ANALIZA KORELACJI

Większość zjawisk w otaczającym nas świecie występuje nie samotnie a w różnorodnych związkach. Odnosi się to również do zjawisk biologiczno-medycznych. O powiązaniach między nimi mówią różnorodne prawa botaniki, zoologii, immunologii, fizjologii, biochemii i innych nauk medycznych formułując przeróżne zależności między występującymi tam zmiennymi. Statystyka dostarcza nam narzędzi, które pozwalają zweryfikować rozpoznane powiązania, jak również pomaga wykryć nierozpoznane dotychczas współzależności. Często słyszymy zdanie „Rak płuc powiązany jest z paleniem papierosów". Mówi ono, że im więcej papierosów się pali, tym bardziej prawdopodobne, że zachoruje się na raka. Mówimy, że im więcej jednego tym więcej drugiego. Zamiast używać nieprecyzyjnych słów takich jak „więcej" lub mało" statystycy wolą oceniać rzeczy używając liczb. Tak powstała matematyczna teoria korelacji i regresji jako narzędzie służące do dokładnego określania stopnia, w jakim zmienne są ze sobą powiązane. Podstawowym problemem statystyki w takich badaniach jest stwierdzenie czy między zmiennymi zachodzi jakiś związek, jakaś zależność i czy związek jest bardziej czy mniej ścisły. Analiza regresji i korelacji to jedna z najważniejszych i najszerzej stosowanych metod statystycznych. Obecny kurs poświęcony jest korelacji.

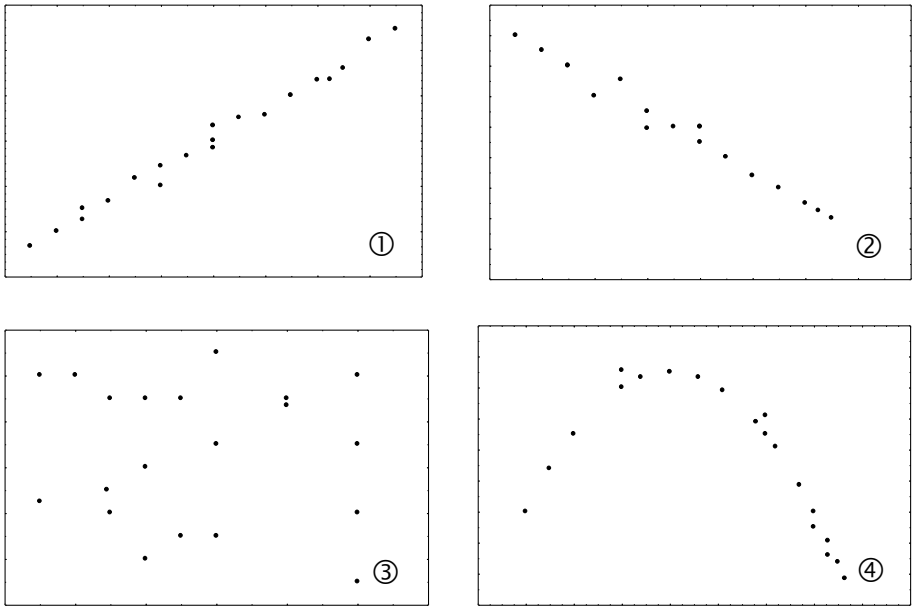
Dwie zmienne mogą być pomiędzy sobą powiązane zależnością funkcyjną lub zależnością statystyczną (korelacyjną). *Związek funkcyjny* odznacza się tym, że każdej wartości jednej zmiennej niezależnej (będziemy ją oznaczać X) odpowiada tylko jedna, jednoznacznie określona wartość zmiennej zależnej (oznaczamy ją przez Y). Wiadomo na przykład, że pole kwadratu jest funkcją jego boku ($P=a^2$). *Związek statystyczny* polega na tym, że określonym wartościom jednej zmiennej odpowiadają ściśle określone średnie wartości drugiej zmiennej. Można zatem obliczyć, jak zmieni się - średnio biorąc - wartość zmiennej zależnej Y w zależności od wartości zmiennej niezależnej X. Zwróćmy też uwagę, że liczbowe stwierdzenie występowania współzależności nie zawsze oznacza występowanie związku przyczynowo-skutkowego między badanymi zmiennymi. Współwystępowanie dwóch zjawisk może również wynikać z bezpośredniego oddziaływania na nie jeszcze innego trzeciego zjawiska.

W analizie korelacji badacz jednakowo traktuje obie zmienne - nie wyróżniamy zmiennej zależnej i niezależnej. Korelacja między X i Y jest taka sama jak między Y i X. Mówi nam ona, na ile obie zmienne zmieniają się równocześnie w sposób liniowy. Precyzyjna definicja zaś brzmi:

Korelacja między zmiennymi X i Y jest miarą siły liniowego związku między tymi zmiennymi.

Analizę związku korelacyjnego między badanymi cechami rozpoczynamy zawsze od sporządzenia wykresu. Wykresy, które reprezentują obrazowo związek pomiędzy zmiennymi nazywane są wykresami rozrzutu (*Scatterplot*). W prostokątnym układzie współrzędnych na osi odciętych zaznaczamy jedną zmienną a na osi rzędnych wartości drugiej zmiennej. Punkty, odpowiadające poszczególnym wartościom cech, tworzą korelacyjny wykres rozrzutu. Rzadko zdarza się, że zaznaczone punkty leżą dokładnie na linii prostej (pełna korelacja), częściej spotykana konfiguracja składa się z wielu zaznaczonych punktów leżących mniej więcej wzdłuż konkretnej krzywej (najczęściej linii prostej). Taka sytuacja przedstawiona jest jako przypadek ① i ② na rysunku 1. Gdy korelacja staje się coraz mniej doskonała, wówczas punkty zaczynają się rozpraszać i przesuwać, aż do bezkształtnej chmury punktów (brak korelacji). Taka sytuacja występuje w przypadku ③ na rysunku 1.
Korelacja dodatnia występuje wtedy, gdy wzrostowi wartości jednej cechy odpowiada wzrost średnich wartości drugiej cechy (przypadek ① na rysunku).

Korelacja ujemna występuje wtedy, gdy wzrostowi wartości jednej cechy odpowiada spadek średnich wartości drugiej cechy (przypadek ② na rysunku).
Czasami układ punktów wskazuje na korelację krzywoliniową – rysunek ④.



① - korelacja liniowa dodatnia ② - korelacja liniowa ujemna ③ - brak korelacji ④- korelacja krzywoliniowa

Rys. 1 Korelacyjne wykresy rozrzutu

Siłę współzależności dwóch zmiennych można wyrazić liczbowo za pomocą wielu mierników. Najbardziej popularny to **współczynnik korelacji liniowej Pearsona** - oznaczony symbolem r_{xy} - i przyjmujący wartości z przedziału [-1, 1]. Należy zwrócić uwagę, że współczynnik korelacji Pearsona wyliczamy, gdy obie zmienne są mierzalne i mają rozkład zbliżony do normalnego oraz zależność jest prostoliniowa. Stąd też nazwa współczynnik korelacji **liniowej** Pearsona. Przy interpretacji współczynnika korelacji liniowej Pearsona należy więc pamiętać, że wartość współczynnika bliska zeru nie zawsze oznacza brak zależności, a jedynie brak zależności liniowej.

Znak współczynnika korelacji informuje nas o kierunku korelacji, natomiast jego bezwzględna wartość o sile związku. Mamy oczywiście równość $r_{xy} = r_{yx}$. Jeśli $r_{xy} = 0$, oznacza to zupełny brak związku korelacyjnego między badanymi zmiennymi X i Y (przypadek ③ na rysunku). Im wartość bezwzględna współczynnika korelacji jest bliższy jedności, tym zależność korelacyjna między zmiennymi jest silniejsza. Gdy $r_{xy} = |1|$, to zależność korelacyjna przechodzi w zależność funkcyjną (funkcja liniowa).

W analizie statystycznej zwykle przyjmuje się następującą skalę:

$r_{xy} = 0$	zmienne nie są skorelowane
$0 < r_{xy} < 0{,}1$	korelacja nikła
$0{,}1 \le r_{xy} < 0{,}3$	korelacja słaba
$0{,}3 \le r_{xy} < 0{,}5$	korelacja przeciętna
$0{,}5 \le r_{xy} < 0{,}7$	korelacja wysoka
$0{,}7 \le r_{xy} < 0{,}9$	korelacja bardzo wysoka
$0{,}9 \le r_{xy} < 1$	korelacja prawie pełna.

Przedstawiona skala jest oczywiście umowna. W literaturze możemy również znaleźć inne określenia.

Tak jak wartość innych parametrów populacji, współczynnik korelacji (w populacji) nie jest znany i musimy go oszacować na podstawie znajomości losowej próby par wyników obserwacji zmiennych X i Y. Tak wyliczony z próby współczynnik r_{xy} jest estymatorem współczynnika korelacji ρ w populacji generalnej, a jego wartość liczbowa stanowi ocenę punktową siły powiązania w całej populacji. Stąd konieczność testowania istotności współczynnika korelacji wyliczonego w oparciu o próbę losową. Weryfikujemy, więc następujący układ hipotez:

$$H_0: \rho = 0$$
$$H_1: \rho \neq 0$$

Weryfikacja tej hipotezy zerowej pomoże nam w ocenie, czy istniejąca zależność między X i Y w próbie jest tylko przypadkowa, czy też jest prawidłowością w populacji.
W dalszych rozważaniach ograniczymy się do przykładowej analizy za pomocą pakietu ***STATISTICA.***

Rozważmy badanie, w którym analizowano powiazanie między obwodem serca a masą ciała dla 15 krów

Masa	641	620	633	651	640	666	650	688	680	670	630	665
Obwód	205	212	213	216	217	218	219	221	226	207	222	212

Chcemy zbadać siłę i kierunek zależności między wiekiem a wzrostem w analizowanej grupie. Po wprowadzeniu danych do programu *STATISTICA* i wykonaniu analizy otrzymujemy następujący arkusz wyników.

Korelacje (Przykład 1)
Oznaczone wsp. korelacji są istotne z p < ,05000
(Braki danych usuwano przypadkami)

Zmn. X & Zmn. Y	Średnia	Odch.st.	r(X,Y)	r2	t	p	Ważnych	Stała zal: Y	Nachyle zal: Y	Stała zal: X	Nachyle zal: X
Obwód serca	215,6667	6,33208									
Waga	652,6000	21,41361	0,789128	0,622723	4,632217	0,000469	15	77,06200	2,668646	63,38385	0,233348
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]

Rys. 2 Okno z wynikami - opcja Dokładna tabela wyników

Poszczególne pola arkusza wyników zawierają:

[1] - Średnie arytmetyczne wybranych zmiennych

[2] - Odchylenia standardowe

[3] - Współczynnik korelacji Pearsona

[4] - Współczynnik determinacji (R^2 kwadrat współczynnika korelacji). Jest to opisowa miara dokładności dopasowania regresji do danych empirycznych. Przyjmuje wartości z przedziału <0, 1> lub w ujęciu procentowym <0, 100%> i informuje, zgodnie z zapisem, jaka część zaobserwowanej w próbie całkowitej zmienności Y została wyjaśniona (zdeterminowana) regresją względem X. Im większe R^2 tym powiązanie jest lepsze i możemy mieć większe zaufanie do ewentualnej linii regresji.

[5] - Wartość statystyki t badającej istotność współczynnika korelacji

[6] – Wartość prawdopodobieństwa testowego p.

[7] - Liczebność grupy

[8] - Wyraz wolny regresji liniowej Y względem X

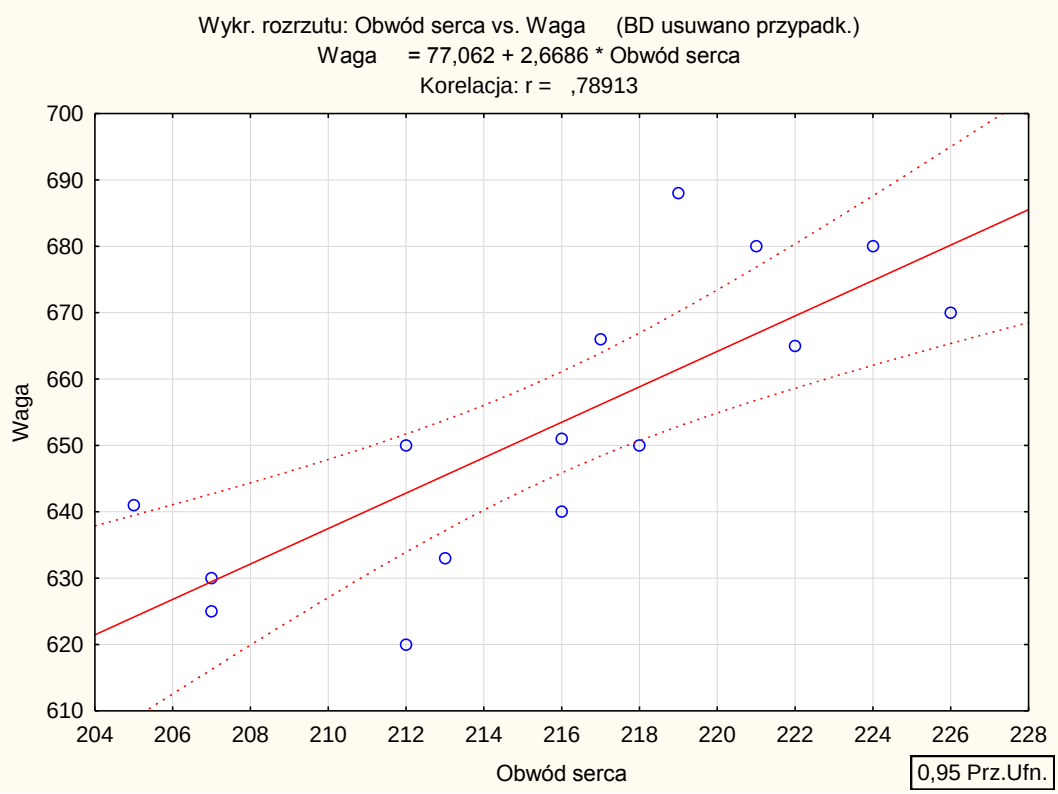
[9] - Współczynnik regresji liniowej zmiennej Y względem zmiennej X

[10] - Wyraz wolny regresji liniowej X względem Y

[11] - Współczynnik regresji liniowej zmiennej X względem zmiennej Y

Punkty od [8] do [11] umożliwiają wyliczenie funkcji regresji zmiennej Y względem X i funkcji regresji zmiennej X względem Y opisujących analityczną postać zależności pomiędzy zmiennymi. Pojęcie regresji zostanie omówione dokładniej w kolejnym odcinku.

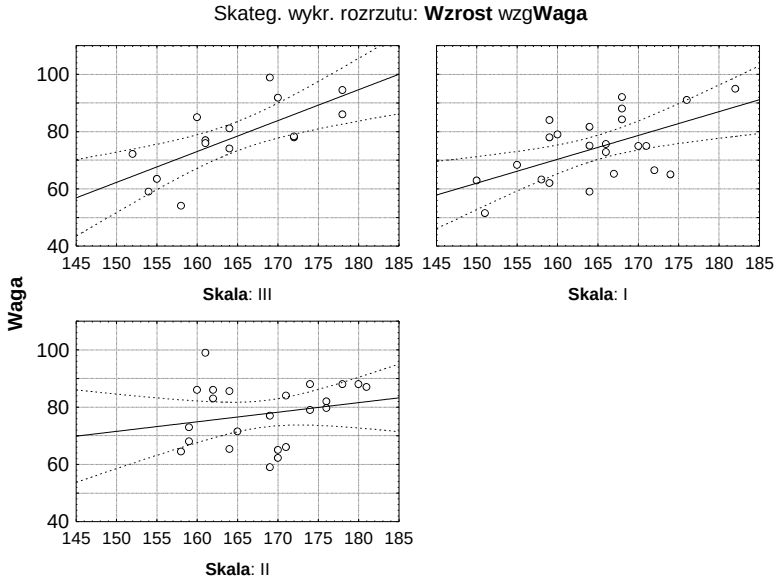
Jak widzimy pomiędzy wagą i wzrostem zachodzi wysoce istotna (p = 0,00000) korelacja. Wartość współczynnika korelacji wynosi 0,789. Ponadto jak mówi o tym współczynnik determinacji zmienność jednej cechy (np. wagi) w 62% jest wyjaśniona zmiennością drugiej (czyli wzrostu). Tą sytuację pokazuje poniższy wykres rozrzutu.



Rys. 3 Wykres rozrzutu danych z analizowanego przykładu

Rysunek zawiera wykres prostej regresji wagi osoby badanej względem wieku. Na wykresie zaznaczono też 95% przedział ufności linii regresji (obszar zaznaczony przerywanymi liniami).

Program *STATISTICA* umożliwia tworzenie skategoryzowanych wykresów rozrzutu, tj. wykresy rozrzutu w podgrupach wyznaczonych przez wybraną zmienną grupującą. Poniższy rysunek zawiera wykresy rozrzutu wagi względem wzrostu w podgrupach wyznaczonych przez zmienną skala natężenia choroby. Możemy sprawdzić jak wygląda ta zależność dla poszczególnych stopni natężenia choroby.



Rys. 10 Skategoryzowany wykres rozrzutu.

Przypominamy, że współczynnik korelacji Pearsona może być wyliczany tylko dla zmiennych mierzalnych, co najmniej na skali przedziałowej. Zmienne te muszą mieć rozkład normalny i związek między nimi powinien być w przybliżeniu liniowy. Jeżeli nie są spełnione założenia musimy skorzystać z nieparametrycznego odpowiednika – współczynnika korelacji rang Spearmana.

Korelacje nieparametryczne

Przedstawiony w poprzednim ćwiczeniu współczynnik korelacji Pearsona może być wyliczany tylko dla zmiennych mierzalnych, co najmniej na skali przedziałowej. Zmienne te muszą mieć rozkład normalny i związek między nimi powinien być w przybliżeniu liniowy. Jeżeli nie są spełnione założenia musimy skorzystać z nieparametrycznych odpowiedników. Należą do nich współczynniki:

- **współczynnik korelacji rang Spearmana**. Współczynnik ten został solidnie opisany i rozpropagowany w 1904 roku przez angielskiego psychologa Charlesa Spearmana. Zauważył on, że w wielu badaniach nie da się zastosować klasycznego współczynnika korelacji lub daje on nieistotne wyniki ze względu na nadmiar obserwacji odstających. Współczynnik korelacji rang Spearmana stosuje się do analizy współzależności obiektów pod względem cechy dwuwymiarowej (X, Y). Zakładając, że badamy n obiektów opisanych za pomocą dwóch cech, należy te obiekty uporządkować ze względu na wartości każdej cechy oddzielnie. Obiektom w każdym z uporządkowań przypisujemy liczbę określającą ich miejsce położenia (dla $x_i - r_{1i}$, a dla $y_i - r_{2i}$) (1,2,3,...,n). Liczby te nazywa się rangami, a procedurę nadawania rang – rangowaniem. Spearman zdefiniował swój współczynnik, jako zwykły współczynnik korelacji Pearsona, liczony dla rang zmiennych (stąd nazwa *współczynnik korelacji rang*). W przypadku, gdy występują **jednakowe** wartości zmiennych, przyporządkowujemy im średnią arytmetyczną obliczoną z ich kolejnych numerów. Mówi się wówczas o występowaniu **rang wiązanych***)*.

Współczynnik korelacji rang Spearmana wyliczany jest według poniższego wzoru:

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

gdzie:
$d_i = r_{1i} - r_{2i}$,
r_{1i} – ranga i-tego obiektu w pierwszym uporządkowaniu,
r_{2i} – ranga i-tego obiektu w drugim uporządkowaniu,
n – liczba badanych obiektów.

Gdy obserwacje każdej zmiennej w próbie powtarzają się, to współczynnik korelacji dla rang jest dodatkowo korygowany ze względu na rangi wiązane.

- **współczynnik tau Kendalla (τ-Kendalla)** - współczynnik ten opiera się na różnicy między prawdopodobieństwem tego, że dwie zmienne układają się w tym samym porządku (dla obserwowanych danych) a prawdopodobieństwem, że ich uporządkowanie się różni. Zaproponowany przez Kendalla (1955 r.) wymaga, aby wartości zmiennych można było uporządkować (zmienne muszą być mierzone co najmniej na skali porządkowej). Współczynnik ten przyjmuje wartości z przedziału <-1, 1>. Wartość 1 oznacza pełną zgodność, wartość 0 brak zgodności uporządkowań, natomiast wartość -1 całkowitą ich przeciwstawność. Współczynnik Kendalla wskazuje, więc nie tylko siłę, lecz również kierunek zależności. Jest doskonałym narzędziem do opisu podobieństwa uporządkowań zbioru danych.
- **statystyka gamma** - współczynnik ten ma podobną konstrukcje i interpretację jak współczynnik R Spearmana lub τ Kendalla. Wymaga też podobnych założeń. Stosuje się go w przypadkach, gdy dane zawierają wiele obserwacji powiązanych (reprezentujących ten sam wariant cechy)

Współczynnik korelacji rang Spearmana przyjmuje wartości z przedziału <-1,1>. Im bliższy jest on liczbie 1 lub -1, tym silniejsza jest analizowana zależność. Zatem jego interpretacja jest podobna do klasycznego współczynnika korelacji Pearsona, z jednym zastrzeżeniem: w odróżnieniu od współczynnika Pearsona, który mierzy liniową zależność między zmiennymi, korelacja rangowa pokazuje dowolną monotoniczną zależność (także nieliniową).

Tak jak wartość innych parametrów populacji, współczynnik rang Spearmana (w populacji) nie jest znany i musimy go oszacować na podstawie znajomości losowej próby par wyników obserwacji zmiennych X i Y. Tak wyliczony z próby współczynnik r_{xy} jest estymatorem współczynnika korelacji ρ w populacji generalnej, a jego wartość liczbowa stanowi ocenę punktową siły powiązania w całej populacji. Stąd konieczność testowania istotności współczynnika korelacji wyliczonego w oparciu o próbę losową. Weryfikujemy, więc następujący układ hipotez:

$$H_0: \rho = 0$$
$$H_1: \rho \neq 0$$

Weryfikacja tej hipotezy zerowej pomoże nam w ocenie, czy istniejąca zależność między X i Y w próbie jest tylko przypadkowa, czy też jest prawidłowością w populacji.

W dalszych rozważaniach ograniczymy się do przykładowej analizy za pomocą bardzo popularnego i przystępnego pakietu *STATISTICA***.** Jak w tym pakiecie przeprowadzić analizę korelacji pokazuje odpowiednia prezentacja. Obecnie omówimy otrzymane wyniki i przedstawimy najciekawsze interpretacje geometryczne otrzymanych wyników.

Przykładowo chcemy ustalić, współzależność między opiniami wydanymi przez dwóch lekarzy o zdrowiu 10 pacjentów. Opinie te zostały ujęte w punktach:

Pacjenci		A	B	C	D	E	F	G	H	I	J
Punkty uzyskane od	**I lekarza**	42	27	36	33	24	47	39	52	43	37
	II lekarza	39	24	35	29	26	47	44	51	39	32

Chcemy zbadać siłę i kierunek zależności między opiniami dwóch lekarzy. Po wprowadzeniu danych (każda zmienna w osobnej kolumnie) do programu *STATISTICA* i wykonaniu analizy otrzymujemy następujący arkusz wyników.

Korelacja porządku rang Spearmana (Arkusz3
BD usuwane parami
Oznaczone wsp. korelacji są istotne z p <,05

Para zmiennych	**N ważnych**	R Spearman	t(N-2)	poziom p
OCENA L1 & OCENA L2	10	0,936175	7,532386	0,000067

[1] [2] [3] [4] [5]

Rys. 11 Arkusz wyników obliczania współczynnika Spearmana

Poszczególne wartości oznaczają:

[1] - nazwy zmiennych

[2] - liczebność grup

[3] - wartość współczynnika R Spearmana

[4] - wartość statystyki t sprawdzaj istotność współczynnika R Spearmana

[5] – wartość prawdopodobieństwa p dla powyższej statystyki t

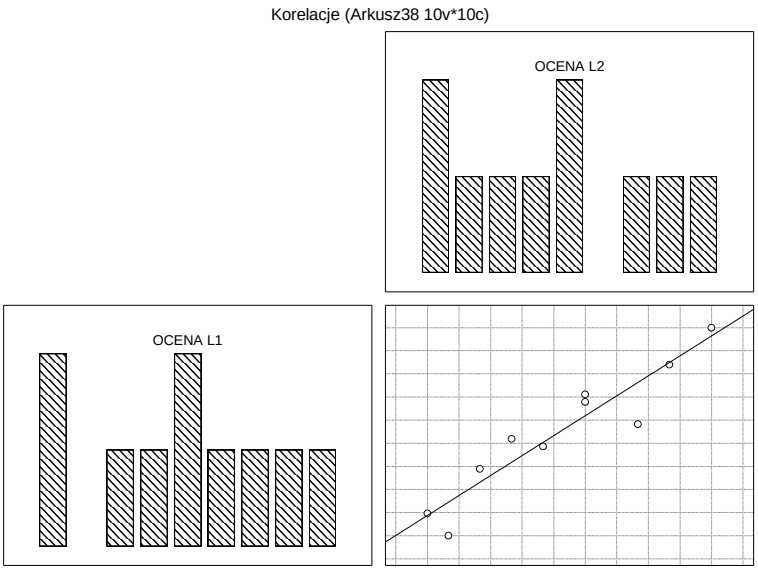
Uzyskany współczynnik R Spearmana R_S = 0,936 wskazuje na silną, istotną (p = 0,000067) współzależność opinii dwóch lekarzy o stanie zdrowia pacjenta. Podobne wyniki otrzymujemy wyliczając współczynnik τ Kendalla. Widoczne są one w poniższym oknie.

Korelacja tau Kendalla (Arkusz38)
BD usuwane parami
Oznaczone wsp. korelacji są istotne z p <,05000

Para zmiennych	**N ważnych**	Tau Kendalla	Z	poziom p	p-dokł jednostr
OCENA L1 & OCENA L2	10	0,809040	3,256323	0,001129	p < .001

Rys. 12 Arkusz wyników obliczania współczynnika tau Kendalla

Nasze obliczenia możemy ilustrujemy na koniec wykresem macierzowym.



Rys.13 Wykres macierzowy dla danych z przykładu