Jak rozumět statistikám a posuzovat rizika

Jungmannova
národní
akademie

Následuj ty, kteří hledají pravdu, ale utíkej od těch, kteří ji našli.

Charles Spencer Chaplin

Co je a co není věda

- vědeckou teorii nelze dokázat (verifikovat), ale pouze vyvrátit (falsifikovat)
- pouze takové tvrzení, které lze testovat a tedy případně i vyvrátit, je vědecké
- vyvrácení teorie otevírá prostor pro další pokrok

sir Karl Raimund Popper

Pravděpodobnost a matematická statistika

- pravděpodobnost „se dívá dopředu“
- statistika „se dívá dozadu“
- statistika je aplikovaná pravděpodobnost
- odhadujeme pravděpodobnost platnosti hypotézy za podmínky, že nastal výsledek testu

Neurčitost

- systémy s mnoha vzájemně propojenými proměnnými nazýváme dynamické
- neurčitost je nedílnou součástí dynamických systémů, malá změna jednoho stavu vede v krátké době k naprosto odlišnému výsledku
- mikrosvět se chová stochasticky, ale to běžně nepozorujeme, tuto variabilitu zakrývá zákon velkých čísel

Podle čeho se rozhodovat

- odhad pravděpodobností je pouze první krok
- rozhodovací matice, v jednodušších případech vektor
- důležitý je i zisk/ztráta v případě rozhodnutí
- tam, kde jde o velké peníze je extrémně obtížné uspět

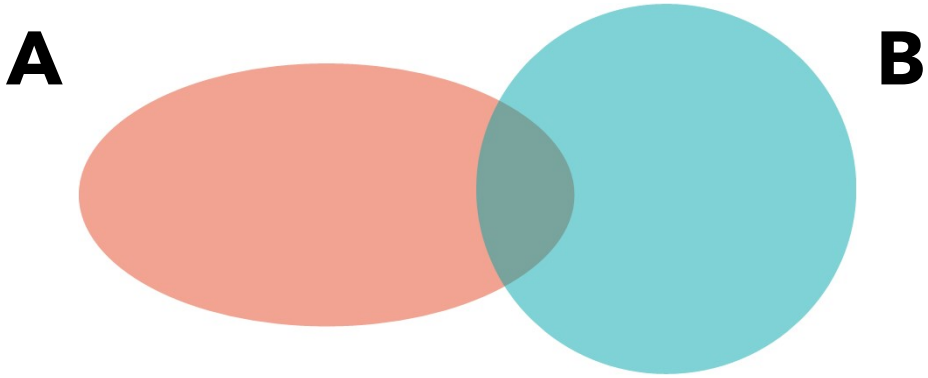
Pravděpodobnost

Kolmogorovova definice 1933

- mějme prostor $\Omega \neq \emptyset$ a systém podmnožin (σ-algebru) $\boldsymbol{S}$, na kterém je definovaná funkce P tak, že
  $\forall\, A \epsilon \boldsymbol{S}$ platí $P(A) \geq 0$;
  $P(\Omega) = 1$;
  $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ pro $A_1, A_2, \ldots \in \boldsymbol{S}$ tak, že $A_i \cap A_j = \emptyset$ pro $i \neq j$
  (σ-algebra je systém podmnožin uzavřený na základní množinové operace tj. průnik, sjednocení a doplněk množin)
- pak platí i $P(\emptyset) = 0$, $0 \leq P(A) \leq 1$, $P(\bar{A}) = 1 - P(A)$ a třeba i $P(A \cup B) + P(A \cap B) = P(A) + P(B)$

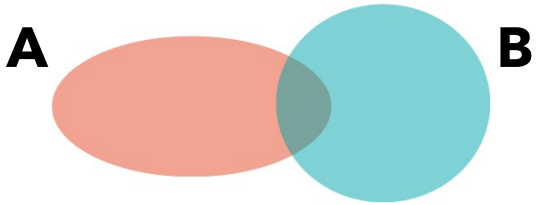
Podmíněná pravděpodobnost

- $P(A|B) = \frac{P(A \cap B)}{P(B)}, P(B) > 0$
- $P(A \cap B) = P(A|B) \times P(B) = P(B|A) \times P(A)$



Bayesova věta

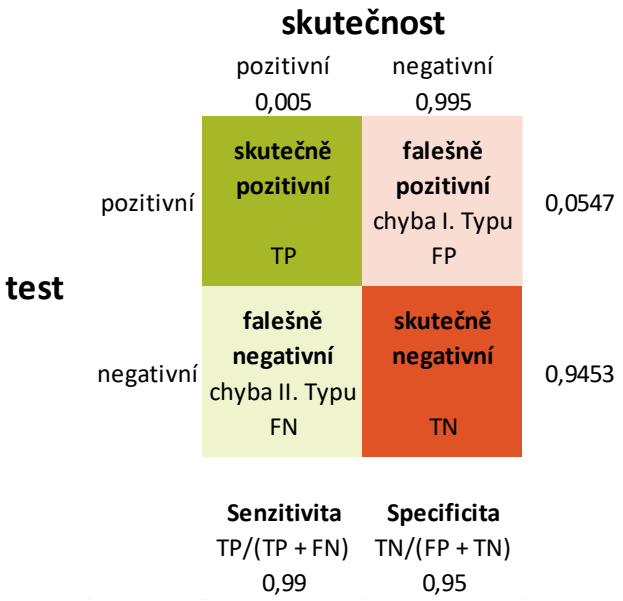
- pokud $P(B) > 0$, tak $P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$



varianta 2

- protože platí $P(B) = P(B \cap A) + P(B \cap \bar{A}) =$
  $= P(B|A) \times P(A) + P(B|\bar{A}) \times P(\bar{A})$
- tak dostáváme $P(A|B) = \frac{P(B|A) \times P(A)}{P(B|A) \times P(A) + P(B|\bar{A}) \times P(\bar{A})}$

Testy, senzitivita, specificita



- uživatelé drogy (TP + FN) = 0,005
- senzitivita 0,99
- specificita 0,95
- neuživatelé drogy (FP + TN) = 0,995 %
- pravděpodobnost pozitivního testu (TP + FP) = 0,99×0,005 + (1-0,95)×0,995 = = 0,0547
- mezi těmi, které test označil je uživatelů drogy TP = 0,99×0,005/0,0547 = 0,0905
- mezi negativně testovanými je uživatelů pouze FN = 0,0000529

Problém 3 dveří (Monty Hall)

- TV show ‚Let's Make a Deal' dávala finalistovi šanci vyhrát automobil, nebo kozu. Soutěžící si volil jedny ze 3 dveří a poté moderátor otevřel jedny z těch dvou zbývajících a ukázal obecenstvu kozu
- soutěžící pak se mohl rozhodnout, zda zůstane u původní volby, nebo ji změní
- jak byste se rozhodli vy a proč?

Řešení

	Dveře A	Dveře B	Dveře C
varianta 1			
varianta 2			
varianta 3			

Data

Kvalitativní

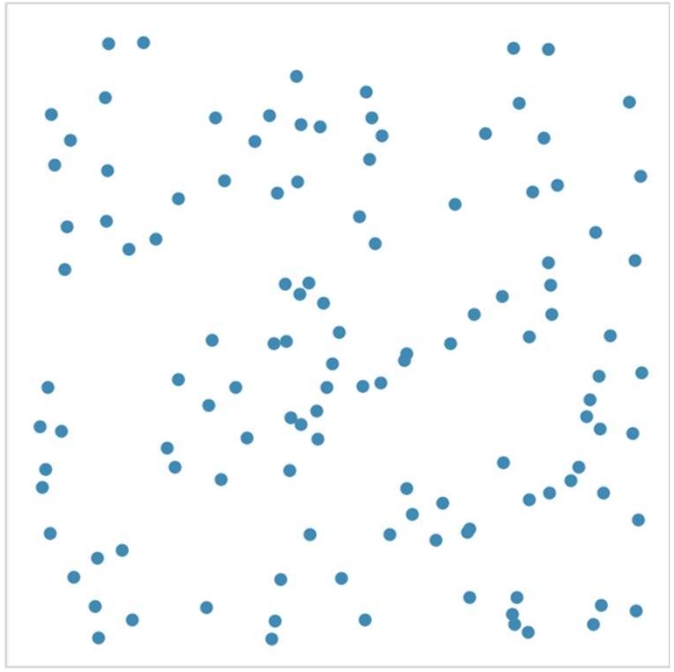
- binární – ano/ne, muž/žena, zná/nezná značku XY
- nominální – kraje, značky automobilů, krevní skupiny
- ordinální – lze porovnávat a řadit – stupeň vzdělání

Kvantitativní

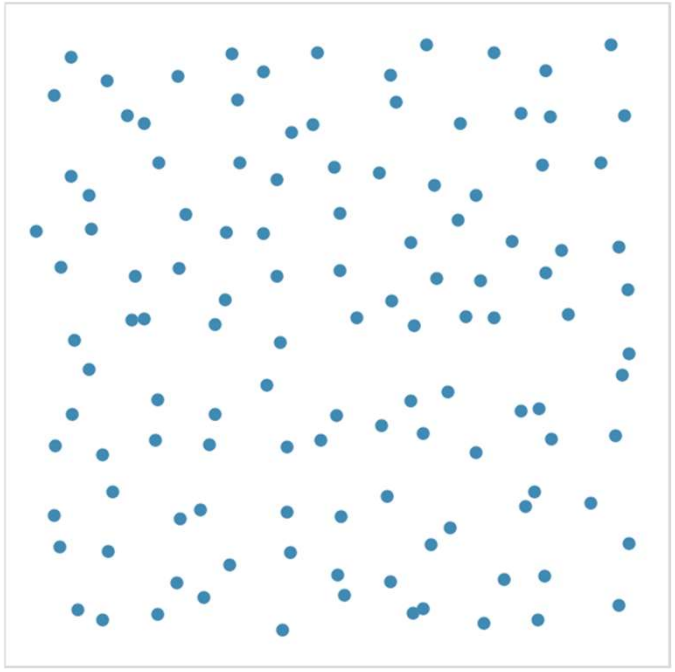
- diskrétní – přirozená čísla – počet dětí, velikost populace, ročník
- spojitá (intervalová) – výška, váha, teplota, přesný věk

Náhodná data

A



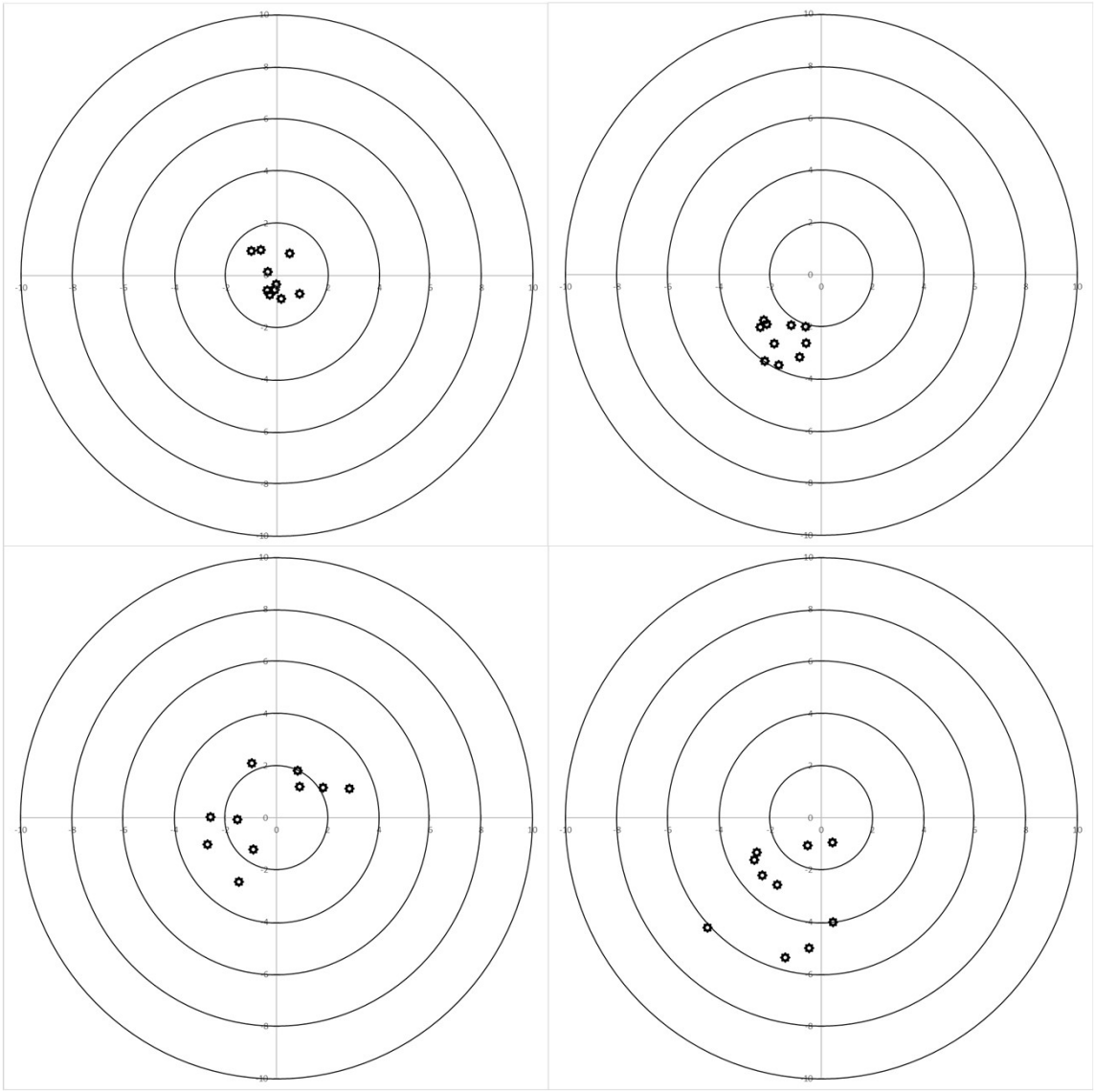
B



Zdroje nepřesnosti dat

Základní dva zdroje nepřesnosti dat

- systematická odchylka
- šum - rozptyl



Náhodné veličiny

- definice pravděpodobnosti na Kolmogorovském prostoru je poněkud abstraktní, proto se místo toho pracuje z náhodnými veličinami, které nabývají hodnot z R, reálných čísel a k těm pak přiřazujeme pravděpodobnosti
- distribuční funkce určuje rozdělení pravděpodobnosti $F(x) = P(X \leq x)$ pro $x \in R$; $0 \leq F(x) \leq 1$
- zprava spojitá; neklesající $x_i < x_j \Rightarrow F(x_i) \leq F(x_j)$; $\lim_{x \to -\infty} F(x) = 0$; $\lim_{x \to \infty} F(x) = 1$

Míry polohy

střední hodnota

- $EX = \sum_I x_i p_i$, kde $P[X = x_i] = p_i$, pro $i \in I$

  $EX = \int_R xf(x)dx$, $f(x)$ je hustota rozdělení X
- $E(c) = c$; $E(aX) = aE(X)$; $E(X + Y) = E(X) + E(Y)$
- $E(XY) = E(X)E(Y)$ pokud X a Y jsou nezávislé

Míry polohy

medián

- $P(X \leq m) \geq \frac{1}{2}$ a zároveň $P(X \geq m) \geq \frac{1}{2}$
  dělí soubor na stejné poloviny

modus

- nejčastější, typická hodnota

Rozptyl a směrodatná odchylka

rozptyl

- $\sigma^2 = var(X) = E\left(\left(X^2 - E(X)\right)^2\right) = E(X^2) - (E(X))^2$

směrodatná odchylka

- $\sigma = \sqrt[2]{var(X)}$

Příklady rozdělení náhodných veličin

- rovnoměrné diskrétní nabývá hodnot 1, 2, …, n s pravděpodobnostmi 1/n; $EX = (n+1)/2$;
- dichotomické nabývá hodnot 0 a 1 s pravděpodobnostmi q a p (q = 1-p)
- binomické – počet úspěchů při n opakováních pokusu s pravděpodobností úspěchu p

- Poissonovo r

Zákon velkých čísel

- průměr ***n*** stejně rozdělených, nezávislých veličin s konečnou střední hodnotou $\mu < \infty$ a konečným rozptylem $\sigma^2 < \infty$ konverguje k μ
- jinými slovy pokud máme ***n*** nezávislých měření X_i, ‚rozumné' náhodné veličiny, pak $\frac{\sum_{i=1}^{n} X_i}{n} \to \mu$

Poznámka: veličiny jsou nezávislé, proto nezáleží na tom, jak dopadlo prvních ***k*** pokusů

Centrální limitní věta

- mějme posloupnost ***n*** stejně rozdělených nezávislých veličin X_i s konečnou střední hodnotou $\mu < \infty$ a konečným rozptylem $\sigma^2 < \infty$ a $X = \sum_{i=1}^{n} X_i$,
  pak $U = \frac{X - n\mu}{\sqrt{n\sigma^2}} \sim N(0,1)$, kde N(0,1) je normované normální rozdělení
- přesnost odhadu roste s $\sqrt{n}$ pro zdvojnásobení přesnosti tedy potřebuji 4× více pozorování

‚Zákon malých čísel‘

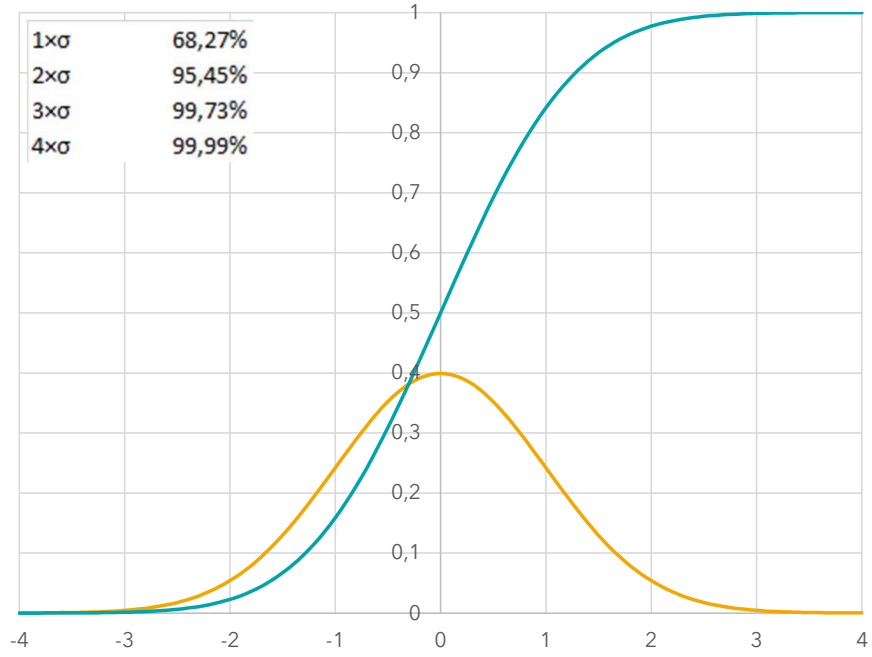
- analýza výskytu rakoviny ledvin v USA ukázala, že nejrizikovější jsou venkovské okresy
- příčinou může být odlišná strava, odlišný životní styl, znečištění prostředí postřiky
- kde byste hledali vy?

‚Zákon malých čísel‘

3,55	3,52	3,52	3,57	3,43	2,58	4,00	3,75	3,83	3,25
kostka_1	kostka_2	kostka_3	kostka_4	kostka_5	kostka_106	kostka_107	kostka_108	kostka_109	kostka_110
5	5	1	4	2	1	3	5	2	2
5	3	4	1	2	5	5	6	2	5
4	4	5	2	4	1	6	1	3	6
1	4	2	4	4	1	5	6	3	2
6	3	2	3	2	3	1	6	4	2
3	2	4	6	1	1	5	2	4	2
3	4	3	5	5	3	1	2	4	2
4	4	2	1	1	3	4	6	5	2
4	4	4	1	4	3	5	1	6	2
6	6	3	2	2	2	5	3	3	3
2	6	4	2	4	6	4	6	5	6
5	3	2	4	4	2	4	1	5	5

Normální rozdělení N(0,1)

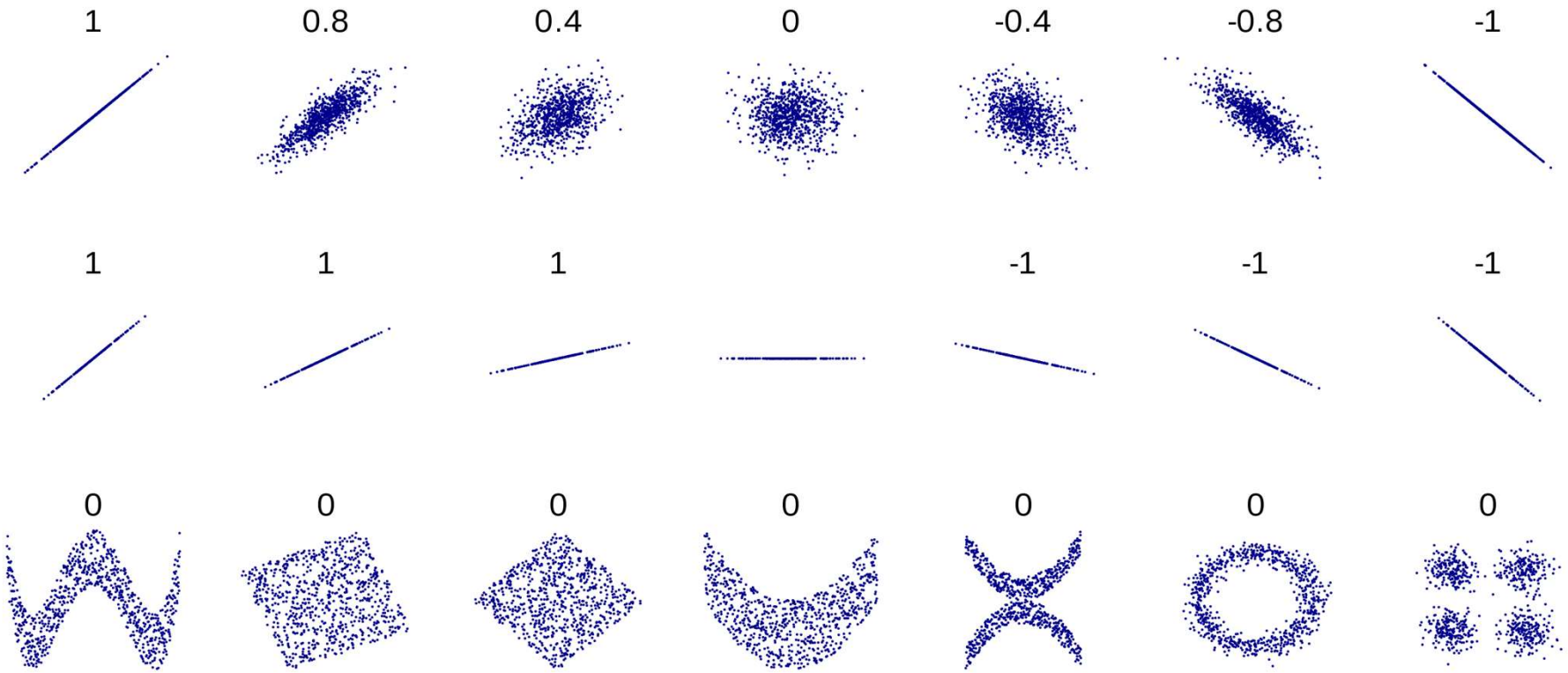
- IQ má rozdělení N(100,15)
- od v rozmezí IQ 85 – 115 leží více, než 2/3 populace
- v ČR je cca 240 000 lidí s IQ 130 a výše (Mensa)
- cca 14 000 s IQ nad 145 a pouze cca 333 s IQ nad 160



Výběrová střední hodnota a rozptyl pro dvě veličiny

- $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim EX = \mu$
- $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2) \sim var(X) = \sigma^2$
- kovariance $cov(X,Y) = E\big((X - EX)(Y - EY)\big) = E(XY) - E(X)E(Y)$
  pro X, Y nezávislé $E(XY) = E(X)E(Y)$, ne naopak!
- Pearsonův korelační koeficient $\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{(E(X^2) - \big(E(X)\big)^2)(E(Y^2) - (E(Y))^2)}}$
- $var(X \pm Y) = var(X) + var(Y) \pm cov(X,Y)$

Korelační koeficient (z Wikipedie)

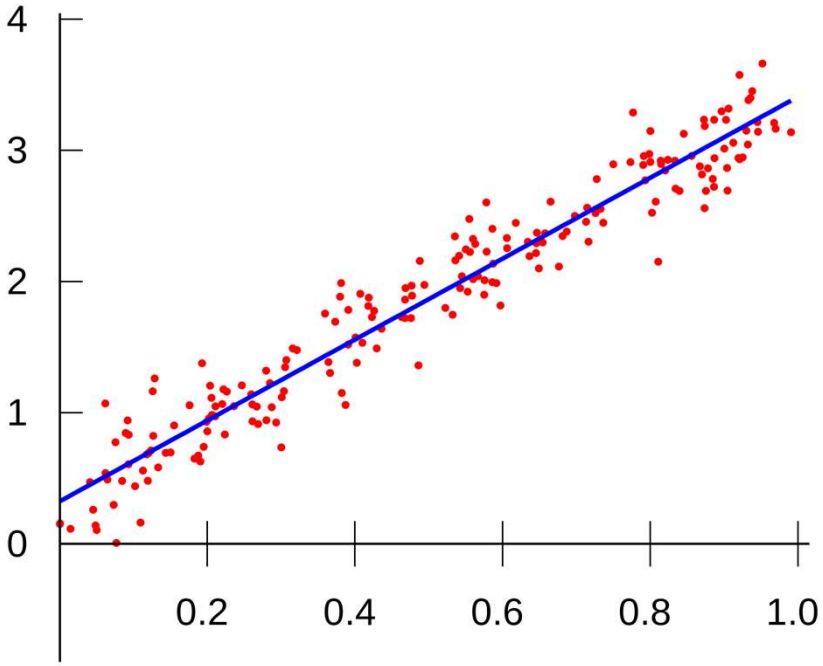


Lineární regrese

- ***n*** dvojic bodů x_i, y_i a hledám koeficienty pro vztah $y_i = kx_i + q + \varepsilon_i$, kde ε_i je reziduální odchylka (šum)
- minimalizujeme součet čtverců odchylek $\sum_{i=1}^{n}(y_i - \bar{y})^2$

$$k = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} \qquad \text{a} \qquad q = \frac{n\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

Ukázka



- má smysl pouze pro vysoký korelační koeficient

Zobecnění

- mnohorozměrná lineární regrese – mám sadu proměnných X_i a pomocí jejich lineární kombinace vysvětluji Y
- nelineární regrese, závislost může být kvadratická, exponenciální, logaritmická atd.
- záleží na objemu dat, který je k dispozici u obřích datových souborů, lze efektivně využívat AI, která dokáže najít v datech libovolné vzory (pattern recognition)

Jak se zbavit šumu

- zvětšením vzorku
- nahrazením lidského rozhodování lineárním modelem
- nahrazením lidského rozhodování AI
- náhodným rozdělením dat na trénovací vzorek (70 %) a testovací vzorek (30 %), model se vyvíjí na trénovacích datech a na testovacích se ověří a vyloučí se náhodné artefakty (overfitting, přetrénování modelu)

Jsou složité modely lepší?

- žádný model není 1:1 s realitou, vždy ho lze vylepšit o další detaily a vztahy
- každý parametr v modelu je třeba statisticky odhadnout a tento odhad sebou nese chybu
- je třeba hledat rovnováhu mezi komplexností modelu a naší schopností dostatečně přesně odhadnout parametry, případné chyby se někdy sčítají, ale někdy třeba násobí

Jsou složité modely lepší?

- existují modely s tisíci parametrů
- pokud tyto parametry popisují lidské chování, je třeba je neustále znovu kalibrovat
- nezbytné Big Data a AI
- redukované modely dávají často lepší výsledky, protože jsou parametry přesněji odhadnuté a omezuje se možnost chyb

Průzkumy veřejného mínění

- kvótní náhodný výběr vybírá respondenty podle hlavních sociodemografických charakteristik
- každá podskupina se může jinak chovat, mít jiné preference, tím, že zachováváme jejich podíl ve vzorku, hlídáme si, aby nedocházelo k systematickému zkreslení
- pokud chceme data analyzovat po skupinách, měla by každá mít nejméně 50 členů
- každý výzkumná zpráva povinně obsahuje protokol, popisující složení vzorku, způsob sběru dat, kdy byla data sbírána a další detaily

Ukázka intervalů spolehlivosti pro průzkumy veřejného mínění

5%	100	150	200	250	300	350	400	450	500	600	700	800	900	1000	1100	1200	1300	1400	1500	1600	1700	1800	1900	2000	5%	N = velikost vzorku
1%	–	–	–	–	–	–	–	–	–	–	–	–	–	0,6%	0,6%	0,6%	0,5%	0,5%	0,5%	0,5%	0,5%	0,5%	0,4%	0,4%	99%	hladina významnosti testu
2%	–	–	–	–	–	–	–	–	1,2%	1,1%	1,0%	1,0%	0,9%	0,9%	0,8%	0,8%	0,8%	0,7%	0,7%	0,7%	0,7%	0,6%	0,6%	0,6%	98%	úroveň měřené veličiny v populaci
3%	–	–	–	–	–	1,8%	1,7%	1,6%	1,5%	1,4%	1,3%	1,2%	1,1%	1,1%	1,0%	1,0%	0,9%	0,9%	0,9%	0,8%	0,8%	0,8%	0,8%	0,7%	97%	výběrová chyba
4%	–	–	–	2,4%	2,2%	2,1%	1,9%	1,8%	1,7%	1,6%	1,5%	1,4%	1,3%	1,2%	1,2%	1,1%	1,1%	1,0%	1,0%	1,0%	0,9%	0,9%	0,9%	0,9%	96%	
5%	–	–	3,0%	2,7%	2,5%	2,3%	2,1%	2,0%	1,9%	1,7%	1,6%	1,5%	1,4%	1,4%	1,3%	1,2%	1,2%	1,1%	1,1%	1,1%	1,0%	1,0%	1,0%	1,0%	95%	
10%	5,9%	4,8%	4,2%	3,7%	3,4%	3,1%	2,9%	2,8%	2,6%	2,4%	2,2%	2,1%	2,0%	1,9%	1,8%	1,7%	1,6%	1,6%	1,5%	1,5%	1,4%	1,4%	1,3%	1,3%	90%	
15%	7,0%	5,7%	4,9%	4,4%	4,0%	3,7%	3,5%	3,3%	3,1%	2,9%	2,6%	2,5%	2,3%	2,2%	2,1%	2,0%	1,9%	1,9%	1,8%	1,7%	1,7%	1,6%	1,6%	1,6%	85%	
20%	7,8%	6,4%	5,5%	5,0%	4,5%	4,2%	3,9%	3,7%	3,5%	3,2%	3,0%	2,8%	2,6%	2,5%	2,4%	2,3%	2,2%	2,1%	2,0%	2,0%	1,9%	1,8%	1,8%	1,8%	80%	
25%	8,5%	6,9%	6,0%	5,4%	4,9%	4,5%	4,2%	4,0%	3,8%	3,5%	3,2%	3,0%	2,8%	2,7%	2,6%	2,5%	2,4%	2,3%	2,2%	2,1%	2,1%	2,0%	1,9%	1,9%	75%	
30%	9,0%	7,3%	6,4%	5,7%	5,2%	4,8%	4,5%	4,2%	4,0%	3,7%	3,4%	3,2%	3,0%	2,8%	2,7%	2,6%	2,5%	2,4%	2,3%	2,2%	2,2%	2,1%	2,1%	2,0%	70%	
35%	9,3%	7,6%	6,6%	5,9%	5,4%	5,0%	4,7%	4,4%	4,2%	3,8%	3,5%	3,3%	3,1%	3,0%	2,8%	2,7%	2,6%	2,5%	2,4%	2,3%	2,3%	2,2%	2,1%	2,1%	65%	
40%	9,6%	7,8%	6,8%	6,1%	5,5%	5,1%	4,8%	4,5%	4,3%	3,9%	3,6%	3,4%	3,2%	3,0%	2,9%	2,8%	2,7%	2,6%	2,5%	2,4%	2,3%	2,3%	2,2%	2,1%	60%	
45%	9,8%	8,0%	6,9%	6,2%	5,6%	5,2%	4,9%	4,6%	4,4%	4,0%	3,7%	3,4%	3,3%	3,1%	2,9%	2,8%	2,7%	2,6%	2,5%	2,4%	2,4%	2,3%	2,2%	2,2%	55%	
50%	9,8%	8,0%	6,9%	6,2%	5,7%	5,2%	4,9%	4,6%	4,4%	4,0%	3,7%	3,5%	3,3%	3,1%	3,0%	2,8%	2,7%	2,6%	2,5%	2,5%	2,4%	2,3%	2,2%	2,2%	50%	