Dr Tadeusz Sozański

Badanie zależności zmiennych porządkowych i interwałowych

Rozważmy dwie *zmienne* x i y określone dla tej samej *zbiorowości* złożonej z n jednostek (wartości zmiennych x i y dla i-tej jednostki badanej będziemy oznaczać x_i i y_i). Załóżmy, że obie zmienne są mierzalne w sposób co najmniej *porządkowy*. *Zależność dodatnią* między x i y określa się jako stan rzeczy polegający na tym, że „wzrostowi wartości zmiennej x na ogół towarzyszy wzrost wartości zmiennej y". Inaczej mówiąc, jeśli dla pewnej pary jednostek i, j mamy $x_i<x_j$, to „na ogół" zachodzi także nierówność $y_i<y_j$. Wówczas, jeśli np. x oznacza „wykształcenie" a y „dochód", osoba i, mająca wykształcenie niższe od osoby j, musiałaby mieć także niższy od niej dochód. Gdyby sytuacja taka powtarzała się bez wyjątków, mielibyśmy do czynienia z *doskonałą zależnością dodatnią*. Z kolei *doskonała zależność ujemna* miałaby miejsce, gdyby nierówność $x_i<x_j$ zawsze pociągała za sobą nierówność $y_i>y_j$. Idealną zależność jednego lub drugiego rodzaju rzadko spotyka się w rzeczywistych danych, dlatego też ocena *kierunku zależności* musi się opierać na porównaniu liczby par, dla których porządek według wartości zmiennej y jest taki sam jak porządek według wartości zmiennej x, i liczby par, dla których porządek ten ulega odwróceniu. Dokładniej, dla dowolnej nieuporządkowanej pary jednostek $\{i,j\}$ zachodzi jeden z pięciu warunków:

(1) $x_i<x_j$ i $y_i<y_j$ lub $x_i>x_j$ i $y_i>y_j$; taką parę nazywamy *zgodnie uporządkowaną*;

(2) $x_i<x_j$ i $y_i>y_j$ lub $x_i>x_j$ i $y_i<y_j$; taką parę nazywamy *niezgodnie uporządkowaną*;

(3) $x_i=x_j$ i $y_i\neq y_j$; mówimy, że para jest *powiązana (zawiera węzeł) ze względu na zmienną x*;

(4) $x_i\neq x_j$ i $y_i=y_j$; mówimy, że para jest *powiązana (zawiera węzeł) ze względu na zmienną y*;

(5) $x_i=x_j$ i $y_i=y_j$; mówimy wtedy, że para jest *podwójnie powiązana*;

Liczbę par typu (1)–(5) oznacza się odpowiednio P, Q, X_0, Y_0, Z_0. Liczba $^1/_2 n(n-1)$ wszystkich nieuporządkowanych par, jakie można utworzyć z n jednostek, jest zatem równa sumie $P+Q+X_0+Y_0+Z_0$. *Zależność dodatnia* to przypadek $P>Q$ (pary zgodnie uporządkowane przeważają nad niezgodnie uporządkowanymi), zaś przypadek $P<Q$ (pary niezgodnie uporządkowane występują częściej od zgodnie uporządkowanych) to *zależność ujemna*; gdy $P=Q$, mówimy o braku zależności. Zauważmy także, że choć jedną ze zmiennych oznaczyliśmy tu symbolem x, a drugą y, pierwszej przypisując niejako rolę zmiennej niezależnej (wyjaśniającej), a drugiej rolę zmiennej zależnej (wyjaśnianej), w istocie obie zmienne występują tu „na równych prawach". Pojęcie zależności jest *symetryczne*, stąd należałoby mówić raczej o *współzależności*.

Opierając się na różnicy $P-Q$, Kendall zdefiniował trzy współczynniki zależności dla zmiennych porządkowych τ_a (tau-a), τ_b (tau-b), τ_c (tau-c). Współczynnik τ_a, który otrzymuje się dzieląc $P-Q$ przez $^1/_2 n(n-1)$, stosuje się przy braku węzłów ($X_0=0$, $Y_0=0$, $Z_0=0$). W badaniach socjologicznych zmienne porządkowe, takie jak np. wykształcenie, mają zwykle niewielką liczbę wartości w porównaniu z liczbą jednostek w próbie, a więc pewne wartości muszą się powtarzać. W tej sytuacji stosuje się zazwyczaj współczynnik τ_b, dany wzorem

$$\tau_b=\frac{P-Q}{\sqrt{(P+Q+X_0)\,(P+Q+Y_0)}} \qquad (1)$$

Wielkość $P+Q+Y_0$ daje się przedstawić inaczej jako $^1/_2 n(n-1)-(X_0+Z_0)$, z kolei X_0+Z_0 to liczba par i,j takich, że $x_i=x_j$. Jeśli zmienna x przyjmuje k różnych wartości $x_1^*,\ldots,x_k^*$, z częstościami $n_1,\ldots,n_k$, wówczas liczba par jednostek przyjmujących wartość x_j^* jest równa $^1/_2 n_j(n_j-1)$, itd. Sumując te wielkości dla $j=1,\ldots,k$ dostaniemy zatem X_0+Z_0. Podobnie oblicza się Y_0+Z_0, a następnie $P+Q+X_0=^1/_2 n(n-1)-(Y_0+Z_0)$.

Pokażemy teraz jak wyznaczyć P i Q na przykładzie tabeli (dane fikcyjne, ale realistyczne), w której przedstawiono łączny rozkład liczebności dla dwu zmiennych porządkowych x i y opisanych jako „wykształcenie" i „stosunek do kary śmierci". Wykształcenie jest tu zmienną o trzech wartościach: 1 (wykształcenie „co najwyżej zawodowe"), 2 („średnie") i 3 („wyższe"). Wartości 1, 2, 3 i 4 zmiennej y przypisanej pytaniu „Czy jesteś za przywróceniem kary śmierci?" oznaczają odpowiednio

odpowiedzi „Zdecydowanie nie”, „Nie”, „Tak” i „Zdecydowanie tak”. Jak widać, cała populacja, licząca 500 jednostek, składa się z 12 klas odpowiadających wszystkim możliwym połączeniom wartości zmiennych *x* i *y*. Liczebności klas umieszczono w „komórkach” tabeli ponumerowanych od (1) do (12).

y Czy jesteś za karą śmierci?	*x* Wykształcenie			Razem
	1 Podst/zawod.	2 Średnie	3 Wyższe	
1 Zd. nie	(1) 20	(2) 20	(3) 10	50
2 Nie	(4) 30	(5) 20	(6) 50	100
3 Tak	(7) 150	(8) 70	(9) 30	250
4 Zd. Tak	(10) 50	(11) 40	(12) 10	100
Razem	250	150	100	500

Rozważmy dowolną parę jednostek i załóżmy, że jeden z dwu elementów należy do klasy (1). Jeśli drugi element pary leży w tej samej komórce, oba elementy nie różnią się wartością zarówno *x* jak i *y* (podwójny węzeł); gdy drugi element pochodzi z którejś z komórek (4), (7) lub (10) leżących w tej samej kolumnie tabeli, mamy węzeł ze względu na *x*, natomiast węzeł ze względu na *y* pojawia się, gdy drugi element weźmiemy z komórki (2) lub (3) leżącej w tym samym wierszu tabeli. Zauważmy teraz, że wszystkie pary o drugim elemencie należącym do którejś z pozostałych komórek (5), (6), (8), (9), (11) i (12) są zgodnie uporządkowane. Par takich, że jeden element pochodzi z komórki (1), a drugi z jakiejś komórki spośród 6 komórek położonych w tabeli poniżej i na prawo od komórki (1) jest 20(20+50+70+30+40+10)=20·220=4400.

Pierwszy element z komórki	Liczba par zgodnych	Liczba par niezgodnych
1	20(20+50+70+30+40+10) = 4400	–
2	20(50+30+10) = 1800	20(30+150+50) = 4600
3	–	10(30+20+150+70+50+40) = 3600
4	30(70+30+40+10) = 4500	–
5	20(30+10) = 800	20(150+50) = 4000
6	–	50(150+70+50+40) = 15500
7	150(40+10) = 7500	–
8	70·10 = 700	70·50 = 3500
9	–	30(50+40) = 2700
	P = 19700	*Q* = 33900

Niech teraz pierwszy element pary pochodzi z komórki (2). Jak poprzednio, pary zgodne powstają przez dodanie elementów z komórek położonych poniżej i na prawo od komórki (2); są to komórki (6), (9) i (12), a liczba takich par jest równa 20(50+30+10)=20·90=1800. Jeśli natomiast dołączyć do elementu z komórki (2) element pochodzący z którejś z komórek (4), (7), (10) położonych poniżej i na lewo od komórki (2), otrzymamy parę niezgodnie uporządkowaną. Par takich jest 20(30+150+50) = 20·230=4600. Zauważmy dalej, że biorąc pierwszy element z komórki (3), a drugi z komórek położonych poniżej i na lewo od tej komórki, otrzymamy zawsze parę niezgodną. Po przejściu do drugiego wiersza sytuacja o tyle się zmienia, że jako drugi element pary moglibyśmy wziąć jednostkę z pierwszego wiersza, jednak otrzymana para została uwzględniona już wcześniej, a zatem aby uniknąć podwójnego liczenia drugi element bierzemy zawsze z komórki położonej niżej w tabeli. Postępując w ten sposób, otrzymamy zestawienie podane w powyższej tabeli.

Tak więc $P-Q$= -14200 i zależność jest ujemna („*im wyższy* poziom wykształcenia, *tym niższa*

akceptacja kary śmierci"). Aby obliczyć współczynnik τ_b, trzeba jeszcze wyznaczyć X_0+Z_0 i Y_0+Z_0. Pierwsza z tych wielkości jest równa $^1/_2250\cdot249+^1/_2150\cdot149+^1/_2100\cdot99=47250$, a druga $^1/_250\cdot49+^1/_2100\cdot99+^1/_2250\cdot249+^1/_2100\cdot99 = 42250$. Mamy także $^1/_2n(n-1) = ^1/_2500\cdot499 = 124750$, a stąd $P+Q+Y_0=^1/_2n(n-1)-(X_0+Z_0) = 124750-47250 = 77500$ i $P+Q+X_0=^1/_2n(n-1)-(Y_0+Z_0)= 124750-42250 = 82500$. Pierwiastek z iloczynu tych wielkości w przybliżeniu równa się 79961, a stąd $\tau_b = -0.18$.

Współczynnik τ_b przyjmuje wartości z przedziału [-1,1], jednak wartości skrajne może osiągać jedynie wtedy, gdy $k=l$, gdzie k i l oznaczają liczby różnych wartości zmiennych odpowiednio x i y (w przykładzie $k=3$ i $l=4$). Aby temu zaradzić, Kendall zaproponował drugi sposób unormowania różnicy P-Q przez podzielenie jej przez $^1/_2(n^2(m-1)/m)$, gdzie m jest mniejszą z liczb k i l. Otrzymany w ten sposób współczynnik τ_c w naszym przykładzie jest równy -0.17, a więc niewiele różni się od τ_b. Oba warianty tau obliczane są w SPSS dla tabeli dwudzielczej w ramach komendy CROSSTABS.

Tamże dostępny jest jeszcze jeden współczynnik zależności zmiennych porządkowych oparty na różnicy $P-Q$, oznaczany literą γ (gamma) i obliczany według wzoru zaproponowanego przez Goodmana i Kruskala.

$$\gamma=\frac{P-Q}{P+Q} \tag{2}$$

Współczynnik ten przyjmuje wyższe wartości od τ_b i τ_c i z tego zapewne względu jest zwykle preferowany przez socjologów. W naszym przykładzie $\gamma=-14200/(19700+33900)= -14200/53600= -0.26$.

Badanie zależności zmiennych porządkowych można oprzeć także na porównaniu liczby jednostek zgodnie i niezgodnie położonych w stosunku do wartości środkowych w szeregach statystycznych x i y. Niech Me_x i Me_y oznaczają mediany zmiennych x i y. Jednostka i jest *zgodnie położona*, gdy $x_i>Me_x$ i $y_i>Me_y$ lub $x_i<Me_x$ i $y_i<Me_y$, a niezgodnie położona, gdy $x_i>Me_x$ i $y_i<Me_y$ lub $x_i<Me_x$ i $y_i>Me_y$. Stosownie do tego, który z trzech możliwych przypadków ma miejsce dla i-tej jednostki: $x_i>Me_x$, $x_i=Me_x$ i $x_i<Me_x$, przypiszmy jej odpowiednio liczbą 1,0 i -1, otrzymując w ten sposób zmienną zwaną *znakiem odchylenia od* Me_x. Podobnie definiuje się znak odchylenia od mediany zmiennej y. Iloczyn znaków odchyleń od median jest zmienną także przyjmującą wartości 1, 0, -1. Jej wartość wskazuje czy jednostka jest zgodnie czy niezgodnie położona, czy znajduje się w środku jednego z szeregów. Średnia arytmetyczna tej zmiennej, równa różnicy między liczbą jednostek zgodnie położonych a liczbą jednostek niezgodnie położonych, zwana *współczynnikiem korelacji ćwiartkowej*, może służyć także jako miara zależności zmiennych porządkowych.

•

Dla zmiennych ilościowych (mierzalnych w sposób co najmniej interwałowy) także rozważa się odchylenia wartości zmiennych od miar tendencji centralnej, tym razem jednak położenie i-tej jednostki w populacji określa się przez porównanie x_i ze średnią arytmetyczną M_x, przy czym uwzględnia się nie tylko *znak* odchylenia od średniej, ale i jego *rozmiar*. Mając dane dwie zmienne x i y o średnich M_x i M_y, możemy obliczyć sumę iloczynów odchyleń od średnich $\sum(x_i-M_x)(y_i-M_y)$ dla $i=1,\ldots,n$. W sumie tej mogą występować iloczyny dodatnie i ujemne oraz iloczyny zerowe (jedna ze zmiennych ma wartość równą średniej dla tej zmiennej), nie wpływające na wartość sumy, której wartość i znak informują zatem, które iloczyny, dodatnie czy ujemne „ważą" bardziej. Tym razem nie musi to oznaczać przewagi liczebnej odchyleń zgodnych nad odchyleniami niezgodnymi co do znaku, jak było w przypadku zmiennych porządkowych, gdyż liczy się teraz także wartość bezwzględna iloczynu odchyleń od średnich. Okazuje się ponadto, że określenie kierunku zależności oparte na odchyleniach od średnich jest równoważne określeniu wykorzystującemu iloczyny różnic wartości zmiennych dla poszczególnych par $\{i,j\}$. Zachodzi mianowicie następująca równość (dowodzi się ją na drodze czysto rachunkowej):

$$n\sum_{i=1}^{n}(x_i-M_x)(y_i-M_y)=\sum_{i<j}(x_i-x_j)(y_i-y_j) \tag{3}$$

Sumowanie po prawej stronie obejmuje wszystkie pary jednostek, a więc liczba składników jest równa

$^1/_2 n(n-1)$ (w parze $\{i,j\}$ elementy zapisuje się w kolejności podyktowanej przez porządek numerów, stąd oznaczenie $i<j$ pod znakiem sumy).

Średnią z iloczynów odchyleń od średnich nazywamy *kowariancją* zmiennych x i y i oznaczamy $Cov(x,y)$. Wzór definicyjny ma zatem postać:

$$Cov(x,y) = M_{(x-M_x)(y-M_y)} = \frac{1}{n}\sum_{i=1}^{n}(x_i-M_x)(y_i-M_y) \qquad (4^*)$$

(Gwiazdą zaznaczono wzory, które należy zapamiętać do egzaminu). Obliczając kowariancję z próby losowej jako oszacowanie nieznanej kowariancji w populacji generalnej (kowariancji dwu zmiennych losowych) sumę iloczynów odchyleń dzieli się nie przez n lecz przez $n-1$. *Kowariancja z próby* tak zdefiniowana jest wówczas „nieobciążonym estymatorem" kowariancji w populacji generalnej (tzn. przy pobieraniu wielu różnych prób z tej samej populacji oszacowanie „średnio" dawać będzie poprawny, nie zaniżony ani nie zawyżony wynik).

Zdefiniujemy teraz *współczynnik korelacji liniowej* r_{xy} (zwany często współczynnikiem Pearsona) zmiennych x i y jako iloraz kowariancji $Cov(x,y)$ i iloczynu odchyleń standardowych s_x i s_y.

$$r_{xy} = \frac{Cov(x,y)}{s_x\, s_y} \qquad (5^*)$$

Oczywiście parametr ten oblicza się jedynie wtedy, gdy żadna ze zmiennych nie jest stałą, tzn. $s_x>0$ i $s_y>0$. Współczynnik korelacji jest podstawową miarą zależności dla zmiennych ilościowych. Z oczywistych wzorów (symetria kowariancji i współczynnika korelacji liniowej)

$$Cov(x,y)=Cov(y,x),\ \ r_{xy} = r_{yx} \qquad (6)$$

wynika, że parametry te charakteryzują *wzajemny* związek dwu zmiennych zarówno co do siły (wartość bezwzględna) jak i znaku (zależność dodatnia lub ujemna). Współczynnik korelacji jest wygodniejszy od kowariancji ze względu na to, że przyjmuje wartości z przedziału [-1,1] (fakt ten udowodniony zostanie później). Do obliczania r_{xy} stosuje się zwykle następujący wzór wynikający wprost z definicji.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-M_x)(y_i-M_y)}{\sqrt{\sum_{i=1}^{n}(x_i-M_x)^2\sum_{i=1}^{n}(y_i-M_y)^2}} \qquad (7)$$

Dla przykładu zbadajmy zależność między ocenami z dwu przedmiotów w grupie 10 osób.

i	x	y	$x-M_x$	$y-M_y$	$(x-M_x)(y-M_y)$	$(x-M_x)^2$	$(y-M_y)^2$
1	4.0	5.0	0.45	1.35	0.6075	0.2025	1.8225
2	3.0	2.0	-0.55	-1.65	0.9075	0.3025	2.7225
3	3.0	4.0	-0.55	0.35	-0.1925	0.3025	0.1225
4	5.0	4.5	1.45	0.85	1.2325	2.1025	0.7225
5	2.0	3.0	-1.55	-0.65	1.0075	2.4025	0.4225
6	3.5	4.0	-0.05	0.35	-0.0175	0.0025	0.1225
7	4.5	4.5	0.95	0.85	0.8075	0.9025	0.7225
8	3.0	3.5	-0.55	-0.15	0.0825	0.3025	0.0225
9	4.0	3.0	0.45	-0.65	-0.2925	0.2025	0.4225
10	3.5	3.0	-0.05	-0.65	0.0325	0.0025	0.4225
	35.5	36.5			4.1750	6.7250	7.5250

Oceny te zestawiono niżej w dwu kolumnach oznaczonych x i y (i w pierwszej kolumnie to numer

jednostki badanej). Zaczynamy od zsumowania wartości zmiennych x i y, otrzymując liczby 35.5 i 36.5, skąd M_x=3.55 i M_y=3.65. Następnie obliczamy odchylenia od średnich dla obu zmiennych (kolumny $x-M_x$ i $y-M_y$), iloczyny odchyleń (kolumna $(x-M_x)(y-My)$ i kwadraty odchyleń od średniej dla x i dla y. Po zsumowaniu trzech ostatnich kolumn otrzymujemy: $\sum(x-M_x)(y-M_y)$=4.175, $\sum(x-M_x)^2$=6.725, $\sum(y-M_y)^2$=7.525, a stąd r_{xy}= 0.59.

Mając dowolne m zmiennych ilościowych określonych w tym samym zbiorze jednostek możemy wyznaczyć dla każdej uporządkowanej pary zmiennych o numerach (g,h) ich współczynnik korelacji, oznaczony r_{gh}, otrzymując w ten sposób tablicę o m wierszach i m kolumnach, zwaną *macierzą korelacji* (macierz ta służy jako „półfabrykat" w rozmaitych analizach wielozmiennowych). Macierz korelacji jest *symetryczna*, tzn. $r_{gh}=r_{hg}$ dla dowolnych g,h=1,…,m, a na *głównej przekątnej* ma same jedynki, tzn. r_{gg}=1 dla g=1,…,m. Łatwo wykazać, że współczynnik korelacji dowolnej zmiennej z nią samą jest zawsze równy 1. Zauważmy najpierw, że kowariancja jest uogólnieniem wariancji na przypadek dwu zmiennych:

$$Cov(x,x)=s_x^2 \tag{8}$$

W konsekwencji r_{xx}=Cov(x,x)/($s_x s_x$)=s_x^2/s_x^2=1. Podamy teraz dalsze własności kowariancji. Dla matematyka najważniejsza jest *dwuliniowość*, która wyraża się w następujących dwu wzorach

$$Cov(x+x',y) = Cov(x,y)+Cov(x',y),\ \ Cov(ax,y) = aCov(x,y) \tag{9}$$

łatwych do udowodnienia w oparciu o elementarne własności średniej arytmetycznej ($M_{x+y}=M_x+M_y$ i $M_{ax}=aM_x$). Dwuliniowość wraz z symetrią oraz następującą własnością kowariancji

$$Cov(x,c)=0 \tag{10}$$

(c jest *stałą*, tzn. $c_i=c$ dla i=,…,n), bardzo ułatwiają wyprowadzanie dalszych własności kowariancji i współczynnika korelacji liniowej. Zacznijmy od wzoru, który umożliwia obliczenie wariancji sumy dwu zmiennych.

$$s^2_{x+x'} = s^2_x + s^2_{x'} + 2Cov(x,x') = s^2_x + s^2_{x'} + 2r_{xx'}s_x s_{x'} \tag{11}$$

Mówimy, że zmienne x i x' są *nieskorelowane*, jeśli $Cov(x,x')$=0 (wtedy też $r_{xx'}$=0). Dla zmiennych nieskorelowanych wariancja sumy zmiennych równa jest sumie wariancji.

Bardzo ważną własnością współczynnika korelacji liniowej jest *niezmienniczość ze względu na przekształcenia liniowe dopuszczalne w systemie pomiaru interwałowego*. Za pomocą przekształceń liniowych $T(x)=ax+b$ i $U(y)=cy+d$, gdzie a>0 i c>0 utwórzmy nowe zmienne $x'=T(x)$ i $y'=U(y)$ (niech np. x i y oznaczają odpowiednio wzrost i wagę wyrażone w calach i funtach, a x' i y' te same wielkości wyrażone w centymetrach i kilogramach). Mamy wówczas

$$r_{xy} = r_{x'y'} \tag{12}$$

czyli wartość współczynnika korelacji się nie zmieni. Jeśli zastosujemy do zmiennej x przekształcenie liniowe takie, że a<0 (np przekształcenie $x'=6-x$, które „odwraca" wartości 1–5 przypisane elementom uporządkowanej 5-punktowej kafeterii dyzjunktywnej), a zmienną y pozostawimy bez zmian, współczynnik korelacji zmieni tylko znak na przeciwny.

Obliczenie kowariancji często upraszcza poniższy wzór, w którym xy oznacza iloczyn zmiennych x i y, czyli zmienną, której wartością dla i-tej jednostki jest iloczyn $x_i y_i$.

$$Cov(x,y) = M_{xy} - M_x M_y \tag{13}$$

Dla x=y wzór ten przyjmuje postać $s_x^2 = M_{x^2} - (M_x)^2$. Korzyść, jaką daje zastosowanie wzoru (13) polega na tym, że oszczędzamy na obliczaniu odchyleń od średnich dla każdej jednostki.

•

Wzór (13) można wykorzystać do wyprowadzenia wzoru na kowariancję dwu *zmiennych zerojedynkowych* *u* i *v*. Niech n_{11}, n_{10}, n_{01} i n_{00} oznaczają odpowiednio: liczbę przypadków, dla których zmienne *u* i *v* przyjęły wartość 1; liczbę przypadków, dla których zmienna *u* przyjęła wartość 1, a *v* wartość 0; liczbę przypadków, dla których zmienna *u* przyjęła wartość 0, a *v* wartość 1; liczbę przypadków, dla których *u* i *v* przyjęły wartość 0. Iloczyn zmiennych zerojedynkowych jest także zmienną zerojedynkową: *uv*=1 wtedy i tylko wtedy, gdy *u*=1 i *v*=1.

Jak wiadomo, średnia arytmetyczna zmiennej zerojedynkowej jest równa częstości względnej przypadków, dla których zmienna przyjmuje wartość 1. Mamy zatem $M_{uv}=n_{11}/n$, $M_u=n_{1.}/n$, $M_v=n_{.1}/n$, gdzie $n_{1.}=n_{11}+n_{10}$, $n_{.1}=n_{11}+n_{01}$ (częstości, z jakimi zaobserwowano wartość 0 dla zmiennych *u* i *v*, dane są podobnymi wzorami: $n_{0.}=n_{01}+n_{00}$, $n_{.0}=n_{10}+n_{00}$). Kowariancja zmiennych *u* i *v* wyraża się zatem wzorem $\frac{n_{11}}{n}-\frac{n_{1.}}{n}\frac{n_{.1}}{n}$. Liczebności n_{11}, n_{10}, n_{01} i n_{00}, oznaczane też *a,b,c,d* zestawia się zwykle w tabeli czteropolowej. Oto przykład, w którym wartości 1 i 0 zmiennych *u* i *v* oznaczają odpowiedzi „tak” i „nie” na dwa pytania dychotomiczne.

u	*v*		Razem
	1 (Tak)	0 (Nie)	
1 (Tak)	80	40	120
2 (Nie)	60	70	130
Razem	140	110	250

n_{11}	n_{10}	$n_{1.}$
n_{01}	n_{00}	$n_{0.}$
$n_{.1}$	$n_{.0}$	n

a	*b*	*a+b*
c	*d*	*c+d*
a+c	*b+d*	*a+b+c+d*

W naszym przykładzie: $n_{11}=a=80$, $n_{10}=b=40$, $n_{01}=c=60$, $n_{00}=d=70$, $n_{1.}=a+b=80+40=120$, $n_{.1}=a+c=80+60=140$, a stąd $M_{uv}=80/250=0.32$, $M_u=120/250=0.48$, $M_v=140/250=0.56$, $Cov(u,v)=0.32-0.48\cdot0.56 = 0.32-0.27 = 0.05$.

Kowariancję można obliczyć także za pomocą równoważnego wzoru $\frac{1}{n^2}(n_{11}n_{00}-n_{10}n_{01})$, w którym pojawia się wielkość $n_{11}n_{00}-n_{10}n_{01}=ad-bc$, zwana *iloczynem krzyżowym*. Obliczając iloczyn krzyżowy (mnożąc przez siebie dwie liczebności na głównej przekątnej \ i dwie liczebności na przeciwległej przekątnej /, po czym odejmując drugi iloczyn od pierwszego) możemy najszybciej sprawdzić czy dwie zmienne zerojedynkowe są niezależne (zerowanie się iloczynu krzyżowego) czy zależne dodatnio lub ujemnie. W naszym przykładzie $ad-bc=80\cdot70-40\cdot60=5600-2400=3200$, a więc zmienne *u* i *v* są zależne dodatnio.

Zauważmy teraz, że wariancje zmiennych *u* i *v* wyrażają się wzorami: $s_u^2=\frac{n_{1.}}{n}(1-\frac{n_{1.}}{n})=\frac{n_{1.}}{n}\frac{n_{0.}}{n}$, $s_v^2=\frac{n_{.1}}{n}(1-\frac{n_{.1}}{n})=\frac{n_{.1}}{n}\frac{n_{.0}}{n}$ (otrzymanymi ze wzoru (13) zastosowanego do pary identycznych zmiennych zerojedynkowych), skąd z kolei wynikają wzory na odchylenia standardowe: $s_u=\frac{\sqrt{n_{1.}n_{0.}}}{n}$, $s_v=\frac{\sqrt{n_{.1}n_{.0}}}{n}$. Teraz gotowe są już wszystkie elementy

potrzebne do wyprowadzenia wzoru na współczynnik korelacji zmiennych zerojedynkowych. Oto ten wzór

$$r_{uv} = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1.}n_{0.}n_{.1}n_{.0}}} \qquad (14)$$

Dla danych zestawionych w tabeli obliczyliśmy już iloczyn krzyżowy, otrzymując wartość 3200. Liczbę tę dzielimy teraz przez pierwiastek z 120·130·140·110= 240240000, równy 15599.7, otrzymując wartość 0.205.

Współczynnik korelacji można obliczać dla dowolnych zmiennych dwupunktowych, także jakościowych, zakładając jedynie, że dla każdej zmiennej jedna z wartości zostaje wyróżniona jako symbol kodowy zajścia pewnego zdarzenia *A* (*B*) interesującego badacza. Współczynnik *r* nazywamy wtedy *współczynnikiem korelacji zdarzeń A* i *B* i oznaczamy r_{AB}. W zależności od znaku pokazuje on czy częstość współwystępowania tych zdarzeń jest wyższa czy niższa niż częstość oczekiwana w sytuacji, gdy łączne zajście obu zdarzeń jest dziełem przypadku. Wartość 1 oznacza, że zdarzenia *A* i *B* zawsze zachodzą równocześnie, natomiast wartość -1 pojawia się wtedy, gdy zajściu *A* zawsze towarzyszy niezajście *B* (wtedy też zajściu *B* towarzyszy niezajście *A*). Przypadek r_{AB}=0 ma miejsce wtedy, gdy *B* zachodzi równie często wtedy, gdy *A* zachodzi, jak wtedy, gdy *A* nie zachodzi. Dodajmy jeszcze, że przypisanie liczb 1 i 0 odpowiednio zajściu i niezajściu wyróżnionego zdarzenia jest czysto arbitralne. Wartość współczynnika nie zmieni się, jeśli liczby te zastąpimy innymi, byle tylko zachowany został porządek. Możemy także zmienić porządek wartości na przeciwny (np. jedynkę zastępując zerem, a zero jedynką) pod warunkiem, że to samo zrobimy z drugą zmienną.

Przykładem zastosowania współczynnika korelacji zdarzeń w metodologii socjologicznej jest definicja *mocy rozdzielczej wskaźnika W dla indicatum I* jako r_{WI}.

Pearsonowski współczynnik wykorzystuje się także do badania zależności zmiennych porządkowych. Okazuje się mianowicie, że współczynnik τ_b Kendalla daje się zdefiniować jako współczynnik korelacji zmiennych sgn_x i sgn_y określonych na zbiorze par (*i*,*j*) takich, że *i*<*j* dla *i*,*j*=1,…,*n*. Wartość zmiennej sgn_x dla pary (*i*,*j*) to znak różnicy, równy -1,0 lub 1 odpowiednio gdy $x_i<x_j$, $x_i=x_j$, $x_i>x_j$. Analogicznie definiuje się sgn_y. Mamy wówczas $\tau_b = r_{(sgnx)(sgny)}$.

Inny współczynnik zależności zmiennych porządkowych, zwany *współczynnikiem korelacji rang*, został zdefiniowany przez Spearmana wprost jako współczynnik zmiennych rangowych r^x i r^y skonstruowanych ze zmiennych *x* i *y*. *Ranga i-tej jednostki ze względu na zmienną x* jest to numer, jaki otrzymuje dana jednostka po przepisaniu szeregu statystycznego w porządku niemalejących wartości zmiennej *x*. Po zmianie numeracji, jednostki przyjmujące tę samą wartość zmiennej *x* następują po sobie. Dla każdej takiej sekwencji przypadków oblicza się średnią arytmetyczną rang i liczbę tę przypisuje wszystkim jednostkom w ramach danej sekwencji, otrzymując w ten sposób *rangi skorygowane*; są to właśnie wartości zmiennej r^x.

(*i*)	x_i	r^x_i
(5)	2.0	1
(2)	3.0	2
(3)	3.0	3
(8)	3.0	4
(6)	3.5	5
(10)	3.5	6
(1)	4.0	7
(9)	4.0	8
(7)	4.5	9
(4)	5.0	10

(*i*)	y_i	r^y_i
(2)	2.0	1
(5)	3.0	2
(9)	3.0	3
(10)	3.0	4
(8)	3.5	5
(3)	4.0	6
(6)	4.0	7
(4)	4.5	8
(7)	4.5	9
(1)	5.0	10

r^x_i	r^y_i
1	1
3	3
3	3
3	3
5.5	5
5.5	6.5
7.5	6.5
7.5	8.5
9	8.5
10	10

Dla zmiennych *x* i *y*, dla których obliczyliśmy wyżej współczynnik korelacji liniowej, obliczymy teraz współczynnik korelacji rang Spearmana. Najpierw wyznaczamy rangi zwykłe i skorygowane ze

względu na x i y. Ponieważ później musimy zestawić obok siebie wartości r^x i r^y dla tych samych jednostek, w kolumnie oznaczonej (i) zapisujemy stare numery jednostek.

Przepisujemy teraz szereg rang skorygowanych ze względu na x w porządku jak wyżej, a następnie dopisujemy przy każdej pozycji rangę skorygowaną ze względu na y, odnajdując ją według starego numeru w szeregu umieszczonym wyżej po prawej stronie.

(i)	r^x	r^y	(a)	(b)	(ab)	(a^2)	(b^2)	(d)	(d^2)
(5)	1.0	3.0	-4.5	-2.5	11.25	20.25	6.25	-2.0	4.00
(2)	3.0	1.0	-2.5	-4.5	11.25	6.25	20.25	2.0	4.00
(3)	3.0	6.5	-2.5	1.0	-2.50	6.25	1.00	-3.5	12.25
(8)	3.0	5.0	-2.5	-0.5	1.25	6.25	0.25	-2.0	4.00
(6)	5.5	6.5	0.0	1.0	0.00	0.00	1.00	-1.0	1.00
(10)	5.5	3.0	0.0	-2.5	0.00	0.00	6.25	2.5	6.25
(1)	7.5	10.0	2.0	4.5	9.00	4.00	20.25	-2.5	6.25
(9)	7.5	3.0	2.0	-2.5	-5.00	4.00	6.25	4.5	20.25
(7)	9.0	8.5	3.5	3.0	10.50	12.25	9.00	0.5	0.25
(4)	10.0	8.5	4.5	3.0	13.50	20.25	9.00	1.5	2.25
	55.0	55.0			49.25	79.50	79.50		60.50

Pozostaje jeszcze obliczyć współczynnik korelacji zmiennych r^x i r^y. Średnie arytmetyczne dla tych zmiennych równe są M_{rx}=55.0/10=5.5, M_{ry}=55.0/10=5.5, zatem odchylenia od średnich oblicza się według wzorów: $a_i=r^x{}_i-5.5$, $b_i=r^x{}_i-5.5$. Mamy następnie: $\sum a_i b_i$=49.25, $\sum a_{i2}$=79.50, $\sum b_{i2}$=79.50, a stąd $r_S=(\sum a_i b_i)/\sqrt{(\sum a_{i2}\ \sum b_{i2})}$= 49.25/79.50=.62.

Jeśli nie występują „rangi wiązane” (nie powtarza się żadna z wartości zmiennych x i y), współczynnik korelacji Spearmana wyraża się wzorem

$$r_s = 1-\frac{6\sum_{i=1}^{n} d_i^{\,2}}{n(n^2-1)} \tag{15}$$

w którym d_i oznacza różnicę rang przypisanych i-tej jednostce ze względu na jedną i drugą zmienną. Wzór (15) stosuje się także przy niewielkiej liczbie węzłów, gdyż daje dość dobre przybliżenie. Jego zastosowanie do rang skorygowanych w powyższym przykładzie (patrz kolumny (d) i (d^2)) daje wynik 0.63. Współczynnikiem Spearmana posługujemy się najczęściej, gdy pierwotne dane mają już postać rangową, jak w przypadku, gdy dwie osoby prosimy o uporządkowanie tej samej listy przedmiotów, np. od najmniej potrzebnego do najbardziej potrzebnego na biwaku (eksperyment Hare'a). Współczynnik r_S obliczony dla pary uporządkowań mierzy wówczas zgodność dwu uporządkowań.

•

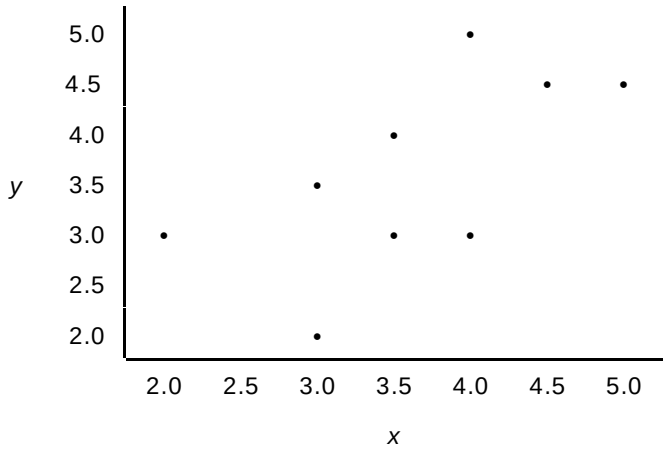
Pearsonowski współczynnik korelacji liniowej zdefiniowany został jako miara zależności dwu zmiennych interwałowych określonych w tej samej populacji (generalnej lub próbie). Opiszemy teraz sens, jaki posiada ten współczynnik w kontekście *analizy regresji dwu zmiennych*, wyjaśniając w szczególności dlaczego parametr ten nosi nazwę współczynnika korelacji *liniowej* (korelacja jest synonimem współzależności).

W naukach ścisłych związki między zmiennymi opisuje się za pomocą *funkcji* liczbo-liczbowych. Zależność funkcyjna między zmiennymi x i y oznacza zatem istnienie pewnej funkcji f, która dla każdego obiektu pozwala wyznaczyć wartość *zmiennej zależnej* y na podstawie znanej wartości *zmiennej niezależnej* x, tzn. $y_i=f(x_i)$ dla i=1,…,n. *Zależność liniowa* ma miejsce, gdy odwzorowanie f ma postać $f(x_i)=a+bx_i$. W analizie statystycznej dwu zmiennych także wychodzi się od pojęcia zależności funkcyjnej, zakłada się jednak, że dane obserwacyjne nie muszą być idealnie zgodne z teoretycznym modelem zależności, co z kolei tłumaczy się działaniem czynników ubocznych o charakterze losowym (w tym błędem pomiaru). Wartość zmiennej y *obserwowaną* dla i-tego

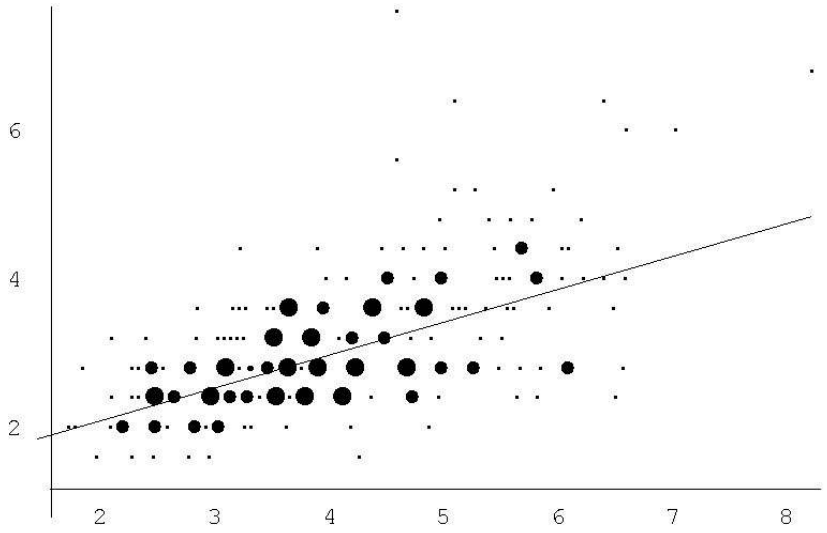
przypadku przedstawia się wówczas jako sumę wartości wyznaczonej dla tego przypadku przez zastosowanie funkcji *f* do obserwowanej wartości zmiennej *x*, oraz błędu (zakłócenia) losowego, symbolicznie $y_i=f(x_i)+e_i$. Zmienną $f(x)$ oznacza się zwykle ŷ, jej wartości nazywa się wartościami *teoretycznymi* albo *przewidywanymi* zmiennej zależnej. O zmiennej *e*, której wartości ($e_i=y_i-\hat{y}_i$) nazywamy *odchyleniami* wartości obserwowanych (empirycznych) od wartości przewidywanych (teoretycznych) zakłada się, że ma średnią 0, co oznacza, że $M_y=M_{\hat{y}}$, oraz jest nieskorelowana z ŷ ($Cov(\hat{y},e)=0$).

Badanie zależności dwu zmiennych zaczynamy zwykle od sporządzenia *wykresu rozrzutu*, przypisując *i*-tej jednostce badanej punkt (x_i,y_i) w dwywymiarowym układzie współrzędnych (na osi poziomej odkłada się wartości zmiennej niezależnej, a na osi pionowej wartości zmiennej zależnej). Jeśli ten sam punkt odpowiada pewnej liczbie przypadków („podwójny węzeł"), fakt ten należy odnotować na wykresie za pomocą stosownych środków graficznych.

Oto wykres rozrzutu sporządzony dla fikcyjnych danych (n=10), dla których obliczaliśmy wyżej współczynnik korelacji.



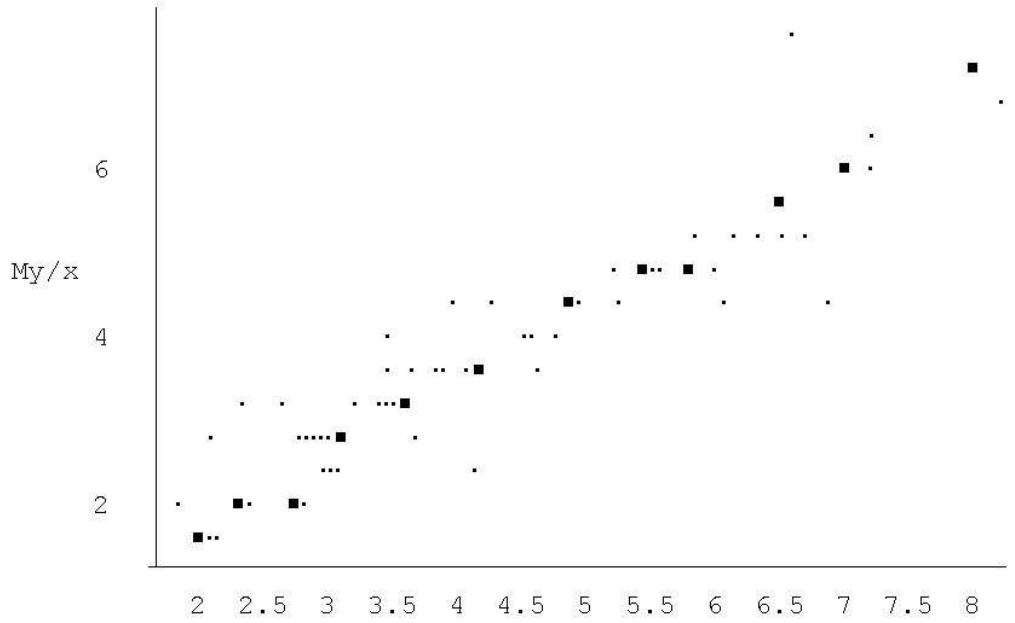
A oto bardziej skomplikowany przykład pochodzący z badań wykonanych w Krakowie w połowie lat 70-tych. Badano wówczas opinie studentów o płacach, zadając respondentom pytania o minimalną i maksymalną płacę, jaką powinien otrzymywać absolwent szkoły wyższej i absolwent szkoły średniej. Na poniższym wykresie zmienne *x* (na osi poziomej) i *y* oznaczają odpowiednio pożądaną płacę po studiach i po szkole średniej wyrażone w tysiącach ówczesnych złotych. Zmienne te otrzymano dzieląc sumę płacy minimalnej i maksymalnej przez 2 (n=704, M_x=4.02 s_x=1.00 M_y=3.00 s_y= 0.73 r_{xy}=0.69).



· 1-5 przypadków; ● 6-30 przypadków; ● 31 i więcej przypadków

Jak widać, punkty tworzą wydłużoną chmurę zagęszczającą się w otoczeniu punktu o współrzędnych (M_x,M_y). Jest oczywiste, że zmienne *x* i *y* są zależne dodatnio. Narysowano także prostą, która w przybliżeniu opisuje ten związek (już za chwilę wyjaśnimy w jaki sposób ją się wyznacza).

Zanim zdecydujemy się na wybór jakiejś funkcji, możemy najpierw przypisać wartościom zmiennej niezależnej *warunkowe średnie arytmetyczne* zmiennej zależnej: warunkową średnią $M_{y/xi}$ otrzymuje się przez zsumowanie wartości y_j dla wszystkich przypadków spełniających warunek $x_j=x_i$, a następnie podzielenie sumy przez liczbę takich przypadków. Wykres funkcji, która x_i przyporządkowuje $M_{y/xi}$, przedstawia poniższy rysunek.



Obraz zależności staje się teraz bardziej przejrzysty, średnia warunkowa dość regularnie rośnie ze wzrostem *x*, choć wykazuje też lokalne wahania, które można „wygładzić", obliczając średnie arytmetyczne *y* w zbiorach jednostek, dla których wartości zmiennej *x* leżą w kolejnych przedziałach klasowych (wyżej utworzono przedziały o długości .5 i środkach 2, 2.5 itd. zaznaczonych na osi). Punkty oznaczone kwadratami, odpowiadające tym średnim, w przybliżeniu układają się w linię prostą, co podsuwa myśl, by do danych zastosować model zależności liniowej.

Znalezienie funkcji „najlepiej pasującej" do danych wymaga, po pierwsze, wskazania klasy funkcji, do której ma należeć rozwiązanie problemu; po drugie, określenia kryterium dopasowania, dyktującego wybór „najlepszej" funkcji w danej klasie. Jako kryterium przyjmuje się minimalizację sumy kwadratów odchyleń wartości obserwowanych zmiennej zależnej od wartości przewidywanych,

tzn. minimalizację wyrażenia $\sum_{i=1}^{n} (y_i-\hat{y}_i)^2$. Postępowanie to nazywa się *metodą najmniejszych kwadratów.* Jeśli „dopuścić do gry" dowolne funkcje, rozwiązaniem okaże się funkcja $f(x_i)=M_{y/xi}$, zwana *funkcją regresji pierwszego rodzaju*. Ze względu na nieregularność tej funkcji zwykle wolimy jednak odwzorowanie bardziej „oddalone" od konkretnych danych, ale za to prostsze, jeśli brać pod uwagę wzór i kształt wykresu. Stąd naturalna decyzja, aby zachowując metodę najmniejszych kwadratów, rozwiązania problemu szukać w klasie funkcji liniowych. Każda taka funkcja wyznaczona jest przez dwa parametry: współczynnik *b*, przez który mnoży się wartość *x*, i wyraz wolny *a*.

Kryterium ma wtedy postać wyrażenia $K(a,b)=\sum_{i=1}^{n} (y_i-a-bx_i)^2$, w którym występują *ustalone*

wartości obu zmiennych oraz *zmienne* parametry *a* i *b*, problem zaś sprowadza się do tego, aby znaleźć takie wartości tych parametrów, przy których wyrażenie to, traktowane jako funkcja dwu zmiennych, osiąga minimum. W podręcznikach statystyki jako metodę rozwiązania *problemu regresji liniowej* podaje się zwykle obliczenie tzw. pochodnych cząstkowych i przyrównanie ich do zera.

Okazuje się jednak, że i bez tak potężnych środków znalezienie rozwiązania nie nastręcza trudności. Zauważmy najpierw, że funkcje $K(a,b)$ i $K'(a,b)=(1/n)K(a,b)$ („średni kwadrat odchylenia obserwowanych wartości y od wartości teoretycznych") przyjmują minimum w tym samym punkcie, można więc tę drugą funkcję wykorzystać jako miarę dopasowania funkcji liniowej do danych. Wyznaczenie minimum funkcji K' okazuje się banalnie łatwe po przedstawieniu $K'(a,b)$ jako sumy trzech składników jak niżej.

$$K'(a,b) = (bs_x - r_{xy}s_y)^2 + (M_y - bM_x - a)^2 + (1 - r_{xy}^2)s_y^2 \tag{16}$$

(rachunkowe sprawdzenie powyższej równości jest wprawdzie dość żmudne, ale nie wymaga znajomości rachunku różniczkowego). W trzecim składniku niewiadome a i b nie występują, natomiast pierwszy i drugi składnik są kwadratami pewnych wyrażeń, więc co najwyżej mogą być równe 0. W konsekwencji otrzymujemy wzór

$$\text{Min } K'(a,b) = (1 - r_{xy}^2)s_y^2 \tag{17}$$

zaś liczby, dla których funkcja K' przyjmuje wartość minimalną muszą spełniać układ równań

$$bs_x - r_{xy}s_y = 0, \quad M_y - a - bM_x = 0 \tag{18}$$

(w literaturze statystycznej równania te nazywają się *równaniami normalnymi*). W pierwszym równaniu występuje tylko jedna niewiadoma b. Rozwiązując je, otrzymujemy wzór

$$b_{yx} = r_{xy}\frac{s_y}{s_x} \tag{19*}$$

Współczynnik b_{yx} nazywamy *współczynnikiem regresji liniowej y względem x*. Po podstawieniu b_{yx} do drugiego równania (18), rozwiązujemy je ze względu na niewiadomą a, otrzymując wynik $a_{yx}=M_y - b_{yx}M_x$. Poszukiwana funkcja liniowa, najlepiej pasująca do danych dana jest zatem wzorem $\hat{y}=b_{yx}x+a_{yx}$. Po wstawieniu $M_y - b_{yx}M_x$ w miejsce a_{yx} i prostych przekształceniach otrzymujemy następujące równanie, zwane *równaniem regresji liniowej*

$$\hat{y} - M_y = b_{yx}(x - M_x) \tag{20*}$$

Jest to równanie prostej przechodzącej przez punkt o współrzędnych (M_x,M_y) i nachylonej do osi x pod kątem zależnym od b_{yx} (dokładniej, $b_{yx}=\text{tg }\beta$, gdzie β jest kątem nachylenia). Znak współczynnika regresji liniowej, który w związku ze wzorem (19) jest taki sam jak znak współczynnika korelacji, informuje o kierunku zależności. Wartość bezwzględna b_{yx} pokazuje, w jakim stosunku pozostaje odchylenie teoretycznej zmiennej zależnej od swojej średniej do odchylenia zmiennej niezależnej od swojej średniej.

W 1885 roku angielski uczony Franciszek Galton opublikował wyniki swoich badań nad zależnością wzrostu dzieci od wzrostu rodziców. Stwierdził, że wprawdzie wysocy ojcowie mają na ogół wysokich synów, a więc zależność jest dodatnia, ale w następnych pokoleniu ma miejsce „cofanie się ku przeciętności" (*regression towards mediocrity*), tzn. wzrost syna leży bliżej średniej w populacji synów niż wzrost jego ojca porównany ze średnią w populacji ojców. Formalnie prawidłowość ta oznacza, że $0<b_{yx}<1$ (sytuacja taka ma miejsce także w naszych danych płacowych, gdzie $b_{yx}=.50$). W taki oto sposób narodził się termin „regresja", który obecnie stosowany jest także, gdy $b_{yx}>1$, i oznacza statystyczną zależność funkcyjną.

Ze wzoru (19) wynika, że wartość współczynnika b_{yx} zależy zarówno od współczynnika korelacji r_{xy} jak i odchyleń standardowych s_x i s_y. Jeżeli zmienne x i y mierzone są w tych samych jednostkach (tak było u Galtona i w naszym przykładzie), wówczas po zmianie jednostki miary, co wymaga użycia tego samego przekształcenia liniowego (dopuszczalnego dla pomiaru interwałowego), współczynnik regresji liniowej nie zmieni wartości. Dla dwu zmiennych, z których każda jest mierzona w innych jednostkach, jak np. wzrost i waga, zmiana np. centymetrów na cale i kilogramów na funty będzie miała wpływ na wartość bezwzględną b_{yx}; po takiej zmianie prosta regresji może być mniej lub

bardziej „stromo” położona w stosunku do osi poziomej, aczkolwiek efekt wznoszenia się lub opadania, zależny tylko od znaku, zachowa się. Nachylenie prostej regresji nie musi mieć związku z „siłą” zależności liniowej poza przypadkiem $b_{yx}=0$ oznaczającym (o czym niżej) „brak zależności liniowej”.

Zapamiętajmy także, że inaczej niż współczynnik korelacji, współczynnik regresji liniowej jest parametrem niesymetrycznym: współczynnik $b_{xy}=r_{xy}(s_x/s_y)$ (*współczynnik regresji liniowej x względem y*) nie musi być równy b_{yx}. Oba współczynniki mają jednak ten sam znak (równy znakowi współczynnika korelacji) a ponadto powiązane są następującym wzorem (otrzymanym przez pomnożenie stronami dwu wariantów wzoru (19))

$$b_{yx}b_{xy} = r_{xy}^2 \tag{21}$$

z którego z kolei wynika, że wartość bezwzględną współczynnika korelacji liniowej można wyznaczyć ze wzoru $|r_{xy}| = \sqrt{b_{yx}b_{xy}}$.

Przekształcając wzór (19) przez wstawienie w miejsce r_{xy} wzoru definicyjnego (5), otrzymujemy równości

$$b_{yx} = \frac{Cov(x,y)}{s_x^2} = \frac{\sum_{i=1}^{n}(x_i-M_x)(y_i-M_y)}{\sum_{i=1}^{n}(x_i-M_x)^2} \tag{22}$$

Ostatnie wyrażenie może służyć jako wzór roboczy do obliczania b_{yx} (analogiczny wzór, mający w mianowniku $\sum(y-M_y)^2$, służy do obliczania b_{xy}).

•

Przedstawmy teraz empiryczną zmienną zależną y jako sumę zmiennej teoretycznej $\hat{y}=a_{yx}+b_{yx}x$ i odchylenia (błędu dopasowania) $y-\hat{y}$ i obliczmy parametry statystyczne dla obu składników takiego rozkładu. Łatwo sprawdzić, że $M_{\hat{y}}=M_y$, skąd wynika natychmiast, że $M_{y-\hat{y}}=0$. Obliczenie wariancji $\hat{y}$ daje następujący wynik $s_{\hat{y}}^2=b_{yx}^2s_x^2=r_{xy}^2s_y^2$. Wariancja zmiennej $y-\hat{y}$, z uwagi na to, że $M_{y-\hat{y}}=0$, to po prostu średnia arytmetyczna zmiennej $(y-\hat{y})^2$, równa $K'(a_{yx},b_{yx})$. Minimum $K'(a,b)$ przypada w punkcie (a_{yx},b_{yx}) i jak wiemy ze wzoru (17) jest równe $(1-r_{xy}^2)s_y^2$. W ten sposób udowodniliśmy następujące formuły

$$s_{\hat{y}}^2=r_{xy}^2s_y^2,\ \ s^2_{y-\hat{y}} = (1-r_{xy}^2)s_y^2 \tag{23}$$

Wariancja $y-\hat{y}$ nosi nazwę *wariancji resztowej*. Podane wyżej wzory implikują równość

$$s_y^2 = s_{\hat{y}}^2 + s^2_{y-\hat{y}} \tag{24*}$$

Jak wiadomo, wariancja sumy dwu zmiennych jest równa sumie ich wariancji wtedy i tylko wtedy, gdy zmienne te są nieskorelowane. Równość (24) implikuje zatem równość $Cov(\hat{y},y-\hat{y})=0$.

Wzór (24) interpretuje się jako rozkład *zmienności całkowitej y* na dwa składniki, z których pierwszy to *zmienność wyjaśniona* przez model zależności liniowej od zmiennej x, a drugi to „reszta”, czyli *zmienność niewyjaśniona* na gruncie przyjętego modelu. Ponieważ wariancja wyjaśniona jest równa $r_{xy}^2s_y^2$, kwadrat współczynnika korelacji liniowej informuje jaką część zmienności całkowitej zmiennej zależnej stanowi zmienność wyjaśniona przez liniowe oddziaływanie zmiennej niezależnej.

Wielkość $(1-r_{xy}^2)s_y^2$ jako równa $s^2_{y-\hat{y}}$ nie może być mniejsza od 0. W konsekwencji $1-r_{xy}^2\geq 0$, a stąd

$$0 \leq r_{xy}^2 \leq 1,\ \ -1 \leq r_{xy} \leq 1 \tag{25*}$$

Jeśli $r_{xy}^2=1$, czyli $r_{xy}=1$ lub -1), mówimy, że między zmiennymi x i y zachodzi *korelacja doskonała*. Wariancja resztowa $s^2_{y-\hat{y}}$ jest wtedy równa 0, to zaś oznacza, że $y-\hat{y}=0$, a stąd $\hat{y}_i=y_i$ dla każdego i. Na

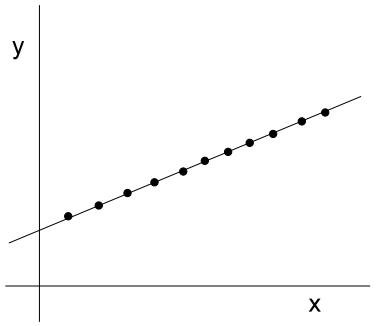
wykresie rozrzutu wszystkie punkty leżą wtedy na prostej regresji.

Jeśli $r_{xy}^2=0$, czyli $r_{xy}=0$, mówimy o całkowitym braku zależności *liniowej*. Mamy wówczas $b_{yx}=0$ i równanie regresji przybiera postać $\hat{y}=M_y$, wyznaczając prostą przebiegającą równolegle do osi *x* i przecinającą oś *y* w punkcie $(0,M_y)$. Na wykresie punkty mogą być rozrzucone chaotycznie wokół tej prostej, mogą jednak także układać się wokół pewnej krzywej. Brak zależności prostoliniowej nie wyklucza zatem innej formy zależności.

Podsumowaniem interpretacji Pearsonowskiego współczynnika korelacji liniowej w kontekście teorii regresji liniowej niech będzie zestawienie pięciu możliwych sytuacji.

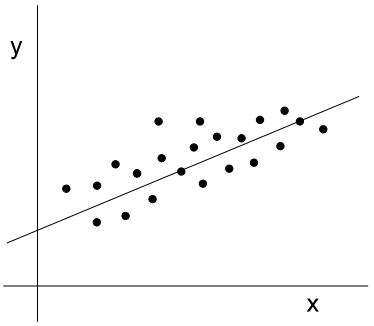
(1) Doskonała dodatnia zależność liniowa

$r_{xy} = 1$



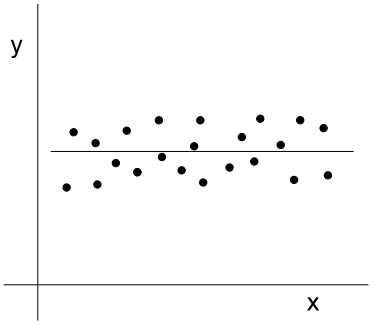
(2) Niedoskonała dodatnia zależność liniowa

$0 < r_{xy} < 1$



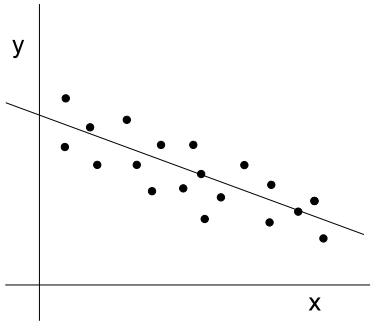
(3) Brak zależności liniowej

$r_{xy} = 0$



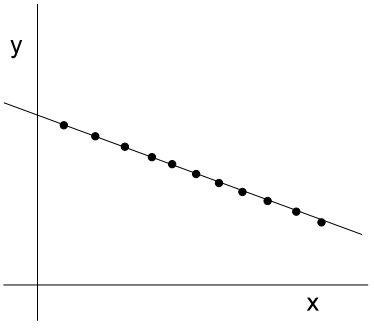
(4) Niedoskonała ujemna zależność liniowa

$-1 < r_{xy} < 0$



(5) Doskonała ujemna zależność liniowa

$r_{xy} = -1$



Na tym zakończę wykład poświęcony badaniu zależności *dwu* zmiennych porządkowych i interwałowych, świadomie ograniczony do zagadnień leżących w polu zainteresowania *statystyki opisowej*. Współczynniki r_{xy} i b_{yx} zwykle oblicza się z próby losowej, a otrzymane wartości traktuje

jako oceny nieznanych parametrów ρ_{XY} i β_{YX} w populacji generalnej (*X* i *Y* oznaczają tu zmienne losowe). Problemy wnioskowania statystycznego dotyczącego tych parametrów omówione zostaną na osobnym wykładzie.

Tekst ten powstał w roku akademickim 2000/2001 na użytek zajęć ze *Statystyki*, które prowadziłem wówczas na kierunku socjologia w WSBiP w Ostrowcu Św. Przygotowując wykład na temat badania zależności zmiennych porządkowych i interwałowych, korzystałem z ogólnie dostępnych podręczników oraz skryptu Grzegorza Lissowskiego, *Zależności statystyczne między uporządkowaniami* (Warszawa 1978: WSNS). W roku akademickim 2008/2009 starą notatkę wykorzystałem ponownie na zajęciach ze statystyki dla studentów socjologii Uniwersytetu Pedagogicznego w Krakowie. Tym razem zapis wykładu, wcześniej dostępny tylko w formie wydruku na papierze, umieściłem do wglądu publicznego na swojej stronie.

http://www.cyf-kr.edu.pl/~ussozans/

26/04/09