

Zadanie 2 – Braki w danych

Eksploracja danych PS1

Artur Tagisow

Tomasz Kołtun

1. Oblicz dla próbek podstawowe statystyki - średnią, odchylenie standardowe, min, max, medianę, kwartyle



The screenshot shows a Jupyter Notebook interface with two tables of statistical data. The first table is for a 'uniform' distribution and the second is for a 'normal' distribution. Both tables have 9 columns: n, śr. arytm., odch. std, mediana, min, max, I kwantyl, and III kwantyl. The 'uniform' table shows values for n=1000, with a mean of 0.511721, standard deviation of 0.289935, median of 0.511873, min of 0.0005, max of 0.999155, 1st quartile of 0.252996, and 3rd quartile of 0.754. The 'normal' table shows values for n=1000, with a mean of 165.354099, standard deviation of 7.530122, median of 165.394064, min of 142.740787, max of 186.315081, 1st quartile of 160.053549, and 3rd quartile of 170.411374.


	n	śr. arytm.	odch. std	mediana	min	max	I kwantyl	III kwantyl
uniform	1000	0.511721	0.289935	0.511873	0.0005	0.999155	0.252996	0.754

	n	śr. arytm.	odch. std	mediana	min	max	I kwantyl	III kwantyl
normal	1000	165.354099	7.530122	165.394064	142.740787	186.315081	160.053549	170.411374

2. Na podstawie oryginalnej próbek stwórz serię próbek usuwając 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80% danych (w każdej próbce powinny być usuwane te same elementy, tylko rozszerzamy zakres usuniętych obserwacji)

3. Oblicz podstawowe statystyki dla tych próbek (tak jak w pkt. 2.)

Z usuniętymi dla rozkładu normalnego: (kolejno od 10% usunięć do 80%)



The screenshot shows a Jupyter Notebook interface with a table of statistical data for a normal distribution. The table has 9 columns: n, śr. arytm., odch. std, mediana, min, max, I kwantyl, and III kwantyl. The rows show results for n=1000 with varying percentages of data removed (10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%). The values for the mean, standard deviation, median, min, max, 1st quartile, and 3rd quartile are shown for each percentage of data removed.

	n	śr. arytm.	odch. std	mediana	min	max	I kwantyl	III kwantyl
10%	1000	165.354099	7.530122	165.394064	142.740787	186.315081	160.053549	170.411374
20%	1000	165.381921	7.586224	165.423063	142.740787	186.315081	160.148013	170.441179
30%	1000	165.52223	7.601543	165.394064	142.740787	186.315081	160.129914	170.379268
40%	1000	165.235749	7.634838	165.336555	142.740787	186.315081	159.836893	170.356941
50%	1000	165.267943	7.655391	165.423063	142.740787	186.315081	159.792076	170.438891
60%	1000	165.289026	7.670437	165.335431	142.740787	186.315081	159.792076	170.375268
70%	1000	164.996481	7.708965	165.163642	142.740787	186.315081	159.28397	170.29424
80%	1000	165.117782	7.830533	165.43376	142.740787	184.0403	159.370156	170.169038
90%	1000	164.547999	7.805184	165.241169	142.740787	184.0403	159.374135	170.143366

Z usuniętymi dla rozkładu jednolitego: (kolejno od 10% usunąć do 80%)

"ROZKŁAD JEDNOLITY"							
	n	śr. arytm.	odch. std	mediana	min	max	I kwantyl
0	1000	0.511721	0.289935	0.511673	0.0005	0.999155	0.262996
							0.764
	n	śr. arytm.	odch. std	mediana	min	max	I kwantyl
0	1000	0.513825	0.288979	0.510343	0.0005	0.999155	0.266094
							0.766125
	n	śr. arytm.	odch. std	mediana	min	max	I kwantyl
0	1000	0.514418	0.288387	0.515241	0.0005	0.995598	0.269012
							0.76902
	n	śr. arytm.	odch. std	mediana	min	max	I kwantyl
0	1000	0.520569	0.285342	0.520761	0.0005	0.995367	0.280257
							0.771697
	n	śr. arytm.	odch. std	mediana	min	max	I kwantyl
0	1000	0.518123	0.282179	0.520761	0.0005	0.992832	0.293143
							0.781696
	n	śr. arytm.	odch. std	mediana	min	max	I kwantyl
0	1000	0.512398	0.281138	0.523276	0.0005	0.992832	0.271896
							0.755354
	n	śr. arytm.	odch. std	mediana	min	max	I kwantyl
0	1000	0.516353	0.280089	0.523276	0.0005	0.992832	0.293143
							0.755354
	n	śr. arytm.	odch. std	mediana	min	max	I kwantyl
0	1000	0.519328	0.274858	0.520527	0.0005	0.992125	0.301433
							0.753215
	n	śr. arytm.	odch. std	mediana	min	max	I kwantyl
0	1000	0.543055	0.271158	0.545469	0.0005	0.992125	0.317927
							0.777952

4. Wypełnij nowe próbki

- wartością średnią (z pkt. 4 odpowiednią dla każdej kolumny, np. dla kolumny z usuniętymi 20% danymi, liczymy średnią z wartościami pustymi i następnie tą średnią uzupełniamy braki w tej kolumnie)
- medianą
- wylosowanymi w sposób losowy nowymi wartościami (najlepiej z tego samego rozkładu jak [dane](#) oryginalne z pierwszej próbki)

5. Oblicz podstawowe statystyki dla wypełnionych danych i przeanalizuj

- jak wpływa na statystyki brak danych (zwiększający się procent braków)
- jak wpływa wypełnianie braków średnią, medianą
- jak wpływa wypełnianie braków losowymi wartościami

Tabela danych jest tylko (3x8), że trudno byłoby je czytać, dlatego proszę o spojrzenie w załączony PDF z Jupyter Notebooka.

Wnioski

Widać, że między wszystkimi badanymi statystykami, najbardziej wrażliwe na zmiany przy wypełnianiu braków w danych są statystyki dotyczące rozkładu (I kwantyl, III kwantyl, odch. std)

Jak wpływa na statystyki brak danych (zwiększający się procent braków):

We wszystkich sprawdzanych statystykach (oprócz min i max) wyniki różnią się o tak małą ilość, że prawdopodobnie mieści się w granicy błędu statystycznego.

Podobnie w przypadku rozkładu jednolitego.

Jak wpływa wypełnianie braków średnią, medianą

Wypełnianie średnią sprawia, że odchylenie standardowe staje się bardzo małe eg. $2.22 \cdot 10^{-16}$ (zarówno dla rozkładu normalnego i jednolitego). Uzupełnianie medianą nie ma takiego efektu. Jest

to zrozumiałe, skoro składnikiem wzoru na odchylenie standardowe jest średnia arytmetyczna w całym zbiorze. Wypełnienie braków w danych właśnie średnią arytmetyczną sprawia, że dla wypełnionych próbek zachodzi $((\text{średnia arytmetyczna}) - (\text{średnia arytmetyczna}))^2$, czyli 0.

Wypełnianie medianą wpływa najbardziej na wartość I i III kwantyla – sprawia, że wraz z zwiększeniem ilości ubytków (a zatem liczby wartości, które wypełniamy medianą) I i III kwantyl zbliżają się do siebie.

Jak wpływa wypełnianie braków losowymi wartościami

Wypełnianie braków losową wartością między min a max w danym zbiorze sprawia, że uzyskiwane podstawowe statystyki są bardzo oddalone od przypadku bazowego.

Różnica między zbiorem jednostajnym a normalnym jest najbardziej widoczna w odchyleniu standardowym.

Najbardziej drastycznie zmienia się odchylenie standardowe w rozkładzie normalnym. Dzieje się tak, ponieważ losowe liczby nie respektują rozkładu normalnego.

Co ciekawe, mediana w przypadku rozkładu jednostajnego jest również mocno zmieniona. Po sporządzeniu sprawozdania zauważyliśmy, że wartości, którymi uzupełnialiśmy zbiór jednostajny mogły przyjmować tylko wartość 0 lub 1 (błąd w kodzie – `math.random(int(min) – int(max))`). Cu automatycznie sprawia, że uzyskany zbiór nie jest w rozkładzie jednostajnym, ponieważ wartości 0 i 1 mają o wiele większe prawdopodobieństwo wystąpienia..

Wraz z wrostem ubytku średnia arytmetyczna maleje, co wskazywałoby na to, że funkcja `math.random` w języku Python faworyzuje liczbę zero, gdy stoi przed wyborem między zero a jeden.