



UNIwersytet WARMIŃSKO-MAZURSKI
W OLSZTYNIE

Modelowanie Statystyczne: „Porównanie kosztów transportu
latem i zimą”

Górski Adam

Wydział Matematyki i Informatyki

Projektowanie Systemów Informatycznych i Sieci Komputerowych.

Matematyczne Modelowanie Systemów.

dr ***** .

Spis treści

1. Opis danych	4
2. Sformułowanie problemu badawczego w oparciu o wybrane cechy z przedstawionego zestawu danych	4
3. Wyznaczenie parametrów statystycznych dla grup Zima i Lato.....	6
3.1 Analiza otrzymanych wyników	7
3.2 Wnioski	7
4. Zbudowanie szeregu rozdzielczego dla obu cech	10
4.1 Wykresy	13
4.2 Wnioski	17
4.2.1 Zbiór Zima.....	17
4.2.2 Zbiór Lato.....	17
4.2.3 Ogólne wnioski dla Lata i Zimy	18
5. Zmiana uporządkowania szeregu rozdzielczego	19
5.2 Porównanie wyników z punktem 4	26
5.3 Wnioski.....	26
6. Weryfikowanie zgodności próby z rozkładem teoretycznym	27
6.1 Obliczenia Zima chi-kwadrat	27
6.1.1 Sformułowanie hipotez dla rozkładu normalnego – Zima	27
6.1.2 Opis testu $\alpha = 0.01$	27
6.1.3 Opis testu $\alpha = 0.05$	27
6.1.4 Wnioski	27
6.2 Obliczenia Lato chi-kwadrat	28
6.2.1 Sformułowanie hipotez dla rozkładu normalnego – Lato	28
6.2.2 Opis testu $\alpha = 0.01$	28
6.2.3 Opis testu $\alpha = 0.05$	28
6.2.4 Wnioski	28
6.3 Obliczenia Zima rozkład Kołmogorowa	29
6.3.1 Sformułowanie hipotez dla rozkładu Kołmogorowa – Zima	29
6.3.2 Opis testu $\alpha = 0.01$	29
6.3.3 Opis testu $\alpha = 0.05$	29
6.3.4 Wnioski	29
6.4 Obliczenia Lato rozkład Kołmogorowa	30
6.4.1 Sformułowanie hipotez dla rozkładu Kołmogorowa – Lato	30
6.4.2 dla $\alpha = 0.01$	30
6.4.3 dla $\alpha = 0.05$	31
6.4.4 Wyniki testów	31
7. Porównanie wyników punktu 6 z punktem 3	31

Spis tabel

Tabela 1 Wyniki obliczeń parametrów statystycznych	6
Tabela 2 Szereg rozdzielnicy dla Zimy	10
Tabela 3 Cechy szeregu dla Zimy	11
Tabela 4 Szereg rozdzielnicy dla Lata.....	12
Tabela 5 Cechy szeregu dla Lata.....	13
Tabela 6 Szereg rozdzielnicy dla Zimy	20
Tabela 7 Cechy szeregu dla Zimy	20
Tabela 8 Szereg rozdzielnicy dla Lata.....	21
Tabela 9 Cechy szeregu dla Lata.....	22
Tabela 10 Wyniki obliczeń rozkład chi-kwadrat-Zima.....	27
Tabela 11 Wyniki obliczeń rozkład chi-kwadrat-Lato.....	28
Tabela 12 Wyniki obliczeń rozkład Kołmogorowa Zima	29
Tabela 13 Wyniki obliczeń rozkład Kołmogorowa Lato	30

Spis rysunków

Rysunek 1 Koszty transportu Zima	8
Rysunek 2 Koszty transportu Lato	8
Rysunek 3 Wykres liczebności dla Zimy	13
Rysunek 4 Wykres liczebności skumulowanej dla Zimy.....	14
Rysunek 5 Wykres prawdopodobieństwa empirycznego dla Zimy	14
Rysunek 6 Wykres dystrybuanty empirycznej dla Zimy	15
Rysunek 7 Wykres liczebności dla Lata	15
Rysunek 8 Wykres liczebności skumulowanej dla Lata	16
Rysunek 9 Wykres prawdopodobieństwa empirycznego dla Lata.....	16
Rysunek 10 Wykres dystrybuanty empirycznej dla Lata.....	17
Rysunek 11 Wykres liczebności dla Zimy	22
Rysunek 12 Wykres liczebności skumulowanej dla Zimy.....	23
Rysunek 13 Wykres prawdopodobieństwa empirycznego dla Zimy	23
Rysunek 14 Wykres dystrybuanty empirycznej dla Zimy	24
Rysunek 15 Wykres liczebności dla Lata.....	24
Rysunek 16 Wykres liczebności skumulowanej dla Lata	25
Rysunek 17 Wykres prawdopodobieństwa empirycznego dla Lata.....	25
Rysunek 18 Wykres dystrybuanty empirycznej dla Lata.....	26

1. Opis danych

Wykorzystane dane pochodzą z witryny internetowej znajdującej się pod adresem: <https://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work>. Zbiór przedstawia realne dane o grupie pracowników oraz o ich przyczynach nieobecności w pracy. Zestaw składa się z 21 kolumn i zawiera szczegółowe informacje na temat nieobecności pracowników ich przyczyn z podziałem na kolumny. Cały zbiór przedstawia wartości liczbowe. Minimalna wartość danych to 0, a maksymalna to 388. Podczas analizy pierwszej kolumny zbioru można znaleźć informację na temat liczby pracowników których dotyczą te dane liczba ich nie jest duża wynosi 36 osób ponieważ maksymalna wartość w kolumnie ID to 36. Oprócz liczby pracowników możemy sprawdzić dzień tygodnia, porę roku, a także uzyskać takie dane jak wiek, poziom wykształcenia oraz wiele danych takich jak waga na podstawie których możemy wyciągać wiele przyczyn i wniosków dlaczego pracownik nie przychodzi do pracy. Cały zbiór liczy 741 wierszy. Z tak pokaźnego zbioru danych wybrałem dwie najbardziej interesujące mnie kolumny, które dotyczą pory roku oraz kosztów transportu. Poniższa analiza statystyczna ma pokazać jaka jest różnica pomiędzy dwoma porami roku zimą i latem w obszarze kosztów transportu.

2. Sformułowanie problemu badawczego w oparciu o wybrane cechy z przedstawionego zestawu danych

Problem badawczy: „Porównanie kosztów transportu latem i zimą”.

W swoich rozważaniach pragnę się skupić na dwóch kolumnach, a mianowicie:

- A. Seasons- która oznacza pory roku,
- B. Transportation expense- oznaczająca koszty transportu.

Z kolumny dotyczącej pór roku wybrałem dane oznaczone literą 2 oraz 4 które oznaczają:

2- lato,

4- zima.

Dla wybranych pór roku przyporządkowałem odpowiadające im koszty transportu dzięki czemu wyodrębniłem zbiór danych o liczbie wierszy równej 195. Na tych danych wykonam analizę opierając się o funkcje statystyczne takie jak: minimum, maksimum, suma, średnia, reguła 3-sigma, a także wiele innych, które pozwolą pokazać jaka jest różnica pomiędzy latem i zimą. Na podstawie tej różnicy postaram się wysnuć wniosek skąd mogą pochodzić te dane,

ponieważ jeśli mamy do czynienia z półkulą wschodnią lub rejonem kół podbiegunowych to różnica pomiędzy latem, a zimą będzie bardzo widoczna, ponieważ ilość paliwa potrzebna przy uruchomieniu silnika będzie znacząca. Koszty zwiększą się również, gdy będziemy musieli podgrzać olej tak aby jego lepkość i gęstość była na odpowiednim poziomie przed włączeniem pojazdu co generuje dodatkowe koszty związane z dojazdem do pracy i wydłuża czas trwania tej czynności. Uważam, że moja analiza jest cenna szczególnie, gdy planowałbym system, który w zależności od rodzaju pory roku, danych pogodowych bądź natężenia ruchu, a także dnia tygodnia oraz miesiąca miałby podpowiadać jak manipulować cenami na stacjach paliw, aby koszt dochodu byłby jak największy, ponieważ nie wystarczy podnieść cenę czasami lepiej jest obniżyć, aby zachęcić ludzi do odwiedzania danej stacji paliw.

3. Wyznaczenie parametrów statystycznych dla grup Zima i Lato

Parametry statystyczne	Zima	Lato
średnia	231,0974359	221,7538462
odchylenie standardowe	69,15909376	67,98424345
wariancja	4807,634787	4645,681364
przedział wartości typowych (typowe 1)- maksimum	300,2565297	289,7380896
przedział wartości typowych (typowe 1)- minimum	161,9383421	153,7696027
procent wartości typowych (typowe 1)	74,35897436	73,84615385
przedział wartości typowych (typowe 2)- maksimum	369,4156234	357,722333
przedział wartości typowych (typowe 2)- minimum	92,77924838	85,78535926
procent wartości typowych (typowe 2)	98,97435897	91,79487179
wartości odstające	2	16
odchylenie przeciętne	53,69893491	53,94445759
współczynnik zmienności	0,299263787	0,306575262
minimum	118	118
kwartyl I	179	179
mediana	225	225
kwartyl III	289	260
maksimum	378	388
rozstęp maksimum – minimum	260	270
rozstęp międzykwartylowy	110	81
odchylenie ćwiartkowe	55	40,5
przedział kwartylowych wartości typowych (typowe 1) -maksimum	280	265,5
przedział kwartylowych wartości typowych (typowe 1) -minimum	170	184,5
Procent	58,97435897	34,87179487
przedział kwartylowych wartości typowych (typowe 2) -maksimum	454	381,5
przedział kwartylowych wartości typowych (typowe 2) -minimum	124	138,5
procent	87,17948718	87,69230769
wartości odstające	25	24
suma	45064	43242
średnia	44153	

Tabela 1 Wyniki obliczeń parametrów statystycznych

3.1 Analiza otrzymanych wyników

Z przeprowadzonych obliczeń wynika, że suma kosztów zimą jest większa co jest wynikiem oczywiście prawidłowym, ponieważ silnik zimą z racji na niższą temperaturę potrzebuje więcej paliwa do jego ogrzania. Różnica nie jest zbyt znaczna, lecz może wpływać na to kilka czynników takich jak:

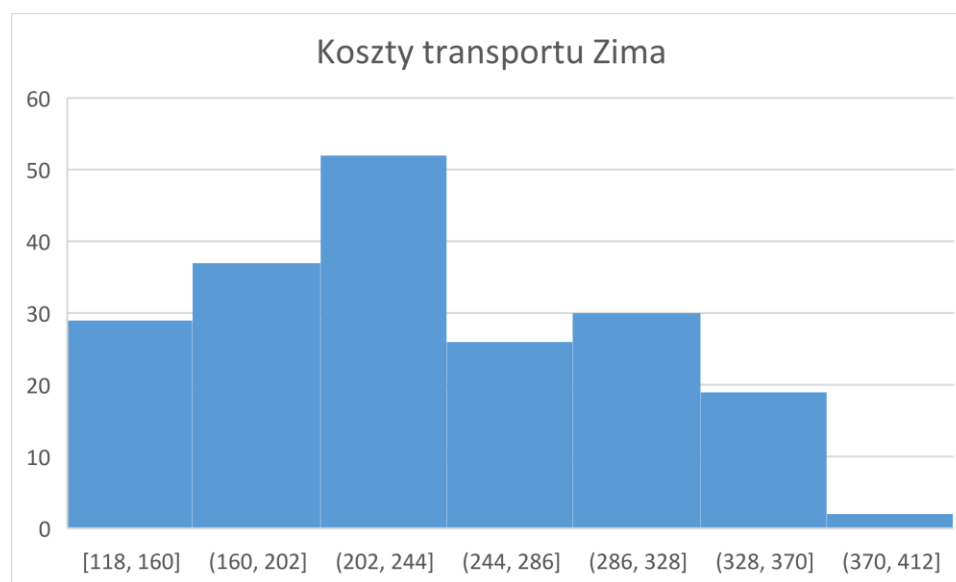
- średnia temperatura powietrza,
- zmienna cena paliw,
- używanie klimatyzacji latem,
- czas włączenia silnika na postoju (zimą olej jest bardziej gęsty potrzebuje więcej czasu na rozprowadzenie),
- większa prędkość pojazdu latem.

Średnie koszty zimą są większe niż średnie koszty latem co oczywiście potwierdza wynik zsumowania tych zbiorów danych. Wartości minimalne latem i zimą są takie same na uwagę zasługuje fakt, że wartość maksymalna latem okazała się większa. Uważam, że takie odchylenie jest wytłumaczalne, ponieważ uruchomiona klimatyzacja, dobra pogoda oraz wysoka prędkość powoduje większe spalanie naszego pojazdu. Ciekawą rzeczą jest fakt, że mediana dla obu zbiorów wyszła identyczna przyjmując wartość 225, a warto zwrócić uwagę, że średnia zimą jest większa od wartości mediany, lecz latem średnia okazała się mniejsza co może mieć wpływ na wygląd wykresu i na jego asymetrię. Analiza zgodności z rozkładem normalnym dla reguły 3-sigma wykazała, że procentowa liczba wartości w obszarze pierwszym dla obu grup utrzymuje się w granicach 74% co nie jest najlepszym wynikiem. Wartość idealna dla obszaru pierwszego to 68%. Dalsza analiza pokazuje, że dla drugiego obszaru wartości są dokładniejsze, ponieważ dla zimy to 98%, a dla lata 91% lecz oba zbiory nie osiągnęły idealnego wyniku równego 95%. Warto również zwrócić swój wzrok na wyniki współczynnika zmienności który osiągnął wartość około 0,3. W procentach uzyskamy wartość na poziomie 30% co oznacza, że obserwujemy tzw. „przeciętną zmienność,” co świadczy, że dane występujące w obu grupach są niejednorodne, co możemy zauważyć otwierając plik arkusza kalkulacyjnego.

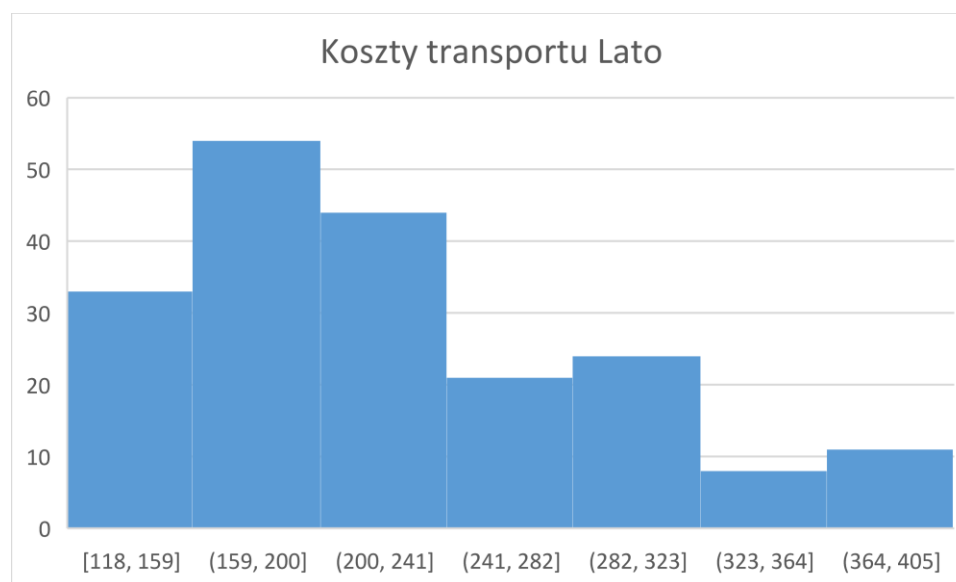
3.2 Wnioski

Z dokonanych obliczeń możemy odczytać kilka ważnych informacji, które mogą już na etapie analizy rozpoznać jak będą rozkładać się dane na wykresie. Pomocne są w tym dwie

wartości średnia i mediana. W zbiorze oznaczającym zimę średnia góruje na medianą co oznacza, że mamy do czynienia z asymetrią prawostronną. Odmierna sytuacja panuje w zbiorze danych lato to mediana góruje nad średnią co oznacza, że asymetria będzie lewostronna. Wyniki te potwierdzają poniższe histogramy.



Rysunek 1 Koszty transportu Zima



Rysunek 2 Koszty transportu Lato

Następną warto uwagi cechą jest wynik oznaczonych jako „typowe 1” oraz „typowe 2” który może dokładnie zobrazować nam w jakim stopniu dane rozkładają się w porównaniu do rozkładu normalnego który króluje i modeluje dane uzyskane w naszym świecie. Dzięki

analizie tego parametru można zauważyć nieprawidłowości oraz wysnuć bardzo interesujące wnioski. Wynik dla „*typowe 1*” to około 73%, a dla drugiego poziomu „*typowe 2*” to 98% dla Zimy i 91% dla Lata co oznacza, że kierowcy mogli podawać dane lekko odbiegające od rzeczywistości w celu uzyskania większych zwrotów. Zazwyczaj w takiej sytuacji każdy zaokrągla w górę, ponieważ każdy woli posiadać więcej niż tracić. Dlatego też wartości procentowe różnią się od tych wzorcowych. Jak wspomniałem wartość współczynnika zmienności osiągnęła 0,3 co przeliczeniu na procenty daje wynik 30%. Na tej podstawie można zauważyć słuszną tezę, że wartości są umiarkowanie zróżnicowane.

4. Zbudowanie szeregu rozdzielczego dla obu cech

przedziały grupowania		liczebność	liczebność skumulowana	prawdopodobieństwo empiryczne	dystribuanta empiryczna	środki przedziałów	obliczenia średniej	obliczenie wariancji	
l	P	ni	ni_sk	fi	Fi	xi	xi*ni	(xi-xsr)^2	ni*(xi-xsr)^2
103	132	25	25	0,128205	0,128205	117,5	2937,5	12842,12	321053
132	161	4	29	0,020513	0,148718	146,5	586	7110,381	28441,53
161	190	37	66	0,189744	0,338462	175,5	6493,5	3060,643	113243,8
190	219	0	66	0	0,338462	204,5	0	692,9044	0
219	248	64	130	0,328205	0,666667	233,5	14944	7,165917	458,6187
248	277	13	143	0,066667	0,733333	262,5	3412,5	1003,427	13044,56
277	306	31	174	0,158974	0,892308	291,5	9036,5	3681,689	114132,4
306	335	2	176	0,010256	0,902564	320,5	641	8041,951	16083,9
335	364	8	184	0,041026	0,94359	349,5	2796	14084,21	112673,7
364	393	11	195	0,05641	1	378,5	4163,5	21808,47	239893,2

Tabela 2 Szereg rozdzielczy dla Zimy

h=	29
x01=	103
liczebność próby	195
średnia z szeregu	230,8231
odchylenie z szeregu	70,12899
procent typowych wartości wyznaczonych z odchylenia i średniej	74%
współczynnik zmienności	0,303821
procent wartości typowych ² wyznaczonych z odchylenia i średniej	94%
liczba wartości odstających po lewej stronie	0
liczba wartości odstających po prawej stronie	2

Tabela 3 Cechy szeregu dla Zimy

przedziały grupowania		liczebność	liczebność skumulowana	prawdopodobieństwo empiryczne	dystribucja empiryczna	środki przedziałów	obliczenia średniej	obliczenia wariancji	
L	p	ni	ni_sk	fi	Fi	xi	xi*ni	(xi-xsr)^2	ni*(xi-xsr)^2
108	137	22	22	0,112821	0,112821	122,5	2695	11733,89	258145,6
137	166	11	33	0,05641	0,169231	151,5	1666,5	6292,151	69213,66
166	195	54	87	0,276923	0,446154	180,5	9747	2532,412	136750,3
195	224	0	87	0	0,446154	209,5	0	454,6736	0
224	253	55	142	0,282051	0,728205	238,5	13117,5	58,93515	3241,433
253	282	10	152	0,051282	0,779487	267,5	2675	1345,197	13451,97
282	311	24	176	0,123077	0,902564	296,5	7116	4313,458	103523
311	340	3	179	0,015385	0,917949	325,5	976,5	8963,72	26891,16
340	369	13	192	0,066667	0,984615	354,5	4608,5	15295,98	198847,8
369	398	3	195	0,015385	1	383,5	1150,5	23310,24	69930,73

Tabela 4 Szereg rozdzielczy dla Lata

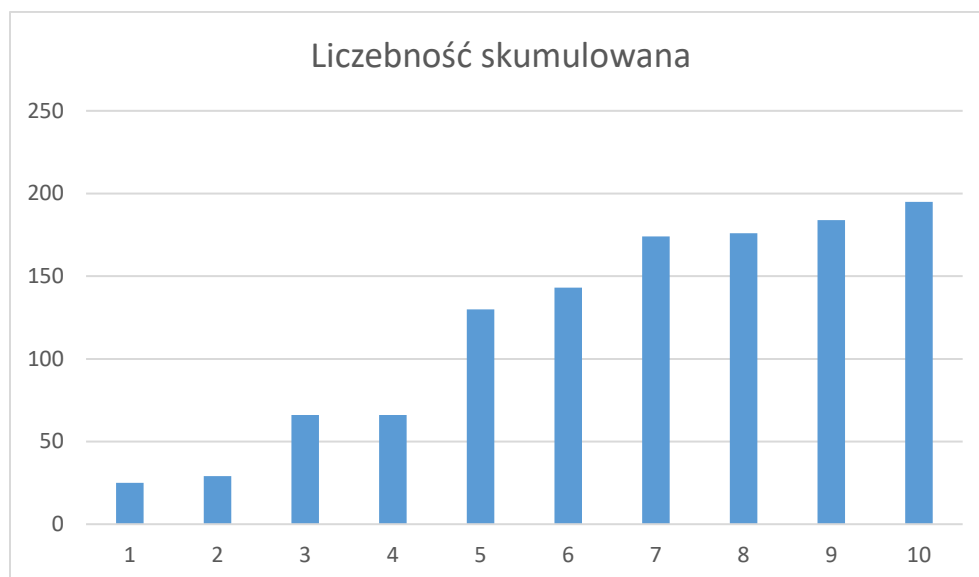
h=	20
x01=	108
liczebność próby	195
średnia z szeregu	224,3718
odchylenie z szeregu	67,17736
procent typowych wartości wyznaczonych z odchylenia i średniej	61%
współczynnik zmienności	0,299402
procent wartości typowych ² wyznaczonych z odchylenia i średniej	98%
liczba wartości odstających po lewej stronie	0
liczba wartości odstających po prawej stronie	16

Tabela 5 Cechy szeregu dla Lata

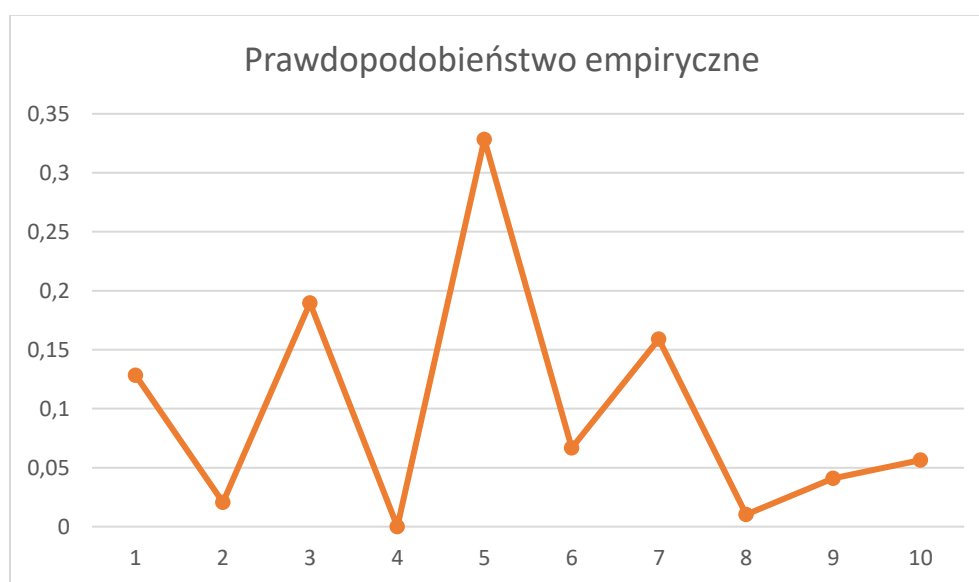
4.1 Wykresy



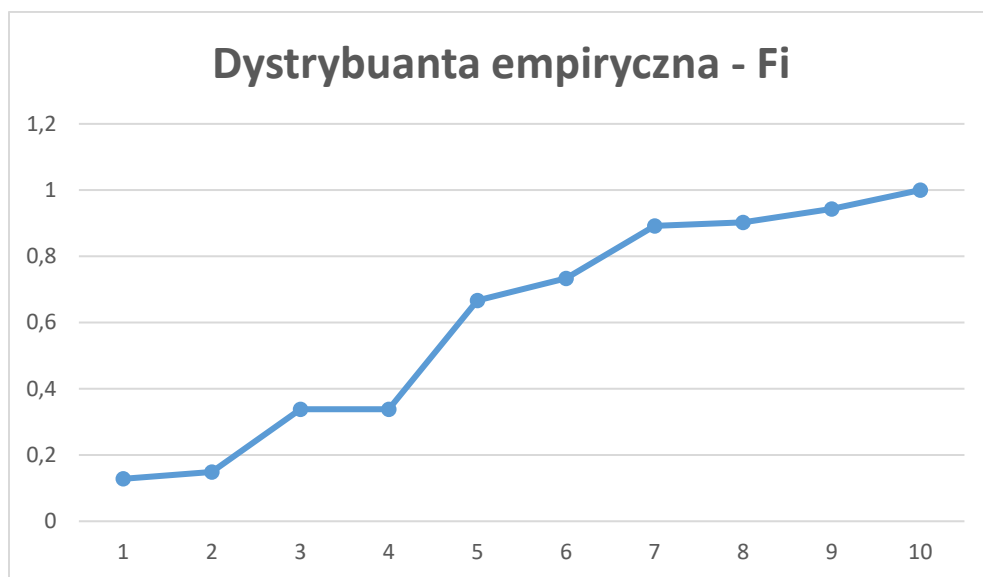
Rysunek 3 Wykres liczebności dla Zimy



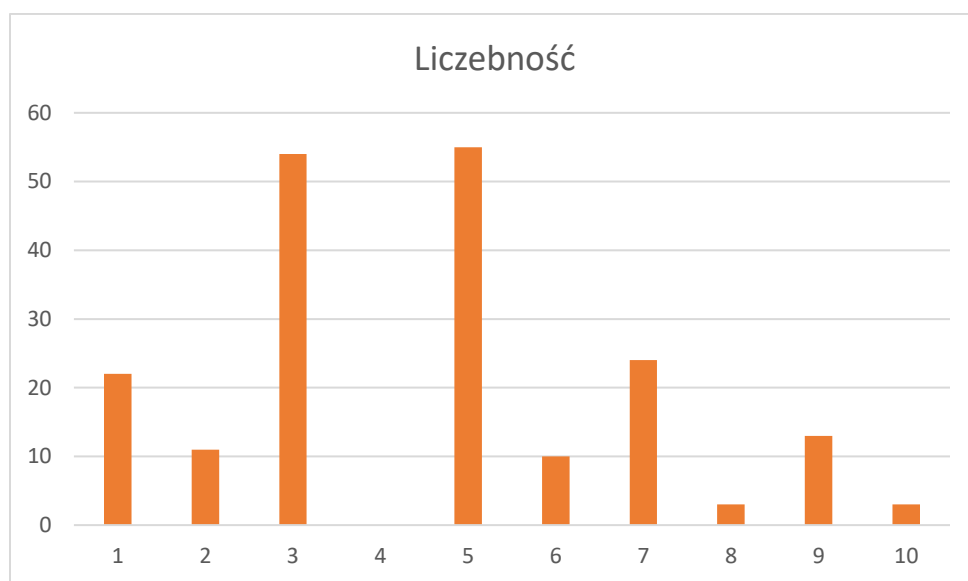
Rysunek 4 Wykres liczebności skumulowanej dla Zimy



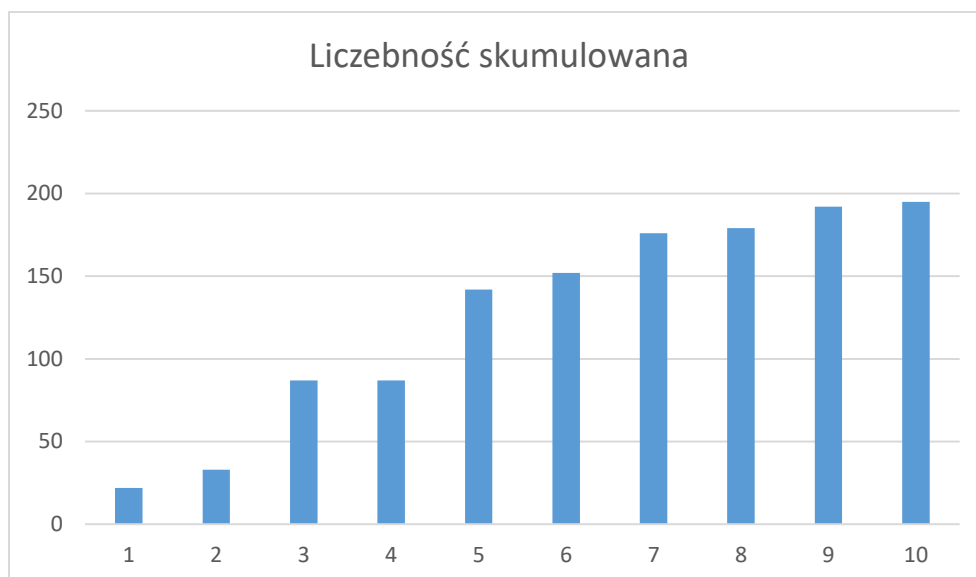
Rysunek 5 Wykres prawdopodobieństwa empirycznego dla Zimy



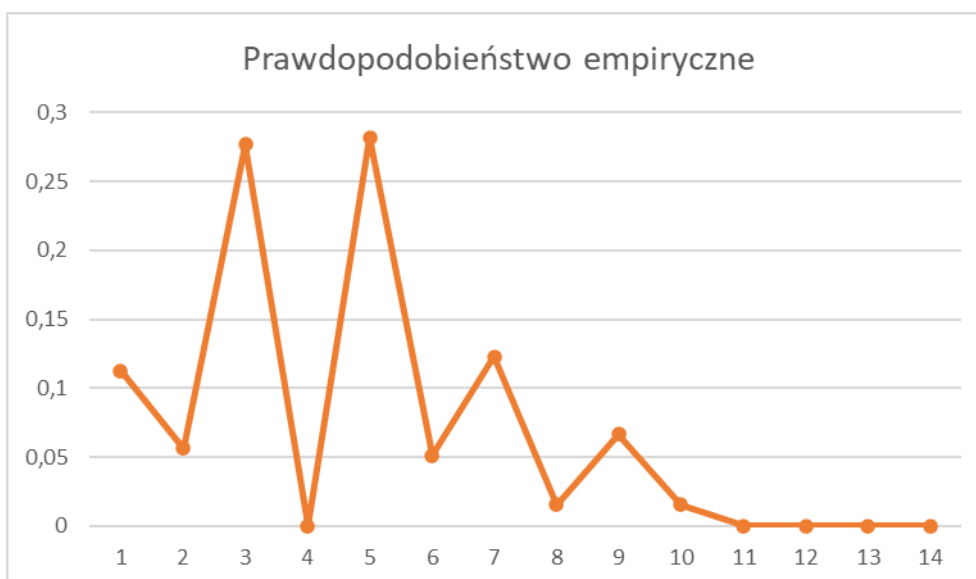
Rysunek 6 Wykres dystrybuanty empirycznej dla Zimy



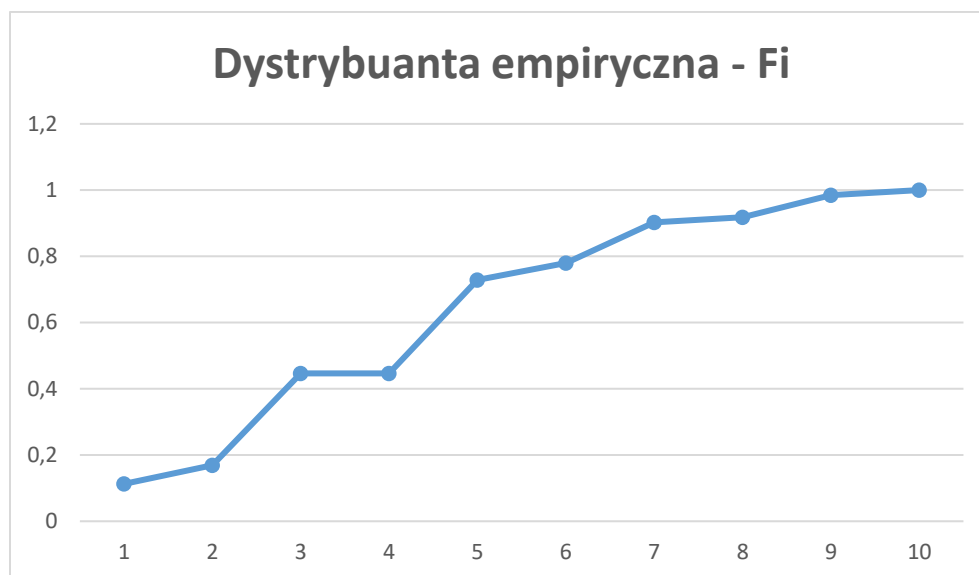
Rysunek 7 Wykres liczebności dla Lata



Rysunek 8 Wykres liczebności skumulowanej dla Lata



Rysunek 9 Wykres prawdopodobieństwa empirycznego dla Lata



Rysunek 10 Wykres dystrybuanty empirycznej dla Lata

4.2 Wnioski

4.2.1 Zbiór Zima

Histogram liczebności dla Lata można zauważyć, że przybiera kształt podobny do wykresu rozkładu normalnego z asymetrią prawostronną pomimo występujących brakujących danych w wykresie. Asymetria prawostronna oznacza, że wartości odstające w dużej mierze będą znajdować się po prawej stronie wykresu. Przedział wartości typowych jeden wynosi 74% jednakże przedział wartości typowych dwa wyniósł 94%. Przedział drugi wartości typowych jest wzorcowy, lecz należy zwrócić uwagę na przedział pierwszy który odbiega od wzorcowego i powinniśmy się zastanowić nad tym czy kierowcy zbyt mocno nie zawyżali poniesionych kosztów. Średnia wartość wyników jest wyższa od mediany.

4.2.2 Zbiór Lato

Histogram liczebności dla Lata różni się od rozkładu normalnego. Średnia wyszła nieznacznie mniejsza niż wartość dominanty więc wykres będzie posiadał asymetrię lewostronną. Asymetria lewostronna oznacza, że wartości odstające w dużej mierze będą znajdować się po lewej stronie wykresu. Przedział wartości typowych jeden wynosi 61% jednakże przedział wartości typowych dwa wyniósł 98%. Przedział drugi wartości typowych jest wzorcowy tak samo przedział pierwszy który nie odbiega zbyt mocno od wzorcowego. Średnia wartość wyników jest niższa od mediany.

4.2.3 Ogólne wnioski dla Lata i Zimy

Histogram liczebności skumulowanej bardzo dobrze pokazuje jaka jest różnica pomiędzy liczebnością poszczególnych klas, gdy widzimy spory skok wykresu słupkowego oznacza to, że dana klasa jest bardzo liczna. Im liczniejsza klasa tym skok większy, a dalej powinniśmy zauważyć stopniowe wypłaszczenie. Wykres ten bardzo przypomina dystrybuantę rozkładu normalnego. Myślę, że jest to dobre porównanie, ponieważ dystrybuantę definiujemy jako prawdopodobieństwo tego, że zmienna losowa X ma wartości mniejsze bądź równe x . Uważam, że powyższy wykres przedstawia to bardzo dobrze. Więc możemy przyjąć, że ostatnia kolumna to prawdopodobieństwo równe 1 więc prawdopodobieństwo uzyskania takiego kosztu lub mniejszego będzie blisko 1.

Dystrybuanta empiryczna pod względem kształtu przypomina wykres wartości skumulowane co oczywiście nie jest bez znaczenia. Jak zauważyłem histogram liczebności skumulowanej bardzo przypomina dystrybuantę rozkładu normalnego tak samo jest w tym przypadku. Dystrybuanta ta, to prawdopodobieństwo tego, że znajdziemy się w poszczególnej grupie, która osiągnęła określone koszty dojazdu do pracy. Myślę, że gdybyśmy poprzedni wykres znormalizowali do wartości 0-1 to otrzymalibyśmy podobne lub identyczne wyniki.

Ostatnim analizowanym wykresem jest prawdopodobieństwo empiryczne. Jak można zauważyć przypomina on bardzo wykres pierwszy dotyczący liczebności. Jednak wskazuje na prawdopodobieństwo tego, że nasza wartość uzyskana znajdzie się w wybranym miejscu. Dodatkowo np. przy asymetrii prawostronnej można wywnioskować, że znajdują się kierowcy, którzy ponadprzeciętnie duże koszty dojazdu. W modelowaniu systemów ten fakt można wykorzystać podczas pracy nad systemem, który pomagałby przy wyznaczaniu, ile powinien zarabiać pracownik pracujący i mieszkający w danym miejscu tak aby nasza oferta była dla niego atrakcyjna i oczywiście odpowiednio dopasowana do wymagań firmy.

5. Zmiana uporządkowania szeregu rozdzielczego

przedziały grupowania		liczebność	liczebność skumulowana	prawdopodobieństwo empiryczne	dystribucja empiryczna	środki przedziałów	obliczenia średniej	obliczenie wariancji	
L	p	ni	ni_sk	fi	Fi	xi	xi*ni	(xi-xsr)^2	ni*(xi-xsr)^2
108	128	25	25	0,128205	0,128205	118	2950	12774,8	319369,9
128	148	0	25	0	0,128205	138	0	8653,77	0
148	168	4	29	0,020513	0,148718	158	632	5332,744	21330,98
168	188	35	64	0,179487	0,328205	178	6230	2811,719	98410,15
188	208	2	66	0,010256	0,338462	198	396	1090,693	2181,386
208	228	33	99	0,169231	0,507692	218	7194	169,6673	5599,022
228	248	31	130	0,158974	0,666667	238	7378	48,64168	1507,892
248	268	13	143	0,066667	0,733333	258	3354	727,616	9459,009
268	288	1	144	0,005128	0,738462	278	278	2206,59	2206,59
288	308	30	174	0,153846	0,892308	298	8940	4485,565	134566,9
308	328	0	174	0	0,892308	318	0	7564,539	0

328	348	2	176	0,010256	0,902564	338	676	11443 ,51	22887 ,03
348	368	8	184	0,041026	0,94359	358	2864	16122 ,49	12897 9,9
368	388	11	195	0,05641	1	378	4158	21601 ,46	23761 6,1

Tabela 6 Szereg rozdzielnicy dla Zimy

h=	20
x01=	108
liczebność próby	195
średnia z szeregu	231,0256
odchylenie z szeregu	71,04043
procent typowych wartości wyznaczonych z odchylenia i średniej	76%
współczynnik zmienności	0,3075
procent wartości typowych ² wyznaczonych z odchylenia i średniej	100%
liczba wartości odstających po lewej stronie	0
liczba wartości odstających po prawej stronie	2

Tabela 7 Cechy szeregu dla Zimy

przedziały grupowani a		liczebnoś ć	liczebność skumulowan a	prawdopo dobieństw o empiryczn e	dystrybua nta empiryczn a	środki przed ziałów	obliczeni e średniej	obliczenie wariancji	
l	P	ni	ni_sk	fi	Fi	xi	xi*ni	(xi-xsr)^2	ni*(xi- xsr)^2
110	130	22	22	0,112821	0,112821	120	2640	12326,69	271187,2
130	150	0	22	0	0,112821	140	0	8285,667	0
150	170	11	33	0,05641	0,169231	160	1760	5044,642	55491,06
170	190	54	87	0,276923	0,446154	180	9720	2603,616	140595,3
190	210	0	87	0	0,446154	200	0	962,5904	0
210	230	32	119	0,164103	0,610256	220	7040	121,5648	3890,072
230	250	23	142	0,117949	0,728205	240	5520	80,53912	1852,4
250	270	9	151	0,046154	0,774359	260	2340	839,5135	7555,621
270	290	15	166	0,076923	0,851282	280	4200	2398,488	35977,32
290	310	10	176	0,051282	0,902564	300	3000	4757,462	47574,62
310	330	3	179	0,015385	0,917949	320	960	7916,437	23749,31
330	350	0	179	0	0,917949	340	0	11875,41	0
350	370	13	192	0,066667	0,984615	360	4680	16634,39	216247
370	390	3	195	0,015385	1	380	1140	22193,36	66580,08

Tabela 8 Szereg rozdzielczy dla Lata

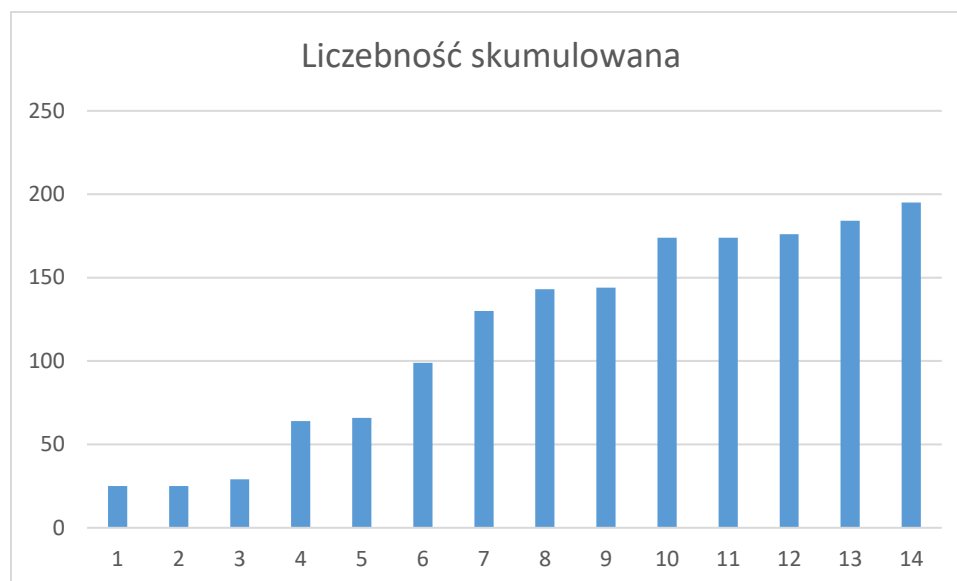
h=	17
x01=	110
liczebność próby	195
średnia z szeregu	220,5128
odchylenie z szeregu	66,82161
procent typowych wartości wyznaczonych z odchylenia i średniej	74%
współczynnik zmienności	0,303028
procent wartości typowych ² wyznaczonych z odchylenia i średniej	92%
liczba wartości odstających po lewej stronie	0
liczba wartości odstających po prawej stronie	16

Tabela 9 Cechy szeregu dla Lata

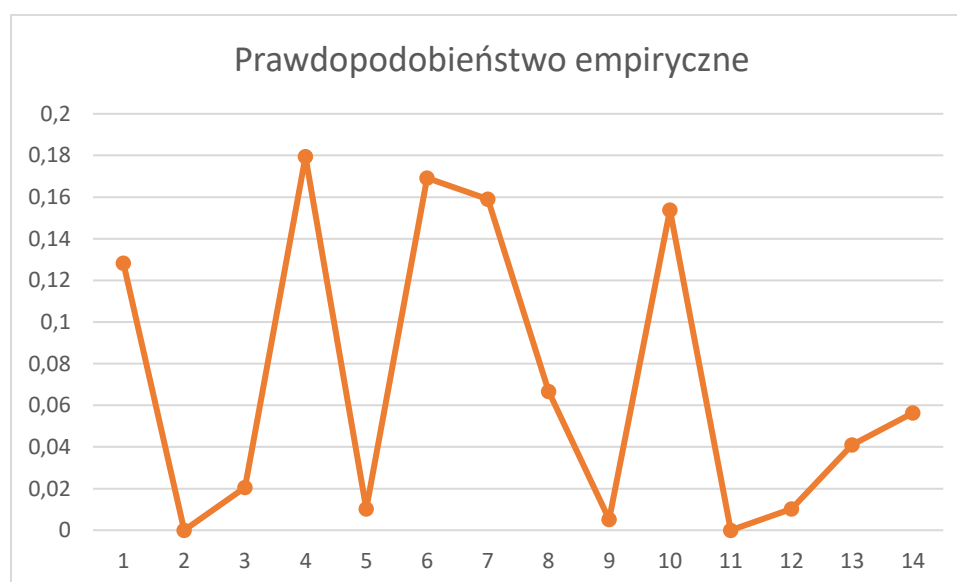
5.1 Wykresy



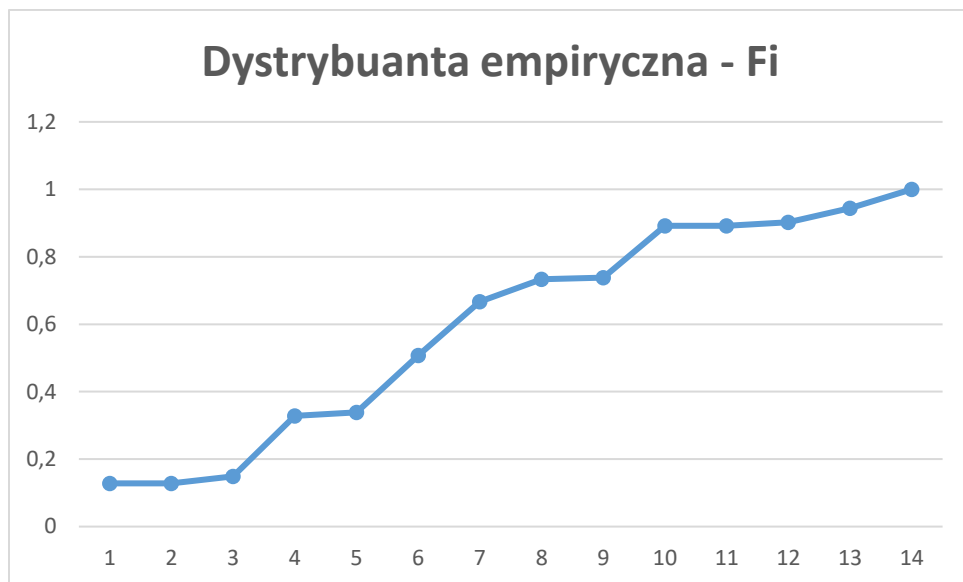
Rysunek 11 Wykres liczebność dla Zimy



Rysunek 12 Wykres liczebności skumulowanej dla Zimy



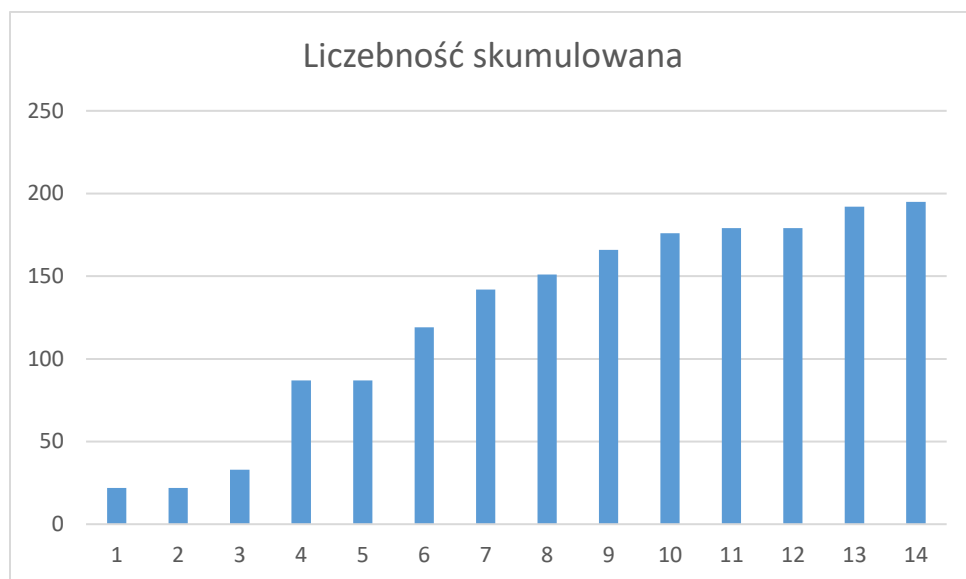
Rysunek 13 Wykres prawdopodobieństwa empirycznego dla Zimy



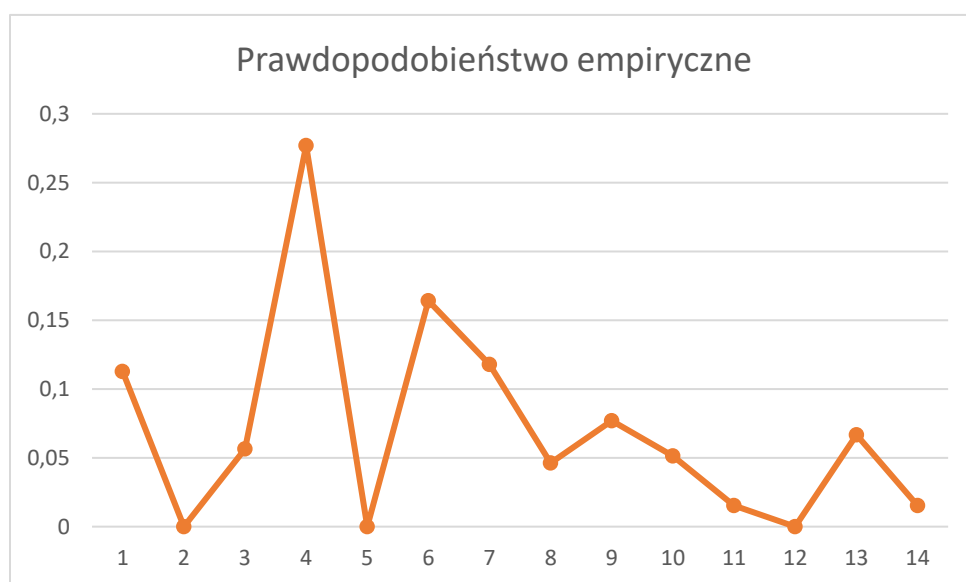
Rysunek 14 Wykres dystrybuanty empirycznej dla Zimy



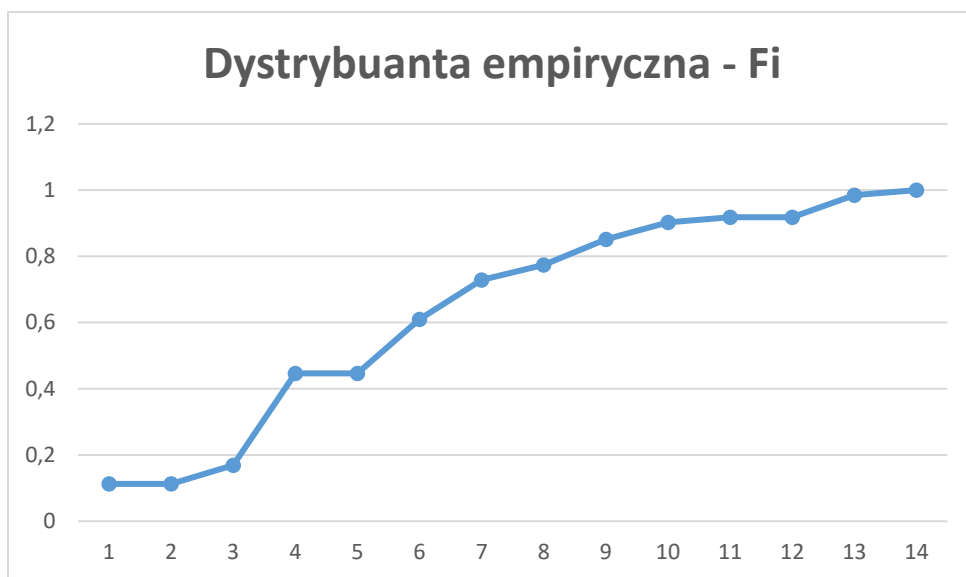
Rysunek 15 Wykres liczebności dla Lata



Rysunek 16 Wykres liczebności skumulowanej dla Lata



Rysunek 17 Wykres prawdopodobieństwa empirycznego dla Lata



Rysunek 18 Wykres dystrybucyjny empirycznej dla Lata.

5.2 Porównanie wyników z punktem 4

W tym punkcie zająłem się manipulacją wcześniej otrzymanego szeregu rozdzielczego manipulując parametrami takimi jak: liczba klas, wartość x_{01} oraz różnicą pomiędzy tymi klasami. Wyniku manipulacji dla grupy Zima otrzymałem wykres, który nie przypominał wykresu rozkładu normalnego. Inaczej jest dla zbioru Lato tam zmiany przyniosły poprawę w jakości wykresu dane w sposób widoczny przypominają rozkład normalny pozostałe wykresy również uległy poprawie. Natomiast dane liczbowe wraz z pogorszeniem jakości wykresu również uległy zmianie. Typowe 1 dla Zimy wyniosło 76%, a typowe 2 100 % co świadczy, że dane nie odzwierciedlają wykresu rozkładu normalnego. Dane w kolumnie Lato uległy zmianie w typowe 1 jest to pogorszenie, lecz typowe dwa jest współmiernie dobrze do wcześniejszego wykresu jednak sama wizualizacja poprawiła się dzięki czemu wykres jest czytelniejszy.

5.3 Wnioski

Podsumowując zmiana parametrów i uporządkowania szeregu rozdzielczego może przynieść korzyści w postaci np. bardziej dopasowanego wykresu do określonego wzorca, lecz należy manipulować danymi ostrożnie w przeciwnym wypadku uzyskamy pogorszenie jakości. Należy pamiętać, że nie zawsze uda się osiągnąć satysfakcjonujący wynik, a w dużej mierze wynik uzależniony jest od danych, którymi dysponujemy. W przypadku grupy Zimowej wykres pogorszył się, a także dane wyniki uległy współmiernemu pogorszeniu co pokazuje, że odpowiedni dobór parametrów jest kluczowy. W zależności od ustawionych parametrów x_{01} , h i ilości klas możemy sprawić, że ten sam wykres raz może przypominać rozkład normalny

i będzie dla nas wiarygodny. W drugim przypadku może być dla nas kompletnie bezużyteczny. Wiadomo, że oko ludzkie jest zawodne warto, dlatego sprawdzać również dane liczbowe które powinny być dla nas ważnym wyznacznikiem prawidłowości w doborze odpowiednich parametrów.

6. Weryfikowanie zgodności próby z rozkładem teoretycznym

6.1 Obliczenia Zima chi-kwadrat

	lewy	prawy	ni	F0(x0i)	F0(x1i)	F0(x1i)- F0(x0i)	npi	(npi-ni)2	(npi-ni)^2/npi
1	108,5	128,5	25	0,037919	0,068705	0,030786	6,003214	360,8779	60,11412
2	128,5	188,5	39	0,068705	0,268966	0,200262	39,05102	0,002603	6,67E-05
3	188,5	228,5	35	0,268966	0,485549	0,216582	42,23356	52,32434	1,238928
4	228,5	268,5	44	0,485549	0,7066	0,221051	43,10494	0,801127	0,018585
5	268,5	308,5	31	0,7066	0,869321	0,162722	31,7307	0,533921	0,016827
6	308,5	388,5	21	0,869321	0,988773	0,119452	23,29315	5,258517	0,225754

Tabela 10 Wyniki obliczeń rozkład chi-kwadrat-Zima

6.1.1 Sformułowanie hipotez dla rozkładu normalnego – Zima

Średnia = 231.

Odchylenie = 69.

H0: F(x) ma rozkład normalny z parametrami 231 i 69.

Hipoteza H1: F(x) nie ma rozkładu normalnego z parametrami 231 i 69.

6.1.2 Opis testu alfa = 0.01

alfa = 0.01,

wartość statystyki testowej = 61.61,

wartość krytyczna = 11,34.

6.1.3 Opis testu alfa = 0.05

alfa = 0.05,

wartość statystyki testowej = 61.61,

wartość krytyczna = 7.81.

6.1.4 Wnioski

Wartość statystyki testowej przy obu wartościach alfa jest większa od wartości krytycznej w związku z tym odrzucamy Hipotezę H0 na rzecz Hipotezy H1.

6.2 Obliczenia Lato chi-kwadrat

	lewy	prawy	ni	$F_0(x_{0i})$	$F_0(x_{1i})$	$F_0(x_{1i}) - F_0(x_{0i})$	n_{pi}	$(n_{pi} - n_i)^2$	$(n_{pi} - n_i)^2 / n_{pi}$
1	108,5	165,5	33	0,048296	0,204592	0,156296	30,47779	6,361529	0,208727
2	165,5	184,5	50	0,204592	0,292337	0,087745	17,1102	1081,739	63,22186
3	184,5	241,5	48	0,292337	0,613979	0,321643	62,72032	216,6879	3,454827
4	241,5	260,5	20	0,613979	0,71514	0,101161	19,72638	0,074866	0,003795
5	260,5	298,5	25	0,71514	0,869914	0,154773	30,18078	26,84051	0,889325
6	298,5	374,5	19	0,869914	0,987488	0,117574	22,92691	15,42062	0,672599

Tabela 11 Wyniki obliczeń rozkład chi-kwadrat-Lato

6.2.1 Sformułowanie hipotez dla rozkładu normalnego – Lato

Średnia = 222.

Odchylenie = 68.

H_0 : $F(x)$ ma rozkład normalny z parametrami 222 i 68.

Hipoteza H_1 : $F(x)$ nie ma rozkładu normalnego z parametrami 222 i 68.

6.2.2 Opis testu alfa = 0.01

alfa = 0.01,

wartość statystyki testowej = 68.45,

wartość krytyczna = 11,34.

6.2.3 Opis testu alfa = 0.05

alfa = 0.05,

wartość statystyki testowej = 68.73,

wartość krytyczna = 7.81.

6.2.4 Wnioski

Wartość statystyki testowej przy obu wartościach alfa jest większa od wartości krytycznej w związku z tym odrzucamy Hipotezę H_0 na rzecz Hipotezy H_1 .

6.3 Obliczenia Zima rozkład Kołmogorowa

	lewy	prawy	ni	ni skumul.	F _n (x)	x _i	F ₀ (x _i)	F _n (x)- F ₀ (x _i)
1	106,5	129,5	25	25	0,128205	118	0,050744	0,077461
2	129,5	152,5	0	25	0,128205	141	0,096058	0,032148
3	152,5	175,5	4	29	0,148718	164	0,165771	0,017053
4	175,5	198,5	37	66	0,338462	187	0,261841	0,076621
5	198,5	221,5	0	66	0,338462	210	0,380431	0,04197
6	221,5	244,5	52	118	0,605128	233	0,511562	0,093566
7	244,5	267,5	25	143	0,733333	256	0,641443	0,09189
8	267,5	290,5	13	156	0,8	279	0,756677	0,043323
9	290,5	313,5	18	174	0,892308	302	0,848257	0,044051
10	313,5	336,5	2	176	0,902564	325	0,913451	0,010887
11	336,5	359,5	0	176	0,902564	348	0,955024	0,05246
12	359,5	382,5	19	195	1	371	0,97877	0,02123

Tabela 12 Wyniki obliczeń rozkład Kołmogorowa Zima

6.3.1 Sformułowanie hipotez dla rozkładu Kołmogorowa – Zima

Średnia = 231.

Odchylenie = 69.

H₀: F(x) ma rozkład normalny z parametrami 231 i 69.

Hipoteza H₁: F(x) nie ma rozkładu normalnego z parametrami 231 i 69.

6.3.2 Opis testu alfa = 0.01

alfa = 0.01,

wartość statystyki testowej = 1.31,

wartość krytyczna = 1.63.

6.3.3 Opis testu alfa = 0.05

alfa = 0.05,

wartość statystyki testowej = 1.31,

wartość krytyczna = 1.36.

6.3.4 Wnioski

Wartość statystyki testowej dla testu Kołmogorowa jest niższa niż wartość krytyczna w obu wartościach alfa więc nie ma podstaw do odrzucenia Hipotezy H₀ na rzecz H₁.

6.4 Obliczenia Lato rozkład Kołmogorowa

	lewy	prawy	ni	ni skumul.	$F_n(x)$	x_i	$F_0(x_i)$	$ F_n(x)-F_0(x_i) $
1	108,5	127,5	22	22	0,112821	118	0,064923	0,047897
2	127,5	146,5	0	22	0,112821	137	0,108361	0,00446
3	146,5	165,5	11	33	0,169231	156	0,169566	0,000335
4	165,5	184,5	50	83	0,425641	175	0,249371	0,17627
5	184,5	203,5	4	87	0,446154	194	0,345662	0,100492
6	203,5	222,5	0	87	0,446154	213	0,453174	0,00702
7	222,5	241,5	44	131	0,671795	232	0,564254	0,10754
8	241,5	260,5	20	151	0,774359	251	0,670457	0,103902
9	260,5	279,5	1	152	0,779487	270	0,764419	0,015069
10	279,5	298,5	24	176	0,902564	289	0,841345	0,061219
11	298,5	317,5	0	176	0,902564	308	0,899624	0,00294
12	317,5	336,5	3	179	0,917949	327	0,940481	0,022532
13	336,5	355,5	0	179	0,917949	346	0,966986	0,049037
14	355,5	374,5	13	192	0,984615	365	0,982898	0,001718
15	374,5	393,5	3	195	1	384	0,991736	0,008264

Tabela 13 Wyniki obliczeń rozkład Kołmogorowa Lato

6.4.1 Sformułowanie hipotez dla rozkładu Kołmogorowa – Lato

Średnia = 222.

Odchylenie = 68.

H_0 : $F(x)$ ma rozkład normalny z parametrami 222 i 68.

Hipoteza H_1 : $F(x)$ nie ma rozkładu normalnego z parametrami 222 i 68.

6.4.2 dla alfy = 0.01

alfa = 0.01,

wartość statystyki testowej = 2,461477554,

wartość krytyczna = 1.63.

6.4.3 dla $\alpha = 0.05$

$\alpha = 0.05$,

wartość statystyki testowej = 2,461477554,

wartość krytyczna = 1.36.

6.4.4 Wyniki testów

Wartość statystyki testowej przy obu wartościach α jest większa od wartości krytycznej w związku z tym odrzucamy Hipotezę H_0 na rzecz Hipotezy H_1 .

7. Porównanie wyników punktu 6 z punktem 3

Porównując wyniki obu punktów należy pod uwagę wziąć z punktu pierwszego dwie wartości:

	Zima	Lato
procent wartości typowych 1	74,35897436	73,84615385
procent wartości typowych 2	98,97435897	91,79487179

Warto też spojrzeć procent wartości kwartylowych:

	Zima	Lato
procent typowych wartości kwartylowych 1	58,97435897	34,87179487
procent typowych wartości kwartylowych 2	87,17948718	87,69230769

Procent wartości typowych może nieidealnie plasuje się, jeśli chodzi o wzorcowe wartości rozkładu normalnego które mają wynik 68% i 95% lecz nie odbiegają od nich zbyt mocno.

Wynik średniej i pozostałych cech statystycznych jest taki sam jak w punkcie 3. Jednak przeprowadzony test w punkcie 6 pokazuje, że dane, mimo że procentowo wyglądają bardzo dobrze to jednak nie są zgodne z rozkładem normalnym. Nawet jeśli weźmiemy dane i w których liczba wartości nie jest duża, a ich występowanie pokrywa się z regułą 3 sigma to na wykresach można było to dostrzec, że obie grupy nie tworzyły wykresu normalnego, choć przyjmowały taki wygląd, lecz zbyt mało danych by można było stwierdzić, że są one z nim zgodne. Dlatego uważam, że obliczenia są prawidłowe i na podstawie ilości oraz różnorodności danych nie można stwierdzić, że są one zgodne z rozkładem normalnym. Na szczególną uwagę zasługuje test Kołmogorowa, który dla Zimy wykazał zgodność, lecz bazując na obliczeniach z pozostałych punktów i na wykresach nie ma wystarczającej ilości danych, aby orzec, że dane są zgodne z rozkładem normalnym.

