Projekt z Analizy Danych

Szymon Sergiel
Politechnika Lubelska
Wydział Matematyki i Informatyki Technicznej
Inżynieria i Analiza Danych

Spis treści

1 Wyznaczenie i zinterpretowanie wartości podstawowych statystyk.

1.1 Interpretacja statystyk

1. Średnia:
   - Wartość centralna lub przeciętna obliczona jako suma wszystkich wartości podzielona przez ich liczbę. Odpowiada za ogólne położenie danych.
2. Minimum(Min):
   - Najniższa wartość w zestawie danych. Pokazuje dolną granicę obserwacji.
3. Maksimum(Max):
   - Najwyższa wartość w zestawie danych. Określa górną granicę obserwacji.
4. Mediana:
   - Wartość środkowa w uporządkowanym zbiorze danych. Dzieli dane na dwie równe części i jest odporna na wartości odstające.
5. Pierwszy kwartyl (Q1):
   - Wartość, poniżej której znajduje się 25 % danych. Odpowiada za dolny ćwiartkowy podział danych.
6. Trzeci kwartyl (Q3):
   - Wartość, powyżej której znajduje się 25% danych (lub poniżej której znajduje się 75%). Odpowiada za górny ćwiartkowy podział danych.
7. Odchylenie Standardowe:
   - Miara rozproszenia danych wokół średniej. Informuje, jak bardzo wartości odbiegają od średniej. Wyższe wartości wskazują na większe zróżnicowanie.
8. Wariancja:
   - Miara zmienności, będąca kwadratem odchylenia standardowego. Informuje o rozrzucie danych względem średniej.

1.2 Wyznaczanie statystyk z wartościami odstającymi

X1	X2	X3	X4	X5	X6	X7	
45,69607	45,31328	95,48987	-44,3919	20,56007	50,52372	100,9561	Średnia
20,00153	10,02169	4,527749	-2974,79	2,801222	32,51864	-896,883	Min
2043,594	1018,777	2042,532	4015,591	1013,957	1058,551	2116,189	Max
44,11436	44,47955	93,35301	-44,5613	20,19612	49,99824	100,0544	Mediana
32,33114	27,44483	59,40515	-95,6587	16,69891	46,6288	93,30335	Q1
56,93303	62,37008	129,2033	7,064569	23,41082	53,24205	106,5273	Q3
47,13857	29,81763	61,95046	131,0155	22,77618	23,08346	56,20648	Odch.Standardowe
2222,045	889,0911	3837,859	17165,06	518,7542	532,846	3159,168	Wariancja

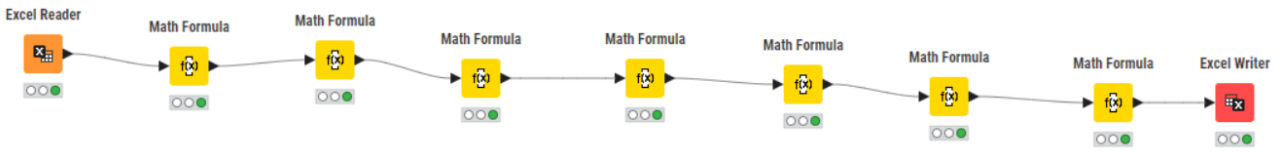
Rysunek 1: Podstawowe wartości statystyk z outlinerami

Interpretacja wartości

Zmienna X4 ma wartości bardzo rozproszone (duże różnice między minimum i maksimum, bardzo duże odchylenie standardowe i wariancję). Zmienna X7 wydaje się mieć względnie mniejsze zróżnicowanie, co można zaobserwować po niższej wariancji i mniejszych różnicach między kwartylami. Wartości kwartylowe (Q1, Q3) pozwalają ocenić asymetrię rozkładów danych — np. w X4 dolny kwartyl (-95,6587) jest znacznie bardziej negatywny niż górny (7,06459). Jak można zauważyć dla wszystkich kolum wartości min i max są bardzo od siebie oddalone przez co mają ogromny wpływ na warość średnią oraz odchylenie standardowe. Wprowadza to zamęt w statystykach więc trzeba się pozbyć wartości odstających. Wykorzystamy do tego program Knime .

1.3 Usuwanie wartości odstających z użyciem programu Knime

Jak to wygląda w Knime:



Rysunek 2: Użyte nody

Wzór na Z-score.

Do wyznaczenia wartości odstających wykorzystałem Z-score.

$$Z = \frac{X - \mu}{\sigma}$$

Gdzie

- Z to Z-score
- X to wartość zmiennej
- μ to średnia populacji
- σ to odchylenie standardowe populacji

Ponadto Z-score musi należeć do przedialu (-3:3) w przeciwnym wypadku jest to wartość odstająca.

W użytych nodach math formula występuje następująca formuła matematyczna.

```
1 if(3<($X1$-COL_MEAN($X1$))/COL_STDDEV($X1$),0,1)
2
```

Rysunek 3: Formuła Matematyczna cz.1

Taki wzór występuje w kazdym kolejnym nodzie tylko z odwołaniem do kolejnej kolumny. Ponieważ ten wzór sprawdza tylko wartości które są powyżej 3 potrzebny jest jeszcze jeden wzór który sprawdzi te poniżej -3.Do tego użyłem tego samego zestawu nodów co w **Rysunek 2**,tylko zmieniłem wzór na:

```
1 if(-3>($X1$-COL_MEAN($X1$))/COL_STDDEV($X1$),0,1)
2
```

Rysunek 4: Formuła Matematyczna cz.2

Usuwanie wierszy w programie Excel.

Po zastosowaniu powyższych wzorów oraz użyciu programu knime aby zapisał to wszystko w Exelu, pojawiły się dodatkowe kolumny z wartośćiami 0 i 1.W celu usuniecia wartości odstających trzeba usunąć wiersze przy których była wartość 0. W tym celu zaznaczałem wszytkie kolumny z wartościami 0 i 1 oraz użyłem skrótu klawiszowego **CTRL+G→Specjalnie→różnice w kolumach** ,który pozwala na przejśćie do konkretnych komórek które są inne od pozostałych. Ponieważ duża większość komórek miały w sobie 1 przekierowywało mnie do tych gdzie była wartość 0 a następnie usuwałem cały wiersz. Z danych zostało usniętych 14 wierszy .

1.4 Wyznaczanie podstawowych wartości bez wartości odstających

Ponownie wyliczamy podstawowe wartości ale tym razem bez wartośći odstających.

X1	X2	X3	X4	X5	X6	X7	
45,09947	44,3645	93,12745	-42,7405	19,90006	49,96588	99,95046	Średnia
20,01024	10,05796	4,564668	-238,477	5,6379	35,65386	71,05499	Min
69,95664	103,4838	211,8873	118,4073	37,08842	68,93629	138,6079	Max
45,40073	44,39988	93,17017	-41,1493	19,74475	49,85367	99,80172	Mediana
32,60088	27,26592	59,05714	-93,5169	16,46553	46,4182	92,93179	Q1
57,78208	60,92254	126,3339	10,36376	23,28843	53,30146	106,6909	Q3
14,53429	19,90248	39,9204	66,13035	4,902078	5,070435	10,19079	Odch.Sta
211,2455	396,1086	1593,639	4373,223	24,03037	25,70931	103,8522	Wariancja

Rysunek 5: Podstawowe wartości statystyk bez outlinerów

Po odjęciu wartości odstających średnie dla większości zmiennych uległy zmniejszeniu, co sugeruje, że wartości odstające były zazwyczaj znacznie wyższe niż przeciętne wartości (np. dla X4 średnia zmniejszyła się z -44,3919 do -42,7405). W przypadku X7 zmiana jest niewielka, co wskazuje na niewielki wpływ wartości odstających na tę zmienną.Wartości minimalne uległy zmianie, np. dla X4 wcześniej było -2974,79, a teraz wynosi -238,477, co oznacza, że ekstremalne wartości zostały usunięte.Maksymalne wartości również się zmniejszyły, np. dla X3 z 2042,532 do 211,8873, co wyraźnie pokazuje, że największe wartości odstające zostały odfiltrowane.Mediana uległa niewielkim zmianom w większości zmiennych (np. dla X2 z 44,47955 do 44,39988), co wskazuje, że wartości środkowe nie były silnie zaburzone przez wartości odstające.Kwartyle również zmieniły swoje wartości, co wskazuje na bardziej zbliżony do symetrycznego rozkład po usunięciu wartości odstających. Przykład: dla X6 Q3 zmieniło się z 53,24205 do 53,30146.Odchylenie standardowe znacznie zmniejszyło się w większości przypadków. Dla X4 zmniejszyło się z 131,0155 do 66,13035, co oznacza mniejsze rozproszenie danych po usunięciu wartości odstających.Wariancja, będąca kwadratem odchylenia standardowego, zmniejszyła się bardzo wyraźnie, np. dla X3 z 3837,859 do 1593,639. To również pokazuje mniejsze zróżnicowanie danych.

Usunięcie wartości odstających znacząco wpłynęło na zakres (min, max), odchylenie standardowe i wariancję, co wskazuje, że pierwotne dane były silnie zniekształcone przez skrajne obserwacje. Największy wpływ usunięcie wartości odstających miało na zmienne X3, X4 i X5, które charakteryzowały się dużymi różnicami między wartościami minimalnymi i maksymalnymi w pierwotnych danych.

2 Tworzenie szeregu rodzielczego dla zmiennej X1 o 10 przedziałach

2.1 Przygotowanie danych potrzebnych do tworzenia szeregu

- Wyliczenie wartości Min i Max dla zmiennej X1 które już posiadamy.
- Obliczenie rozpiętości czyli długości szeregu który wyliczamy poprzez odjęcie od wartośći Max wartość Min.
- Ustalenie liczby przedziałów. W tym wypadku wynosi ona 10.
- Obliczenie kroku czyli długością między lewym końcem przedziału a prawym końcem przedziału. Wylicza sie go poprzez podzielenie rozpiętości przez liczbę przedziałów.

Dane wyglądają następująco :

X1min	20,01024
X1max	69,95664
Rozpiętość	49,9464
Przedziały	10
Krok	4,99464

Rysunek 6: Dane potrzebne do tworzenia szeregu

2.2 Tworzenie szeregu rozdielczego

Szereg rozdzielczy tworzymy w następujący sposób :

1. Pierwszy lewy koniec przedziału to wartość Min.
2. Aby wyliczyć pierwszy prawy koniec przedziału należy dodać do lewego końca przedziału Krok.
3. Pierwszy prawy koniec przedziału staje sie drugim lewym końcem przedziału.
4. Powtarzamy czynności z 2 i 3 punktu aż dojdziemy do dziesiątego prawego końca przedziału w tym przypadku wpisujemy tam wartość Max.

Mój szereg wygląda następująco :

	lewy koniec	prawy konie
1	20,010235	25,004875
2	25,004875	29,999515
3	29,999515	34,994156
4	34,994156	39,988796
5	39,988796	44,983436
6	44,983436	49,978076
7	49,978076	54,972716
8	54,972716	59,967356
9	59,967356	64,961996
10	64,961996	69,95664

Rysunek 7: Szereg Rozdzielczy

2.3 Wyznaczanie wartości statystyk

Częstość(n_i):

Wylicza się to za pomocą funkcji CZĘSTOŚĆ w Exel. W oknie tablica dane zaznaczamy całą koulmnę ze zmiennymi X1 a w oknie tablica przedziały wszystkie prawe końce przedziału z naszego szeregu. Funkcja ta wylicza ile danych z naszej zaznaczonej kolumny znajduje sie w poszczególnych przedziałach.

Średnia z przedziałów (x_i):

Obliczamy tą statystykę funkcją ŚREDNIA. Zaznaczamy pierwszy lewy i prawy koniec przedziału i robimy tak samo z następnymi.

Pozostałe statystyki to iloczyn **średniej i częstości(x_i*n_i)** oraz iloczyn **kwadratu średniej i częstości(x_i^2*n_i)**. Te statystki są potrzebne później do wyliczenia Średniej , Wariacji oraz Odchylenia standardowego . Dla wszystkich satystyk które otrzymaliśmy wyliczymy **Sumę**.

Do wyliczenia **Średniej** używamy sumy z iloczynów średniej z przedziałów oraz częstości (x_i*n_i) dzieloną na sumę częstości.

$$S = \frac{\sum x\_i * n\_i}{\sum n\_i}$$

Wariację wyliczamy z sumy (x_i^2*n_i) podzieloną przez częstośc od której odejmujemy kwadrat obliczonej wcześniej średniej.

$$Var = \frac{\sum x\_i^2 * n\_i}{\sum n\_i} - S^2$$

Odchylenie standardowe obliczamy poprzez spierwiastkowanie Wariancji.

n_i	x_i	x_i*n_i	x_i^2*n_i
209	22,50756	4704,079	105877,3
181	27,5022	4977,897	136903,1
211	32,49684	6856,832	222825,4
200	37,49148	7498,295	281122,1
179	42,48612	7605,015	323107,5
211	47,48076	10018,44	475683,1
166	52,4754	8710,916	457108,7
227	57,47004	13045,7	749736,7
196	62,46468	12243,08	764759,8
205	67,45932	13829,16	932905,7
0			

Rysunek 8: Wartośći Statystyk cz.1

SUMA	1985	449,8344	89489,41	4450029
Średnia	45,08283			
Wariancja	209,3673			
Odch.Stand	14,46953			

Rysunek 9: Wartośći Statystyk cz.2

3 Uzupełnianie danych

Do uzupełniania danych użyłem mediany z kolumn zmiennych. Mediana jest dobrym wyborem do uzupełniania brakujących danych, ponieważ jest bardziej odporna na wartości odstające niż średnia. Oto szczegółowe powody, dlaczego warto używać mediany:

1. Odporność na wartości odstające
   - Mediana to wartość środkowa w uporządkowanym zbiorze danych, co oznacza, że nie zależy od ekstremalnie dużych lub małych wartości.
   - Średnia (np. arytmetyczna) jest podatna na odchylenia, jeśli w danych występują skrajne wartości. Może to prowadzić do nieprawidłowego wypełnienia braków.
2. Zachowanie charakterystyki danych
   - Mediana lepiej reprezentuje centralną tendencję danych w przypadku rozkładów niesymetrycznych (skośnych).
   - Pozwala to uniknąć przesuwania danych w stronę wartości dominujących w outlierach.
3. Łatwość obliczeń przy brakujących wartościach
   - Przy uzupełnianiu braków medianą nie musimy martwić się o to, jak wartości odstające wpływają na całość zbioru – wystarczy sortowanie i znalezienie wartości środkowej.
4. Uniwersalność
   - Mediana działa dobrze zarówno dla małych, jak i dużych zestawów danych, niezależnie od ich rozkładu.
   - Można ją stosować także przy dużych brakach danych, gdzie wypełnianie średnią może być mniej reprezentatywne.

Aby uzupełnić dane znów wykorzystałem skrót klawiszowy **CTRL+G→Specjalnie→ puste** tylko tym razem zaznaczyłem puste . W ten sposób program szukał mi pustych kolumn które po znalezieniu są od razu są zaznaczone do edytowania zatem wystrczy wpisać do nich medianę . Po całym zabiegu znów wyliczyłem podstawowe statystyki które wyglądają następująco :

X1	X2	X3	X4	X5	X6	X7	
45,09962	44,36455	93,1275	-42,7397	19,89991	49,96588	99,95023	Średnia
20,01024	10,05796	4,564668	-238,477	5,6379	35,65386	71,05499	Min
69,95664	103,4838	211,8873	118,4073	37,08842	68,93629	138,6079	Max
45,40073	44,39988	93,17017	-41,1493	19,74475	49,85367	99,80172	Mediana
32,60174	27,29095	59,06764	-93,5159	16,47258	46,4182	92,94332	Q1
57,76329	60,88678	126,2286	10,21518	23,28464	53,30146	106,688	Q3
14,53063	19,88743	39,90029	66,1137	4,89961	5,070435	10,18309	Odch.Sta
211,1391	395,5099	1592,033	4371,022	24,00618	25,70931	103,6953	Wariancja

Rysunek 10: Wartośći Statystyk po uzupełnianiu danych

Porównując do porzednich wartości nie zaszły bardzo widoczne zmiany a w niektórych praktycznie żadne.

4 Podsumowanie

1. Cel projektu
    - Raport dotyczy analizy danych zmiennych X1–X7 w celu uzupełnienia braków danych oraz stworzenie szeregu rodzielczego dla zmiennej X1
2. Podstawowe statystyki
    - Analiza wykazała, że zmienne mają zróżnicowane charakterystyki, w tym duże rozbieżności w wartościach maksymalnych i minimalnych. Dlatego musiałem przeprowadzić usuwanie wartości odstających aby zmniejszyć tą rozbieżność oraz aby dane były bardzej podatne na analizę .Po usunieciu zmiennych wartości minimalne i maksymalne się zmieniły. Mediana okazała się bardziej reprezentatywna niż średnia, szczególnie w przypadku zmiennych X3 i X4, które charakteryzują się znaczną zmiennością (wariancja wynosi odpowiednio 1593,639 i 4373,223).
3. Uzupełnienie braków
    - Brakujące dane zostały uzupełnione za pomocą mediany, co pozwoliło zachować reprezentatywność zmiennych mimo obecności wartości odstających. Mediana była bardziej odpowiednia niż średnia, ponieważ lepiej odzwierciedlała centralną tendencję danych.
4. Wnioski
    - Dane zostały oczyszczone z wartości odstających i braków, co poprawiło ich spójność i umożliwia przeprowadzenie dalszej analizy. Zaleca się jednak dodatkową weryfikację zmiennych X3 i X4, które cechują się wysoką wariancją, mogącą wpływać na wyniki kolejnych etapów.