

Statystyka medyczna

Część 2

Przedziały ufności

Gdzie jest Krzyś?



Na 80% jest u siebie w domu na kanapie
(powiemy: na poziomie istotności 0.2)



Na 90% jest gdzieś w Lublinie
(powiemy: na poziomie istotności 0.1)



Na 99% jest gdzieś w Polsce
(powiemy: na poziomie istotności 0.01)

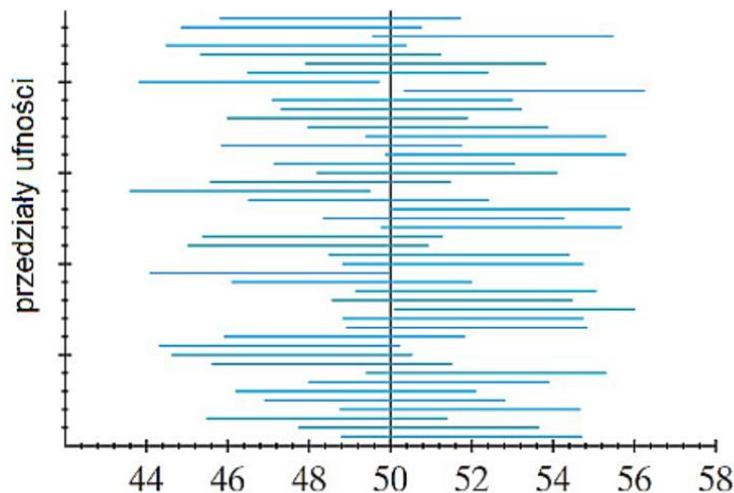
Przedział ufności

- “rozsądny” przedział, w którym dany parametr (np. średnia, odchylenie standardowe) może się znajdować -> estymacja przedziałowa
- w estymacji przedziałowej trzeba wybrać jeszcze pewien tzw. poziom ufności mierzony w %, tj. miarę pewności, że dany przedział zawiera estymowany parametr
- jeśli wybierzemy np. 95% poziom ufności, to będziemy mieli 95% pewności, że skonstruowany na podstawie próby losowej przedział zawiera estymowany parametr

Co znacza stwierdzenie, że

"przedziałem ufności dla średniej populacji μ jest $[a,b]$ na poziomie istotności α "

- prawdopodobieństwo, że tak wyliczony przedział ufności zawiera prawdziwą wartość średniej populacji, μ , wynosi $(1 - \alpha)$;
- gdyby brać wiele przedziałów ufności z próbek, to $(1 - \alpha) \cdot 100\%$ z nich zawierałoby średnią:



(a) 90% przedział ufności, σ^2 znana

Przedziałem ufności dla średniej μ populacji na poziomie istotności α jest

1

$\left[\bar{x} - z_{\alpha} \frac{\sigma}{\sqrt{n}}; \bar{x} + z_{\alpha} \frac{\sigma}{\sqrt{n}} \right]$ gdy odchylenie standardowe populacji σ jest znane

2

$\left[\bar{x} - z_{\alpha} \frac{s^*}{\sqrt{n}}; \bar{x} + z_{\alpha} \frac{s^*}{\sqrt{n}} \right]$ gdy odchylenie standardowe populacji σ nie jest znane
lecz próbka jest duża

3

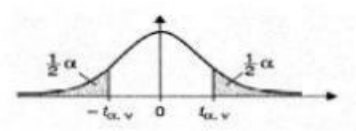
$\left[\bar{x} - t_{\alpha} \frac{s^*}{\sqrt{n}}; \bar{x} + t_{\alpha} \frac{s^*}{\sqrt{n}} \right]$ gdy odchylenie standardowe populacji σ nie jest znane
i próbka nie jest duża

$t(\alpha, n-1)$

$t(\alpha, n-1)$

Poziom ufności	α	z_{α}
90%	0.1	$\Phi^{-1}(0.95) = 1.645$
95%	0.05	$\Phi^{-1}(0.975) = 1.960$
99%	0.01	$\Phi^{-1}(0.995) = 2.576$

Tablica 3. Wartości $t_{\alpha, \nu}$ rozkładu t Studenta



$$P(|t| \geq t_{\alpha, \nu}) = \alpha$$

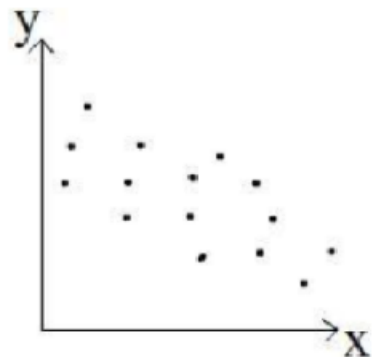
α ν	0,2	0,1	0,05	0,02	0,01	0,002	0,001
1	3,078	6,314	12,706	31,821	63,657	318,309	636,619
2	1,886	2,920	4,303	6,965	9,925	22,327	31,599
3	1,638	2,353	3,182	4,541	5,841	10,215	12,924
4	1,533	2,132	2,776	3,747	4,604	7,173	8,610
5	1,476	2,015	2,571	3,365	4,032	5,893	6,869
6	1,440	1,943	2,447	3,143	3,707	5,208	5,959
7	1,415	1,895	2,365	2,998	3,499	4,785	5,408
8	1,397	1,860	2,306	2,896	3,355	4,501	5,041
9	1,383	1,833	2,262	2,821	3,250	4,297	4,781
10	1,372	1,812	2,228	2,764	3,169	4,144	4,587
11	1,363	1,796	2,201	2,718	3,106	4,025	4,437
12	1,356	1,782	2,179	2,681	3,055	3,930	4,318
13	1,350	1,771	2,160	2,650	3,012	3,852	4,221
14	1,345	1,761	2,145	2,624	2,977	3,787	4,140
15	1,341	1,753	2,131	2,602	2,947	3,733	4,073
16	1,337	1,746	2,120	2,583	2,921	3,686	4,015
17	1,333	1,740	2,110	2,567	2,898	3,646	3,965
18	1,330	1,734	2,101	2,552	2,878	3,610	3,922
19	1,328	1,729	2,093	2,539	2,861	3,579	3,883

20	1,325	1,725	2,086	2,528	2,845	3,552	3,850
21	1,323	1,721	2,080	2,518	2,831	3,527	3,819
22	1,321	1,717	2,074	2,508	2,819	3,505	3,792
23	1,319	1,714	2,069	2,500	2,807	3,485	3,768
24	1,318	1,711	2,064	2,492	2,797	3,467	3,745
25	1,316	1,708	2,060	2,485	2,787	3,450	3,725
26	1,315	1,706	2,056	2,479	2,779	3,435	3,707
27	1,314	1,703	2,052	2,473	2,771	3,421	3,690
28	1,313	1,701	2,048	2,467	2,763	3,408	3,674
29	1,311	1,699	2,045	2,462	2,756	3,396	3,659
30	1,310	1,697	2,042	2,457	2,750	3,385	3,646
40	1,303	1,684	2,021	2,423	2,704	3,307	3,551
60	1,296	1,671	2,000	2,390	2,660	3,232	3,460
120	1,289	1,658	1,980	2,358	2,617	3,160	3,373

Współczynnik korelacji liniowej Pearsona

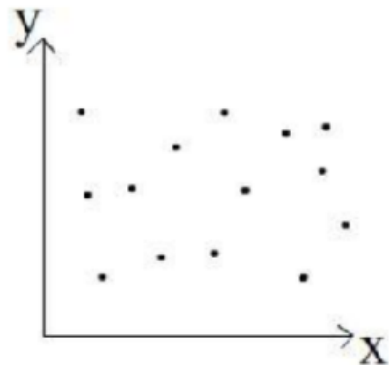
- Mając dwa **równoliczne** zbiory obserwacji dwóch różnych cech, często interesuje nas, czy istnieje jakaś zależność pomiędzy tymi dwiema cechami
- Np. mając masę ciała grupy ludzi i wysokość ich ciśnienia skurczowego, może nas interesować, czy wraz ze wzrostem masy ciała wzrasta również ciśnienie krwi (lub czy np. wyższej masie ciała odpowiada wyższy poziom cholesterolu)
- Używając współczynnika korelacji Pearsona badamy **zależność liniową** (ale przy pomocy innych technik można badać inne rodzaje zależności, np. kwadratowa, wykładnicza...)

$$r_{(X,Y)} = \frac{\sum_{i=1}^n (x_i - \bar{x}) (y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n (y_i - \bar{y})^2}}$$



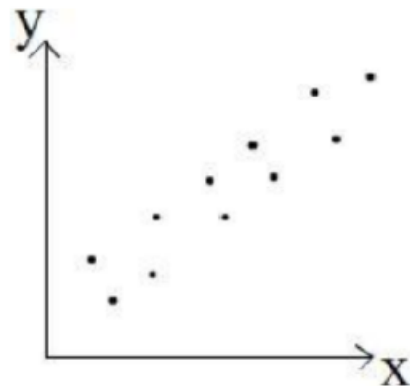
korelacja ujemna

$$r < 0$$



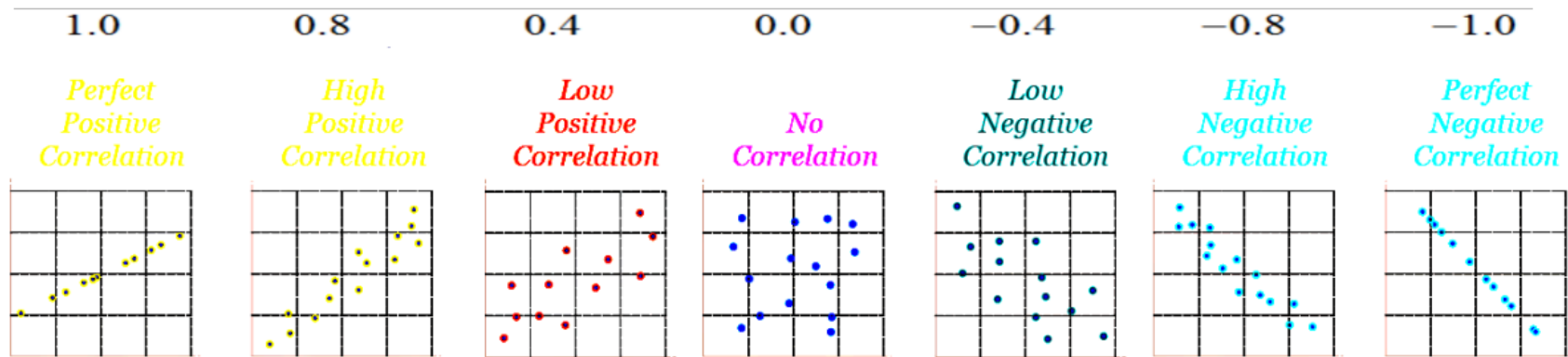
brak korelacji

$$r = 0$$



korelacja dodatnia

$$r > 0$$



r	Interpretacja
$r > 0,8$	Bardzo silna pozytywna korelacja
$0,4 < r < 0,8$	Silna korelacja dodatnia
$0,2 < r < 0,4$	Słaba korelacja dodatnia
$r < -0,8$	Bardzo silna negatywna korelacja
$-0,8 < r < -0,4$	Silna ujemna korelacja
$-0,4 < r < -0,2$	Słaba ujemna korelacja

gdy $r=0$ brak liniowej zależności między zmiennymi X i Y

Własności współczynnika korelacji liniowej Pearsona

- $-1 \leq r \leq 1$
- Jeśli $r > 0$ to większym wartościom jednej cechy odpowiadają (średnio) większe wartości drugiej cechy.
- Jeśli $r < 0$ to większym wartościom jednej cechy odpowiadają (średnio) mniejsze wartości drugiej cechy.
- obie zmienne na skali interwałowej lub ilorazowej, dane z obu zmiennych mają rozkłady normalne
- Obie zmienne mają ten sam status, tzn. X nie musi zależeć od Y i na odwrót
- Jest symetryczny względem zmiennych X i Y (tzn. $r(X,Y)=r(Y,X)$)
- Korelacja nie implikuje związku przyczynowego (choć może się zdarzyć).

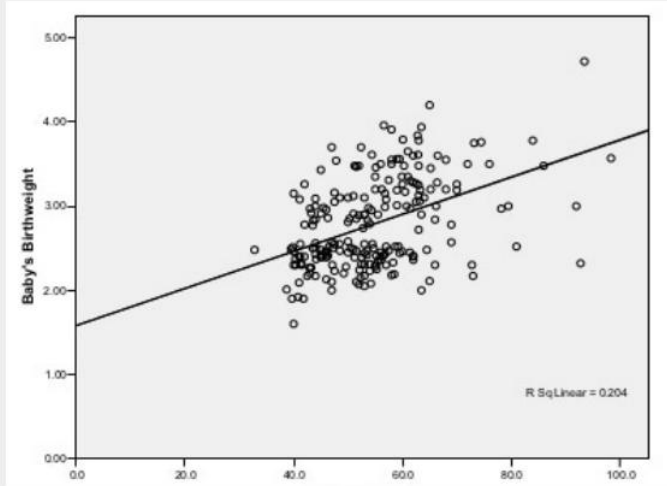


Sleeping with my
shoes always gives
me a headache.

Eh!?
Weren't you
drunk last
night?

Correlation does not imply causation

Współczynnik determinacji R^2



Wykres punktowy pokazuje związek między wagą matki (zmienna X) a wagą noworodka (zmienna Y)

Współczynnik korelacji wynosi $r = 0.452$.

Współczynnik determinacji wynosi $R^2 = 0.204$.

Oznacza to, że 20% zmienności wagi dziecka wyjaśnia waga matki.

Współczynnik determinacji R^2 wyjaśnia, ile procent zmienności jednej zmiennej tłumaczy się zmiennością drugiej zmiennej czyli wyjaśnia „moc” obliczonego współczynnika korelacji.

Współczynnik korelacji rang Spearmana r_s

Współczynnik korelacji Pearsona można obliczać tylko dla danych typu ilościowego, dlatego istnieją jeszcze inne współczynniki korelacji, które można obliczać dla danych typu porządkowego, np. współczynnik korelacji Spearmana.

Wykorzystywany do opisu siły korelacji dwóch cech, w przypadku, gdy :

- Cechy mają charakter jakościowy, pozwalający na uporządkowanie ze względu na siłę tej cechy
- Cechy mają charakter ilościowy, ale ich liczebność jest niewielka

Dla zmiennych X i Y na skali:

- nominalnej,
- porządkowej
- lub liczbowej o niskiej liczebności

Współczynnik korelacji rang Spearmana wynosi:

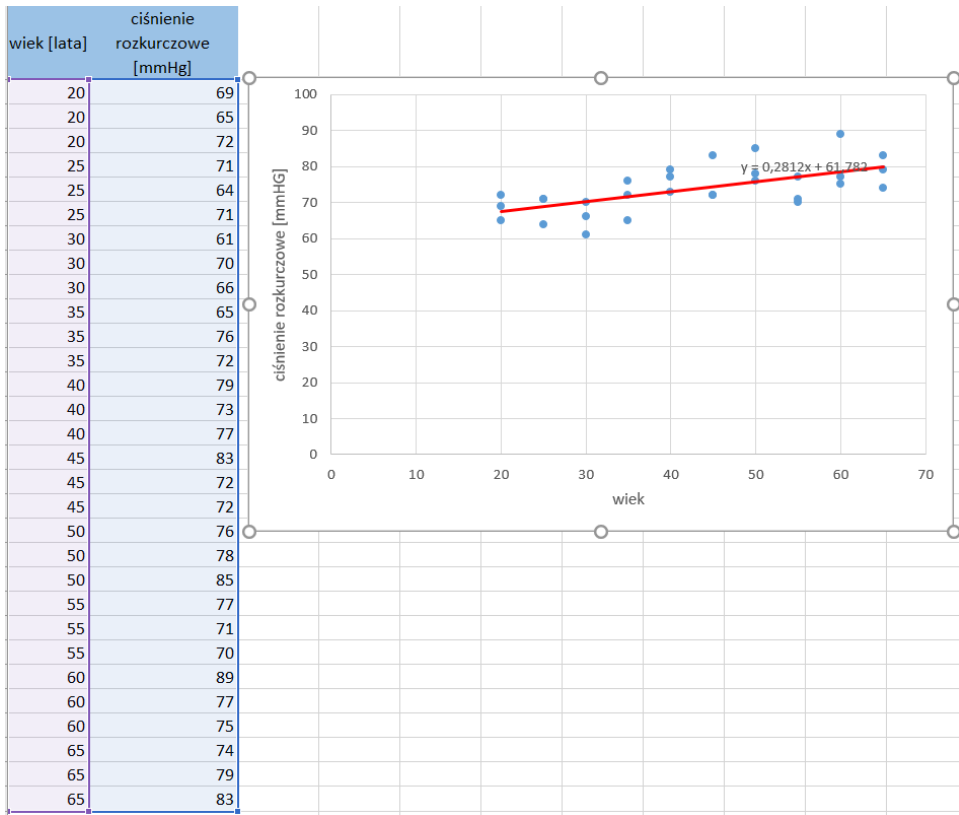
$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

gdzie d_i^2 – kwadraty różnic pomiędzy rangami odpowiadających sobie wartości cech x_i i y_i .

n – liczba par danych (wierszy w tabeli).

r_s	interpretacja
poniżej 0,2	korelacja słaba (praktycznie brak związku)
0,2-0,4	korelacja niska (zależność wyraźna)
0,4-0,6	korelacja umiarkowana (zależność istotna)
0,6-0,8	korelacja wysoka (zależność znaczna)
0,8-0,9	korelacja bardzo wysoka (zależność bardzo duża)
0,9-1,0	zależność praktycznie pełna

Prosta regresja liniowa

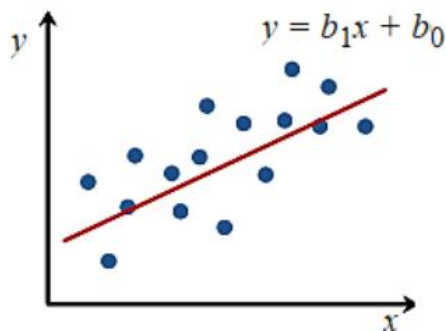


- Interesuje nas związek między X i Y
- Nad X mamy kontrolę i jej wartości mogą być z góry ustalone (np. wiek)
- Czy istnieje związek pomiędzy wiekiem osoby a jej rozkurczowym ciśnieniem krwi? Czy można ten związek opisać w postaci jakiegoś wzoru?

Prosta regresja liniowa - dopasowana linia lub krzywa regresji, która reprezentuje oszacowaną wartość oczekiwaną **zmiennej Y** przy konkretnych wartościach innej zmiennej lub **zmiennych X**.

$$y = b_1x + b_0$$

- Zmienne X i Y **nie są** równoprawne.
- Zmienna Y jest tradycyjnie nazywana zmienną **objaśnianą** lub **zależną**.
- Zmienne X nazywa się **zmiennymi objaśniającymi** lub **niezależnymi**.



$$y = b_1x + b_0$$

gdzie

$$b_1 = \frac{\sum (X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$$b_0 = \bar{Y} - b_1 \cdot \bar{X}$$

X_i, Y_i wartości zmiennych X i Y

$\bar{X}, \bar{Y}$ - średnie zmiennych X i Y

Współczynnik korelacji Pearsona

- Wykrywa liniową zależność między zmiennymi, wartość z zakresu $[-1, 1]$
- Jest symetryczny w odniesieniu do obu zmiennych
- Nie zakładamy, że jakakolwiek zmienna zależy od drugiej, czy jak
- Nie przewiduje przyszłości, nie określa rodzaju związku między zmiennymi

Regresja liniowa

- Wykrywa linię jak najbliższą danych punktów (obserwacji)
- Nie jest symetryczna w odniesieniu do obu zmiennych.
- Zmienną X nazywamy zmienną niezależną (objaśniającą), a Y zmienną zależną (objaśnianą)
- Możemy jej użyć do przewidzenia przyszłych wartości zmiennej Y na podstawie zmiennej X

Uwaga: oprócz regresji liniowej możemy znaleźć regresję wykładniczą, wielomianową, logarytmiczną itp.