MiNI AKADEMIA

Od fragmentu po całość, czyli o sztuce uczenia się z danych

Warsztaty – 4 listopada 2017 r.

I. Podstawowe wykresy i analiza częstości

Zadanie 1. W tabelach 1 i 2 zamieszczono dane dotyczące koloru oczu (wg skali Martina) kobiet i mężczyzn z całej Polski, badanych w latach 1955-1956 (dane pochodzą ze strony *antropologia-fizyczna.pl*). Przedstawić dane zamieszczone w obu tabelach na wykresach kołowych i słupkowych. Omówić przydatność tych wykresów do ilustracji danych zamieszczonych w tabeli 1 i 2.

Tabela 1

Kolory oczu	Mężczyźni	Kobiety
ciemne 1. 2	0.09%	0.3%
ciemnobrązowe 3	1.19%	1.89%
piwne 4	4.04%	5.56%
jasnobrązowe 5	5.5%	6.14%
jasnobrązowe 6	4.96%	3.32%
ciemnozielone 7	9.90%	12.24%
jasnozielone 8	11.56%	15.25%
ciemnoszare 9	2.37%	3.73%
ciemnoszare 10	4.7%	6.13%
jasnoszare 11	5.38%	7.15%
jasnoszare 12	7.44%	5.97%
niebieskie 13	7.19%	5.91%
niebieskie 14	7.31%	4.68%
niebieskie 15	15.48%	10.44%
jasnoniebieskie 16	12.89%	11.28%

Tabela 2

Kolory oczu	Mężczyźni	Kobiety
jasne (12-16)	50.31%	38.28%
mieszane (7-11)	33.91%	44.5%
ciemne (1-6)	15.78%	17.22%

Tabela 3

Kolory oczu	Mężczyźni	Kobiety
jasne (9, 10 i 12-16)	42.64%	37.11%
mieszane (7, 8 i 11)	45.78%	48.05%
ciemne (1-6)	11.58%	14.83%

Zadanie 2. Według późniejszych badań rozkład koloru oczu przedstawia się tak, jak podano w tabeli 3.

a) Porównując wyniki zamieszczone w tabelach 1 oraz 3 podać, o ile procent wzrosła frakcja mężczyzn o mieszanym kolorze oczu? (Uwaga! W drugim badaniu przyjęto inną definicję koloru mieszanego).

b) O ile punktów procentowych wzrosła frakcja mężczyzn o mieszanym kolorze oczu?

Zadanie 3. W tabeli 4 zawarto informacje o pięciu najczęściej występujących literach w językach wietnamskim, islandzkim, norweskim, holenderskim, walijskim i galicyjskim. Poniżej w wymienionych językach zapisano twierdzenie Pitagorasa. Twoim zadaniem jest odgadnąć, które sformułowanie napisane jest w którym z tych języków. (Częstości liter na podstawie *https://www.sttmedia.com/characterfrequencies*. Uwaga: znaki diakrytyczne, czyli litery z kropkami, ogonkami itp., są odróżnialne od liter bez dodatków.)

Tabela 4: Najczęściej występujące litery w wybranych językach.

wietnamski		islandzki		norweski		holenderski		walijski		galicyjski	
N	11.01%	A	10.22%	E	16.63%	E	19.06%	D	9.88%	E	13.17%
H	7.95%	R	8.17%	N	8.14%	N	9.91%	A	9.36%	A	12.35%
C	6.71%	I	7.53%	T	7.79%	A	7.66%	Y	8.49%	O	10.29%
T	6.60%	E	7.50%	R	7.52%	T	6.42%	E	8.31%	R	7.02%
I	5.71%	N	7.28%	A	6.05%	I	6.29%	N	8.12%	S	7.01%

a) Ef gefinn er rétthyrndur þríhyrningur segir reglan til um að ef lögð eru saman önnur veldi skammhliða þríhyrningsins jafngildi sú summa öðru veldi langhliðarinnar.

b) In een rechthoekige driehoek is de som van de kwadraten van de lengtes van de rechthoekszijden gelijk aan het kwadraat van de lengte van de schuine zijde.

c) Mewn unrhyw driongl ongl sgwâr, mae arwynebedd y sgwâr sydd ag ochr yr hypotenws, yn hafal i swm arwynebau y sgwariau a'u hochrau eraill (sydd yn cwrdd ar yr ongl sgwâr).

d) En todo triángulo rectángulo, a hipotenusa ao cadrado é igual á suma dos cadrados dos catetos.

e) I en rettvinklet trekant er summen av kvadratene på katetene lik kvadratet på hypotenusen.

Zadanie 4. Odczytać poniższą wiadomości wiedząc, że została ona zakodowana za pomocą tzw. szyfru Cezara. Skorzystać w tym celu z informacji podanych w tabeli 5.

ÓŹĆAĘRAC HĆLŹAU SĘOF ĆŚĘHĆKŚŃK ŹCSGUPŃHĆŚK Ń ĘPKMUZCHĆŚK. ŚĘ YZĆCÓREJ ÓBEJZĘA, JĆŃLÓŃ ŹBKO ŹAĘGŃPŚUŻHŃ Ń AZBĘRUŻHŃ ŹAĘR ŹŃL ŹCSGUPKS ĆŃKSŃ Ń SĘAKZŃŃ. YUŚŃKBĘE SĘ HĆAKZC ÓFAC, SUEK GCI AKE ŚĘBŃFĆĘŚŃKS JU AĘÓŃHN „HĆBWZKÓ" OĘÓ ÓŃKZĄŚÓŃ ŻBŃĘAE, ECBŃURC Ń YUZC ZUÓĄ.

Tabela 5: Częstości występowania liter w języku polskim.

A	Ą	B	C	Ć	D	E	Ę
8.37%	0.79%	1.93%	3.89%	0.60%	3.35%	8.68%	1.13%
F	G	H	I	J	K	L	Ł
0.26%	1.46%	1.25%	8.83%	2.28%	3.01%	2.24%	2.38%
M	N	Ń	O	Ó	P	R	S
2.81%	5.69%	0.16%	7.53%	0.79%	2.87%	4.15%	4.13%
Ś	T	U	W	Y	Z	Ź	Ż
0.72%	3.85%	2.06%	4.11%	4.03%	5.33%	0.08%	0.93%

Wskazówka: W sposobie szyfrowania, zwanym szyfrem Cezara, każda litera pierwotnego tekstu jest zastępowana inną, oddaloną od niej o stałą liczbę pozycji w alfabecie (w przypadku, gdy podczas „przesuwania" kończy się alfabet, przeskakujemy na początek do litery A).

II. Charakterystyki liczbowe próbki

Zadanie 5. Wykazać, że średnia arytmetyczna posiada następujące własności:

a) Jeśli $\overline{x}$ jest średnią z liczb $x_1, \ldots, x_n$, to $(x_1 - \overline{x}) + \ldots + (x_n - \overline{x}) = 0$.

b) Jesteśmy zainteresowani znalezieniem takiej liczby a, która minimalizuje sumę kwadratów odchyleń od poszczególnych obserwacji liczbowych $x_1, \ldots, x_n$, czyli wyrażenie $(x_1 - a)^2 + \ldots + (x_n - a)^2$. Wykazać, że poszukiwaną liczbą jest średnia arytmetyczna, tzn. $a = \overline{x}$.

Podać interpretacje powyższych dwóch własności.

Zadanie 6. Policzono średnią arytmetyczną z pięciu pomiarów pewnej wielkości fizycznej i otrzymano 22. Wartość kolejnego, szóstego, pomiaru wyniosła 16.

a) Obliczyć średnią arytmetyczną z sześciu uzyskanych obserwacji.

b) Wyprowadzić wzór rekurencyjny na średnią arytmetyczną z $n + 1$ obserwacji, jako funkcję średniej z n obserwacji oraz $(n + 1)$-szej obserwacji.

Zadanie 7. Rozważyć następujące trzy sytuacje:

a) Pan Abacki jedzie samochodem z miejscowości A do B z prędkością 30 km/h. Jak szybko musi jechać z powrotem z B do A, aby średnia prędkość podczas całej podróży wyniosła 60km/h?

b) Pan Babacki jedzie przez godzinę z prędkością 30 km/h, a przez następną godzinę z prędkością 90 km/h. Jaka jest średnia prędkość podczas całej podróży?

c) Pan Cabacki jedzie samochodem z miejscowości A do B z prędkością 30 km/h, a z powrotem, z miejscowości B do A z prędkością 90 km/h. Jaka jest średnia prędkość podczas całej podróży?

Sformułować wnioski odnośnie stosowania różnych średnich.

Zadanie 8. Pan Oszczędny otworzył lokatę w banku na okres 4 lat, wpłacając 10000 złotych. Załóżmy, że była to lokata o zmiennym oprocentowaniu, przy czym w pierwszym roku oprocentowanie wynosiło 4%, a w kolejnych trzech latach, odpowiednio, 6%, 7% i 5%. Ile wynosi średnie oprocentowanie tej lokaty?

Zadanie 9. Wykazać, że poniższe wzory na wariancję próbkową są równoważne:

a) $s^2 = \frac{1}{n-1}\Big((x_1 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2\Big)$,

b) $s^2 = \frac{1}{n-1}\Big[\big(x_1^2 + \ldots + x_n^2\big) - n\overline{x}^2\Big]$.

Zadanie 10. Niech $x_1, \ldots, x_n$ oznacza próbkę oraz niech $y_i = x_i + C$ dla $i = 1, \ldots, n$. Wykazać, że $\overline{y} = \overline{x} + C$ oraz że $s_y^2 = s_x^2$.

Zadanie 11. W poniższej tabeli podano liczę klientów, która odwiedziła pewien sklep w ciągu kolejnych dziesięciu dni roboczych. Ilu klientów odwiedzało średnio ów sklep w ciągu dnia? Wyznaczyć wariancję liczby klientów odwiedzających ten sklep. (Wskazówka: posłużyć się wzorami podanymi w zadaniu 10.)

Dzień	pon.	wt.	śr.	czw.	pt.	pon.	wt.	śr.	czw.	pt.
Liczba klientów	136	130	137	134	138	135	137	139	142	132

III. Metoda Monte Carlo

Zadanie 12. Posługując się metodą Monte Carlo wyznaczyć pola powierzchni następujących figur:

a) trójkąta o wierzchołkach w punktach $(0, 0)$, $(1, 0)$, $(0.5, 1)$;

b) figury opisanej wzorem $\{(x, y) : 0 \leqslant x \leqslant 1, 0 \leqslant y \leqslant x^2\}$.

Zadanie 13. Zastosować metodę Monte Carlo do wyznaczenia liczby π.

IV. Metoda najmniejszych kwadratów

Zadanie 14. W tabelce podano liczby obserwujących profil kilku popularnych piłkarzy na Twitterze oraz liczby polubień ich oficjalnych stron na Facebooku.

Piłkarz	Facebook [mln]	Twitter [mln]
Cristiano Ronaldo	122.0	62.3
Neymar Jr.	60.0	34.2
Luis Suárez	18.0	12.0
Thomas Müller	9.4	3.7
Manuel Neuer	9.2	4.0
Andrés Iniesta	27.0	19.4
Arturo Vidal	2.0	3.6
Mario Balotelli	10.0	3.9
Mesut Özil	31.0	19.0
Gianluigi Bufon	4.8	2.8
Wayne Rooney	25.0	16.5
Marcelo	19.0	8.9
Antoine Griezmann	7.0	4.6
Ronaldinho Gaúcho	34.0	16.2
Cesc Fabregas	9.6	7.2

a) Posługując się ołówkiem i linijką, dopasować prostą, która opisuje zależność liczby fanów obserwujących na Twitterze profil popularnych piłkarzy od liczby polubień oficjalnej strony tychże piłkarzy na Facebooku.

b) Oficjalną stronę Messiego na Facebooku polubiło 89.2 mln osób. Ile osób zaczęłoby obserwować jego profil na Twitterze, gdyby go założył? (Wskazówka: skorzystając z wykresu otrzymanego w poprzednim punkcie zadania.)

c) Wyznaczyć liczbę potencjalnych obserwatorów profilu Messiego na Twitterze, jaka otrzymalibyśmy posługując się metodą najmniejszych kwadratów.

d) Stronę Roberta Lewandowskiego na Facebooku lubi 9.2 mln osób, a jego profil na Twitterze obserwuje 1 mln osób. Zaznaczyć na wykresie punkt odpowiadający Lewandowskiemu, a następnie obliczyć, o ile faktyczna liczba obserwujących profil piłkarza na Twitterze różni się od wartości wynikającej z przyjętego modelu.