Marek Cieciura, Janusz Zacharski

PODSTAWY PROBABILISTYKI Z PRZYKŁADAMI ZASTOSOWAŃ W INFORMATYCE

CZĘŚĆ II
STATYSTYKA OPISOWA

Warszawa, wrzesień 2011

Data ostatniej aktualizacji: czwartek, 20 października 2011, godzina 17:20

Statystyka jest bardziej sposobem myślenia lub wnioskowania niż pęczkiem recept na młócenie danych w celu odsłonięcia odpowiedzi - Calyampudi Radhakrishna Rao

Podręcznik:

PODSTAWY PROBABILISTYKI Z PRZYKŁADAMI ZASTOSOWAŃ W INFORMATYCE

publikowany jest w częściach podanych poniżej

Nr	Tytuł
I.	Wprowadzenie
II.	Statystyka opisowa
III.	Rachunek prawdopodobieństwa
IV.	Statystyka matematyczna
V.	Przykłady zastosowań w informatyce
VI.	Wybrane twierdzenia z dowodami
VII.	Tablice statystyczne

Autorzy proszą o przesyłanie wszelkich uwagi i propozycji dotyczących zawartości podręcznika z wykorzystaniem formularza kontaktowego zamieszczonego w portalu http://cieciura.net/mp/

Publikowane części będą na bieżąco poprawiane, w każdej będzie podawana data ostatniej aktualizacji.

SPIS TREŚCI

2. CHARAKTERYSTYKI LICZBOWE

2.1. Uwagi wstępne

Niech $x_1, x_2, ..., x_n$ będą wartościami cechy X wszystkich elementów populacji albo próby. Są to tzw. *dane statystyczne.*

Charakterystyki liczbowe (opisowe) są to liczby charakteryzujące rozkład cechy populacji. Charakterystyki liczbowe cechy X, podobnie jak parametry rozkładu zmiennej losowej, dzielimy na

- Charakterystyki położenia (średnia, mediana, dominanta);
- Charakterystyki rozproszenia (wariancja, odchylenie standardowe, współczynnik zmienności, odchylenie przeciętne, rozstęp);
- Charakterystyki asymetrii (współczynnik asymetrii, wskaźnik asymetrii);
- Charakterystyki spłaszczenia (kurtoza).

2.2. Charakterystyki położenia

Inne nazwy charakterystyk położenia to: charakterystyki/miary przeciętne, średnie, tendencji centralnej

2.2.1. Średnia arytmetyczna danych statystycznych

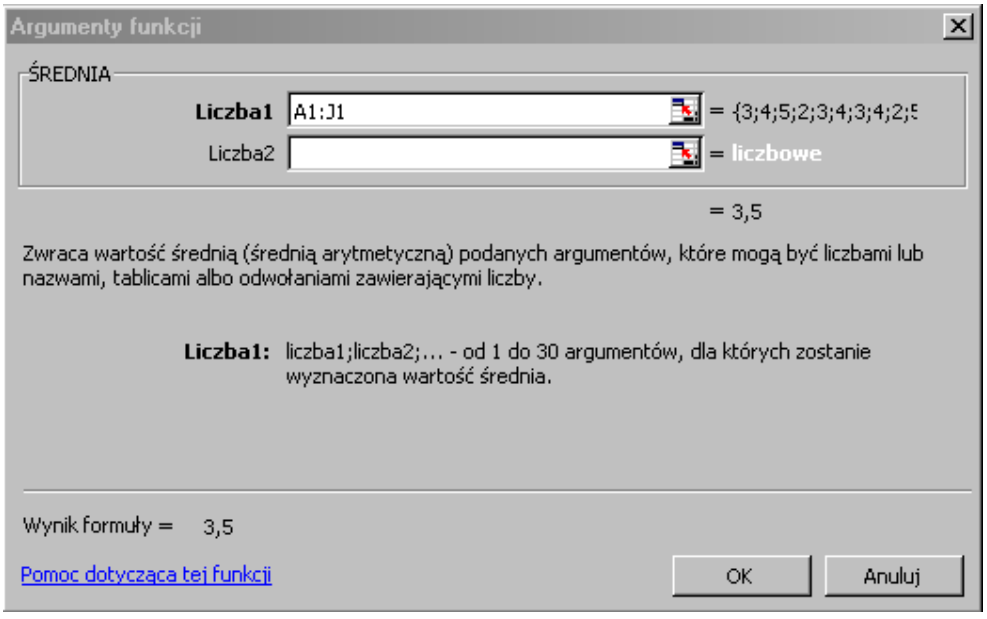
$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Przykład 2.1

Z pewnego egzaminu uzyskano następujące oceny: 3, 4, 5, 2, 3, 4, 3, 4, 2, 5. Należy obliczyć ich średnia arytmetyczną.

$$\overline{x} = \frac{3+4+5+2+3+4+3+4+2+5}{10} = 3,5$$

Średnią arytmetyczną można obliczyć korzystając z arkusza kalkulacyjnego Excel co ilustruje poniższy rysunek.



Wykorzystano funkcję statystyczną ŚREDNIA wpisując wcześniej dane w komórki A1:J1. ■

Własności średniej arytmetycznej danych statystycznych $(x_1, x_2, ..., x_n)$ [1]

1. $x_{min} \le \overline{x} \le x_{max}$
2. $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$
3. $\sum_{x_i > \overline{x}}(x_i - \overline{x}) = \sum_{x_i < \overline{x}}(\overline{x} - x_i)$ zwraca się uwagę, że w nawiasach są wartości dodatnie
4. Wyrażenie $\sum_{i=1}^{n}(x_i - c)^2$ ma wartość najmniejszą gdy c= $\overline{x}$

2.2.2. Mediana danych statystycznych

Uporządkujmy dane statystyczne od najmniejszej do największej:

$$x_{(1)},\ x_{(2)},\ ...,\ x_{(n)}$$

Mediana danych statystycznych jest to liczba

$$m_e = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{gdy n jest liczbą nieparzystą} \\ \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n+2}{2}\right)}}{2} & \text{gdy n jest liczbą parzystą} \end{cases}$$

Przykład 2.2

Wyznaczymy medianę dla danych statystycznych w dwóch przypadkach

a) 3, 0, 2, 1, 6, 7, 4, 2, 5

b) 3, 0, 2, 1, 6, 7, 4, 2

Rozwiązanie

a) Porządkujemy dane statystyczne od najmniejszej do największej 0, 1, 2, 2, 3, 4, 5, 6, 7.
Ponieważ liczba danych statystycznych jest n = 9 (liczba nieparzysta}, więc

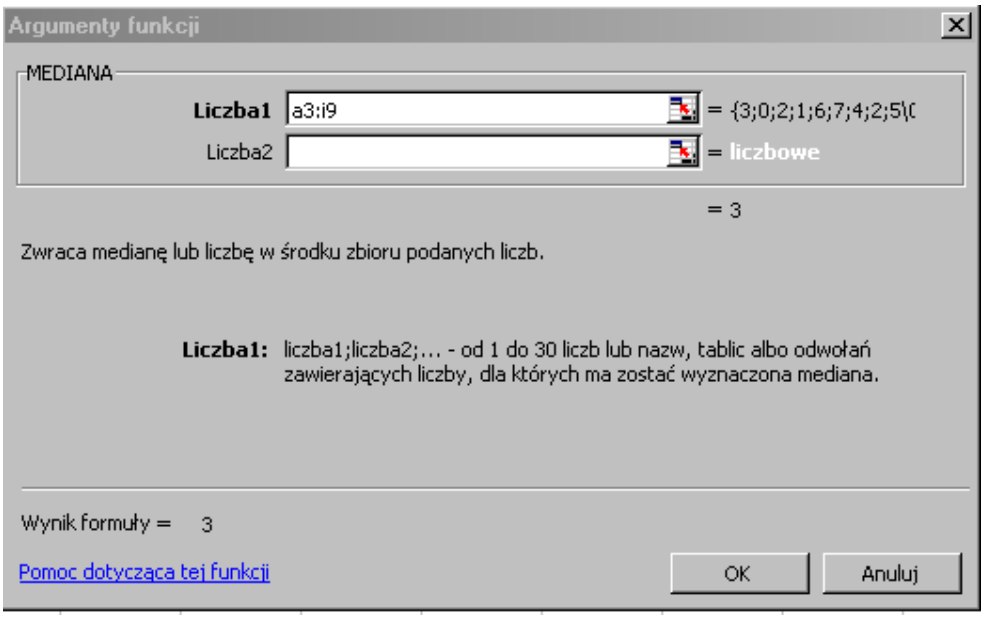
$$m_e = x_{\left(\frac{n+1}{2}\right)} = x_{(5)} = 3$$

b) Porządkujemy dane statystyczne od najmniejszej do największej 0, 1, 2, 2, 3, 4, 6, 7
Ponieważ liczba danych statystycznych jest n = 8 (liczba parzysta}, więc

$$m_e = \frac{1}{2}\left[x_{(\frac{n}{2})} + x_{(\frac{n+2}{2})}\right] = \frac{1}{2}\left[x_{(4)} + x_{(5)}\right] = \frac{1}{2}[2+3] = 2{,}5$$

Medianę można obliczyć korzystając z arkusza kalkulacyjnego Excel co dla pierwszego przypadku ilustruje poniższy rysunek.

[1] Patrz punkt 19.1. części VII Wybrane twierdzenia z dowodami



Wykorzystano funkcję statystyczną MEDIANA wpisując wcześniej dane w komórki a3:i9. ■

2.2.3. Dominanta danych statystycznych

Jest to najczęściej występująca dana statystyczna (o ile istnieje), oznacza się litera d. Dominanta jest także nazywana *modą*.

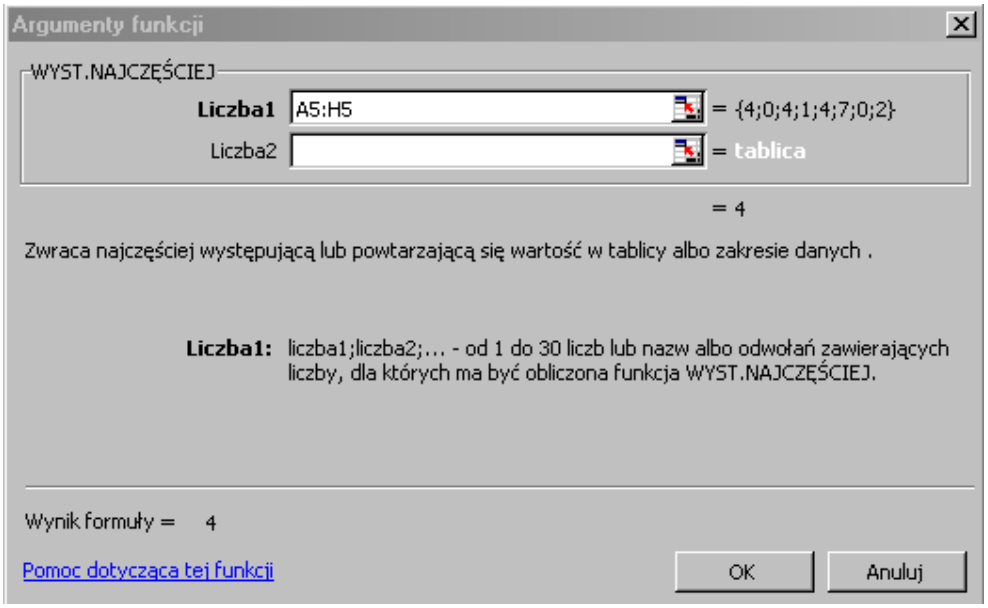
Przykład 2.3

Wyznaczymy dominantę dla danych statystycznych w dwóch przypadkach:
a) 4, 0, 4, 1, 4, 7, 0, 2 b) 3, 0, 2, 1, 6, 7, 4, 2, 1, 4, 2, 1

Rozwiązanie

a) Najczęściej występującą daną statystyczną jest liczba 4 (występuje 3 razy), zatem d = 4.
b) Nie ma danej statystycznej występującej najczęściej. Dominanta tych danych nie istnieje.

Dominantę można obliczyć korzystając z arkusza kalkulacyjnego Excel co dla pierwszego przypadku ilustruje poniższy rysunek.



Wykorzystano funkcję statystyczną WYST.NAJCZESCIEJ wpisując dane w komórki a3:i9 ■

Interpretacja charakterystyk położenia

Średnia arytmetyczna, mediana i dominanta są przykładami tzw. *charakterystyk położenia*, czyli wielkości informujących o przeciętnej wielkości cechy populacji. Wokół tych wielkości skupiają się na ogół wartości cechy populacji. Inaczej wyrażamy to mówiąc, że poznane charakterystyki są miarami *tendencji centralnej wartości cechy populacji.*

Średnia arytmetyczna jest liczbą informującą o tym, jaką wartość cechy powinny mieć elementy populacji, gdyby wszystkie dane statystyczne były sobie równe i suma tych wartości byłaby taka sama (podział wielkości na n równych części).

Mediana dzieli zbiór danych statystycznych na dwa równoliczne podzbiory: do jednego z nich należą dane mniejsze lub równe medianie, zaś do drugiego dane większe lub równe medianie.

Dominanta jest najbardziej typową daną statystyczną.

Jak określać przeciętny poziom cechy

Przykład 2.4

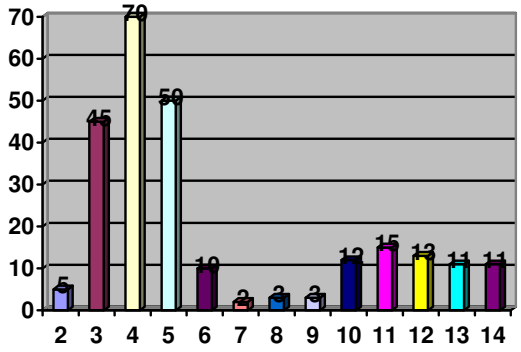
W pewnej firmie postanowiono przeanalizować zarobki pracowników. Dane w tys. zł. dotyczące wszystkich 250 pracowników przedstawia poniższa tabela:

Zarobki	2	3	4	5	6	7	8	9	10	11	12	13	14	Razem
Liczba pracowników	5	45	70	50	10	2	3	3	12	15	13	11	11	250

Tabelę otrzymano zliczając takie same zarobki w analizowanych danych przy pomocy funkcji statystycznej „Występowanie – ile razy" arkusza Excel – tak samo postąpiono w 2 kolejnych przykładach.

Innym sposobem jest wykorzystanie narzędzia analizy „Histogram" z pakietu „Analiza danych" - Analysis ToolPak arkusza Excel – otrzymuje się od razu liczby pracowników dla **wszystkich** poziomów zarobków.

C	D	E	F	G	H
Zbiór danych (kos	Częstość	Łączna wa	Zbiór dany	Częstość	Łączna wartość %
2	5	2,00%	4	70	28,00%
3	45	20,00%	5	50	48,00%
4	70	48,00%	3	45	66,00%
5	50	68,00%	11	15	72,00%
6	10	72,00%	12	13	77,20%
7	2	72,80%	10	12	82,00%
8	3	74,00%	13	11	86,40%
9	3	75,20%	14	11	90,80%
10	12	80,00%	6	10	94,80%
11	15	86,00%	2	5	96,80%
12	13	91,20%	8	3	98,00%
13	11	95,60%	9	3	99,20%
14	11	100,00%	7	2	100,00%

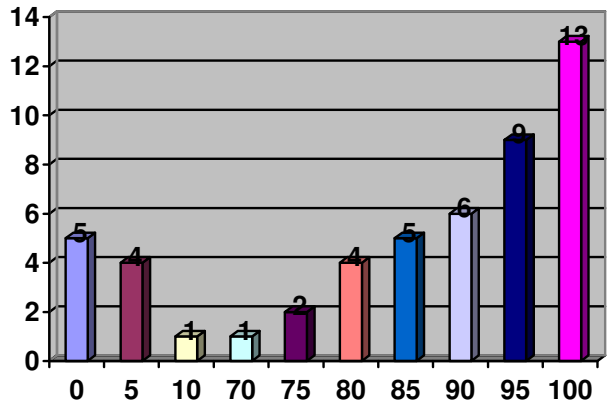
Ilustracja graficzna otrzymane przy pomocy modułu Microsoft Graph	Wyniki otrzymane przy pomocy funkcji statystycznych arkusza Excel
	6,152 Średnia 5 Mediana 4 Dominanta

Chcemy określić przeciętne wynagrodzenie w firmie. Średnia arytmetyczna wynosi 6152 zł, a 180 pracowników, czyli 72% otrzymuje wynagrodzenia poniżej średniej arytmetycznej. W tym przypadku jako przeciętne wynagrodzenie należy przyjąć medianę, która w tym przypadku wynosi 5 tys. zł. Zwraca się uwagę, że najczęściej występującym wynagrodzeniem, czyli dominantą, jest pensja w wysokości 4 tys. zł.

Przykład 2.5

Wykładowca postanowił przeanalizować wyniki testu z „Metod probabilistycznych". Dane dotyczące liczby zdobytych punktów przez 50 studentów przedstawia poniższa tabela.

Liczba punktów	0	5	10	70	75	80	85	90	95	100	Razem
Liczba studentów	5	4	1	1	2	4	5	6	9	13	50

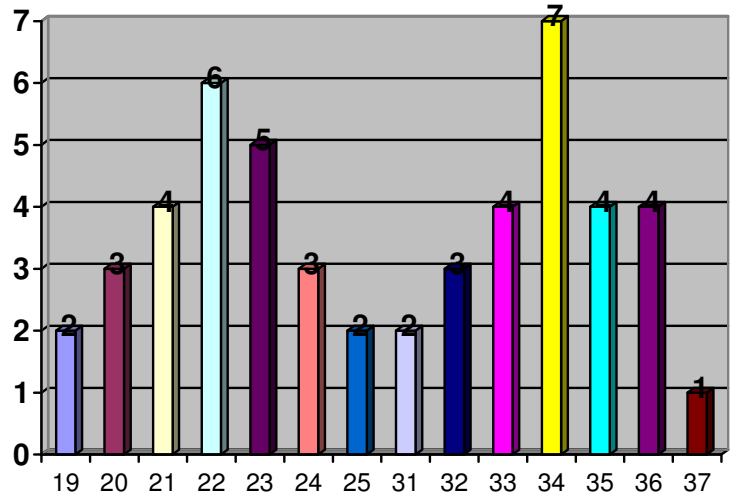
Ilustracja graficzna otrzymane przy pomocy modułu Microsoft Graph	Wyniki otrzymane przy pomocy funkcji statystycznych arkusza Excel
	73,8 Średnia 90 Mediana 100 Dominanta

Średnia arytmetyczna wyników testu wynosi 73,8 czyli dotyczy jedynie 3 wyników - spowodowane jest to tym, że 5 studentów tzn. 10% wypadło bardzo słabo otrzymując 0, 5 lub 10 punktów. Stąd jako przeciętny wynik testu należy przyjąć medianę, która jest równa 90 punktów. Zwraca się uwagę, że najczęściej występującym wynikiem, czyli dominantą, jest maksymalna liczba punktów równa 100.

Przykład 2.6

W pewnej uczelni postanowiono przeanalizować wiek studentów na specjalności bazy danych – w sumie 50 studentów. Dane przedstawia poniższa tabela:

Wiek	19	20	21	22	23	24	25	31	32	33	34	35	36	37	Razem
Liczba studentów	2	3	4	6	5	3	2	2	3	4	7	4	4	1	50

Ilustracja graficzna otrzymane przy pomocy modułu Microsoft Graph	Wyniki otrzymane przy pomocy funkcji statystycznych arkusza Excel
	28 Średnia 28 Mediana 34 Dominanta

Średnia arytmetyczna wieku studentów jest równa 28. Zauważmy, że nie ma ani jednego studenta o takim wieku, a także wieku zbliżonego (brak studentów o wieku 26, 27, 28, 29 i 30). W tym przypadku dla określenia przeciętnego wieku studentów należy podać dwa najczęściej występujące poziomy wieku: 22 i 34 – być może dotyczą one przeciętnego wieku studentów studiów stacjonarnych i niestacjonarnych. W tym przypadku podanie średniej arytmetycznej i mediany jest mylące.

Podsumowanie – jak określać przeciętny poziom cechy

- Średnia arytmetyczna - jeżeli rozkład jest symetryczny z jedną modą
- Mediana - jeżeli rozkład jest niesymetryczny z jedną modą
- Moda – jeżeli rozkład jest wielo modalny, podając ją dla każdego obszaru zmienności

Inne charakterystyki położenia

2.2.4. Średnia ważona danych statystycznych

z odnoszącymi się do ich nieujemnymi wagami $w_1, w_2, ..., w_n$ z których co najmniej jedna jest dodatnia, jest określona przez:

$$\overline{x}_w = \frac{w_1x_1 + w_2x_2 + ... + w_nx_n}{w_1 + w_2 + ... + w_n} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

W ten sposób dane którym przypisano większe wagi mają większy udział w określeniu średniej ważonej niż dane, którym przypisano mniejsze wagi.

Jeśli wszystkie wagi są równe, wówczas średnia ważona jest równa średniej arytmetycznej.

Przykład 2.7

W pewnej uczelni ocenę ukończenia studiów stanowi suma:

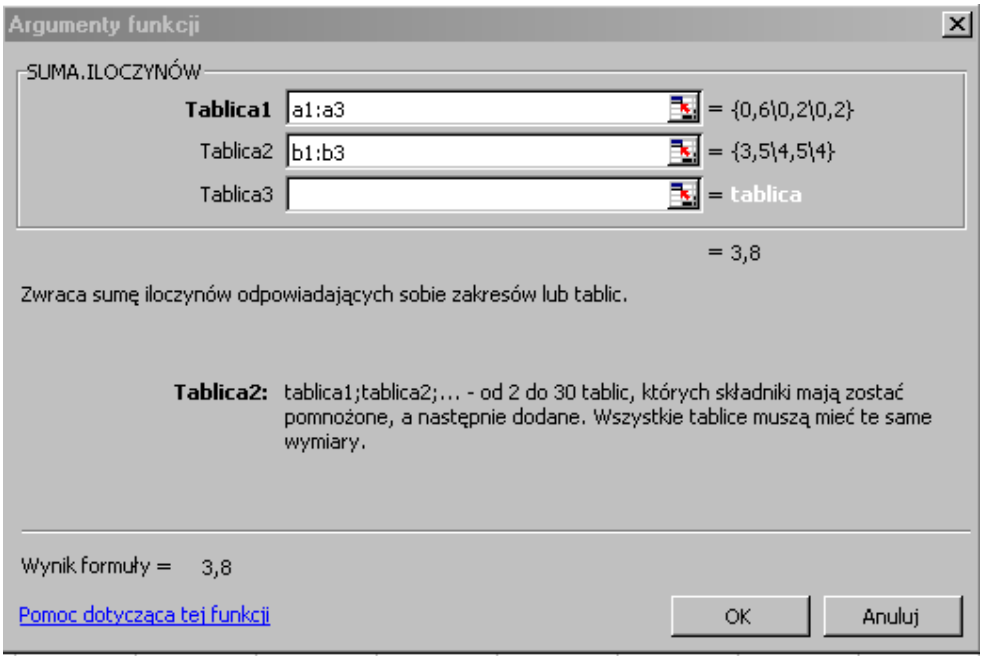
- 0,6 średniej wszystkich ocen x_1 z egzaminów i zaliczeń - cały okres studiów
- 0,2 oceny x_2 pracy dyplomowej,
- 0,2 oceny x_3 egzaminu dyplomowego.

Jest to przykład średniej ważonej:

$$\overline{x}_w = \frac{0,6x_1 + 0,2x_2 + 0,2x_3}{0,6 + 0,2 + 0,2} = 0,6x_1 + 0,2x_2 + 0,2x_3$$

Niech x1=3,5 x2=4,5 x3=4,0. Wtedy $\overline{x}_w = 0,6 \cdot 3,5 + 0,2 \cdot 4,5 + 0,2 \cdot 4,0 = 2,1 + 0,9 + 0,8 = 3,8$

Średnią ważoną można obliczyć korzystając z arkusza kalkulacyjnego Excel co ilustruje poniższy rysunek.



Wykorzystano funkcję matematyczną SUMA.ILOCZYNÓW wpisując wcześniej dane w komórki a1:a3 oraz b1:b3. W ogólnym przypadku (kiedy suma wag jest różna od 1) wynik należy podzielić przez sumę wag, którą można obliczyć z wykorzystaniem funkcji matematycznej SUMA. ■

2.2.5. Średnia ucinana danych statystycznych

Inne nazwy to: ***średnia obcięta lub średnia trymowana.*** Jest innych średnich, mody i mediany jedną z miar statystycznych tendencji centralnej.

Najprostszym przykładem jest sędziowanie zawodów sportowych przez 5 sędziów. Odrzuca się najniższą i najwyższą ocenę, a pozostałe sumuje się.

Przy obliczaniu średniej ucinanej obserwacje porządkuje się od najmniejszej do największej, odrzuca się mały procent najbardziej ekstremalnych obserwacji na obu krańcach (wartości najmniejsze oraz największe w próbce), na ogół równej liczności, a następnie oblicza się średnią z pozostałych obserwacji. Na ogół odrzuca się minimum i maksimum z próbki lub wartości poniżej 25 centyla i powyżej 75 centyla.

Wartości poniżej 25 centyla	Wartości poniżej 50 centyla	Wartości poniżej 75 centyla	Wartości poniżej 100 centyla
Odrzucanie	Obliczanie średniej		Odrzucanie

Rysunek 2.1.

Średnia ucinana jest charakterystyką mało wrażliwą na wartości odstające.

Średnia ucinana wykorzystywana jest do oceny zawodników w różnych konkurencjach, odrzuca się wtedy ocenę najwyższą i najniższą, a następnie z pozostałych oblicza się średnią arytmetyczną.

Przykład 2.8

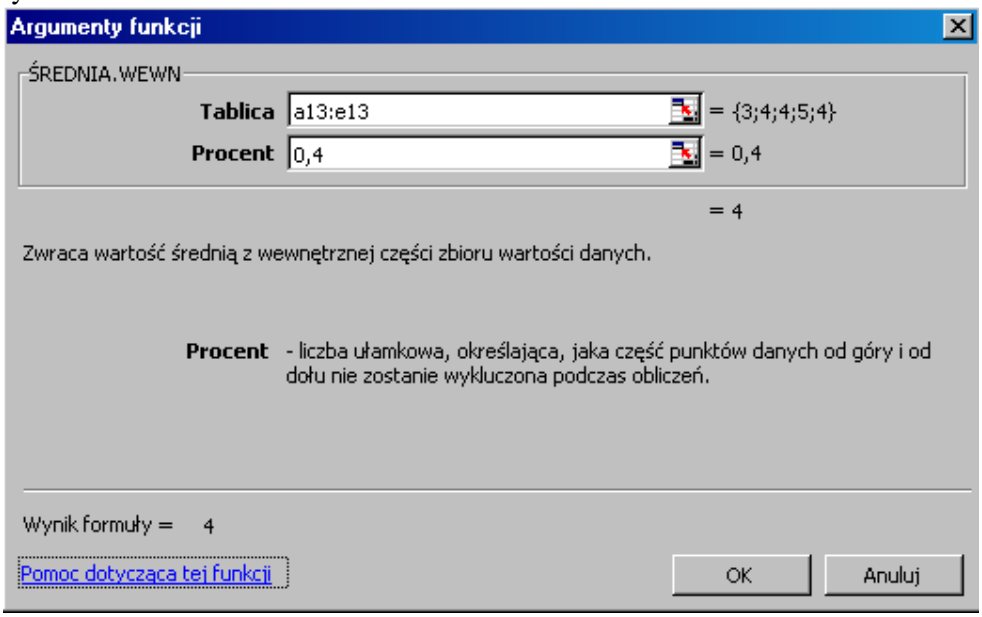
Pięciu sędziów oceniło skok do wody pewnego zawodnika wystawiając oceny: 3, 4, 4, 5, 4.

Obliczyć średnią ocen po odrzuceniu oceny najniższej i najwyższej.

Rozwiązanie

Ocena najniższa to 3, a ocena najwyższa 5. Pozostałe oceny to 4, zatem ich średnia wynosi 4.

Średnią ucinaną można obliczyć korzystając z arkusza kalkulacyjnego Excel co ilustruje poniższy rysunek.



Wykorzystano funkcję statystyczną ŚREDNIA.WEWN wpisując wcześniej dane w komórki a13:e13 oraz określając, że 40% danych ma być odrzuconych (20% najmniejszych i 20% największych). ■

2.2.6. Średnia geometryczna danych statystycznych

$$\overline{x}_g = \sqrt[n]{x_1 \cdot x_2 \cdots x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

Średnia geometryczna znajduje zastosowanie w badaniu średniego tempa zmian zjawisk, których rozwój jest przedstawiony w postaci szeregów dynamicznych, np. do uśredniania indeksów łańcuchowych[2].

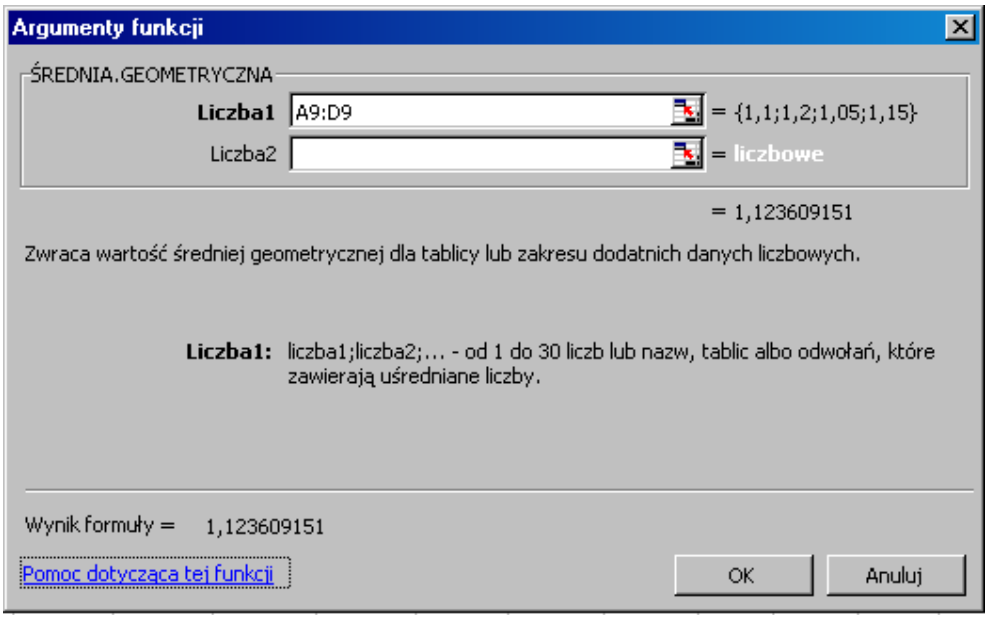
Przykład 2.9

Roczny procentowy przyrost przychodów pewnej firmy informatycznej w kolejnych czterech latach wynosił: 10%, 20%, 5%, 15%. Jaki był średni przyrost w tym okresie?

$$\overline{x}_g = \sqrt[4]{1{,}1 \cdot 1{,}2 \cdot 1{,}05 \cdot 1{,}15} = \sqrt[4]{1{,}5939} = \sqrt[2]{\sqrt[2]{1{,}5939}} = \sqrt[2]{1{,}2625} = 1{,}1236$$

Średnia geometryczna powyższych danych wynosi 12,5%.

Średnią geometryczną można obliczyć korzystając z arkusza kalkulacyjnego Excel co ilustruje poniższy rysunek.



Wykorzystano funkcję statystyczną ŚREDNIA.GEOMETRYCZNA wpisując wcześniej dane w komórki a9:d9. ■

2.2.7. Średnia harmoniczna danych statystycznych

$$\overline{x}_h = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}} = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}$$

Tak więc jest średnia harmoniczna (dla danych statystycznych różnych od zera) jest odwrotnością średniej arytmetycznej odwrotności danych statystycznych.

Średnią harmoniczną stosuje się w przypadku gdy wartości zmiennej podane są w jednostkach względnych (np. m/s, cm/osoba).

[2] Indeks łańcuchowy - iloraz poziomu zjawiska w okresie badanym, do poziomu zjawiska w okresie poprzedzającym okres badany.

Przykład 2.10

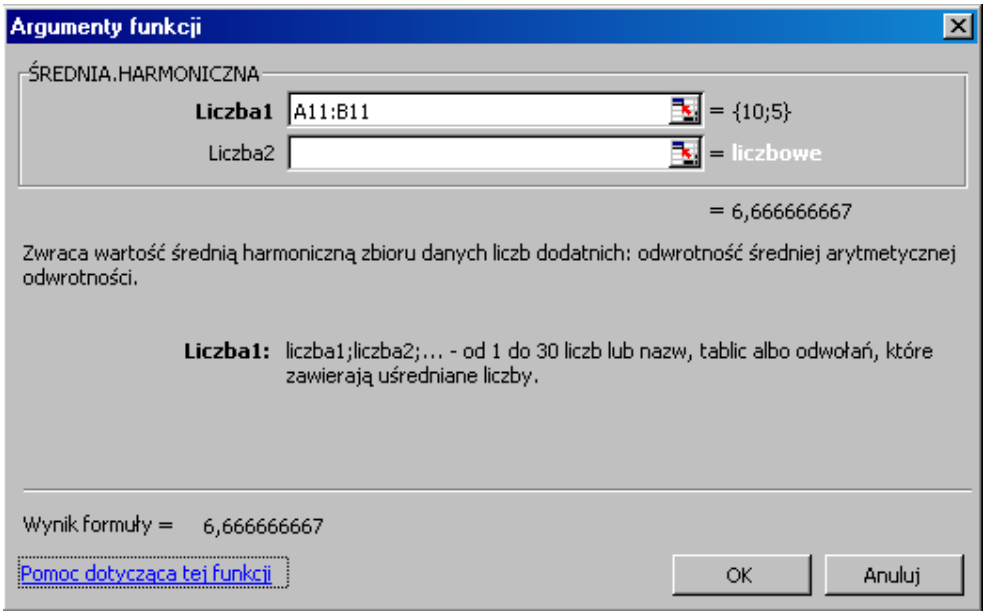
Odległość z miasta A do B rowerzysta przejeżdża z prędkością 10 km/godz, z powrotem jedzie z prędkością 5 km/godz. Jaka była prędkość średnia rowerzysty?

$$\text{Średnia arytmetyczna } \bar{x} = \frac{10+5}{2} = 7,5$$

$$\text{Średnia harmoniczna } x_h = \frac{2}{\frac{1}{10}+\frac{1}{5}} = \frac{2}{\frac{1+2}{10}} = \frac{20}{3} = 6,67$$

Załóżmy, że odległość pomiędzy miastami wynosi 10 km. Zatem czas przejazdu z A do B wynosi 1 godz., a powrotem 2 godz. Sumaryczna odległość wynosi 20 km, sumaryczny czas przejazdu 3 godz., zatem średnia prędkość wynosi 20/3 = 6,67 km/godz i pokrywa się ze średnią harmoniczną.

Średnią harmoniczną można obliczyć korzystając z arkusza kalkulacyjnego Excel co ilustruje poniższy rysunek.



Wykorzystano funkcję statystyczną ŚREDNIA.HARMONICZNA wpisując wcześniej dane w komórki A11:B1. ■

2.2.8. Średnia kwadratowa danych statystycznych

$$\bar{x}_k = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$$

Można wykazać prawdziwość zależności pomiędzy elementami próby $(x_1, x_2, ..., x_n)$ [3]:

$$\frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}} \le \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} \le \frac{1}{n}\sum_{i=1}^{n} x_i \le \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$$

Zwraca się uwagę, że elementy powyższej zależności liczone od lewej to: średnia harmoniczna, średnia geometryczna, średnia arytmetyczna i średnia kwadratowa.

[3] Patrz punkt 19.2. części VI Wybrane twierdzenia z dowodami

2.3. Charakterystyki rozproszenia

Inne nazwy charakterystyk rozproszenia to: charakterystyki/miary zróżnicowania, dyspersji

2.3.1. Wariancja danych statystycznych

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Wariancję można wyznaczyć ze wzoru[4]

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - (\frac{1}{n}\sum_{i=1}^{n}x_i)^2 = \overline{x_k^2} - (\overline{x})^2$$

Wzór ten umożliwia obliczenie wariancji w jednym przebiegu.

Przykład 2.11

Obliczyć wariancję wyników egzaminu podanych w przykładzie 2.1[5].

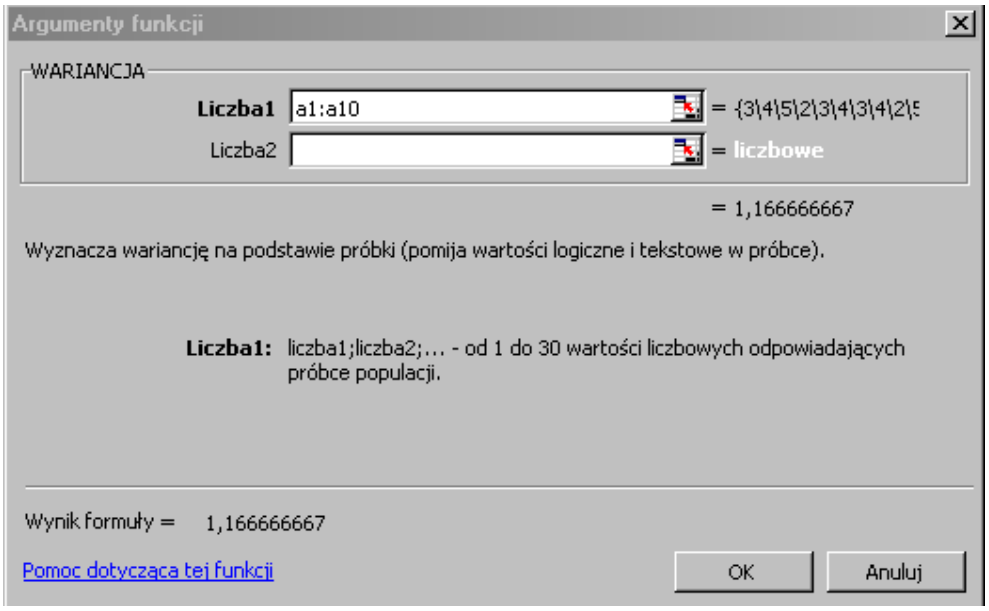
Pierwszy etap obliczeń zgodnie z powyższym wzorem przedstawiono w poniższej tabeli.

i	1	2	3	4	5	6	7	8	9	10	Suma	Suma/10
x_i	3	4	5	2	3	4	3	4	2	5	35	3,5
x_i^2	9	16	25	4	9	16	9	16	4	25	133	13,3

Zatem

$$s_x^2 = \overline{x_k^2} - (\overline{x})^2 = 13,3 - 3,5^2 = 13,3 - 12,25 = 1,05$$

Wariancję można obliczyć korzystając z arkusza kalkulacyjnego Excel co ilustruje poniższy rysunek.



Wykorzystano funkcję statystyczną WARIANCJA wpisując wcześniej dane w komórki A1:A10.

[4] Patrz punkt 19.3. części VII Wybrane twierdzenia z dowodami

[5] Rekomenduje się przeprowadzenie obliczeń z wykorzystaniem arkusza Excel

Zwraca się uwagę na różnicę w wynikach. Spowodowane jest to tym, że w arkuszu Excel we wzorze według którego obliczana jest wariancja zamiast $\frac{1}{n}$ występuje $\frac{1}{n-1}$ po to, aby zapewnić nieobciążoność wariancji, pojęcie zostanie wyjaśnione w statystyce matematycznej. Powody zostaną wyjaśnione przy omawianiu Statystyki matematycznej. ■

2.3.2. Odchylenie standardowe danych statystycznych

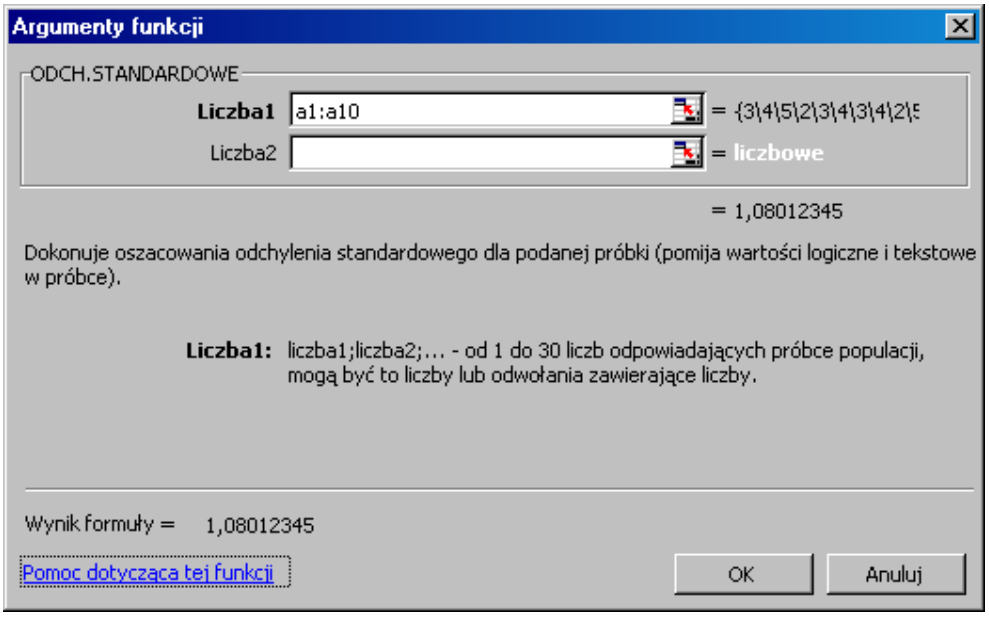
Odchylenie standardowe wyznaczane jest jako pierwiastek z wariancji.

$$s_x = \sqrt{s_x^2}$$

Przykład 2.12

Obliczyć odchylenie standardowe wyników egzaminu podanych w przykładzie 2.1.

Odchylenie można obliczyć korzystając z arkusza kalkulacyjnego Excel co ilustruje poniższy rysunek.



Wykorzystano funkcję statystyczną ODCH.STANDARDOWE wpisując wcześniej dane w komórki A1:A10.

2.3.3. Współczynnik zmienności danych statystycznych

$$v_x = \frac{s_x}{|\overline{x}|}100\%$$

przy założeniu, że $\overline{x} \neq 0$.

2.3.4. Rozstęp danych

$$r_0 = x_{max} - x_{min}$$

gdzie: x_{min} najmniejsza dana statystyczna, x_{max} – największa dana statystyczna.

Rozstęp można wyznaczyć jako różnicę wyników uzyskiwanych za pomocą dwóch funkcji statystycznych arkusza Excel: MAX i MIN.

2.3.5. Przedział typowych jednostek populacji

$$\left\langle \overline{x} - s_x ; \overline{x} + s_x \right\rangle$$

Interpretacja charakterystyk rozproszenia

Wariancja, odchylenie standardowe, współczynnik zmienności i rozstęp są przykładami charakterystyk rozproszenia (zmienności, zróżnicowania).

Każda z tych charakterystyk ma wartość równą zeru tylko w przypadku równych wszystkich danych statystycznych (nie ma wtedy zróżnicowania danych) i ma coraz większą wartość, gdy dane są bardziej zróżnicowane.

Wariancja i odchylenie standardowe mierzą rozproszenie danych statystycznych od ich średniej arytmetycznej.

Jeśli dane statystyczne są wyrażone w pewnych jednostkach, to wariancja jest wyrażona w tej jednostce do kwadratu. Tej niedogodności nie ma odchylenie standardowe.

Współczynnik zmienności wyraża, jaki procent stanowi odchylenie standardowe względem wartości średniej arytmetycznej. Jest wielkością niemianowaną (bez jednostki). Nadaje się więc do porównywania zróżnicowania cech populacji wyrażonych w różnych jednostkach.

Rozstęp wyraża długość najkrótszego przedziału, do którego należą wszystkie dane statystyczne.

2.3.5. Kwantyle

Kwantylem rzędu p (p-tym kwantylem) cechy X populacji nazywamy liczbę (oznaczenie k_p) taką, że co najmniej p procent danych statystycznych jest mniejszych lub równych tej liczbie oraz co najmniej 1-p procent danych statystycznych jest większych lub równych tej liczbie, przy czym liczba $p \in (0; 1)$.

Kwartyle q_1, q_2, q_3 pierwszy, drugi oraz trzeci są to kwantyle odpowiednio rzędu 0,25 , 0,50, 0,75. Kwartyl drugi q_2 jest oczywiście medianą cechy X.

Kwintyl to kwantyl rzędu 1/5 (pierwszy kwintyl, dolny kwintyl[6]), 2/5, 3/5 lub 4/5 (czwarty kwintyl, górny kwintyl). 20% obserwacji ma wartości poniżej dolnego kwintyla, a 20% powyżej górnego kwintyla.

Decyle d_1, d_2, … , d_9 pierwszy, drugi itd. do dziewiątego są to kwantyle odpowiednio rzędów 0,1, 0,2, … , 0,9.

Centyle c_1, c_2, … , c_{99} pierwszy, drugi itd. oraz dziewięćdziesiąty dziewiąty są to kwantyle odpowiednio rzędu 0,01 , 0,02, … , 0,99 – centyl jest więc wielkością, poniżej której padają wartości zadanego procentu próbek. Używa się także nazwy percentyl.

Kwartyle, kwintale, decyle i centyle dzielą dane statystyczne na odpowiednio cztery, dziesięć oraz sto równolicznych podzbiorów, co wykorzystuje się, gdy danych statycznych jest dużo.

Przykład 2.13

Badano wydajność 20 serwisantów. Otrzymane dane, dotyczące czasu usuwania określonej awarii, uporządkowano niemalejąco

48, 52, 53, 54, 56, 64, 65, 68, 68, 68, 70, 72, 72, 73, 74, 76, 83, 87, 89, 120

Obliczymy kwantyle rzędu 0,15 i rzędu 0,28.

Obliczamy 15% liczebności danych statystycznych n = 20,

$$l = 0{,}15 \cdot 20 = 3$$

Zatem $k_{0,15} = x_{(3)} = 53$ (trzeci wyraz w uporządkowanym niemalejąco ciągu danych statystycznych)

Sprawdzimy, czy otrzymany wynik jest zgodny z definicją kwantyla $k_{0,15}$.

[6] Przy pomocy kwintyli często redaguje się zasadę Pareto: dolny kwintyl obiektów generuje 80% zasobów.

Danych statystycznych co najwyżej równych 53 mamy 3, czyli 15% wszystkich danych, natomiast danych co najmniej równych 53 mamy 18, czyli 90%, co jest większe od 100% – 15% wszystkich danych. Kwantyl $k_{0,15}$ został zatem wyznaczony poprawnie.

Obliczamy 28% liczebności danych statystycznych

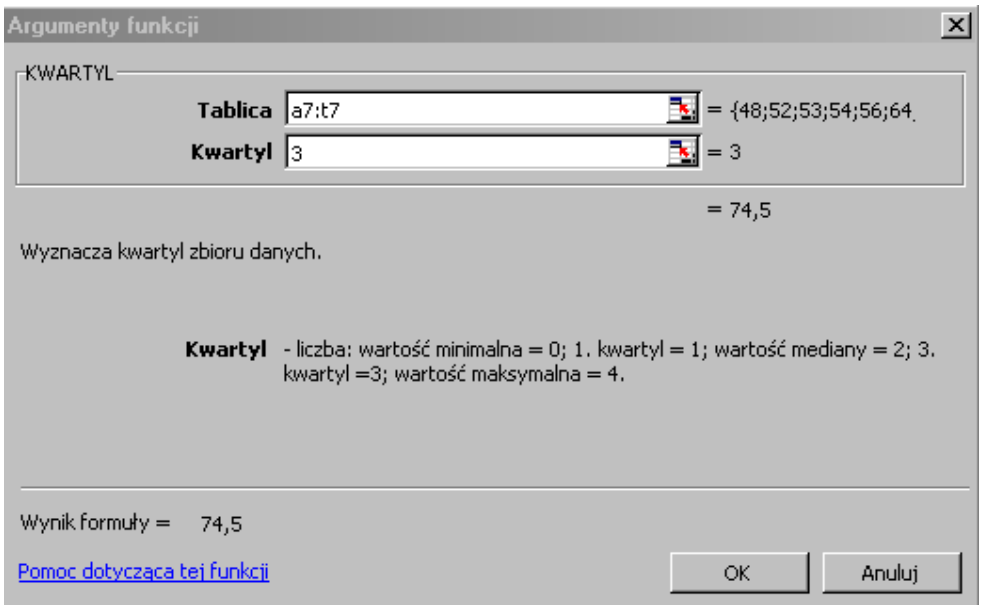
$$l = 0{,}28 \cdot 20 = 5{,}6 \approx 6$$

Przyjmujemy, że $k_{0,28} = x_{(6)} = 64$.

Rzeczywiście, danych co najwyżej równych 64 mamy 6, co stanowi 30% wszystkich danych. Jest to więcej niż 28%. Z drugiej strony danych co najmniej równych 64 mamy 15, co stanowi 75% wszystkich danych. Jest to więcej niż 100% - 28%. Zatem kwantyl $k_{0,28}$ został wyznaczony poprawnie. Zauważmy, że w tym przypadku każda liczba z przedziału (56; 64> jest kwantylem $k_{0,28}$.

Obliczymy teraz trzeci kwartyl q_3. Ponieważ 75% liczby 20 wynosi 15, to $q_3 = x_{(15)} = 73$.

Kwartyle można obliczyć korzystając z arkusza kalkulacyjnego Excel co dla trzeciego ilustruje poniższy rysunek.



Wykorzystano funkcję statystyczną KWARTYL wpisując wcześniej dane w komórki a7:t7. ■

2.3.6. Wskaźnik struktury

Rozważmy cechę X i pewien wariant tej cechy.

Wskaźnik struktury wariantu cechy X populacji jest to stosunek liczby danych statystycznych równych wariantowi do liczby wszystkich danych statystycznych

$$w = \frac{k}{n}$$

k – liczba danych statystycznych równych danemu wariantowi,
n – liczba wszystkich danych statystycznych.

Przykład 2.14

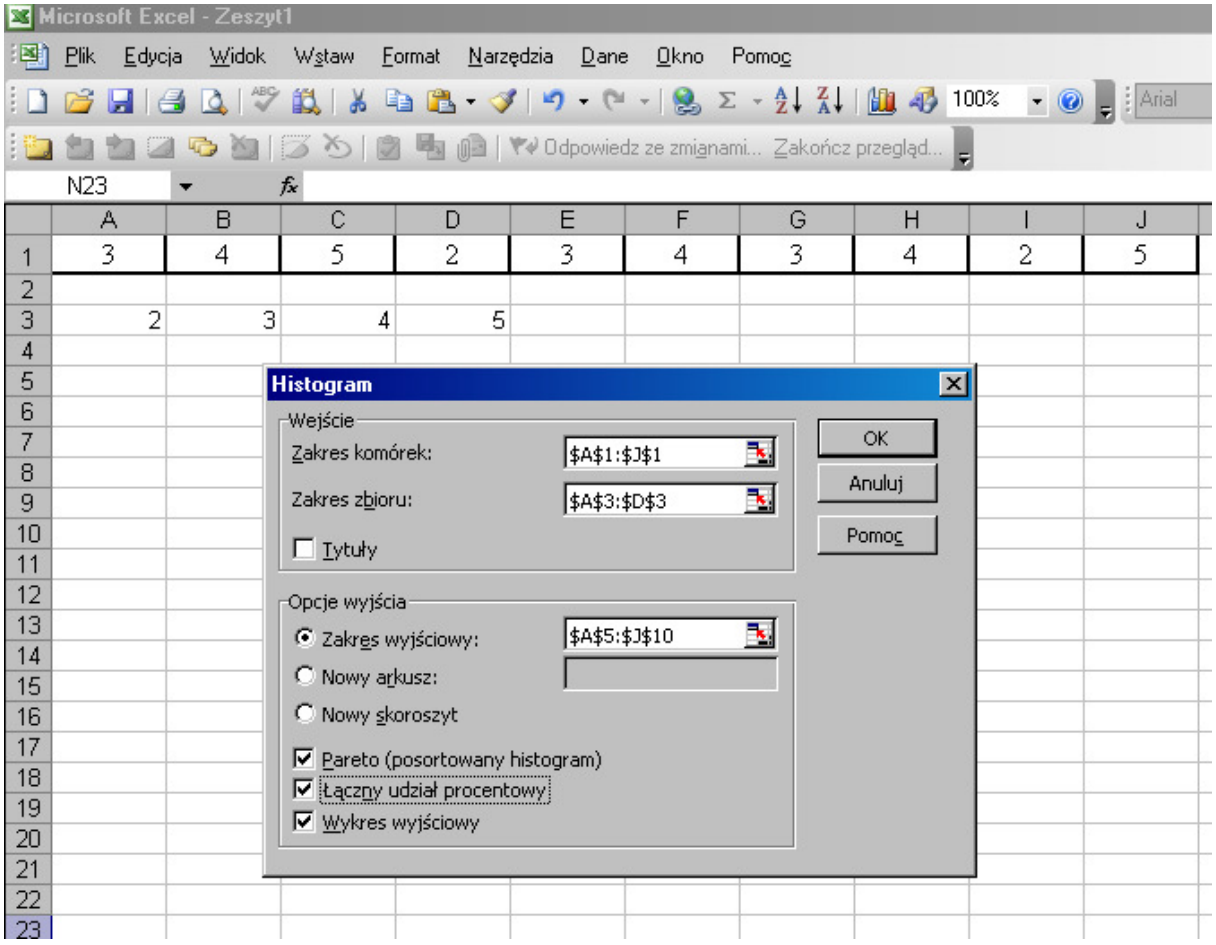
Populacja: Partia towaru licząca 1000 sztuk w tym 30 wadliwych.
Cecha populacji X: zmienna losowa przyjmująca 1, gdy sztuka jest wadliwa i wartość 0, gdy sztuka jest dobra. Wskaźnik struktury wariantu 1 (sztuka wadliwa) jest równy

$$w = \frac{30}{1000} = 3\%$$

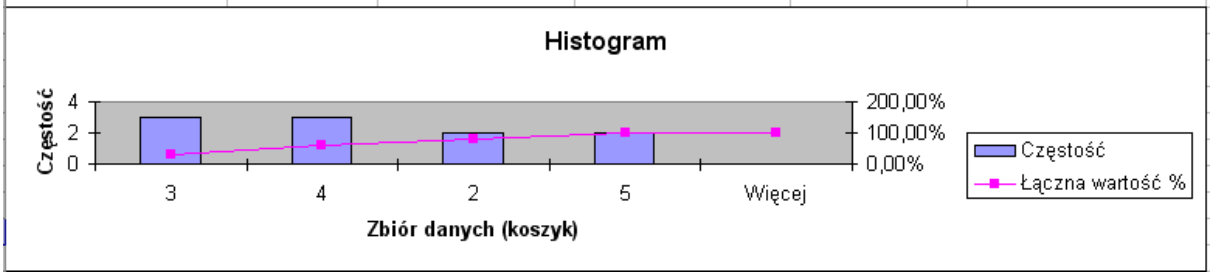
i w rozważanej sytuacji nazywa się *wadliwością towaru*, oznacza procent sztuk wadliwych w całej partii. ■

Przykład 2.15

Obliczymy częstości występowania wyników egzaminu podanych w przykładzie 2.1, korzystając
z arkusza kalkulacyjnego Excel co ilustrują poniższe rysunki.



Zbiór danych (koszyk)	Częstość	Łączna wartość %	Zbiór danych (koszyk)	Częstość	Łączna wartość %
2	2	20,00%	3	3	30,00%
3	3	50,00%	4	3	60,00%
4	3	80,00%	2	2	80,00%
5	2	100,00%	5	2	100,00%
Więcej	0	100,00%	Więcej	0	100,00%



2.4. Charakterystyki asymetrii[7]

2.4.1. Współczynniki asymetrii

Współczynnik asymetrii

$$a_k = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{s_x^3}$$

gdzie s jest odchyleniem standardowym, zaś licznik nazywa się momentem centralnym rzędu 3,

Wskaźnik asymetrii

$$a_s = \frac{\overline{x} - d}{s_x}$$

gdzie $\overline{x}$, d, s są odpowiednio średnią, dominantą i odchyleniem standardowym cechy X.

Jest to tzw. klasyczny miernik asymetrii standaryzowany.

Jeśli a_k i a_s są równe 0, to rozkład cechy X jest symetryczny, jeśli są różne od zera, to rozkład jest *asymetryczny*, przy czym, jeśli są dodatnie, to asymetria rozkładu jest *prawostronna*, jeśli są ujemne, to asymetria jest *lewostronna*.

Wartość bezwzględna współczynnika i wskaźnika asymetrii mierzy siłę asymetrii, im jest większa tym asymetria jest silniejsza.

Współczynnik i wskaźnik asymetrii są jednostkami niemianowanymi, mogą więc służyć do porównywania asymetrii cech populacji wyrażonych w różnych jednostkach.

Uwaga:
W pakiecie Excel współczynnik asymetrii można obliczyć za pomocą funkcji statystycznej SKOŚNOŚĆ w której stosowany jest nieco zmieniony wzór na współczynnik asymetrii

$$a'_K = \frac{1}{(n-1)(n-2)} \frac{\sum_{i=1}^{n}(x_i - \overline{x})^3}{s_x^3}$$

po to, aby zapewnić nieobciążoność współczynnika, pojęcie zostanie wyjaśnione w statystyce matematycznej.

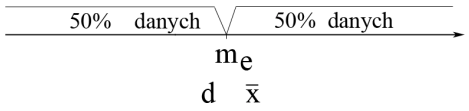
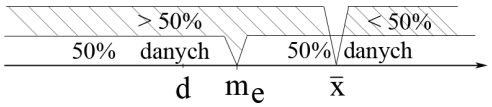
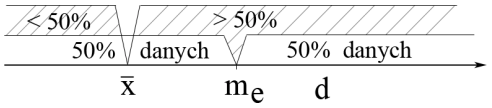
[7] Używana jest nazwa skośność.

2.4.2. Interpretacja symetrii w przypadku rozkładu jednomodalnego[8]

W tym przypadku mediana jest zawarta między średnią i dominantą, czyli prawdziwa jest jedna z poniższych nierówności podwójnych

$$\bar{x} \le m_e \le d \quad \text{lub} \quad d \le m_e \le \bar{x}$$

Zatem:

Jeśli cecha X populacji ma rozkład symetryczny, to średnia arytmetyczna, mediana i dominanta tej cechy są sobie równe $\bar{x} = m_e = d$, tzn. w ciągu uporządkowanych monotonicznie danych statystycznych, na lewo i na prawo od średniej jest tyle samo tych danych oraz średnia jest równa najczęściej występującej danej statystycznej (rys. 2.2).	 **Rys. 2.2. Rozkład symetryczny**
• Jeśli cecha X populacji ma rozkład asymetryczny o asymetrii prawostronnej (dodatniej), nazywany także rozkładem prawostronnie skośnym, to jest więcej danych statystycznych mniejszych od średniej niż danych statystycznych większych od tej średniej oraz najczęściej występująca dana statystyczna jest mniejsza od średniej (rys. 2.3).	 **Rys. 2.3. Rozkład o asymetrii prawostronnej**
• Jeśli cecha X populacji ma rozkład asymetryczny o asymetrii lewostronnej (ujemnej), nazywany także rozkładem o lewostronnie skośnym, to jest więcej danych statystycznych większych od średniej niż danych statystycznych mniejszych od tej średniej, oraz najczęściej występująca dana statystyczna jest większa od średniej (rys. 2.4).	 **Rys. 2.4.Rozkład o asymetrii lewostronnej**

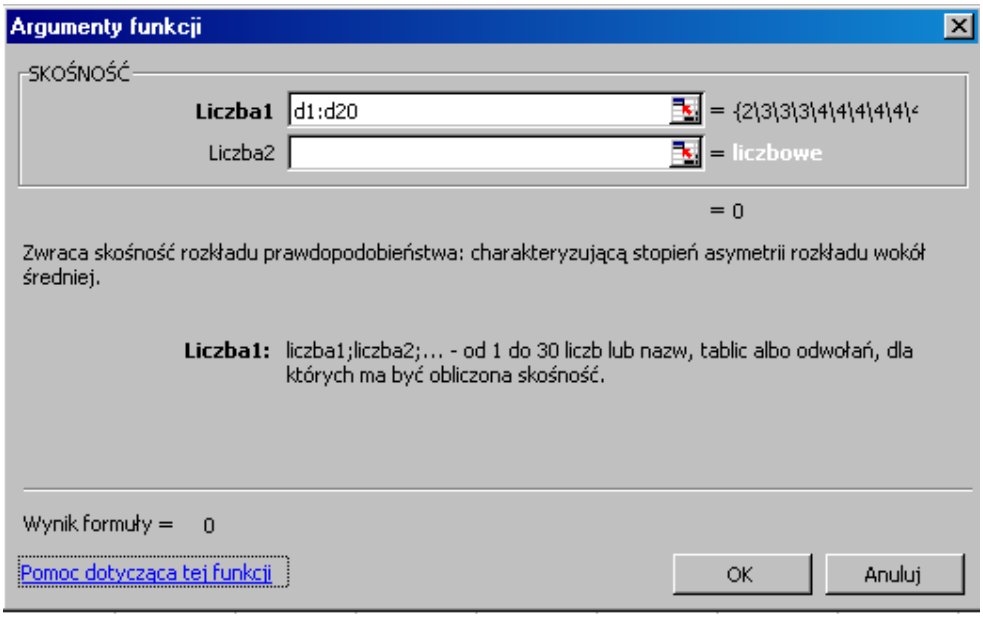
Przykład 2.16

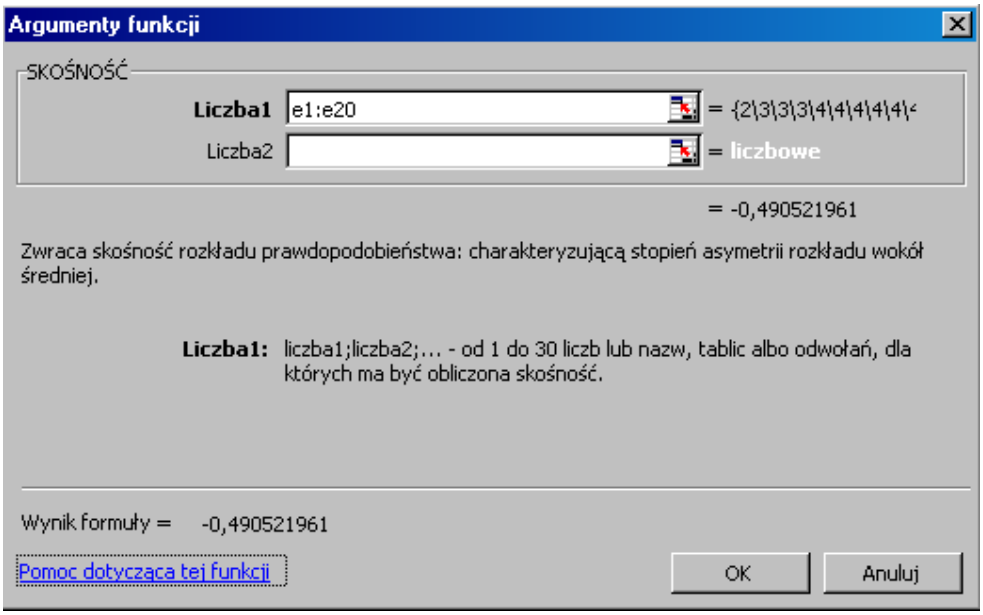
Ocenić kurtozę rozkładu ocen z egzaminu w dwóch grupach, które podano w poniższej tabeli.

	2	3	4	5	6
Grupa 1	1	3	12	3	1
Grupa 2	1	3	6	8	2

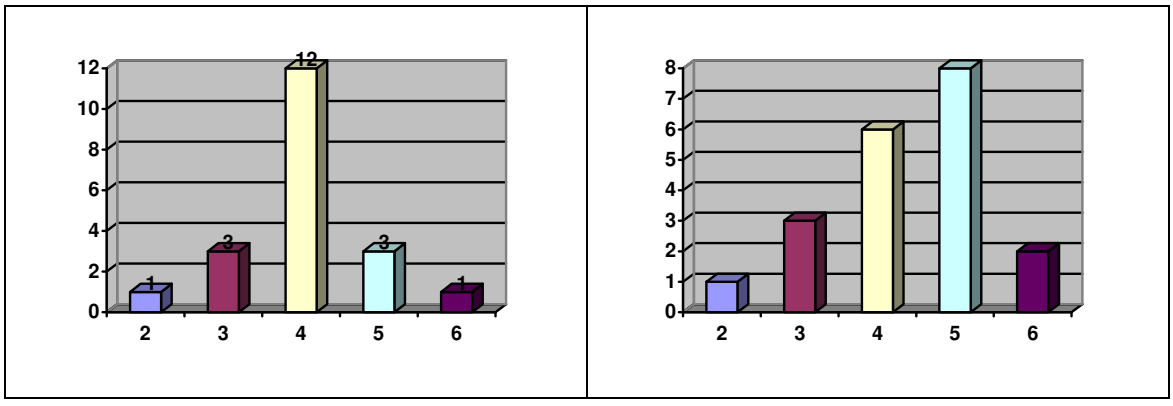
Wyniki obliczeń z wykorzystaniem funkcji statystycznej SKOSNOŚĆ.

[8] Rozkładu z tylko jedną dominującą wartością.





Rozkłady ocen przedstawiają poniższe rysunki.



I w końcu wyniki obliczenia charakterystyk liczbowych korzystając z narzędzia „Statystyka opisowa" pakietu Analysis ToolPak.

Kolumna1		*Kolumna2*	
Średnia	4	Średnia	4,35
Błąd standardowy	0,191943	Błąd standardowy	0,232549
Mediana	4	Mediana	4,5
Tryb	4	Tryb	5
Odchylenie standardowe	0,858395	Odchylenie standardowe	1,03999
Wariancja próbki	0,736842	Wariancja próbki	1,081579
Kurtoza	1,516807	Kurtoza	-0,03136
Skośność	0	Skośność	-0,49052
Zakres	4	Zakres	4
Minimum	2	Minimum	2
Maksimum	6	Maksimum	6
Suma	80	Suma	87
Licznik	20	Licznik	20
Największy(1)	6	Największy(1)	6
Najmniejszy(1)	2	Najmniejszy(1)	2
Poziom ufności(95,0%)	0,401741	Poziom ufności(95,0%)	0,48673

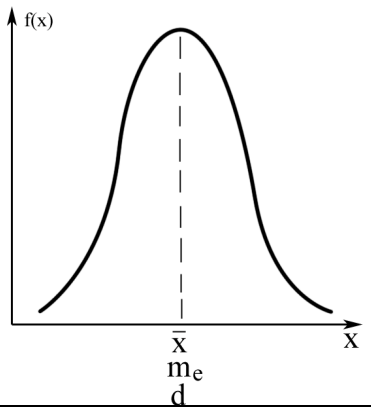
Rekomenduje się Czytelnikowi przeanalizowanie powyższych wyników. ■

2.4.3. Interpretacja asymetrii za pomocą wykresu szeregu rozdzielczego

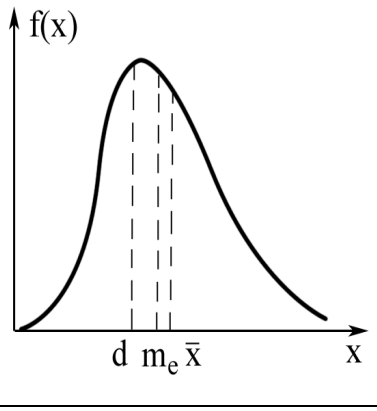
Za pomocą wykresu szeregu rozdzielczego łatwo określić istnienie asymetrii i jej znak, mianowicie:

- Jeśli wykres szeregu rozdzielczego cechy populacji jest symetryczny względem pewnej prostej prostopadłej do osi odciętych (prostej o równaniu postaci x = a), to cecha ta ma rozkład symetryczny - patrz rys. 2.2 i 2.5 (średnia, mediana i dominanta są równe a).
- Jeśli wykres szeregu rozdzielczego cechy populacji nie jest symetryczny względem żadnej prostej prostopadłej do osi odciętych i jego prawa część jest wydłużona, to cecha ta ma rozkład asymetryczny o asymetrii dodatniej, czyli prawostronnej (patrz rysunki 2.3, 2.6).
- Jeśli wykres szeregu rozdzielczego cechy populacji nie jest symetryczny względem żadnej prostej prostopadłej do osi odciętych i jego lewa część jest wydłużona, to cecha ta ma rozkład asymetryczny o asymetrii ujemnej, czyli lewostronnej patrz (rysunki 2.4. i 2.7).

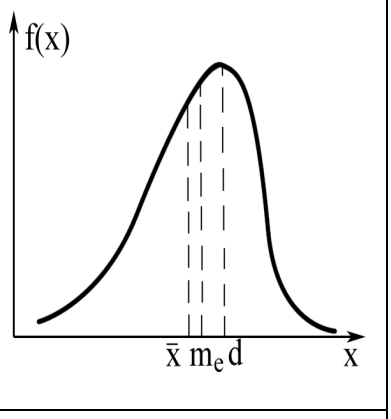
Poniższe trzy wykresy szeregów rozdzielczych dotyczą odpowiednio cechy o rozkładzie symetrycznym, asymetrycznym o asymetrii dodatniej i asymetrycznym o asymetrii ujemnej.



Rys. 2.5. Rozkład symetryczny



Rys. 2.6. Rozkład asymetryczny o asymetrii prawostronnej (dodatniej)



Rys. 2.7. Rozkład asymetryczny o asymetrii lewostronnej (ujemnej)

2.5. Charakterystyki spłaszczenia [9]

Miernik spłaszczenia

$$m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4$$

Współczynnik spłaszczenia (kurtoza)

$$k = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{s_x^4} - 3$$

Kurioza jest miarą skupienia wokół średniej arytmetycznej, im większa jest jej wartość, tym bardziej wartości zmiennej koncentrują się wokół średniej – miarą odniesienia jest rozkład normalny. Jeśli kurioza jest ujemna, to rozkład jest bardziej spłaszczony od normalnego[10], jeśli dodatnia, to rozkład jest bardziej wysmukły niż normalny.

Uwaga:

W pakiecie Excel współczynnik asymetrii można obliczyć za pomocą funkcji statystycznej KURTOZA w której stosowany jest nieco zmieniony wzór

$$k' = \frac{n(n+1)}{(n-1)(n-2)(n-3)}\frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{s_x^4} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

po to, aby zapewnić nieobciążoność współczynnika, pojęcie zostanie wyjaśnione w statystyce matematycznej.

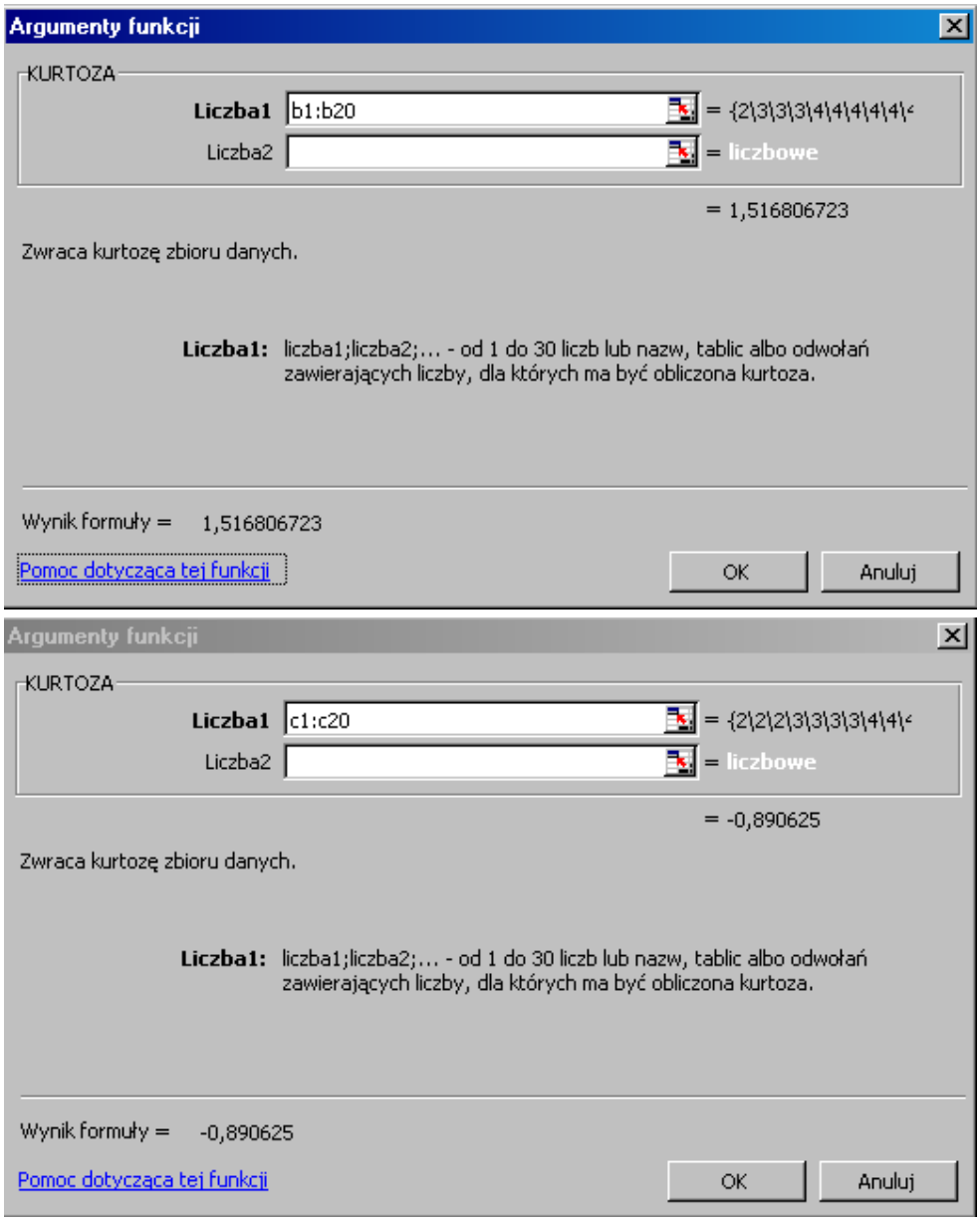
Przykład 2.17

Ocenić kurtozę rozkładu ocen z egzaminu w dwóch grupach, które podano w poniższej tabeli.

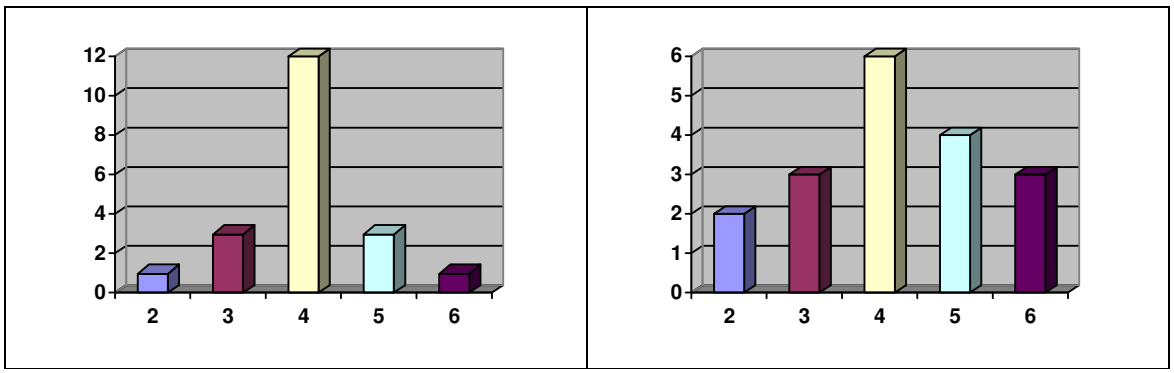
	2	3	4	5	6
Grupa 1	1	3	12	3	1
Grupa 2	2	3	6	4	3

[9] Inna nazwa to charakterystyki ekscesu.
[10] Patrz punkt 8.2.2.

Wyniki obliczeń z wykorzystaniem funkcji statystycznej KURTOZA.



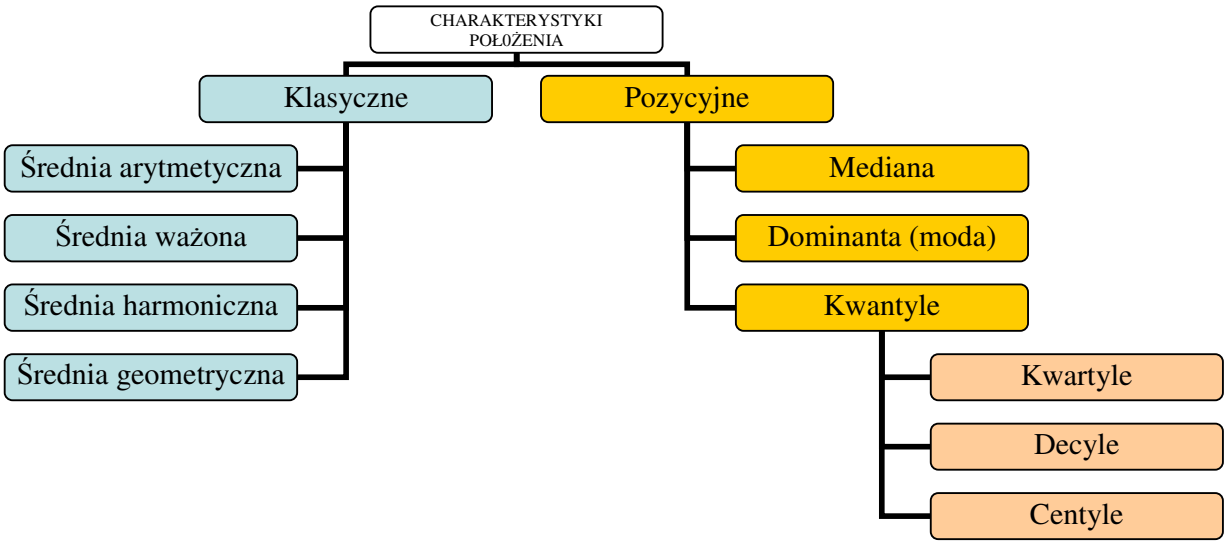
Rozkłady ocen przedstawiają poniższe rysunki.



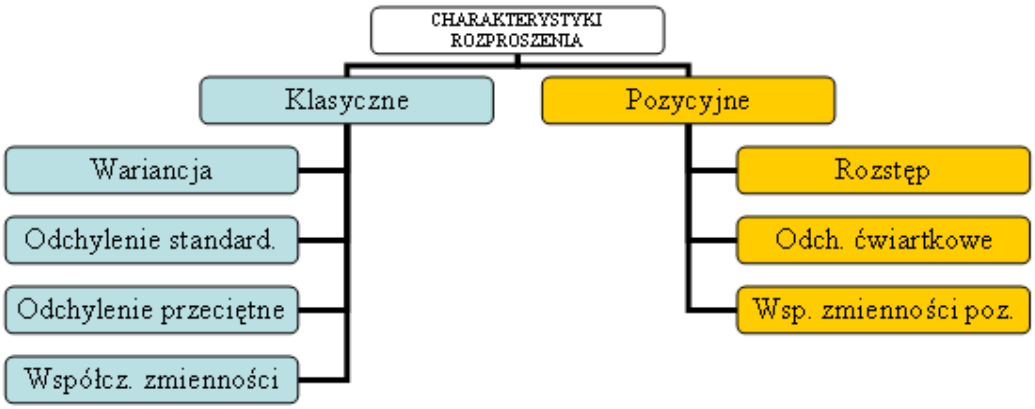
■

2.6. Podsumowanie

2.6.1. Wybrane charakterystyki liczbowe w postaci graficznej



Rysunek 2.8. Charakterystyki położenia



Rysunek 2.9. Charakterystyki rozproszenia

2.6.2. Możliwości obliczania charakterystyk liczbowych w zależności od skali

RODZAJ CHARAKTERYSTYKI	NAZWA CHARAKTERYSTYKI	SKALA		
		Nominalna	Porządkowa[11]	Przedziałowa
Miary położenia	Średnia arytmetyczna			+
	Średnia harmoniczna			+
	Średnia geometryczna			+
	Dominanta (moda)	+	+	+
	Kwantyle		+	+
	Mediana		+	+
Miary zróżnicowania	Wariancja			+
	Odchylenie standardowe			+
	Odchylenie przeciętne			+
	Rozstęp		+	+
Miary asymetrii (skośności)	Miernik asymetrii klasyczny			+
Miary spłaszczenia	Współczynnik spłaszczenia			+

2.6.3. Możliwości obliczania charakterystyk liczbowych w arkuszu Excel

Lp	Charakterystyki liczbowe	Funkcje statystyczne	STATYSTYKA OPISOWA
1.	Średnia arytmetyczna	ŚREDNIA	+
2.	Mediana	MEDIANA	
3.	Dominanta	WYST.NAJCZESCIEJ	+
4.	Średnia ważona	SUMA.ILOCZYNÓW	
5.	Średnia ucinana	ŚREDNIA.WEWN	
6.	Średnia geometryczna	ŚREDNIA.GEOMETRYCZNA	
7.	Średnia harmoniczna	ŚREDNIA.HARMONICZNA	
8.	Wariancja	WARIANCJA	+
9.	Odchylenie standardowe	ODCH.STANDARDOWE	+
10.	Kwartle	KWARTYL	+
11.	Współczynnik asymetrii	SKOŚNOŚĆ	+
12.	Współczynnik spłaszczenia	KURTOZA	+

[11] Działania na rangach nie maja uzasadnienia

2.7. Przykłady analizy statystycznej danych

Zakładamy, że cecha X populacji jest mierzalna. Aby poznać strukturę tej cechy należy zgromadzić i opracować dane statystyczne. Opracowanie danych statystycznych polega na ich prezentacji (tabelarycznej i graficznej) oraz obliczeniu charakterystyk liczbowych.

Podamy przykłady analizy gdy cecha X jest skokowa o umiarkowanej liczbie wariantów (do 25)[12]. Danych statystycznych jest znacznie więcej niż wariantów. Z powyższych założeń wynika, że niektóre warianty cechy muszą się powtarzać.

Oznaczenia

X - cecha populacji,
r - liczba wariantów,
$w_1, w_2, ..., w_r$ - warianty cechy X,
n - liczba danych statystycznych,
n_i - liczebność wariantu w_i (ile razy powtarza się wariant w_i)

Prezentacja danych statystycznych

- Tabelaryczna - za pomocą szeregu statystycznego punktowego

Wariant w_i	Liczebność n_i
w_1	n_1
w_2	n_2
…	…
w_r	n_r
Suma	n

- graficzna - wykres szeregu punktowego

Charakterystyki liczbowe

Wzory na średnią arytmetyczną, wariancję i współczynnik asymetrii przybierają teraz postać:

Średnia arytmetyczna	Wariancja	Współczynnik asymetrii
$\overline{x} = \frac{1}{n}\sum_{i=1}^{r} n_i w_i$	$s_x^2 = \frac{1}{n}\sum_{i=1}^{r} n_i (w_i - \overline{x})^2$	$a_k = \frac{\frac{1}{n}\sum_{i=1}^{r} n_i \left(w_i - \overline{x}\right)^3}{s_x^3}$

Przykład 2.18

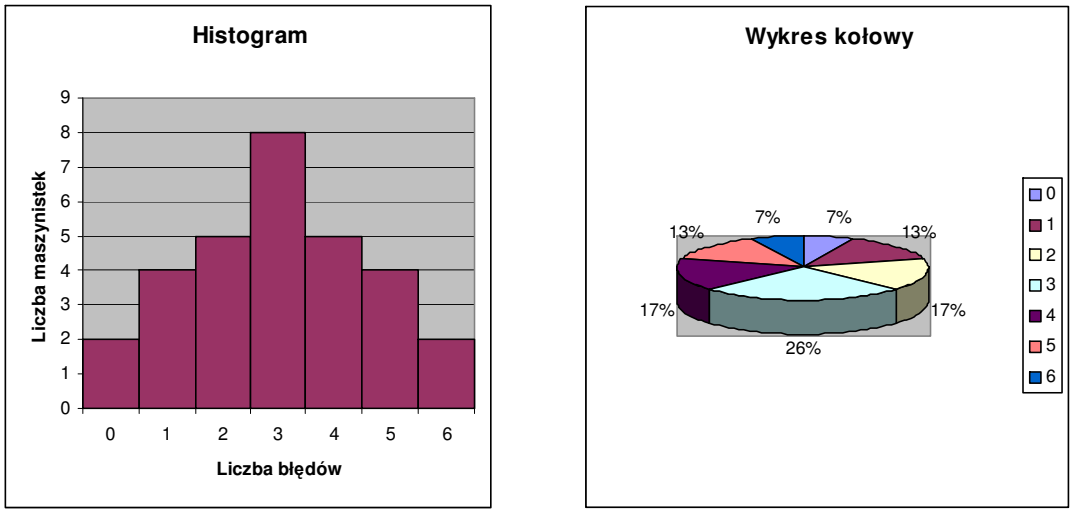
Badano liczbę błędów w kodzie źródłowym 30 programistów (cecha X populacji). Otrzymano następujące wyniki: 3, 2, 1, 3, 4, 5, 3, 1, 0, 2, 6, 3, 4, 5, 3, 1, 5, 3, 0, 1, 2, 2, 4, 3, 4, 4, 3, 2, 6, 5.
Opracujemy te dane.

Prezentacja tabelaryczna: Szereg statystyczny punktowy

Liczba błędów w_i	0	1	2	3	4	5	6	Razem
Liczebność n_i	2	4	5	8	5	4	2	30

[12] Jeżeli cecha ma rozkład skokowy i wariantów jest dużo lub ma rozkład ciągły dane statystyczne grupujemy w klasach, których liczba zależy od ilości danych. Tym przypadkiem nie będziemy się zajmować.

Prezentacja graficzna



Rys. 2.10 Prezentacje graficzne danych

Charakterystyki liczbowe

Liczba błędów w_i	Liczebność n_i	$n_i w_i$	Liczebność skumulowana[13] s_i	$n_i(w_i - \overline{x})^2$
0	2	0	2	18
1	4	4	6	16
2	5	10	11	5
3	8	24	19	0
4	5	20	24	5
5	4	20	28	16
6	2	12	30	18
Razem	30	90		78

Charakterystyki tendencji centralnej	Charakterystyki zróżnicowania
$\overline{x} = \frac{1}{n}\sum_{i=1}^{r} n_i w_i = \frac{90}{30} = 3$	$s_x^2 = 2{,}6$
$m_e = \frac{1}{2}\left[x_{(15)} + x_{(16)}\right] = \frac{1}{2}[3+3] = 3$ - patrz [14]	s = 1,61
d = 3	$r_0 = 6 - 0 = 6$
	v = 53,3 %

Przedział typowych jednostek populacji <1,39 ; 4,61>. Do tego przedziału należą programiści, którzy popełnili 2, 3 lub 4 błędy. Jest ich 18.

Rozkład cechy jest symetryczny, bo $\overline{x} = m_e = d$, więc wskaźnik asymetrii $a_1 = 0$.

Histogram jest symetryczny względem prostej x = 3. ■

[13] Suma liczebności danych statystycznych równych wariantowi w_i oraz liczebności wszystkich wariantów < w_i.

[14] $x_{(15)}$ i $x_{(16)}$ oznaczają piętnasty i szesnasty wynik w ciągu uporządkowanych niemalejąco danych. Z czwartej kolumny tabeli wynika , że $x_{(12)}$ do $x_{(19)}$ są równe 3.

Przykład 2.19

Dane dotyczące działu pewnej firmy informatycznej. Przyjęte oznaczenia.

Płeć	Wykształcenie	Specjalność	Stanowisko	Ocena roczna
1 - kobieta 2 - mężczyzna	1 – średnie 2 – studia 1 stopnia 3 – studia 2 stopnia 4 – studia 3 stopnia	1– Tester 2 – Grafik 3 – Programista 4 – Analityk 5 – Projektant	1 – pracownik 2 – kierownik 3 – dyrektor	1 – niedostateczna 2 – dostateczna 3 – dobra 4 – bardzo dobra 5 – wzorowa

Dane

Lp	Wiek	Płeć	Wykształcenie	Specjalność	Staż	Stanowisko	Zarobki	Ocena roczna
1.	23	2	1	1	4	1	2000	3
2.	25	2	2	1	2	1	2000	2
3.	24	2	1	2	1	1	2500	3
4.	30	2	3	3	4	1	3000	5
5.	26	2	3	3	2	1	3000	4
6.	25	1	2	3	3	1	3000	4
7.	27	2	3	3	5	1	3000	3
8.	31	2	3	3	5	2	4000	5
9.	26	1	2	3	3	2	4000	4
10.	35	2	3	4	8	1	3000	4
11.	37	1	3	4	10	1	3500	4
12.	37	1	4	4	7	3	5000	4
13.	38	2	3	5	9	1	3500	4
14.	39	1	3	5	2	1	3500	3
15.	39	2	4	5	5	2	4000	4

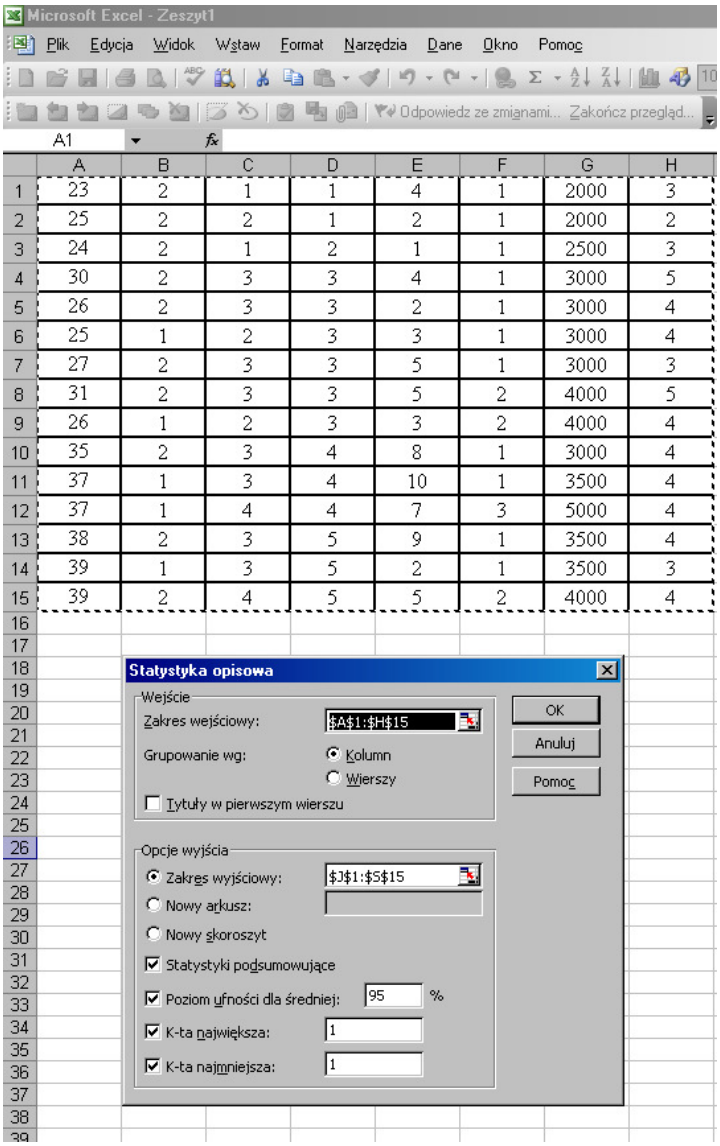
Obliczanie w Excelu wariancji oraz skośności i kurtozy

	Wiek	Wiek-sr	(Wiek-sr)**2	(Wiek-sr)**3	(Wiek-sr)**4	Wiek-sr
	23	-7,800000	60,840000	-474,552000	3701,505600	7,800000
	25	-5,800000	33,640000	-195,112000	1131,649600	5,800000
	24	-6,800000	46,240000	-314,432000	2138,137600	6,800000
	30	-0,800000	0,640000	-0,512000	0,409600	0,800000
	26	-4,800000	23,040000	-110,592000	530,841600	4,800000
	25	-5,800000	33,640000	-195,112000	1131,649600	5,800000
	27	-3,800000	14,440000	-54,872000	208,513600	3,800000
	31	0,200000	0,040000	0,008000	0,001600	0,200000
	26	-4,800000	23,040000	-110,592000	530,841600	4,800000
	35	4,200000	17,640000	74,088000	311,169600	4,200000
	37	6,200000	38,440000	238,328000	1477,633600	6,200000
	37	6,200000	38,440000	238,328000	1477,633600	6,200000
	38	7,200000	51,840000	373,248000	2687,385600	7,200000
	39	8,200000	67,240000	551,368000	4521,217600	8,200000
	39	8,200000	67,240000	551,368000	4521,217600	8,200000
Suma	462	0,000000	516,400000	570,960000	24369,808000	80,800000
			36,885714	0,210058	-1,800917051	5,386667
			Wariancja /(n-1)	Skosnosc nieobciazona	Kurtoza nieobciążona	Odch. przec
			6,073361037			
			Odchylenie			

Wyniki dotyczące wieku otrzymane funkcjami statystycznymi Excela

Srednia	30,8		Kurtoza	-1,80092
Mediana	30		Skosnosc	0,210058
Minimum	23		Moda	25
Maksimum	39		Percentyl 0,5	30
Wariancja	36,88571		Licznosc	15

Obliczymy teraz charakterystyki liczbowe korzystając z narzędzia „Statystyka opisowa” pakietu Analysis ToolPak. Dane i wpisane parametry narzędzia podano poniżej.



	A	B	C	D	E	F	G	H
1	23	2	1	1	4	1	2000	3
2	25	2	2	1	2	1	2000	2
3	24	2	1	2	1	1	2500	3
4	30	2	3	3	4	1	3000	5
5	26	2	3	3	2	1	3000	4
6	25	1	2	3	3	1	3000	4
7	27	2	3	3	5	1	3000	3
8	31	2	3	3	5	2	4000	5
9	26	1	2	3	3	2	4000	4
10	35	2	3	4	8	1	3000	4
11	37	1	3	4	10	1	3500	4
12	37	1	4	4	7	3	5000	4
13	38	2	3	5	9	1	3500	4
14	39	1	3	5	2	1	3500	3
15	39	2	4	5	5	2	4000	4

Część 1 otrzymanych wyników.

Wiek		*Płeć*		*Wykształcenie*		*Specjalność*	
Średnia	30,8	Średnia	1,666666667	Średnia	2,666666667	Średnia	3,266667
Błąd standardowy	1,568135077	Błąd standardowy	0,125988158	Błąd standardowy	0,232310684	Błąd standardowy	0,330464
Mediana	30	Mediana	2	Mediana	3	Mediana	3
Moda	25	Moda	2	Moda	3	Moda	3
Odchylenie standardowe	6,073361037	Odchylenie standardowe	0,487950036	Odchylenie standardowe	0,899735411	Odchylenie standardowe	1,279881
Wariancja próbki	36,88571429	Wariancja próbki	0,238095238	Wariancja próbki	0,80952381	Wariancja próbki	1,638095
Kurtoza	-1,800917051	Kurtoza	-1,615384615	Kurtoza	-0,005589566	Kurtoza	-0,35601
Skośność	0,210057532	Skośność	-0,788226982	Skośność	-0,578350018	Skośność	-0,3393
Zakres	16	Zakres	1	Zakres	3	Zakres	4
Minimum	23	Minimum	1	Minimum	1	Minimum	1
Maksimum	39	Maksimum	2	Maksimum	4	Maksimum	5
Suma	462	Suma	25	Suma	40	Suma	49
Licznik	15	Licznik	15	Licznik	15	Licznik	15
Największy(1)	39	Największy(1)	2	Największy(1)	4	Największy(1)	5
Najmniejszy(1)	23	Najmniejszy(1)	1	Najmniejszy(1)	1	Najmniejszy(1)	1
Poziom ufności(95,0%)	3,363315227	Poziom ufności(95,0%)	0,270217723	Poziom ufności(95,0%)	0,498256861	Poziom ufności(95,0%)	0,708774

STATYSTYKA OPISOWA

Część 2 otrzymanych wyników.

Staż		*Stanowisko*		*Zarobki*		*Ocena roczna*	
Średnia	4,666667	Średnia	1,333333	Średnia	3266,667	Średnia	3,7333333
Błąd standardowy	0,708228	Błąd standardowy	0,159364	Błąd standardowy	206,2515	Błąd standardowy	0,2062515
Mediana	4	Mediana	1	Mediana	3000	Mediana	4
Moda	2	Moda	1	Moda	3000	Moda	4
Odchylenie standardowe	2,742956	Odchylenie standardowe	0,617213	Odchylenie standardowe	798,8086	Odchylenie standardowe	0,7988086
Wariancja próbki	7,52381	Wariancja próbki	0,380952	Wariancja próbki	638095,2	Wariancja próbki	0,6380952
Kurtoza	-0,5407	Kurtoza	2,625	Kurtoza	0,39314	Kurtoza	0,379646
Skośność	0,666485	Skośność	1,791551	Skośność	0,294102	Skośność	-0,4153717
Zakres	9	Zakres	2	Zakres	3000	Zakres	3
Minimum	1	Minimum	1	Minimum	2000	Minimum	2
Maksimum	10	Maksimum	3	Maksimum	5000	Maksimum	5
Suma	70	Suma	20	Suma	49000	Suma	56
Licznik	15	Licznik	15	Licznik	15	Licznik	15
Największy(1)	10	Największy(1)	3	Największy(1)	5000	Największy(1)	5
Najmniejszy(1)	1	Najmniejszy(1)	1	Najmniejszy(1)	2000	Najmniejszy(1)	2
Poziom ufności(95,0%)	1,518999	Poziom ufności(95,0%)	0,341801	Poziom ufności(95,0%)	442,3655	Poziom ufności(95,0%)	0,4423655

Dla każdej cechy wyznaczone zostały takie same wyniki, nie wszystkie mają sens – patrz „Możliwość obliczania charakterystyk liczbowych zbiorowości”

2.8. Analiza danych przedstawionych w postaci szeregu rozdzielczego przedziałowego

Cecha populacji X ma rozkład skokowy i wariantów jest dużo (>25) lub ma rozkład ciągły.

2.8.1. Prezentacja danych statystycznych

Prezentacja tabelaryczna - szereg rozdzielczy przedziałowy. Dane statystyczne grupujemy w r klasach

Klasa $<a_i ; a_{i+1})$	Liczebność klasy n_i
$<a_1 ; a_2)$	n_1
$<a_2 ; a_3)$	n_2
...	...
$<a_r ; a_{r+1})$	n_r
Suma	n

2.8.2. Charakterystyki liczbowe

Charakterystyki położenia	Charakterystyki rozproszenia
Średnia arytmetyczna $\bar{x}=\frac{1}{n}\sum_{i=1}^{r} n_i \tilde{x}_i$ gdzie $\tilde{x}_i$ - środek klasy o numerze i	Wariancja $s^2=\frac{1}{n}\sum_{i=1}^{r} n_i (\tilde{x}_i-\bar{x})^2$
Mediana $m_e=a_k+\frac{b}{n_k}\left(\frac{n}{2}-s_{k-1}\right)$ a_k - lewy koniec klasy mediany[15] b - długość klasy mediany, n_k - liczebność klasy mediany, s_{k-1} - liczebność skumulowana klasy poprzedzającej klasę mediany	Odchylenie standardowe $s=\sqrt{s^2}$ Współczynnik zmienności $v=\frac{s}{\bar{x}}$ Rozstęp $r_o=a_{r+1}-a_1$
Dominanta $d=a_k+b\frac{n_k-n_{k-1}}{2n_k-n_{k-1}-n_{k+1}}$ a_k - lewy koniec klasy dominanty, b - długość klasy dominanty, n_k - liczebność klasy dominanty, n_{k-1} - liczebność klasy poprzedzającej klasę dominanty, n_{k+1} - liczebność klasy następującej po klasie dominanty.	

Przedział typowych jednostek populacji

$$<\bar{x}-s\ ;\ \bar{x}+s>$$

Asymetria - wskaźnik asymetrii

$$a_1=\frac{\bar{x}-d}{s}$$

Współczynnik asymetrii

$$a_1=\frac{\frac{1}{n}\sum_{i=1}^{r} n_i\left(\tilde{x}_i-\bar{x}\right)^3}{s^3}$$

[15] tj. klasy, do której należy mediana

Uwagi.

1. Stosując powyższe wzory obliczamy jedynie w przybliżeniu wartości charakterystyk opisowych, gdyż obliczenia nie są wykonywane przy pomocy indywidualnych danych statystycznych.
2. Dominanta nie może być obliczona z podanego wzoru gdy najbardziej liczna jest klasa pierwsza lub ostatnia, a także w przypadku, gdy klasa najliczniejsza nie istnieje.

Przykład 2.20

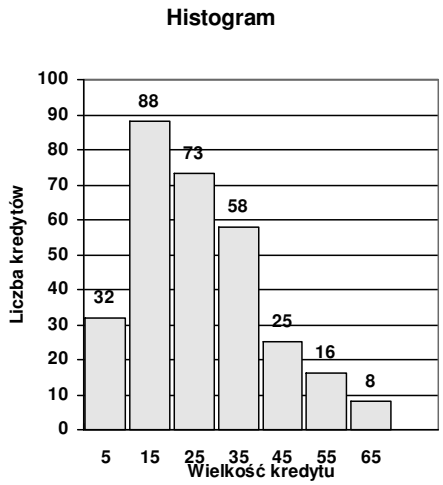
Badano wysokości kredytów w tysiącach złotych udzielonych przez pewien bank w ciągu lutego 2005 r. Otrzymane dane są przedstawione w szeregu rozdzielczym przedziałowym.

Wysokość kredytu	0 –10	10 – 20	20 – 30	30 – 40	40 – 50	50- 60	60 – 70	Razem
Liczba kredytów	32	88	73	58	25	16	8	300

Opracujemy te dane.

Rozwiązanie

Prezentacja graficzna



Rys. 2.11 Histogram wielkości kredytów

Charakterystyki liczbowe

Nr klasy i	Klasa $<a_i;a_{i+1})$	Liczebność n_i	Liczebność skumulowana S_i	Środek klasy $\tilde{x}_i$	$n_i\tilde{x}_i$	$n_i(\tilde{x}_i-\overline{x})^2$
1	0 - 10	32	32	5	160	14382,08
2	10 - 20	88	120	15	1320	11038,72
3	20 - 30	73	193	25	1825	105,12
4	30 - 40	58	251	35	2030	4491,52
5	40 - 50	25	276	45	1125	8836,00
6	50 - 60	16	292	55	880	13271,04
7	60 - 70	8	300	65	520	12043,52
	Suma	300			7860	64168

Charakterystyki tendencji centralnej

Średnia arytmetyczna $\bar{x} = 26{,}2$ tys. zł

Mediana $m_e = 20 + \frac{10}{73}\left(\frac{300}{2} - 120\right) =$ 24,11 tys. zł

Dominanta $d = 10 + 10 \cdot \frac{88 - 32}{2 \cdot 88 - 32 - 73} = 17{,}89$ tys. zł

Miary zróżnicowania

Wariancja $s^2 = 213{,}89$ tys. zł^2

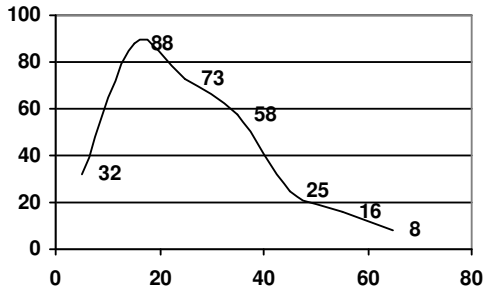
Odchylenie standardowe $s = 14{,}63$ tys. zł

Rozstęp $r_0 = 70 - 0 = 70$ tys. zł

Współczynnik zmienności $v = 56\%$

Przedział typowych wielkości kredytów $< 11{,}6 \ ; \ 40{,}8>$

Wskaźnik asymetrii $a_1 = 0{,}57$ (asymetria prawostronna – wydłużona prawa część wykresu rys. poniżej – także rysunek 10.2).



Rys. 2.12 Wykres liczebności kredytobiorców dla wyróżnionych wysokości kredytów ■

3. BADANIE ZALEŻNOŚCI CECH POPULACJI

3.1. Wprowadzenie

Badamy populację ze względu na dwie cechy X i Y, czyli ze względu na parę cech (X, Y). Przedstawimy metody badania pozwalające stwierdzić, czy cechy te są zależne i jak silnie (analiza korelacji) oraz jakim wzorem można przedstawić tę zależność o ile ona istnieje i jest dostatecznie silna (analiza regresji).

Omówimy najpierw stosowne pojęcia.

3.1.1. Dane statystyczne dwóch cech populacji

Badamy populację ze względu na parę cech (X,Y). Każdemu elementowi populacji lub próby przyporządkowujemy parę (x, y), gdzie x jest wartością cechy X, y wartością cechy Y badanego elementu. Pary te nazywamy danymi statystycznymi pary (X,Y) cech populacji.

Ograniczymy nasze rozważania do przypadku, w którym danych statystycznych jest skończenie wiele. Ich liczbę oznaczymy n.

3.1.2. Prezentacja danych statystycznych pary cech populacji

Prezentacji tabelarycznej danych statystycznych pary cech (X, Y) dokonujemy za pomocą szeregu statystycznego lub tablicy korelacyjnej, natomiast prezentacji graficznej za pomocą wykresu szeregu statystycznego lub wykresu tablicy korelacyjnej.

Szereg statystyczny pary cech (X, Y) jest to tabela

i	x_i	y_i
1	x_1	y_1
2	x_2	y_2
…	…	…
n	x_n	y_n

w której występują wszystkie dane statystyczne i są uporządkowane wg pewnego kryterium.

Tablica korelacyjna pary cech (X,Y), gdzie X i Y są skokowe o niezbyt dużej liczbie wariantów (do 20) i jest wiele danych statystycznych, to tablica postaci

w_i \ v_j	v_1	v_2	...	v_s
w_1	n_{11}	n_{12}	…	n_{1s}
w_2	n_{21}	n_{22}	…	n_{2s}
w_r	n_{r1}	n_{r2}	…	n_{rs}

gdzie:

r - liczba wariantów cechy X,

$w_1, w_2, \ w_r$ - warianty cechy X,

s - liczba wariantów cechy Y,

$v_1, v_2, \ v_s$ - warianty cechy Y,

n_{ij} - liczba danych statystycznych równych parze wariantów (w_i, v_j).

Tak więc w boczku tablicy korelacyjnej znajdują się warianty cechy X, natomiast w główce warianty cechy Y, zaś w komórkach - liczby danych statystycznych, których wartość cechy X

jest równa wariantowi znajdującemu się w wierszu komórki oraz wartość cechy Y jest równa wariantowi znajdującemu się w kolumnie komórki.

Jeśli cechy X i Y populacji są ciągłe lub skokowe o dużej liczbie wariantów (powyżej 20), to warianty tych cech dzielimy na klasy zgodnie z procedurą przedstawioną w punkcie 10.3. Wówczas w boczku i główce tablicy korelacyjnej umieszcza się klasy poszczególnych cech. Za pomocą szeregu statystycznego można prezentować dane statystyczne niezależnie od ich rodzaju i liczebności danych statystycznych. Jednak, gdy tych danych jest dużo (ponad 20), to prezentacja ta nie jest przejrzysta. Dlatego dane statystyczne przedstawiamy wtedy w tablicy korelacyjnej.

W poniższych przykładach pokazujemy konstrukcję szeregów statystycznych i tablic korelacyjnych oraz prezentację graficzną danych statystycznych.

Przykład 2.21

W 15 osobowej grupie studentów informatyki przeprowadzono badanie ze względu na parę cech (X, Y), X- ocena końcowa z matematyki, Y - ocena końcowa ze statystyki. Otrzymano wyniki: (3,4), (4,4), (5,5), (5,4), (2,2), (2,3), (2,2), (3,4), (3,3), (3,2), (2,3), (4,5), (3,3), (2,2), (4,4).

Dokonamy prezentacji tabelarycznej i graficznej otrzymanych danych statystycznych.

Szereg statystyczny

Porządkując dane niemalejąco ze względu na wartości cechy X i niemalejąco ze względu na wartości cechy Y, gdy wartości cechy X są jednakowe, otrzymujemy szereg statystyczny

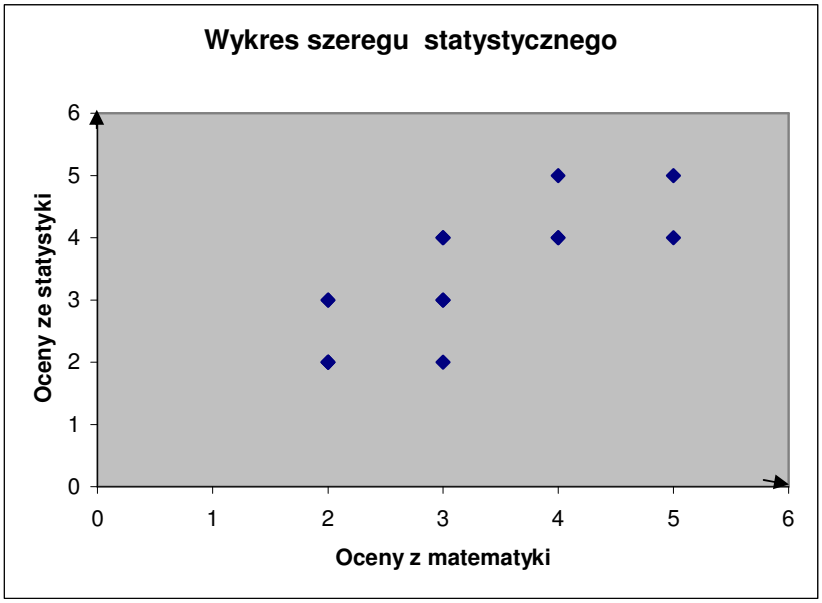
x_i	2	2	2	2	2	3	3	3	3	3	4	4	4	5	5
y_i	2	2	2	3	3	2	3	3	4	4	4	4	5	4	5

Tablica korelacyjna

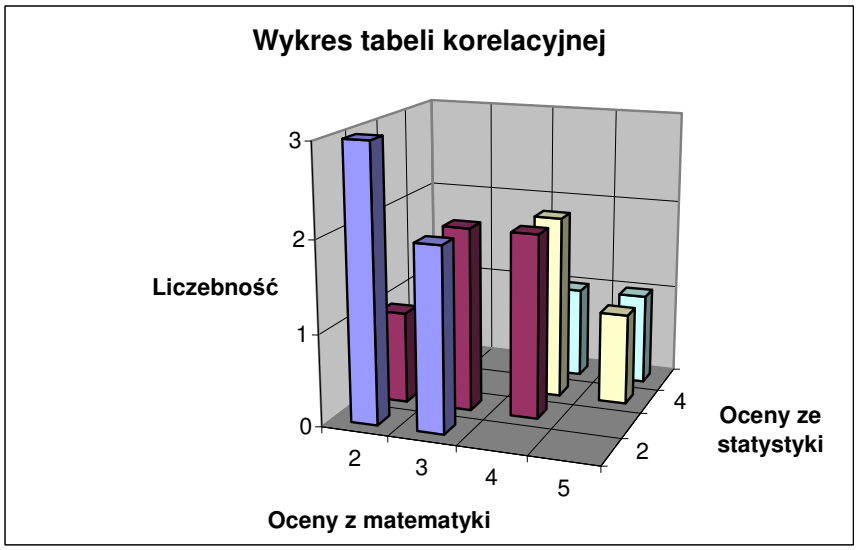
Ponieważ wiele danych jest jednakowych, można więc utworzyć tablicę korelacyjną pomimo małej liczby wszystkich danych statystycznych.

w_i \ v_j	2	3	4	5
2	3	2		
3	1	2	2	
4			2	1
5			1	1

Prezentacja graficzna



Rys. 2.13. Wykres szeregu statystycznego dla danych z przykładu 2.21



Rys. 2.14. Wykres tablicy korelacyjnej dla danych z przykładu 2.21

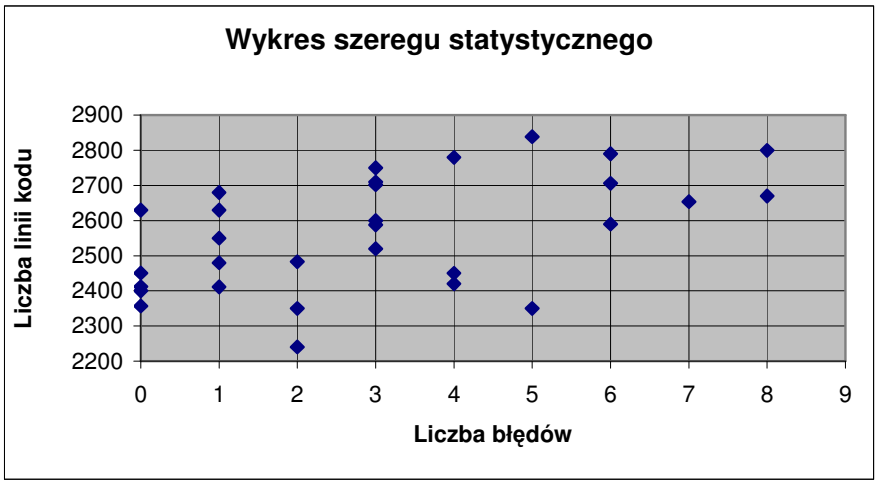
■

Przykład 2.22

Populację 30 testerów oprogramowania, które w ciągu czasu T sprawdzali poprawność postaci źródłowej oprogramowania, badano ze względu na parę cech (X,Y), X – liczba niewykrytych błędów w sprawdzanym programie, Y – liczba sprawdzonych linii kodu. Otrzymane wyniki przedstawione są w poniższym szeregu statystycznym.

x_i	2	2	5	0	0	1	0	4	4	0	1	2	3	1	3
y_i	2240	2350	2350	2357	2400	2411	2412	2420	2450	2451	2480	2483	2520	2550	2588
x_i	6	3	0	1	7	8	1	3	6	3	3	4	6	8	5
y_i	2590	2600	2630	2630	2654	2670	2680	2702	2706	2710	2750	2780	2790	2800	2839

Przedstawimy te dane w tablicy korelacyjnej. W tym celu najpierw podzielimy warianty cechy Y na 6 klas.



Rys 2.15. Wykres szeregu dla danych z przykładu 12.12

	Cecha Y
Liczba klas	$r = 6$
Najmniejsza wartość	$y_{min} = 2240$
Największa wartość	$y_{max} = 2839$
Rozstęp	$r_o = y_{max} - y_{min} = 599$
Dokładność danych	$\alpha = 1$
Długość klasy	$b = r_o / r \approx 100$ [16]
Lewy koniec 1 klasy	$a_1 = y_{min} + \alpha/2 = 2239,5$
Prawy koniec 1 klasy	$a_2 = a_1 + b = 2339,5$

Końce pozostałych klas otrzymujemy dodając kolejno do końców a_1 i a_2 pierwszej klasy długość przedziału b.

Ω_i \ B_j	2239,5:2339,5	2339,5:2439,5	2439,5:2539,5	2539,5:2639,5	2639,5:2739,5	2739,5:2839,5
0	1	3	1	1	1	
1		1	1	2		
2		1	1	2		
3			1		2	1
4		1	1			1
5		1				1
6				1	1	1
7					1	
8					1	1

■

[16] Do wyniku dzielenia dodano dokładność danych, czyli 50 (uogólniona zasada zaokrąglania w górę), powoduje to, że wszystkie dane zmieszczą się w wyznaczonych klasach.

3.2. Zależność cech populacji

Jak już było powiedziane jednym z głównych zadań statystyki, przy badaniu populacji ze względu na parę cech, jest wypracowanie metod badania pozwalających stwierdzić, czy cechy te są zależne. Wymaga to jednak ścisłego określenia tego pojęcia. Okazuje się, że pojęcie zależności cech może być rozumiane rozmaicie.

3.2.1. Zależność funkcyjna cech populacji

Mówimy, że cechy X i Y są zależne funkcyjne, jeśli istnieje taka funkcja f, że

$$Y = f(X), \text{ lub } X = f(Y),$$

czyli wszystkie dane statystyczne należą do wykresu tej funkcji.

Zależność funkcyjna ma duże znaczenie zarówno teoretyczne jak również praktyczne, pozwala bowiem wyznaczyć szereg rozdzielczy jednej cechy na podstawie szeregu rozdzielczego drugiej cechy, obliczyć charakterystyki liczbowej jednej cechy na podstawie charakterystyk drugiej cechy, a także, co jest szczególnie ważne, wyznaczyć wartość jednej cechy, gdy znana jest wartość drugiej cechy. Jednak w zagadnieniach praktycznych zależność funkcyjna występuje niezmiernie rzadko. Dlatego istnieje potrzeba wprowadzenia ogólniejszych definicji zależności cech populacji i ustalenia zasad, kiedy taka zależność może być przybliżana z małym błędem zależnością funkcyjną.

3..2.2. Zależność stochastyczna (statystyczna) cech populacji

Rozważmy szeregi rozdzielcze warunkowe cechy postaci $X/Y=v_j$ dla wszystkich wariantów v_j. Jeśli w każdym z tych szeregów dowolny wariant w_i cechy X występuje z jednakową częstością, to *cechę X nazywamy cechą stochastycznie niezależną od cechy Y.*

Analogicznie definiuje się niezależność stochastyczną cechy Y od cechy X. Mówimy, że *cechy X i Y są niezależne stochastycznie*, jeśli cecha X nie zależy stochastycznie od cechy Y i Y nie zależy w tym sensie od X.

Niezależność stochastyczna bywa nazywana także *niezależnością statystyczną*.

Niezależność stochastyczna cech X i Y oznacza, że przyjęcie przez jedną cechę dowolnej wartości nie ma wpływu na wielkość częstości, z którą przyjmowane są wartości przez drugą cechę.

Cechy X i Y są zależne stochastycznie, jeśli przynajmniej w dwóch szeregach warunkowych nie wszystkie warianty mają jednakową częstość. Zależność stochastyczna oznacza więc, że fakt przyjęcia przez jedną cechę pewnej wartości może mieć wpływ na częstości przyjmowania wartości przez drugą cechę.

3.2.3. Zależność korelacyjna cech populacji

Cecha X populacji jest niezależna korelacyjnie od cechy Y, jeśli warunkowa wartość oczekiwana cechy $X/Y=v_j$ jest dla dowolnego wariantu v_j cechy Y taka sama, czyli gdy

$$\overline{x}(v_1) = \overline{x}(v_2) = \ldots = \overline{x}(v_s)$$

Analogicznie definiuje się niezależność korelacyjną cechy Y od cechy X. Jeśli cechy X i Y oraz Y i X są niezależne w powyższym sensie, to mówimy, że są one *niezależne korelacyjnie*.

Cechy są zależne korelacyjnie, jeśli przynajmniej w dwóch szeregach warunkowych średnie warunkowe są różne.

Przykład 2.23

W 15 osobowej grupie studentów informatyki przeprowadzono badanie ze względu na parę cech (X, Y), X - ocena końcowa z matematyki, Y - ocena końcowa ze statystyki – patrz przykład 2.21. Przedstawiono w tym przykładzie tablicę korelacyjną:

w_i \ v_j	2	3	4	5
2	3	2		
3	1	2	2	
4			2	1
5			1	1

a) Wyznaczymy szeregi brzegowe b) Wyznaczymy szeregi warunkowe $X/Y=v_j$ c) Wyznaczymy szeregi warunkowe $Y/X=w_i$ d) Obliczymy warunkowe średnie e) Stwierdzimy, w jakim sensie cechy X i Y są zależne.

Rozwiązanie

a) Szeregi brzegowe

Szereg brzegowy cechy X
Struktura ocen z matematyki

Oceny z matematyki w_i	Liczebności ocen $n_{i\bullet}$
2	5
3	5
4	3
5	2
Suma	15

Szereg brzegowy cechy Y
Struktura ocen ze statystyki

Oceny ze statystyki v_j	Liczebności ocen $n_{\bullet j}$
2	4
3	4
4	5
5	2
Suma	15

b) Szeregi warunkowe $X/Y=v_j$

Szereg warunkowy X/Y=2
Struktura ocen z matematyki studentów mających ocenę 2 ze statystyki

Ocena z matematyki w_i	Liczebność n_{i1}
2	3
3	1
Suma	4

Szereg warunkowy X/Y=3
Struktura ocen z matematyki studentów mających ocenę 3 ze statystyki

Ocena z matematyki w_i	Liczebność n_{i2}
2	2
3	2
Suma	4

Szereg warunkowy X/Y=4
Struktura ocen z matematyki studentów mających ocenę 4 ze statystyki

Ocena z matematyki w_i	Liczebność n_{i3}
3	2
4	2
5	1
Suma	5

Szereg warunkowy X/Y=5
Struktura ocen z matematyki studentów mających ocenę 5 ze statystyki

Ocena z matematyki w_i	Liczebność n_{i4}
4	1
5	1
Suma	2

c) Szeregi warunkowe $Y/X=w_i$

Szereg warunkowy Y/X=2
Struktura ocen ze statystyki studentów mających ocenę 2 z matematyki

Ocena ze statystyki v_j	Liczebność n_{1j}
2	3
3	2
Suma	5

Szereg warunkowy Y/X=3
Struktura ocen ze statystyki studentów mających ocenę 3 z matematyki

Ocena ze statystyki v_j	Liczebność n_{2j}
2	1
3	2
4	2
Suma	5

Szereg warunkowy Y/X=4
Struktura ocen ze statystyki studentów mających ocenę 4 z matematyki

Ocena ze statystyki v_j	Liczebność n_{3j}	Częstość Wariantu v_j $n_{3j}/n_{\bullet j}$
4	2	2/3=0,67
5	1	1/3=0,33
Suma	3	1

Szereg warunkowy Y/X=5
Struktura ocen ze statystyki studentów mających ocenę 5 z matematyki

Ocena ze statystyki v_j	Liczebność n_{4j}	Częstość Wariantu v_j $n_{4j}/n_{\bullet j}$
4	1	1/2=0,5
5	1	1/2=0,5
Suma	2	1

Ostatnie dwie tabele rozszerzono o kolumnę częstości warunkowych wariantów. Widzimy, że częstość otrzymania oceny 4 ze statystyki, gdy student z matematyki otrzymał także 4 nie jest równa częstości otrzymania tej oceny ze statystyki, gdy z matematyki otrzymał 5. Oznacza to, że ocena ze statystyki zależy stochastycznie od otrzymanej oceny z matematyki. Zatem cechy X i Y są zależne stochastycznie.

d) Warunkowe średnie

Warunkowe średnie obliczymy na podstawie powyższych szeregów warunkowych.
Warunkowe średnie cech $X/Y=v_j$

$$\overline{x}(v_j) = \text{średnia warunkowa cechy } X/Y = v_j$$

$$\overline{x}(2) = \frac{2\cdot 3+3\cdot 1}{4} = 2,25 \qquad \overline{x}(3) = \frac{2\cdot 2+3\cdot 2}{4} = 2,5 \qquad \overline{x}(4) = \frac{3\cdot 2+4\cdot 2+5\cdot 1}{5} = 3,8$$

$$\overline{x}(5) = \frac{4\cdot 1+5\cdot 1}{2} = 4,5$$

Warunkowe wartości oczekiwane cechy $Y/X=w_i$

$$\overline{y}(w_i) = \text{średnia warunkowa cech } Y/X = w_i$$

$$\overline{y}(2) = \frac{2\cdot 3+3\cdot 2}{5} = 2,4 \qquad \overline{y}(3) = \frac{2\cdot 1+3\cdot 2+4\cdot 2}{5} = 3,2 \qquad \overline{y}(4) = \frac{4\cdot 2+5\cdot 1}{3} = 4,3$$

$$\overline{y}(5) = \frac{4\cdot 1+5\cdot 1}{2} = 4,5$$

e) Zależność cech

Cechy X i Y są zależne stochastycznie, co zostało wykazane w punkcie a).

Cechy X i Y są zależne korelacyjnie, gdyż warunkowe średnie cech postaci $X/Y=v_j$ nie są sobie równe.

Cechy X i Y nie są zależne funkcyjnie (patrz tablica korelacyjna w przykładzie 12.12 i rysunek 2.11). ■

3.3. Charakterystyki liczbowe dwóch cech

3.3.1. Charakterystyki liczbowe dwóch cech, gdy dane przedstawione są w szeregu statystycznym

Badamy populację ze względu na parę cech (X,Y).

Zakładamy, że dane statystyczne przedstawione są w szeregu statystycznym

x_i	x_1	x_2	…	x_n
y_i	y_1	y_2	…	y_n

Przedstawimy najważniejsze charakterystyki liczbowe tych cech.

Nazwa charakterystyki	Określenie charakterystyki	Nr
Średnia cechy X i średnia cechy Y	$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$	(2.3.1)
Moment rzędu 2 cechy X i cechy Y	$\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2, \qquad \overline{y^2} = \frac{1}{n}\sum_{i=1}^{n} y_i^2$	(2.3.2)
Wariancja cechy X i wariancja cechy Y	$s_x^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2, \quad s_y^2 = \frac{1}{n}\sum_{i=1}^{n} (y_i - \overline{y})^2$	(2.3.3)

Odchylenie standardowe cechy X i cechy Y	$s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2}, \quad s_Y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2}$	(2.3.4)
Średnia iloczynu cech X i Y	$\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$	(2.3.5)
Kowariancja cech X i Y	$cov_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})$	(2.3.6)
Współczynnik korelacji cech X i Y	$r = \frac{cov_{xy}}{s_x s_y}$	(2.3.7)

Związki między charakterystykami

$$s_x^2 = \overline{x^2} - (\overline{x})^2 \quad s_y^2 = \overline{y^2} - (\overline{y})^2 \tag{2.3.8}$$

$$cov_{xy} = \overline{xy} - \overline{x}\cdot\overline{y} \tag{2.3.9}$$

$$r = \frac{\overline{xy} - \overline{x}\cdot\overline{y}}{\sqrt{\overline{x^2} - (\overline{x})^2}\sqrt{\overline{y^2} - (\overline{y})^2}} \tag{2.3.10}$$

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}} \tag{2.3.11}$$

Nowymi charakterystykami są te, które dotyczą obu cech łącznie. Najważniejsza z nich to współczynnik korelacji. Omówimy jego własności.

3.3.2. Własności współczynnika korelacji

1. Współczynnik korelacji r cech X i Y jest liczbą z przedziału domkniętego <-1 ; 1>

$$-1 \le r \le 1$$

2. Współczynnik korelacji r jest równy 1 lub -1 wtedy i tylko wtedy, gdy cechy X i Y są zależne liniowo tzn., gdy istnieją liczby a i b takie, że Y = aX + b, przy czym, jeśli r = 1, to a>0 (zależność jest liniowa rosnąca), jeśli r = -1, to a<0 (zależność jest liniowa malejąca).

3. Jeśli cechy X i Y są niezależne stochastycznie, to współczynnik korelacji jest równy 0.

Uwaga: Twierdzenie odwrotne do własności 3 nie jest prawdziwe, bowiem z faktu, iż r = 0 nie wynika, że cechy X i Y są niezależne stochastycznie.

3..3.3. Interpretacja współczynnika korelacji

Współczynnik korelacji r cech X i Y jest miarą siły zależności liniowej tych cech. Im wartość bezwzględna r jest bliższa 1, tym zależność stochastyczna mniej różni się od zależności liniowej, przy czym dla r>0 upodabnia się do zależności liniowej rosnącej, natomiast dla r<0 do zależności malejącej. Dla |r| =1 staje się zależnością liniową.

Nazwy cech w zależności od wielkości współczynnika korelacji r

Wielkość współczynnika korelacji r	**Nazwa cech**
$r \neq 0$	Cechy skorelowane
$r = 0$	Cechy nieskorelowane
$r > 0$	Cechy skorelowane dodatnio
$r < 0$	Cechy skorelowane ujemnie

Niektórzy praktycy przyjmują następującą zasadę określania siły korelacji (liniowej) cech populacji za pomocą współczynnika korelacji r tych cech.

Wielkość współczynnika korelacji r	**Siła korelacji cech**
$0 < \lvert r \rvert < 0,3$	Cechy skorelowane słabo
$0,3 \leq \lvert r \rvert < 0,5$	Cechy skorelowane średnio
$0,5 \leq \lvert r \rvert < 0,7$	Cechy skorelowane mocno
$\lvert r \rvert \geq 0,7$	Cechy skorelowane bardzo mocno

Przykład 2.24

W 15 osobowej grupie studentów przeprowadzono badanie ze względu na parę cech (X, Y), X - ocena końcowa z matematyki, Y - ocena końcowa ze statystyki (patrz przykład 2.21). Otrzymane wyniki przedstawione są w szeregu statystycznym

Zbadamy siłę związku liniowego obu cech obliczając współczynnik korelacji tych cech.

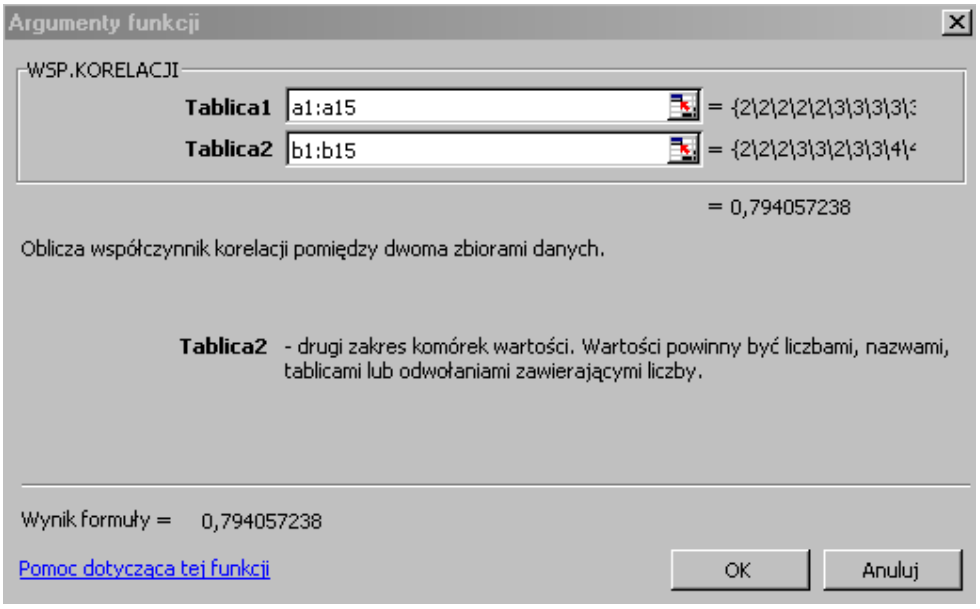
	x_i	y_j	x_i^2	y_i^2	$x_i y_j$
	2	2	4	4	4
	2	2	4	4	4
	2	2	4	4	4
	2	3	4	9	6
	2	3	4	9	6
	3	2	9	4	6
	3	3	9	9	9
	3	3	9	9	9
	3	4	9	16	12
	3	4	9	16	12
	4	4	16	16	16
	4	4	16	16	16
	4	5	16	25	20
	5	4	25	16	20
	5	5	25	25	25
Suma	47	50	163	182	169

Korzystamy ze wzoru (2.11)

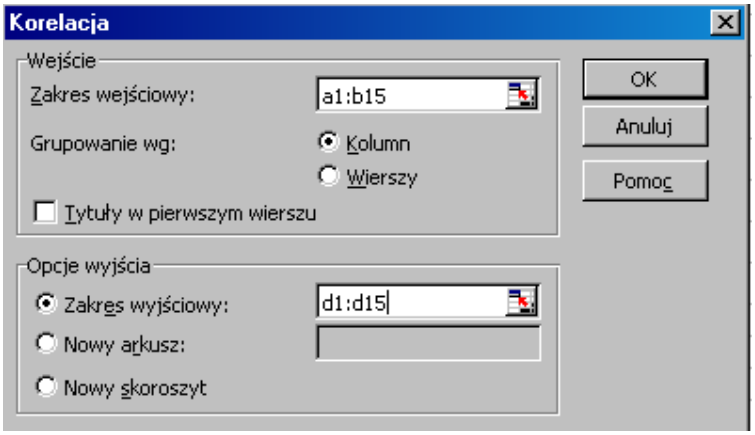
$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}} = \frac{15\cdot 169 - 47\cdot 50}{\sqrt{15\cdot 163 - 47^2}\sqrt{15\cdot 182 - 50^2}} = 0{,}79$$

Współczynniki korelacji można wprost obliczyć z wykorzystaniem arkusza Excel na dwa sposoby (dane i wynik podane są w arkuszu 5 pliku Przykłady) korzystając z:

- Funkcji statystycznej WSP. KORELACJI



- Narzędzia analizy z „Korelacja” z pakietu „Analiza danych” - Analysis ToolPak arkusza Excel



	A	B	C	D	E	F
1	2	2			*Kolumna 1*	*Kolumna 2*
2	2	2		Kolumna 1	1	
3	2	2		Kolumna 2	0,794057	1
4	2	3				
5	2	3				
6	3	2				
7	3	3				
8	3	3				
9	3	4				
10	3	4				
11	4	4				
12	4	4				
13	4	5				
14	5	4				
15	5	5				

Wnioski: Cechy nie są zależne liniowo, bo $|r| \neq 1$ są bardzo silnie skorelowane dodatnio, bo r>0 i |r| >0,7. Zatem można z niewielkim błędem aproksymować powyższą zależność zależnością liniową. ■

3.3.4. Współczynnik korelacji Spearmana

Przy podstawieniu we wzorze na współczynnik korelacji ( ) zamiast wyników x_i oraz y_i ich rangi c_i oraz d_i, przy czym $c_i \in \{1,...,n\}$, $d_i \in \{1,...,n\}$ otrzymuje się tzw. współczynnik korelacji Spearmana[17]

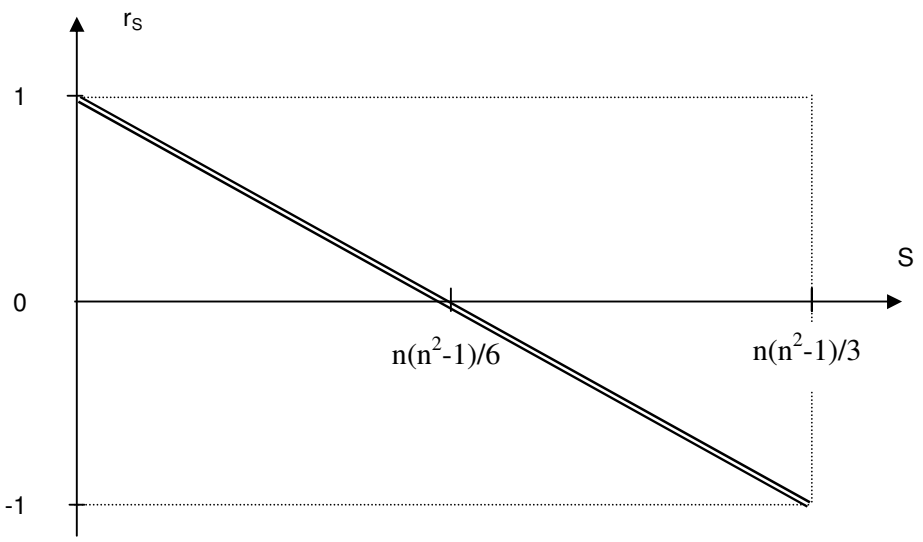
$$r_S = 1 - \frac{6\sum_{i=1}^{n}(c_i - d_i)^2}{n(n^2-1)}$$

Kluczowym elementem wzoru na współczynnik korelacji Spearmana jest suma kwadratów różnic pomiędzy rangami $S=\sum_{i=1}^{n}(c_i - d_i)^2 \geq 0$. Przy jej wykorzystaniu otrzymujemy wzór.

$$r_S = 1 - \frac{6S}{n(n^2-1)} = 1 - cS \quad \text{gdzie:} \quad c = \frac{6S}{n(n^2-1)} > 0$$

- Zależność współczynnika korelacji r_S od sumy S jest liniowa, przy czym wartość współczynnika korelacji maleje ze wzrostem wartości tej sumy.
- Współczynnik korelacji przyjmuje wartość maksymalną, jeżeli S=0, wartość ta jest równa jeden. Sytuacja ta występuje wtedy, jeżeli rangi są parami równe. W tym przypadku uporządkowanie wyników obu prób jest takie samo.
- Kiedy uporządkowania elementów pierwszej próby jest odwrotne do uporządkowania elementów drugiej próby współczynnik korelacji jest równy -1.

[17] Patrz punkt 19.4. części VII Wybrane twierdzenia z dowodami

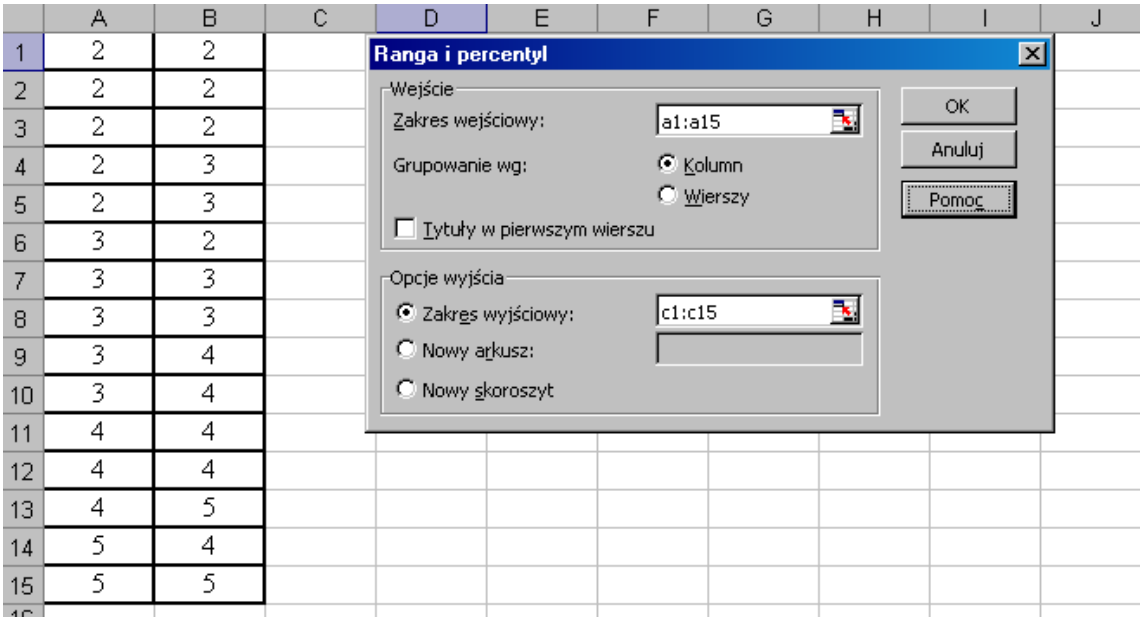


Rys 2.16. Zależność współczynnika korelacji Spearmana od sumy S

Przykład 2.15a

Wyznaczymy współczynnik korelacji Spearmana dla danych z przykładu 2.15

Do rangowania dwukrotnie wykorzystamy narzędzie analizy „Ranga i percentyl" z pakietu „Analiza danych" - Analysis ToolPak .



Wykorzystując obliczone rangi obliczamy wartość współczynnika, co przedstawiono poniżej.

	A	B	C	D	E	F	G	H	I	J	K
1	2	2	*Kolumna1*	*Ranga*	*Kolumna1*	*Ranga*					
2	2	2	5	1	5	1	0	0			
3	2	2	5	1	5	1	0	0			
4	2	3	4	3	4	3	0	0			
5	2	3	4	3	4	3	0	0			
6	3	2	4	3	4	3	0	0			
7	3	3	3	6	4	3	3	0			
8	3	3	3	6	4	3	3	9			
9	3	4	3	6	3	8	-2	9			
10	3	4	3	6	3	8	-2	4		252	6*suma
11	4	4	3	6	3	8	-2	4		15	licznosc
12	4	4	2	11	3	8	3	4		225	kwadrat licznosci
13	4	5	2	11	2	12	-1	9		224	kwadrat licznosci -1
14	5	4	2	11	2	12	-1	1		3360	iloczyn
15	5	5	2	11	2	12	-1	1		0,075	dopełnienie do 1
16			2	11	2	12	-1	1		0,925	wsp. korelacji
17							Roznica rang	42			
18								Kwadrat i suma			
19											

3.4. Regresja

3.4.1. Pojęcie regresji I rodzaju

Dotychczas zajmowaliśmy się przedstawianiem metod badania istnienia i siły zależności cech populacji. Teraz podamy metody aproksymacji zależności cech zależnością funkcyjną i oceną dokładności tej aproksymacji.

Wydaje się, że naturalne jest przedstawić cechę Y jako funkcję cechy X przyporządkowując każdemu wariantowi w_i cechy X średnią warunkową $\bar{y}(w_i)$. Można udowodnić, że postępowanie to jest optymalne (pod pewnym względem).

Regresją I rodzaju cechy Y względem cechy X nazywamy przyporządkowanie każdemu wariantowi w_i cechy X warunkowej średniej cechy Y/X= w_i. Oznaczamy ją symbolem $\hat{Y} = \overline{Y/X = w}$.

Krzywa regresji I rodzaju cechy Y względem cechy X jest to wykres tej regresji, czyli zbiór wszystkich punktów płaszczyzny postaci

$$(w_i, \bar{y}(w_i)) \qquad \text{dla } i = 1, 2, \ldots, q;$$

w_i wariant cechy X, $\bar{y}(w_i)$ - średnia warunkowa cechy Y/X= w_i.

Regresją I rodzaju cechy X względem cechy Y nazywamy przyporządkowanie każdemu wariantowi v_j cechy Y warunkowej wartości oczekiwanej cechy X/Y=v_j. Oznaczamy ją symbolem $\hat{X} = \overline{X/Y = v}$.

Krzywa regresji I rodzaju cechy X względem cechy Y jest to wykres tej regresji, czyli zbiór wszystkich punktów płaszczyzny postaci

$$(\bar{x}(v_j), v_j) \qquad \text{dla } j = 1, 2, \ldots s;$$

v_j wariant cechy Y, $\bar{x}(v_j)$ - średnia warunkowa cechy X/Y= v_j.

3.4.2. Pojęcie regresji II rodzaju

Regresję I rodzaju cechy Y względem cechy X wybiera się, zgodnie z zasadą najmniejszych kwadratów, ze zbioru wszystkich funkcji. Jednak posługiwanie się tą regresją jest niewygodne, gdyż nie można na ogół przedstawić jej wzorem zależnym od parametrów, co utrudnia przewidywanie wartości cechy Y dla ustalonej wartości cechy X (dla której regresja nie jest określona). Dlatego bardzo często ograniczamy wybór funkcji do pewnej klasy K.

Niech K będzie klasą funkcji określonych wspólnym wzorem zależnym od parametrów.

Regresją II rodzaju cechy Y względem X w klasie K nazywamy funkcję $\hat{Y} = h(X)$, gdzie funkcja h jest wybrana zgodnie z zasadą najmniejszych kwadratów spośród funkcji należących do klasy K.

3.4.3. Liniowa regresja II rodzaju

Liniowa regresja II rodzaju cechy Y względem cechy X jest to regresja II rodzaju cechy Y względem X w klasie K wszystkich funkcji liniowych postaci $y = ax + b$.

Miarą aproksymacji jest wzór

$$f(a,b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

Zgodnie z zasadą najmniejszych kwadratów należy wyznaczyć punkt (a_y, b_y), w którym funkcja f ma wartość najmniejszą. Można wykazać[18], że funkcja f ma wartość najmniejszą w punkcie (a_y, b_y), gdzie

$$a_y = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{s_y}{s_x} r \qquad (12.40)$$

$$b_y = \overline{y} - a_y \overline{x}$$

Zatem

$\hat{Y} = a_y X + b_y$ - regresja II rodzaju liniowa cechy Y względem cechy X ,

$\hat{y} = a_y x + b_y$ - równanie prostej regresji II rodzaju cechy Y względem cechy X,

$a_y = \frac{s_y}{s_x} r$, $b_y = \overline{y} - a\overline{x}$ - współczynniki regresji liniowej II rodzaju cechy Y względem cechy X.

Interpretacja współczynników regresji

a_y – średnia zmiana cechy Y, gdy cecha X wzrosła o jednostkę

b_y – rzędna punktu przecięcia prostej regresji $\hat{y} = a_y x + b_y$ z osią Ox.

Analogicznie definiujemy regresję II rodzaju liniową cechy X względem cechy Y.

$\hat{X} = a_x Y + b_x$ - regresja II rodzaju liniowa cechy X względem cechy Y ,

$\hat{x} = a_x y + b_x$ - równanie prostej regresji II rodzaju cechy X względem cechy Y,

a_x, b_x - współczynniki regresji liniowej II rodzaju cechy Y względem cechy X,

[18] Patrz punkt 19.5. części VII Wybrane twierdzenia z dowodami

gdzie

$$a_x = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2} = \frac{s_x}{s_y} r$$

$$b_x = \overline{x} - a_x \overline{y}$$

Funkcję regresji charakteryzują następujące własności[19]:

- Suma różnic pomiędzy wartościami zmiennej zależnej i wartościami funkcji regresji jest równa zeru $K = \sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$
- Suma odchyleń dodatnich od funkcji regresji jest równa sumie odchyleń ujemnych $\sum_{y_i > \hat{y}_i} (y_i - \hat{y}_i) = \sum_{y_i < \hat{y}_i} (\hat{y}_i - y_i)$

Przykład 2.25

W 15 osobowej grupie studentów informatyki przeprowadzono badanie ze względu na parę cech (X,Y), X - ocena końcowa z matematyki, Y - ocena końcowa ze statystyki (patrz przykład 2.21).

W przykładzie 2.15 obliczyliśmy, że

$$\sum_{i=1}^{15} x_i = 47, \quad \sum_{i=1}^{15} y_i, = 50, \quad \sum_{i=1}^{15} x_i^2 = 163, \quad \sum_{i=1}^{15} y_i^2 = 182, \quad \sum_{i=1}^{15} x_i y_i = 169$$

więc

$$a_y = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{15 \cdot 169 - 47 \cdot 50}{15 \cdot 163 - 47^2} = 0,78$$

$$b_y = \overline{y} - a_y \overline{x} = \frac{50}{15} - 0,78 \cdot \frac{47}{15} = 0,88$$

$a_y = 0,78,\ b_y = 0,88$ - współczynniki regresji II rodzaju liniowej cechy Y względem cechy X,

$\hat{Y} = 0,78X + 0,88$ - regresja II rodzaju liniowa cechy Y względem cechy X,

$\hat{y} = 0,78x + 0,88$ - równanie prostej regresji II rodzaju cechy Y względem cechy X.

Można obliczyć, że

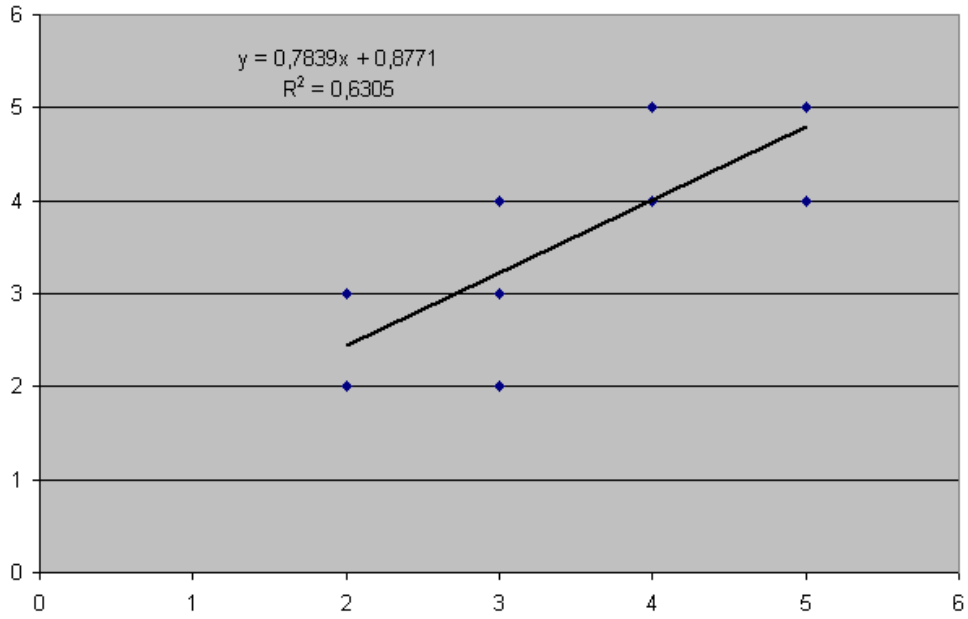
$\hat{x} = 0,80y + 0,45$ - równanie prostej regresji II rodzaju cechy X względem cechy Y.

Współczynniki regresji można obliczyć z wykorzystaniem arkusza Excel na cztery sposoby (dane i wynik podane są w arkuszu 5 pliku Przykłady) korzystając z:

- Obliczając na podstawie danych poszczególne elementy wzoru na współczynnik korelacji, takie jakie obliczono w przykładzie ….i na ich podstawie obliczyć współczynnik korelacji.
- Kreatora wykresu:
  1. Wybiera się typ wykresu XY punktowy
  2. Wybiera się myszą jeden z punktów wykresu, naciska prawy przycisk myszy, wybiera z menu opcję Dodaj linie trendu i wybiera Trend liniowy

[19] Patrz punkt 19.6. części VII Wybrane twierdzenia z dowodami

3. Wskazuje się myszą linię trendu, wybiera opcję Formatuj linię trendu i zaznacza Wyświetl równanie na wykresie oraz Wyświetl wartości R-kwadrat na wykresie. Należy wyjaśnić, że R jest współczynnikiem korelacji pomiędzy analizowanymi wartościami Y, a wartościami Y obliczonymi z równania regresji na podstawie wartości X – podano je w arkuszu 5 pliku Przykłady. Pod wartościami Y podano obliczona wartość R i jego kwadrat.



- Funkcji statystycznej REGLINP. W tym celu należy:
  1. Wyselekcjonować obszar na wyniki: 1 wiersz i 2 kolumny, ponieważ chcemy otrzymać tylko wartości współczynników regresji
  2. Wpisać do tego obszaru nazwę funkcji z zakresem danych – w naszym przypadku =REGLINP(B1:B15;A1:A15;1;1)
  3. Równocześnie nacisnąć przyciski CTRL+SHIFT+ENTER

0,783898305	0,877118644
0,166428483	0,548625294
0,630526898	0,660143005
22,18524059	13
9,668079096	5,665254237

W otrzymywanych wynikach zacieniowane te które otrzymano na wykresie, pozostałe zostaną omówione w ramach analizy statystycznej.

- Narzędzia analizy z „Regresja” z pakietu „Analiza danych” - Analysis ToolPak

Statystyki regresji	
Wielokrotność R	0,794057238
R kwadrat	0,630526898
Dopasowany R kwadrat	0,60210589
Błąd standardowy	0,660143005
Obserwacje	15

ANALIZA WARIANCJI

	df	*SS*	*MS*	*F*	*Istotność F*
Regresja	1	9,668079096	9,668079096	22,18524059	0,000408
Resztkowy	13	5,665254237	0,435788787		
Razem	14	15,33333333			

	Współczynniki	*Błąd standardowy*	*t Stat*	*Wartość-p*	*Dolne 95%*	*Górne 95%*	*Dolne 95,0%*	*Górne 95,0%*
Przecięcie	0,877118644	0,548625294	1,598757209	0,133885924	-0,30811	2,062352	-0,30811	2,062352
Zmienna X 1	0,783898305	0,166428483	4,71012108	0,000407701	0,424351	1,143445	0,424351	1,143445

W otrzymywanych wynikach zacieniowane te które otrzymano na wykresie, pozostałe zostaną omówione w ramach analizy statystycznej. ■