

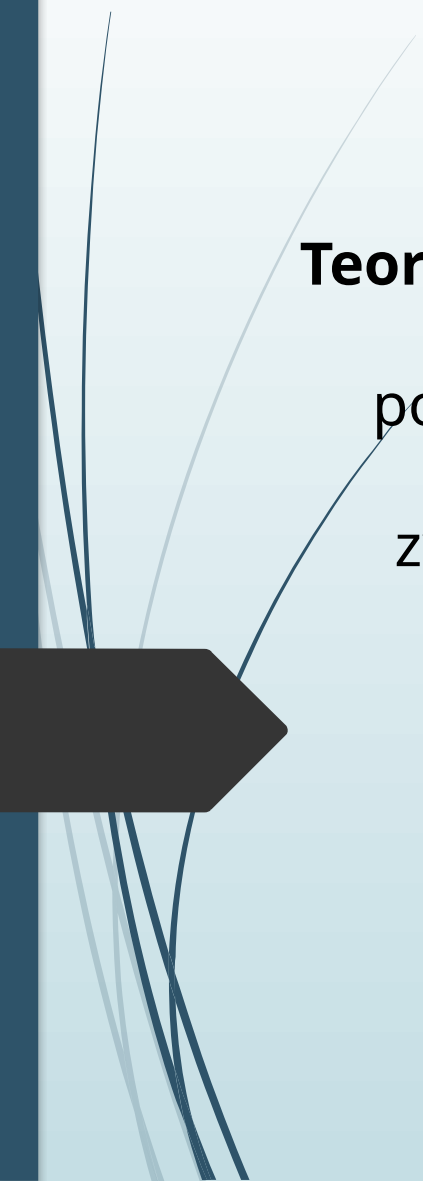


Analiza współzależności – korelacja

Teoria i praktyka

Teoria współzależności

Teoria współzależności – dział statystyki, w którym podejmuje się problemy związane z badaniem związku pomiędzy dwoma lub większą liczbą zmiennych.



Teoria współzależności

Jednostki tworzące zbiorowość statystyczną charakteryzowane są zazwyczaj za pomocą wielu cech zmiennych, które nierzadko pozostają ze sobą w pewnym związku.

Występująca zależność może mieć charakter funkcyjny lub stochastyczny.

Teoria współzależności

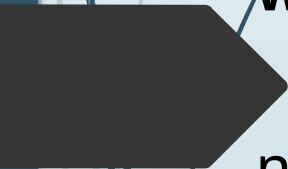
Charakter funkcyjny zależności powoduje, że zmiany jednej zmiennej związane są ściśle ze zmianami drugiej zmiennej.

Charakter funkcyjny zależności oznacza, że możliwe jest określenie funkcji $Y=f(X)$, która **każdej wartości zmiennej X przyporządkowuje ściśle jedną wartość zmiennej Y** ; np. pole koła, im większa średnica tym zawsze większe pole.

Teoria współzależności

W przypadku **zależności stochastycznej** zmiany jednej zmiennej wywołują średnie zmiany drugiej zmiennej.

Oznaczyć to można symbolicznie przez $Y'=f(X)$, gdzie Y' oznacza wartość średnią. Zatem: **danej wartości zmiennej X odpowiadać może więcej niż jedna wartość zmiennej Y ;**



np. przypuszczać można, że im więcej czasu studenci poświęcają na naukę tym lepszą otrzymają ocenę, jednak przy tej samej ilości czasu poświęconego na naukę możliwe jest, że różni studenci będą otrzymywać różne

Teoria współzależności

Określenie siły, kierunku oraz kształtu tego związku możliwe jest dzięki analizie współzależności zjawisk.

Zakres analizy współzależności zjawisk:

- analiza korelacji
- analiza regresji

Analiza korelacji

Analiza korelacji pozwala określić siłę zależności między zmiennymi, a w przypadku zależności liniowej dwóch zmiennych – także **kierunek zależności**.

Siłę związku między zmiennymi określa się za pomocą szeregu miar, których wybór zależy od tego, czy zmienne mają charakter mierzalny, czy też niemierzalny.

Analiza korelacji

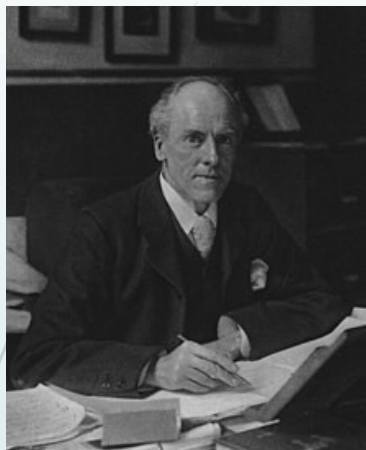
Zależności przyczynowo-skutkowe, to takie, w których zmienne rzeczywiście na siebie oddziałują. Mogą być to zależności **jednostronne**, jak w przypadku wpływu czasu przygotowania do egzaminu na ocenę, lub **dwustronne**.

Analiza korelacji

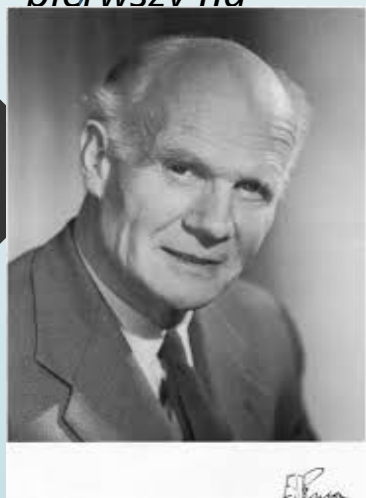
Ważnym problemem związanym z oceną współzależności jest rozróżnienie zależności pozornych od zależności przyczynowo-skutkowych.

Zależności pozorne nie dają się wytłumaczyć w sposób logiczny, np. wysokość drzewa i wysokość PKB

Sławni statystycy



pierwszyna



Londynie.

Karl Pearson – (ur. 1857, zm. 1936) – angielski matematyk, prekursor statystyki matematycznej.

W roku 1898 otrzymał Medal Darwina za jego pracę nad podejściem do problemów biologicznych.

W 1911 roku na University College London utworzył świecie uniwersytecki Wydział Statystyki.

Egon Sharpe Pearson (ur. 1895, zm. 1980) – brytyjski

syn Karla Pearsona. Zajmował się problematyką testowania statystycznych. W latach 1955-1957 pełnił funkcję Królewskiego Towarzystwa Statystycznego w

Analiza korelacji

Dobór metody oceny współzależności zależy od charakteru ocenianych zmiennych. Dla zmiennych o charakterze ilościowym podstawową metodą oceny zależności jest **współczynnik korelacji liniowej Pearsona** oraz **współczynnik korelacji rang Spearmana**, natomiast dla zmiennych o charakterze jakościowym miary zależności oparte na **statystyce** χ^2 .

Teoria współzależności

Korelacja jest inaczej zapisaniem w postaci liczbowej związków zachodzących pomiędzy dwiema zmiennymi.

Możemy za jej pomocą określić m.in. jak liczba pracowników w danym przedsiębiorstwie powiązana jest z generowanym przez nią przychodem.

Niezbędne dane:

- 1) ilość pracowników
- 2) przychody firmy.

Im więcej tych danych, tym dokładniejsze będą wyniki pomiaru.

Analiza korelacji

Analiza powiązań zmiennych ilościowych pozwala na ocenę **kierunku i siły zależności**.

Jeżeli dwie zmienne podążają średnio w tym samym kierunku, tj. jednocześnie na ogół rosną albo maleją to taką korelację określa się jako dodatnią.

Jeśli podążają w różnych kierunkach, tj. wyższym wartościom jednej zmiennej odpowiadają niższe wartości drugiej zmiennej to taką korelację określa się jako ujemną.

Analiza korelacji

Analiza powiązań zmiennych ilościowych pozwala na ocenę kierunku i siły zależności.

Siłę korelacji określa się poprzez podobieństwo zależności empirycznej do zależności funkcyjnej.

Im zależności empirycznej bliżej do zależności typu funkcyjnego, tym korelacja jest silniejsza.

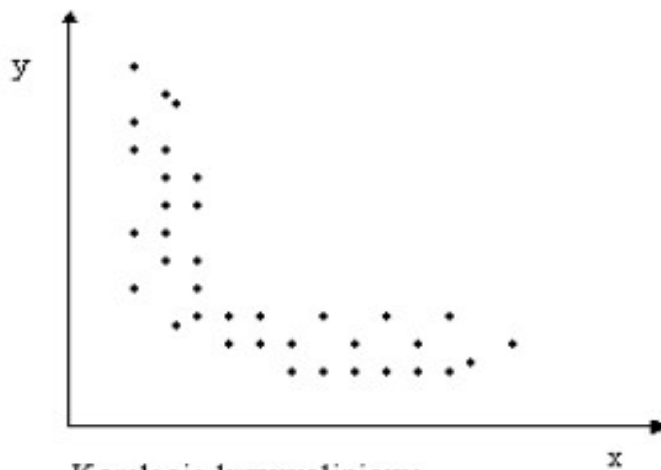
Diagram rozrzutu



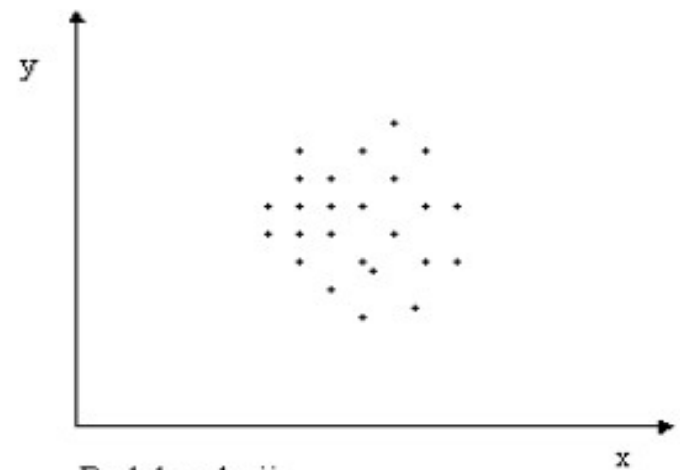
Korelacja liniowa dodatnia
 $0 < r_{xy} \leq 1$



Korelacja liniowa ujemna
 $-1 \leq r_{xy} < 0$



Korelacja krzywoliniowa



Brak korelacji
 $r_{xy} = 0$

Analiza korelacji

Współczynnik korelacji liniowej Pearsona - jest miarą pozwalającą na określenie kierunku i siły zależności prostoliniowej pomiędzy dwoma zmiennymi mającymi charakter ilościowy (mierzalny).

O **zależności prostoliniowej** mówi się, gdy jednostkowym przyrostom jednej zmiennej towarzyszy średnio stały przyrost (dodatni lub ujemny) drugiej zmiennej.

Analiza korelacji

Miarą pozwalającą na określenie kierunku zależności jest **kowariancja**:

$$s_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

\bar{x}, \bar{y} wartości zmiennych X i Y ;

\bar{xy} średnie arytmetyczne zmiennych X i Y ;

średnia arytmetyczna iloczynu zmiennych X i Y ,

$$\bar{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i$$

Kowariancja przyjmuje wartości z przedziału

, gdzie $S(x)$ i $S(y)$ są odchyleniami standardowymi zmiennych X i Y .

Kowariancja przyjmuje wartości z przedziału nieunormowanego, dlatego też nie można na jej podstawie porównywać siły zależności różnych par zmiennych

Analiza korelacji

Natomiast **współczynnik korelacji liniowej Pearsona** określony wzorem:

$$r_{xy} = \frac{\text{COV}(X, Y)}{S(X) \cdot S(Y)}$$

jest miarą unormowaną. Przyjmuje wartości z przedziału $[-1; 1]$.

Znak współczynnika określa kierunek zależności:

„+” – **zależność dodatnia**

„-” – **zależność ujemna**

Analiza korelacji

Współczynnik korelacji linowej Pearsona jest współczynnikiem symetrycznym, tj. jego wartość nie zależy od tego, która ze zmiennych jest zmienną zależną, a która zmienną niezależną:

$$r_{xy} = r_{yx}$$

Analiza korelacji

Z kolei wartość bezwzględna współczynnika określa siłę zależności. Im wartość bezwzględna współczynnika bliższa 1 tym zależność liniowa silniejsza, im bliższa 0 tym zależność liniowa słabsza.

W interpretacji można posłużyć się poniższym wzorem:

- ☐ $|r_{xy}| \approx 1$ – pełna zależność liniowa;
- ☐ $|r_{xy}| \approx 0.8$ – bardzo silna zależność liniowa;
- ☐ $|r_{xy}| \approx 0.6$ – dość silna zależność liniowa;
- ☐ $|r_{xy}| \approx 0.4$ – przeciętna zależność liniowa;
- ☐ $|r_{xy}| \approx 0.2$ – słaba zależność liniowa;
- ☐ $|r_{xy}| \approx 0$ – bardzo słaba zależność liniowa;
- ☐ $|r_{xy}| \approx 0$ – brak zależności liniowej.

Przykład:

Badając zależność między wielkością zbiorów truskawek, a ilością nawozu uzyskano na 10 (n=10) plantacjach wyniki:
Określić siłę i kierunek zależności.

Ilość nawozu (kg) x	Zbiory (t) y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
0	18	-90	-2,91		8100	8,4681
20	20,3	-70	-0,61		4900	0,3721
40	20,5	-50	-0,41		2500	0,1681
60	20,4	-30	-0,51		900	0,2601
80	21,2	-10	0,29		100	0,0841
100	21,7	10	0,79		100	0,6241
120	21,3	30	0,39		900	0,1521
140	21,6	50	0,69		2500	0,4761
160	22,2	70	1,29		4900	1,6641
180	21,9	90	0,99		8100	0,9801
900	209,1			571	33000	13,249

$$\bar{x} = \frac{900}{10} = 90$$

$$\bar{y} = \frac{209,1}{10} = 20,91$$

$$S_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}} = \sqrt{\frac{33000}{10}} = 57,45$$

$$S_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n}} = \sqrt{\frac{13,249}{10}} = 1,15$$

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n S_x S_y} = \frac{571}{10 \cdot 57,45 \cdot 1,15} = 0,86$$

$$r_{xy}^2 = 0,74$$

Analiza korelacji - przykład

Tabela przedstawia liczbę pracowników i generowane przychody przedsiębiorstwa. Jak na ich podstawie zbadać korelację zachodzącą pomiędzy nimi?

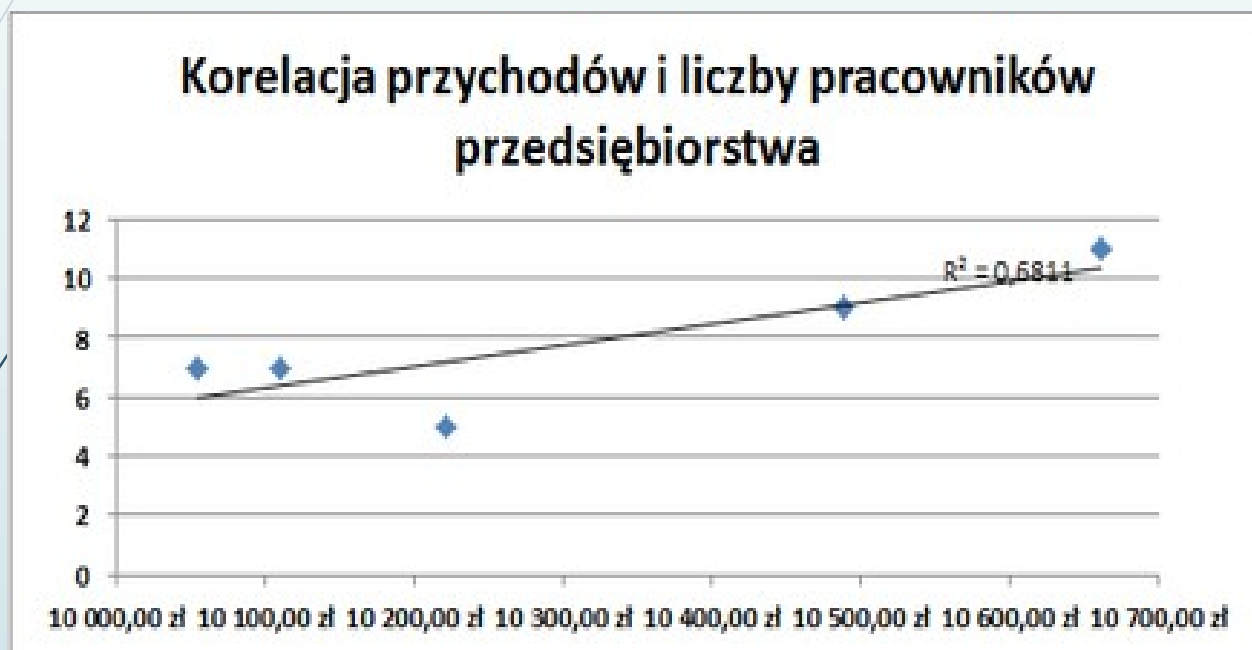
Rok	2000	2001	2002	2003	2004
Liczba pracowników	5	7	7	9	11
Generowane przychody	10 222,00 zł	10 055,00 zł	10 110,00 zł	10 489,00 zł	10 661,00 zł

X – liczba pracowników

Y – generowane przychody ze sprzedaży.

Rozwiązanie: $r = 0,83$

Analiza korelacji – przykład



Współczynnik korelacji cząstkowej

Szczególnym zastosowaniem współczynnika korelacji cząstkowej może być **rozdzielenie korelacji pozornych od rzeczywistych**.

Jeżeli pomiędzy dwoma zmiennymi występuje wyraźna korelacja, ale **nie można ustalić dla nich związku przyczynowo-skutkowego**, to oznacza, że może istnieć trzecia zmienna, która oddziałuje jednocześnie na dwie pozostałe, powodując, że powstaje pomiędzy nimi korelacja pozorna. Podstawową trudnością w takiej sytuacji jest określenie trzeciej zmiennej wpływającej jednocześnie na dwie pozostałe.

Współczynnik korelacji cząstkowej

Współczynnik korelacji cząstkowej pozwala na ocenę siły i kierunku zależności liniowej pomiędzy dwoma zmiennymi po wyłączeniu wpływu trzeciej zmiennej na ten związek.

Przyjmuje on wartości z przedziału $[-1; 1]$, a jego interpretacja jest analogiczna do interpretacji współczynnika korelacji liniowej Pearsona.

Współczynnik korelacji cząstkowej

Współczynnik korelacji cząstkowej pomiędzy zmiennymi X i Y po wyłączeniu wpływu zmiennej Z na ten związek wyraża się wzorem:

$$r_{xy \cdot z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{1 - r_{xz}^2} \sqrt{1 - r_{yz}^2}}$$

gdzie r_{xy} , r_{xz} , r_{yz} są współczynnikami korelacji liniowej Pearsona wyznaczonymi dla każdej pary zmiennych X , Y i Z .

Współczynnik korelacji wielorakiej

Współczynnik korelacji wielorakiej stosowany jest w sytuacji, gdy przedmiotem zainteresowania jest związek pomiędzy wartościami jednej zmiennej a zespołem złożonym z wielu innych zmiennych. W przypadku, gdy ocenie podlega związek pomiędzy jedną zmienną a zespołem złożonym z dwóch innych zmiennych współczynnik korelacji wielorakiej można wyznaczyć ze wzoru:

$$R_{xy} = \sqrt{\frac{r_{xy}^2 + r_{xz}^2 - r_{yz}^2}{1 - r_{yz}^2}}$$

gdzie r_{xy} , r_{xz} , r_{yz} są współczynnikami korelacji liniowej Pearsona wyznaczonymi dla każdej pary zmiennych X, Y i Z.

Współczynnik korelacji wielorakiej przyjmuje wartości z przedziału [0;1] i pozwala na ocenę siły zależności. Im wartość współczynnika korelacji wielorakiej bliższa 1 tym oceniana zależność jest silniejsza, im wartość współczynnika korelacji wielorakiej bliższa 0 tym oceniana zależność jest słabsza. Wartość 1 oznacza zależność funkcyjną, a wartość 0 całkowity brak zależności. Współczynnik korelacji wielorakiej nie pozwala na ocenę kierunku zależności.

Współczynnik korelacji rang Spearmana

Współczynnik R Spearmana wykorzystywany jest do opisu siły korelacji dwóch cech, w przypadku gdy:

- ❖ **cechy mają charakter jakościowy**, pozwalający na uporządkowanie ze względu na siłę tej cechy,
np. stopień zaangażowania w trening sportowy
- ❖ **cechy mają charakter ilościowy, ale ich liczebność jest niewielka**

np. odległość w skokach narciarskich

Współczynnik korelacji rang Spearmana

Ranga (d) jest to liczba odpowiadająca miejscu w uporządkowaniu każdej z cech.

Jeśli w badanej zbiorowości jest więcej jednostek z identycznym natężeniem badanej cechy, to jednostkom tym przypisuje się identyczne rangi, licząc średnią arytmetyczną z rang przynależnych tym samym jednostkom.

Współczynnik korelacji rang Spearmana

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

d_i^2 - różnica pomiędzy rangami odpowiadających sobie wartości cech x_i i y_i

Współczynnik korelacji rang przyjmuje wartości z przedziału **[-1; 1]**.

Interpretacja jest podobna do współczynnika korelacji liniowej Pearsona.

Przykład

Spożycie mięsa	Dochód na osobę	Rangi		d_i	d_i^2
		SM	DnO		
20	9	1	1	0	0
24	11	2	2	0	0
25	12	3	3	0	0
26	18	4	6,5	-2,5	6,25
27	15	5	4	1	1
28	22	6	8	-2	4
29	17	7,5	5	2,5	6,25
29	18	7,5	6,5	1	1
31	24	9	9	0	0
32	28	10	10	0	0
		Suma		18,5	

$$r_s = 1 - \frac{\sum_{i=1}^n d_i^2}{n(n^2 - 1)} = 1 - \frac{6 \cdot 18,5}{10 \cdot 10^2 - 1} = 0,89$$

Własności:

- cechy jakościowe i ilościowe
- zbiorowość nieliczna
- współczynnik determinacji R^2

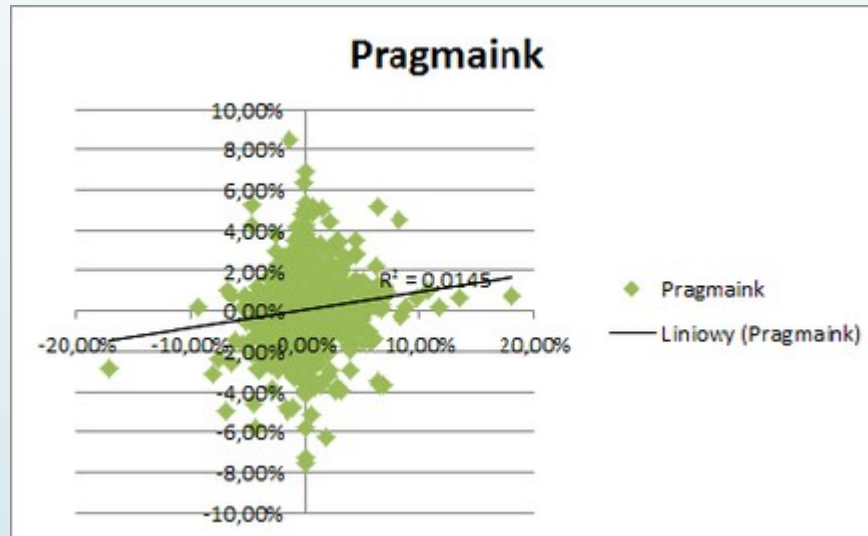
Wykorzystanie współczynnika korelacji w praktyce – na GPW

Wskaźnik powinno się wykorzystywać w podejmowaniu decyzji inwestycyjnych.

Analiza portfelowa – nauka zajmująca się statystyczno-matematycznym ujęciem inwestycji giełdowych.

Przy doborze danego papieru wartościowego do portfela zaleca się obliczenie jego współczynnika korelacji z obecnie posiadanymi już papierami wartościowymi.

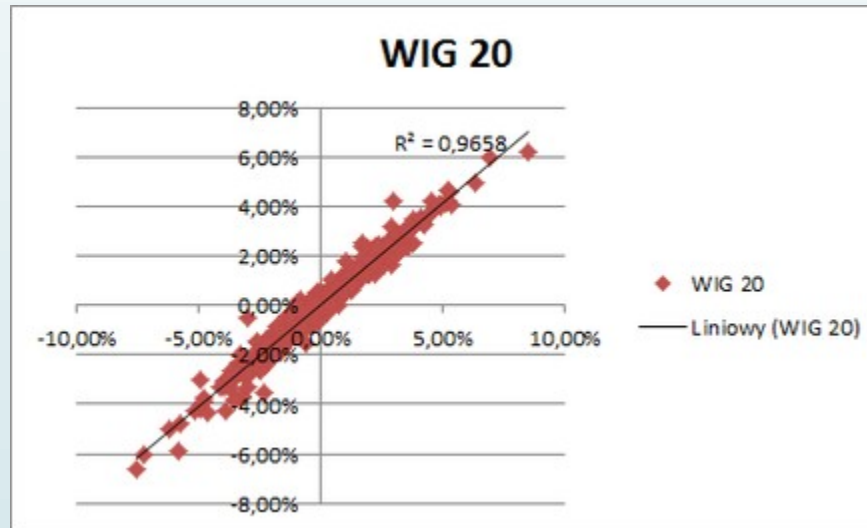
Analiza korelacji



Korelacja: Spółka Pragma Inkaso S.A. – Indeks WIG.

Słaba korelacja wyraża się bardzo dużym rozproszeniem punktów.

Analiza korelacji



Korelacja: Indeks WIG20 – Indeks WIG.


Korelacja silna to punkty ułożone w ciąg - przypominające linię.

Wykorzystanie współczynnika korelacji w praktyce – na GPW

Aby dokonać pomiaru korelacji walorów swojego portfela inwestycyjnego zaleca się stworzenie tzw. „Macierzy korelacyjnej”.

Macierz korelacyjna					
	<i>KGHM</i>	<i>Patentus</i>	<i>Tauron</i>	<i>Zastal</i>	<i>Graal</i>
<i>KGHM</i>	1	-0,21101	0,33174	0,370468	0,190101
<i>Patentus</i>	-0,21101	1	0,720598	0,760161	0,384229
<i>Tauron</i>	0,33174	0,720598	1	-0,2135	-0,2135
<i>Zastal</i>	0,370468	0,760161	0,66993	1	0,419973
<i>Graal</i>	0,190101	0,384229	-0,2135	0,419973	1

Proponowany wybór spółki dominującej z danego sektora, która ma najlepsze perspektywy rozwoju i stabilną kondycję finansową.



Wykorzystanie współczynnika korelacji w praktyce – na GPW

Warto przy podejmowaniu decyzji inwestycyjnych kierować się współczynnikiem korelacji, głównie po to, aby zminimalizować ryzyko inwestycyjne przez odpowiednią dywersyfikację statystyczną swojego portfela.

Przykład

Przeprowadzono badanie zależności między ceną biletów oferowanych przez sześć przedsiębiorstw transportowych za przejazd 100 km a liczbą pasażerów korzystających z usług tych przedsiębiorstw w ciągu tygodnia.

W oparciu o powyższe informacje należy:

- 1) sporządzić diagram korelacyjny i zinterpretować rozkład punktów,
- 2) określić kierunek i siłę związku korelacyjnego; zinterpretować

otrzy

Cena biletu za przejazd 100 km (w zł)	15	20	25	30	35	40
Liczba pasażerów	440	430	430	380	360	350