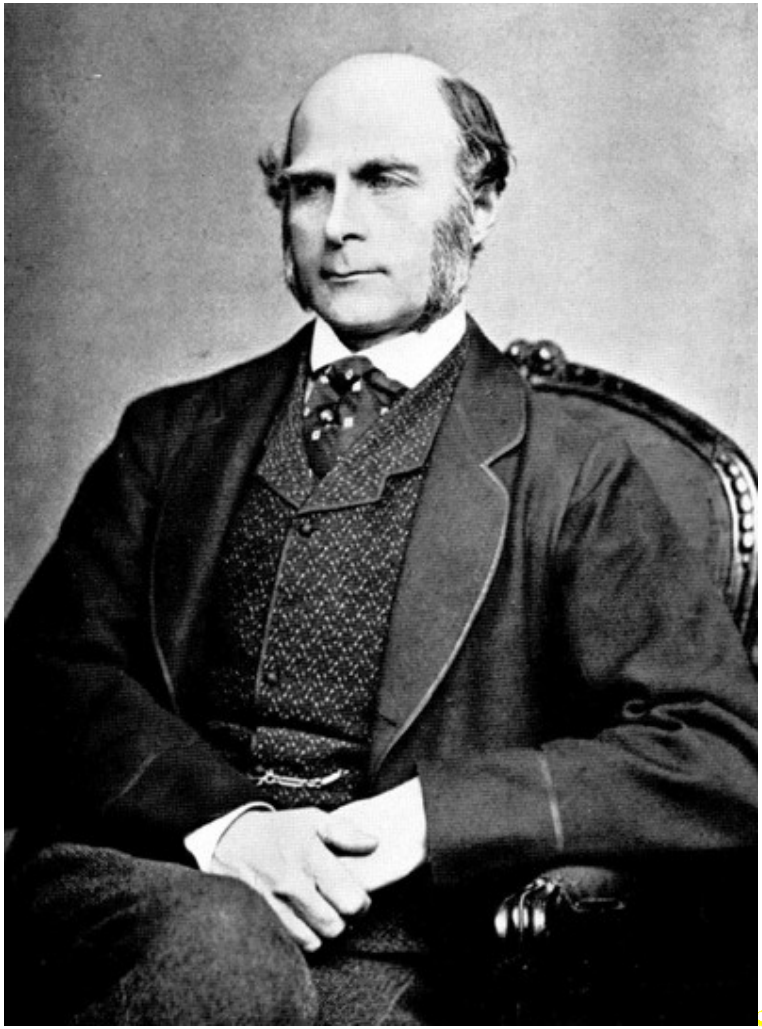
Korelační analýza

souvislost, souběžnost, příčinnost

Francis Galton (1822 – 1911)

Francis Galton (1822 – 1911)
polyhistor
psycholog, antropolog, meteorolog,
geograf, genetik, biolog,
kriminolog, psychometrik, statistik
'Hereditary Genius'
první mapa počasí
otisky prstů
dotazníky pro subjektivní názory
koncept korelace
regrese k průměru
bratranec Charlese Darwina

"Men who leave their mark on the world are very often those who, being gifted and full of nervous power, are at the same time haunted and driven by a dominant idea, and are therefore within a measurable distance of insanity"



Karl Pearson (1857 – 1936)

Karl Peason (1857 – 1936)
matematik
filozof
zakladatel matematické statistiky
zakladatel biometrie
'Gramatika vědy'
redaktor Biometriky
korelační koeficient
chí-kvadrát testy
momentová metoda

"The day must come when the biologist will, without being a mathematician, not hesitate to use mathematical analysis when he requires it."
Karl Pearson

Úlohy a otázky:

A) Souvisí spolu výskyt proměnné X a proměnné Y tak, že s vyššími hodnotami X se pojí vyšší hodnoty Y (a nižšími nižší), či naopak s vyššími hodnotami X se pojí nižší hodnoty Y (a s nižšími X vyšší Y)?

B) Můžeme v datech zjistit souběžnost resp. protiběžnost hodnot dvou číselných proměnných?

C) Je hodnota Y důsledkem hodnoty X? Reprezentuje proměnná X příčinu pro důsledek Y?

D) Jsou X a Y nositeli (částečně) stejné informace?

E) Vylučují se (resp. doplňují se) X a Y nebo naopak jedno předpokládá druhé?

Korelační analýza zkoumá vztah dvou číselných proměnných.

Statistika

- ***statistika*** **zkoumá variabilitu dat:**
  - ***popisuje ji***
  - ***vysvětluje ji***
  - ***predikuje ji***
- ***korelační analýza*** **zkoumá společnou variabilitu (kovariabilitu):**
  - ***popisuje ji***
  - ***používá ji pro vysvětlení***
  - ***používá ji pro predikci***

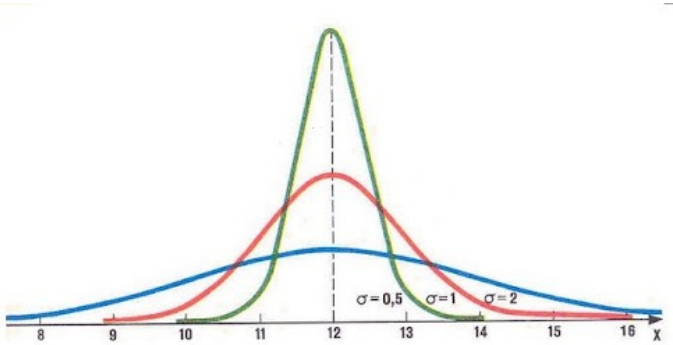
Variabilita

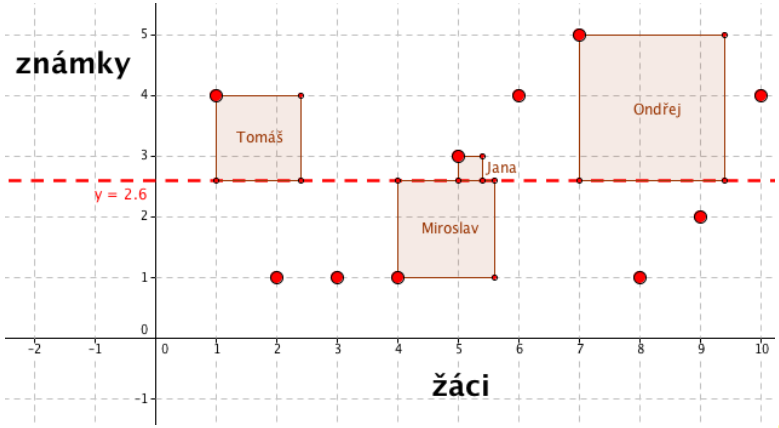
- **Variabilitu proměnné popisujeme rozptylem**

$$\text{var } X = \frac{\sum (X_i - \bar{X})^2}{N - 1}$$

$$s = \sqrt{\text{var } X}$$

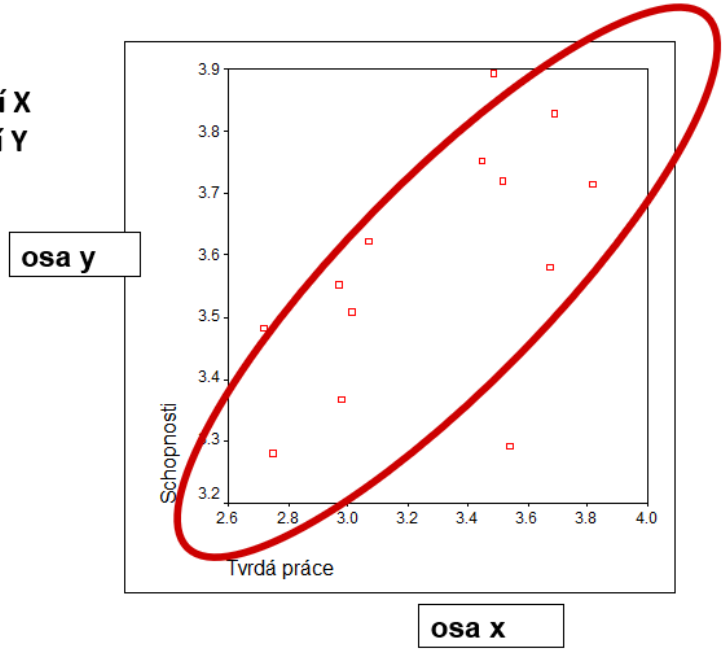
$$\text{var } X = \frac{\sum_{i \neq j} (X_i - X_j)^2}{2N(N-1)}$$





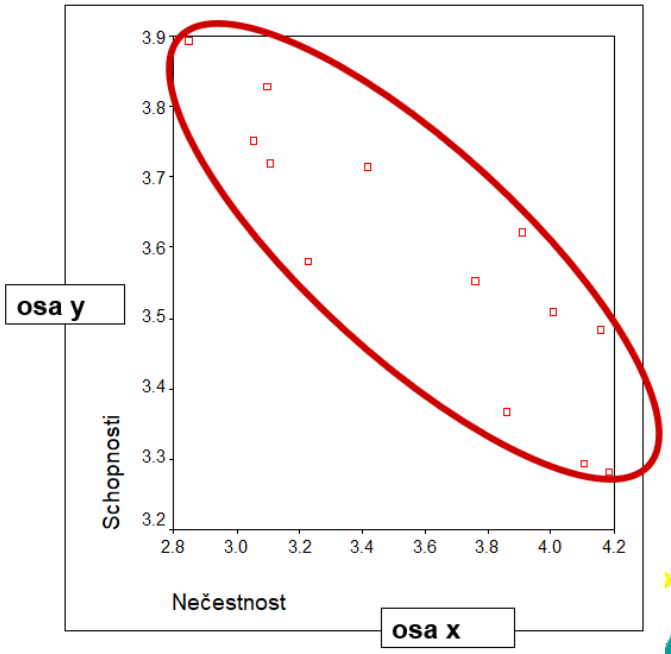
Souběh a protiběh variabilit

čím vyšší X
tím vyšší Y



čím vyšší X
tím nižší Y

čím nižší X
tím vyšší Y



Popis kovariability

- **kovariance: souběh variabilit dvou proměnných**

$$cov(X,Y) = \frac{1}{N-1}\sum (X_i - \bar{X}) * (Y_i - \bar{Y})$$

$$cov(X,Y) = \frac{1}{2N(N-1)}\sum_{i \neq j} (X_i - X_j) * (Y_i - Y_j)$$

Kovariance

- **Kovariance = Souběh variabilit dvou proměnných**
  - Statistická míra lineární závislosti dvou veličin
  - Je vyjádřena v jednotkách X a Y

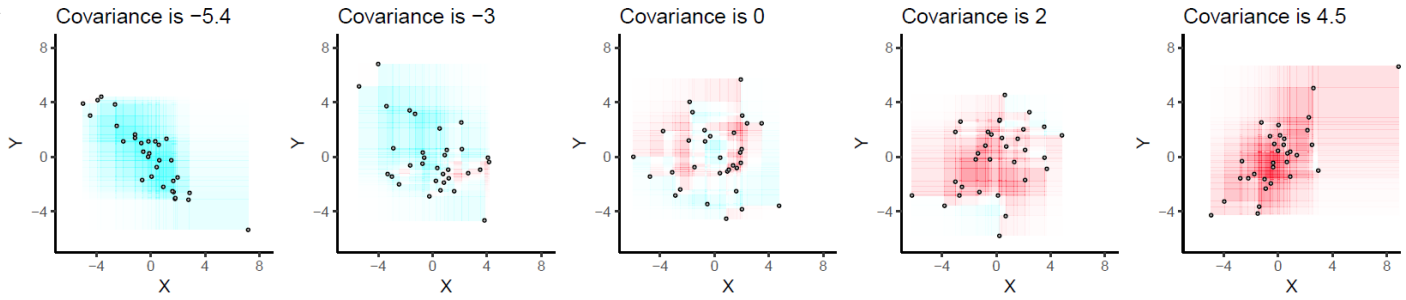
$$\operatorname{cov}(X,Y) = \mathrm{E}\left[(X - \bar{x})(Y - \bar{y})\right]$$

$$\operatorname{cov}(X,Y) = \mathrm{E}\left[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])\right]$$

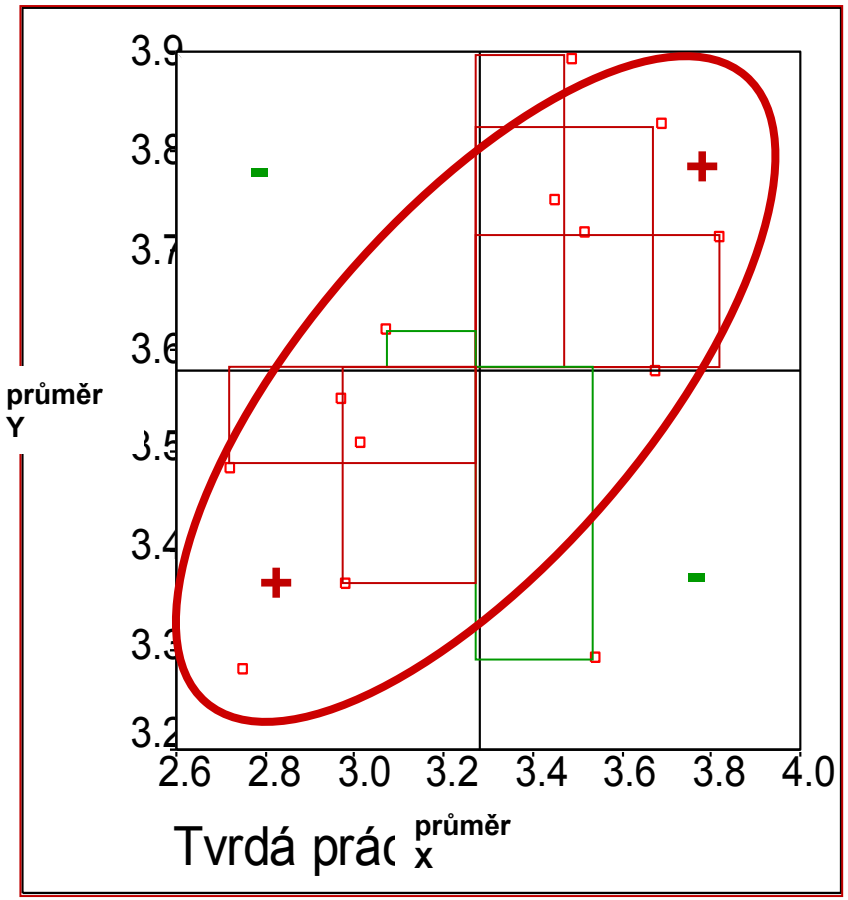
$$cov(X,Y) = \frac{1}{N-1}\sum (X_i - \bar{X}) * (Y_i - \bar{Y})$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

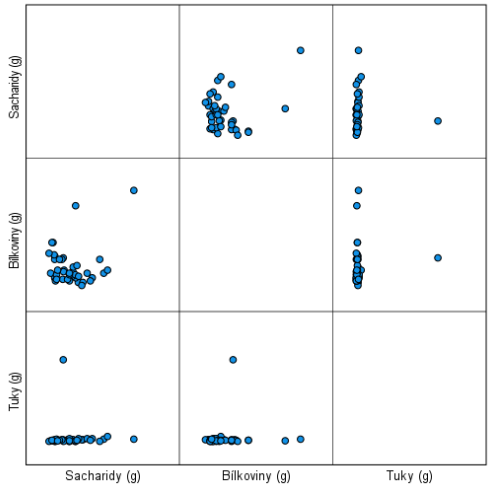
- cov(XY) > 0 -> souvislost mezi veličinami X a Y je pozitivní (čím větší X tím větší Y a naopak)
- cov(XY) < 0 -> souvislost mezi veličinami X a Y je negativní (čím větší X tím menší Y a naopak)
- nezávislé veličiny mají cov(XY) = 0, ale neplatí, že by cov(XY) = 0 znamenalo, že X a Y jsou nezávislé
- kovaria



Kovariance



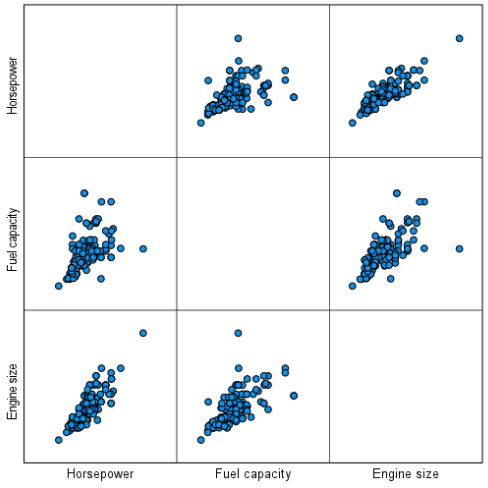
Kovariance



Covariance

	Sacharidy (g)	Bílkoviny (g)	Tuky (g)
Sacharidy (g)	44,093	1,773	-,526
Bílkoviny (g)	1,773	1,543	,217
Tuky (g)	-,526	,217	5,506

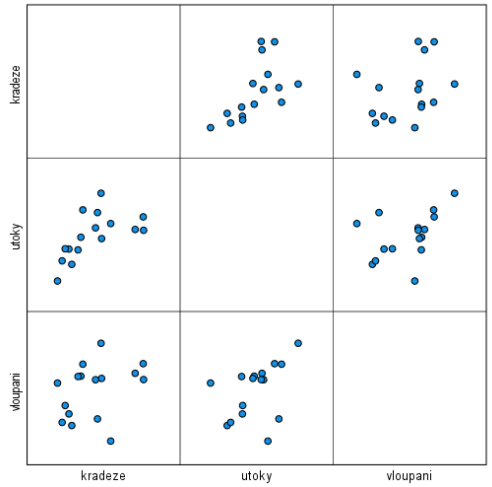
	Název	Sacharidy	Bílkoviny	Tuky
1	Ananas	10,9	,4	,4
2	Avokádo	5,3	2,0	15,3
3	Banán	20,4	1,0	,5
4	Broskev	9,9	,7	,1



Covariance

	Engine size	Horsepower	Fuel capacity
Engine size	1,091	49,599	2,693
Horsepower	49,599	3214,926	110,203
Fuel capacity	2,693	110,203	15,116

	model	engine_s	horsepow	fuel_cap
1	Integra	1,8	140	13,2
2	TL	3,2	225	17,2
3	CL	3,2	225	17,2
4	RL	3,5	210	18,0



Covariance

	kradeze	utoky	vloupani
kradeze	24718,533	8052,600	16315,633
utoky	8052,600	7131,667	10818,683
vloupani	16315,633	10818,683	93055,563

	mesto	kradeze	utoky	vloupani
1	Atlanta	106	147	1112
2	Boston	122	90	982
3	Chicago	340	242	808
4	Dallas	184	293	1668

Korelace – kovariance v poměru k rozptylům

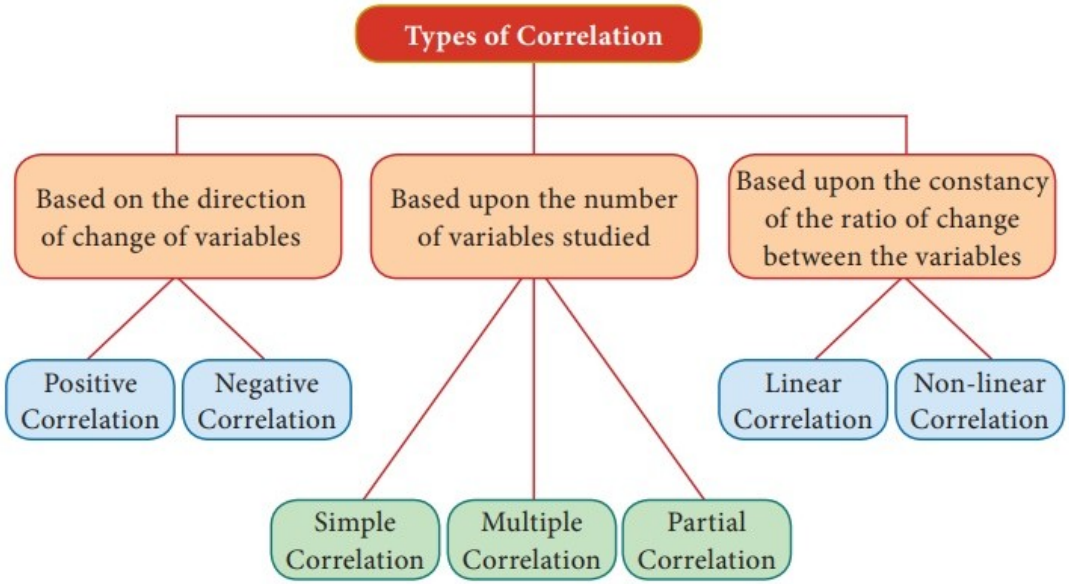
korelace **= vztah dvou proměnných**

= kovariance standardizovaná k rozptylům obou proměnných

= měří *vztah* dvou variabilit, nikoliv jejich velikost

Korelace - typy

- **Párová korelace**
- **Parciální korelace**
- **Mohonásobná korelace**



Párová korelace - Pearsonův lineární korelační koeficient

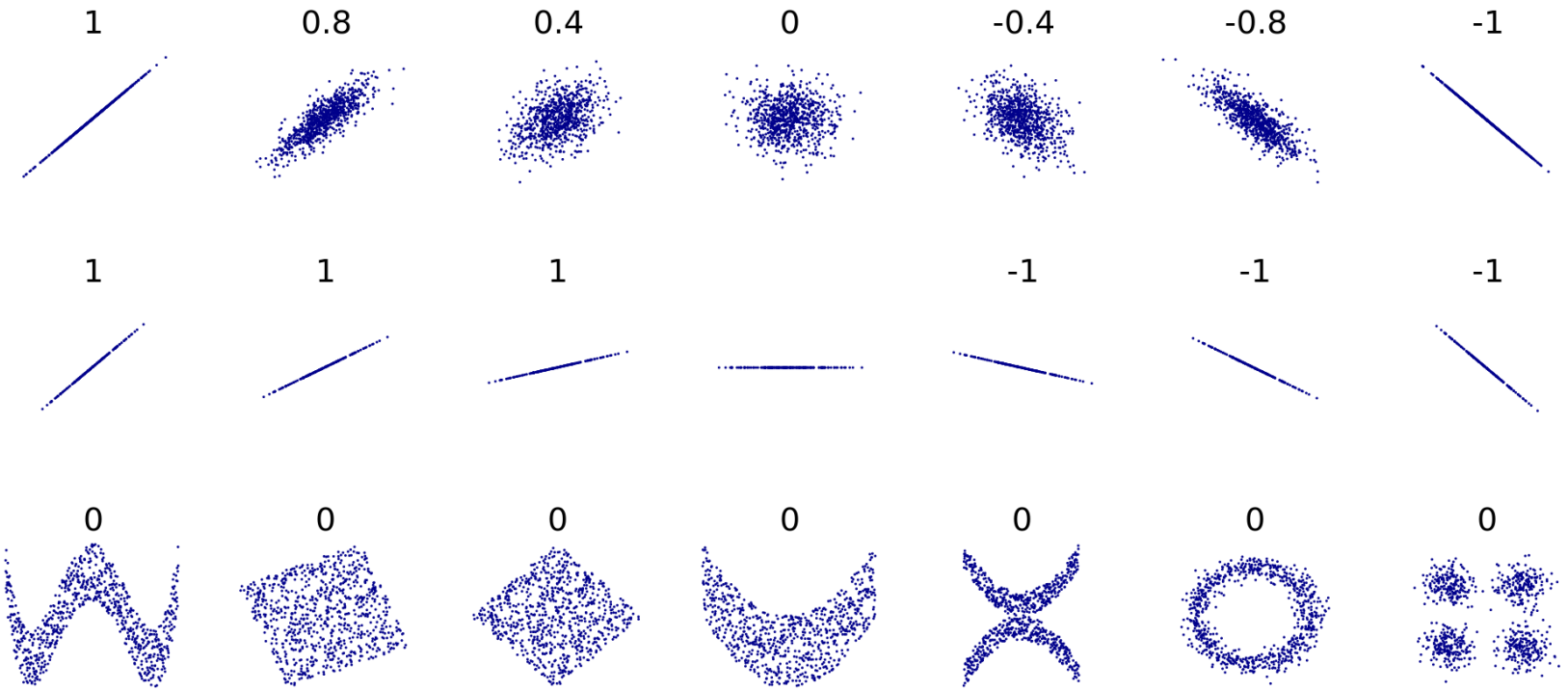
- **Korelace**
  - Měří vztah dvou proměnných
  - Jedná se o kovarianci standardizovanou k rozptylům obou proměnných

Vlastnosti:

- **r definován, pro n > 1**
- **r definován pro nenulové variability;**

  nesmí platit $s_x = 0$ nebo $s_y = 0$
- **r = 1, právě když body jsou seřazeny v nějaké přímce s nenulovým kladným spádem**
- **r = -1, právě když body jsou seřazeny v nějaké přímce s nenulovým záporným spádem**
- **čím více se r blíží k +1, tím více se body shlukují kolem stoupající přímky; čím více se r blíží k –1, tím více se body shlukují kolem klesající přímky**
- **jestliže v mraku bodů nelze vystopovat žádný *lineární* trend, r = 0**

Pearsonův lineární korelační koeficient



Pearsonův lineární korelační koeficient

Další vlastnosti:

- **r se nezmění, když se**

  - posune škála jedné nebo obou proměnných o libovolnou konstantu (změna počátku)

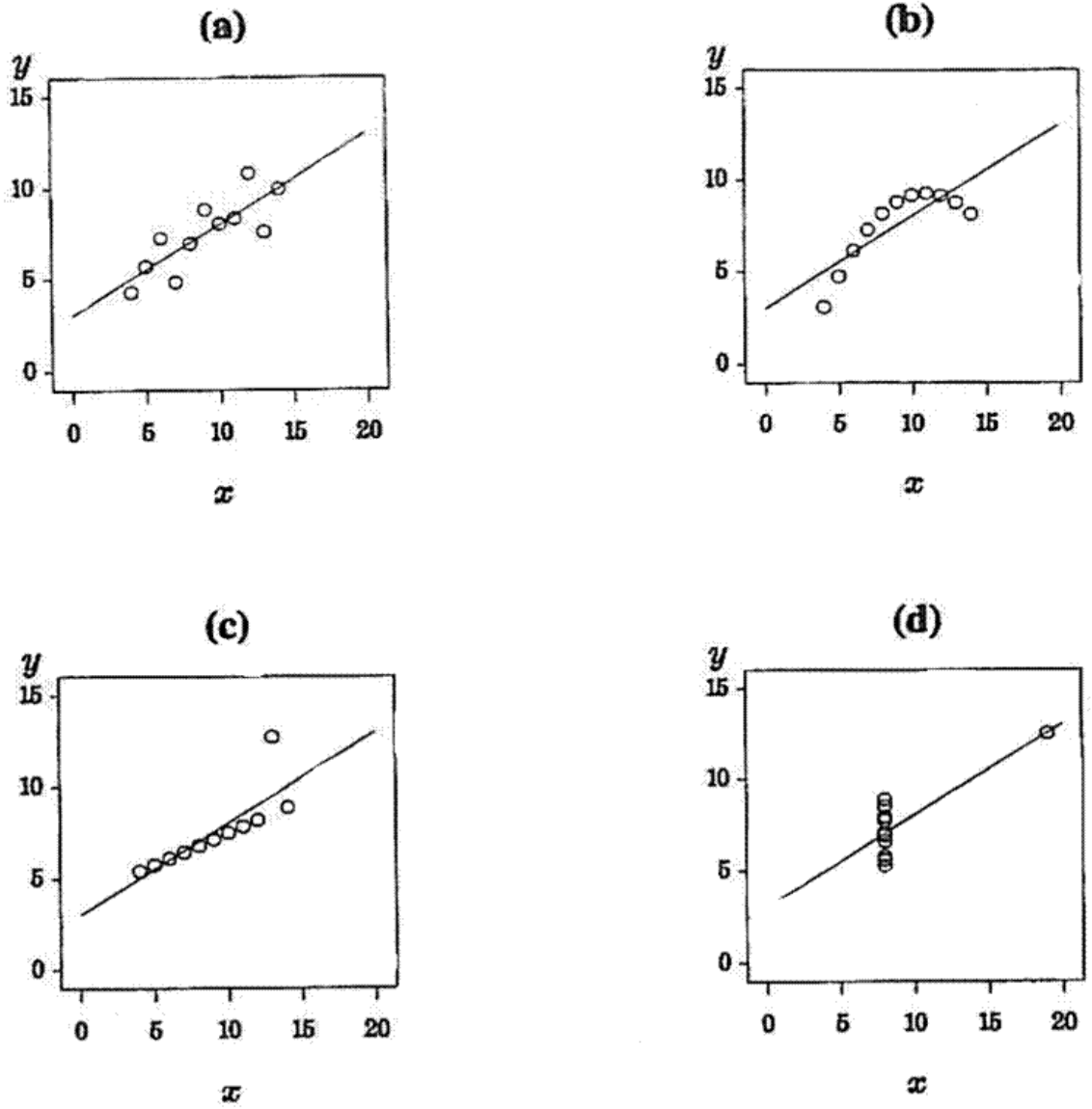
  - změní škála jedné nebo obou proměnných násobkem libovolnými činiteli (změna měřítka)

- **Nulové r**
  - mrak bodů tvoří pravidelný kruhový útvar
  - přímka, kolem které se shlukují body, je vodorovná nebo svislá
  - body leží symetricky kolem osy procházející průměrem X a to i když odpovídají úplné závislosti Y na X, ***např. $Y = (X - 4)^2$***
  - silné shlukování kolem rostoucí/klesající přímky je zkresleno bodem vzdáleným od mraku
  - kříží se kladný a záporný trend – překrytí dvou bodových mraků

Zkreslení koeficientu korelace

- **vzdálený bod – mrak bodů ukazuje na silnou/slabou korelaci, ale vzdálený bod ji uměle sníží/zvýší**
- **dvě skupiny nulové korelace umístěné v rovině vykazují vyšší korelaci**
- **číselné proměnné mají diskretní povahu (škála celých čísel od 1 do K) a přesné seskupení hodnot kolem přímky není plně možné**
- **jsou-li rozložení X a/nebo Y výrazně šikmá s dlouhým koncem**

Zkreslení koeficientu korelace



všechny situace mají stejný korelační koeficient = **0.816**

(Anscombe 1973)

r^2 – koeficient determinace

- **procento společné variability**
- **procento společné informace**

- **uvádíme i v procentech: $100*r^2$ %**

- **koeficient *indeterminace*: $1 - r^2$ (též i v %)**

- **hranice pro zabarvení matice (libovolné, ale užitečné):**
  R= .9, .7, .5, .3,
  odpovídá zhruba determinacím:
  R^2 = 80%, 50%, 25% a 10%

Poučky o velikosti koeficientů

Hodnota korelace v absolutní hodnotě	Interpretace souvislosti
0,01 – 0,09	**triviální, žádná**
0,10 – 0,29	**nízká až střední**
0,30 – 0,49	střední až podstatná
0,50 – 0,69	podstatná až velmi silná
0,70 – 0,89	velmi silná
0,90 – 0,99	téměř perfektní

De Vaus: 2002

Další míry korelace

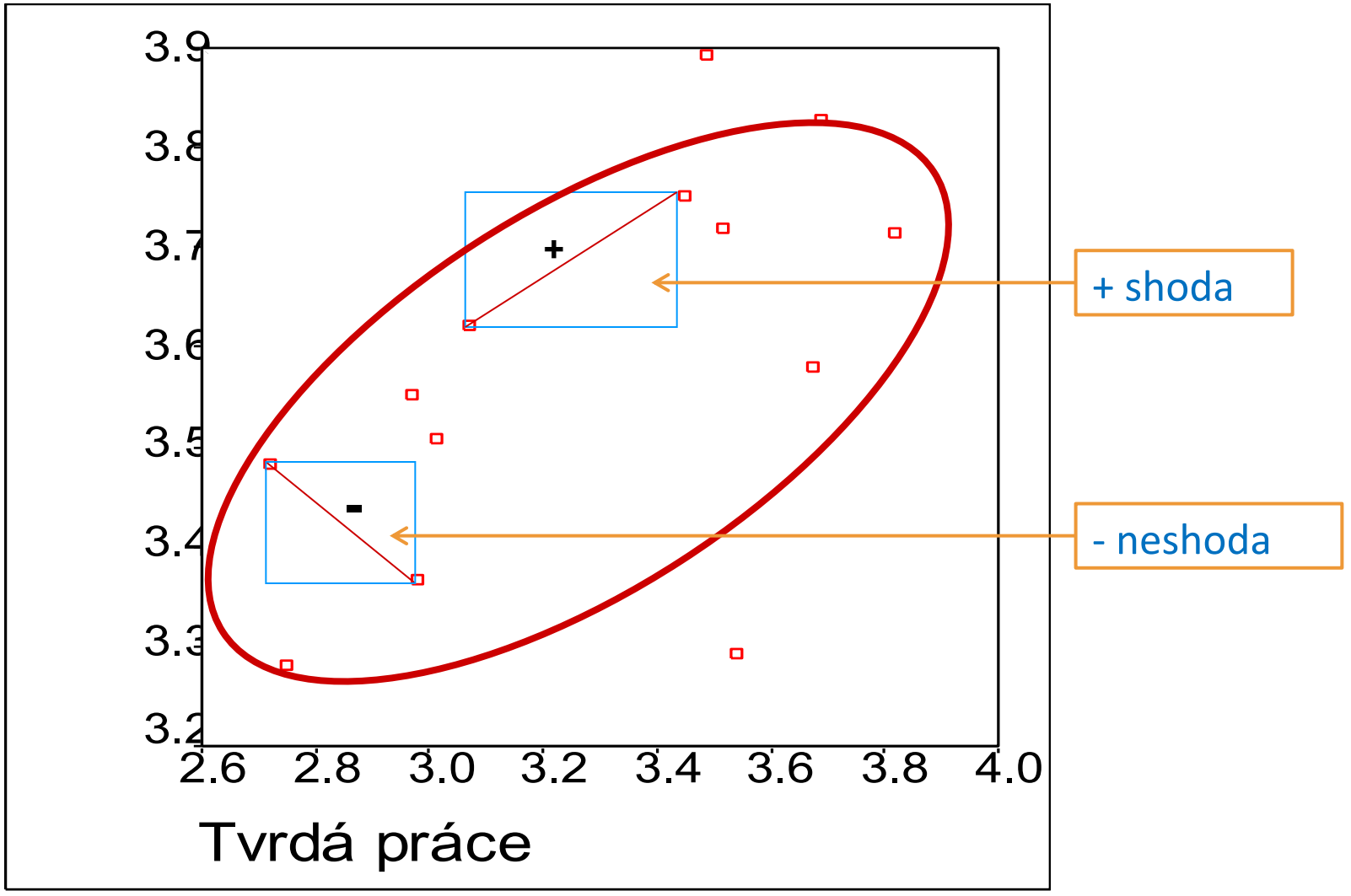
- jiná data (pořadí)
- šikmá rozložení
- vzdálená pozorování
- zvyklosti oboru
- komparace s jinými výstupy

Pořadová korelace – Spearmanovo r

- **Spearmanův koeficient pořadové korelace ρ vznikne tak, že se do vzorečku pro Pearsonův lineární korelační koeficient dosadí místo hodnot X a Y jejich pořadí v řadě**
- **též lze počítat přímo z pořadí; vychází ze vzdálenosti/nepodobnosti pořadí**

- **ρ = 1, pokud jsou řady zcela shodné**
- **ρ = -1, pokud jsou řady zcela protichůdné**
- **ρ = 0, pokud mezi řadami není žádná tendence ke shodě či protichůdnosti, ale pořadí jsou k sobě zcela náhodně**

Pořadová korelace – Kendallovo τ

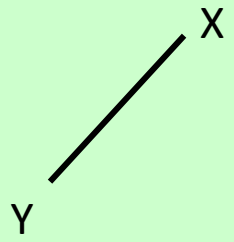


Pořadová korelace Kendallovo t

τ = (počet shod – počet neshod)/(počet shod + počet neshod)

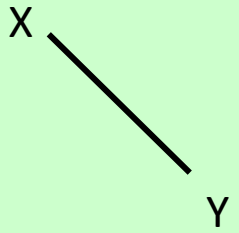
shoda

bodu X odpovídá vyšší hodnota u obou proměnných než má Y



neshoda

bod X má hodnotu u jedné proměnné vyšší než Y a u druhé nižší



- **dvojice, které mají stejné hodnoty jedné nebo obou proměnných se nepočítají ani do shod ani do neshod**

- **τ = 1, pokud jsou řady zcela shodné**
- **τ = -1, pokud jsou řady zcela protichůdné**
- **τ = 0, pokud mezi řadami není žádná tendence ke shodě či protichůdnosti, ale pořadí jsou k sobě zcela náhodně**

Pozn.: pro menší soubory je nutno spočítat signifikance přesně

Testy hypotéz

- **u statistických řad, jejichž souběžnost zjišťujeme, a které reprezentují obecnější situaci, výsledky procesů či širší základní soubor je podstatné vědět, zda o korelovanosti vůbec můžeme mluvit**

– **základní otázka: můžeme považovat dvě řady za korelované, nebo koeficient korelace zachycuje pouze náhodně vzniklé souladnosti v řadách?**

– **tedy: Je možné, že souběh/protiběh řad reprezentuje nenáhodný vztah, nebo mohl vzniknout jen působením náhody?**

Testy hypotéz

- **u korelačních koeficientů je základní dvojicí hypotéz, které testujeme:**

> ***H_0: korelační koeficient je nulový***
> ***H_A: korelační koeficient je nenulový***

- **prokazujeme, že naše spočtená míra je signifikantně nenulová, tedy, že v datech se projevuje nějaký nenáhodně vzniklý vztah (hypotéza H_A)**
- **platí pro Pearsonův, Spearmanův, Kendallův i Blomquistův koeficient – nenulová signifikance ukazuje na vlastnosti charakterizované koeficientem (linearní trend, monotonní trend, diagonální rozmístění dat)**

Testy hypotéz H_0: r = 0

t – test:

- **Studentovo t-rozložení – tabulky nebo výpočet dosažené signifikance**
- **dosadíme-li za t kritické hodnoty, dostaneme kritické hodnoty pro *r* přímo**

Testy hypotéz H_0: r = 0

z – test:

- **má normální rozdělení, tj. kritickou hodnotu pro alfa = 0.05 je 1.96**
- **ve skutečnosti je test vychýlen**

Význam signifikance

- **signifikance znamená pouze *nenulovost a nenáhodnost koeficientu*, rozhodnutí o tom, zda je hodnota zajímavá *provádí analytik***
- **signifikantní *neznamená* tedy ještě, že hodnota koeficientu, tj. *síla vztahu je dostatečná* na to, abychom ji považovali za interpretačně zajímavou. tj. za věcně interpretovatelnou**
- **i nízké hodnoty koeficientu korelace mohou být zajímavé**
  - **a) ukazují na trend, který se *začíná objevovat* a prosazovat, ukazují na nové procesy**
  - **b) ukazují trendy, které jsou přehlušeny *velkými šumy*, ale existují**
  - **c) naznačují zprostředkovanou vazbu**
- ***POZOR: test signifikance vychází pouze z hodnoty r ne z celkové struktury bodů: proto signifikance může být způsobena odlehlými pozorováními***

Význam nesignifikance

signifikance ukazuje na působení nenáhodných faktorů

x

nesignifikance (tj. nepřijatelně vysoké riziko přijetí korelovanosti proměnných) může znamenat:

- **správný závěr: korelace je nulová nebo nepatrná**
- **chybu 2. druhu (vztah existuje, ale neprokázali jsme ho):**
  - **korelace není prokázána vzhledem k *malému počtu případů/pozorování* (nemáme dost statistické informace k prokázání existujícího vztahu)**
  - **trend v datech je zahlušen *velkými chybami měření***
  - **trend je rušen *odlehlými pozorováními***
  - **trend je rušen *pozorováními, která do zkoumaného vztahu nepatří***
  - **trendy jsou v datech dva ve dvou podsouborech a *kříží a ruší se navzájem***

Korelace a směr závislosti

nezaměstnanost **volební přízeň**

schopnost **tvrdá práce**

počet čápů na 1000 obyv. **počet dětí na 1000 obyvatel**

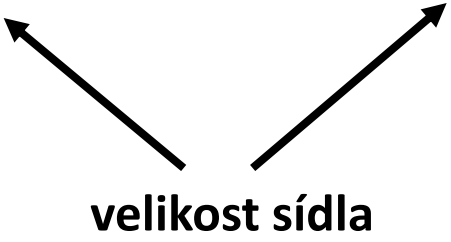
?

Parciální korelace – společná příčina

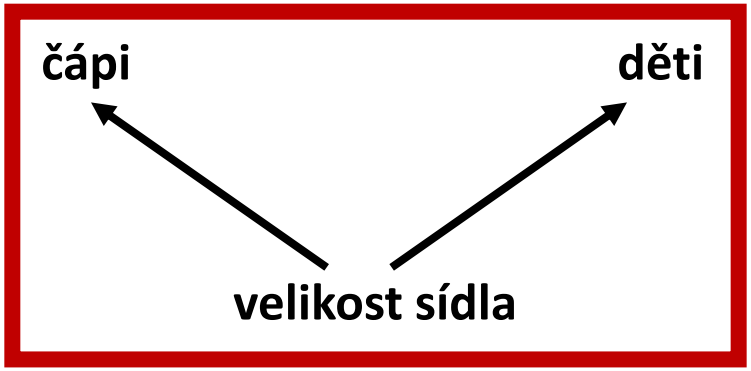
čápi ⟶ děti r = .56

čápi ⟷ děti r = .56

čápi ⟵ - - ⟶ děti r = .56



r(v,č) = .81 r(v,d) = .69

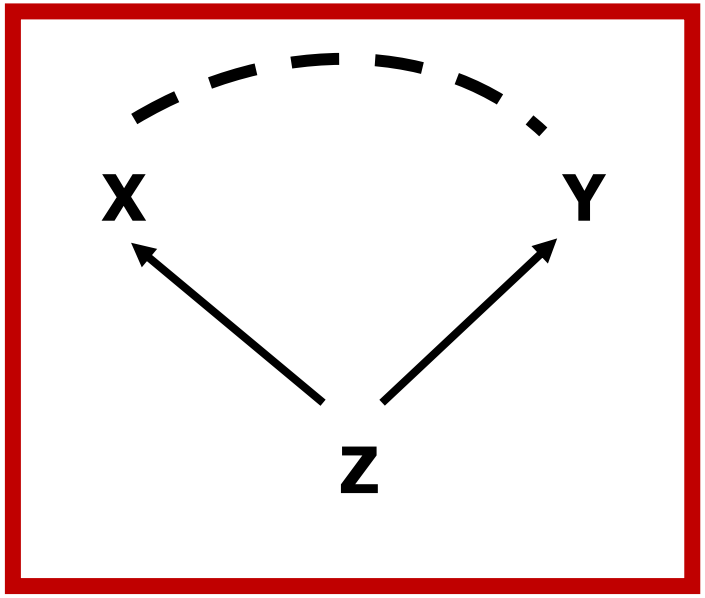


r(č,d/v) = 0.0

r(v,č) = .81 r(v,d) = .69

Schéma nepravé korelace

- **příčina Z ovlivňuje hodnoty proměnných X a Y**



- **otázka: jaká je korelovanost X a Y po očištění od vlivu Z?**
- **úloha: jak změřit tuto korelovanost a vliv Z na vztah X a Y ?**

Parciální korelace

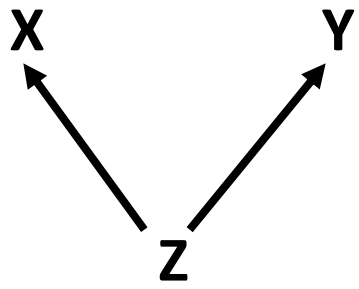
- **X, Y, Z (tři známé proměnné v datech, jejichž vzájemné korelace jsou známy):**

- **je-li parciální korelační koeficient roven nebo blízko k nule (nesignifikantní), znamená to, že proměnná Z plně vysvětluje korelaci mezi X a Y**

- **pokud se parciální koeficient jen podstatně redukuje, znamená to, že Z ovlivňuje vztah X a Y , ale není samo**

Parciální korelace

Model 1: Společná příčina X, Y, Z:



Z = společná příčina

$r(X,Y) \neq 0$, $r(X,Z) \neq 0$, $r(Y,Z) \neq 0$,

$r(X,Y/Z) = 0$

Model 2: Zprostředkující vlastnost

X ⟶ Z ⟶ Y

$r(X,Y) \neq 0$, $r(X,Z) \neq 0$, $r(Y,Z) \neq 0$,

$r(X,Y/Z) = 0$

Modely 1 a 2 se nedají statisticky odlišit

Korelační analýza

- **korelační analýza je první stupeň analýzy vztahů, po ní následují:**
  - ***regresní analýza*** **– tvar (model) orientovaných závislostí,**
  - ***faktorová analýza*** **– přehledná struktura vztahů a hledání společných příčin mnoha proměnných,**
  - ***analýza kovariančních struktur a kauzálních sítí*** **– kauzální modelování**
  - ***škálování*** **– grafické zobrazování vztahů, analýza pořadových a nelineárních koeficientů**
  - **speciální analýzy …**