

KORELACJE

KORELACJA

KORELACJA

- współzależność, wzajemny związek pomiędzy dwoma zmiennymi
- oznacza współwystępowanie, nie powinna być interpretowana jako zależność przyczynowo-skutkowa
- znając wartość jednej zmiennej da się z pewnym prawdopodobieństwem przewidzieć wartość drugiej
- w statystyce korelacja jest miarą powiązania pomiędzy zmiennymi. Jest określana na podstawie współczynnika korelacji

ZNACZENIE

- **atrybucja** - wyjaśnianie obserwowanych zjawisk, ustalanie jakie są przyczyny – atrybucja oparta na wiedzy naukowej
- **predykcja** - przewidywanie przyszłych zdarzeń (wartości zmiennych) – analiza regresji

WSPÓŁCZYNNIK KORELACJI

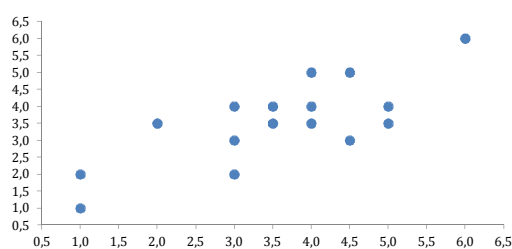
WSPÓŁCZYNNIK KORELACJI

- jest miarą powiązania pomiędzy zmiennymi
- liczbową miarą związku
- im silniejsza korelacja (wyższa wartość współczynnika korelacji) tym lepiej potrafimy przewidzieć wartość jednej zmiennej na podstawie znajomości wartości drugiej lub wyjaśnić związek między nimi
- informuje o sile i kierunku zależności

3

Czy istnieje korelacja pomiędzy ocenami z języka polskiego i matematyki?

Obserwacja	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Język polski:	1	1	2	3	3	3	3,5	3,5	4	4	4	4,5	4,5	5	5	6
Matematyka:	2	1	3,5	4	3	2	3,5	4	3,5	5	4	5	3	3,5	4	6



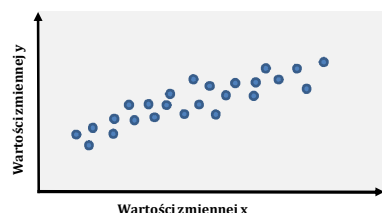
Siła: czy punkty na wykresie tworzą wyraźną smugę?

Kierunek : czy wartości jednej zmiennej rosną czy maleją wraz ze wzrostem wartości drugiej zmiennej?

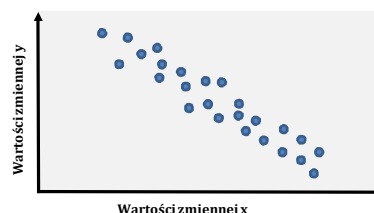
Kształt: czy punkty na wykresie układają się wokół pewnej linii?

4

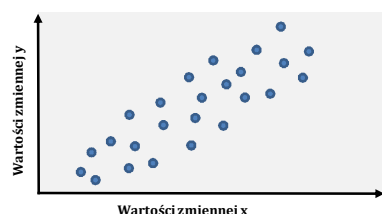
INTERPRETACJA GRAFICZNA



Korelacja liniowa dodatnia (zależność silna)



Korelacja liniowa ujemna (zależność silna)



Korelacja liniowa dodatnia (zależność umiarkowana)



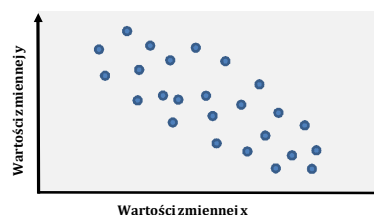
Korelacja liniowa ujemna (zależność umiarkowana)

5

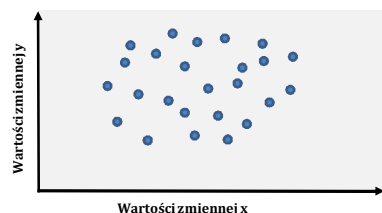
INTERPRETACJA GRAFICZNA



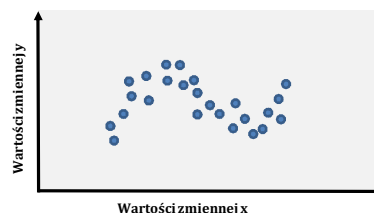
Korelacja liniowa dodatnia (zależność słaba)



Korelacja liniowa ujemna (zależność słaba)



Brak korelacji



Korelacja krzywoliniowa

6

WSPÓŁCZYNNIK KORELACJI

- Wartość współczynników korelacji mieści się między $-1,1$.

KIERUNEK KORELACJI

- Korelacja dodatnia (ujemna) oznacza, że wraz ze wzrostem wartości jednej zmiennej wartości drugiej zmiennej rosną (maleją)

SIŁA ZWIĄZKU (wartości bezwzględne współczynnika korelacji)

- 0,00-0,10 bardzo niska
- 0,11-0,30 niska
- 0,31-0,50 umiarkowana
- 0,51-0,70 wysoka
- 0,71-1,00 bardzo wysoka

W badaniach społecznych wartość współczynnika korelacji powyżej 0,7 powinna budzić wątpliwości!

7

WSPÓŁCZYNNIK KORELACJI

WYBÓR WSPÓŁCZYNNIKA KORELACJI

Zastosowanie konkretnego współczynnika korelacji jest uzależnione od

- liczby korelowanych zmiennych
- skali pomiarowej zmiennych (nominalna, porządkowa, ilościowa)
- liczby wartości przyjmowanych przez korelowane zmienne
- natury związku (korelacyjny czy funkcjonalny)
- testowania szczegółowych warunków zastosowania danego współczynnika

PARAMETRYCZNY WSPÓŁCZYNNIK KORELACJI

- r Pearsona
- oparty na wartościach zmiennej, do jego wyznaczenia wykorzystujemy parametry – średnie i odchylenia standardowe
- zakładamy normalność rozkładu zmiennych

NIEPARAMETRYCZNE WSPÓŁCZYNNIK KORELACJI

- rho-Spearmana, tau-Kendalla, d-Somersa
- korelacje oparte na rangach, do ich wyznaczenia wykorzystujemy uporządkowanie zmiennych
- brak założeń dotyczących rozkładu

8

WSPÓŁCZYNNIK KORELACJI

	PRZEDZIAŁOWA STOSUNKOWA	PORZĄDKOWA	NOMINALNA
PRZEDZIAŁOWA STOSUNKOWA	r Pearsona	↑	eta
PORZĄDKOWA	←	ρ Spearmana (rho) τ-b Kendalla (tau-b) τ-c Kendalla (tau-c) d Sommersa	↑
NOMINALNA	eta	←	C – kontyngencji Yula (phi) V Cramera

WYBÓR WSPÓŁCZYNNIKA KORELACJI

- jeśli obie korelowane zmienne mierzone są na tej samej skali wybieramy współczynnik „na przekątnej”
- jeśli jedna ze zmiennych jest nominalna, druga ilościowa wybieramy współczynnik eta
- jeśli zmienne mierzone są na różnych skalach wybieramy współczynnik dedykowany dla zmiennych mierzonych na słabszej skali

9

R - PEARSONA

r PEARSONA

- współczynnik korelacji liniowej / współczynnik wg momentu iloczynowego /parametryczny współczynnik korelacji
- opisuje związek LINIOWY między zmiennymi
- suma iloczynów wartości standaryzowanych par zmiennych podzielona przez liczbę par

$$r = \frac{\sum z_x z_y}{n}$$

- z_x - standaryzowana wartość zmiennej x
- z_y - standaryzowana wartość zmiennej y
- n - liczba par zmiennych (liczba obserwacji)

$$z_x = \frac{x - \bar{x}}{S_x} \quad z_y = \frac{y - \bar{y}}{S_y}$$

$$r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{n S_x S_y}$$

- w przypadku wyliczania r dla próby należy licznik podzielić przez (n-1) zamiast przez n.

10

R - PEARSONA

WŁASNOŚCI

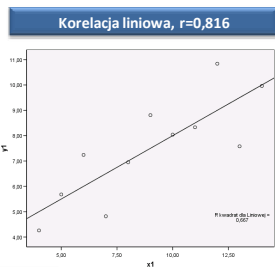
$$r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{n S_x S_y}$$

- obie zmienne muszą być mierzone na skalach ilościowych
- zmienne muszą być zróżnicowane (odchylenie różne od zera)
- wrażliwe na zmienne odstające (zaniżona wartość r)
- rozkład zmiennych powinien być normalny ($A_s = 0$, $K = 0$), ale akceptowalna jest umiarkowana skośność: $|A| < 1,0$, $|K| < 3,0$
- odchylenia powinny być równe (dla $N > 30$ założenie to traci na znaczeniu, można zaniedbać)
- Jeżeli r Pearsona = 0, ściśle oznacza to, że model liniowy (liniowa zależność między zmiennymi) w ogóle nie pasuje do danych, bo:
 - nie ma żadnej zależności między zmiennymi, albo
 - zależność ma inny kształt niż liniowy
- Współczynnik korelacji jest „wrażliwy” na zakres zmienności. Im mniejszy zakres tym mniejsza wartość bezwzględna współczynnika - łatwiej stwierdzić korelację jeśli jest większy zakres zmienności
- Przy liczeniu korelacji im „dłuższa” skala tym lepiej

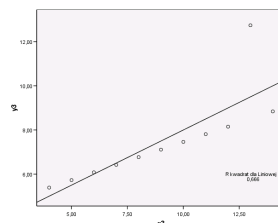
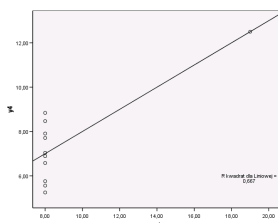
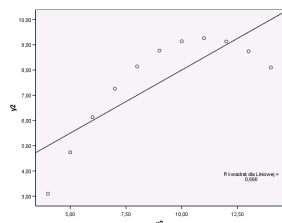
11

KWARTET ANSCOMBE'A

- cztery zestawy 11 par x i y , takie, że w każdym zestawie:
 - dla zmiennej x : $M=9$, $S^2=11$
 - dla zmiennej y : $M=7,5$, $S^2=11$,
 - **współczynnik korelacji: $r = 0,816$**
- dopiero ilustracja graficzna pozwala dostrzec różnice!
- współczynnik korelacji zaniża opis siły związku w przypadku zależności nieliniowych



Brak korelacji liniowej, $r=0,816$



12

WSPÓŁCZYNNIK DETERMINACJI

WSPÓŁCZYNNIK DETERMINACJI

- r^2 – kwadrat współczynnika korelacji liniowej
- jest wskaźnikiem stopnia odchylenia pomiarów przewidywanych od pomiarów rzeczywistych
- $r^2 * 100\%$ – odsetek wspólnej wariancji
- informuje o tym jaką część zmienności jednej zmiennej możemy wyjaśnić zmiennością drugiej zmiennej

$r = 0,1$	$r^2 = 0,01 = 1\%$
$r = 0,2$	$r^2 = 0,04 = 4\%$
$r = 0,3$	$r^2 = 0,09 = 9\%$
$r = 0,4$	$r^2 = 0,16 = 16\%$
$r = 0,5$	$r^2 = 0,25 = 25\%$
$r = 0,6$	$r^2 = 0,36 = 36\%$
$r = 0,7$	$r^2 = 0,49 = 49\%$
$r = 0,8$	$r^2 = 0,64 = 64\%$
$r = 0,9$	$r^2 = 0,81 = 81\%$
$r = 1,0$	$r^2 = 1,00 = 100\%$

- wzrost siły korelacji taką samą wartością (np. o 0,1) nie powoduje takiego samego wzrostu wyjaśnionej wariancji
- związek między zmiennymi nie jest dwa razy silniejszy jeśli współczynniki korelacji wynoszą $r = 0,6$ i $r = 0,3$
- korelacja $r = -0,3$ jest silniejsza niż korelacja $r = 0,2$

13

INTERPRETACJA

PRZYKŁAD 1

Współczynniki korelacji między wiekiem a czasem korzystania z Internetu wynosi $r = -0,4$

- korelacja między wiekiem a czasem korzystania z Internetu jest umiarkowana i ujemna co oznacza, że im starszy badany tym krócej korzysta z sieci
- 16% ($0,4^2$) zmienności (różnicowania) czasu korzystania z Internetu można wyjaśnić zmiennością wieku.
- pozostałe 84% zmienności czasu przeznaczanego na korzystanie z Internetu zależy od innych niebadanych czynników (ceny dostępu, posiadania komputera, ilości wolnego czasu, liczby posiadanych przyjaciół, itp.)

PRZYKŁAD 2

Korelacja między IQ w dzieciństwie a IQ w wieku dorosłym $r = 0,75$

- Istnieje bardzo silny dodatni związek między IQ w dzieciństwie i IQ w wieku dorosłym.
- Im wyższy iloraz IQ w dzieciństwie tym wyższy iloraz IQ w wieku dorosłym.
- Wspólna wariancja (zmienność) wynosi 56,25% ($0,75^2$)
- 56,25% zmienności IQ w wieku dorosłym można wyjaśnić zmiennością IQ w dzieciństwie

14

WSPÓŁCZYNNIKI KORELACJI RANGOWEJ

ZAŁOŻENIA

- obie zmienne są mierzone przynajmniej na skali porządkowej (rangowej)
- im większa liczba wartości, które przyjmuje każda ze zmiennych tym lepiej
- wartości współczynników wahają się w granicach od - 1 do + 1
- współczynniki korelacji rangowej **nie** są wyliczane na podstawie wartości zmiennej, ale na podstawie ich pozycji (rangi) wartości w szeregu statystycznym

15

RANGOWANIE

RANGOWANIE

- polega na przydzieleniu poszczególnym obserwacjom odpowiedniej rangi - uszeregowanie obserwacji w określonym porządku niezależnie od różnicy wielkości między nimi
- rangi wyraża się liczbami całkowitymi 1, 2, 3, 4... N

PRZYKŁAD

- Wartości zmiennej: 35, 13, 17, 15, 22, 16, 10, 8, 19, 21
- Wartości zmiennej uporządkowane:

WARTOŚĆ	8	10	13	15	16	17	19	21	22	35
Pozycja	1	2	3	4	5	6	7	8	9	10
RANGA	1	2	3	4	5	6	7	8	9	10

16

RANGI WIĄZANE

- Wartości zmiennej: 22, 10, 13, 17, 13, 22, 22, 10, 8, 13, 13, 19, 17, 30
- Wartości zmiennej uporządkowane:

WARTOŚĆ	8	10	10	13	13	13	13	17	19	22	22	22	30
Pozycja	1	2	3	4	5	6	7	8	9	10	11	12	13
RANGA	1	2,5	2,5	5,5	5,5	5,5	5,5	8	9	11	11	11	13

- Obliczenie rang wiązanych (wartości zmiennej powtarzają się):** dodajemy pozycje na których występują te same wartości i dzielimy sumę przez ich liczbę czyli liczymy średnią z pozycji
- obliczanie rangi wartości zmiennej „10”: „10” znajdują się na pozycji 2 i 3, zatem dodajemy numery pozycji i dzielimy przez ich liczbę $(2 + 3)/2 = 2,5$
- obliczanie rangi wartości zmiennej „13”: „13” znajdują się na pozycji 4,5,6,7 więc $(4+5+6+7)/4=5,5$
- obliczanie rangi wartości zmiennej „22”: „22” znajdują się na pozycji 10,11,12 więc $(10 + 11 + 12)/3=11$

17

RHO SPEARMANA

$$\rho = 1 - \frac{6 \sum_{i=1}^N d_i^2}{N * (N^2 - 1)}$$

d^2 - podniesiona do kwadratu różnica między rangami w parach

współczynnik rho spełnia warunki

- $\rho = 1$ jeśli uporządkowania są takie same
- $\rho = -1$ jeśli uporządkowania są przeciwne
- $\rho = 0$ jeśli brak uporządkowania (kolejność przypadkowa)

18

RHO SPEARMANA

ZASTOSOWANIE I WŁASNOŚCI

jest rangowym odpowiednikiem r Pearsona – obliczamy dla zmiennych ilościowych:

- jeśli związek między zmiennymi ilościowymi nie jest liniowy
- jeśli nie są spełnione założenia obliczenia r Pearsona (brak normalności rozkładu)
- jeśli występują wartości odstające (ρ jest odporne na wartości odstające – „uwzględnia” ich rangi, a nie wartości)
- jeśli związek jest liniowy wartości r i ρ są równe
- obliczając ρ wg wzoru zakładamy, że rangi są liczbami całkowitymi
- jeśli są rangi wiązane (nie są wartościami całkowitymi) to ρ traci swoją wartość w miarę wzrostu liczby rang wiązanych, co najwyżej $1/3$ wyników może być uwikłana w rangi wiązane
- jeśli zbiory wartości korelowanych zmiennych są „krótkie” (zmienna przyjmuje tylko kilka wartości) to pojawia się zbyt wiele rang wiązanych i współczynnik ρ nie jest odpowiednią miarą korelacji

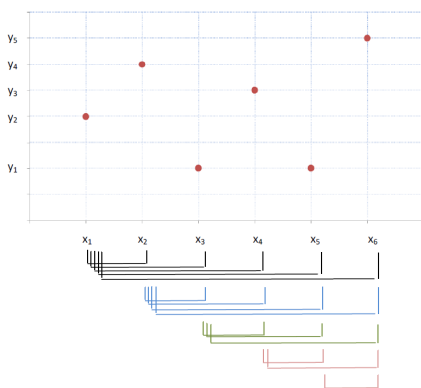
PRZYKŁAD

Współczynniki korelacji między stażem pracy za liczbą dni urlopu $\rho = 0,3$

- zależność nie jest liniowa
- korelacja między stażem a wysokością zarobków jest niska i dodatnia, wraz ze wzrostem długości zatrudnienia rośnie liczba dni urlopu

19

UPORZĄDKOWANIE PAR



Sprawdzamy co się dzieje z wartościami drugiej zmiennej (y) jeśli wartości jednej zmiennej (x) rosną

- analizujemy parę $x_1 - x_2$ ($x_1 < x_2$) odpowiada jej para $y_2 - y_4$ przy czym $y_2 < y_4$ czyli wartości x wzrosły i wartości y wzrosły – odnotowujemy wzrost (+)
- analizujemy parę $x_1 - x_3$ ($x_1 < x_3$) odpowiada jej para $y_2 - y_1$ przy czym $y_2 > y_1$ czyli wartości x wzrosły, wartości y zmalały – odnotowujemy spadek (-)
- analizujemy parę $x_1 - x_4$ ($x_1 < x_4$) odpowiada jej para $y_2 - y_3$ przy czym $y_2 < y_3$ czyli wartości x wzrosły, wartości y też wzrosły – odnotowujemy wzrost (+)
- analizujemy parę $x_1 - x_5$ ($x_1 < x_5$) odpowiada jej para $y_2 - y_1$ przy czym $y_2 > y_1$ czyli wartości x wzrosły, wartości y zmalały – odnotowujemy spadek (-)
- analizujemy parę $x_1 - x_6$ ($x_1 < x_6$) odpowiada jej para $y_2 - y_5$ przy czym $y_2 < y_5$ czyli wartości x wzrosły, wartości y wzrosły – odnotowujemy spadek (-)
-
• analizujemy parę $x_3 - x_5$ ($x_3 < x_5$) odpowiada jej para $y_1 - y_1$ przy czym $y_1 = y_1$ czyli wartości x wzrosły, wartości y nie zmieniły się (wiązane rangi zmiennej y) – odnotowujemy brak zmiany (0)
-

Procedurę należy powtórzyć dla wszystkich par zmiennej x

- liczba wszystkich (+) oznacza sytuację, kiedy wzrostowi wartości jednej zmiennej towarzyszy wzrost wartości drugiej
- liczba wszystkich (-) oznacza sytuację, kiedy wzrostowi wartości jednej zmiennej towarzyszy spadek wartości drugiej zmiennej
- Liczba wszystkich (0) oznacza sytuację, kiedy wzrostowi wartości jednej zmiennej nie towarzyszą zmiany wartości drugiej zmiennej

Wszystkich par zmiennych jest $n(n-1)/2$

20

TAU KENDALLA

WARTOŚĆ TAU

- jest miarą podobieństwa uporządkowań
- informuje, jak bardzo liczba par o ustalonym porządku (np. rosnących) przewyższa liczbę par o porządku przeciwnym (malejących) czyli która sytuacja występuje częściej jak rosną wartości jednej zmiennej czy (1) częściej wartości drugiej zmiennej rosną (tau dodatnie) czy (2) maleją (tau ujemne).
- to różnica między prawdopodobieństwem tego, że dwie zmienne układają się w tym samym porządku (obie maleją lub rosną) w obrębie obserwowanych danych a prawdopodobieństwem, że ich uporządkowanie się różni (jedna maleje, druga rośnie lub odwrotnie)
- uwzględnia rangi wiązane - na wyniki duży wpływ ma częstotliwość ich występowania
- jest porównywalne z rho Spearmana co do siły statystycznego wnioskowania
- logika leżąca u podstaw definicji rho i tau, a także same formuły obliczeniowe są różne, stąd dają one nieco inne wyniki
- zaleca się by zmienna była mierzona przynajmniej na 5 stopniowej skali porządkowej. W przypadku krótszych skal porządkowych należy skorzystać ze współczynników korelacji dla zmiennych nominalnych

21

WSPÓŁCZYNNIKI RANGOWE

TAU B KENDALLA

- należy stosować do zmiennych, które mają jednakowe liczby różnych wartości

TAU-C KENDALLA

- zaleca się stosować do zmiennych, które mierzone skalach o niejednakowej liczbie wartości

D SOMMERSA

- można stosować w przypadku analizy związków niesymetrycznych – przyczynowo-skutkowych

GAMMA

- niezalecany, ignoruje rangi wiązane i zawyża korelację

PRZYKŁAD

Współczynniki korelacji między miejscem zamieszkania a częstością chodzenia do kina $\tau - b = 0,6$

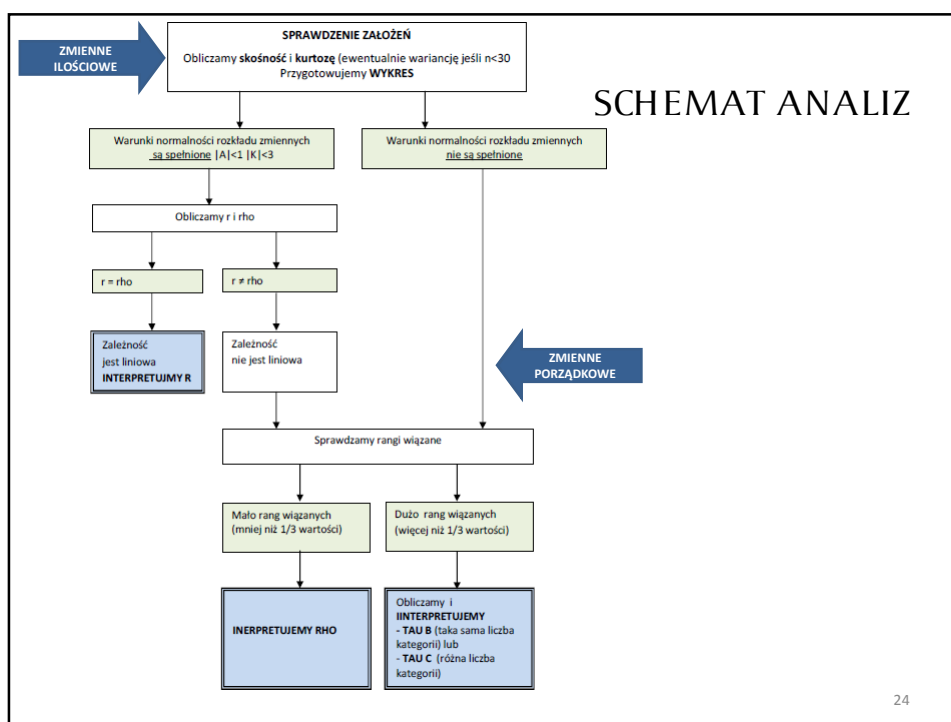
- istnieje dodatnia i silna korelacja między miejscem zamieszkania a częstością chodzenia do kina
- im większa miejscowość zamieszkania tym częściej badani chodzą do kina

22

WYBÓR WSPÓŁCZYNNIKA KORELACJI DLA ZMIENNYCH PORZĄDKOWYCH

	SYMETRYCZNY ZWIĄZEK	PRZYCZYNOWO-SKUTKOWY ZWIĄZEK
MAŁA LICZBA RANG WIĄZANYCH	rho Spearmana	d Sommersa
TAKA SAMA LICZBA WARTOŚCI X i Y	tau-b	d Sommersa
DYSPROPORCJA W LICZBIE WARTOŚCI X i Y	tau-c	tau-c

23



24

ISTOTNOŚĆ WSPÓŁCZYNNIKA KORELACJI

ISTOTNOŚĆ STATYSTYCZNA

- Współczynnik korelacji interpretujemy zaczynając od sprawdzenia czy jest **istotny statystycznie** czyli czy można uzyskany wynik mówiący o korelacji między zmiennymi (wartość współczynnika korelacji) **uogólnić na populację**, z której pochodzi próba.
- Jeśli korelacja nie jest istotna statystycznie to nie wiemy czy zachodzi w populacji
- Jeśli korelacja stwierdzona w próbie nie jest istotna statystycznie to nie ma powodów by się nią zajmować ☺
- O istotności korelacji orzekamy na podstawie tzw. p-wartości wyliczanej przez program statystyczny

25

P - WARTOŚĆ

P-WARTOŚĆ

- istotność statystyczna, dokładna informacja o tym, jakie jest prawdopodobieństwo błędu (pomyłki) przy uogólnienia wyniku z próby na populację (*a dokładnie rzecz biorąc przy odrzucaniu prawdziwej hipotezy zerowej – to informacja na przyszłość*) czyli
- p-wartość interpretujemy jako prawdopodobieństwo pomyłki przy podejmowaniu decyzji - uogólnić czy nie uzyskany wynik na populację.

- jeżeli **$p < 0,05$** to korelacja **jest istotna** statystycznie
- jeżeli **$p > 0,05$** to korelacja **nie jest istotna** statystycznie

- jeśli współczynnik korelacji **jest istotny** – podajemy wartość współczynnika, kierunek zależności i siłę związku oraz dokonujemy interpretacji (**$p < 0,5$**)
- Jeśli współczynnik korelacji **nie jest istotny** to można stwierdzić, że w (tych!) badaniach nie potwierdzono istnienia korelacji, **co absolutnie nie oznacza, że ona nie istnieje ☺ ($p > 0,5$)**
- $p < 0,05$ oznacza to, że wyniki z próby można uogólnić na populację, z której została wylosowana, dopuszczamy przy tym, że w 5 przypadkach na 100 podejmiemy błędną decyzję stwierdzając korelację w populacji!

26

PROCEDURA INTERPRETACJI

ETAPY INTERPRETACJI WSPÓŁCZYNNIKA KORELACJI W NAUKACH SPOŁECZNYCH

1. Istotność statystyczna (podstawa do uogólnienia na populację)
2. Kierunek związku (+ lub -)
3. Siła zależności (wartość bezwzględna współczynnika korelacji)
4. Współczynnik determinacji (dla r Pearsona)

27

PRZYKŁAD r -Pearsona

Korelacje			
		staz	zarobki
staz	Korelacja Pearsona	1	,567**
	Istotność (dwustronna)		,000
	N	160	160
zarobki	Korelacja Pearsona	,567**	1
	Istotność (dwustronna)	,000	
	N	160	160



** Korelacja istotna na poziomie 0.01 (dwustronnie).

INTERPRETACJA

Korelacja jest istotna statystycznie $p < 0,05$ (a nawet $p < 0,001$, ale nie wolno napisać, że $p = 0,0!$)

Korelacja jest dodatnia, co oznacza, że wraz ze wzrostem stażu pracy rosną liniowo zarobki.

Korelacja między stażem pracy a zarobkami jest wysoka ($r = 0,567$). Współczynnik determinacji wynosi $r^2 = 0,3215$ czyli 32,15% zmienności zarobków można wyjaśnić długością zatrudnienia. Wspólna wariancja wynosi 32,15%. Pozostałe 67,85% zmienności zarobków zależy od innych czynników np. zaangażowania w pracę, zajmowanego stanowiska, wykształcenia, itd.

28

PRZYKŁAD ρ -Spermana

Korelacje

			zm_12 liczba dni w roku poświęconych na naukę języka obcego	zm_13 Wynik testu z języka obcego (w %)
rho Spearmana	zm_12 liczba dni w roku poświęconych na naukę języka obcego	Współczynnik korelacji	1,000	,472**
		Istotność (dwustronna)	.	,000
		N	123	123
	zm_13 Wynik testu z języka obcego (w %)	Współczynnik korelacji	,472**	1,000
		Istotność (dwustronna)	,000	.
		N	123	123

WARTOŚĆ P

**. Korelacja istotna na poziomie 0.01 (dwustronnie).

INTERPRETACJA

Korelacja między czasem poświęcanym na naukę języka obcego a wynikiem testu z języka obcego jest istotna statystycznie ($p < 0,001$). Współczynnik korelacji $\rho = 0,472$. Korelacja jest dodatnia, co oznacza, że wraz ze wzrostem liczby dni nauki rośnie wynik z testu. Korelacja między liczbą dni nauki a wynikiem testu jest umiarkowana.

29

PRZYKŁAD τ - B

Miary symetryczne

		Wartość	Błąd standardowy asymptotyczny ^a	Przybliżone T ^b	Istotność przybliżona
Porządkowa przez Porządkowa	tau-b Kendalla	,497	,063	7,721	,000
	tau-c Kendalla	,480	,062	7,721	,000
	Korelacja Spearmana	,611	,072	7,987	,000 ^c
Przedziałowa przez Przedziałowa	R Pearsona	,632	,067	8,434	,000 ^c
N ważnych obserwacji		109			

WARTOŚĆ P

a. Nie zakładając hipotezy zerowej.

b. Użyto asymptotycznego błędu standardowego przy założeniu hipotezy zerowej.

c. W oparciu o aproksymację rozkładu normalnego.

INTERPRETACJA

zmienne: pozytywne widzenie przyszłości, koncentracja na planach (obie zmienne mierzone na skali porządkowej, które mają taką samą liczbę kategorii)

Korelacja między zmiennymi jest istotna statystycznie ($p < 0,001$). Korelacja jest dodatnia i umiarkowana ($\tau_b = 0,497$), co oznacza, że wraz z bardziej pozytywnym widzeniem przyszłości częściej rośnie poziom koncentracji na planach. Wzrost koncentracji na planach częściej współwystępuje ze wzrostem pozytywnego postrzegania przyszłości. Wzrostowi pozytywnego postrzegania przyszłości towarzyszy wzrost koncentracji na planach.

30

WSPÓŁCZYNNIKI KORELACJI DLA ZMIENNYCH NOMINALNYCH

- Brak założeń dotyczących rozkładu
- Liczba wartości obu zmiennych decyduje o wyborze współczynnika korelacji
- Nie ma interpretacji kierunku związku. Interpretacja wymaga wglądu w tabelę krzyżową.
- Zakres przyjmowanych wartości <0 – brak związku, 1 – >
- Łatwo interpretuje się 0 i 1, są trudności w interpretacji wartości pośrednich
- Współczynniki te są najczęściej wykorzystywane jako ilościowe miary siły związku/zależności między zmiennymi czyli mierzą tzw. „wielkość efektu”

Kategorie siły związku/efektu wg COHENA

- od 0,2 – efekt mały
- od 0,5 – efekt przeciętny,
- od 0,8 – efekt duży

31

CHI KWADRAT

Rozkład empiryczny O
obserwowany (z badań)

Platforma internetowa	Studenci	Niestudujący	Razem
Amazon	20	10	30
Zalando	40	10	50
Allegro	40	80	120
Suma	100	100	200

Rozkład teoretyczny E
oczekiwany (obliczony)

Platforma internetowa	Studenci	Niestudujący	Razem
Amazon	15	15	30
Zalando	25	25	50
Allegro	60	60	120
Suma	100	100	200

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

O – liczebność empiryczna komórki
E – liczebność teoretyczna komórki

(O - E)
komórka (1w, 1k) = 20 - 15
komórka (1w, 2k) = 10 - 15
.....

komórka (3w, 2k) = 80 - 60

32

CHI KWADRAT

CHI KWADRAT

- Statystyka chi kwadrat jest miarą różnicy między tabelą rozkładu empirycznego i teoretycznego (rozkładami zmiennych). Obliczenie jak bardzo różnią się od siebie tabele rozkładów polega na wyznaczeniu statystyki χ^2
- Statystyka bazuje na porównaniu ze sobą wartości obserwowanych (otrzymanych w badaniu) a wartości teoretycznych - pokazuje, jak bardzo liczebności obserwowane odbiegają od liczebności oczekiwanych
- Duże różnice wskazują na istnienie zależności pomiędzy zmiennymi.
- Im większe rozbieżności między rozkładami tym większa wartość chi kwadrat
- Jeżeli wartość chi kwadrat jest bliska zero między zmiennymi nie ma związku, zmienne nie są skorelowane

LICZEBNOŚCI

- obserwowane (empiryczne) – uzyskane w badaniach
- oczekiwane (teoretyczne) – obliczone przy założeniu, że pomiędzy zmiennymi nie ma żadnego związku

33

WSPÓŁCZYNNIKI KORELACJI DLA ZMIENNYCH NOMINALNYCH

ϕ – PHI YULA

- jest miarą korelacji między dwiema w tabeli 2×2
- jest miarą koncentracji przypadków na przekątnej
- w SPSS phi może być ujemne

$$\phi = \sqrt{\frac{\chi^2}{N}}$$

C – KONTYNGENCJI

- Stosuje się go tylko wtedy, kiedy dwie zmienne mają taką samą liczbę kategorii (czyli do tablic kwadratowych)
- Liczba kategorii w ramach każdej zmiennej może być może być równa 2 lub więcej
- Maksymalna wartość zależy od rozmiaru tabeli.
- C może osiągnąć wartość 1 jedynie dla nieskończonej liczby kategorii
dla n=2, max C=0,707; dla n=3, max C= 0,816; dla n=4, max C=0,866; n=∞, C=1

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

V – CRAMERA

- Stosuje się do zmiennych nominalnych o nierównej liczbie kategorii (czyli do tabel niesymetrycznych)

$$V = \sqrt{\frac{\chi^2}{N(k-1)}}$$

N – liczba obserwacji w próbie
 χ^2 – wartość statystyki chi kwadrat
 k – mniejsza liczba wierszy lub kolumn

34

ETA η

ETA

- jedna zmienna ilościowa a druga nominalna
- **eta²** jest miarą, która mówi, jaka część wariancji (zmienności) zmiennej ilościowej jest wyjaśniana przez przynależność do kategorii drugiej zmiennej.

r_{pbi} – PUNKTOWO-BISERYJNE

- współczynnik korelacji pomiędzy zmienną dwukategorialną (dychotomiczną) a zmienną ilościową
- Zmienna ilościowa może być sztucznie zdychotomizowana
- Szczególny przypadek eta kiedy zmienna kategorialna ma 2 wartości
- Najczęstsze zastosowanie: korelacja pomiędzy pozycją testową typu Tak/Nie a wynikiem testu
- Ogólniejsze zastosowanie: korelacja zmiennej dychotomicznej z ilościową, np. płeć a wynik testu, posiadanie konta na FB a czas korzystania z internetu
- Im bardziej liczebności obu kategorii odbiegają od stosunku 50%:50%, tym mniejsze wartości przyjmuje r_{pbi}

35

PROPORCJONALNA REDUKCJA BŁĘDU

PRE proporcjonalna redukcja błędu (PRE - ang. Proportional reduction of error)

- metody przewidywania wartości jednej zmiennej na podstawie drugiej zmiennej
- mając informację o współczynniku korelacji można poprawić przewidywania wartości jednej zmiennej na podstawie wartości drugiej zmiennej
- im związek zmiennymi jest silniejszy tym, większa jest redukcja błędu przewidywania

PRZYKŁAD

- płeć: kobiety i mężczyźni; wynik eksperymentu: pomaga, nie pomaga
- przewidując czy dana osoba pomoże (bez wiedzy czy jest kobietą czy mężczyzną) wynik przewidujemy losowo (połowa osób pomoże, druga połowa nie pomoże)
- Jeśli wiemy jaki jest związek między zmiennymi (np. wiemy, że częściej pomagają kobiety) to przewidując czy dana osoba pomoże czy nie, jeśli znamy jej płeć możemy poprawić przewidywanie jej decyzji
- jeśli wiemy, że pomogła, to możemy poprawić przewidywanie tego, że jakiej jest płci

Współczynniki mające interpretację w kategoriach PRE

- phi Yula, tau b, d Sommersa
- **Lambda λ** - współczynnik lambda odnosi się do tego, na ile znajomość rozkładu jednej zmiennej poprawia predykcję rozkładu drugiej zmiennej
- po podniesieniu do kwadratu: r , ρ , eta

36

ZWIĄZEK PRZYCZYNOWY

- związek między zmiennymi
- odpowiedni porządek w czasie
- wykluczenie alternatywnych wyjaśnień (poprzez kontrolę wpływu innych zmiennych i wpływu błędów próby)

37

WSPÓŁCZYNNIKI ZGODNOŚCI *

W KENDALLA

- Miara zgodności wyników kilku osób
- Wykorzystywany do oceny zgodności opinii różnych osób, dotyczących tej samej rzeczy, tego samego zjawiska
- Jest wykorzystywany do oceny zgodności ocen dokonywanych przez sędziów kompetentnych

KAPPA COHENA

- stosowany dla oszacowania zgodności dwóch ocen (np. wydawanych przez dwóch sędziów/ekspertów, którzy oceniają te same obiekty, ocena zgodności opinii małżonków, itp.)
- dane dotyczące oceny tego samego przedmiotu/zjawiska pochodzą od dwóch osób
- oceny powinny być wystawiane na tej samej skali (kategorie ocen powinny być identyczne)

38