



Fundusze Europejskie
dla Rozwoju Społecznego



Rzeczpospolita
Polska

Dofinansowane przez
Unię Europejską



Politechnika Świętokrzyska
Wydział Zarządzania i Modelowania Komputerowego

Kierunek studiów:
Inżynieria danych

Katarzyna Brzozowska-Rup

Materiały dydaktyczne do przedmiotu

STATYSTYKA

opracowane w ramach realizacji Projektu
**„Dostosowanie kształcenia
w Politechnice Świętokrzyskiej do potrzeb
współczesnej gospodarki”**
FERS.01.05-IP.08-0234/23

Kielce, 2025





Spis treści

1.	Wstęp	3
2.	Statystyka opisowa.....	3
2.1.	Podstawowe pojęcia.....	3
2.2.	Rozkład empiryczny cechy, jego opis statystyczny i prezentacja graficzna.....	5
2.3.	Miary statystyczne.....	13
3.	Zmienna losowa jednowymiarowa	23
3.1.	Pojęcie zmiennej losowej oraz jej charakterystyki	23
3.2.	Wybrane typy rozkładów.....	31
4.	Centralne twierdzenie graniczne Lindeberga-Levy'ego.....	37
5.	Elementy wnioskowania statystycznego.....	46
5.1.	Istota wnioskowania statystycznego.....	46
5.2.	Estymacja przedziałowa.....	45
5.3.	Weryfikacja wybranych hipotez statystycznych.....	56
6.	Literatura.....	63
7.	Wzory i tablice.....	64



Katarzyna Binaszka-Ryś

Materiały dydaktyczne objęte licencją Creative Commons BY 4.0.

Licencja dostępna pod adresem: <https://creativecommons.org/licenses/by/4.0>



1. Wstęp

Prezentowany materiał jest kontynuacją opracowania „Rachunek prawdopodobieństwa” opracowanego w ramach realizacji Projektu „Dostosowanie kształcenia w Politechnice Świętokrzyskiej do potrzeb współczesnej gospodarki”. Rachunek prawdopodobieństwa i statystyka są szczególnymi działami matematyki, które dostarczają narzędzi do badania i opisu zjawisk, uwzględniając ich losowy charakter. Rozwój technologiczny powoduje, że każdego dnia tworzona jest ogromna ilość danych. Tempo wzrostu ich ilości rośnie z każdym rokiem zwiększając zapotrzebowanie na usługi statystyczne w różnorodnych dziedzinach działalności (zarówno praktyczne, jak i teoretyczne). Nieodzownym narzędziem analizy danych jest statystyka, która stanowi fundament badań naukowych, ekonomicznych, społecznych, przyrodniczych i technicznych. Jej metody pozwalają nie tylko na opis rzeczywistości, identyfikację prawidłowości, ale także na przewidywanie przyszłych zjawisk oraz podejmowanie optymalnych decyzji. Celem prezentowanego materiału jest wsparcie studentów w osiągnięciu intuicji statystycznej oraz przyswojeniu materiału teoretycznego, niezbędnego do umiejętnego korzystania z metod analizy statystycznej z zastosowaniem arkusza kalkulacyjnego MS Excel.

Materiał stanowi uzupełnienie wykładu oraz zajęć praktycznych (laboratoryjnych i projektowych) z przedmiotu Statystyka. Obejmuje zagadnienia z zakresu statystyki opisowej oraz wnioskowania statystycznego. Dedykowany jest studentom pierwszego roku kierunku inżynieria danych. Treści zostały dobrane tak, aby zilustrować kluczowe koncepcje teoretyczne oraz ich zastosowania w rzeczywistych sytuacjach.

2. Statystyka opisowa

2.1. Podstawowe pojęcia

Słowo *statystyka* ma obecnie wiele znaczeń. Jako dyscyplina naukowa zajmuje się metodami ilościowymi badania (zbierania, opracowywania i analizy danych) zjawisk masowych opisując je najczęściej za pomocą liczb w celu ustalenia pewnych regularności i/lub tendencji charakteryzujących ich kształtowanie się.



Dlaczego warto uczyć się statystyki?

Świat nie działa w deterministyczny sposób. Pogoda, giełda, zachowania ludzi, ... itp. rządzą się pewnymi prawami, a statystyka dostarcza „narzędzi”, które umożliwiają odkrywanie ogólnych wzorców i prawidłowości, rozwijanie umiejętności krytycznego myślenia. Można śmiało stwierdzić, że statystyka to tarcza przed manipulacją. Kiedy ktoś mówi „80% ludzi...”, to dzięki statystyce możesz zapytać: *ile osób było w próbie?, jaki był rozkład?, czy to istotne statystycznie, czy tylko medialnie?*

Etapy badania statystycznego:

1. przygotowanie badania – określenie **jednostki statystycznej¹** oraz **zbiorowości statystycznej (populacji generalnej/populacji statystycznej)**, tj. zbioru wszystkich elementów objętych badaniem ze względu na określoną cechę lub cechy. Należy pamiętać, że badana zbiorowość powinna być jednorodna.
W przypadku, gdy badania całej populacji generalnej nie są możliwe (np. ze względu na koszt lub czasochłonność) wówczas bada się podzbiór populacji zwany **próbą**. Próba powinna być losowa² i reprezentatywna³;
2. pozyskanie danych – inaczej obserwacja statystyczna. Obserwacji statystycznej podlegają wybrane właściwości statystyczne nazywane cechami statystycznymi. Cechy te mogą być **stałe** (wspólne dla wszystkich jednostek danej zbiorowości i nie podlegające badaniu, a jedynie decydujące o zaliczeniu jednostki do określonej zbiorowości) lub **zmienne**, tj. właściwości, które różnicują poszczególne jednostki statystyczne i to one głównie są celem obserwacji;
3. opracowanie i prezentacja materiału statystycznego w formie tabelarycznej lub graficznej;
4. opis lub wnioskowanie statystyczne.

¹ **Jednostki statystyczne**, to poszczególne elementy składowe badanej zbiorowości

² **Próba** jest **losowa**, jeżeli każda jednostka populacji generalnej ma takie samo prawdopodobieństwo znalezienia się w próbie i jeżeli istnieje możliwość ustalenia prawdopodobieństwa znalezienia się w próbie dla każdego zespołu elementów populacji generalnej.

³ Próba jest **reprezentatywna**, jeżeli jej struktura nie różni się od struktury zbiorowości generalnej (całej populacji). Każda próba ma określoną **liczebność (liczność)**. Jeżeli liczebność próby losowej jest $n \leq 30$ wówczas mówimy, że próba losowa jest mała, jeżeli liczebność próby losowej jest $n > 30$ wówczas mówimy, że próba losowa jest duża.



Szczegóły na temat realizacji badania statystycznego można znaleźć w pozycjach umieszczonych w literaturze opracowania.

Cechy statystyczne mogą mieć dwojaki charakter:

- ilościowy (**mierzalny**) – właściwości, które można zmierzyć i wyrazić za pomocą odpowiednich jednostek fizycznych (np. długość w kilometrach, centymetrach);
- jakościowy (**niemierzalny**) – zwykle określane słownie, poprzez opis (np. wykształcenie, typ szkoły, kolor oczu, płeć).

Cechy mierzalne dzieli się na:

- **skokowe (dyskretne)**, czyli takie, które przyjmują skończony lub przeliczalny zbiór wartości na danej skali liczbowej, najczęściej jest to zbiór liczb całkowitych dodatnich (np. liczba osób w grupie laboratoryjnej, liczba usterek, itp.);
- **ciągłe**, które mogą przyjąć każdą wartość z określonego przedziału liczbowego $[a, b]$.

2.2. Rozkład empiryczny cechy, jego opis statystyczny i prezentacja graficzna

W prezentowanym materiale przyjmujemy, że w wyniku pomiaru n obiektów otrzymujemy wartości cechy: $x_1, x_2, \dots, x_n$, które określają **empiryczny rozkład cechy**⁴ (będący podstawą analizy statystycznej).

Opis statystyczny sprowadza się do wyznaczenia pewnych liczbowych parametrów określających badany zbiór danych. Zbiór wyników obserwacji uporządkowanych według określonych cech (kryteriów) nazywamy **szeregiem statystycznym**. Na wykładzie omówiliśmy szczegółowo ich rodzaje, tj. szereg szczegółowy (wyliczający) oraz **szereg rozdzielczy** punktowe i przedziałowe.

⁴ Określenie empirycznego rozkładu cechy polega na przyporządkowaniu uszeregowanym rosnąco wartościom, przyjmowanym przez tę cechę, odpowiednio zdefiniowanych częstości ich występowania.



Przypomnijmy

- **Szereg rozdzielczy punktowy** jest stosowany do klasyfikacji danych statystycznych w przypadku, gdy mamy do czynienia z niewielką liczbą wartości cechy. Szeregi rozdzielcze punktowe budowane są dla cechy skokowej;
- **Szereg rozdzielczy przedziałowy** – jest stosowany do klasyfikacji danych statystycznych w przypadku, gdy mamy do czynienia z dużą liczbą wartości cechy ($k > 12$) oraz dużą próbą ($n > 30$). Szeregi rozdzielcze przedziałowe budowane są zarówno dla cechy ciągłej i skokowej.

Przykład 2.1 (dane BAEL)

Proces starzenia się społeczeństwa, obserwowany od lat, stanowi jeden z kluczowych trendów demograficznych. Porównanie liczby urodzeń żywych według kolejności urodzeń⁵ i struktury wieku pozwala lepiej zrozumieć dynamikę zmian demograficznych.

W tabelach 2.1 i 2.2 podano przykłady rozkładów empirycznych cechy dyskretnej i ciągłej. Dane prezentują sytuację w województwie świętokrzyskim w roku 2007 i 2024 ich analizą zajmiemy się w kolejnych rozdziałach opracowania.

⁵ „Kolejność urodzenia dziecka - określa, którym z kolei dzieckiem urodzonym przez matkę jest noworodek zarejestrowany w danym roku, przy uwzględnieniu wszystkich poprzednich urodzeń noworodków żywych i martwych. (...) Poziom urodzeń według poszczególnych, charakteryzujących to zjawisko cech oraz natężenie urodzeń są odzwierciedleniem (w dłuższych okresach) rozwoju społecznego kraju będącego bezpośrednio lub pośrednio wynikiem: postępu w rozwoju medycyny, poziomu funkcjonowania systemu ochrony zdrowia, a także zmian wyboru, jakiego dokonują kolejne pokolenia w zakresie rozwoju rodziny. Urodzenia są także składnikiem bilansu liczby i struktury ludności, elementem tablic trwania życia, a także stanowią podstawę opracowania analiz i prognoz demograficznych (szczególnie znaczenie w tym przypadku mają prawdopodobieństwa urodzeń dzieci poszczególnych kolejności oraz struktury kobiet według generacji i liczby urodzonych dzieci).” (Zeszyt metodologiczny Ruch naturalny. Bilanse ludności, 2018)

Tabela 2.1. Szereg rozdzielczy
punktowy

Urodzenia żywe według kolejności urodzenia dziecka	Liczba urodzeń	
	2007	2024
1	6 150	3176
2	3 963	1873
3	1 207	629
4	369	153
5	148	48
6 i dalsze	134	18

* W podziale według kolejności urodzenia nie uwzględniono urodzeń o nieustalonej kolejności urodzenia

Tabela 2.2. Szereg rozdzielczy
przedziałowy.

Wiek -zadany przedziałowo	Liczba ludności wg grup wieku	
	2007	2024
0-4	56 604	39 821
5-9	60 757	54 414
10-14	75 726	58 214
15-19	92 708	59 832
20-24	106 383	55 459
25-29	100 634	58 886
30-34	89 712	68 571
35-39	78 793	81 837
40-44	77 248	94 964
45-49	93 002	91 264
50-54	102 246	79 739
55-59	91 306	70 425
60-64	61 039	75 892
65-69	49 899	86 095
70-74	50 280	74 683
75-79	44 196	50 863
80-84	28 502	26 694
85 i więcej	16 515	30 338

Źródło: <https://bdl.stat.gov.pl/bdl/dane/podgrup/temat>

Zauważmy, że dane zostały zaprezentowane w taki sposób, że nie mamy informacji jaką wartość ma każda jednostka badanej zbiorowości (tj. nie dysponujemy danymi indywidualnymi), a przedstawiono je w postaci pogrupowanej, tj. poszczególnym wartością cechy x_i (tabela 2.1) lub przedziałom wartości cechy (tabela 2.2) przyporządkowano liczebności n_i .



W praktyce przy określaniu rozkładu empirycznego zamiast liczebności n_i stosuje się

(1) **częstości**⁶ określone jako

$$p_i = \frac{n_i}{n} \quad (2.1)$$

gdzie $n = \sum_{i=1}^k n_i$, k – liczba wartości/ przedziałów jakie przyjmuje cecha, oznaczające udział jednostek o określonej wartości cechy x_i , lub wartości z określonego przedziału w ogólnej liczebności badanej zbiorowości

lub

(2) **częstości skumulowane** – ważne w statystyce pojęcie **dystrybuanty empirycznej**

$$F(x) = \begin{cases} 0 & \text{dla } x < x_1 \\ \sum_{j=1}^i p_j & \text{dla } x_i \leq x < x_{i+1} \\ 1 & \text{dla } x \geq x_k \end{cases} \quad (2.2)$$

Przykład 2.2

Na rys. 2.1 przedstawiono rozkład empiryczny za pomocą częstości oraz dystrybuanty dla liczby urodzeń żywych według kolejności urodzenia dziecka w 2024 roku (dane z przykładu 2.1).

[Tekst alternatywny. Zrzut ekranu przedstawiający rozkład szeregu rozdzielczego punktowego zadany przez częstość oraz dystrybuantę. Przedstawiono formułę wykorzystywaną do wyznaczania dystrybuanty]

	D	E	F	G	H
31		Urodzenia żywe wg kolejności	Liczba urodzeń		
32		x_i	n_i	p_i	$F(x_i)$
33		1	3176	0,539	0,539
34		2	1873	0,318	=H33+G34
35		3	629	0,107	0,963
36		4	153	0,026	0,989
37		5	48	0,008	0,997
38		6 i dalsze	18	0,003	1,000

Rys. 2.1. Rozkład danych z przykładu 2.2.

⁶ W literaturze wielkości p_i nazywane są również częstościami względnymi w odróżnieniu od częstości absolutnych (liczebności n_i)



Na przykład, wartość $F(3)$ zaprezentowana w tabeli oznacza, iż w roku 2024 odsetek rodzin, w których liczba urodzonych dzieci nie przekroczyła trzech, wynosił 96,3% ogółu analizowanych przypadków. Natomiast jedynie 3 na 1000 urodzeń dotyczyły szóstego lub kolejnego dziecka – co stanowi zaledwie 0,3% – co jednoznacznie wskazuje na zanikanie rodzin wielodzietnych w analizowanej populacji.

Graficzna prezentacja rozkładu empirycznego

Najpopularniejszymi formami graficznej prezentacji szeregu rozdzielczego są: histogram, wielobok liczebności oraz krzywa liczebności.

Histogram (inaczej diagram kolumnowy) to wykres zbudowany z prostokątów, w którym podstawa każdego prostokąta oznacza wartość badanej zmiennej, a jego wysokość — liczebność lub częstość względną. Powierzchnie prostokątów są proporcjonalne do liczebności klas. Z technicznego punktu widzenia (wykonania histogramu w Excelu) nie ma znaczenia, czy badana cecha ma charakter ciągły czy skokowy.

Wielobok liczebności/częstości to linia łamana, której odcinki łączą punkty współrzędnych $(x_i; n_i)$ lub $(x_i; p_i)$, przy czym w przypadku cechy ciągłej wartości x_i należy rozumieć jako środki przedziałów.

Krzywa liczebności jest granicą, do której dąży histogram, przy założeniu, że rozpiętość przedziałów zmniejsza się, a jednocześnie wzrasta liczba obserwacji.

Sposób wykonania powyższych wykresów w arkuszu kalkulacyjnym przedstawimy w poniższych przykładach

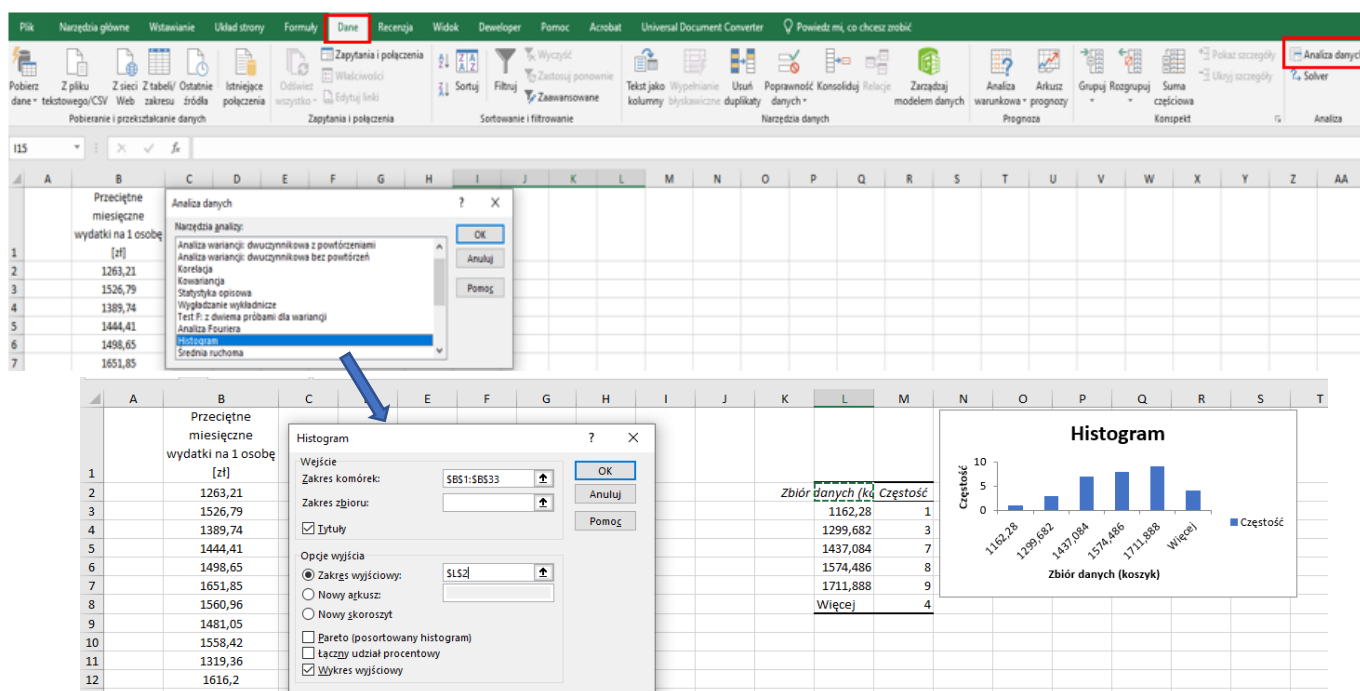
Przykład 2.3 – tworzenie histogramu za pomocą funkcji dostępnej w Analizie danych

W przypadku, gdy dysponujemy danymi w postaci szeregu szczegółowego wygodną formą prezentacji rozkładu danych jest funkcja **Histogram** dostępna w MS Excel pozycji **Dane -> Analiza danych**.

Wykres wykonamy dla danych pozyskanych z Banku Danych Lokalnych, dotyczących przeciętnych wydatków na 1 osobę w województwach w Polsce w latach 2022 i 2023: 1263,21; 1526,79; 1389,74; 1444,41; 1498,65; 1651,85; 1560,96; 1481,05; 1558,42; 1319,36; 1616,2; 1266,97; 1399,56; 1162,28; 1285,05; 1672,67; 1435,47; 1702,49;

1600,89; 1597,97; 1619,92; 1849,29; 1699,17; 1631,42; 1776,65; 1340,05; 1774,58; 1457,42; 1558,16; 1320,27; 1324,01; 1849,11. Przy czym czas w prezentowanym badaniu nie odgrywa istotnej roli.

[Tekst alternatywny. Zrzut ekranu przedstawiający instrukcję wykonania Histogramu za pomocą opcji dostępnej w pozycji Dane -> Analiza danych -> Histogram. W drugiej części rysunku przedstawiono również wynik w postaci rozkładu cechy ciągłej (tj. szeregu rozdzielczego przedziałowego) i jego prezentacji graficznej]

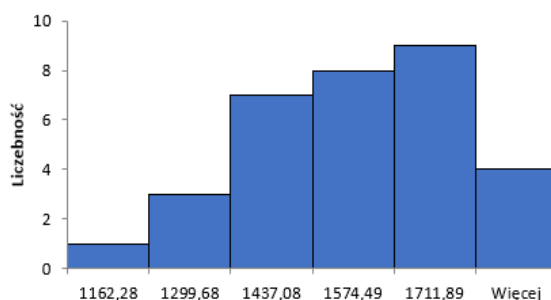


Rys.2.2. Instrukcja tworzenia histogramu (Analiza danych)

Analizowana cecha (przeciętne miesięczne wydatki na 1 osobę) ma charakter cechy ciągłej w związku z powyższym kolumny histogramu powinny się łączyć.

W tym celu należy pojedynczym kliknięciem na kolumnie wyselekcjonować opcję: **formatuj serię** danych następnie zmienić szerokość odstępu do zera. Zauważmy, że w wyniku powyższej procedury oś pionowa podpisana jest jako **Częstość**, podczas gdy na osi wyświetla się informacja o liczebności. W związku z powyższym wskazana jest zmiana opisu osi (jak na rys. 2.3).

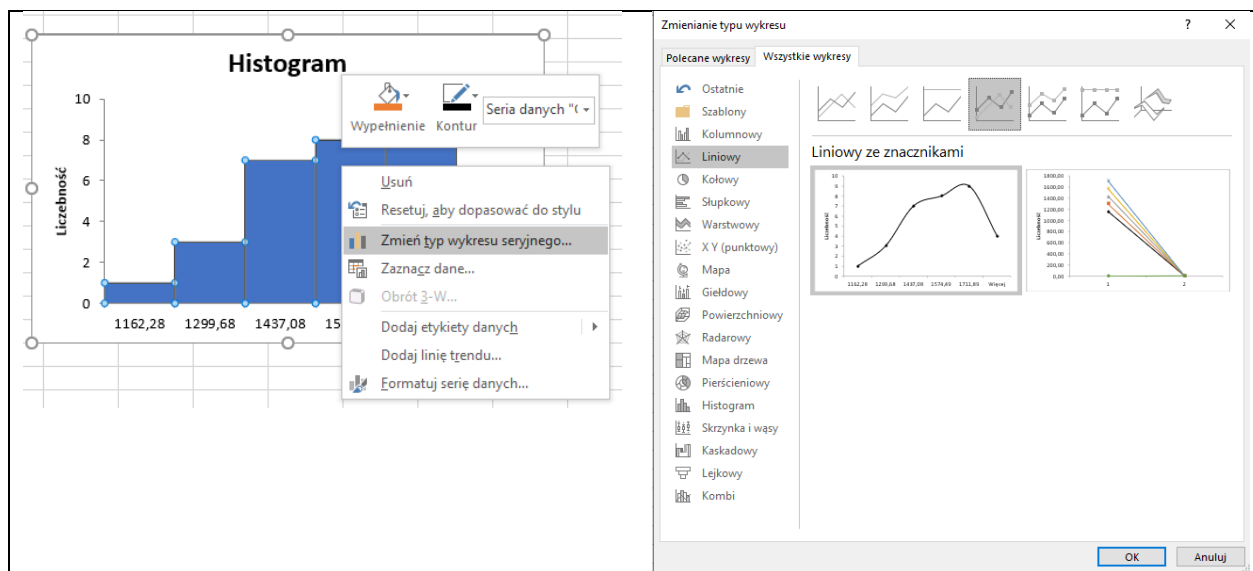
[Tekst alternatywny. Rysunek składa się z niebieskich kolumn (prostokątów) takiej samej szerokości, różnej wysokości równej liczebność]



Rys.2.3. Histogram liczebności – cecha ciągła

Aby wyznaczyć krzywą liczebności w obszarze kreślenia wykresu należy kliknąć prawym przyciskiem myszy na obszarze wykresu. Z rozwiniętego menu podręcznego wybrać polecenie Typ wykresu. Na liście typów wykresu wskazać typ liniowy.

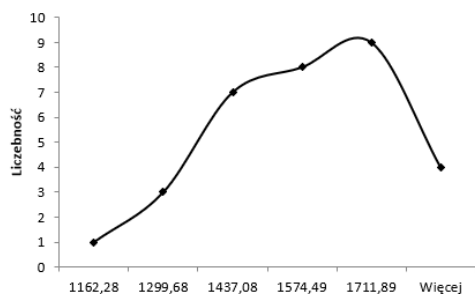
[Tekst alternatywny. Zrzut ekranu przedstawiający instrukcję Zmiany typu wykresu.]



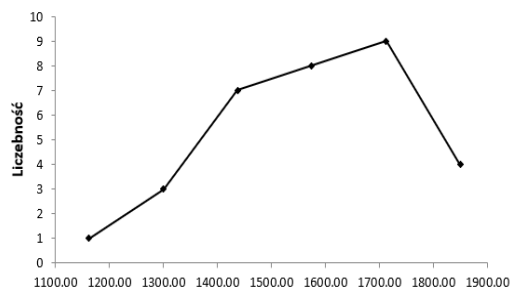
Rys. 2.4. Okno edycji wykresu



[Tekst alternatywny. Wykresy przedstawiają prezentację graficzną krzywej i wieloboku liczebności.]



Rys. 2.5. Krzywa liczebności

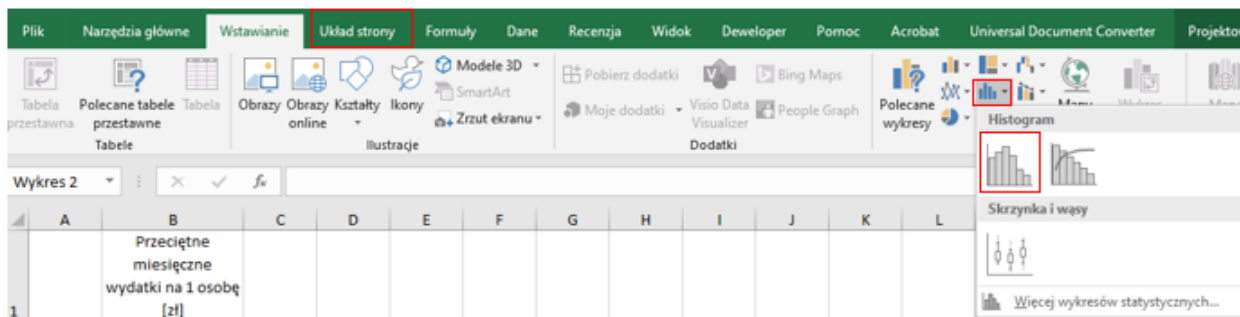


Rys. 2.6. Wielobok liczebności

Wielobok liczebności uzyskujemy w zasadzie tak samo - wybierając **wykres liniowy**, ale nie wygładzony.

Innym sposobem wykonania histogramu jest wykorzystanie pozycji **Wstaw -> Wstaw wykres statystyczny -> Histogram** dostępnej na pasku arkusza MS Excel.

[Tekst alternatywny. Zrzut ekranu przedstawiający sposób wykonania histogramu z wykorzystaniem pozycji **Wstaw -> Wstaw wykres statystyczny -> Histogram** dostępnej z menu głównego arkusza MS Excel.]

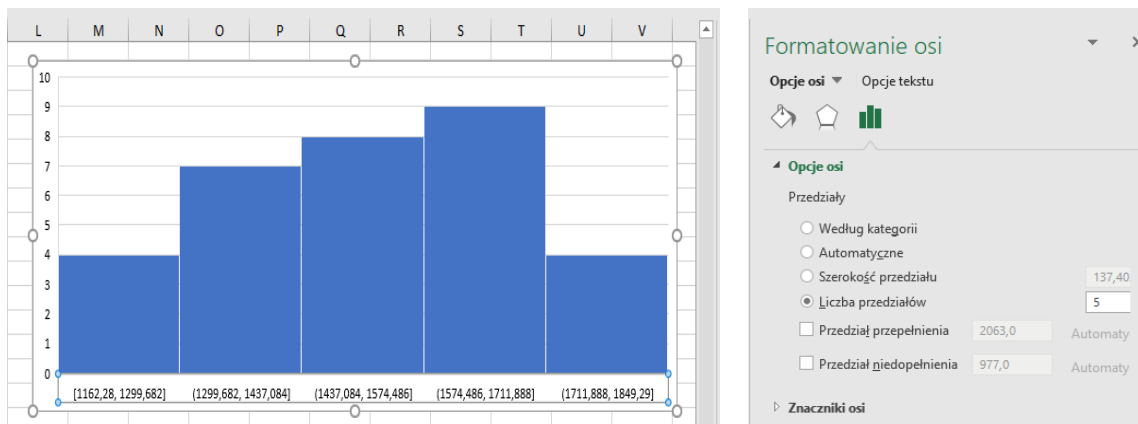


Rys.2.7. Okno edycji wykresu

Aby ustalić zadaną liczbą przedziałów histogramu należy kliknąć prawym przyciskiem myszy oś poziomą wykresu, wybrać polecenie **Formatuj oś**, a następnie wybierać pozycję **Opcje osi** i podać **liczbę przedziałów**.



[Tekst alternatywny. Zrzut ekranu przedstawiający instrukcję zmiany liczby przedziałów w Histogramie]



Rys.2.8. Okno edycji opcji wykresu

2.3. Miary statystyczne

Analiza danych statystycznych polega na przedstawieniu wyników badań za pomocą odpowiednich charakterystyk liczbowych, zwanych parametrami statystycznymi. **Parametry statystyczne** to liczby służące do syntetycznego opisu struktury populacji generalnej (całej zbiorowości statystycznej).

Parametry te dzielą się na:

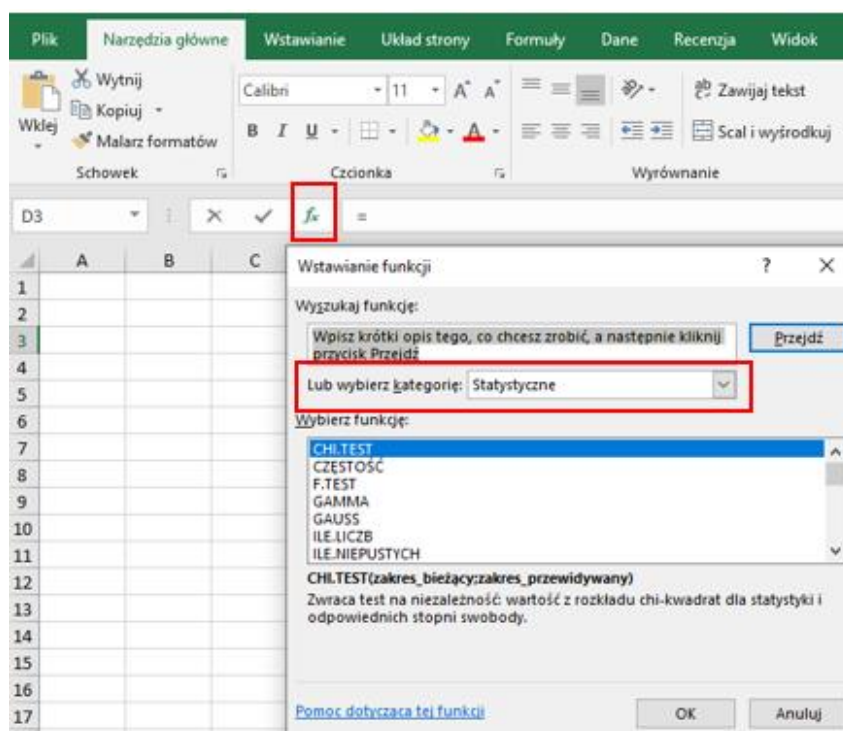
- miary położenia, tj. średnia, mediana, dominanta, kwantyle;
- miary zmienności (rozproszenia, dyspersji, zróżnicowania), tj. wariancja, odchylenie standardowe, współczynnik zmienności, rozstęp, rozstęp międzykwartylowy;
- miary asymetrii, tj. trzeci moment centralny, współczynniki asymetrii i skośności;
- miary koncentracji (skupienia), tj. czwarty moment centralny, kurtoza.



Obszerne objaśnienia dotyczące istoty, obliczania i interpretacji powyższych miar podane są na wykładzie, można je również znaleźć w podręcznikach podanych w literaturze tego opracowania.

W arkuszu kalkulacyjnym MS Excel obliczanie miar statystycznych dla szeregu szczegółowego sprowadza się do zastosowania odpowiedniej funkcji, natomiast dla szeregów rozdzielczych (danych pogrupowanych) **nie ma funkcji** bezpośrednio wyznaczających te miary. W celu ich obliczenia należy zastosować odpowiednie wzory – wzory te zostały zebrane na końcu tego opracowania w rozdziale 7.

[Tekst alternatywny. Zrzut ekranu. Okno dialogowe **Wstawianie funkcji**, kategoria statystyczne.]



Rys.2.9. Funkcje statystyczne w programie MS Excel

Łatwo dostępnym i intuicyjnym wsparciem w analizie i podsumowaniu podstawowych miar statystycznych danych prezentowanych w postaci szeregu szczegółowego jest narzędzie **Statystyka opisowa** dostępna w pozycji **Dane -> Analiza danych** arkusza kalkulacyjnego.



[Tekst alternatywny. Zrzut ekranu. Wywołanie i wyznaczenie podstawowych parametrów statystycznych za pomocą narzędzia Statystyka opisowa.]

Rok	urodzenia żywe (ogółem) woj. świętokrzyskie
2007	11971
2008	12736
2009	12741
2010	12445
2011	11600
2012	11510
2013	10720
2014	10756
2015	10579
2016	10610
2017	10915
2018	10721
2019	10220
2020	9488
2021	8717
2022	8157
2023	7238
2024	6472

urodzenia żywe (ogółem) woj. świętokrzyskie	
Średnia	10422
Błąd standardowy	424,26
Mediana	10720,5
Tryb	#N/D!
Odchylenie standardowe	1799,988
Wariancja próbeki	3239958
Kurtoza	0,102713
Skośność	-0,81
Zakres	6269
Minimum	6472
Maksimum	12741
Suma	187596
Licznik	18

Rys.2.10. Wywołanie i ustawienie parametrów opcji: Statystyka opisowa w MS Excel



Jedną z najczęściej stosowanych miar tendencji centralnej w statystyce i ekonometrii jest **średnia arytmetyczna**. Jej powszechność wynika zarówno z intuicyjnego charakteru, jak i z silnych podstaw matematycznych oraz praktycznego znaczenia w analizie ilościowej (reprezentuje typowy poziom analizowanej cechy, jest łatwa do obliczenia i zrozumienia, jest wyjściowym parametrem do wyznaczania innych bardziej skomplikowanych miar statystycznych). Są takie sytuacje, w których **średnia arytmetyczna nie odzwierciedla w sposób właściwy przeciętnej wartości**. Do takich przypadków należy **analiza zjawisk w ujęciu wskaźników oraz dynamiki zjawisk**.

Średnia harmoniczna jest klasyczną miarą stosowaną do analizy zjawisk w ujęciu wskaźników, tj., gdy wartości badanej cechy x_i podane są w przeliczeniu na stałą jednostkę innej cechy (np. prędkość w kilometrach na godzinę [km/h]; wydajność w sztukach na godzinę [sztuk/h]; gęstość zaludnienia w liczbie osób na km² [osoba/km] spożycie w kilogramach na osobę [kg/osoba]). Wyznaczamy ją ze wzoru

$$\bar{x}_H = \frac{n}{\sum_{i=1}^k \frac{1}{x_i} n_i} \quad (2.3)$$

gdzie $n = \sum_{i=1}^k n_i$.

W przypadku szeregu rozdzielczego przedziałowego zamiast x_i podstawiamy środek i-tego przedziału klasowego.

Prostym przykładem, kiedy należy wykorzystać średnią harmoniczną jest wyznaczenie średniej prędkości (we wzorze 2.3 należy przyjąć, że x_i jest prędkością na i -tym odcinku, tj. $x_i = v_i$). Przypuśćmy, że odcinek $s = 60$ km przejeżdżamy z prędkością $v_1 = 120$ km/h, następnie w drodze powrotnej przejeżdżamy ten sam odcinek z prędkością $v_2 = 50$ km/h (przypadek, gdy w drodze powrotnej nie możemy skorzystać z drogi szybkiego ruchu). Jaka była średnia prędkość? Zauważmy: jadąc z prędkością 120 km/h podróż zabrała nam $t_1 = \frac{s}{v_1} = 30 \text{ min} = 0,5 \text{ h}$.

Droga powrotna trwała $t_2 = \frac{s}{v_2} = 72 \text{ min} = 1,2 \text{ h}$.

Stąd cała podróż – odcinek 120 km zajęła 1,7 h, co daje średnią prędkość

$$\bar{x}_H = \frac{2}{\frac{1}{v_1} + \frac{1}{v_2}} = 70.59 \text{ km/h}.$$



Zauważmy, że powyższy wzór możemy przekształcić i uzyskamy średnią prędkość wyrażoną w postaci

$$\bar{x}_H = \frac{2s}{\frac{s}{v_1} + \frac{s}{v_2}} = \frac{2s}{t_1 + t_2}$$

Analizując zmiany zachodzące w czasie - gdy zjawiska zmieniają się w ujęciu dynamicznym, średnie tempo zmian można opisać za pomocą **średniej geometrycznej**, którą wyznacza się jako pierwiastek n -tego stopnia z iloczynu n wartości zmiennej X_i

$$\bar{x}_G = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdot \dots \cdot x_n} = \sqrt[n]{\prod_{i=1}^n x_i} \quad (2.4)$$

Z takimi zjawiskami można spotkać się na przykład w demografii, przy okazji analizy wyników inwestycyjnych czy badaniach dotyczących wzrostu lub spadku dochodu narodowego.

Przykład 2.4 – średnia harmoniczna

Oblicz średnią gęstość zaludnienia dla trzech województw o największej powierzchni w km² (tj. woj. mazowieckie (35 559 km²), wielkopolskie (29 827 km²) i lubelskie (25 122 km²) oraz trzech o najmniejszej powierzchni (tj. woj. opolskie (9 411 km²), świętokrzyskie (11 709 km²) i śląskie (12 334 km²). Wykorzystaj podane poniżej dane o gęstości zaludnienia województw oraz liczbie ludności w tysiącach (dane pozyskano ze strony BDL, dotyczą 2024 roku).



Tabela 2.3. Dane do analizy

Województwo	gęstość zaludnienia (ludność na 1 km ²)	Powierzchnia ogółem w km ²
DOLNOŚLĄSKIE	143,8	19 947
KUJAWSKO-POMORSKIE	110,4	17 971
LUBELSKIE	79,5	25 122
LUBUSKIE	69,3	13 988
ŁÓDZKIE	128,8	18 219
MAŁOPOLSKIE	225,8	15 184
MAZOWIECKIE	154,9	35 559
OPOLSKIE	98,8	9 411
PODKARPACKIE	115,6	17 845
PODLASKIE	56,1	20 187
POMORSKIE	120,7	19 547
ŚLĄSKIE	347,9	12 334
ŚWIĘTOKRZYSKIE	98,9	11 709
WARMIŃSKO-MAZURSKIE	55,8	24 174
WIELKOPOLSKIE	116,7	29 827
ZACHODNIOPOMORSKIE	70,8	22 909

Źródło: <https://bdl.stat.gov.pl/bdl/dane/podgrup/temat>

Średnia gęstość zaludnienia (ludzi/km²) dla kilku regionów **nie jest zwykłą średnią arytmetyczną gęstości**, bo każdy region ma inną powierzchnię.

Dlatego stosuje się wzór na średnią harmoniczną ważoną:

$$\bar{x}_H = \frac{\sum_{i=1}^k P_i}{\sum_{i=1}^k \frac{P_i}{x_i}}$$

gdzie :

P_i – powierzchnia i-tego regionu

x_i – gęstość zaludnienia w tym regionie

k – liczba regionów



Średnia harmoniczna dla 3 największych województw

$$\bar{x}_{H_najw} = \frac{35\,559 + 29\,827 + 25\,122}{\frac{35\,559}{154,9} + \frac{29\,827}{116,7} + \frac{25\,122}{79,5}} \approx 113$$

Podobnie wyznaczamy średnią harmoniczną dla trzech województw o najmniejszej powierzchni

$$\bar{x}_{H_najm} \approx 134$$

Należy zwrócić uwagę, że wyznaczone średnie nie określają średniej gęstości zaludnienia w Polsce (taką średnią uzyskalibyśmy obliczając iloraz całkowitej liczby ludności do całkowitej powierzchni Polski), wynosi ona około 119 osób na 1 km².

Obserwacja

Analiza danych dotyczących gęstości zaludnienia w Polsce prowadzi do interesujących, a zarazem pouczających wniosków na temat właściwego interpretowania wartości średnich.

Zestawienie powyższych średnich harmoniczných pokazuje wyraźnie, że średnia krajowa nie jest reprezentatywna dla wszystkich regionów. Mimo znacznie wyższej gęstości zaludnienia w najmniejszych województwach, ich wpływ na wartość średnią jest ograniczony ze względu na niewielki udział powierzchniowy w ogólnej strukturze kraju. Z kolei największe województwa, mimo że charakteryzują się tylko nieznacznie wyższą od średniej krajowej gęstością, ze względu na swój znaczny udział terytorialny, w istotnym stopniu determinują wartość średnią dla Polski.

Wnioski metodologiczne:

1. Średnia globalna może maskować zróżnicowanie regionalne - uogólnianie wartości średniej na poziomie kraju może prowadzić do błędnych interpretacji, jeśli nie uwzględnia się struktury danych (np. udziału powierzchni województw lub liczby ludności).
2. Konieczne jest uwzględnienie wag w analizie, szczególnie przy danych przestrzennych, gdzie jednostki analizy różnią się istotnie pod względem rozmiaru lub znaczenia (w tym przypadku: powierzchni województw).

Rekomendacja dydaktyczna:

Zaleca się, aby podczas analizy danych nie ograniczać się wyłącznie do wartości średnich nieważonych. W sytuacjach, gdy jednostki obserwacji różnią się pod



względem istotnych cech (np. powierzchnia, liczba ludności, udział w PKB), analiza powinna być uzupełniona o wartości ważone, miary rozproszenia (np. odchylenie standardowe), a także analizę rozkładu cechy w populacji.

Przykład 2.5 średnia geometryczna

Na rys. 2.11 zaprezentowano liczbę urodzeń żywych w województwie świętokrzyskim w latach 2007-2024 (dane pozyskano z BDL) oraz indeksy łańcuchowe $x_i = \frac{y_i}{y_{i-1}}$ (indeks łańcuchowy jest miarą, która informuje o ile procent zmieniła się wartość zjawiska w danym okresie badawczym w stosunku do okresu poprzedniego).

[Tekst alternatywny. Zrzut ekranu przedstawiający rozkład szeregu rozdzielczego punktowego zadany przez częstość oraz dystrybuantę. Przedstawiono formułę obliczania indeksu łańcuchowego]

	A	B	C	D
13				
		Rok	urodzenia żywe (ogółem) y_i	indeks łańcuchowy x_i
14				
15		2007	11971	-
16		2008	12736	=C16/C15
17		2009	12741	1,00
18		2010	12445	0,98
19		2011	11600	0,93
20		2012	11510	0,99
21		2013	10720	0,93
22		2014	10756	1,00
23		2015	10579	0,98
24		2016	10610	1,00
25		2017	10915	1,03
26		2018	10721	0,98
27		2019	10220	0,95
28		2020	9488	0,93
29		2021	8717	0,92
30		2022	8157	0,94
31		2023	7238	0,89
32		2024	6472	0,89

Rys 2.11. Zrzut ekranu, dane oraz dynamika ich zmian



Średnia roczna liczba urodzeń w województwie wynosi 10 422 dzieci.

Bardzo ważnym parametrem charakteryzującym kształtowanie się zjawiska w czasie jest **średnie tempo zmian** tego zjawiska. Przez **tempo zmian** rozumie się relatywną zmianę wartości cechy w stosunku do jej wartości poprzedniej.

Aby ocenić średnie tempo zmian w trzeciej kolumnie wyznaczono wartości indeksu dynamiki (tzw. indeksu łańcuchowego), określonego za pomocą ilorazu $x_i = \frac{y_i}{y_{i-1}}$, gdzie y_i oznacza poziom badanego zjawiska. Indeksy posłużyły do wyznaczenia średniej geometrycznej

$$\bar{x}_G = \sqrt[15]{1,06 \cdot 1,00 \cdot 0,98 \cdot 0,93 \cdot \dots \cdot 0,89} = \sqrt[15]{0,54} = 0,96$$

Na jej podstawie wyznaczono średnie tempo zmian

$$G = \bar{x}_G - 1 = -0,04$$

Uzyskana wartość informuje o tym, że w latach 2007-2024 liczba urodzeń w województwie świętokrzyskim malała przeciętnie o 4% z roku na rok.

Dodatkowo warto zauważyć, że największy spadek (o 11%) liczby urodzeń żywych w województwie zaobserwowano w roku 2023 i 2024, odpowiednio w porównaniu do 2022 i 2023 roku.

W rozdziale 2.2 omówiliśmy najpopularniejsze formy prezentacji graficznej rozkładu empirycznego. Do ich sporządzenia konieczna była znajomość liczebności (częstości) rozkładu empirycznego. W przypadku szeregów szczegółowych do oceny rozkładu badanej cechy, jej zmienności i asymetrii wygodnie jest stosować prezentację graficzną. Szczególnym sposobem prezentacji rozkładu cechy jest **wykres pudełkowy** (nazywany również wykresem ramka-wąsy, ang. box-plot).

Do jego sporządzenia potrzebna jest jedna oś liczbowa, na której zaznacza się wartości badanej cechy. Wzdłuż niej umieszcza się prostokąt, którego wysokość wyznaczają kwartyle Q_1 (kwartył pierwszy) i Q_3 (kwartył trzeci) szerokość prostokąta jest dowolna. Wewnątrz prostokąta zaznacza się medianę. Wyznaczony prostokąt uzupełniają dwa odcinki (tzw. wąsy). Jeden odcinek łączy Q_1 z wartością minimalną, a drugi odcinek łączy kwartył Q_3 z wartością maksymalną. Na podstawie wykresu pudełkowego można dokonać „szybkiej” oceny symetrii lub asymetrii rozkładu oraz identyfikacji obserwacji nietypowych w zbiorze danych.

W literaturze przedmiotu istnieje co najmniej kilka sposobów ich definiowania, np. za obserwacje odstające uważa się wartości mniejsze od kwartyła pierwszego i większe od

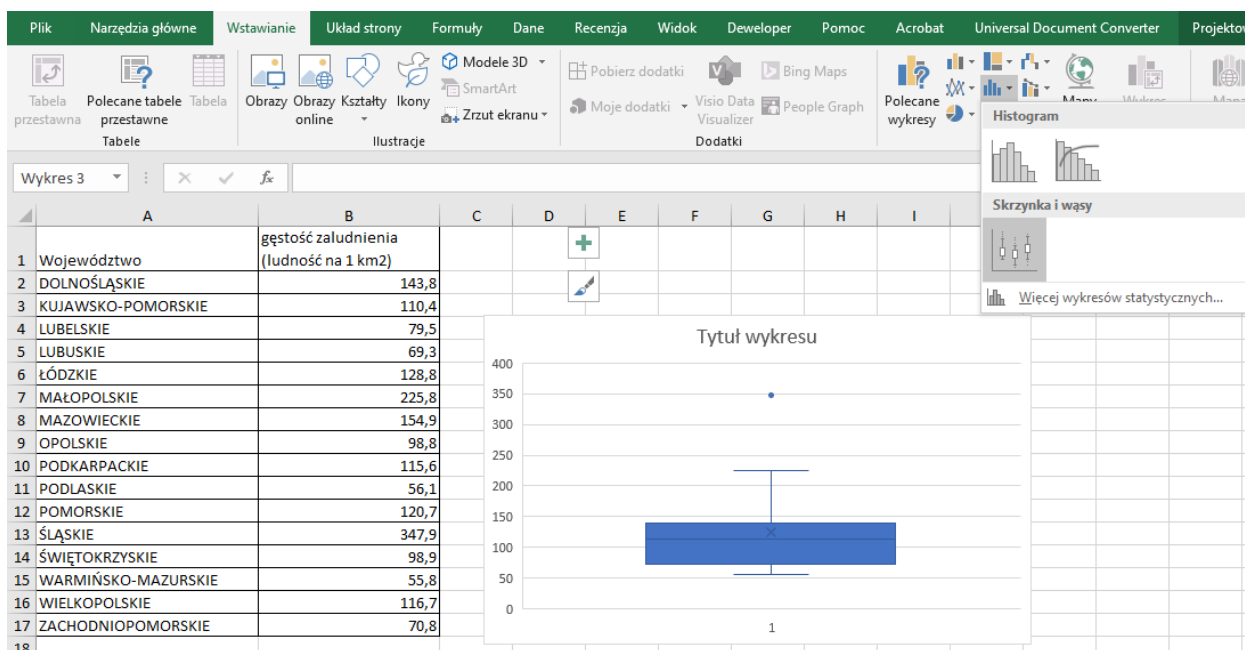


kwartyła trzeciego o więcej niż półtora przedziału ćwiartkowego (tj., półtora długości pudełka).

Przykład 2.6.

Wykres pudełkowy wykonamy dla danych z przykładu 2.4 (gęstość zaludnienia w Polsce z podziałem na województwa).

[Tekst alternatywny. Zrzut ekranu przedstawiający dane dotyczące gęstości zaludnienia w województwach w Polsce w 2024 r oraz instrukcję wykonania wykresu pudełkowego. W zakładce **Wstawianie** (menu główne), wybierz sekcję **Wykresy**, następnie wybierz **Wykres pudełkowy**.]



Rys 2.12. Okno opcji wstawiania wykresu i wykres pudełkowy



Na podstawie powyższego wykresu pudełkowego można sformułować następujące wnioski

- gęstość zaludnienia połowy województw mieści się w przedziale ograniczonym przez kwartyle $Q_1 \approx 73$ i $Q_3 \approx 140$. Oznacza to, że typowe obszary w analizowanej grupie charakteryzują się umiarkowanym zagęszczeniem ludności.
- mediana gęstości zaludnienia ($Q_2 \approx 113$) jest zlokalizowana bliżej kwartyla trzeciego, co wskazuje lewostronną asymetrię⁷ rozkładu badanej cechy. Jednak wnioskując o asymetrii rozkładu w oparciu o kwartyle należy pamiętać o tym, że kwartyle są odporne na skraje wartości – są oparte na 50% danych wokół środka.
- W rozważnym przykładzie mamy do czynienia z **klasyczną pułapką w interpretacji skośności na podstawie miar pozycyjnych**, ponieważ wyznaczając współczynnik skośności na podstawie funkcji *Skośność*() w MS Excel uzyskujemy wartość 2,15, która świadczy o prawostronnej asymetrii rozkładu gęstości ludności w Polsce z podziałem na województwa. To oznacza, że połowa jednostek ma gęstość niższą niż wskazuje na to średnia (około 119 osób/km²).

Ta rozbieżność interpretacji związana jest z występowaniem obserwacji odstającej po prawej stronie (woj. śląskie, które jest wysoce zurbanizowane i wyraźnie odbiega od pozostałych województw).

3. Zmienna losowa jednowymiarowa

3.1. Pojęcie zmiennej losowej oraz jej charakterystyki

Prawie wszystkie wielkości, z którymi mamy do czynienia, mają (mniej lub bardziej) losowy charakter. Każda zmienna losowa ma swój specyficzny charakter, który można scharakteryzować rozkładem prawdopodobieństwa oraz pewnymi standardowymi parametrami.

Pojęcie zmiennej losowej umożliwia przejście od modelu probabilistycznego (Ω, Σ, P) do typowych pojęć matematycznych, jakimi są liczby i funkcje.

Zmienna to atrybut lub właściwość, która opisuje osobę, rzecz lub zjawisko.

⁷ Przypomnijmy:

- w rozkładzie lewostronnie skośnym: $Q_3 - Q_2 < Q_2 - Q_1$;
- w rozkładzie prawostronnie skośnym: $Q_3 - Q_2 > Q_2 - Q_1$



Przypomnijmy

Zmienne losowe oznaczamy wielkimi literami: $X, Y, Z, \dots$, natomiast ich wartości (tzw. realizacje) małymi $x, y, z, \dots$

Rozróżnia się dwa typy zmiennych losowych:

- zmienne losowe typu skokowego (dyskretnego),
- zmienne losowe typu ciągłego.

Zacniemy od przykładu zmiennej losowej o rozkładzie dyskretnym.

Przykład 3.1

Firma ubezpieczeniowa chce ustalić wysokość rocznej składki za ubezpieczenie OC kierowców. Pomimo, że wysokość tej składki zależy od wielu czynników, zarówno związanych z kierowcą, jak i pojazdem (między innymi od wieku, doświadczenia kierowcy, miejsca zamieszkania, marki i modelu samochodu, pojemności silnika, a także sposobu użytkowania pojazdu) firma postanowiła oprzeć swoje badania na wysokości wypłaconych odszkodowań w ciągu jednego roku.

Niech X oznacza wysokość wypłaconego odszkodowania. X jest zmienną losową, gdyż:

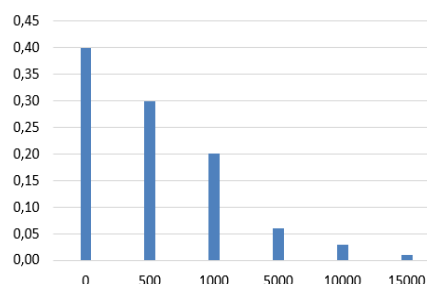
- Nie wiadomo z góry, czy w ogóle będzie szkoda (czyli $x = 0$),
- Jeśli będzie, to ile będzie kosztować – np. 500 zł, 1000 zł, 5 000 zł, 10 000 zł albo 150 000 zł.

W oparciu o dane historyczne firma przeprowadziła analizę wysokości wypłat odszkodowań w ciągu roku. W wyniku analizy wyznaczono rozkład zmiennej losowej X (tj. prawdopodobieństwa realizacji wypłaty odszkodowania o danej wysokości w ciągu roku). Dane (fikcyjne) zebrano w tabeli 3.1 i zilustrowano na rys. 3.1.

Tabela 3.1. Rozkład zmiennej
dyskretnej

X_i	p_i
0	0,40
500	0,30
1 000	0,20
5 000	0,06
10 000	0,03
15 000	0,01

[Tekst alternatywny. Histogram częstości dyskretnej zmiennej losowej. Opis osi X: wartości analizowanej zmiennej losowej (kwota wypłat odszkodowań), opis osi Y – prawdopodobieństwu zajścia danej wartości analizowanej zmiennej losowej. Na wykresie znajdują się niebieskie linie pionowe.]

Rys. 3.1. Wykres rozkładu prawdopodobieństwa
wysokości wypłaconego odszkodowania

Na podstawie powyższych danych wyznacz

- prawdopodobieństwo, że wypłata odszkodowania przekroczy 1000 zł;
- wartość oczekiwaną szkody,
- odchylenie standardowe (czyli jak bardzo kwoty odszkodowań mogą różnić się od średniej).
- dystybuantę, przedstaw ją graficznie oraz zinterpretuj jej wartość w punkcie $x_i = 5\,000$.

Dane można wykorzystać między innymi do wyceny ryzyka oraz ustalenia wysokości składki ubezpieczeniowej (w zależności od liczby wykupionych polis), która z jednej strony pokryje oczekiwane wypłaty, a z drugiej umożliwi firmie osiągnięcie zysku z prowadzonej działalności.

Rozwiązanie:

Ad a) $P(X > 1\,000) = 1 - P(X \leq 1\,000) = 1 - 0,4 - 0,2 = 0,4$

Ad b) Wartość oczekiwana zmiennej losowej dyskretnej wyznaczana jest za pomocą wzoru



$$E(X) = \sum_{i=1}^k x_i p_i \quad (3.1)$$

Obliczenia przeprowadzimy w arkuszu kalkulacyjnym MS Excel

[Tekst alternatywny. Zrzut ekranu przedstawiający dane oraz formułę wyznaczania wartości oczekiwanej.]

	A	B	C	D	E	F
1						
2		X_i	p_i	X_i^2	$F(x_i)$	
3		0	0,40	0	0,40	
4		500	0,30	250000	0,70	
5		1000	0,20	1000000	0,90	
6		5000	0,06	25000000	0,96	
7		10000	0,03	100000000	0,99	
8		15000	0,01	225000000	1,00	
9						
10		E(X)	=SUMA.ILOCZYNÓW(B3:B8;C3:C8)			=1100

Rys. 3.2. Fragment arkusza przedstawiający sposób wyznaczenia wartości oczekiwanej

Wartość oczekiwana wskazuje, że przeciętna kwota roszczeń w ciągu roku wynosi 1100 zł.

Ad c) Odchylenie standardowe

Ze względu na fakt, że odchylenie standardowe wyznaczone jest jako pierwiastek z wariancji najpierw wyznaczamy wariancję. W tym celu korzystamy ze wzoru

$$D^2(X) = \sum_{i=1}^k (x_i - E(X))^2 p_i = \sum_{i=1}^k x_i^2 p_i - (E(X))^2 \quad (3.2)$$



[Tekst alternatywny. Zrzut ekranu przedstawiający dane oraz formuły wyznaczania wariancji oraz odchylenia standardowego dyskretnej zmiennej losowej.]

	A	B	C	D	E	F	G
1							
2		X_i	p_i	X_i^2	$F(x_i)$		
3		0	0,40	0	0,40		
4		500	0,30	250000	0,70		
5		1000	0,20	1000000	0,90		
6		5000	0,06	25000000	0,96		
7		10000	0,03	100000000	0,99		
8		15000	0,01	225000000	1,00		
9							
10		E(X)	1100				
11		$D^2(X)$	=SUMA.ILOCZYNÓW(D3:D8;C3:C8)-C10^2				=5815000,00
12		$D(X)$	=PIERWIASTEK(C11)		=2411,43		

Rys. 3.3. Fragment arkusza przedstawiający sposób wyznaczenia wariancji i odchylenia standardowego

Zauważmy, że wartości badanej zmiennej losowej charakteryzują się wysoką zmiennością, tj. odchylenie standardowe stanowi ponad 219% wartości oczekiwanej. W związku z powyższym ustalenie wysokości rocznej składki za ubezpieczenie OC nie powinno bazować na wartości oczekiwanej. W tym przypadku lepiej wykorzystać wartość mediany m_e (wartość środkowa, odporna na wartości odstające i nie wymaga symetryczności rozkładu) lub kwartyła trzeciego $q_{0,75} \equiv Q_3$ (to będzie opcja bardziej opłacalna dla firmy ubezpieczeniowej)

$$P(X \leq m_e) = P(X > m_e) = 0,5 \Rightarrow m_e = 500$$

$$P(X \leq q_{0,75}) = 0,75 \Rightarrow q_{0,75} = 1000$$

Ad d) Do opisu rozkładu zmiennej losowej, zamiast funkcji rozkładu prawdopodobieństwa ($p_i = P(X = x_i)$) można użyć funkcji nazywanej dystrybuantą, która w przypadku zmiennej losowej dyskretnej wyznaczana jest ze wzoru

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} p_i \quad (3.3)$$

Wartości funkcji zostały podane na rys. 3.4 gdzie

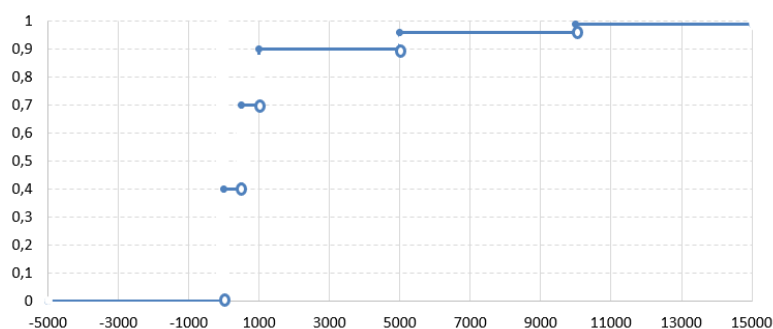


$$F(0) = P(X \leq 0) = P(X = 0) = 0,4$$

$$F(500) = P(X \leq 500) = P(X = 0) + P(X = 500) = 0,4 + 0,3 = 0,7 \quad itd.$$

Wartość dystrybuanty w punkcie 5000 $F(5000) = 0,96$ oznacza, że 96% wypłacanych przez firmę odszkodowań nie przekracza kwoty 5000 zł (zaledwie w 4 przypadkach na 100 firma wypłaca odszkodowania większe niż 5000 zł).

[Tekst alternatywny. Wykres dwuwymiarowy. Opis osi X: wartości analizowanej zmiennej losowej (kwota wypłat odszkodowań), opis osi Y – wartość dystrybuanty. Na wykresie znajdują się niebieskie linie poziome (tzw. funkcja schodkowa lewostronnie ciągła), skoki występują w punktach odpowiadających wartościom zmiennej losowej. Wysokość każdego skoku jest równa prawdopodobieństwu zajścia danej wartości analizowanej zmiennej losowej.]



Rys. 3.4. Wykres dystrybuanty wysokości wypłaconego odszkodowania

Odpowiedź:

Z przeprowadzonych obliczeń wynika, że wysokość rocznej składki za ubezpieczenie OC powinna mieścić się w przedziale (500 zł; 1000 zł).

Do zadania wrócimy w podrozdziale dotyczącym centralnego twierdzenia granicznego (CTG).

W przypadku, gdy zmienna losowa przyjmuje nieprzeliczalnie wiele wartości (np. przyjmuje wartości z pewnego przedziału albo sumy przedziałów zbioru liczb rzeczywistych) mamy do czynienia ze zmienną losową ciągłą. Rozkład takiej zmiennej określany jest przez funkcję gęstości rozkładu $f(x)$ i/lub dystrybuantę, która określana jest wzorem



$$F(a) = \int_{-\infty}^a f(x) dx \quad (3.4)$$

Przykład 3.2

W pewnej przychodni, ustalona godzina wizyty jest jedynie godziną orientacyjną. Ostateczna godzina przyjęcia zależy od ilości i czasu trwania poprzedzających wizyt, co wynika z indywidualnych potrzeb każdego pacjenta, dlatego w przypadku przesunięcia godziny przyjęć o kolejności wejścia do gabinetu decyduje kolejność zapisu. Z obserwacji wynika, że czas oczekiwania na przyjęcie jest zmienną losową i wynosi od 5 do 35 minut.

Przyjmując, że prawdopodobieństwo przyjęcia jest jednakowe w tym czasie, określić funkcję gęstości dla rozkładu czasu oczekiwania na przyjęcie. Oblicz

- wartość oczekiwaną oraz odchylenie standardowe czasu oczekiwania na przyjęcie przez lekarza;
- prawdopodobieństwo, że czas oczekiwania będzie dłuższy niż 10 minut, ale nie przekroczy 20 minut;
- kwantyl rzędu 0,8.

Rozwiązanie:

Z treści zadania wynika, że czas oczekiwania na przyjęcie jest zmienną losową i wynosi od 5 do 35 minut, to oznacza, że właściwą postacią rozkładu badanej zmiennej będzie rozkład jednostajny, tj. rozkład zadany funkcją gęstości postaci

$$f(x) = \begin{cases} a, & \text{dla } x \in [5, 35] \\ 0, & \text{w pozostałych przypadkach} \end{cases}$$



Z własności zmiennej losowej typu ciągłego wynika, że funkcja gęstości musi spełniać następujące warunki:

- (1) $\forall x \in R, f(x) \geq 0$,
- (2) $\int_{-\infty}^{\infty} f(x) dx = 1$,
- (3) funkcja f jest funkcją ciągłą poza co najwyżej skończoną liczbą punktów nieciągłości.

W rozważanym przykładzie własność (1) jest spełniona dla $a \geq 0$.
Korzystając z własności (2) otrzymujemy

$$(1) \int_{-\infty}^{\infty} f(x) dx = \int_{-\infty}^5 0 dx + \int_5^{35} a dx + \int_{35}^{\infty} 0 dx = ax \Big|_5^{35} = 30a$$

stąd $a = \frac{1}{30}$.

Zauważmy, że spełniony jest również warunek (3), gdyż funkcja f ma dwa punkty nieciągłości.

Ad a) średni czas oczekiwania na przyjęcie przez lekarza wynosi

$$E(X) = \int_{-\infty}^{\infty} xf(x) dx = \int_5^{35} \frac{1}{30} x dx = \frac{x^2}{60} \Big|_5^{35} = 20$$

wariancja

$$D^2(X) = E(X^2) - [E(X)]^2 = \int_{-\infty}^{\infty} x^2 f(x) dx - 20^2 = \int_5^{35} \frac{x^2}{30} dx - 400 = \frac{x^3}{90} \Big|_5^{35} - 400 = 75$$

odchylenie standardowe

$$D(X) = \sqrt{D^2(X)} = \sqrt{75} = 8,66.$$

Ad b) Z funkcji gęstości prawdopodobieństwo, że czas oczekiwania będzie dłuższy niż 10 minut, ale nie przekroczy 20 minut wynosi

$$P(10 < X \leq 20) = \int_{10}^{20} \frac{1}{30} dx = \frac{x}{30} \Big|_{10}^{20} = \frac{1}{3}.$$

c) Kwantyl rzędu 0,8 oznaczamy jako $q_{0,8}$. W przypadku zmiennej losowej ciągłej jest to wartość, dla której dystrybuenta wynosi $F(q_{0,8}) = 0,8$.

Aby wyznaczyć kwantyl należy rozwiązać równanie



$$F(q_{0,8}) = \int_{-\infty}^{q_{0,8}} f(x)dx = \int_5^{q_{0,8}} \frac{1}{30} dx = \frac{q_{0,8}-5}{30} = 0,8 \Leftrightarrow q_{0,8} = 29.$$

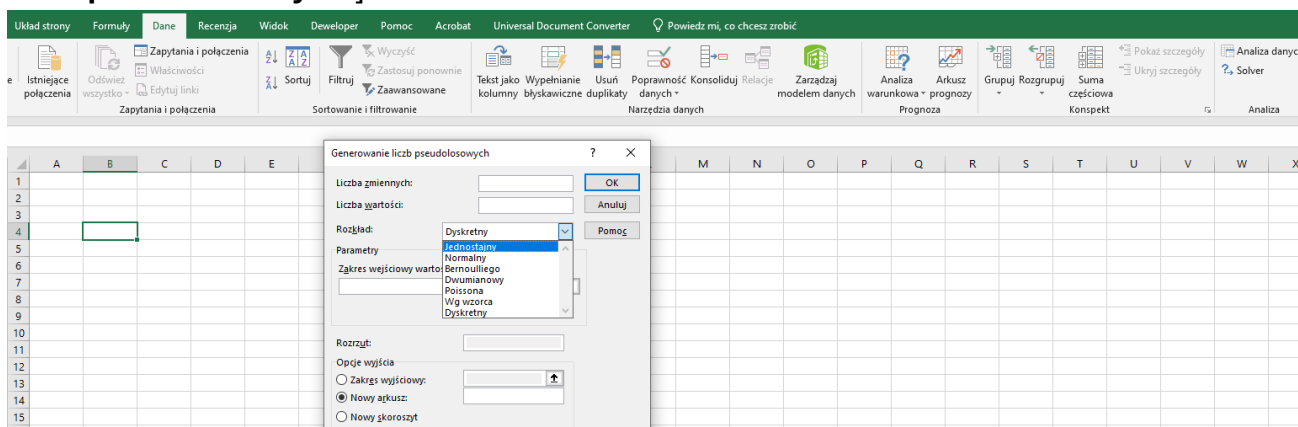
Uzyskana wartość informuje o tym, że 80% pacjentów będzie oczekiwać na przyjęcie przez lekarza nie dłużej niż 29 minut.

Własności zmiennych losowych zostały podane na wykładzie, może je również znaleźć np. w książkach podanych w literaturze.

3.2. Wybrane typy rozkładów

W arkuszu kalkulacyjnym MS Excel dostępne są funkcje umożliwiające generowanie liczb pseudolosowych o zadanym rozkładzie prawdopodobieństwa. Pozwalają one w prosty sposób symulować przebieg zjawisk losowych, badać własności wybranych rozkładów teoretycznych oraz porównywać dane empiryczne z idealnymi wzorcami probabilistycznymi.

[Tekst alternatywny. Zrzut ekranu. Wywołanie opcji **generowanie liczb pseudolosowych** dostępnej w zakładce **Dane -> Analiza danych -> generowanie liczb pseudolosowych**.]



Rys. 3.5. Wybrane typy rozkładów teoretycznych

W tym rozdziale skoncentrujemy się na dwóch teoretycznych rozkładach prawdopodobieństwa: rozkładzie dwumianowym (rozkład dyskretny) oraz rozkładzie normalnym, znany również jako rozkład Gaussa (rozkład ciągły).

W praktyce bardzo często analiza problemów w obszarze inżynierii danych dotyczy sytuacji, w których są możliwe tylko dwa wyniki, opisywane w analizie statystycznej



przez zmienne binarne (najczęściej kodowane jako zmienne zero-jedynkowe: 1- sukces; 0- porażka). Prostym uogólnieniem rozkładu zero-jedynkowego na n niezależnych doświadczeń jest **rozkład dwumianowy**. Rozkład ten dotyczy wielokrotnej realizacji doświadczenia, w wyniku którego sukces pojawi się z prawdopodobieństwem p , (a porażka z prawdopodobieństwem $1 - p$).

Przykład 3.3

Elementem towarzyszącym każdej decyzji, w tym decyzji inwestycyjnej, jest ryzyko. Wynika ono z faktu, że decyzje opierają się na prognozie co do przyszłych warunków, a nie na informacjach pewnych. Poniesienie straty może być zdefiniowane jako sytuacja, w której stopa zwrotu z inwestycji jest mniejsza od zera.

Inwestor zastanawia się nad opłacalnością inwestycji w złoto. W tym celu ze strony <https://stoog.pl/> pozyskał informacje o miesięcznym kursie złota XAUPLN (jest to cena złota za uncję trojańską złota wyrażona w polskiej walucie) z ostatniego roku (dane przedstawiono w tabeli). Wyznaczył stopy zwrotu i oszacował prawdopodobieństwo sukcesu p (tj. sytuacji, w której stopa zwrotu jest większa od zera).

Przyjmując, że rozkład stóp zwrotu jest rozkładem dwumianowym z prawdopodobieństwem sukcesu p , przy liczbie $n = 12$ (n oznacza liczbę stóp zwrotu wykorzystywaną w eksperymencie) wyznaczyć:

- rozkład liczby sukcesów;
- wartość oczekiwaną oraz odchylenie standardowe liczby sukcesów;
- prawdopodobieństwo, że co najmniej przez 8 miesięcy stopa zwrotu będzie dodatnia.



[Tekst alternatywny. Zrzut ekranu przedstawiający dane wejściowe, stopy zwrotu oraz formułę przypisującą danej pozycji sukces lub porażkę.]

	B	C	D	E	F	G	H	I	J	K
		Kurs XAUPLN								
1	Data	Zamknięcie	Stopa zwrotu	sukces/porażka						
2	30.06.2024	9358,86								
3	31.07.2024	9697,54	3,62%	=JEŻELI(D3>0;1;0)						
4	31.08.2024	9694,63	-0,03%							
5	30.09.2024	10137	4,56%	1						
6	31.10.2024	10972,8	8,25%	1						
7	30.11.2024	10757,5	-1,96%	0			p	=SUMA(E3:E14)/12=		0,67
8	31.12.2024	10840,1	0,77%	1						
9	31.01.2025	11392,2	5,09%	1						
10	28.02.2025	11545,9	1,35%	1						
11	31.03.2025	12089,1	4,70%	1						
12	30.04.2025	12422,5	2,76%	1						
13	31.05.2025	12328,6	-0,76%	0						
14	30.06.2025	11891,8	-3,54%	0						
15										

Rys. 3.6. Tabela zawierająca kurs złota XAUPLN (kolumna B) w poszczególnych miesiącach, wyznaczoną stopę zwrotu (kolumna D) oraz prawdopodobieństwo sukcesu

p

Rozwiązanie

Ad a)

Definiujemy pomocniczą zmienną

$$X_i = \begin{cases} 1, & \text{gdy stopa zwrotu jest większa od 0,} \\ 0, & \text{w przeciwnym przypadku,} \end{cases} \quad \begin{matrix} p = 0,67 \\ q = 0,33 \end{matrix}$$

oraz zmienną o rozkładzie dwumianowym określającą liczbę sukcesów w $n = 12$ próbach $S_{12} = \sum_{i=1}^{12} X_i$



[Tekst alternatywny. Zrzut ekranu przedstawiający sposób wyznaczania rozkładu zmiennej losowej o rozkładzie dwumianowym z zadanymi parametrami. Zastosowanie funkcji **ROZKŁ.DWUM()**.]

	B	C	D	E	F	G	H
15							
16							
17			$P(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}$				
18		S_{12}	$P(S_{12} = k)$				
19			$0 = \text{ROZKŁ.DWUM}(C19;12;0,67;0)$				
20			1	ROZKŁ.DWUM(liczba_s; próby; prawdopodob_s; skumulowany)			
21			2	0,00			
22			3	0,00			
23			4	0,01			
24			5	0,05			
25			6	0,11			
26			7	0,19			
27			8	0,24			
28			9	0,22			
29			10	0,13			
30			11	0,05			
31			12	0,01			

Rys. 3.7. Rozkład liczby sukcesów w 12 próbach (tj. zmiennej losowej S_{12})

Ad b)

Wartość oczekiwana liczby sukcesów w 12 próbach wyznaczana jest ze wzoru

$$E(S_{12}) = np \text{ i wynosi } E(S_{12}) = 12 \cdot 0,67 = 8,04$$

natomiast odchylenie standardowe $D(S_{12}) = \sqrt{npq}$ wynosi

$$D(S_{12}) = \sqrt{12 \cdot 0,67 \cdot 0,33} = 2,65$$

Ad c)

$$P(S_{12} \geq 8) =$$

$$P(S_{12} = 8) + P(S_{12} = 9) + P(S_{12} = 10) + P(S_{12} = 11) + P(S_{12} = 12) = 0,24 + 0,22 + 0,13 + 0,05 + 0,01 = 0,64$$

Obserwacja: Inwestycja w złoto od lat uważana jest za jedną z najbezpieczniejszych metod ochrony pieniędzy przed inflacją i zawirowaniami gospodarczymi. Dane w powyższej tabeli (rys. 3.7) potwierdzają ten fakt, o czym świadczy bliskie zero prawdopodobieństwo, że w ciągu roku odnotujemy zysk tylko co najwyżej przez 4 razy (tzn. przez 2/3 roku odnotowano dodatnią stopę zwrotu). Najbardziej prawdopodobny jest scenariusz, że w ciągu roku 8 razy uzyskamy dodatnią stopę zwrotu.

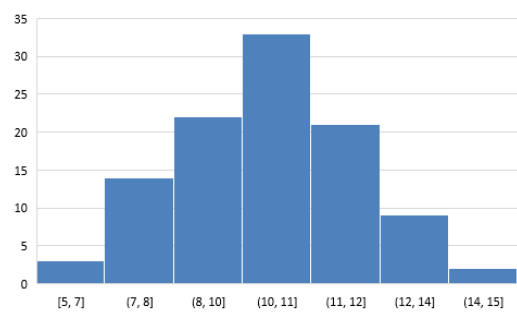
Osobom zainteresowanym polecam stronę <https://www.mennica.com.pl/produkty-inwestycyjne/dlaczego-warto-inwestowac-w-zloto>



Specjalne miejsce w teorii statystyki i rachunku prawdopodobieństwa zajmuje **rozkład normalny**. Popularność tego rozkładu wynika z faktu, że przy nieograniczonym wzroście liczby niezależnych doświadczeń, rozkład sumy zmiennych losowych o dowolnym rozkładzie teoretycznym (dyskretnym i ciągłym) jest do niego zbieżny (tzw. centralne twierdzenie graniczne). W praktyce ze zmiennymi losowymi o rozkładzie normalnym spotykamy się w przypadkach, gdy na wartość badanej zmiennej losowej wpływa duża ilość niezależnie działających czynników losowych, z których żaden nie ma dominującego wpływu.

Przykład 3.4

W pewnej firmie przeprowadzono badanie jakości obsługi klienta. Z uzyskanych danych wynika, że czas oczekiwania na obsługę ma w przybliżeniu rozkład normalny ze średnią wynoszącą 10 minut i odchyleniem standardowym równym 3 minuty. Zebrane dane przedstawiono na rysunku 3.8.



Rys.3.8. Histogram liczebności czasu obsługi

Wyznacz:

- a) jaki odsetek klientów czeka na obsługę co najwyżej 2 minuty;
- b) jaka jest szansa, że czas oczekiwania będzie dłuższy niż 5 minut, ale krótszy niż 15 minut;
- d) jaki czas oczekiwania na obsługę będzie miało 1% najdłużej oczekujących klientów?

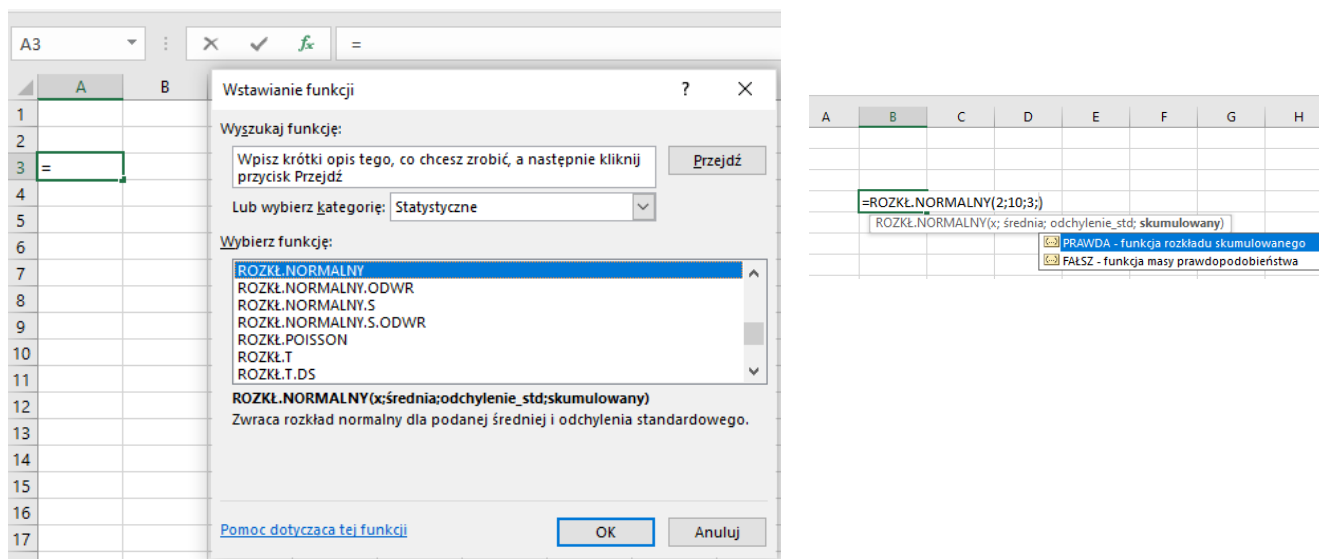
Rozwiązanie

Niech X będzie zmienną losową określającą czas oczekiwania na obsługę, zadaną rozkładem normalnym $N(10; 3)$.

Ad a)

W celu wyznaczenia jaki procent klientów czeka na obsługę nie dłużej niż 2 minuty wykorzystamy dystrybucję rozkładu normalnego

[Tekst alternatywny. Zrzut ekranu przedstawiający sposób wyznaczania dystrybuanty zmiennej losowej o rozkładzie normalnym. Zastosowanie funkcji **ROZKŁ.NORMALNY()**.]



Rys. 3.9. Wywołanie funkcji rozkładu normalnego

$$P(X \leq 2) = \text{ROZKŁ.NORMALNY}(2; 10; 3; 1) = 0,004$$

Zaledwie 0,4% klientów ma szansę na by czekać na obsługę krócej niż 2 minuty.

Ad b)

$$P(5 < X < 15) = \text{ROZKŁ.NORMALNY}(15; 10; 3; 1) - \text{ROZKŁ.NORMALNY}(5; 10; 3; 1) \\ = 0,95 - 0,05 = 0,9$$

90% klientów będzie oczekiwać na obsługę dłużej niż 5, ale krócej niż 15 minut.

Ad c) Należy wyznaczyć kwantyl rzędu 0,99 $q_{0,99}$, w tym celu należy znaleźć argument dystrybuanty dla którego wartość dystrybuanty wynosi 0,99.

Funkcją, która zwraca argument dystrybuanty dla zadanego prawdopodobieństwa p , czyli kwantyl rzędu p rozkładu normalnego jest funkcja

$$q_{0,99} = \text{ROZKŁ.NORMALNY.ODW}(0,99; 10; 10) = 16,98 \approx 17$$



	A	B	C	D	E	F	G	H
1								
2								
3								
4		=ROZKŁ.NORMALNY.ODWR()						
5		ROZKŁ.NORMALNY.ODWR(prawdopodobieństwo; średnia; odchylenie_std)						
6								

Rys. 3.10. Zrzut ekranu. Okno funkcji statystycznej

Wniosek: jeden klient na 100 będzie oczekiwał na obsługę dłużej niż 17 minut.

4. Centralne twierdzenie graniczne Lindeberga - Levy'ego

Centralne twierdzenie graniczne (CTG) jest najważniejszym twierdzeniem granicznym. Zanim przejdziemy do jego zastosowań przypomnijmy jego treść.

Jeżeli $X_1, X_2, X_3, \dots$ jest ciągiem niezależnych zmiennych losowych o identycznym rozkładzie mającym skończoną wartość oczekiwaną $E(X_i) = m$ i wariancją $D^2(X_i) = \sigma^2$, to rozkład prawdopodobieństwa zmiennej losowej Z_n postaci

$$Z_n = \frac{\sum_{i=1}^n X_i - nm}{\sigma\sqrt{n}} \quad (4.1)$$

jest zbieżny do standardowego rozkładu Gaussa $N(0, 1)$

Wniosek 1

Z powyższego twierdzenia wynika, że dla określonych wyżej zmiennych losowych X_k , zmienna losowa zadana sumą $S_n = \sum_{i=1}^n X_i$ (duże n) ma w przybliżeniu rozkład normalny $N(nm; \sigma\sqrt{n})$.

Wniosek 2 (Aczel, 2017)

Jeżeli pobieramy próbę z populacji o średniej m i skończonym odchyleniu standardowym σ , to rozkład średniej z próby $\bar{S}_n = \frac{1}{n} \sum_{i=1}^n X_i$, dąży do rozkładu normalnego $N(m; \frac{\sigma}{\sqrt{n}})$, gdy liczebność próby wzrasta nieograniczenie, czyli dla „dostatecznie dużych n ”.



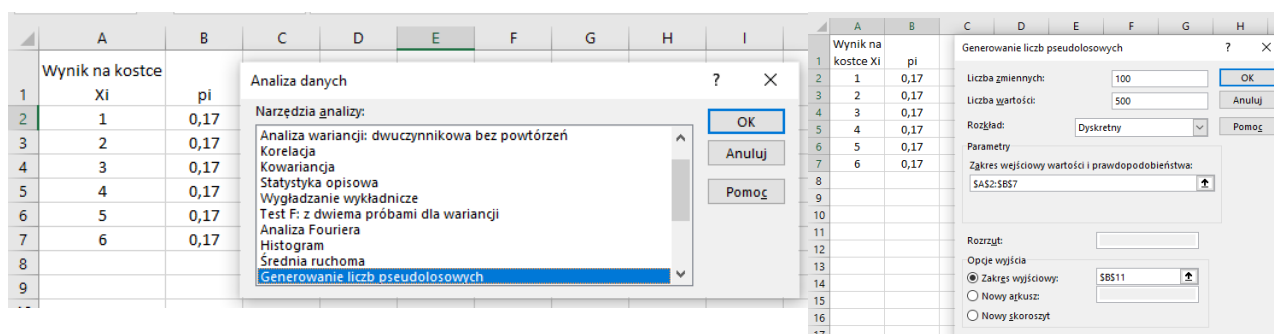
Przykład 4.1 – idea CTG

W arkuszu MS Excel wykonaj następujące kroki

1. wylosuj n wartości z dowolnego rozkładu (tj. mamy n niezależnych realizacji zmiennej losowej o tym rozkładzie);
2. wyznacz sumy tych realizacji (zmienną S_n);
3. wyznacz histogram zmiennej S_n .

Opis eksperymentu: Rzucamy 500 razy 6-ścienną symetryczną kostką do gry. Doświadczenie powtarzamy 100 razy. Zliczamy sumę oczek.

[Tekst alternatywny. Zrzut ekranu przedstawiający uruchomienie narzędzia Generowanie liczb pseudolosowych.]



Rys. 4.1. Okno funkcji generowanie serii zmiennych losowych o zadanym rozkładzie dyskretnym z pozycji Dane -> Analiza danych -> generowanie liczb pseudolosowych

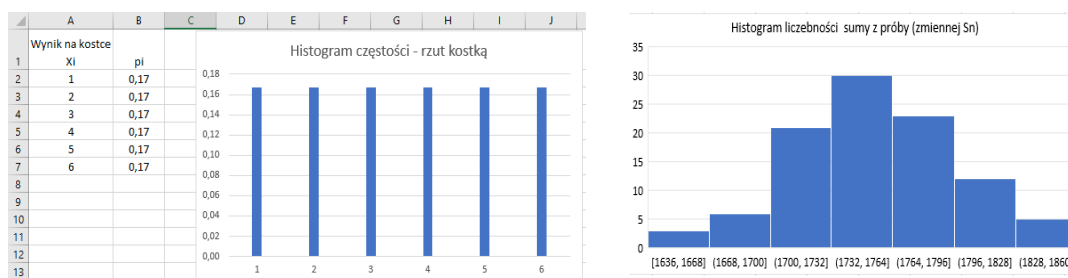
[Tekst alternatywny. Zrzut ekranu przedstawiający dane wejściowe oraz fragment tabeli realizacji 100 prób losowych wygenerowanych z zadanego rozkładu (zgodnie z ideą CTG). Linia przerywaną zaznaczono miejsce ukrytych wierszy]

	A	B	C	D	E	CV	CW	CX
1	Wynik na kostce Xi	pi						
2	1	0,17						
3	2	0,17						
4	3	0,17		E(Xi)	3,50			
5	4	0,17		D(Xi)	1,71			
6	5	0,17						
7	6	0,17						
8								
9								
10	lp		realizacja 1	realizacja 2	realizacja 3	...	realizacja 99	realizacja 100
11	1		3	4	1		4	4
12	2		5	6	6		3	3
13	3		1	3	2		3	2
508	498		5	1	5		6	6
509	499		5	1	6		5	4
510	500		3	1	1		4	2
511	Sn		1743	1763	1794		1727	1773
512	Zn		=(C511-500*E4)/(E5*PIERWIASTEK(500))					



Rys. 4.2. Zrzut ekranu

[Tekst alternatywny. Zrzut ekranu przedstawiający rozkład populacji w przypadku rzutu kostką oraz rozkład sumy z próby. Wyznaczenie rozkładu zmiennej losowej o rozkładzie dwumianowym z zadanymi parametrami.]



Rys. 4.3. Rozkład w populacji i rozkład sumy z próby
rozważanych zmiennych losowych X_i (wynik rzutu kostką)

Możemy zauważyć, że pobierając próbę losową z populacji o rozkładzie jednostajnym⁸ na zbiorze wartości $\{1, 2, 3, 4, 5, 6\}$ (wyniki na kostce), a następnie wyznaczając możliwe wartości sumy z próby i ich prawdopodobieństwa rozkład tej sumy ma w przybliżeniu rozkład normalny ze średnią $n \cdot m$ (w rozważanym przykładzie $500 \cdot 3,5 = 1750$), oraz odchyleniem standardowym $\sigma\sqrt{n}$ ($1,71\sqrt{500} = 38,24$).

Przykład 4.2

Wirusowe zapalenie wątroby typu B (wzw B) to jedna z najgroźniejszych chorób zakaźnych człowieka. Do zakażenia dochodzi przez kontakt z zakażoną krwią, kontakty seksualne z zakażonymi. Standardowe szczepienie zapewnia ochronę u ponad 90% zdrowych osób dorosłych. Niższą skuteczność szczepionki (60-70%) zaobserwowano w grupach osób starszych oraz osób przewlekle chorych

(<https://szczepienia.pzh.gov.pl/szczepionki/wzw-b/>)

Producent nowej szczepionki przeciw wzw B twierdzi, że jej skuteczność (bez względu na wiek i stan zdrowia osoby szczepionej) wynosi 83%. Zakładając, że nowa szczepionka będzie testowana na 100 osobach znaleźć przybliżone prawdopodobieństwo, że pożądaną odporność uzyska więcej niż 80 osób.

⁸ Inaczej mówiąc, mamy do czynienia z populacją, w której występują liczby od 1 do 6 przy jednakowej częstości występowania każdej z nich.



Rozwiązanie

Niech

$$X_i = \begin{cases} 1, & \text{osoba uzyskała odporność } p = 0,83 \\ 0, & \text{w przeciwnym przypadku } q = 0,17 \end{cases}$$

definiujemy zmienną $S_{100} = X_1 + X_2 + \dots + X_{100}$ określającą liczbę osób spośród 100 testowanych, które uzyskają odporność (S_{100} jest zmienną o rozkładzie dwumianowym)

Zauważmy: $E(X_i) = 0,83$, $D(X_i) = \sqrt{0,83 \cdot 0,17} = 0,38$ stąd $E(S_{100}) = 83$, $D(S_n) = 3,8$
Szukane prawdopodobieństwo, wyznaczmy na dwa sposoby

- (1) gdy do dyspozycji mamy tablice rozkładu normalnego, tj. należy przeprowadzić standaryzację zmiennej S_{100} (zgodnie ze wzorem 4.1)

$$\begin{aligned} P(S_{100} > 80) &= 1 - P(S_{100} \leq 80) = 1 - P\left(\frac{S_{100} - 83}{3,8} \leq \frac{80 - 83}{3,8}\right) \\ &= 1 - P(Z_{100} \leq -0,79) = 1 - \Phi_{N(0,1)}(-0,79) \\ &= 1 - (1 - \Phi_{N(0,1)}(0,79)) = 0,785 \end{aligned}$$

gdzie $\Phi_{N(0,1)}(\cdot)$ oznacza dystrybucję standardowego rozkładu normalnego.
Tablice rozkładu umieszczono w rozdziale 7 Wzory.

- (2) gdy dysponujemy dostępem do arkusza kalkulacyjnego, tj. możemy wyznaczyć wartość dystrybucyjny rozkładu normalnego o dowolnych parametrach

$$\begin{aligned} P(S_{100} > 80) &= 1 - P(S_{100} \leq 80) = 1 - \text{ROZKŁ. NORMALNY}(80; 83; 3,8; 1) \\ &= 1 - 0,215 \approx 0,785. \end{aligned}$$

Przykład 4.3

Młody inwestor rozważa inwestowanie w akcje spółek: KGHM Polska Miedź S.A.⁹; PKN Orlen Spółka Akcyjna¹⁰ oraz PKO Bank Polski¹¹. Przed podjęciem ostatecznej decyzji chce się dowiedzieć, **czy przeciętna dzienna stopa zwrotu z tej inwestycji jest stabilna oraz jakie jest prawdopodobieństwo, że przekroczy 2,5%.**

Rozwiązanie

⁹ Spółka strategiczna, jedna z największych polskich spółek z udziałem Skarbu Państwa, jeden z czołowych producentów miedzi i srebra rafinowanego na świecie.

¹⁰ Polski koncern wielobranżowy, operujący na rynkach biopaliw, gazownictwa, energii elektrycznej i ciepłej, petrochemii, prasowym, produkcji nawozów sztucznych, poszukiwań, wydobywania i przerobu ropy naftowej oraz sprzedaży detalicznej.

¹¹ Największy bank uniwersalny w Polsce, spółka akcyjna notowana od 2004 na Giełdzie Papierów Wartościowych w Warszawie.



Inwestor może uzyskać wgląd w kształt rozkładów badanych stóp zwrotu, co pozwoli mu ocenić potencjalne ryzyko i szanse, a tym samym podjęcie decyzji inwestycyjnych.

W poszukiwaniu odpowiedzi na postawione wyżej pytanie wykorzystamy metody symulacyjne oraz CTG. Poniżej przedstawiony jest schemat postępowania.

Schemat postępowania:

- Krok 1** Ze strony stooq.pl pobierz informację o dziennych cenach zamknięcia akcji: KGHM Polska Miedź S.A., PKN Orlen Spółka Akcyjna oraz PKO Bank Polski (PEO) (od 1 stycznia 2022 roku do 30 czerwca 2025 roku – łącznie 873 obserwacji).

[Tekst alternatywny. Zrzut ekranu przedstawiający sposób pozyskania danych ze strony stooq.pl.]

The screenshot shows the stooq.pl website. The top part displays a list of various stock indices under the 'Notowania' (Quotes) section. A blue arrow points from the 'Akcje' (Stocks) link in the left sidebar to the search bar. The bottom part shows the search results for 'KGHM Polska Miedź SA (KGH)' with a table of historical data. A blue arrow points from the 'Dane historyczne' link in the left sidebar to the 'Dane historyczne' section of the search results.

Wartości historyczne KGH

Nr.	Data	Otwarcie	Najwyższy	Najniższy	Zamknięcie	Zmiana	Wolumen
873	30 cze 2025	130.1	131	128.3	128.9	-0.08%	458,871
872	27 cze 2025	130	131.75	127	129	-0.77%	681,046
871	26 cze 2025	124.65	130	124.6	130	+5.01%	1,083,298
870	25 cze 2025	125.25	125.7	123.65	123.8	-0.80%	348,818
869	24 cze 2025	123	125.3	123	124.8	+3.57%	541,089

Rys. 4.5. Schemat postępowania w procedurze pozyskania danych ze strony stooq.pl



[Tekst alternatywny. Zrzut ekranu przedstawiający fragment tabeli dziennych kursów zamknięcia wybranych do analizy akcji. Linia przerywaną zaznaczono miejsce ukrytych wierszy.]

	A	B	C	D	E
1	Ip	Data	KGHM	PKN Orlen	PKO BP
2	1	03.01.2022	133,472	61,5418	90,8446
3	2	04.01.2022	136,475	62,5244	91,027
4	3	05.01.2022	139,193	64,7025	89,3873
873	872	27.06.2025	129	81,61	182
874	873	30.06.2025	128,9	82,1	185
875					

Rys. 4.6. Tabela dane dzienne kursy zamknięcia

Krok 2 Wyznacz proste stopy zwrotu, zgodnie ze wzorem $r_t = \frac{P_t - P_{t-1}}{P_{t-1}}$, gdzie P_t oznacza cenę zamknięcia.

[Tekst alternatywny. Zrzut ekranu przedstawiający fragment tabeli stóp zwrotu analizowanych akcji. Linia przerywaną zaznaczono miejsce ukrytych wierszy.]

	A	B	C	D	E
1	Ip	Data	KGHM	PKN Orlen	PKO BP
2	2	04.01.2022	2,25%	1,60%	0,20%
3	3	05.01.2022	1,99%	3,48%	-1,80%
871	871	26.06.2025	5,01%	0,90%	2,76%
872	872	27.06.2025	-0,77%	-1,33%	-0,08%
873	873	30.06.2025	-0,08%	0,60%	1,65%

Rys. 4.7. Tabela dane stopy zwrotu

Krok 3 Określ wielkość próby - w tym przykładzie założmy, że każda próbka składa się ze 100 dziennych zwrotów. W sposób losowy wybieramy 30 próbek po 100 dziennych zwrotów z oryginalnego zestawu danych.

Próby losujemy wykorzystując funkcje

- **LOS.ZAKR(2, 873)** – funkcja zwraca losową liczbę całkowitą z wybranego zakresu liczb;
- **ADRES()** – funkcja umożliwia uzyskanie adresu komórki w arkuszu, podając określony numer wiersza i kolumny (np. funkcja ADRES(1;3) zwraca adres \$C\$1;



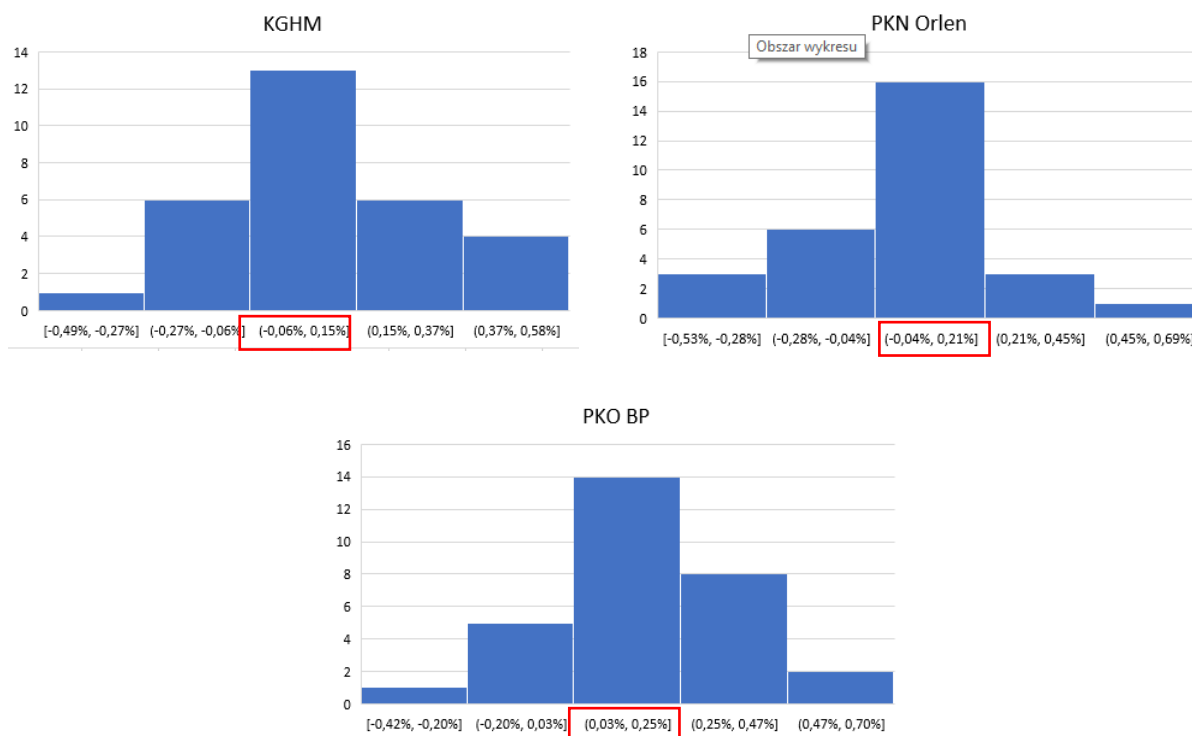
- **ADR.POŚR()** - służy do pozyskania wartości z komórki, o zadanym adresie.

[Tekst alternatywny. Zrzut ekranu przedstawia formułę umożliwiającą losowanie 30 prób, z których każda składa się ze 100 stóp zwrotu wybranych z dostępnej puli 873 obserwacji.]

ROZKŁ.NO... : X ✓ fx =ADR.POŚR(ADRES(LOS.ZAKR(2;873);3))											
	A	B	C	D	E	F	G	H	AH	AI	AJ
1	lp	Data	KGHM		lp	Próba 1	Próba 2	Próba 3	Próba 29	Próba 30	
2	2	04.01.2022	0,022499		1	=ADR.POŚR(ADRES(LOS.ZAKR(2;873);3))					
3	3	05.01.2022	0,019916		2	ADR.POŚR(adres_tekst; [a1])					
4	4	07.01.2022	-0,03048		3						
5	5	10.01.2022	-0,01944		4						
6	6	11.01.2022	0,030984		5						

Rys. 4.8. Schemat losowania próby

- Krok 4 Dla każdej próby obliczamy średnią dziennych stóp zwrotu
- Krok 5 Wykreśl histogramy dla średnich stóp zwrotu. Sprawdź, czy rozkład średnich z próby jest zbliżony do rozkładu normalnego. (Jest to kluczowy krok, ponieważ CTG zakłada, że średnie z próby będą zbiegać się do rozkładu normalnego, niezależnie od pierwotnego rozkładu populacji).



Rys. 4.9. Histogramy losowej próby stóp zwrotu badanych akcji

Każdy z histogramów kształtem przypomina rozkład normalny. Aby sformułować wnioski dotyczące oceny opłacalności inwestycji w badane akcje do przykładu wrócimy w rozdziale 5, w którym wyznaczymy podstawowe parametry statystyczne oraz przedziały ufności (umożliwiające ocenę niepewności) dla średnich dziennych stóp zwrotu oraz sprawdzimy czy średnia dzienna stopa zwrotu badanych akcji jest statystycznie większa niż 0?

Uzupełniając analizę statystyczną o podstawową wiedzę na temat analizowanych spółek, warto zauważyć, że Inwestor podejmując decyzję powinien wziąć pod uwagę następujące fakty:

- KGHM Polska Miedź - sektor: surowce (miedź, srebro, metale przemysłowe), profil ryzyka: wysoki; sektor silnie powiązana z globalnym popytem na surowce i Chinami, zyski zależą od cen miedzi i kursu USD/PLN, atrakcyjność inwestycji związana jest z rosnącym zapotrzebowaniem na miedź wynikającym z zielonej transformacji. KGHM może być dobrą inwestycją długoterminową przy wzroście cen surowców;



- PKN Orlen - sektor: paliwa, petrochemia, energia, profil ryzyka: średni, stabilne zyski, dywersyfikacja działalności (gaz, prąd, paliwa alternatywne). Ryzyko inwestycji związane jest między innymi z ryzykiem decyzji politycznych, ryzykiem ciągłości dostaw, inwestowaniem w kolejne obszary działalności, konkurencją, zmiennością marż rafineryjnych, niedostosowaniem się Grupy ORLEN do tempa zmian technologicznych;
- PKO BP - sektor: bankowy, profil ryzyka: umiarkowany, sektor zależny od stóp procentowych i gospodarki, wydaje się najatrakcyjniejszą opcją na najbliższe 6–12 miesięcy, ale obecnie to raczej spekulacyjna gra.

Korzyści i ograniczenia centralnego twierdzenia granicznego (nie tylko dla inwestorów)

Korzyści

- W praktyce – w szczególności w ekonomii i finansach – **nie znamy rzeczywistego rozkładu populacji**. Jeśli nie wiemy nic (lub prawie nic) o kształcie rozkładu zmiennej losowej „CTG zasługuje na uwagę, ponieważ stwierdza zmierzanie rozkładu średniej z próby do rozkładu normalnego niezależnie od rozkładu w populacji, z której próba została pobrana. Z twierdzenia tego wynikają probabilistyczne stwierdzenia o tym, jaki jest zakres, w którym średnia z próby powinna się mieścić. Pozwala ustalić prawdopodobieństwo odbiegania wartości średniej z próby na określoną odległość od średniej populacji, której średnia z próby jest estymatorem¹²” (Aczel, 2017).
- CTG pozwala twierdzić: *„średnia z próby jest bliska średniej populacji – z dużym prawdopodobieństwem i przy znanym błędzie estymacji”* (tj. dostarcza wiarygodnych szacunków dzięki czemu umożliwia lepszą ocenę ryzyka i pomaga zmniejszyć niepewność.)

Ograniczenia

- Wymagana dotyczące odpowiednio dużej liczebności próby.
- Założenie niezależności zmiennych losowych, co w praktyce implikuje losowy dobór prób, jednakże np. w przypadku szeregów czasowych (tj. stopy zwrotu) między kolejnymi obserwacjami może występować **autokorelacja**.
- CTG zakłada, że zmienna losowa ma **skończoną wariancję** (jeśli analizujemy dane z rozkładów o **nieskończonej wariancji** (jak np. rozkład Cauchy’ego, czy

¹² Pojęcie estymatora zostało omówione na wykładzie.



niektóre przypadki rozkładów stóp zwrotu modelowanych jako procesy z α -stabilnymi rozkładami).

W sytuacji naruszenia wymaganych założeń CTG rozkład średnich z prób niekoniecznie będzie zbliżony do normalnego, co może mieć wpływ na wiarygodność wyników.

Należy podkreślić, że CTG nie jest narzędziem do identyfikowania rozkładu populacji stwierdza jedynie, że **rozkład średnich z prób** (czyli estymatora średniej) **zbliża się do normalnego** – a nie, że analizowane dane opisywane są przez rozkład normalny.

5. Elementy wnioskowania statystycznego

5.1. Istota wnioskowania statystycznego

Wnioskowanie statystyczne stanowi fundament nowoczesnych badań empirycznych – niezależnie od tego, czy analizujemy stabilność modeli makroekonomicznych, czy badamy preferencje konsumenckie lub skuteczność terapii medycznej.

Polega na uogólnianiu wniosków (wnioskowaniu) o całej zbiorowości statystycznej na podstawie znajomości tylko części badanej zbiorowości (tj. próby), przy jednoczesnym zachowaniu świadomości niepewności (Balicki, Makać, 2010).

Istotnym elementem wnioskowania statystycznego jest sposób doboru próby. W tym opracowaniu nie będziemy zajmować się tym zagadnieniem, szczegóły można znaleźć w pozycjach umieszczonych w literaturze oraz w podręcznikach poświęconych metodzie reprezentacyjnej.

Ważne jest, aby próba była pobierana losowo z całej populacji w taki sposób, że wszystkie jednostki populacji generalnej mają jednakowe prawdopodobieństwo dostania się do próby i prawdopodobieństwo to nie zmienia się w trakcie losowania. Taką próbę nazywamy **próbą losową prostą**.

Metody wnioskowania statystycznego obejmują:

- (1) estymację parametrów zbiorowości generalnej,
- (2) weryfikację hipotez statystycznych.

Zaletą wnioskowania statystycznego jest możliwość określenia błędu, z jakimi zostały sformułowane uogólnione wnioski.



Na wykładzie zostały podane i omówione pojęcia związane z działem wnioskowania statystycznego, takie jak: statystyka, estymator, własności estymatorów, przedział ufności, poziom ufności i istotności oraz schemat weryfikacji hipotez statystycznych. W opracowaniu skoncentrujemy się na przeprowadzeniu serii przykładów z pełnym rozwiązaniem wykorzystując arkusz kalkulacyjny MS Excel oraz wzory podane w rozdziale 7.

5.2. Estymacja przedziałowa

Pojęcie przedziału ufności zostało wprowadzone do statystyki przez matematyka i statystyka polskiego pochodzenia Jerzego Sławę-Neymana¹³ (1894-1981). W wyniku pobrania próby losowej z populacji i obliczenia na tej podstawie wartości estymatora szacowanego parametru populacji θ uzyskujemy tzw. **estymator punktowy** parametru $\hat{\theta}$. Estymator punktowy to zmienna losowa, której konkretna realizacja (np. średnia z próby jest estymatorem średniej w populacji) zależy od doboru jednostek w próbie. W związku z tym, każda pojedyncza wartość estymatora, którą uzyskujemy w badaniu, jest tylko jedną z wielu możliwych – a prawdopodobieństwo tego, że dokładnie pokrywa się z rzeczywistym, nieznanym parametrem populacji, jest równe zero (w sensie ciągłym). Z tego powodu w praktyce wnioskowania statystycznego stosuje się estymację przedziałową, której celem jest określenie takiego przedziału liczbowego, który – z określonym prawdopodobieństwem, tj. poziomem ufności (np. 95%) – zawiera nieznaną wartość parametru populacyjnego. Estymacja przedziałowa nie tylko dostarcza przybliżenia, ale również informuje o precyzji i niepewności tego oszacowania, umożliwiając interpretację wyniku w kategoriach probabilistycznych.

Podstawowymi parametrami, które szacowane są dla populacji generalnej są: wartość oczekiwana (średnia), wariancja populacji oraz wskaźnik struktury. Do wyznaczenia przedziałów ufności konieczna jest znajomość estymatorów punktowych, które stanowią podstawę do określenia granic przedziału ufności, a ich rozkład jest kluczowy dla określenia poziomu ufności. W związku z powyższym zaprezentowano je w tabeli poniżej.

Tabela 5.1 Nieobciążone estymatory punktowe parametrów populacji

¹³ Sława-Neyman jest uważany za jednego z ojców nowoczesnej statystyki.

Zainteresowanym polecam artykuł Profesora Czesława Domańskiego „Początki twórczości naukowej Jerzego Sławy-Neymana”, Wiadomości Statystyczne. The Polish Statistician, vol. 66, 2021, 12, s. 98-113.



Parametr	Estymator	Średni błąd szacunku
Wartość oczekiwana m	$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$	σ - znane $D(\bar{x}) = \frac{\sigma}{\sqrt{n}}$
		σ - nieznane $D(\bar{x}) = \frac{s}{\sqrt{n}}$
Odchylenie standardowe σ	m – znane $s = \frac{1}{n} \sum_{i=1}^n (x_i - m)^2$	-
	m – nieznane $s = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$	
Wskaźnik struktury p	Częstość względna w próbie $\hat{p} = \frac{k}{n}$	$D(\hat{p}) = \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$

Przykład 5.1

Rozważmy dane z przykładu 4.3. Poszukując odpowiedzi, w które akcje warto zainwestować dla każdej z akcji wyznaczmy przedziały ufności oraz błąd szacunku

- dla średnich stóp zwrotu;
- dla frakcji dodatnich stóp zwrotu¹⁴.

Na podstawie prób losowych składających się z $n = 200$ dziennych stóp zwrotu z każdej akcji uzyskano następujące estymatory punktowe parametrów populacji

Tabela 5.2 Podstawowe statystyki wyznaczone na podstawie próby (dane w %)

Statystyki	KGHM	PKN Orlen	PKO BP
Średnia	0,15	0,29	0,32
Odchylenie standardowe s	2,38	2,00	2,23
Wskaźnik struktury $\hat{p}$	51,00	56,50	56,00
Współczynnik zmienności V	1586,67	689,66	696,88

Z powyższej tabeli wynika, że w analizowanej próbie akcje PKO BP odnotowały najwyższą średnią dzienną stopę zwrotu, co odpowiada najwyższej oczekiwanej stopie zysku. Wynik ten jest spójny z prezentacją graficzną (histogram na rys. 4.9). Wszystkie

¹⁴ Frakcja (inaczej wskaźnik struktury, proporcja) dodatnich stóp zwrotu dla danej akcji to odsetek okresów (w analizowanym przykładzie dni notowań), w których stopa zwrotu była większa od zera.



analizowane instrumenty charakteryzowały się wysoką zmiennością (mierzoną odchyleniem standardowym), przy czym największe wahania zaobserwowano w przypadku akcji KGHM. Najniższą zmienność, przy relatywnie wysokim udziale okresów z dodatnią stopą zwrotu, wykazały akcje PKN Orlen. We wszystkich przypadkach udział dni z dodatnią stopą zwrotu przekroczył 50%, co wskazuje na przewagę sesji wzrostowych nad spadkowymi.

Przedziały ufności zbudujemy zakładając poziom ufności $1 - \alpha = 0,95$. Dodatkowo przeprowadzimy ocenę precyzji dokonanego szacunku obliczając błąd szacunku.

Uwaga

W każdym z przypadków wzór na przedział ufności dobieramy uwzględniając fakt, że nie znamy rozkładu populacji i analizowana próba jest duża ($n = 200$).

Ad a) W związku z powyższą uwagą do wyznaczenia przedziałów ufności dla średnich stóp zwrotu wykorzystamy wzór

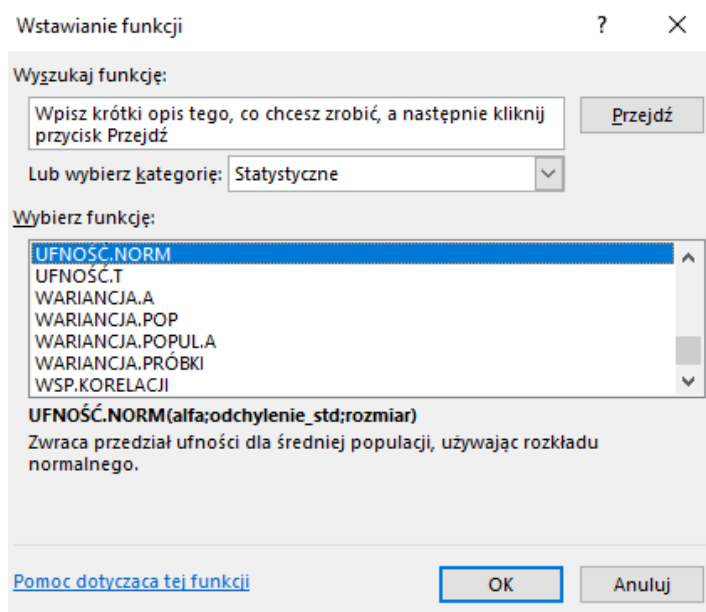
$$\bar{x} - u_{1-\alpha/2} \frac{s}{\sqrt{n}} < m < \bar{x} + u_{1-\alpha/2} \frac{s}{\sqrt{n}} \quad (5.1)$$

gdzie $u_{1-\alpha/2}$ jest kwantylem rozkładu normalnego wyznaczanym za pomocą funkcji statystycznej **ROZKŁAD.NORMALNY.S.ODW()**.

Wygodnym (szybszym) sposobem wyznaczania przedziałów ufności w arkuszu kalkulacyjnym MS Excel jest wykorzystanie funkcji **UFNOŚĆ()** z uwzględnieniem właściwego rozkładu dostępnej w oknie dialogowym funkcji, kategoria Statystyczne.



[Tekst alternatywny. Zrzut ekranu. Wywołanie funkcji statystycznej umożliwiającej wyznaczenie kwantyla rozkładu normalnego - **UFNOŚĆ.NORM()**.]



Rys. 5.1. Okno dialogowe **Wstawianie funkcji**, kategoria: statystyczne

Funkcja **UFNOŚĆ.NORM(alfa; odchylenie_std; rozmiar)**

zwraca wartość błędu oszacowania średniej w populacji (bezwzględny błąd oszacowania). Określa połowę długości przedziału ufności (5.1), zadana jest wzorem

$$Ufność = d(\bar{x}) = u_{1-\alpha/2} \frac{s}{\sqrt{n}} \quad (5.2)$$

Argumenty funkcji określają odpowiednio:

- alfa - poziom istotności używany do obliczania poziomu ufności.
- odchylenie_std - odchylenie standardowe dla zakresu danych,
- rozmiar - wielkość próby.

Na rysunku poniżej przedstawiono dwa pomocnicze zrzuty ekranu, prezentujące funkcję i formuły wykorzystywane w kolejnych etapach, wyznaczania przedziałów ufności dla średnich stóp zwrotu analizowanych akcji.

[Tekst alternatywny. Zrzut ekranu. Wyznaczenie przedziałów ufności dla średniej.]



	AC	AD	AE	AF	AG	AH
1						
2						
3			KGHM	PKN Orlen	PKO BP	
4		średnia z próby	0,15	0,29	0,32	
5		odchylenie standardowe z próby	2,38	2	2,23	
6		poziom ufności	0,95	0,95	0,95	
7		n	200	200	200	
8		ufność	=UFNOŚĆ.NORM(1-AE6;AE5;AE7)			
9		przedział ufności dla średniej	UFNOŚĆ.NORM(alfa; odchylenie_std; rozmiar)			
10		lewy koniec przedziału	-0,18	0,01	0,01	
11		prawy koniec przedziału	0,48	0,57	0,63	

	AC	AD	AE	AF	AG	AH
1						
2						
3			KGHM	PKN Orlen	PKO BP	
4		średnia z próby	0,15	0,29	0,32	
5		odchylenie standardowe z próby	2,38	2	2,23	
6		poziom ufności	0,95	0,95	0,95	
7		n	200	200	200	
8		ufność	0,330	0,277	0,309	
9		przedział ufności dla średniej				
10		lewy koniec przedziału	=AE4-AE8	0,01	0,01	
11		prawy koniec przedziału	=AE4+AE8	0,57	0,63	

Rys. 5.2. Rysunek pomocniczy do przykładu 5.1.a

Ze względu na fakt, że w komórkach: lewy i prawy koniec przedziału dla średniej stopy zwrotu akcji KGHM (m_{KGHM}) podane są formuły, ostateczne wartości podamy poniżej

$$-0,18 < m_{KGHM} < 0,48$$

Można zatem przypuszczać, (z prawdopodobieństwem 0,95), że średnia dzienna stopa zwrotu akcji KGHM zawiera się w przedziale od $-0,18\%$ do $0,48\%$, z błędem szacunku (ufność) 0,33%.

Wnioski praktyczne

Jeżeli przedział ufności zawiera zero (KGHM), nie ma podstaw do stwierdzenia, że średnia stopa zwrotu jest dodatnia w ujęciu statystycznym (na poziomie ufności 95% nie można wykluczyć, iż w populacji średnia dzienna stopa zwrotu KGHM była ujemna (potencjalna strata).

W przypadku PKN Orlen i PKO BP cały przedział znajduje się powyżej zera, co wskazuje na statystyczne potwierdzenie dodatniej wartości oczekiwanej stopy zwrotu



w badanym okresie.

Przedział KGHM sugeruje większe ryzyko realizacji ujemnych średnich stóp zwrotu, natomiast PKO BP prezentuje wyższą dolną granicę i mniejszą niepewność co do kierunku zysku.

Ad b) W celu głębszego zbadania opłacalności inwestycji wyznaczmy również przedział ufności dla frakcji. W tym celu wykorzystamy wzór

$$\hat{p} - u_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + u_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad (5.3)$$

Podobnie jak wyżej sposób wyznaczania przedziałów ufności przedstawimy na rysunku prezentując zrzuty ekranu z wykorzystywanymi funkcjami i formułami.

	AC	AD	AE	AF	AG	AH
19						
20			KGHM	PKN Orlen	PKO BP	
21		n	200	200	200	
22		frakcja elementów z próby $\hat{p}$	0,51	0,57	0,56	
23		poziom ufności $(1 - \alpha)$	0,95	0,95	0,95	
24		kwantyl $u_{1-\alpha/2}$	=ROZKŁ.NORMALNY.S.ODWR(1-(1-AE23)/2)			
25		przedział ufności dla średniej	ROZKŁ.NORMALNY.S.ODWR(prawdopodobieństwo)			
26		lewy koniec przedziału	0,44	0,50	0,49	
27		prawy koniec przedziału	0,58	0,63	0,63	

	AC	AD	AE	AF	AG	AH
19						
20			KGHM	PKN Orlen	PKO BP	
21		n	200	200	200	
22		frakcja elementów z próby $\hat{p}$	0,51	0,57	0,56	
23		poziom ufności $(1 - \alpha)$	0,95	0,95	0,95	
24		kwantyl $u_{1-\alpha/2}$	1,96	1,96	1,96	
25		przedział ufności dla średniej				
26		lewy koniec przedziału	=AE22-AE24*PIERWIASTEK(AE22*(1-AE22)/AE21)			
27		prawy koniec przedziału	=AE22+AE24*PIERWIASTEK(AE22*(1-AE22)/AE21)			

Rys. 5.3. Rysunek pomocniczy do przykładu 5.1.b

Na podstawie przeprowadzonej analizy, otrzymaliśmy granice losowego przedziału pokrywającego nieznaną, rzeczywistą frakcję tych dni (w populacji wszystkich notowań), w których stopa zwrotu jest dodatnia z prawdopodobieństwem 95%. Badana frakcja mieści się w granicach 44–58% dla akcji KGHM, 50–63% dla PKN Orlen oraz 49–63% dla PKO BP.



Przedział ufności dla frakcji w przypadku akcji PKN Orlen, wskazuje, że wzrosty kursów występowały częściej niż spadki. Dodatkowo górna granica przedziału ufności w przypadku tych akcji wykazywała najwyższą wartość sugerując większą regularność sesji wzrostowych w porównaniu z pozostałymi akcjami uwzględnionymi w badaniu.

Z przeprowadzonej analizy możemy wnioskować, że najmniej opłacalną i najbardziej ryzykowną inwestycją będzie inwestycja w akcje KGHM.

Należy jednak zwrócić uwagę, że

- Częstotliwość dni, w których stopa zwrotu jest dodatnia sama w sobie nie oznacza opłacalności — należy uwzględnić także wielkość stóp zwrotu, ich zmienność i koszty transakcyjne.
- Rozpiętość przedziałów ufności odzwierciedla niepewność estymacji - większa niepewność decyzji (niepewność ta m.in. zależy od rozmiaru próbki).

Problem minimalnej liczebności próby

W praktyce bardzo często badacz staje przed problemem jak duża powinna być próba losowa, aby zachowana została z góry ustalona precyzja oszacowania.

Żądania w zakresie precyzji oszacowania można wyrazić jako maksymalny dopuszczalny błąd bezwzględny (d). W przypadku szacowania średniej (przedział ufności określony wzorem (5.1)) błąd bezwzględny zadany jest wzorem (5.2). Stąd ustalając maksymalny błąd szacunku na poziomie d , przy danym z góry poziomie ufności $(1 - \alpha)$ liczebność próby wyznaczana jest na podstawie relacji

$$u_{1-\alpha/2} \frac{s}{\sqrt{n}} \leq d$$

Rozwiązując powyższą nierówność ze względu n otrzymujemy wzór na liczebność próby

$$n \geq \left(\frac{s u_{1-\alpha/2}}{d} \right)^2 \quad (5.4)$$

Wielkość $n = \left(\frac{s u_{1-\alpha/2}}{d} \right)^2$ jest **minimalną liczebnością próby**, która zapewnia, że maksymalny błąd szacunku nie przekroczy ustalonej precyzji d .

Przeprowadzając analogiczne rozumowanie uzyskujemy wzór na liczebność próby dla szacowania proporcji p z bezwzględną dokładnością d przy prawdopodobieństwie $(1 - \alpha)$ niepopelnienia błędu większego niż założony



$$n \geq \frac{u_{1-\alpha/2}^2 \hat{p}(1 - \hat{p})}{d^2} \quad (5.5)$$

Zauważmy, że dla każdego $\hat{p} \in (0, 1)$ zachodzi relacja

$$\hat{p}(1 - \hat{p}) \leq \frac{1}{2} \left(1 - \frac{1}{2}\right) = \frac{1}{4}$$

Zatem dla każdego $\hat{p}$

$$d \leq \frac{u_{1-\alpha/2}}{2\sqrt{n}}$$

Aby otrzymać przedział z maksymalnym błędem d na zadanym poziomie ufności $(1 - \alpha)$ należy wziąć próbę losową o liczebności

$$n \geq \left(\frac{u_{1-\alpha/2}}{2d}\right)^2 \quad (5.6)$$

Wykorzystanie wzoru (5.6) szczególnie poleca się w przypadku, gdy zachodzi obawa co do poprawności przyjęcia za p oszacowania $\hat{p}$.

Przykład 5.2

Dla danych z przykładu 5.1 wyznaczmy minimalną liczebność próby, tak aby maksymalny błąd szacunku, na poziomie ufności 0,95

- dla średniej m wynosił 0,1 punktu procentowego.
- dla parametru p wynosił 0,04.

Ad a) W celu wyznaczenia minimalnej liczebności próby, obliczenia przeprowadzimy w arkuszu kalkulacyjnym MS Excel. Dane, wykorzystaną formułę oraz wyniki przedstawiono na rys. 5.4.

[Tekst alternatywny. Zrzut ekranu, prezentujący dane oraz funkcję służącą do wyznaczenia minimalnej liczebności próby, która gwarantuje oszacowanie średniej na zadanym poziomie ufności oraz z ustalonym maksymalnym błędem szacunku.]



	R	S	T	U	V	W
22						
23			KGHM	PKN Orlen	PKO BP	
24		średnia	0,15	0,29	0,32	
25		odchylenie standardowe	2,28	2,00	2,23	
26		poziom ufności	0,95	0,95	0,95	
27		kwantyl $u_{1-\alpha/2}$	1,96	1,96	1,96	
28		d	0,10	0,10	0,10	
29		liczebność minimalna formuła	=ZAOKR.DO.CAŁK((T25*T27/T28)^2)			
30			ZAOKR.DO.CAŁK(liczba)			
31		liczebność minimalna formuła	1996	1536	1910	

Rys. 5.4. Zrzut ekranu

Na podstawie powyższych obliczeń możemy stwierdzić, że dla uzyskania estymacji średniej dziennej stopy zwrotu z błędem maksymalnym nieprzekraczającym 0,1 punktu procentowego, konieczne jest wylosowanie co najmniej odpowiednio: 1996 obserwacji dla akcji KGHM, 1536 obserwacji dla akcji PKN Orlen oraz 1910 obserwacji dla akcji PKO BP.

W konsekwencji, w celu zwiększenia rzetelności oraz mocy inferencyjnej wniosków przedstawionych w przykładzie 5.1, zasadne byłoby powtórzenie zaprezentowanych obliczeń z wykorzystaniem prób o istotnie większej liczebności.

Ad b)

Na rysunku 5.5 przedstawiono dane, wykorzystaną formułę oraz uzyskane oszacowania minimalnej liczebności próby, która gwarantuje szacowanie frakcji dni, w których stopa zwrotu jest dodatnia z ustaloną precyzją.



[Tekst alternatywny. Zrzut ekranu. Wyznaczanie minimalnej liczebności próby dla frakcji.]

	Z	AA	AB	AC	AD	AE
1						
2			KGHM	PKN Orlen	PKO BP	
3		frakcja $\hat{p}$	0,51	0,57	0,56	
4		poziom ufności	0,95	0,95	0,95	
5		kwantyl $u_{1-\alpha/2}$	1,96	1,96	1,96	
6		d	0,04	0,04	0,04	
7		liczebność minimalna formuła	=ZAOKR.DO.CAŁK($AB5^2 \cdot AB3 \cdot (1-AB3) / AB6^2$)			
8			ZAOKR.DO.CAŁK(liczba)			
9		liczebność minimalna	600	588	591	

Rys. 5.5. Zrzut ekranu

Zauważmy, iż dla oszacowania frakcji przy zadanym poziomie ufności oraz określonej precyzji wymagana jest istotnie mniejsza liczebność próby niż w przypadku estymacji średnich stóp zwrotu. Niemniej jednak, obliczona minimalna liczebność w tym przypadku jest trzykrotnie większa od liczebności próby zastosowanej w przykładzie 5.1. Tym samym, wyniki te dodatkowo potwierdzają zasadność powtórzenia wspomnianych obliczeń w oparciu o próby o znacznie większej liczebności.

5.2. Weryfikacja wybranych hipotez statystycznych

Zacznijmy od przypomnienia definicji hipotezy statystycznej.

Hipotezą statystyczną nazywamy dowolne przypuszczenie (sąd) odnoszące się do parametru rozkładu lub jego postaci. Hipotezy mogą dotyczyć np. wartości średniej, wariancji, odchylenia standardowego, udziału (proporcji), współczynnika korelacji czy też samego kształtu rozkładu badanej cechy. Wspomniane przypuszczenie może być sformułowane na podstawie teorii badanego zjawiska, w oparciu o informacje zawarte w próbie, ale także na podstawie intuicji lub wręcz *ad hoc*.

Z decyzją o przyjęciu lub odrzuceniu hipotezy wiąże się **ryzyko popełnienia błędu**.

Przypomnijmy błąd polegający na odrzuceniu hipotezy zerowej (H_0), która



w rzeczywistości jest prawdziwa nazywamy **błędem pierwszego rodzaju** (i oznaczamy symbolem α). Natomiast błąd polegający na przyjęciu hipotezy zerowej, gdy jest fałszywa nazywamy **błędem drugiego rodzaju**. Ryzyko popełnienia tych błędów można jednak z góry określić. W większości przypadków przeprowadzając weryfikację hipotez uwzględnia się prawdopodobieństwo popełnienia błędu pierwszego rodzaju.

Kolejność działań przy weryfikacji hipotez statystycznych:

1. sformułować hipotezę zerową (H_0) i alternatywną (H_1)
np. $H_0: m = 0$, $H_1: m > 0$;
2. ustalić postać statystyki testowej i obliczyć jej wartość na podstawie próby
wzory statystyk zostały podane w rozdziale 7;
3. określić poziom istotności α (najczęściej: 0,1; 0,05; 0,01);
4. wyznaczyć obszar odrzucenia hipotezy zerowej (obszar krytyczny) lub krytyczny poziom istotności (nazywany też wartością-p lub p-value);
5. podjęcie decyzji (z określonym prawdopodobieństwem błędu)
 - **w oparciu o obszar krytyczny:**
jeżeli wyznaczona na podstawie próby wartość statystyki testowej należy do obszaru krytycznego wówczas należy odrzucić hipotezę zerową jako nieprawdziwą i wnioskować, że hipoteza alternatywna jest prawdziwa. W przeciwnym przypadku należy stwierdzić brak podstaw do odrzucenia hipotezy zerowej.
 - **w oparciu o wartość-p**
wartość – p jest to prawdopodobieństwo, które określa najniższy poziom istotności, przy którym można odrzucić hipotezę zerową (tj. szansę, przy prawdziwej hipotezie zerowej otrzymania wartości statystyki testowej co najmniej tak dużej, jak ta uzyskana z próby,
np. niech U_{obl} będzie wartością odpowiedniej statystyki testującej
wówczas wartość $-p = P(U \geq U_{obl})$.

Decyzja:

Jeśli wartość $-p \leq \alpha$, to odrzucamy H_0 .

Jeśli wartość $-p > \alpha$, to nie ma podstaw do odrzucenia H_0

(gdy wartość $-p$ jest duża, to sprawdzana hipoteza wydaje się wiarygodna.

Oznacza to niestety tylko tyle, że hipoteza zerowa może być prawdziwa, a nie że jest prawdziwa).

Testowanie hipotez jest jednym z najważniejszych narzędzi w analizie danych - uzyskane wyniki mają wpływ na podejmowanie decyzji w wielu dziedzinach.



Omawiając testy istotności rozpatrzmy kilka przypadków w zależności od posiadanych informacji na temat populacji generalnej.

Przykład 5.3

Wykorzystując dane z przykładu 5.2 inwestor sądzi się, że inwestycja w akcje PKO BP średnio w każdym dniu przynosiła zysk na poziomie 0,32%. Z losowo wybranych 600 dni uzyskano średnią stopę zwrotu 0,11% z odchyleniem standardowym 2,40%.

Czy przypuszczenie inwestora jest zasadne? Przyjmijmy poziom istotności $\alpha = 0,1$.

Rozwiązanie

Dysponujemy następującymi informacjami

$n = 600$; $\bar{x} = 0,11\%$, $s = 2,40\%$, badana próba jest duża, nie znamy rozkładu badanej cechy

Stawiamy hipotezy

$$H_0: m = 0,32$$

$$H_1: m \neq 0,32$$

W oparciu o wzory w rozdziale 7 wybieramy statystykę testową $U = \frac{\bar{x} - \mu_0}{s} \sqrt{n}$.

Po podstawieniu zebranych danych otrzymujemy

$$U_{obl} = \frac{0,11 - 0,32}{2,40} \sqrt{600} = -2,14$$

Przy przyjętym poziomie istotności 0,1 oraz dwustronnym obszarze krytycznym $K = (-\infty, -u_{1-\alpha/2}) \cup (u_{1-\alpha/2}, \infty)$ końce obszaru K określone są przez kwantyl rozkładu normalnego $u_{1-0,1/2}$, którego wartość odczytywana jest z tablicy dystrybuanty rozkładu normalnego, lub za pomocą funkcji ROZKŁ.NORMALNY.S.ODW(0,95) ($u_{1-0,1/2} = 1,64$). Obliczona wartość testu $U_{obl} = -2,14$ znajduje się w obszarze krytycznym, co powoduje odrzucenie hipotezy zerowej. W praktyce oznacza to, że inwestor nie ma racji, można bowiem sądzić, z 90% prawdopodobieństwem, że badana stopa zwrotu jest różna od 0,32%.

Uzyskany wynik zachęca do dalszego badania i sprawdzenia, czy badana stopa jest mniejsza od 0,32%.



Fundusze Europejskie
dla Rozwoju Społecznego



Rzeczpospolita
Polska

Dofinansowane przez
Unię Europejską



W tym celu modyfikujemy hipotezę alternatywną przyjmując:

$$H_0: m = 0,32$$

$$H_1: m < 0,32$$

Wartość statystyki testowej nie ulega zmianie.



Zmienia się natomiast postać obszaru krytycznego $K = (-\infty, -u_{1-\alpha})$, $-u_{1-\alpha} = -1,28$. Ponownie wartość statystyki należy do obszaru krytycznego, w związku z tym na przyjętym poziomie istotności możemy twierdzić, że średnia dzienna stopa zwrotu akcji PKO BP jest niższa od 0,32%.

Ze względu na fakt, że z dotychczasowej analizy nie wynika jednoznacznie czy bardziej opłacalną inwestycją będzie zakup akcji PKO BP czy PKN Orlen w kolejnym przykładzie skoncentrujemy się na sprawdzeniu statystycznej istotności różnicy między średnimi stopami zwrotu z obu akcji.

Przykład 5.4.

Zgodnie z powyższym opisem, w przykładzie będziemy kontynuować problem, przed którym stoi inwestor. Celem analizy jest porównanie wartości oczekiwanych z dwóch niezależnych populacji o nieznanym rozkładzie. Przyjmijmy, podobnie jak w poprzednim przykładzie poziom istotności równy 0,1.

Rozwiązanie

Stawiamy hipotezy

$$H_0: m_1 = m_2$$

$$H_1: m_1 \neq m_2$$

gdzie $m_1 = m_{PKO BP}$, $m_2 = m_{PKN Orlen}$.

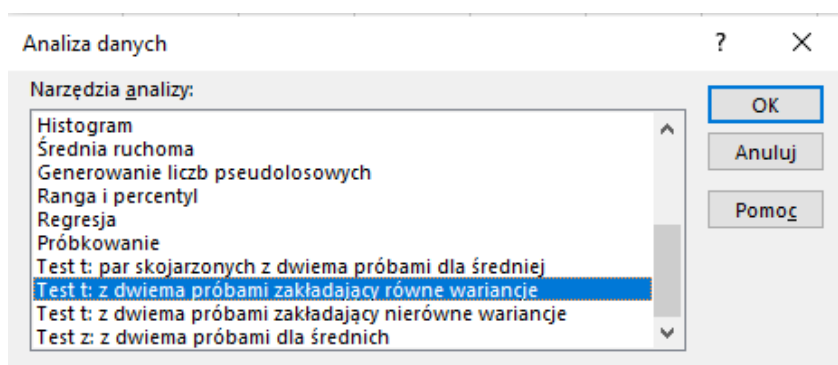
W celu weryfikacji tej hipotezy zerowej z każdej populacji wylosowano próbę losową o liczebności $n_1 = n_2 = 600$. Na podstawie prób wyznaczono średnie i odchylenia standardowe przedstawione w tabeli 5.3.

Tabela 5.3. Dane [%]

	PKN Orlen	PKO BP
średnia	-0,03	0,07
odchylenie standardowe	2,11	2,56

Przeprowadzenie weryfikacji powyższej hipotezy zerowej wymaga przeprowadzenia testu równości wariancji w badanych próbach.

[Tekst alternatywny. Zrzut ekranu. Okno dialogowe opcji Analizy danych.]



Rys. 5.6. Okno dialogowe Analiza danych – wybór odpowiedniego testu średnich

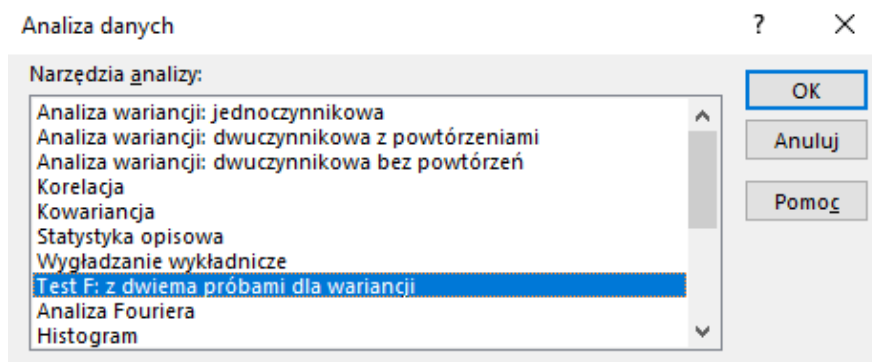
Stawiamy hipotezy

$$H_0: \sigma_1 = \sigma_2$$

$$H_1: \sigma_1 > \sigma_2$$

Weryfikacji wariancji dokonamy na podstawie **testu F z dwiema próbami**, który dostępny jest w Analizie danych.

[Tekst alternatywny. Zrzut ekranu. Okno dialogowe opcji Analizy danych.]



Rys. 5.6. Okno dialogowe Analiza danych – wybór odpowiedniego testu średnich

W ramach pracy samodzielnej zachęca się Studenta do przeprowadzenia prezentowanych w opracowaniu analiz na samodzielnie wybranym zbiorze danych.

[Tekst alternatywny. Zrzut ekranu. Okno dialogowe Testu F: z dwiema próbami dla wariancji oraz uzyskane wyniki]



	G	H	I	J	K	L	M	N	O	P	Q	R	S
1													
2		PKN Orlę	PKO BP										
3		-1,09%	0,87%										
4		0,51%	-2,93%										
5		2,25%	1,45%										
6		-1,42%	2,61%										
7		0,48%	1,74%										
8		-2,87%	-3,04%										
9		-1,33%	2,89%										
10		1,35%	-4,30%										
11		0,92%	1,77%										
12		-0,43%	1,55%										

	PKO BP	PKN Orlen
Średnia	0,07%	-0,03%
Wariancja	0,07%	0,04%
Obserwacje	600,00	600,00
df	599,00	599,00
F	1,47	
P(F<=f) jednostron	0,00	
Test F jednostronn	1,11	

Rys. 5.7. Zrzut ekranu

Z przeprowadzonej analizy wynika, że zmienność (wariancja) dziennych stóp zwrotu akcji PKO BP jest istotnie statystycznie większa od zmienności dziennych stóp zwrotu akcji PKN Orlen.

Z tego względu do weryfikacji hipotezy o równości średnich zastosujemy **test t: z dwiema próbami zakładającymi nierówne wariancje**

[Tekst alternatywny. Zrzut ekranu. Okno dialogowe Testu t: z dwiema próbami zakładającymi nierówne wariancje oraz uzyskane wyniki]



	G	H	I	J	K	L	M	N	O	P	Q	R
573	0,40%	0,43%										
574	-0,12%	-1,67%										
575	0,33%	0,10%										
576	-0,96%	-2,89%										
577	0,69%	2,39%										
578	-2,40%	-1,49%										
579	0,90%	0,25%										
580	0,61%	-0,49%										
581	0,51%	1,99%										
582	1,59%	-1,61%										
583	1,16%	-1,52%										
584	0,86%	1,67%										
585	2,62%	2,75%										
586	1,37%	-0,23%										

Test t: z dwiema próbami zakładający nierówne wariancje

Wejście
Zakres zmiennej 1: SH52:SH5602
Zakres zmiennej 2: SG52:SG5602
Różnica średnich wg hipotezy: 0
☒ Tytuły
Alfa: 0,1
Opcje wyjścia
☒ Zakres wyjściowy: SQ5575
☐ Nowy arkusz:
☐ Nowy skoroszyt

OK
Anuluj
Pomoc

	PKO BP	PKN Orlen
Średnia	0,07%	-0,03%
Wariancja	0,07%	0,04%
Obserwacje	600	600
Różnica średnich	0	
df	1156	
t Stat	0,74	
P(T<=t) jednostron	0,23	
Test T jednostron	1,28	
P(T<=t) dwustron	0,46	
Test t dwustronny	1,65	

Rys. 5.8. Zrzut ekranu



Przeprowadzona analiza statystyczna z wykorzystaniem testu t-Studenta dla prób niezależnych (wariant Welcha, niewymagający założenia homogeniczności wariancji) nie dostarczyła podstaw do odrzucenia hipotezy zerowej zakładającej równość średnich, dziennych stóp zwrotu akcji spółek PKN Orlen oraz PKO BP w badanym okresie od 4 stycznia 2022 roku do 30 czerwca 2025 roku.

Otrzymana wartość statystyki testowej oraz odpowiadająca jej wartość p (0,23 i 0,46, odpowiednio dla testu jednostronnego i dwustronnego) wskazują, iż obserwowane różnice średnich mogą być wynikiem przypadkowych wahań rynkowych, a nie przejawem istotnej przewagi jednej spółki nad drugą. Z punktu widzenia wnioskowania inwestycyjnego oznacza to brak statystycznie potwierdzonej przesłanki, aby preferować inwestycję w akcje jednej z analizowanych spółek wyłącznie na podstawie ich średnich, dziennych stóp zwrotu w rozpatrywanym horyzoncie czasowym.

Podsumowując, warto podkreślić, iż omawiane procedury analityczne mają charakter uniwersalny i mogą być z powodzeniem zastosowane do różnorodnych zbiorów danych. Dlatego też, w ramach pracy własnej, zachęcam studentów do przeprowadzenia analogicznych analiz na danych rzeczywistych odnoszących się do zagadnień, które są dla nich szczególnie interesujące. Takie podejście pozwoli nie tylko utrwalić wiedzę z zakresu statystyki, lecz również umożliwi praktyczne sprawdzenie, w jaki sposób metody te znajdują zastosowanie w badaniach rzeczywistych zjawisk gospodarczych i społecznych.



6. Literatura

1. Aczel M. (2017), *Statystyka w zarządzaniu*, Wydawnictwo, PWN, Warszawa
2. Balicki A., Makać W., (2010), *Metody wnioskowania statystycznego*, Wydawnictwo Uniwersytetu Gdańskiego, Gdańsk.
3. Krysicki W., Bartos J., Dyczka W., Królikowska K., Wasilewski M. (2007), *Rachunek prawdopodobieństwa i statystyka matematyczna w zadaniach, część I*, Wydawnictwo PWN Warszawa.
4. Józwiak J., Podgórski J. (2001), *Statystyka od podstaw*, Wydawnictwo PWN, Warszawa
5. Koronacki J., Mielniczuk J. (2001), *Statystyka dla studentów kierunków technicznych i przyrodniczych*, Wydawnictwa Naukowo-Techniczne, Warszawa
6. Ombach J. (2006), *Wprowadzenie do metod probabilistycznych wspomagane komputerowo - MAPLE*, Wydawnictwo PWSZ, Nowy Sącz.
7. Praca zbiorowa pod redakcją H. Kassyk-Rokickiej (2005), *Statystyka, zbiór zadań*, Wydawnictwo PWE, Warszawa.
8. Sobczyk M. (2003), *Statystyka*, Wydawnictwo UMCS, Lublin.
9. Starzyńska W. (2005), *Statystyka praktyczna*, Wydawnictwo Naukowe PWN, Warszawa.
10. Walesiak M., Gatnar E. (2009), *Statystyczna analiza danych z wykorzystaniem programu R*, PWN, Warszawa.
11. Zeszyt metodologiczny Ruch naturalny. Bilanse ludności, (2018), Główny Urząd Statystyczny, Warszawa.



7. Wzory

7.1 Statystyka opisowa

1. Miary położenia

Średnia arytmetyczna		
Dla szeregu szczegółowego	Dla szeregu rozdzielczego punktowego	Dla szeregu rozdzielczego przedziałowego
$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$	$\bar{x} = \frac{1}{n} \sum_{i=1}^k x_i n_i = \sum_{i=1}^k x_i p_i$	$\bar{x} = \frac{1}{n} \sum_{i=1}^k \dot{x}_i n_i = \sum_{i=1}^k \dot{x}_i p_i$
gdzie n- długość szeregu analizowanego, n_i – liczebność, p_i - częstość i- tego przedziału, $\dot{x}_i$ -środek przedziału, k-liczba przedziałów (grup)		
Dominanta (modalna)		
Dla szeregu rozdzielczego punktowego	Dla szeregu rozdzielczego przedziałowego	
$d_0 = x_D$ dla której $n_D = \max\{n_i\}$	$d_0 = x_d + \frac{n_d - n_{d-1}}{(n_d - n_{d-1}) + (n_d - n_{d+1})} h_d$ gdzie x_d – lewy koniec przedziału, w którym znajduje się dominanta n_d – liczebność przedziału z dominantą n_{d-1} – liczebność przedziału poprzedzającego przedział z dominantą n_{d+1} -liczebność przedziału, następującego po przedziale z dominantą h_d – długość przedziału z dominantą	
Kwantyl rzędu p		
Dla szeregu rozdzielczego punktowego	Dla szeregu rozdzielczego przedziałowego	
$q_p := \min \{x : P(X \leq x) \geq p\}$ $q_p = x_{[n \cdot p] + 1}$ Mediana =Kwantyl rzędu 0,5 (kwartl 2) $m_e = q_{0,5} = Q_2$ $= \begin{cases} \frac{1}{2} \left(x_{\frac{n}{2}} + x_{\frac{n}{2} + 1} \right) & \text{dla } n \text{ parzystych} \\ x_{\left(\frac{n+1}{2}\right)} & \text{dla } n \text{ nieparzystych} \end{cases}$	$q_p = x_p + \frac{h_p}{n_p} \left(n \cdot p - \sum_{i=1}^{p-1} n_i \right)$ gdzie x_p – lewy koniec przedziału, w którym znajduje się kwantyl rzędu p n_p – liczebność przedziału z kwantylem $\sum_{i=1}^{p-1} n_i$ – liczebność skumulowana do przedziału poprzedzającego przedział z kwantylem h_d – długość przedziału z kwantylem Mediana= Kwantyl rzędu 0,5 $m_e = x_m + \frac{h_m}{n_m} \left(0,5n - \sum_{i=1}^{m-1} n_i \right)$	



2. Miary zróżnicowania

Wariancja	
Dla szeregu rozdzielczego punktowego	Dla szeregu rozdzielczego przedziałowego
$s^2(x) = \frac{1}{n} \sum_{i=1}^k (x_i - \bar{x})^2 n_i = \sum_{i=1}^k (x_i - \bar{x})^2 p_i$	$s^2(x) = \frac{1}{n} \sum_{i=1}^k (\dot{x}_i - \bar{x})^2 n_i$
Odchylenie standardowe	$s(x) = \sqrt{s^2(x)}$
Współczynnik zmienności	$V = \frac{s(x)}{\bar{x}} 100\%$
Rozstęp	$R = \max\{x_i\} - \min\{x_i\}$
Rozstęp międzykwartylowy	$R_Q = Q_3 - Q_1$

3. Miary asymetrii (skośności)

Trzeci moment centralny	
Dla szeregu rozdzielczego punktowego	Dla szeregu rozdzielczego przedziałowego
$\mu_3 = \frac{1}{n} \sum_{i=1}^k (x_i - \bar{x})^3 n_i$	$\mu_4 = \frac{1}{n} \sum_{i=1}^k (\dot{x}_i - \bar{x})^3 n_i$
Współczynnik asymetrii	$A_1 = \frac{\mu_3}{(s(x))^3}$
Współczynnik skośności (wskaźnik asymetrii Pearsona)	$A = \frac{\bar{x} - d_0}{s(x)}$
Pozycyjny współczynnik asymetrii (Yule'a-Kendalla'a)	$A_{YK} = \frac{(Q_3 - m_e) - (m_e - Q_1)}{Q_3 - Q_1}$



4. Miary koncentracji (spłaszczenia rozkładu)

Czwarty moment centralny	
Dla szeregu rozdzielczego punktowego	Dla szeregu rozdzielczego przedziałowego
$\mu_4 = \frac{1}{n} \sum_{i=1}^k (x_i - \bar{x})^4 n_i$	$\mu_4 = \frac{1}{n} \sum_{i=1}^k (\dot{x}_i - \bar{x})^4 n_i$
Kurtoza (współczynnik ekscesu)	$K = \frac{\mu_4}{(S(x))^4} - 3$

7.2. Wnioskowanie statystyczne

1. Estymacja przedziałowa

 $(1 - \alpha)$ - poziom ufności,

$$P(\theta_L(X_1, \dots, X_n) < \theta < \theta_P(X_1, \dots, X_n)) = F(\theta_P) - F(\theta_L) = 1 - \alpha,$$

Szacowany parametr	Założenie dotyczące rozkładu	Liczebność próby	Ocena parametru-estymator punktowy	Przedział ufności	Wartość odczytywana z tablic
Wartość oczekiwana m	$X \sim N(m, \sigma)$ σ - znane	n -dowolne	$\hat{m} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$	$\bar{x} - u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} < m < \bar{x} + u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$	$u_{1-\alpha/2}$ kwantyl rozkładu $N(0, 1)$ $u_{1-\frac{\alpha}{2}} = \phi^{-1}(1 - \frac{\alpha}{2})$
	$X \sim N(m, \sigma)$ $\sigma = ?$	$n \leq 30$		$\bar{x} - t_{\alpha, n-1} \frac{S}{\sqrt{n}} < m < \bar{x} + t_{\alpha, n-1} \frac{S}{\sqrt{n}}$	$t_{\alpha, n-1}$ kwantyl rozkładu t-Studenta o $(n - 1)$ stopniach swobody $P(t \geq t_{\alpha}) = \alpha$
		$n > 30$		$\bar{x} - u_{1-\alpha/2} \frac{S}{\sqrt{n}} < m < \bar{x} + u_{1-\alpha/2} \frac{S}{\sqrt{n}}$	$u_{1-\alpha/2}$ kwantyl rozkładu $N(0, 1)$ j.w
	$X \sim ??$	$n > 120$		$\bar{x} - u_{1-\alpha/2} \frac{S}{\sqrt{n}} < m < \bar{x} + u_{1-\alpha/2} \frac{S}{\sqrt{n}}$	
Wariancja σ^2	$X \sim N(m, \sigma)$	$n \leq 30$	$\hat{\sigma}^2 = S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$	$\frac{(n-1)\hat{S}^2}{\chi_{\frac{\alpha}{2}, n-1}^2} < \sigma^2 < \frac{(n-1)\hat{S}^2}{\chi_{1-\frac{\alpha}{2}, n-1}^2}$	$P(\chi^2 \geq \chi_{\frac{\alpha}{2}, n-1}^2) = \frac{\alpha}{2}$ $P(\chi^2 \geq \chi_{1-\frac{\alpha}{2}, n-1}^2) = 1 - \frac{\alpha}{2}$
		$n > 30$		$\left(S - u_{1-\alpha/2} \frac{S}{\sqrt{2n}}\right)^2 < \sigma^2 < \left(S + u_{1-\alpha/2} \frac{S}{\sqrt{2n}}\right)^2$	$u_{1-\alpha/2}$ kwantyl rozkładu $N(0, 1)$ j.w
	$X \sim ??$	$n > 50$		$\left(\frac{\sqrt{2n}S}{\sqrt{2n-3} + u_{\alpha}}\right)^2 < \sigma^2 < \left(\frac{\sqrt{2n}S}{\sqrt{2n-3} - u_{\alpha}}\right)^2$	
Wskaźnik struktury p	częstość sukcesów w n -elementowej próbie $X \sim B(n, p)$	$n > 100$	$\hat{p} = \frac{k}{n}$	$\hat{p} - u_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + u_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$u_{1-\alpha/2}$ kwantyl rozkładu $N(0, 1)$ j.w



Fundusze Europejskie
dla Rozwoju Społecznego



Rzeczpospolita
Polska

Dofinansowane przez
Unię Europejską



2. Weryfikacja hipotez statystycznych – wybrane testy parametryczne

α - poziom istotności. Niech θ oznacza dowolny parametr rozkładu ogólna formuła testu przyjmuje postać

Hipoteza zerowa	$H_0: \theta = \theta_0$		
Hipoteza alternatywna	$H_1: \theta \neq \theta_0$ (test dwustronny)	$H_1: \theta < \theta_0$ (test lewostronny)	$H_1: \theta > \theta_0$ (test prawostronny)
Obszar krytyczny (obszar odrzuceń hipotezy zerowej)	$K = (-\infty, -q_\alpha) \cup (q_\alpha, \infty)$	$K = (-\infty, -q_{2\alpha})$	$K = (q_{2\alpha}, \infty)$

q_α - jest kwantylem z odpowiedniego rozkładu (szczegóły tabele poniżej)



Testy istotności dla jednej populacji

Założenie dotyczące rozkładu	Liczebność próby	Statystyka testująca	Wartość odczytywana z tablic
$H_0: m = m_0$			
$X \sim N(\mu, \sigma)$ σ – znane m – nie jest znane	n -dowolne	$U = \frac{\bar{x} - m_0}{\sigma} \sqrt{n}$	$q_\alpha = u_{1-\alpha/2}$ kwantyl rozkładu $N(0,1)$ j.w
$X \sim N(\mu, \sigma)$ σ – nie jestznane m – nie jest znane	$n \leq 30$	$T = \frac{\bar{x} - m_0}{s} \sqrt{n}$ gdzie $s = \sqrt{\frac{1}{n-1} \sum_{j=1}^n (x_j - \bar{x})^2}$	$q_\alpha = t_{\alpha, n-1}$ kwantyl rozkładu t-Studenta o $(n - 1)$ stopniach swobody $P(t \geq t_\alpha) = \alpha$
	$n > 30$	$U = \frac{\bar{x} - m_0}{s} \sqrt{n}$	$q_\alpha = u_{1-\alpha/2}$ kwantyl rozkładu $N(0,1)$ j.w
$X \sim ??(\mu, \sigma)$ σ – nie jest znane m – nie jest znane	$n > 120$	$U = \frac{\bar{x} - m_0}{s} \sqrt{n}$	
$H_0: \sigma^2 = \sigma_0^2$!!! z reguły stosujemy test prawostronny $H_1: \sigma > \sigma_0$			
$X \sim N(\mu, \sigma)$ σ – nie jest znane m – nie jest znane	$n \leq 50$	$\chi^2 = \frac{(n - 1)s^2}{\sigma_0^2}$	$q_\alpha = \chi_{\alpha, n-1}^2$ kwantyl rozkładu chi-kwadrat o $(n - 1)$; $P(\chi^2 \geq \chi_{\alpha, n-1}^2) = \alpha$
$X \sim ??$ σ – nie jest znane m – nie jest znane	$n > 50$	$U = \sqrt{2\chi^2} - \sqrt{2n - 3}$	$q_{2\alpha} = u_{1-\alpha}$ kwantyl rozkładu $N(0,1)$ $u_{1-\alpha} = \Phi^{-1}(1 - \alpha)$
$H_0: p = p_0$			
X posiada rozkład dwupunktowy $\hat{p} = \frac{k}{n}$	$n > 100$	$U = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$	$q_\alpha = u_{1-\alpha/2}$ kwantyl rozkładu $N(0,1)$ j.w



Testy istotności dla dwóch populacji

Założenie dotyczące rozkładu	Dodatkowe informacje	Statystyka testująca	Wartość odczytywana z tablic
$H_0: m_1 = m_2$ (zakładamy, że badane są dwie niezależne populacje) H_1 może przyjmować jedną z postaci przedstawionych powyżej Wybór testu zależy od posiadanych informacji na temat badanych populacji			
$X_1 \sim N(m_1, \sigma_1)$, $X_2 \sim N(m_2, \sigma_2)$, σ_1, σ_2 – znane m_1, m_2 – nie są znane	n_1, n_2 – dowolne	$U = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	$q_\alpha = u_{1-\alpha/2}$ kwantyl rozkładu $N(0,1)$ j.w
$X_1 \sim N(m_1, \sigma_1)$, $X_2 \sim N(m_2, \sigma_2)$, $m_1, m_2, \sigma_1, \sigma_2$ – nie są znane	$\sigma_1 = \sigma_2$	$T = \frac{\bar{x}_1 - \bar{x}_2 - (m_1 - m_2)}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$ $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2},$ gdzie s_i^2 jest estymatorem wariancji w i-tej populacji $s_i^2 = \frac{1}{n_i-1} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2, i = 1, 2$	$q_\alpha = t_{\alpha, v}$ kwantyl rozkładu t-Studenta o v stopniach swobody, $v = n_1 + n_2 - 2$; $P(t \geq t_\alpha) = \alpha$
	$\sigma_1 \neq \sigma_2$	$T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ gdzie s_1^2 i s_2^2 – tak jak wyżej oznaczają wariancje w badanych populacjach.	$q_\alpha = t_{\alpha, v}$ kwantyl rozkładu t-Studenta o v stopniach swobody, $v \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}$
$X \sim ??$ $m_1, m_2, \sigma_1, \sigma_2$ – nie są znane	$n_1, n_2 > 100$  (Ctrl)	$U = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ gdzie s_1^2 i s_2^2 – tak jak wyżej oznaczają wariancje w badanych populacjach.	$q_\alpha = u_{1-\alpha/2}$ kwantyl rozkładu $N(0,1)$ j.w

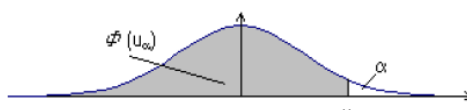


Testy istotności dla dwóch populacji c.d.

Założenie dotyczące rozkładu	Dodatkowe informacje	Statystyka testująca	Wartość odczytywana z tablic
$H_0: \sigma_1^2 = \sigma_2^2$			
$X_1 \sim N(m_1, \sigma_1),$ $X_2 \sim N(m_2, \sigma_2),$ $\sigma_1, \sigma_2, m_1, m_2$ – nie są znane	$H_1: \sigma_1^2 \neq \sigma_2^2$	$F = \frac{s_1^2}{s_2^2}$ gdzie $s_1^2 > s_2^2$	Obszar krytyczny w tym teście należy wyznaczyć z dwóch relacji $P\left(F \geq F_{\frac{\alpha}{2}}\right) = \frac{\alpha}{2},$ (prawostronna część obszaru krytycznego) $P\left(F \leq F_{\frac{\alpha}{2}}\right) = \frac{\alpha}{2},$ (lewostronna część obszaru krytycznego) gdzie $F_{\frac{\alpha}{2}}$ jest kwantylem rozkładu F o $(n_1 - 1)$ i $(n_2 - 1)$ stopniach swobody
	$H_1: \sigma_1^2 > \sigma_2^2$ W przypadku gdy interesuje nas $H_1: \sigma_1^2 < \sigma_2^2$ należy przenumerać populacje tak, aby w efekcie uzyskać prawostronną hipotezę alternatywną.		Prawostronny obszar krytyczny $P(F \geq F_{\alpha}) = \alpha$ gdzie F_{α} jest kwantylem rozkładu F o $(n_1 - 1)$ i $(n_2 - 1)$ stopniach swobody
$H_0: p_1 = p_2$			
X_1, X_2 posiadają rozkłady dwupunktowe, frakcje dla próby wyznaczamy $\hat{p}_1 = \frac{k_1}{n_1}, \hat{p}_2 = \frac{k_2}{n_2}$	$n_1, n_2 > 100$	$U = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ gdzie $\hat{p} = \frac{k_1 + k_2}{n_1 + n_2}$	$q_{\alpha} = u_{1-\alpha/2}$ kwantyl rozkładu $N(0,1)$ j.w



Tablice wybranych rozkładów



1 Dystrybuanta standaryzowanego rozkładu normalnego

$$\Phi(u_\alpha) = P(u \leq u_\alpha) = 1 - \alpha$$

u_α	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09	u_α
0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359	0
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753	0,1
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141	0,2
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517	0,3
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879	0,4
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224	0,5
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549	0,6
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852	0,7
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133	0,8
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389	0,9
1	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621	1
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830	1,1
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015	1,2
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177	1,3
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319	1,4
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441	1,5
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545	1,6
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633	1,7
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706	1,8
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767	1,9
2	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817	2
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857	2,1
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890	2,2
2,3	0,9893	0,9896	0,9898	0,9901	0,9903	0,9906	0,9908	0,9911	0,9913	0,9915	2,3
2,4	0,9918	0,9920	0,9922	0,9924	0,9926	0,9928	0,9930	0,9932	0,9934	0,9936	2,4
2,5	0,9937	0,9939	0,9941	0,9943	0,9945	0,9947	0,9949	0,9951	0,9953	0,9955	2,5
2,6	0,9957	0,9959	0,9961	0,9963	0,9965	0,9967	0,9969	0,9971	0,9973	0,9975	2,6
2,7	0,9977	0,9979	0,9981	0,9983	0,9985	0,9987	0,9989	0,9991	0,9993	0,9995	2,7
2,8	0,9997	0,9999	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	2,8
2,9	0,9997	0,9999	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	2,9
3	0,9997	0,9999	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	3
3,1	0,9997	0,9999	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	3,1
3,2	0,9997	0,9999	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	3,2
3,3	0,9997	0,9999	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	3,3
3,4	0,9997	0,9999	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	3,4
3,5	0,9997	0,9999	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	3,5
3,6	0,9997	0,9999	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	3,6
3,7	0,9997	0,9999	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	3,7
3,8	0,9997	0,9999	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	3,8
3,9	0,9997	0,9999	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	3,9
$\Phi(4) = 0,9999$											
$\Phi(5) = 0,9999$											
$\Phi(6) = 0,9999$											

**2 Wartości krytyczne rozkładu t-Studenta**

$$P(|t| \geq t_{\alpha,k}) = \alpha \quad (k\text{-liczba stopni swobody})$$

$\alpha \backslash k$	0,2	0,1	0,05	0,02	0,01	0,002	0,001	$\alpha \backslash k$
1	3,078	6,314	12,706	31,821	63,66	318,29	636,58	1
2	1,886	2,920	4,303	6,965	9,925	22,328	31,600	2
3	1,638	2,353	3,182	4,541	5,841	10,214	12,924	3
4	1,533	2,132	2,776	3,747	4,604	7,173	8,610	4
5	1,476	2,015	2,571	3,365	4,032	5,894	6,869	5
6	1,440	1,943	2,447	3,143	3,707	5,208	5,959	6
7	1,415	1,895	2,365	2,998	3,499	4,785	5,408	7
8	1,397	1,860	2,306	2,896	3,355	4,501	5,041	8
9	1,383	1,833	2,262	2,821	3,250	4,297	4,781	9
10	1,372	1,812	2,228	2,764	3,169	4,144	4,587	10
11	1,363	1,796	2,201	2,718	3,106	4,025	4,437	11
12	1,356	1,782	2,179	2,681	3,055	3,930	4,318	12
13	1,350	1,771	2,160	2,650	3,012	3,852	4,221	13
14	1,345	1,761	2,145	2,624	2,977	3,787	4,140	14
15	1,341	1,753	2,131	2,602	2,947	3,733	4,073	15
16	1,337	1,746	2,120	2,583	2,921	3,686	4,015	16
17	1,333	1,740	2,110	2,567	2,898	3,646	3,965	17
18	1,330	1,734	2,101	2,552	2,878	3,610	3,922	18
19	1,328	1,729	2,093	2,539	2,861	3,579	3,883	19
20	1,325	1,725	2,086	2,528	2,845	3,552	3,850	20
21	1,323	1,721	2,080	2,518	2,831	3,527	3,819	21
22	1,321	1,717	2,074	2,508	2,819	3,505	3,792	22
23	1,319	1,714	2,069	2,500	2,807	3,485	3,768	23
24	1,318	1,711	2,064	2,492	2,797	3,467	3,745	24
25	1,316	1,708	2,060	2,485	2,787	3,450	3,725	25
26	1,315	1,706	2,056	2,479	2,779	3,435	3,707	26
27	1,314	1,703	2,052	2,473	2,771	3,421	3,689	27
28	1,313	1,701	2,048	2,467	2,763	3,408	3,674	28
29	1,311	1,699	2,045	2,462	2,756	3,396	3,660	29
30	1,310	1,697	2,042	2,457	2,750	3,385	3,646	30
40	1,303	1,684	2,021	2,423	2,704	3,307	3,551	40
60	1,296	1,671	2,000	2,390	2,660	3,232	3,460	60
120	1,289	1,658	1,980	2,358	2,617	3,160	3,373	120
∞	1,282	1,645	1,960	2,326	2,576	3,090	3,290	∞