dwie metryki:

- miejską: $d_M(p, q) = |x_1 - x_2| + |y_1 - y_2|$
- euklidesową: $d_E(p, q) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$

Chcemy ustalić przyporządkowanie klasy dla nowego punktu $P = (0, 0)$ za pomocą algorytmu kNN dla różnych wartości k i przyjętej metryki.



- dla $k = 5$ i metryki d_M punkt P otrzyma klasę A
- dla $k = 3$ i metryki d_M punkt P otrzyma klasę A
- dla $k = 5$ i metryki d_E punkt P otrzyma klasę A
- dla $k = 3$ i metryki d_E punkt P otrzyma klasę A

Rozwiązanie: T, F, F, T

Zadanie 6 [Autor: DCK, gr 1] (2 pkt): Wskaż, czy podane zdania są prawdziwe:

- Poziom istotności testu α jest równy prawdopodobieństwu popełnienia błędu pierwszego rodzaju w tym teście.
- Im wyższy poziom ufności $1 - \alpha$ przyjmujemy, tym szerszy przedział ufności otrzymamy.
- Poprawka Bonferroniego polega na mnożeniu pierwotnie założonego poziomu istotności pojedynczego testu przez liczbę testów, które wykonujemy.
- Moc testu jest równa prawdopodobieństwu popełnienia błędu drugiego rodzaju β w tym teście.

Rozwiązanie: T, T, F, F

Zadanie 6 [Autor: DCK, gr 2] (2 pkt): Wskaż, czy podane zdania są prawdziwe:

- Poziom istotności testu α jest równy prawdopodobieństwu niepopełnienia błędu drugiego rodzaju w tym teście.
- Im wyższy poziom ufności $1 - \alpha$ przyjmujemy, tym węższy przedział ufności otrzymamy.
- Korzystając z poprawki Holma, dla czterech testów, każda z uzyskanych p-wartości porównywana byłaby z $\alpha/4$, gdzie α to pierwotnie założony poziom istotności pojedynczego testu.
- Moc testu $1-\beta$ równa 0,7 oznacza, że z prawdopodobieństwem równym 0,7 odrzucimy fałszywą hipotezę zerową.

Rozwiązanie: F, F, F, T

Zadanie 7 [Autor: DCK, gr 1] (2 pkt): Niech X_1, X_2, X_3 będą próbą prostą z rozkładu normalnego $N(\mu, \sigma)$, gdzie zarówno μ, jak i σ są nieznane. Rozważamy następujące estymatory dla μ:

$$\hat{\mu}_1 = \frac{X_1 + X_2 + X_3}{3}, \quad \hat{\mu}_2 = \frac{2X_1 + 2X_2 + X_3}{5}$$

- Oba estymatory są nieobciążone.
- $Var(\hat{\mu}_1) > Var(\hat{\mu}_2)$ dla każdego μ i σ
- $MSE(\hat{\mu}_1) \leqslant MSE(\hat{\mu}_2)$ dla każdego μ i σ.
- Estymator $\hat{\mu}_1$ jest estymatorem największej wiarogodności dla tego przypadku.

Rozwiązanie: T, F, T, T

Zadanie 7 [Autor: DCK, gr 2] (2 pkt): Niech X_1, X_2, X_3 będą próbą prostą z rozkładu normalnego $N(\mu, \sigma)$, gdzie zarówno μ, jak i σ są nieznane. Rozważamy następujące estymatory dla μ:

$$\hat{\mu}_1 = \frac{X_1 + X_2 + X_3}{3}, \quad \hat{\mu}_2 = \frac{3X_1 + 2X_2 + X_3}{6}$$

- Co najmniej jeden z tych estymatorów jest obciążony.
- Wariancje podanych estymatorów nie zależą od parametru σ.
- $Var(\hat{\mu}_1) < Var(\hat{\mu}_2)$ dla każdego μ i σ.
- Estymator $\hat{\mu}_2$ nie jest estymatorem największej wiarogodności dla tego przypadku.

Rozwiązanie: F, F, T, T

Zadanie 8 [Autor: SN, gr 1] (2 pkt) Malutki Jaś od jakiegoś czasu każdego ranka w swoich **dwóch** butach odnajduje po nocy 0, 1 albo 2 cukierki, w taki sposób umieszczone, że w każdym bucie jest co najwyżej jeden cukierek. Jaś jest uzdolniony matematycznie i wobec tego jest przekonany, że to skrzaty podrzucają mu w nocy cukierki, umieszczając w danym bucie (każdym niezależnie) z prawdopodobieństwem $0 < p < 1$ jeden cukierek, zaś z prawdopodobieństwem $1-p$ nie umieszczają w nim cukierka. Od szesnastu nocy liczy, ile cukierków łącznie danej nocy dostał, i jego obserwacje są następujące: $2, 0, 0, 2, 1, 0, 0, 2, 2, 2, 2, 1, 0, 0, 1, 1$. Słowem, było 6 nocy, kiedy otrzymał dwa cukierki, 4 noce kiedy otrzymał jeden cukierek i 6 nocy, kiedy w ogóle nie otrzymał cukierków.

Jaś zechciał sprawdzić prawdziwość swoich przekonań: najpierw wyestymował wartość nieznanego prawdopodobieństwa p estymatorem największej wiarogodności $\hat{p}$, a następnie przeprowadził test zgodności χ^2-Pearsona, z hipotezą H_0, że liczba otrzymanych cukierków każdej nocy pochodzi ze schematu Bernoulliego z prawdopodobieństwem sukcesu $\hat{p}$ i o dwóch powtórzeniach.

Poniżej podano tabelę wartości dystrybuanty rozkładu χ^2 z jednym i dwoma stopniami swobody:

$x =$	1.0	2.0	3.0	4.0	5.0	6.0	7.0
$F_{\chi^2(1)}(x) =$	0.683	0.843	0.917	0.954	0.975	0.986	0.992
$F_{\chi^2(2)}(x) =$	0.393	0.632	0.777	0.865	0.918	0.950	0.970

Tabela 5: Tabela wartości dystrybuanty rozkładu χ^2

Rozstrzygnij prawdziwość odpowiedzi:

- Statystyka testowa ma rozkład χ^2 o jednym stopniu swobody.
- Należy odrzucić H_0 na poziomie istotności $\alpha = 0.01$.
- Należy odrzucić H_0 na poziomie istotności $\alpha = 0.05$.
- Aby zastosować test zgodności χ^2-Pearsona, trzebaby w H_0 założyć z góry (tzn. przed rozpoczęciem eksperymentu) prawdopodobieństwa dla każdej klasy, nie można nic estymować z danych, tak jak zrobił Jaś.

Rozwiązanie: T, F, T, F

Zadanie 8 [Autor: SN, gr 2] (2 pkt) Malutki Jaś od jakiegoś czasu każdego ranka w swoich **dwóch** butach odnajduje po nocy 0, 1 albo 2 cukierki, w taki sposób umieszczone, że w każdym bucie jest co najwyżej jeden cukierek. Jaś jest uzdolniony matematycznie i wobec tego jest przekonany, że to skrzaty podrzucają mu w nocy cukierki, umieszczając w danym bucie (każdym niezależnie) z prawdopodobieństwem $0 < p < 1$ jeden cukierek, zaś z prawdopodobieństwem $1-p$ nie umieszczają w nim cukierka. Od szesnastu nocy liczy, ile cukierków łącznie danej nocy dostał, i jego obserwacje są następujące: $1, 0, 0, 2, 1, 0, 0, 2, 2, 2, 2, 1, 0, 0, 1, 2$. Słowem, było 6 nocy, kiedy nie otrzymał cukierków, 4 noce kiedy otrzymał jeden cukierek i 6 nocy, kiedy otrzymał dwa cukierki.

Jaś zechciał sprawdzić prawdziwość swoich przekonań: najpierw wyestymował wartość nieznanego prawdopodobieństwa p estymatorem największej wiarogodności $\hat{p}$, a następnie przeprowadził test zgodności χ^2-Pearsona, z hipotezą H_0, że liczba otrzymanych cukierków każdej nocy pochodzi ze schematu Bernoulliego z prawdopodobieństwem sukcesu $\hat{p}$ i o dwóch powtórzeniach.

Poniżej podano tabelę wartości dystrybuanty rozkładu χ^2 z jednym i dwoma stopniami swobody:

$x =$	1.0	2.0	3.0	4.0	5.0	6.0	7.0
$F_{\chi^2(1)}(x) =$	0.683	0.843	0.917	0.954	0.975	0.986	0.992
$F_{\chi^2(2)}(x) =$	0.393	0.632	0.777	0.865	0.918	0.950	0.970

Tabela 6: Tabela wartości dystrybuanty rozkładu χ^2

Rozstrzygnij prawdziwość odpowiedzi:

- Statystyka testowa ma rozkład χ^2 o dwóch stopniach swobody.
- Nie ma podstaw, by odrzucić H_0 na poziomie istotności $\alpha = 0.05$.
- Nie ma podstaw, by odrzucić H_0 na poziomie istotności $\alpha = 0.01$.
- Aby zastosować test zgodności χ^2-Pearsona, trzebaby w H_0 założyć z góry (tzn. przed rozpoczęciem eksperymentu) prawdopodobieństwa dla każdej klasy, nie można nic estymować z danych, tak jak zrobił Jaś.

Rozwiązanie: F, F, T, F

Zadanie 9 [Autor: SN, gr 1] (2 pkt) Rozstrzygnij, które implikacje są **zawsze** prawdziwe:

- dla klasyfikacji binarnej: jeśli w metodzie kNN z $k = 2$ dla badanego punktu nie ma remisu i przeważa klasa c, to metoda kNN z $k = 3$ dla tego punktu również zwróci klasę c.
- dla klasyfikacji binarnej: jeśli metoda kNN z $k = 3$ dla badanego punktu zwróci klasę c, to metoda kNN z $k = 2$ dla tego punktu również zwróci klasę c.
- dla klasyfikacji binarnej: jeśli metoda kNN z $k = 3$ dla badanego punktu zwróci klasę c, to metoda kNN z $k = 4$ dla tego punktu również zwróci klasę c.
- dla klasyfikacji dla trzech klas: jeśli w metodzie kNN z $k = 3$ dla badanego punktu nie ma remisu i przeważa klasa c, to metoda kNN z $k = 2$ dla tego punktu również zwróci klasę c.

Rozwiązanie: T, F, F, F

Zadanie 9 [Autor: SN, gr 2] (2 pkt) Rozstrzygnij, które implikacje są **zawsze** prawdziwe:

- dla klasyfikacji binarnej: jeśli metoda kNN z $k = 3$ dla badanego punktu zwróci klasę c, to metoda kNN z $k = 4$ dla tego punktu również zwróci klasę c.
- dla klasyfikacji binarnej: jeśli w metodzie kNN z $k = 4$ dla badanego punktu nie ma remisu i przeważa klasa c, to metoda kNN z $k = 5$ dla tego punktu również zwróci klasę c.
- dla klasyfikacji binarnej: jeśli metoda kNN z $k = 3$ dla badanego punktu zwróci klasę c, to metoda kNN z $k = 2$ dla tego punktu również zwróci klasę c.
- dla klasyfikacji dla trzech klas: jeśli w metodzie kNN z $k = 2$ dla badanego punktu nie ma remisu i przeważa klasa c, to metoda kNN z $k = 3$ dla tego punktu również zwróci klasę c.

Rozwiązanie: F, T, F, T

Zadanie 10 [Autor: PP, gr 1] (2 pkt) Rozważmy model liniowy $y = X\beta + \varepsilon$ o danej deterministycznej macierzy planu X wymiaru $n \times p$, która jest pełnego rzędu p. Załóżmy, że elementy wektora błędów ε są niezależnymi zmiennymi losowymi o wartości oczekiwanej równej 0 i wariancji równej σ^2. Załóżmy ponadto, że X rozłożono na macierz kolumnowo ortonormalną Q wymiaru $n \times p$ i górnotrójkątną R wymiaru $p \times p$ oraz obliczono estymator najmniejszych kwadratów $\hat{\beta}$. Wskaż, czy następujące zdania są prawdziwe:

- $\| y - X\hat{\beta} \|^2 = \| y \|^2 - \| Q^T y \|^2$.
- $\mathbf{E} \| y - X\hat{\beta} \|^2 = \sigma^2(n-p)$.
- $\mathbf{V}(\hat{\beta}) \neq \sigma^2 R^{-1}(R^{-1})^T$.
- $\mathbf{V}(X\hat{\beta}) = \sigma^2 QQ^T$.

Rozwiązanie: T, T, F, T

Zadanie 10 [Autor: PP, gr 2] (2 pkt) Rozważmy model liniowy $y = X\beta + \varepsilon$ o danej deterministycznej macierzy planu X wymiaru $n \times p$, która jest pełnego rzędu p. Załóżmy, że elementy wektora błędów ε są niezależnymi zmiennymi losowymi o wartości oczekiwanej równej 0 i wariancji równej σ^2. Załóżmy ponadto, że X rozłożono na macierz kolumnowo ortonormalną Q wymiaru $n \times p$ i górnotrójkątną R wymiaru $p \times p$ oraz obliczono estymator najmniejszych kwadratów $\hat{\beta}$. Wskaż, czy następujące zdania są prawdziwe:

- $\mathbf{E} \| y - X\hat{\beta} \|^2 = \mathrm{trace}(\mathbf{V}(y - QQ^T y))$.
- $\mathbf{E} \| X\beta - X\hat{\beta} \|^2 = \sigma^2 p$.
- $\mathbf{E} R^{-1} Q^T y = \beta$.
- $\| Q^T y \| \neq \| Q^T QQ^T y \|$.

Rozwiązanie: T, T, T, F