Přednáška 10

Korelace a regrese

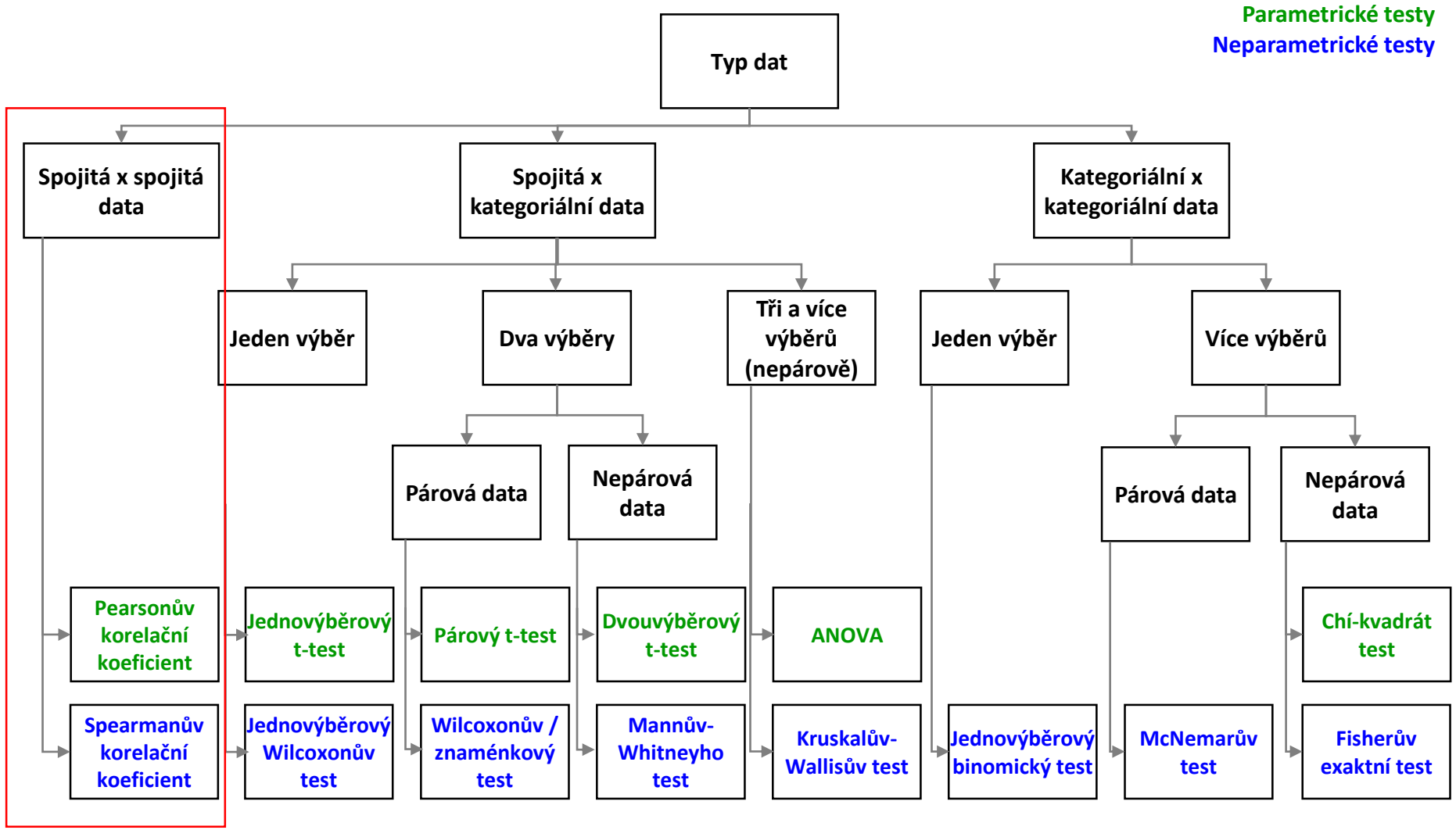
Parametrická a neparametrická korelace

Lineární regrese

Anotace

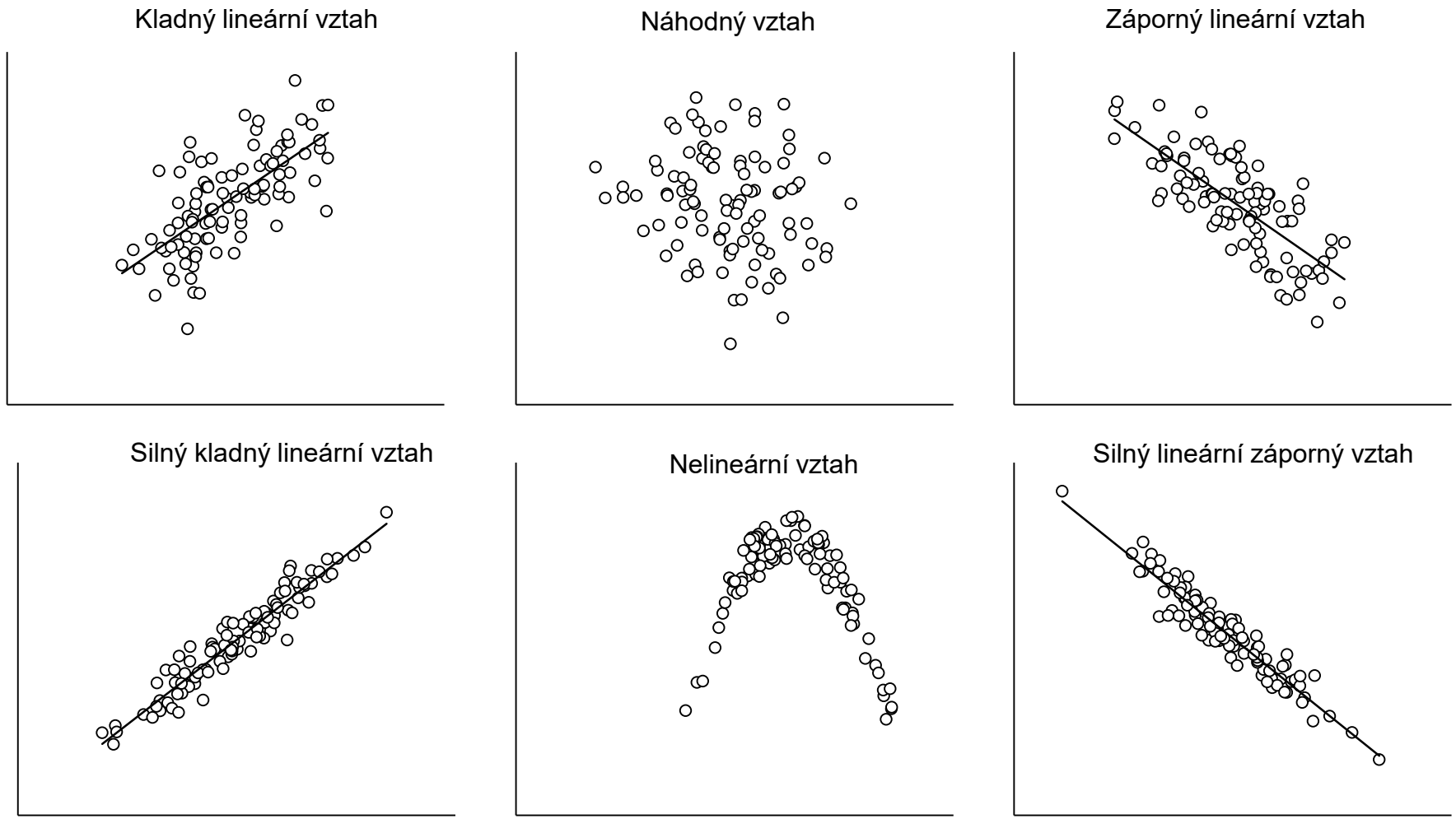
- Korelační analýza je využívána pro vyhodnocení míry vztahu dvou spojitých proměnných.
- Obdobně jako jiné statistické metody, i korelace mohou být parametrické nebo neparametrické
- Regresní analýza vytváří model vztahu dvou nebo více proměnných, tedy jakým způsobem jedna proměnná (vysvětlovaná) závisí na jiných proměnných (prediktorech).
- Regresní analýza je obdobně jako ANOVA nástrojem pro vysvětlení variability hodnocené proměnné

Základní rozhodování o výběru statistických testů



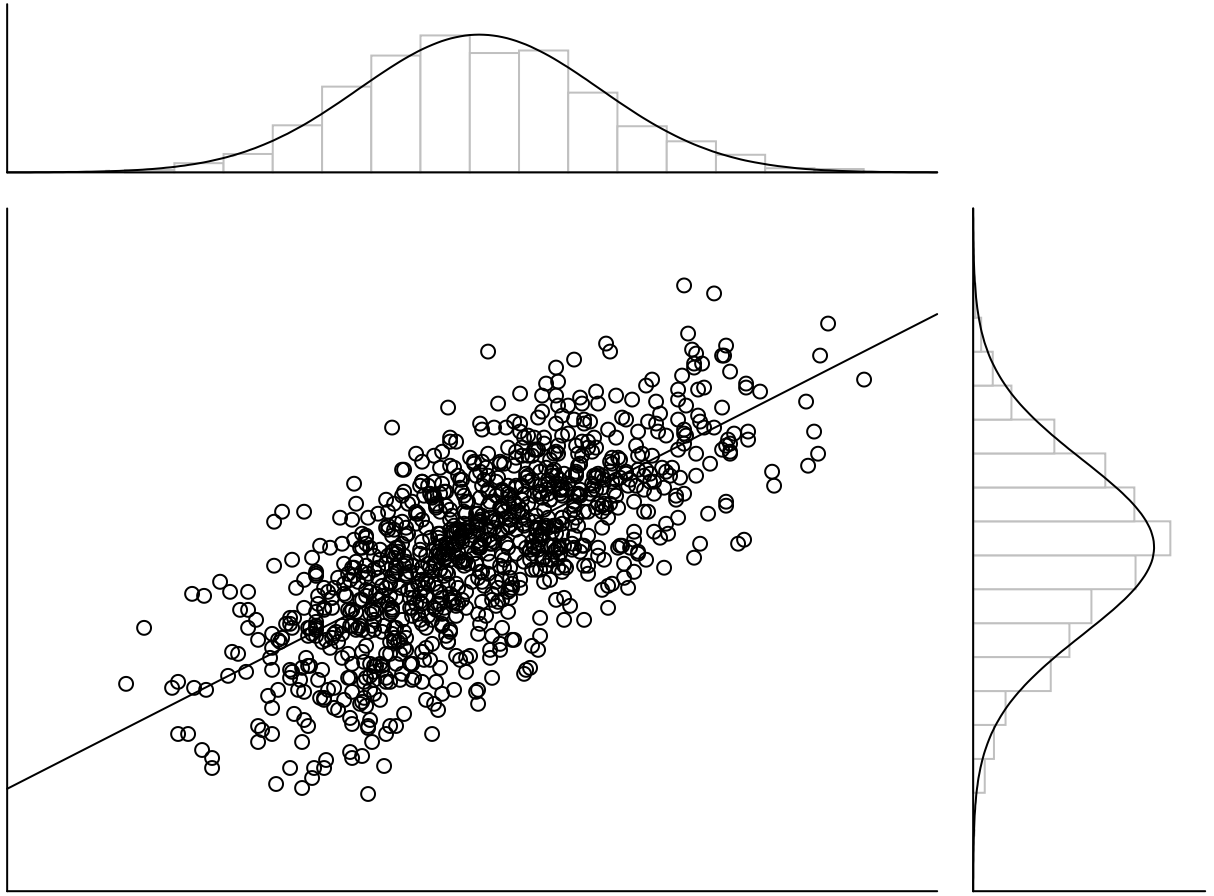
Popis vztahu spojitých proměnných

- Základním nástrojem popisu vztahu spojitých proměnných je XY graf umožňující posoudit typ a sílu jejich vztahu.



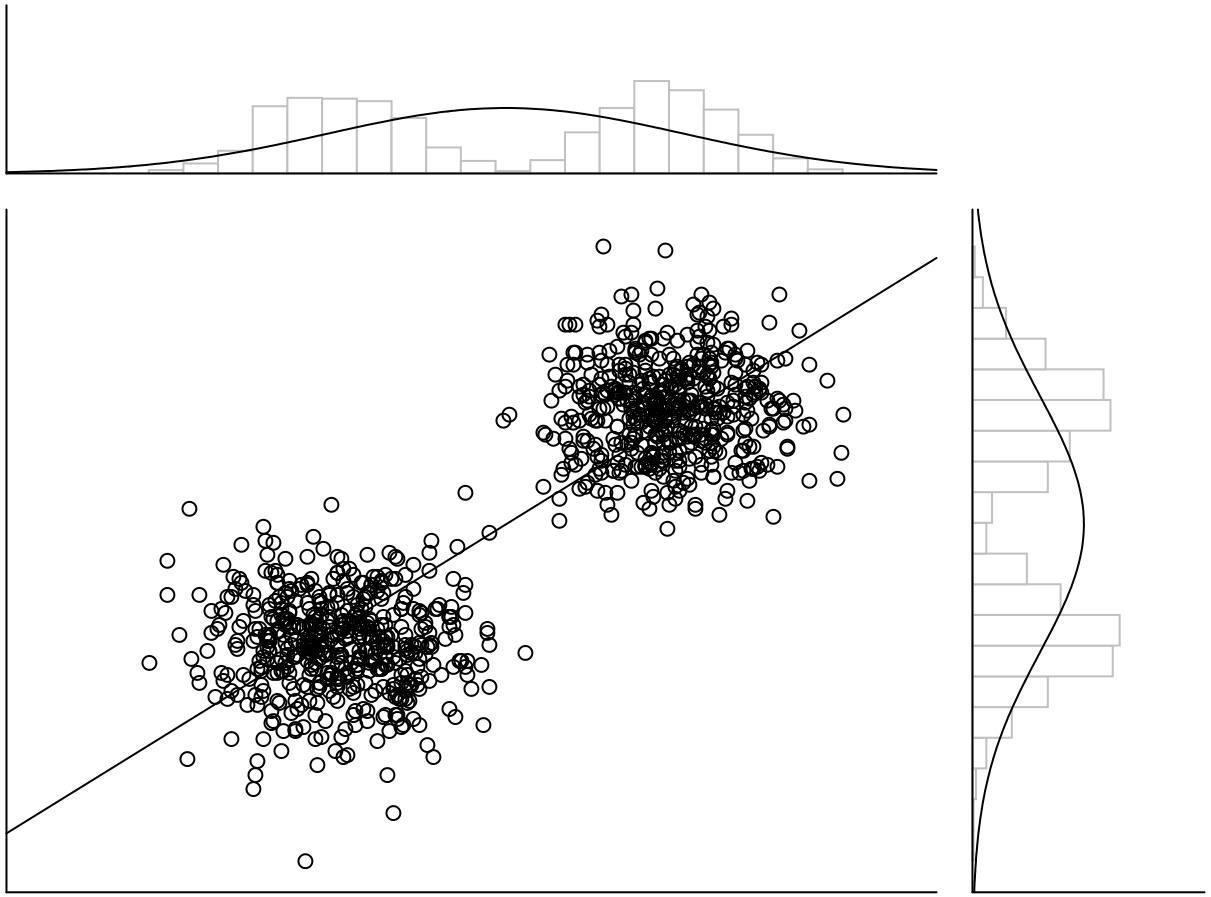
Předpoklady parametrické korelační analýzy

- Korektní interpretace parametrické korelační analýzy předpokládá lineární vztah mezi proměnnými a normální rozložení hodnot obou proměnných.



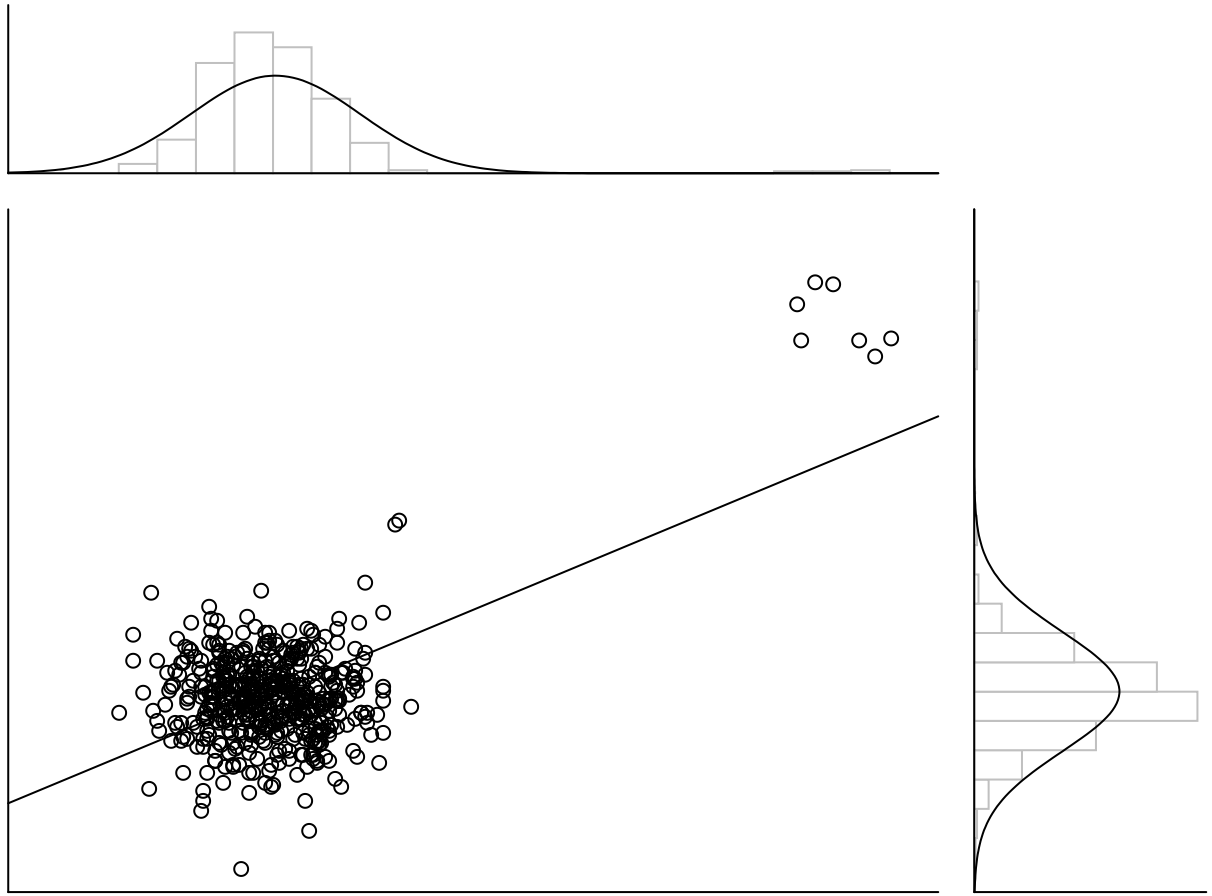
Bimodální rozložení hodnot vstupujících do korelační analýzy

- V případě bimodálního rozložení hodnot vstupujících do korelační analýzy není vhodné korelační analýzu počítat; výsledek není možné interpretovat jako popis lineárního vztahu spojitých proměnných, ale jako důsledek existence podskupin objektů v datech.



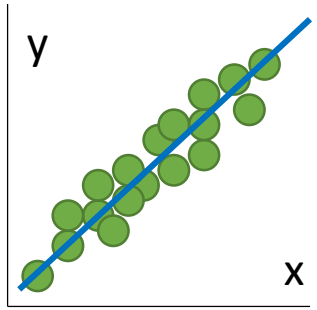
Přítomnost odlehlých hodnot v datech vstupujících do korelační analýzy

- V případě přítomnosti odlehlých hodnot v datech vstupujících do korelační analýzy není vhodné korelační analýzu počítat; výsledek není možné interpretovat jako popis lineárního vztahu spojitých proměnných, ale jako důsledek přítomnosti odlehlých hodnot v datech.

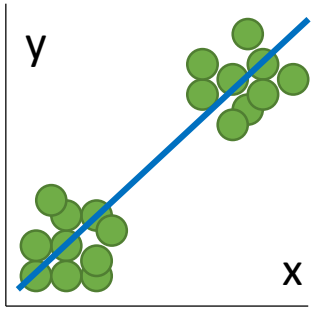


Korelace a kovariance – parametrické míry vztahu spojitých proměnných

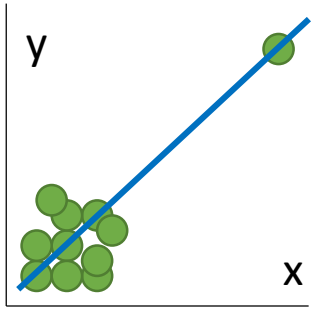
- Kovariance a Pearsonův korelační koeficient jsou základní metody pro popis lineárního vztahu spojitých proměnných
- Předpokladem výpočtu kovariance a Pearsonova korelačního koeficientu je:
  - Normalita dat v obou dimenzích
  - Linearita vztahu proměnných



Lineární vztah – bezproblémové použití kovariance nebo Pearsonova korelačního koeficientu

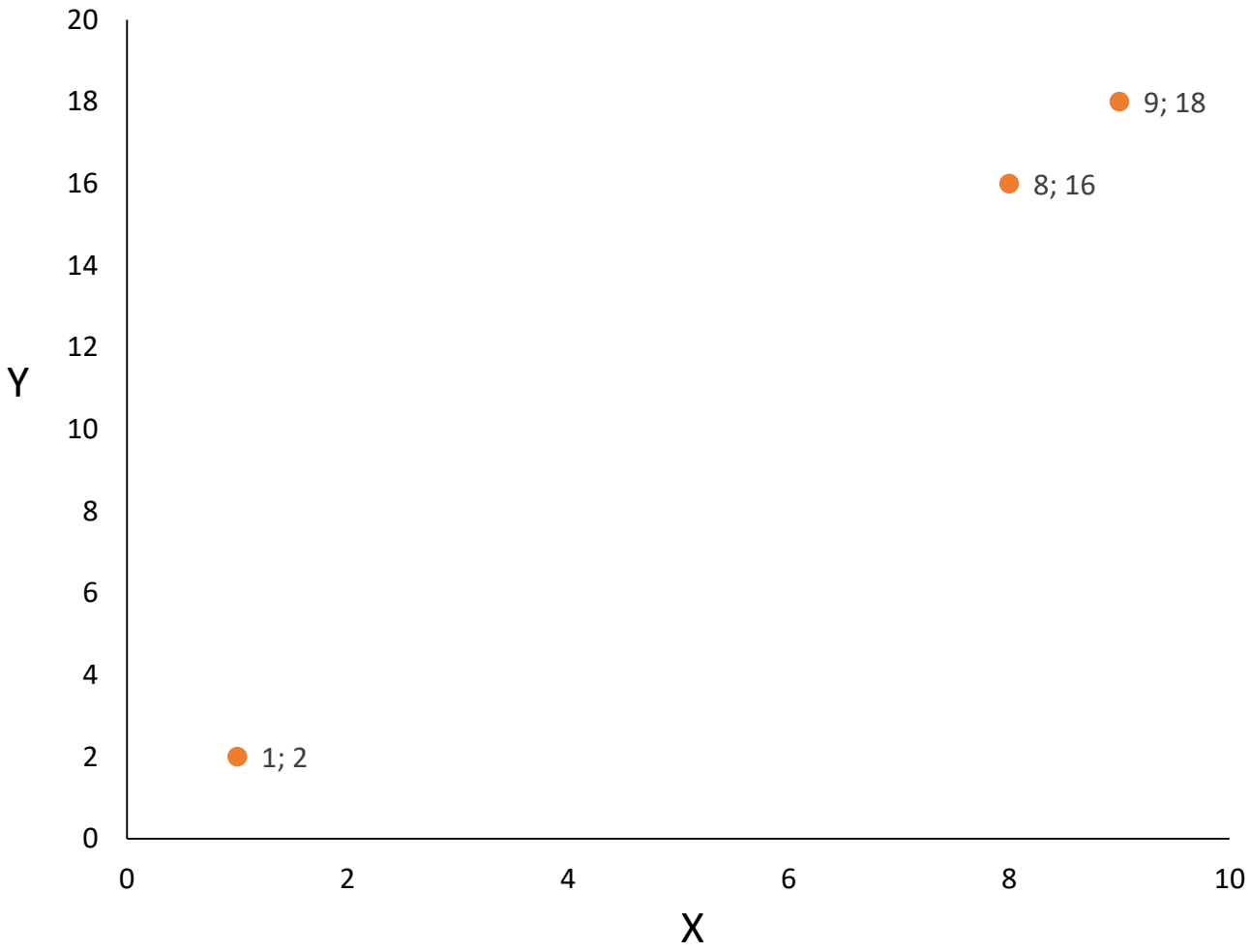


Korelace je dána dvěma skupinami hodnot – vede k identifikaci skupin objektů v datech



Korelace je dána odlehlou hodnotu – analýza popisuje pouze vliv odlehlé hodnoty

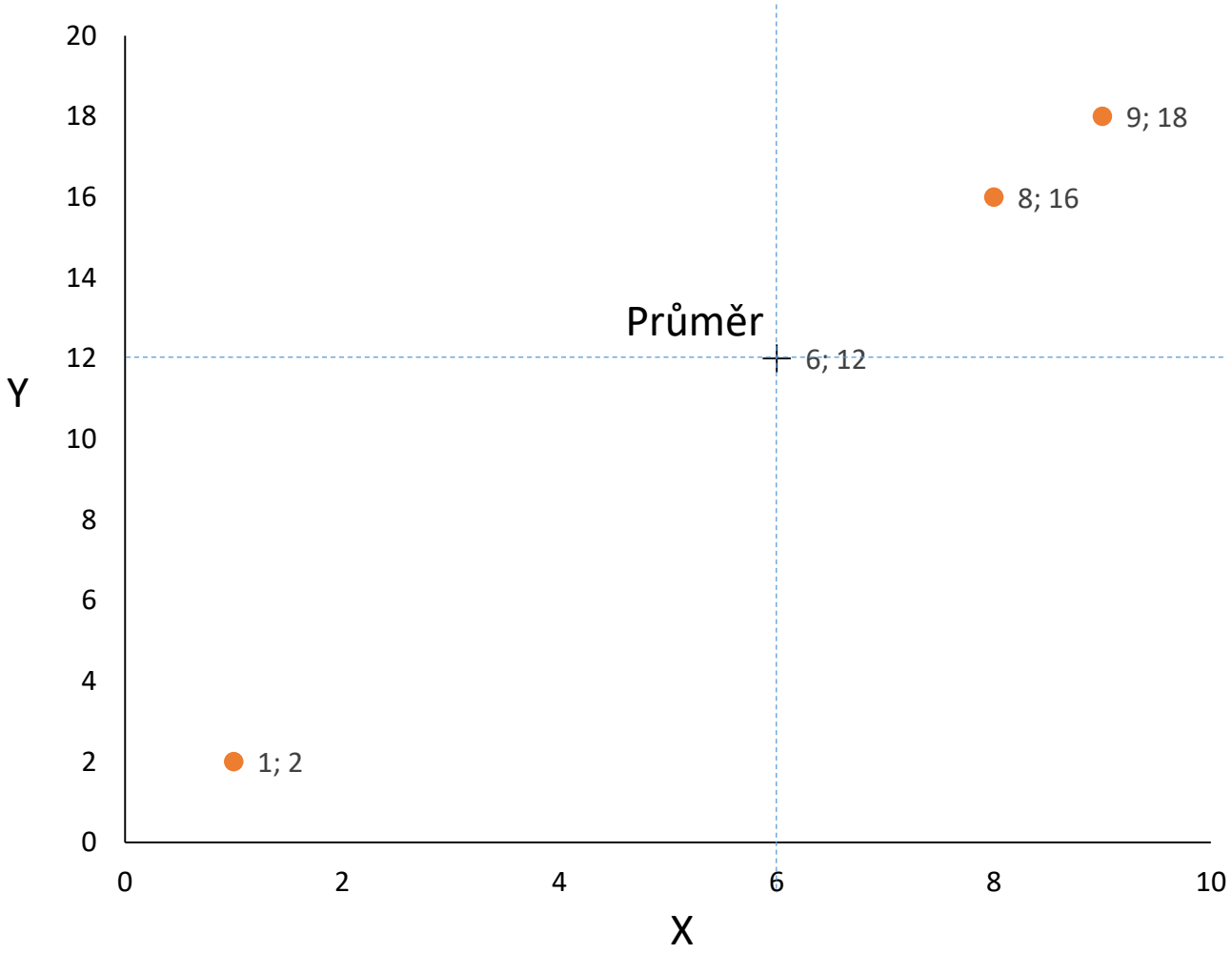
Výpočet kovariance I



Kovariance = sdílený rozptyl

Jak číselně popsat vztah proměnných?

Výpočet kovariance II

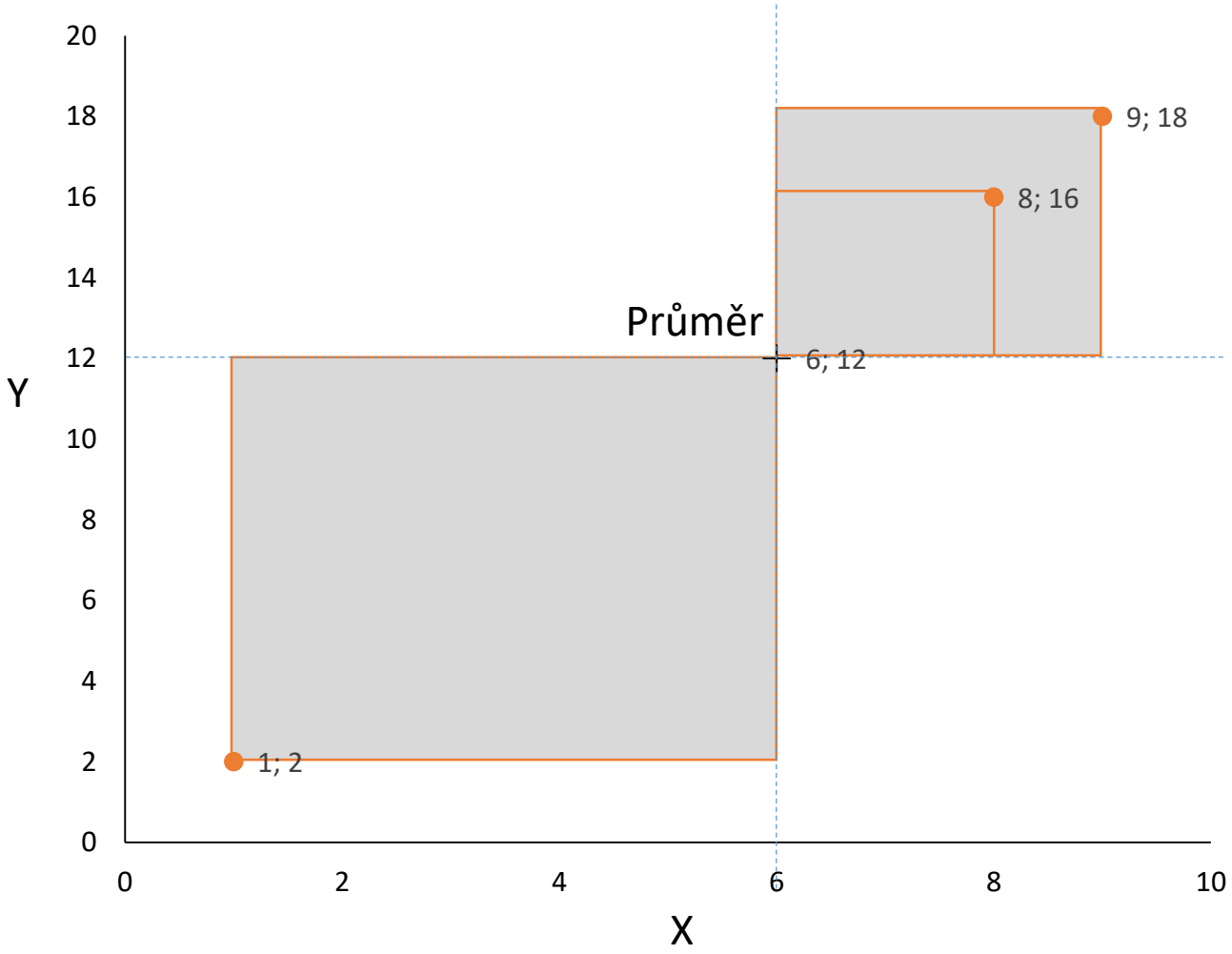


Kovariance = sdílený rozptyl

Jak číselně popsat vztah proměnných?

Data se vyskytují v různých kvadrantech dle průměru !

Výpočet kovariance III



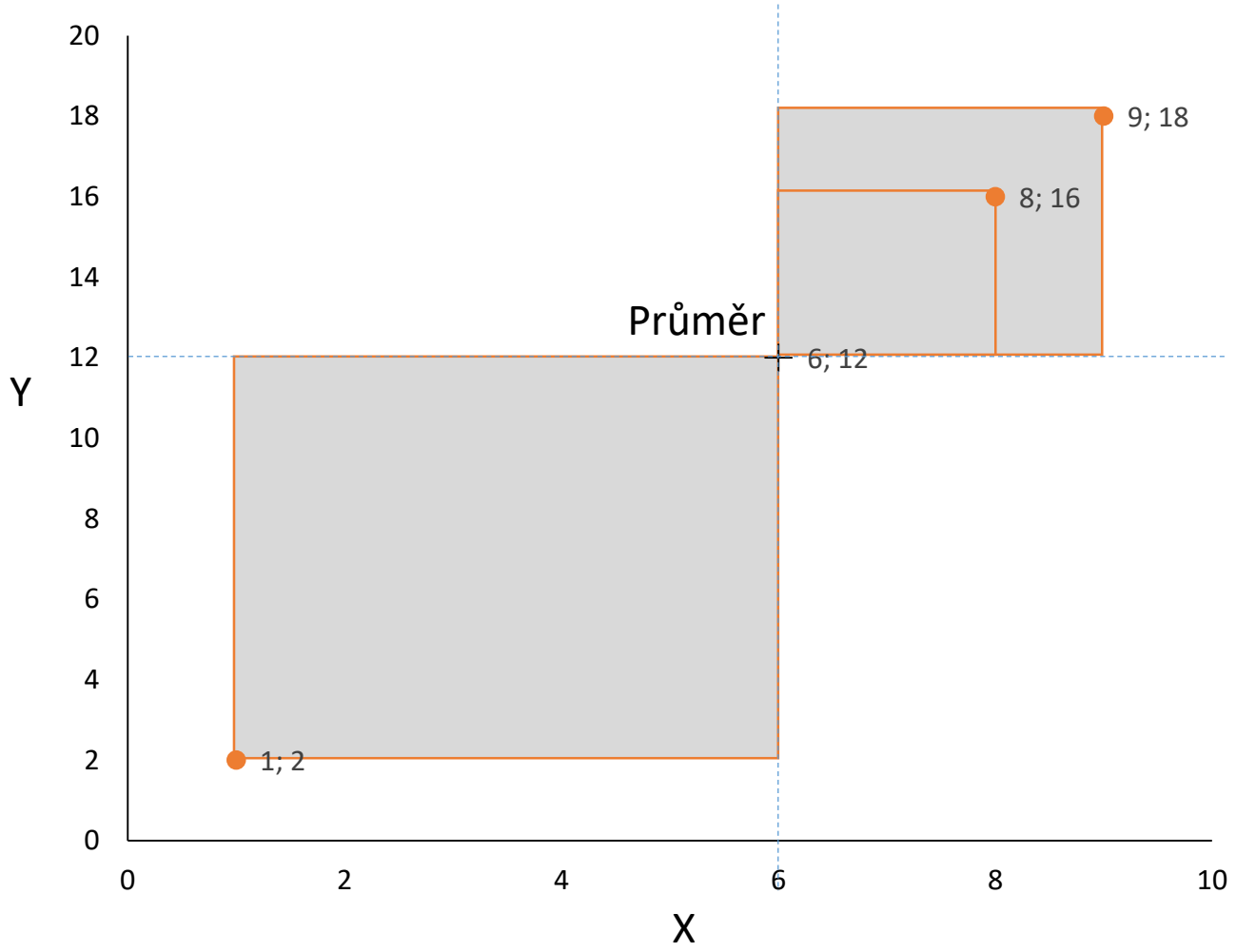
Kovariance = sdílený rozptyl

Jak číselně popsat vztah proměnných?

Data se vyskytují v různých kvadrantech dle průměru !

Sdílený rozptyl počítejme obdobně jako rozptyl !!

Výpočet kovariance IV



Kovariance = sdílený rozptyl

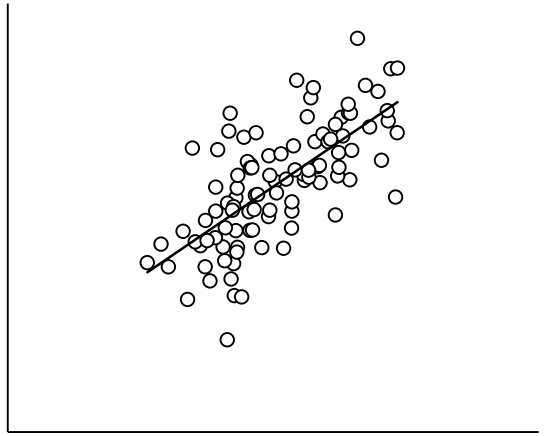
Jak číselně popsat vztah proměnných?

Data se vyskytují v různých kvadrantech dle průměru !

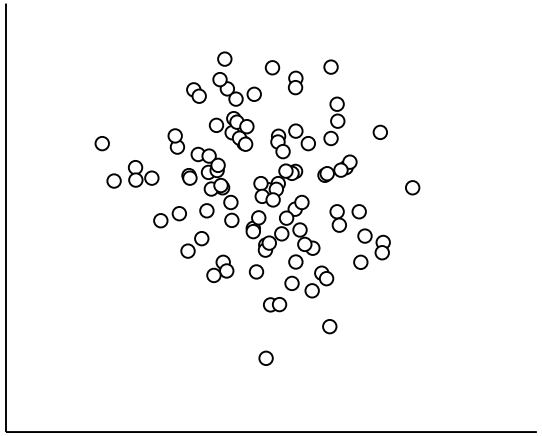
Sdílený rozptyl počítejme obdobně jako rozptyl !!

$$Cov(x,y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x}) * (y_i - \bar{y})}{N-1}$$

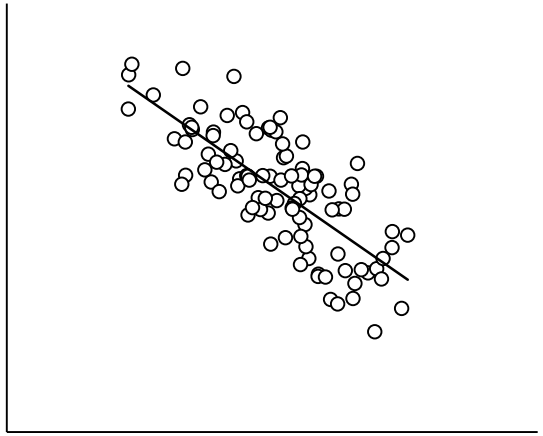
Výpočet kovariance IV



Cov = ?

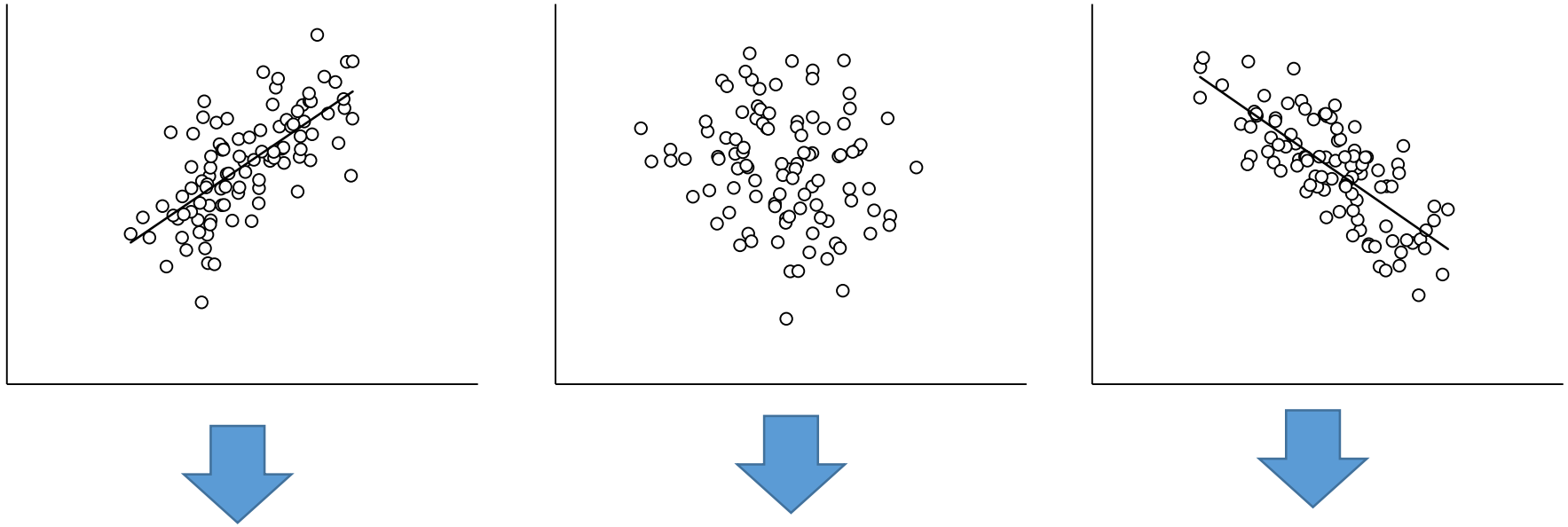


Cov = ?



Cov = ?

Výpočet kovariance IV



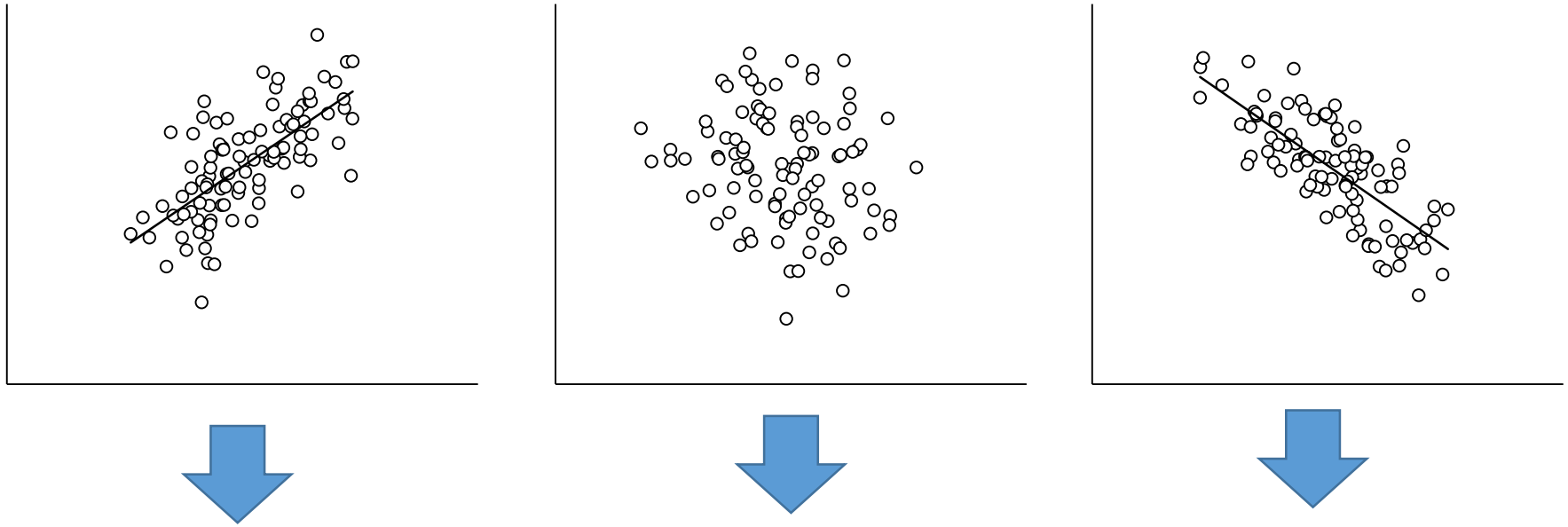
Cov = kladné číslo

Cov = 0

Cov = záporné číslo

Existuje nějaké dané minimum a maximum kovariance?

Výpočet kovariance IV



Cov = kladné číslo

Cov = 0

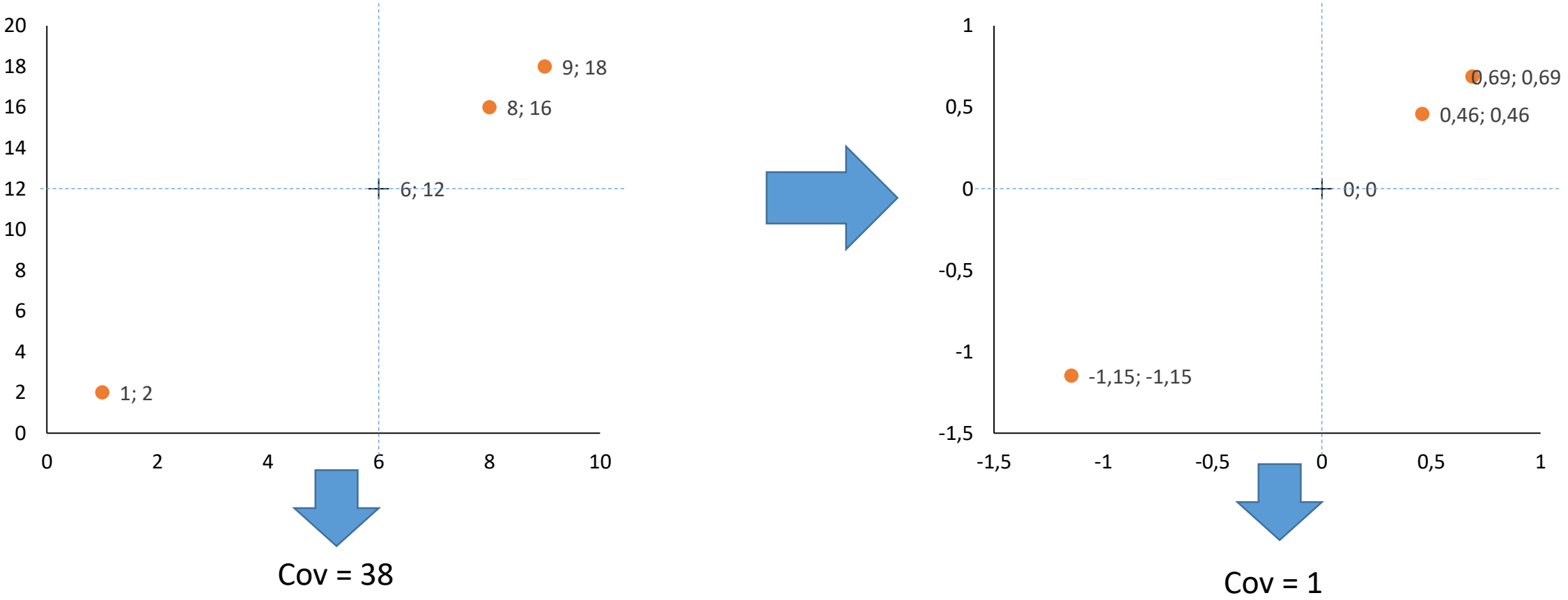
Cov = záporné číslo

Existuje nějaké dané minimum a maximum kovariance?

Neexistuje, teroeticky může být kovariance od - ∞ do +∞; nevýhoda při interpretaci

Kovariance standardizovaných dat

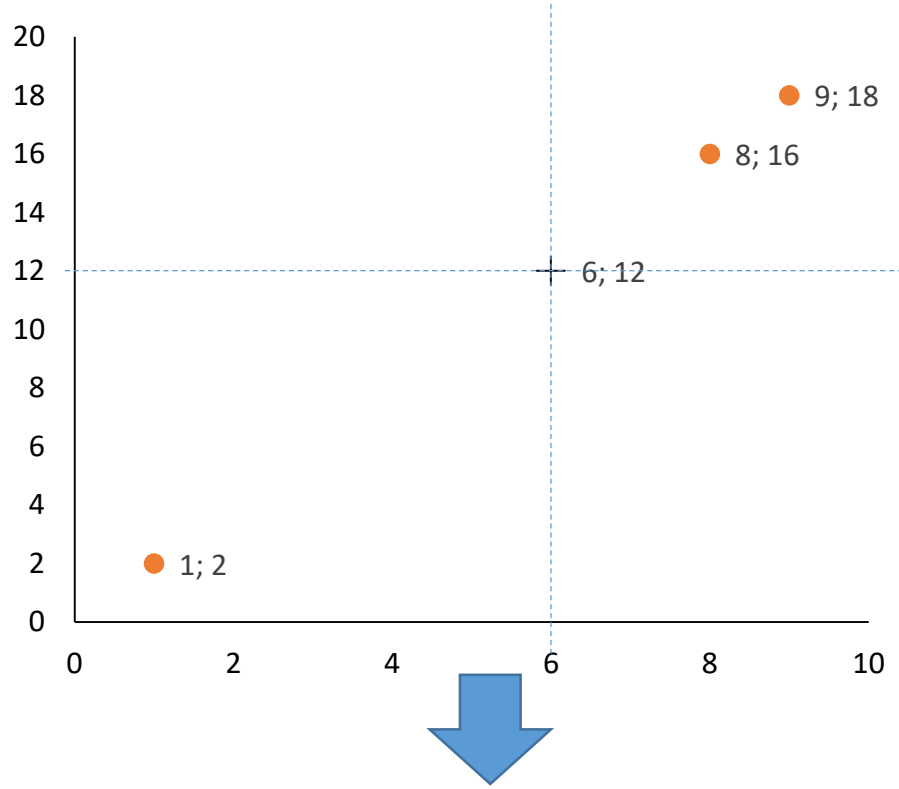
- Jak dopadne výpočet kovariance na datech se standardním normálním rozložením (průměr = 0, rozptyl =1)?



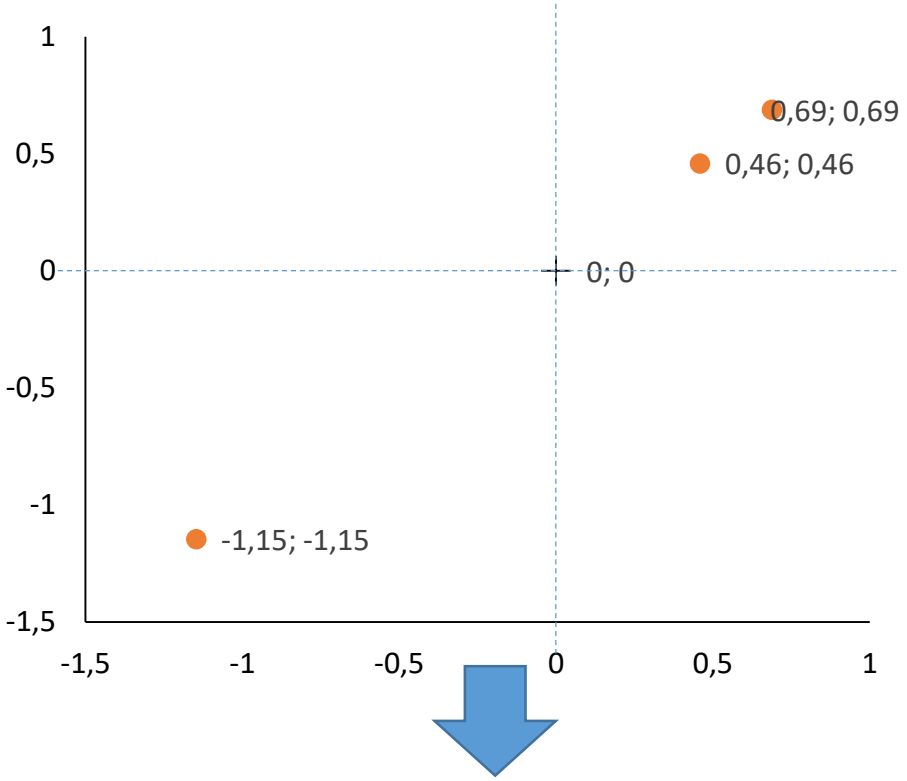
Výpočet Pearsonova korelačního koeficientu

- Pearsonův korelační koeficient představuje standardizovanou formu kovariance

$$r(x,y) = \frac{Cov(x,y)}{s_x s_y}$$

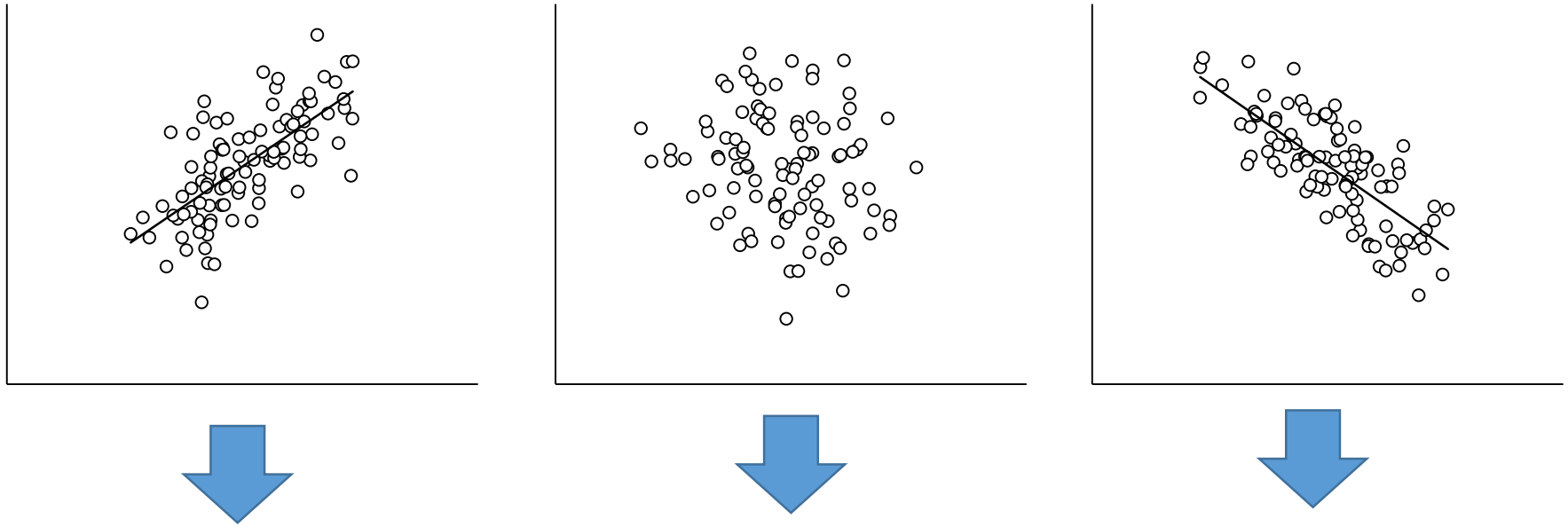


Cov = 38; r = 1



Cov = 1; r=1

Výpočet Pearsonova korelačního koeficitu



r = kladné číslo ≤ 1

r = 0

r = záporné číslo ≥ -1

Existuje nějaké dané minimum a maximum Pearsonova korelačního koeficientu?

Ano, Pearsonův korelační koeficient se pohybuje v rozsahu <-1;1>

Testování Pearsonova korelačního koeficientu

P_I (zem)	10	14	15	32	40	20	16	50
P_I (rostl.)	19	22	26	41	35	32	25	40

$$I = 1,.....,n; n = 8; v = 6$$

$$r = \frac{Cov(x,y)}{S_x.S_y} = \frac{\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i}{\sqrt{\left[\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2\right]\left[\sum y_i^2 - \frac{1}{n}\left(\sum y_i\right)^2\right]}} = 0,7176$$

I. $H_0 : \rho = \phi : \alpha = 0,05$

$\mathbf{tab}: r(v = 6) = 0,7076$

II. $H_0 : \rho = \phi$

$$t = \left[\frac{r}{\sqrt{1-r^2}}\right] \cdot \sqrt{n-2} \qquad v = n-2$$

$$\left.\begin{array}{l} t = \dfrac{0,7176}{0,6965} \cdot \sqrt{6} = 2,524 \\ \mathbf{tab}: \; t_{0,975}^{(n-2)} = 2,447 \end{array}\right\} \; P \leq 0,05$$

Srovnání dvou korelačních koeficientů (r)

1. $n_1 = 1258$ $r_1 = 0{,}682$

2. $n_2 = 462$ $r_2 = 0{,}402$

Krevní tlak x koncentrace kysl. radikálů

$$Z_i = 1.1513 \cdot \log \frac{(1 + r_i)}{(1 - r_i)}$$

$Z_1 = 0{,}833$ $Z_2 = 0{,}426$

Test: $H_0 : \rho_1 = \rho_2$; $\alpha = 0{,}05$

$$Z = \frac{Z_1 - Z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} = \frac{0{,}407}{0{,}0545} = 7{,}461$$

tabulky : $Z_{0{,}975} = 1{,}96$

7,461 >> 1,96 => P << 0,01

Neparametrická korelace (Spearmanův korelační koeficient - rs)

P_I v půdě	1	2	3	6	7	5	4	8
P_I v rostl.	1	2	4	8	6	5	3	7
d_I	0	0	1	2	-1	0	-1	-1

i = 1, ….. n; n = 8 => v = 6

$$r_s = 1 - \frac{6 \cdot \sum di^2}{n(n^2 - 1)} = 0{,}9048$$

$$\textbf{tab}: \; r_s(v = 6) = 0{,}89$$

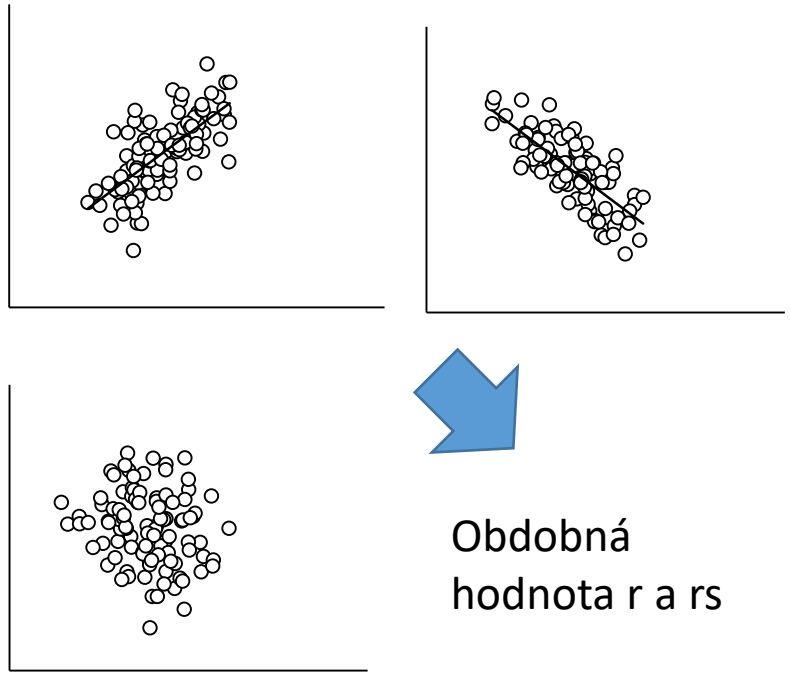
Pacient č.	1	2	3	4	5	6	7
Lékař 1	4	1	6	5	3	2	7
Lékař 2	4	2	5	6	1	3	7
d_I	0	-1	1	-1	2	-1	0

$$r_s = 1 - \frac{6 \cdot 8}{7(49 - 1)} = 0{,}857$$

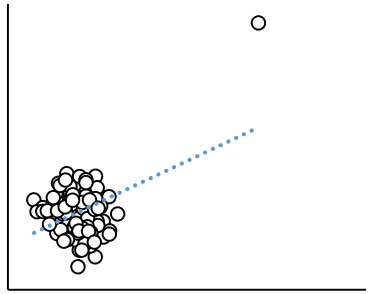
P = 0,358

Pearsonův a Spearmanův korelační koeficient

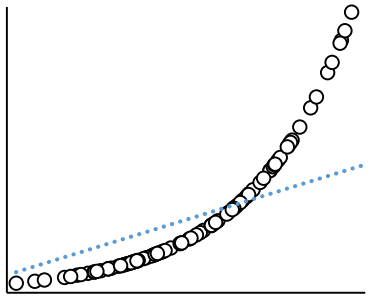
- Porovnání hodnot Pearsonova (r) a Spearmanova (rs) korelačního koeficientu umožňuje posoudit typ vztahu promenných



Obdobná hodnota r a rs

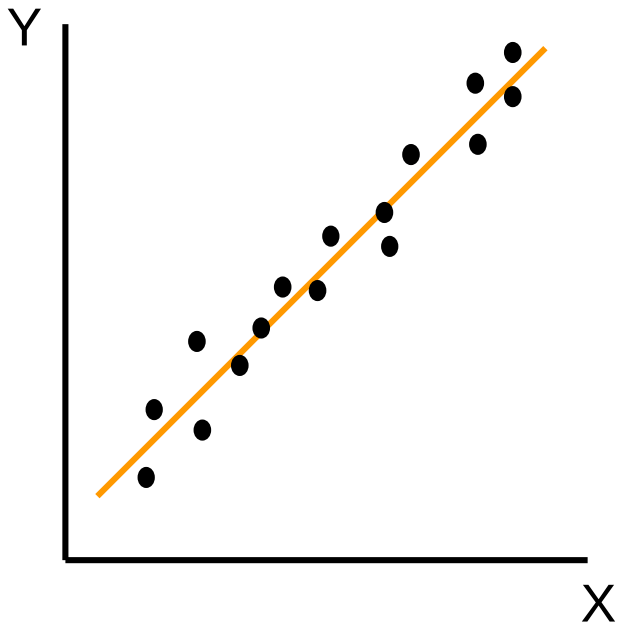


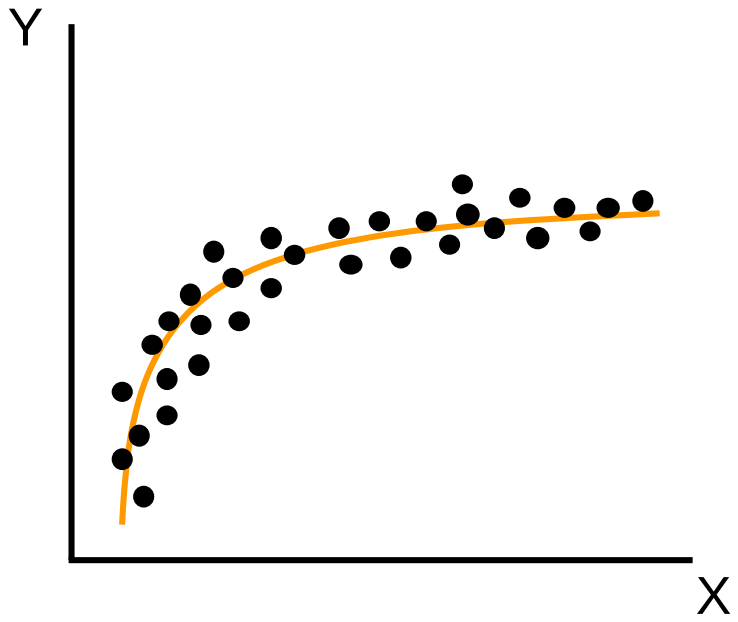
Vysoké r (díky odlehlé hodnotě) a nízké rs (odlehlá hodnota odstraněna transformací na pořadí)



Nízké r (díky nelinearitě vztahu) a vysoké rs (v pořadích jde o silný vztah obou proměnných)

Korelace v grafech I.





Vztahy velmi často implikují funkční vztah mezi Y a X.

$$Y = a + b \cdot X$$

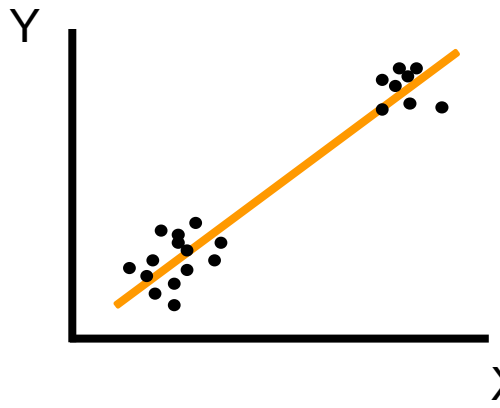
$$Y = a + b_1 \cdot X_1 + b_2 \cdot X_2 + b_3 \cdot X_3$$

$$Y = a + b_1 \cdot X_1 + b_2 \cdot X_2$$

$$Y = a + b_1 \cdot X_1 + b_2 \cdot X_2 + b_3 \cdot X_1 \cdot X_2$$

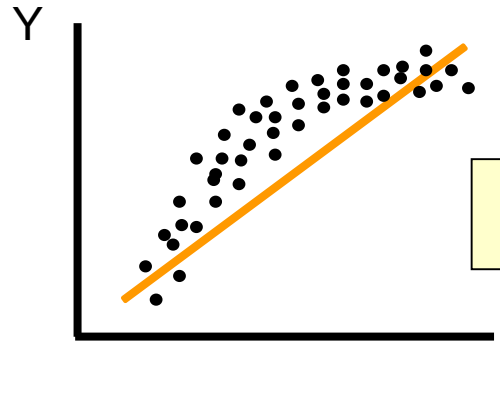
Korelace v grafech II.

Problém rozložení hodnot



r = 0,981
(p < 0,001)

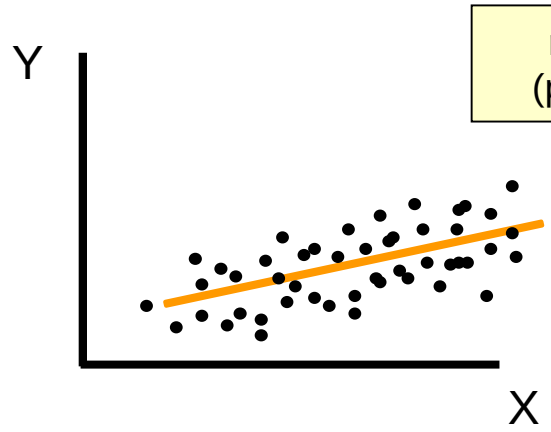
Problém typu modelu



r = 0,761
(p < 0,032)

Problém velikosti vzorku

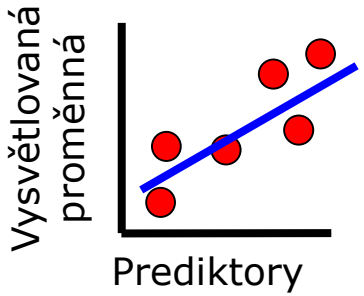
r = 0,891
(p < 0,214)



r = 0,212
(p < 0,008)

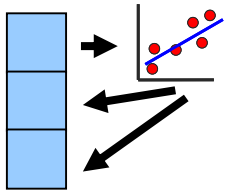
Vytváření modelů

1.Tvorba modelu



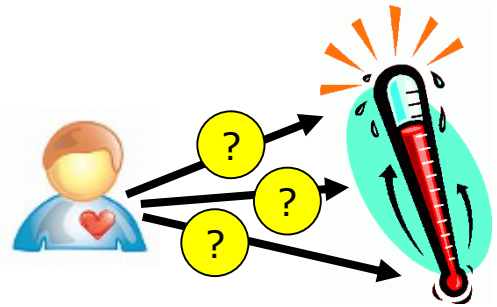
•Parametry ovlivňující vysvětlovanou charakteristiku pacienta
• Rovnice umožňující predikci
• Platnost modelu pouze v rozsahu prediktorů

2.Validace modelu



• Nebezpečí „přeučení" modelu
• Testování modelu na známých datech
•Krosvalidace

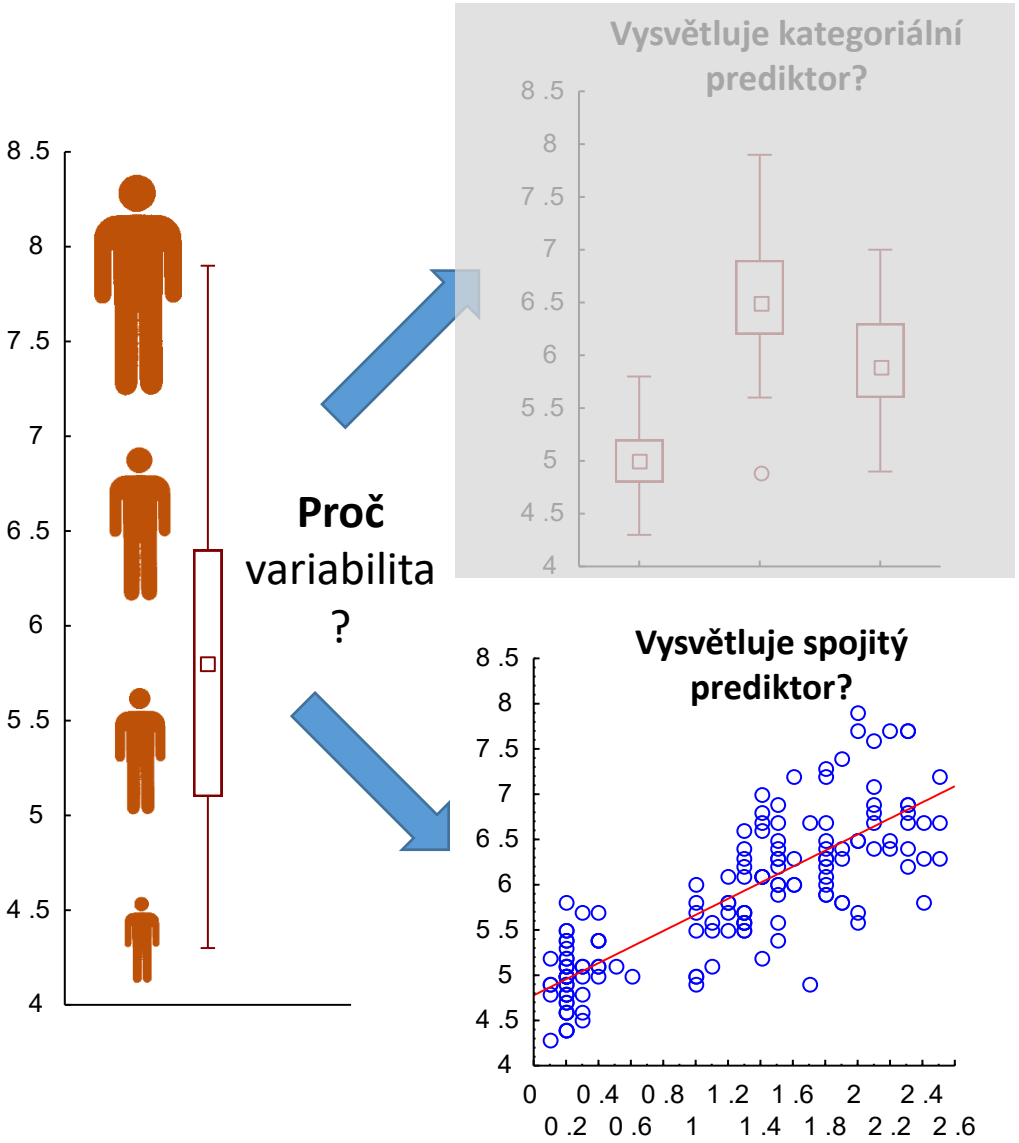
3. Aplikace modelu



• Individuální predikce stavu nenámých pacientů
• Model musí být podložen korektní statistikou a rozsáhlými daty

Cíl stochastického modelování

- Obecným cílem je snaha **vysvětlit variabilitu predikované proměnné** (endpoint, Y) pomocí **prediktorů** (vysvětlující proměnná, faktor, X)
- Jak predikovaná proměnná, tak prediktor mohou být různého typu
  - Binární
  - Kategoriální
  - Ordinální
  - Spojitá
  - Cenzorovaná (-> analýza přežití)
- Kombinace datového typu predikované proměnné a prediktoru určuje použitou metodu analýzy



Základy regresní analýzy

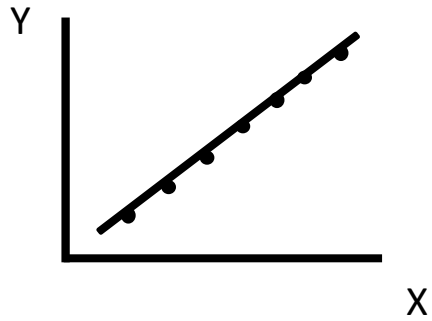
- Regrese - funkční vztah dvou nebo více proměnných

Jednorozměrná
y = f(x)

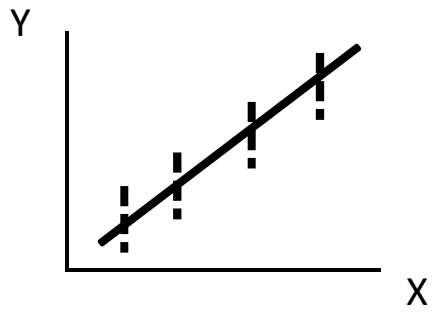
Vícerozměrná
y = f(x1, x2, x3, ……xp)

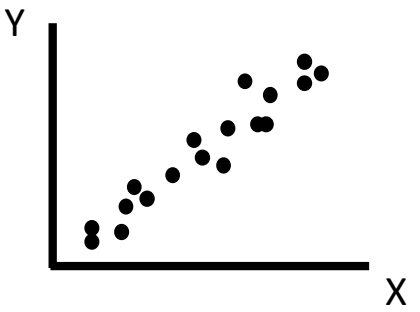
Vztah x, y

Deterministický



Regresní, stochastický





Pro každé x existuje pravděpodobnostní rozložení y

Lineární regrese I

$$Y = a + b \cdot x + e \quad \approx \quad \alpha + \beta \cdot X + \varepsilon$$

y

- $\alpha \approx a$ **(intercept):** $a = \bar{y} - b \cdot \bar{x}$
- $\beta \cdot X \approx b \cdot x$ **(sklon; slope)**
- $\varepsilon \approx e$ **- náhodná složka:** $\mathrm{N}(0; \sigma_e^2) = N(0; \sigma_y^2 x)$

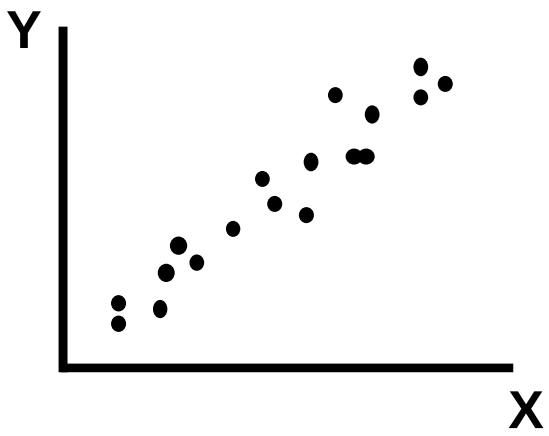
} Komponenty tvořící y se sčítají

ε - náhodná složka modelu přímky = rezidua přímky

$$\sigma_e^2 \left(\sigma_{y \cdot x}^2 \right) \Rightarrow \textbf{rozptyl reziduí}$$

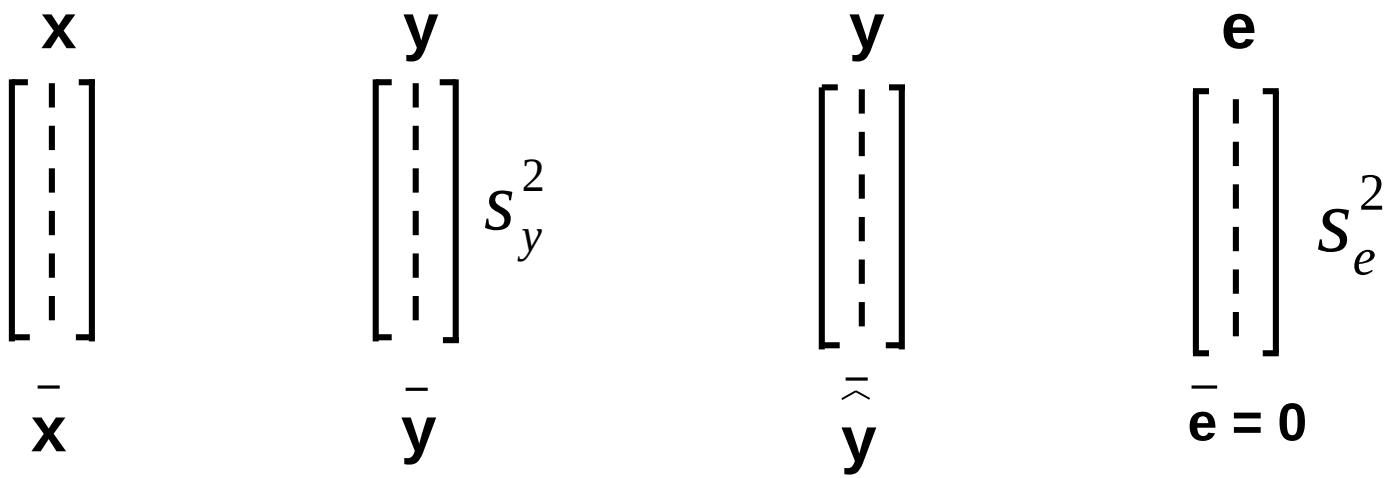
Lineární regrese II

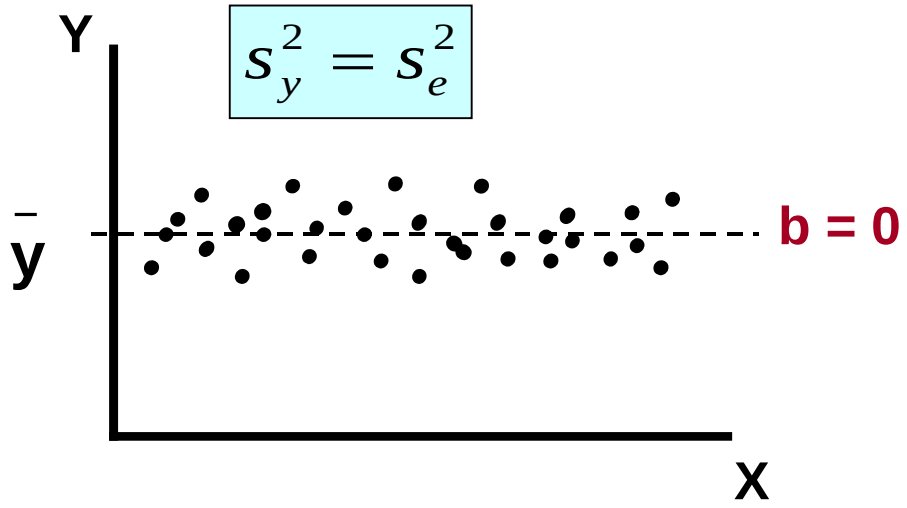
$$\begin{matrix}1\\ \vdots\\ n\end{matrix}\quad \begin{matrix}x\\ \begin{bmatrix}\vdots\end{bmatrix}\end{matrix} \qquad \begin{matrix}1\\ \vdots\\ n\end{matrix}\quad \begin{matrix}y\\ \begin{bmatrix}\vdots\end{bmatrix}\end{matrix}$$

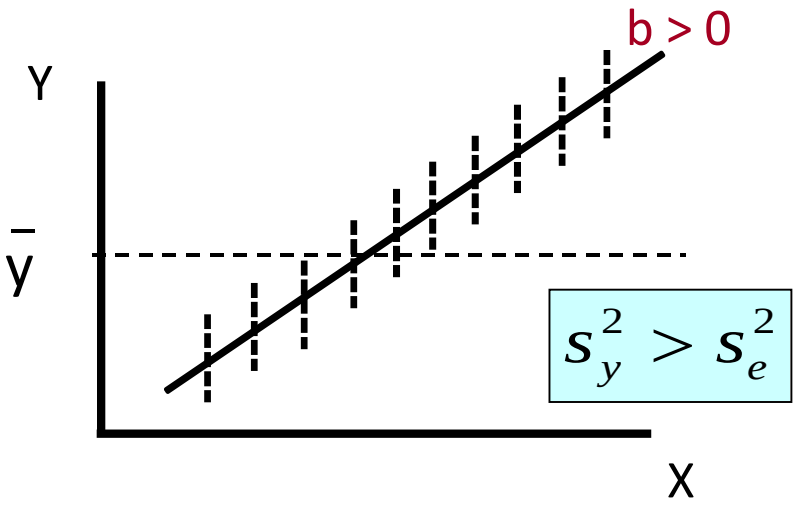


$$\begin{matrix}1\\ \vdots\\ n\end{matrix}\quad \begin{matrix}\hat{y}\\ \begin{bmatrix}\vdots\end{bmatrix}\end{matrix} = a + b \cdot \begin{matrix}x\\ \begin{bmatrix}\vdots\end{bmatrix}\end{matrix} \rightarrow \begin{matrix}y\\ \begin{bmatrix}\vdots\end{bmatrix}\end{matrix} - \begin{matrix}\hat{y}\\ \begin{bmatrix}\vdots\end{bmatrix}\end{matrix} = \begin{matrix}e\\ \begin{bmatrix}\vdots\end{bmatrix}\end{matrix}$$

Lineární regrese III

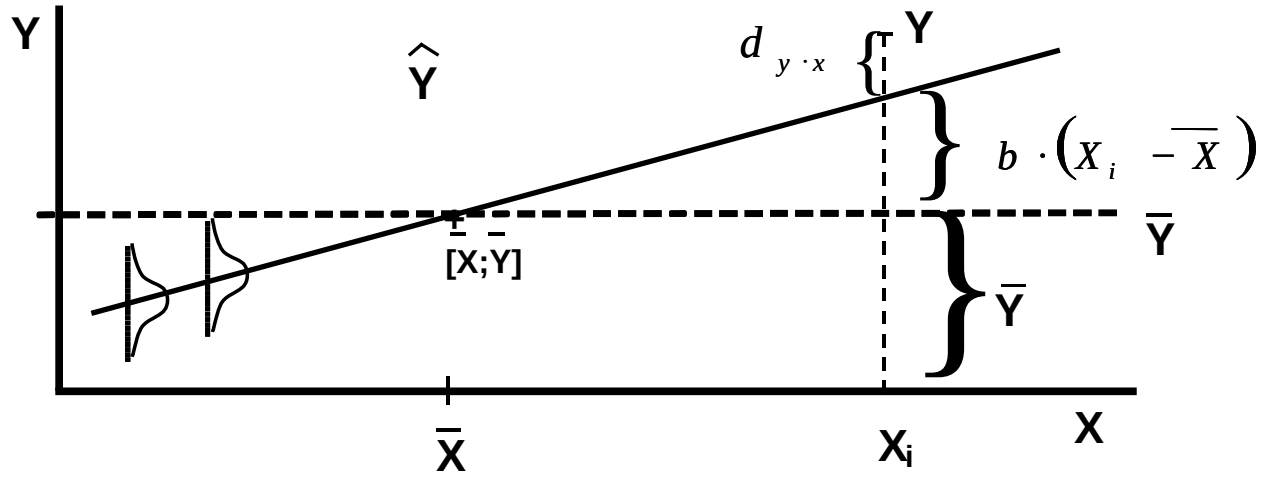






Lineární regrese IV

- Metoda nejmenších čtverců
  - X: Pevná, nestochastická proměnná
  - Rozložení hodnot y pro každé x je normální
  - Rozložení hodnot y pro každé x má stejný rozptyl
  - Rezidua jsou navzájem nezávislá a mají normální rozložení



$$d_{y\cdot x} = y - \hat{y} \qquad \boxed{d_{y\cdot x} = y - y - b(X_i - \bar{X})} \qquad \hat{y} = \bar{y} + b(X_i - \bar{X})$$

Smysl proložení přímky
minimalizace odchylek

$$d^2_{y\cdot x} \rightarrow \sum \left[y - \hat{\alpha} - \hat{\beta}(X_i - \bar{X}) \right]$$

Lineární regrese V

I. $b \sim \beta$: $b = \dfrac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$ $\qquad S_b^2 \sim \sigma_\beta^2$: $\dfrac{1}{\sum (X_i - \bar{X})^2} \cdot S_{y \cdot x}^2$

$S_{y \cdot x}^2 =$ **mean squared deviation from regression**

$S_{y \cdot x} =$ **sample standard deviation from regression**

$$S_{y \cdot x}^2 = \frac{\sum d_{y \cdot x}^2}{n-2} = \frac{\sum Y_i^2 - \dfrac{\sum Y_i^2}{n} - b^2 \cdot \sum (X_i - \bar{X})^2}{n-2}$$

II. $a \sim \alpha$: $a = \bar{Y} - b \cdot \bar{X}$ $\qquad S_a^2 \sim \sigma_\alpha^2$ $\qquad S_\alpha^2 = \left[\dfrac{1}{n} + \dfrac{\bar{X}^2}{\sum X^2} \right] \cdot S_{y \cdot x}^2$

intercept

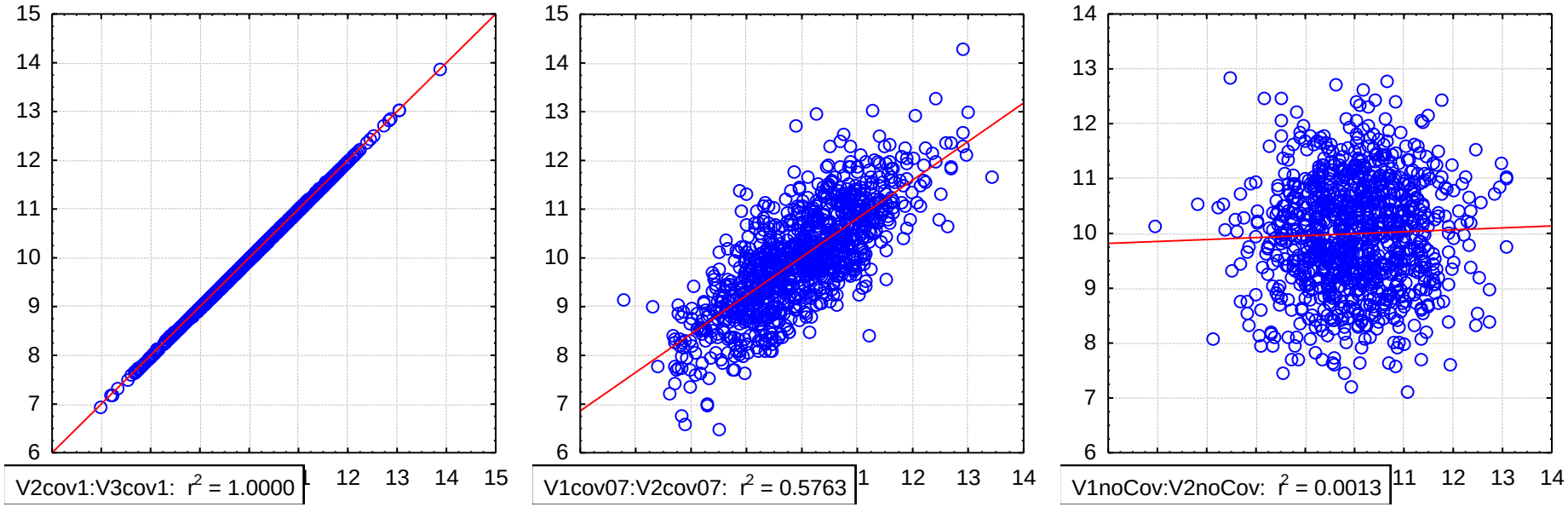
III. $\hat{Y}$: **modelová hodnota**

$$\widehat{Y}_i = a - b \cdot X_i$$

$$S_{\hat{y}_i} = (S_{y \cdot x}) \cdot \sqrt{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum X^2}}$$

Vyčerpaná variabilita a její statistická významnost

- Základním ukazatelem kvality modelu je množství varibility, které je modelem vysvětleno
- Obecně se značí R^2 a uvádí se v procentech nebo podílu celkové variability (v případe lineární regrese jde o Pearsonův korelačnín koeficient na druhou)
- Statistickou významnost vyčepané variability je možné testovat pomocí analýzy rozptylu



Analýza rozptylu v regresi

- Výpočet statistické významnosti rozptylu vyčerpaného regresním modelem

Celková ANOVA → SS_B/SS_T (variance ratio)

→ $MS_B/MS_E = F$

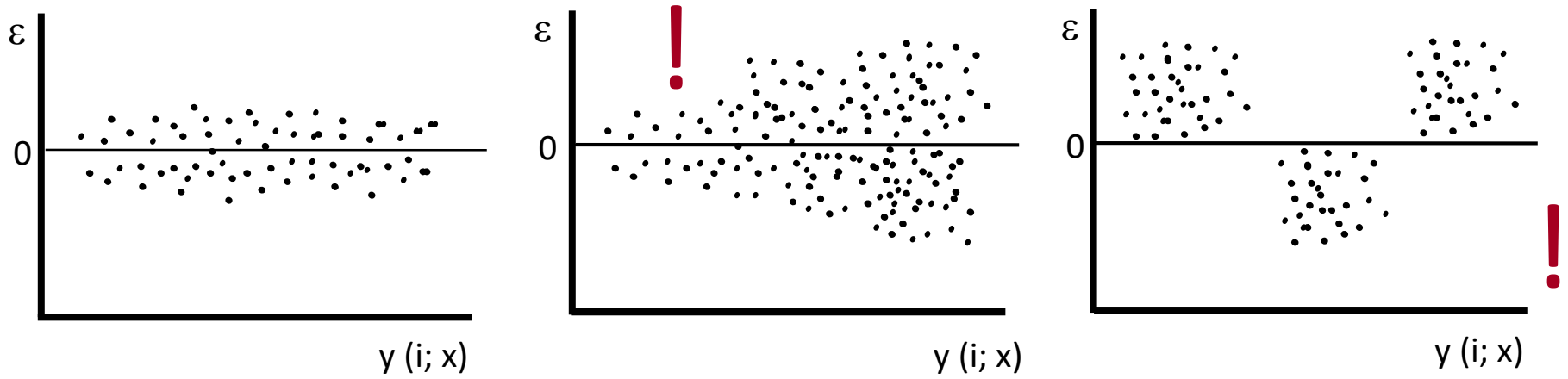
Analýza rozptylu regresního modelu (zde přímky)

Zdroj rozptylu	st.v.	SS	MS	F
Model (přímka)	1	SS_{MOD}	MS_{MOD}	MS_{MOD} / MS_R
Residuum	na - 2	SS_R	MS_R	
celkem	na - 1	SS_T		

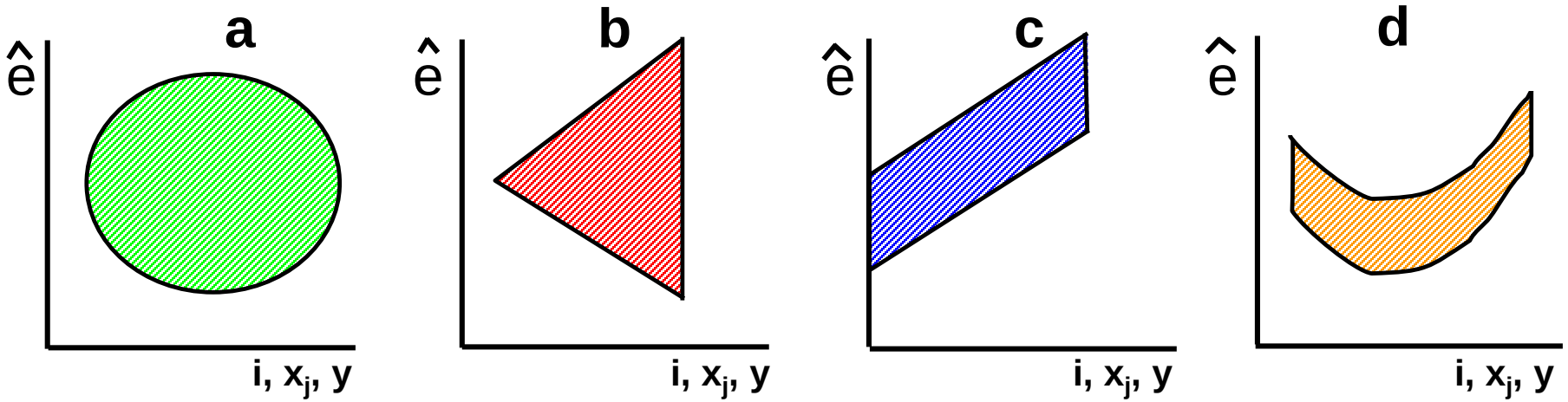
$(SS_{MOD}/SS_T) . 100$ = % rozptylu Y "vyčerpaného" přímkou = koeficient determinace (R^2)

Lineární regrese: analýza reziduí

Grafy residuí modelů (příklady)

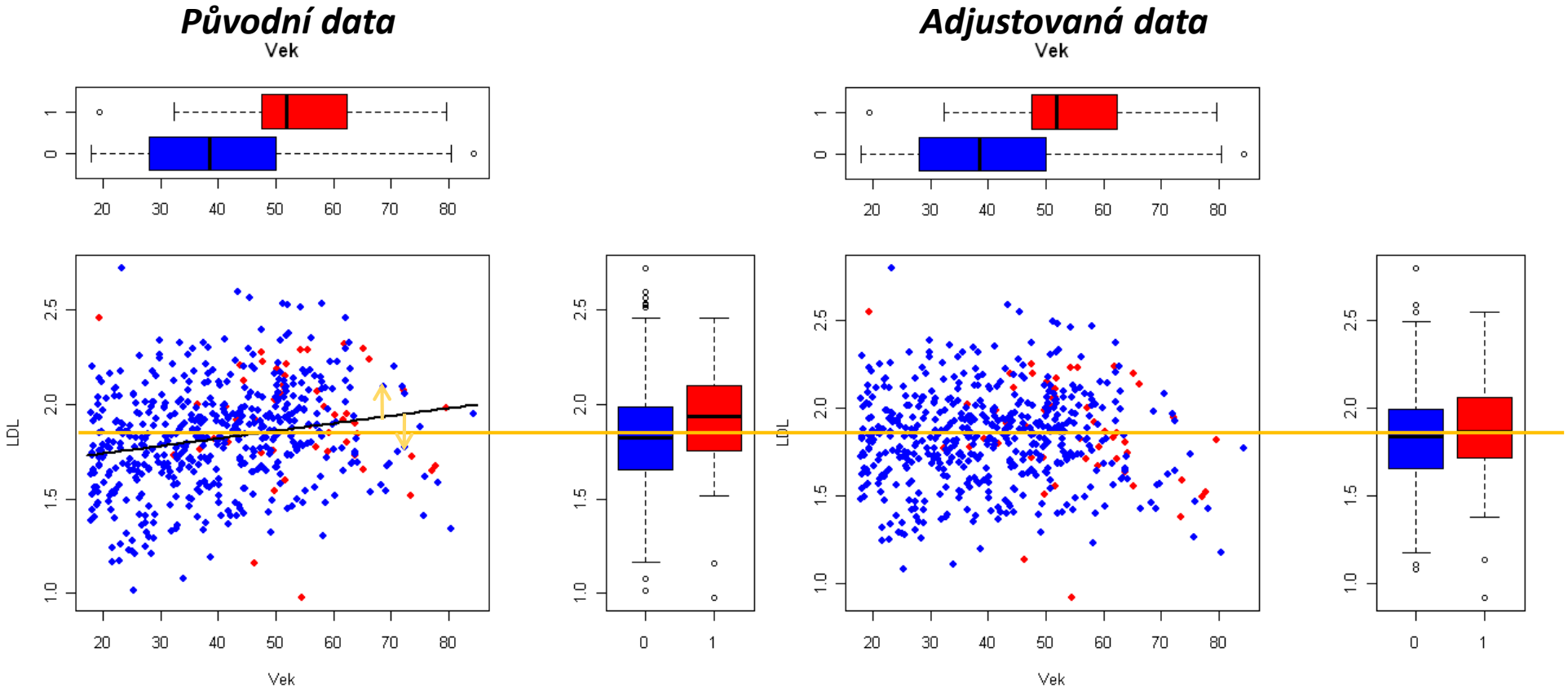


Obecné tvary residuí modelů (schéma)



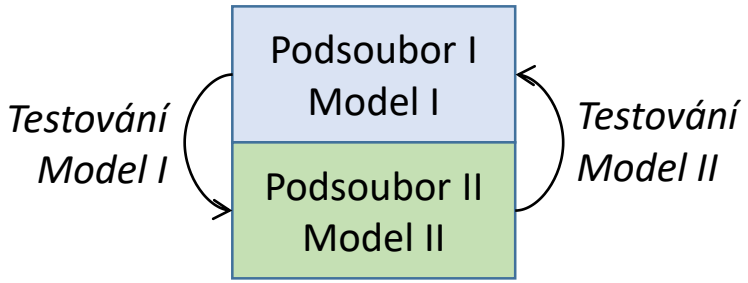
Adjustace proměnných na vliv jiných proměnných

1. V prvním kroku definujeme regresní model vztahu věku a adjustovaného parametru
2. Pro každého pacienta je vypočteno jeho reziduum od regresní přímky
3. Reziduum (představující hodnotu parametru po odečtení vlivu věku, jeho průměr je 0) je přičteno k průměrné hodnotě parametru
4. Výsledná adjustovaná hodnota má odečten vliv věku, ale zároveň není změněna číselná hodnota parametru



Ověření modelu na nezávislém souboru

- Při tvorbě modelů může dojít k problému, kdy vytvořený model je perfektně „vycvičen“ řešit danou úlohu na datovém soubor na němž byla vytvořena
- Z tohoto důvodu je problematické testovat výsledky modelu na stejném souboru, na němž byla vytvořena -> jde o důkaz kruhem
- Řešením je testování výsledků modelu na souboru se známým výsledkem (zde známým zařazením objektů do skupin), který se nepodílel na definici modelu
  - Krosvalidace
    - datový soubor je náhodně rozdělen na několik podsouborů (2 nebo více)
    - Na jednom podsouboru je vytvořen model a jeho výsledky testovány na zbývajících podsouborech
    - Výpočet je proveden postupně na všech podsouborech
  - One out leave out
    - Model je vytvořen na celém souboru bez jednoho objektu
    - na tomto objektu je model testován
    - postup je zopakován pro všechny objekty
  - Permutační metody
    - Jackknife, bootstrap – model je postupně vytvářen na náhodných podvýběrech souboru a testován na zbytku dat



Statistická významnost vs. praktické využití modelu

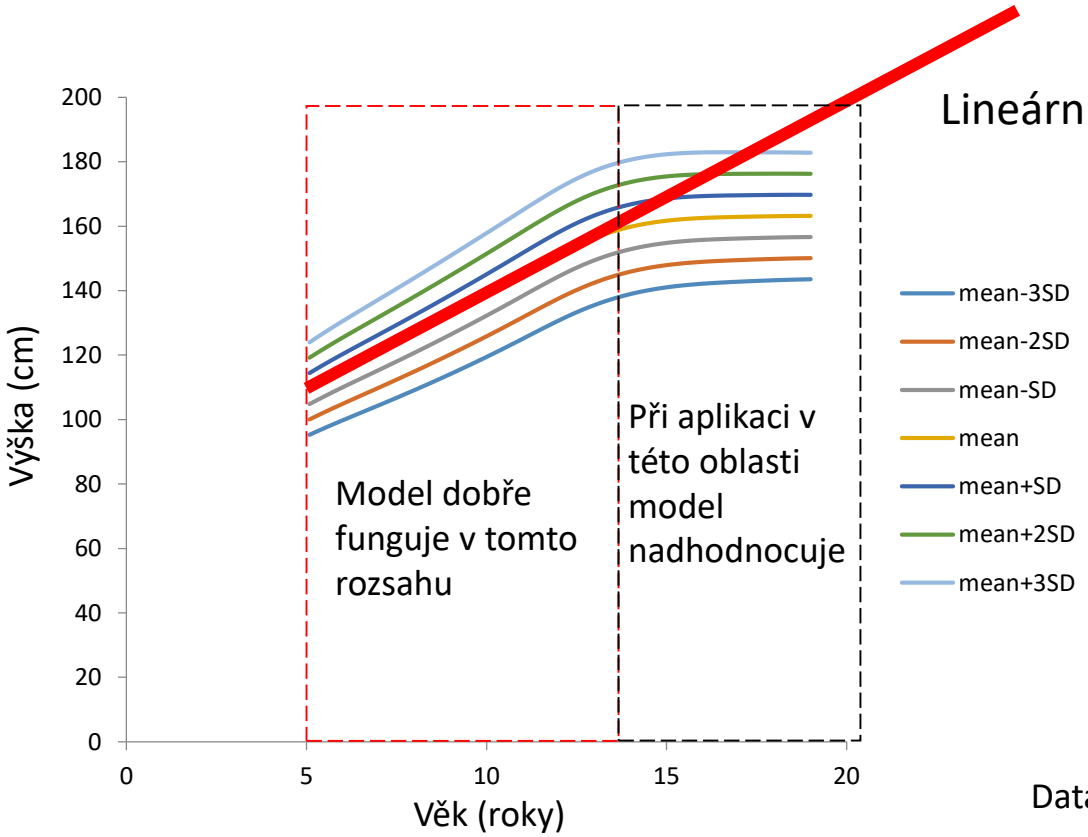
- Při aplikaci modelu v praxi je třeba zohlednit jak zjištěné statistické významnosti, tak praktický význam výstupů modelu
- Jde o analogii k statistické vs. praktické významnosti rozdílů např. v t –testu

- Statistická významnost = vztah mezi proměnnými, rozdíl mezi skupinami není pouhá náhoda (respektivě je dostatečně nízká pravděpodobnost, že nejde o náhodu)

- Praktický význam modelu
  - Z hlediska prediktorů: změna predikované hodnoty při změně prediktoru je prakticky významná (např. velikost nárůstu krevního tlaku při změně věku o 10 let)
  - Z hlediska objektů: Individuální predikce pacienta je dostatečně přesná aby byla prakticky využitelná (predikce různých událostí – hospitalizace, úmrtí, vznik komplikací, výsledek léčby atd.)

Rozsah aplikovatelnosti modelu

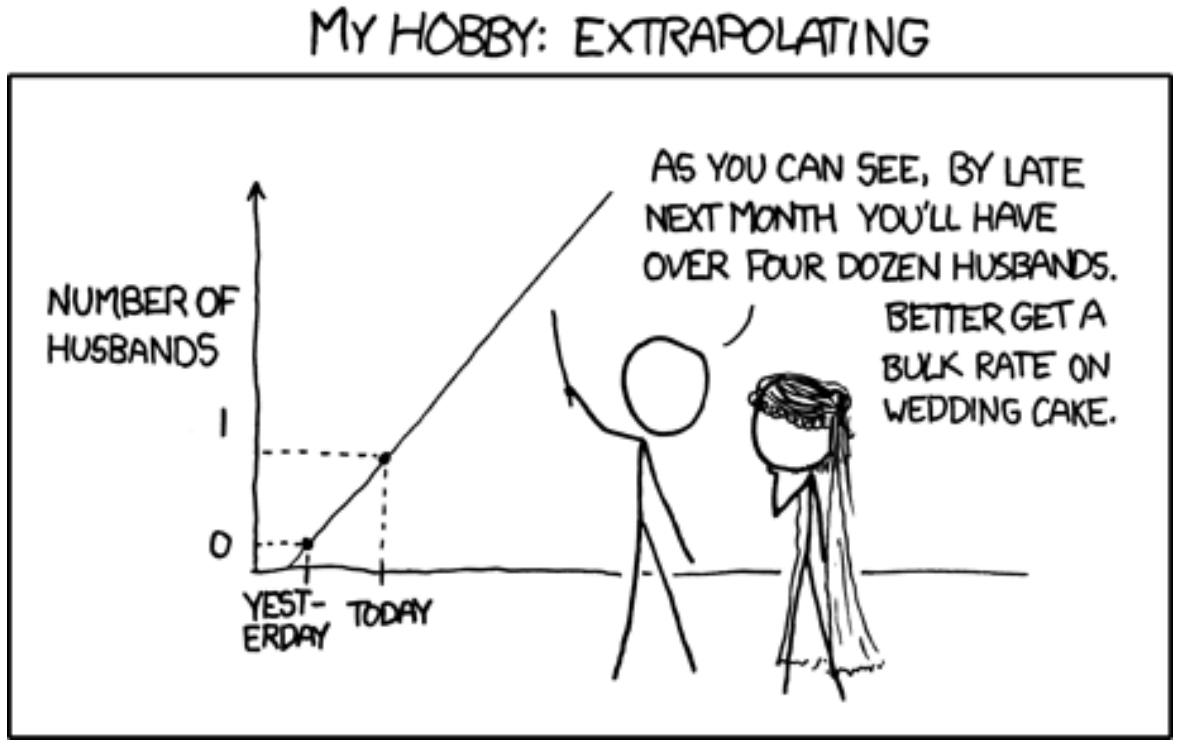
- Modely je možné aplikovat pouze v rozsahu prediktorů, na nichž byly vyvinuty
- Důvodem je naše neznalost chování vztahů mezi prediktory a predikovanou proměnnou mimo hranice v nichž byl model definován (typickými příklady jsou např. křivky dávka-odpověď, růst dětí v závislosti na věku, růst baktérií v závislosti na substrátu apod.)



Lineární model odvozený z části dat

Data: WHO Growth reference 5-19 years

Rozsah aplikovatelnosti modelu: příklad



Obecné zásady tvorby predikčních modelů

- Požadavky na kvalitní predikční model
  - Maximální predikční síla
  - Maximální interpretovatelnost
  - Minimální složitost
- Tvorba modelů
  - Neobsahuje redundantní proměnné
  - Je otestován na nezávislých datech
- Výběr proměnných
  - Algoritmy typu dopředné a zpětné eliminace jsou pouze pomocným ukazatelem při výběru proměnných finálního modelu
  - Při výběru proměnných se uplatní jak klasické statistické metody (ANOVA), tak expertní znalost významu proměnných a jejich zastupitelnosti

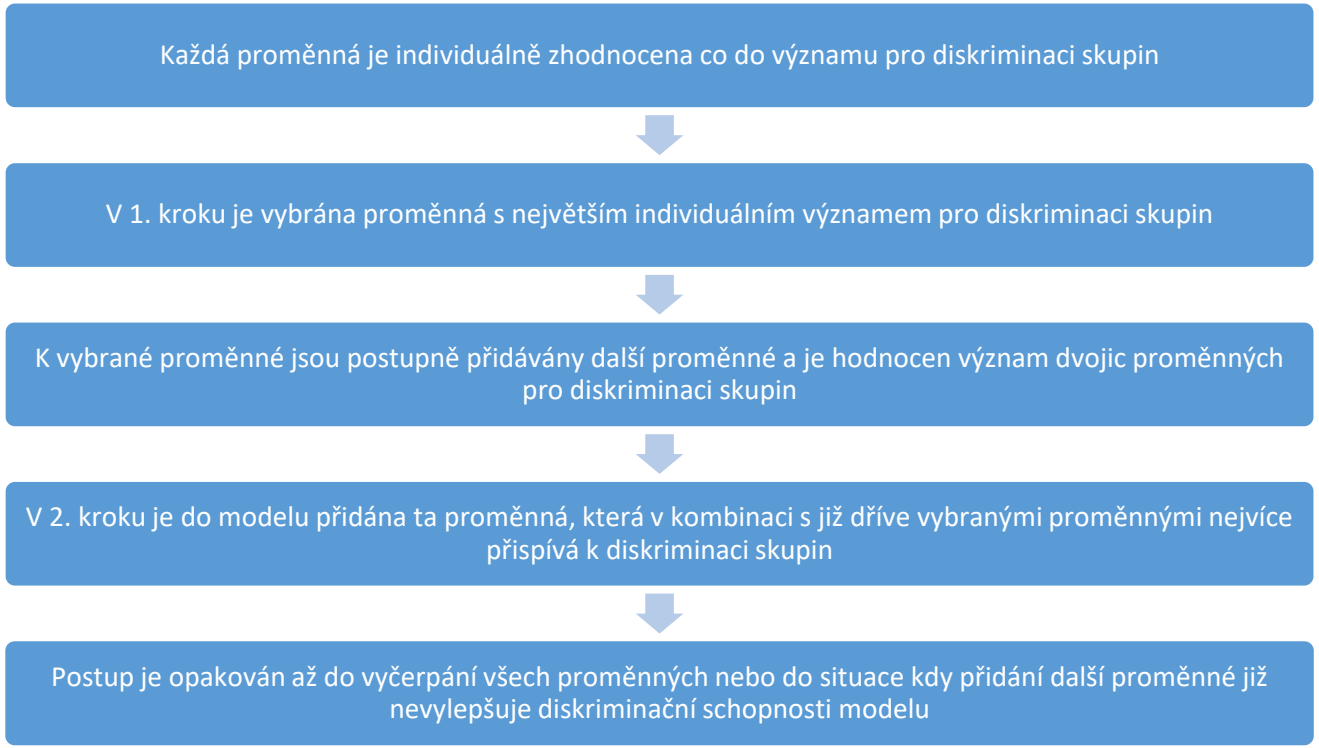
Dopředná a zpětná eliminace

- Dopředná a zpětná eliminace proměnných z modelu (forward, backward stepwise) je obecná technika používaná při tvorbě regresních, diskriminačních a jiných modelů
- Proměnné jsou do modelu postupně přidávány (ubírány) podle jejich významu v modelu

Schéma dopředné eliminace proměnných v modelu

V případě zpětné eliminace začíná proces od modelu se všemi proměnnými a postupně jsou vyřazovány proměnné s nejmenším příspěvkem k diskriminační síle modelu

Proces je třeba expertně kontrolovat, riziková je např. přítomnost redundantních proměnných



Kroky regresní analýzy

- Regresní analýza (a obecně i jiné stochastické modely) by měla probíhat v následujících krocích
  1. Ověření obecných předpokladů – normalita dat, linearita vztahu
  2. Výpočet modelu
  3. Analýza reziduí modelu umožňující ověřit vhodnost aplikace lineárního nebo jiného modelu
  4. Analýza vyčepané variability testující, zda model variabilitu dat významně vysvětluje
  5. Testování regresních koeficientů
     1. Posouzení významnosti komponent modelu
     2. Praktická smysluplnost modelu
  6. Závěr o využitelnosti a smysluplnosti modelu

Predikce binárních endpointů

ROC analýza

Logistická regrese

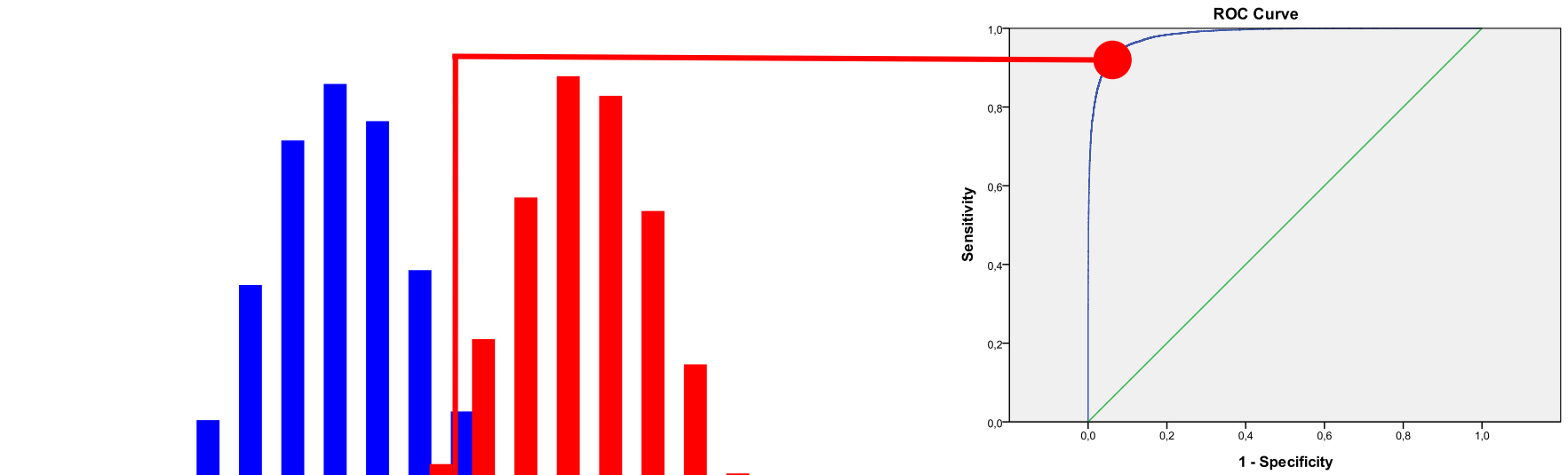
ROC analýza

- Nástroj pro identifikaci cut-off (hranice rozdělení spojitých dat) ve spojitých datech vzhledem k co nejlepšímu odlišení binárního endpointu
- Výsledkem je binarizace spojité proměnné, která je často lépe interpretovatelná než výsledky na spojitých datech
- Identifikace konkrétního cut-off souvisí s preferencí buď sensitivity nebo specificity pro identifikaci endpointu
- Upřednostnění sensitivity nebo specificity je do určité míry subjektivní dle reálného cíle analýzy
  - Vysoká sensitivita – screeningový test, kdy je třeba zachytit všechny možné nemocné (např. závažné onemocnění, které je třeba zachytit v počátečním stadiu)
  - Vysoká specificita – pokud je nezbytné odchytit pouze skutečně nemocné pacienty (např. nechceme vystavovat pacienty zbytečné léčbě málo závažného onemocnění)

ROC analýza

- Identifikace cutt offs pro kategorizaci spojitých proměnných aby při jejich užití v modelech byla maximalizována jejich sensitivita a specificita

Kde leží optimální hranice mezi skupinami?



Identifikace hranice s nejvyšší sensitivitou a specificitou pro odlišení skupin

Sensitivita a specificita

- Klíčové pojmy v popisu vztahu dvou binárních proměnných = situace kdy predikujeme binární endpoint binárním prediktorem

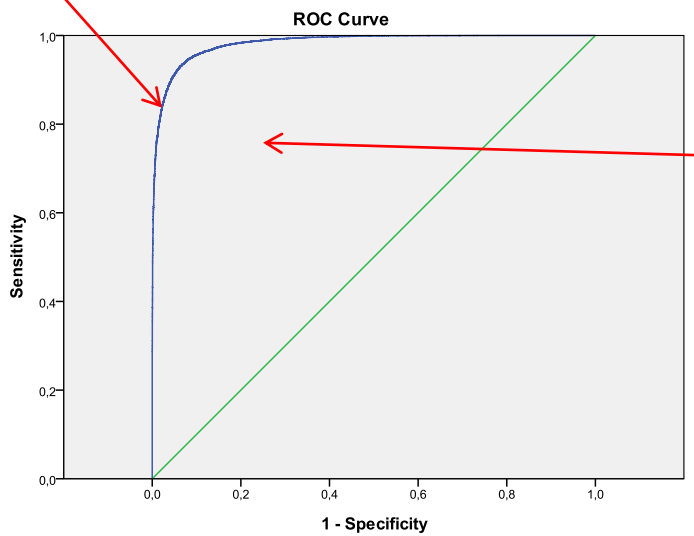
	1 – nemocný	0 - zdravý
1 – riziková skupina	Skutečně pozitivní	Falešně pozitivní
0 – neriziková skupina	Falešně negativní	Skutečně negativní

$$sensitivita = \frac{skutečně\ pozitivní}{skutečně\ pozitivní + falešně\ negativní}$$

$$specificita = \frac{skutečně\ negativní}{skutečně\ negativní + falešně\ pozitivní}$$

Výstupy ROC

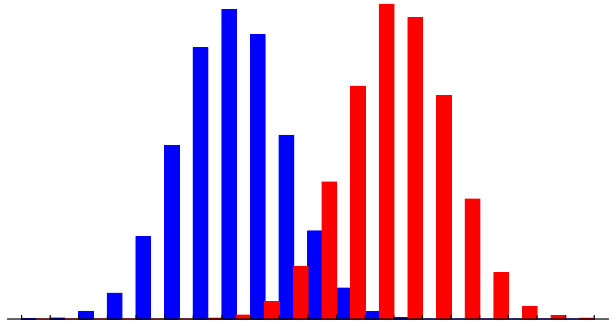
- **Sensitivita a specificita v každém bodě křivky – mohou být doplněny o IS**
- Nejlepší kombinace sensitivity a specificity určuje příslušný dělící bod spojité proměnné
- Při identifikaci cut-off je třeba také kontrolovat, aby výsledná riziková skupina neobsahovala pouze minimum hodnot (cut-off oddělující jednoho pacinta nemá téměř smysl)

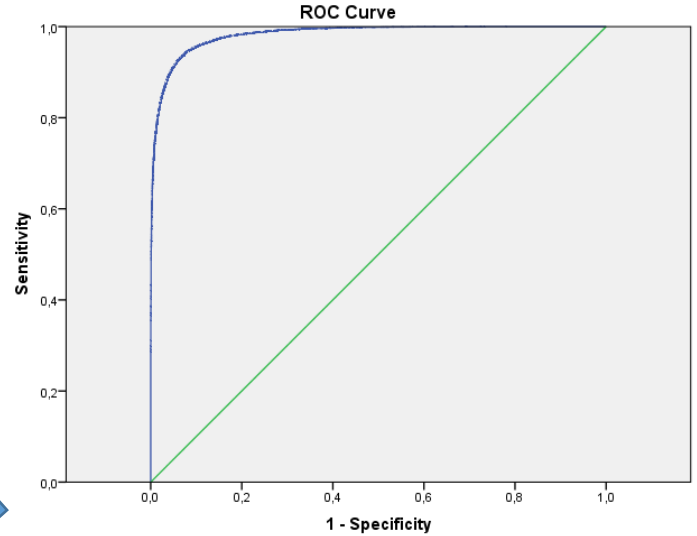


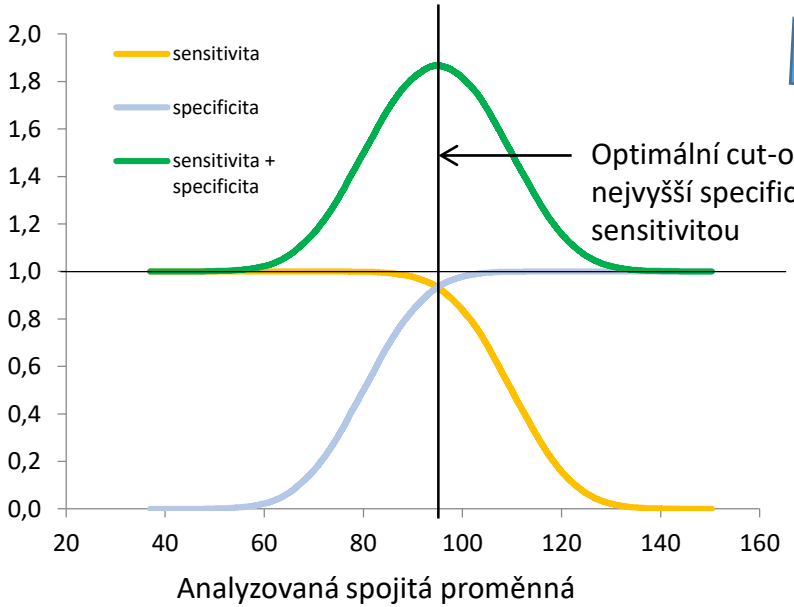
AUC (plocha pod křivkou) + IS
Čím odlišnější od 0.5, tím lepší identifikace endpointu
Testování významnosti AUC

ROC – příklad

Odlišení dvou skupin pacientů
(modří=zdraví; červení=nemocní)







Optimální cut-off s nejvyšší specificitou a sensitivitou

Area Under the Curve

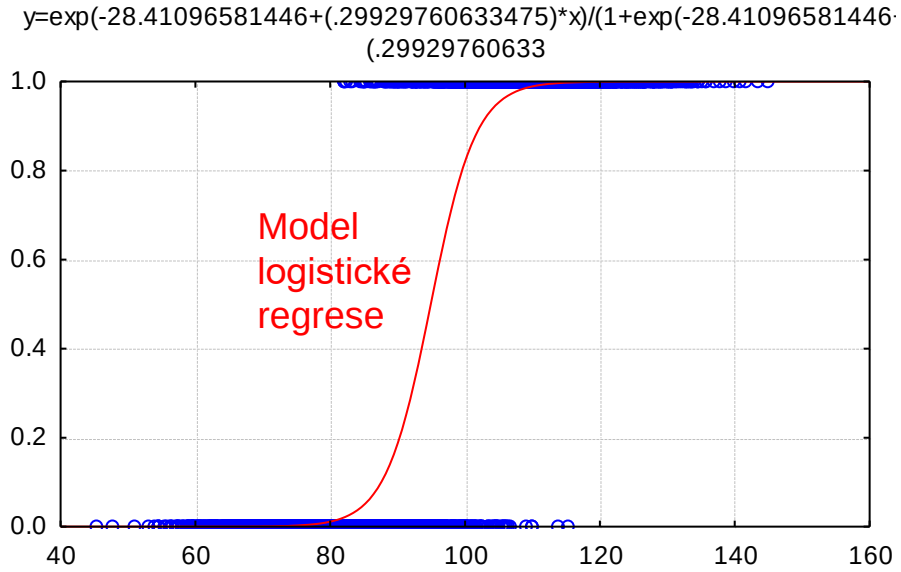
Test Result Variable(s):Var1

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
,983	,000	,000	,982	,984

a. Under the nonparametric assumption
b. Null hypothesis: true area = 0.5

Logistická regrese

- Logistická regrese je základním nástrojem pro analýzu závislosti binárního endpointu (úmrtí, komplikace, výskyt taxonu, příslušnost do kategorie atd.) na spojitých nebo binárních prediktorech
- Cílem analýzy je:
  - Identifikace vztahů mezi prediktory a endpointem a jejich popis (odds ratio)
  - Vytvoření predikčního modelu umožňujícího zařazení pacientů do hodnocených skupin
- Logistická regrese patří do skupiny zobecněných lineárních modelů (lineární statistické modely s linkovací funkcí)



Příklad logistické regrese: predikce binární charakteristiky (osa y) za pomoci spojité proměnné (osa x)

Princip logistické regrese

- V logistické regresi modelujeme vliv spojitých nebo binárních prediktorů na endpoint s binomickým rozdělením - > není možné použít klasickou lineární regresi
- Predikujeme pravděpodobnost výskytu jevu pomocí rovnice:

$$P(x) = \frac{\exp(a + b * x)}{1 + \exp(a + b * x)}$$

- Kde $\frac{\exp(rovnice)}{1 + \exp(rovnoce)}$ je tzv. logit, linkovací funkce pro logistickou regresi a rovnice a+b*x je použitý lineární model
- Pojem linkovací funkce je spjat se zobecněnými lineárními modely, kdy linkovací funkce převádí problém nelineární závislosti y na x na lineární model
- Zjednodušeně řečeno „nelineární vztah=linkovací funkce(lineární model)“
- Zobecněný lineární model s linkovací funkcí „identita“ = lineární model

Odds ratio a logistická regrese

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	SEPALLEN	5,140	1,007	26,080	1	,000	170,773	23,748	1228,028
	Constant	-27,831	5,434	26,236	1	,000	,000		

a. Variable(s) entered on step 1: SEPALLEN.

- Popisuje míru rizika spjatou:
  - **U spojitých proměnných** se změnou hodnoty o 1 (z tohoto důvodu se spojité proměnné často převádí na interpretovatelné jednotky – např. věk po destiletích, koncentrace po stovkách jednotek)
  - **U binárních proměnných** spjatých s výskytem vlastnosti (kódováno jako 1)
    - U klasických dummies jde o riziko vůči všem ostatním pacientům bez dané vlastnosti
    - U binárních proměnných kódovaných vůči referenční kategorii jde o nárůst oproti pacientům v referenční kategorii
- Odds ratio je exponenciální hodnota koeficientu regresní rovnice

Logistická regrese: shrnutí

- Základní nástroj pro identifikaci faktorů ovlivňujících výskyt binárních endpointů a predikci individuální pravděpodobnosti výskytu endpointů
- Použitelná jako obdoba diskriminační analýzy pro 2 skupiny
- Popisuje míru rizikovosti prediktorů pro binární endpoint ve formě odds ratia

- Pro vícerozměrné modely je důležité analyzovat redundanci parametrů a stabilitu vícerozměrných modelů
- Pro praktické nasazení modelů je nezbytná jejich krosvalidace, popřípadě jiné metody testování nasazení modelů na nezávislých datech
- Neumí pracovat s cenzorovanými daty (analýza přežití)

- **Standardní metodika analýzy rizikových faktorů pro binární endpointy (výskyt něčeho – úmrtí, taxon atd.)**

Vícerozměrná analýza dat: úvod

Principy a využití vícerozměrné analýzy dat

Anotace

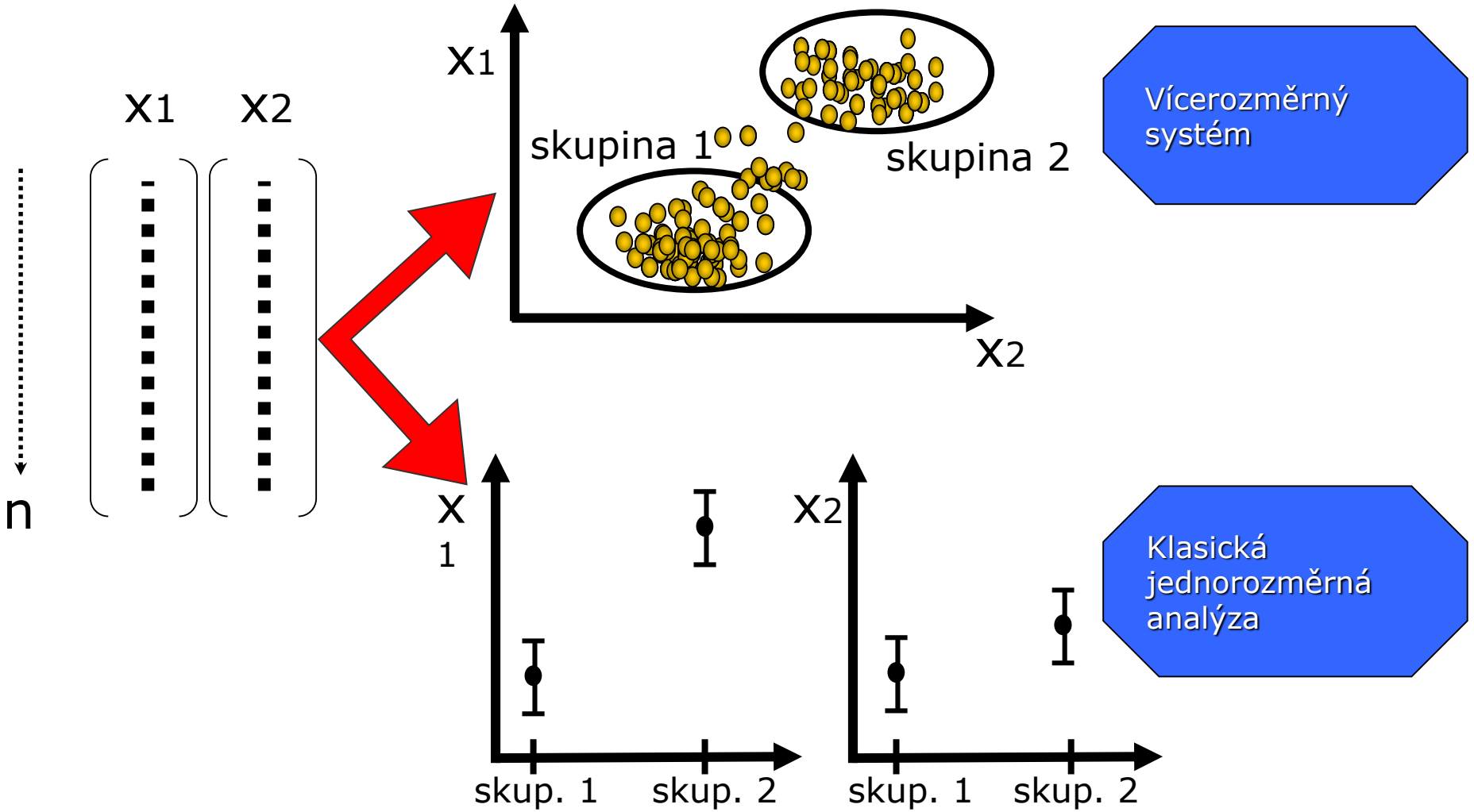
- Vícerozměrná analýza dat představuje nadstavbu nad klasickou, jednorozměrnou statistikou a je zvláště vhodná pro biologická a medicínská data, která jsou vícerozměrná již svou podstatou
- Při vícerozměrné analýze je nicméně nezbytné si uvědomit, že povětšinou vychází ze stejných principů jako jednorozměrné analýzy a tedy i zde je nezbytné dodržovat předpoklady na nichž je výpočet založen. Tento fakt je důležité si uvědomit zejména vzhledem k relativní dostupnosti vícerozměrných analýz v moderních statistických software.

Vztah klasické a vícerozměrné statistiky

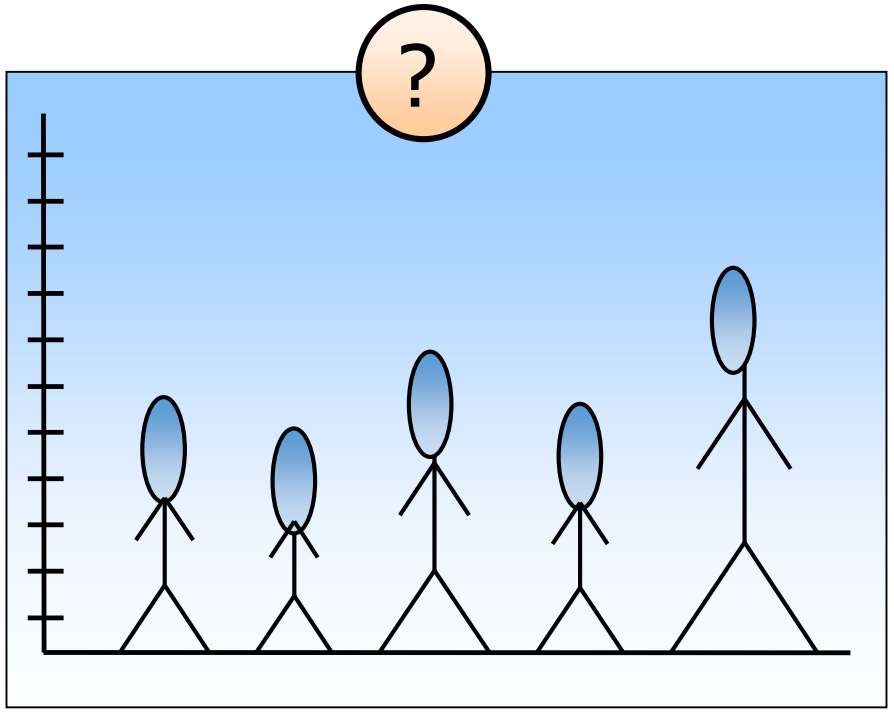
- Vícerozměrná analýza dat využívá přístupů klasické statistiky
- Zároveň je citlivá i na jejich problémy

- Agregace dat přes sumární statistiku nebo kontingenční tabulky – korespondenční analýza
- Korelace – analýza hlavních komponent, faktorová analýza, diskriminační analýza

Vícerozměrné vnímání skutečnosti - nová kvalita analýzy dat



Běžná sumarizace dat „likviduje“ individualitu jedince

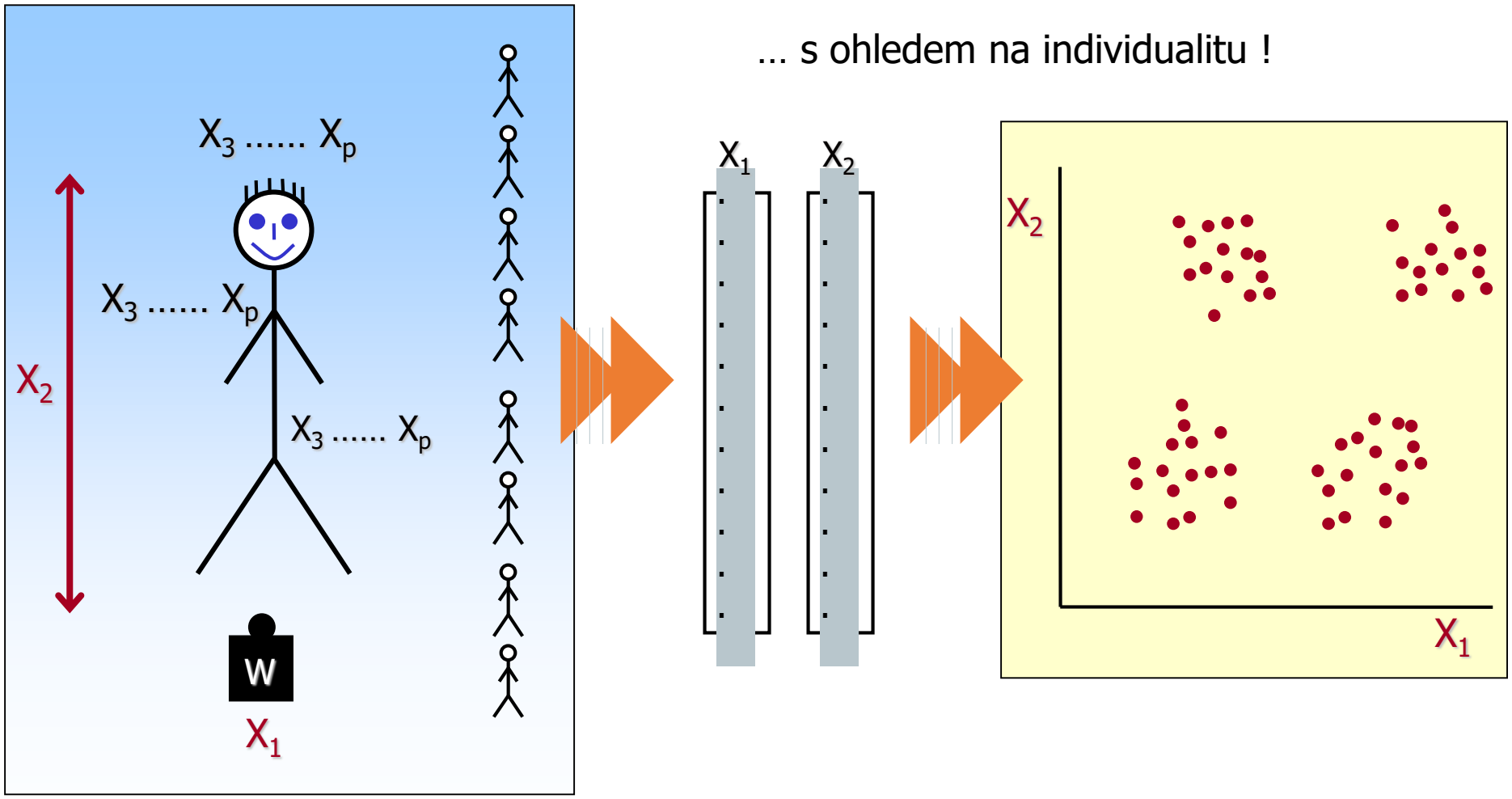


Průměr ± SE

BĚŽNÁ STATISTICKÁ SUMARIZACE

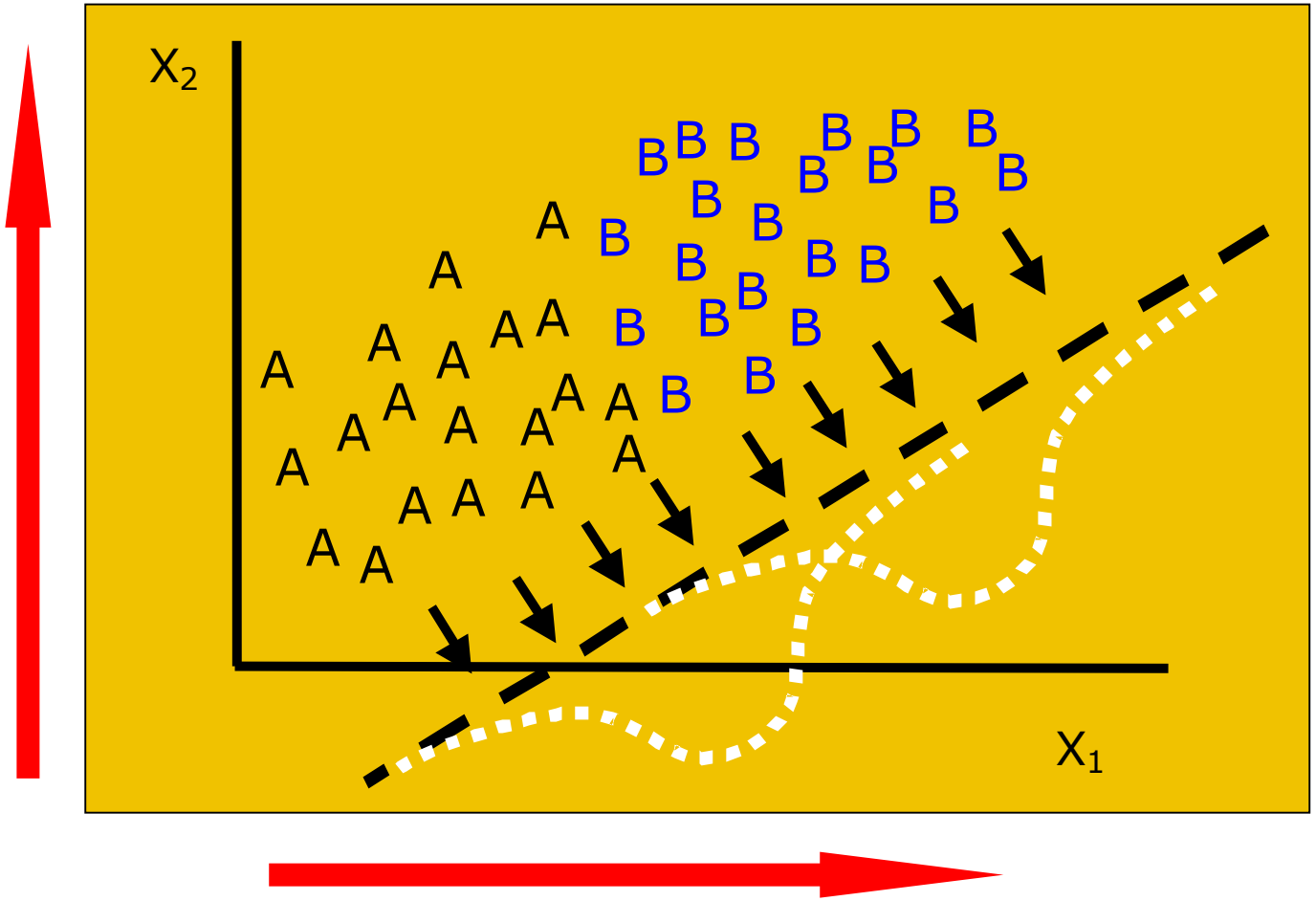
- ✓ Zpřehlednění dat
- ✓ Neodliší původní měření

Vícerozměrné hodnocení



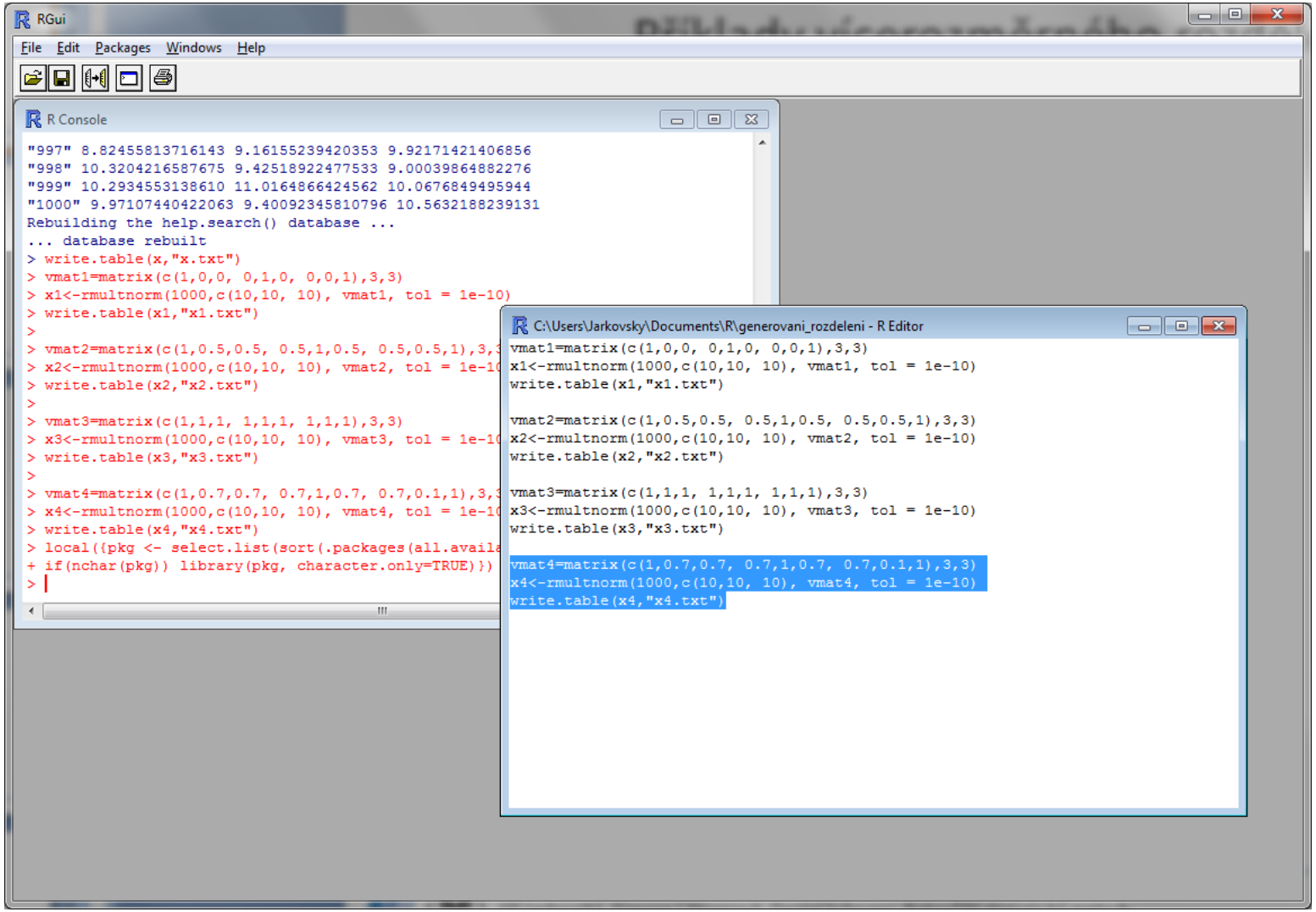
Vícerozměrné hodnocení – nová kvalita

Pouze kombinované parametry mají odpovídající informační sílu



Příklady vícerozměrného rozdělení

- R – knihovna MSBVAR



Vícerozměrné charakteristiky rozdělení

- Základní charakteristikou vícerozměrného rozdělení je vektor středních hodnot (vektor průměrů)

$$\mathrm{E}(\mathbf{X}) = \begin{bmatrix} \mathrm{E}(X_1) \\ \mathrm{E}(X_2) \\ \vdots \\ \mathrm{E}(X_p) \end{bmatrix}$$

- a kovariační matice

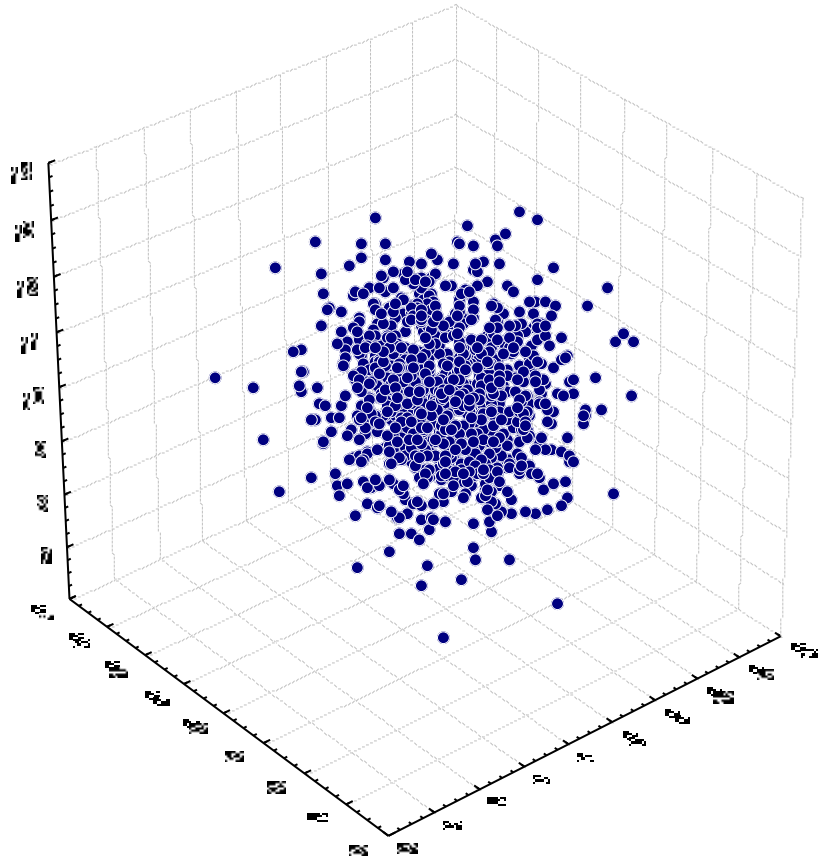
$$\boldsymbol{\Sigma} = \mathrm{var}(\mathbf{X}) = \mathrm{cov}(\mathbf{X}) = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2 & \cdots & \sigma_1\sigma_p \\ \sigma_2\sigma_1 & \sigma_2^2 & \cdots & \sigma_2\sigma_p \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_p\sigma_1 & \sigma_p\sigma_2 & \cdots & \sigma_p^2 \end{pmatrix}$$

- kde je σ_{ij} kovariance dvou náhodných veličin, tj.

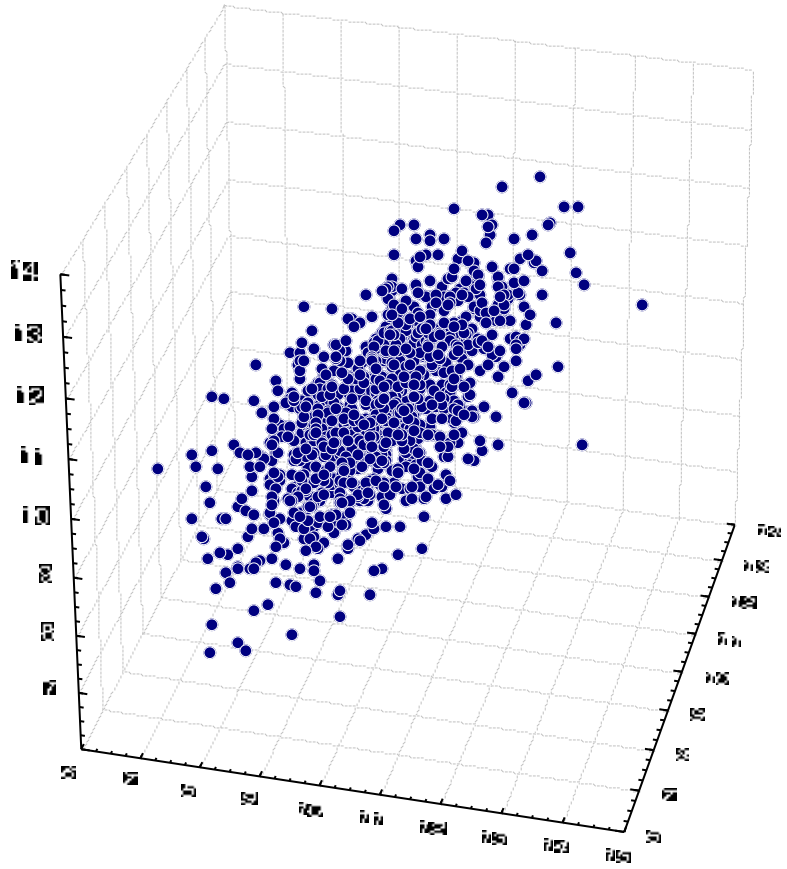
$$\sigma_{ij} = \mathrm{cov}(X_i, X_j) = \mathrm{E}(X_i - \mathrm{E}(X_i))(X_j - \mathrm{E}(X_j))$$

Příklad vícerozměrného rozdělení I

```
vmat1=matrix(c(1,0,0, 0,1,0, 0,0,1),3,3)
x1<-rmultnorm(1000,c(10,10, 10), vmat1, tol = 1e-10)
write.table(x1,"x1.txt")
```

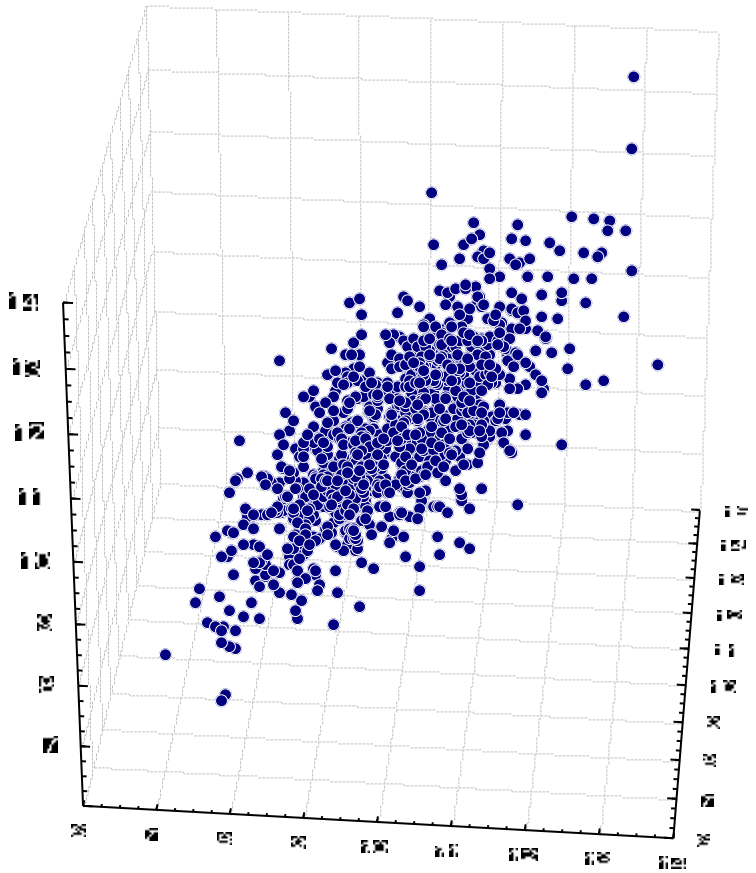


```
vmat2=matrix(c(1,0.5,0.5, 0.5,1,0.5, 0.5,0.5,1),3,3)
x2<-rmultnorm(1000,c(10,10, 10), vmat2, tol = 1e-10)
write.table(x2,"x2.txt")
```

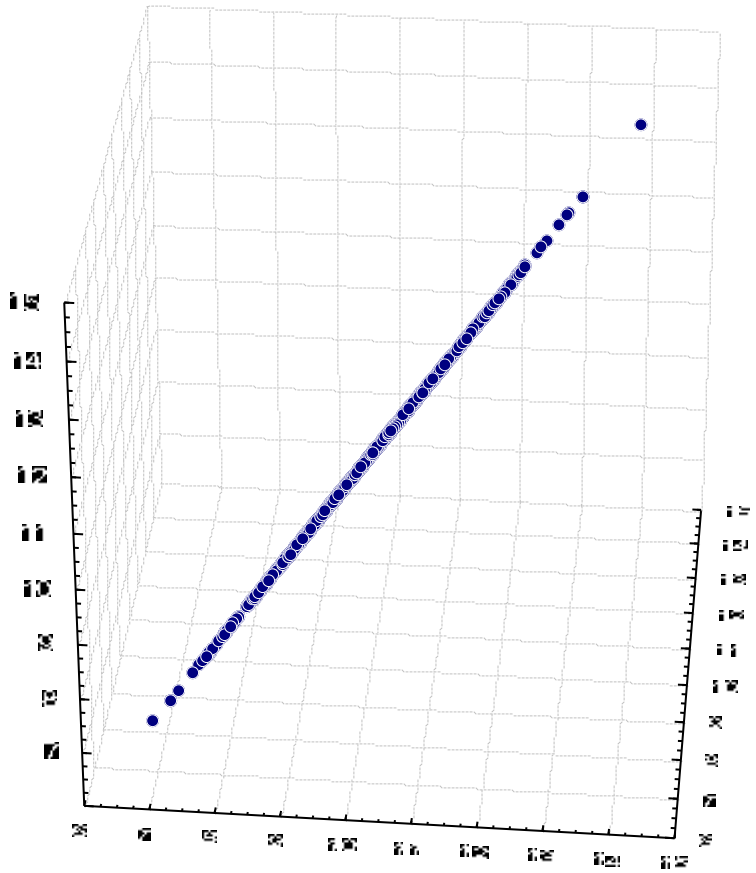


Příklad vícerozměrného rozdělení II

```
vmat4=matrix(c(1,0.7,0.7, 0.7,1,0.7, 0.7,0.1,1),3,3)
x4<-rmultnorm(1000,c(10,10, 10), vmat4, tol = 1e-10)
write.table(x4,"x4.txt")
```

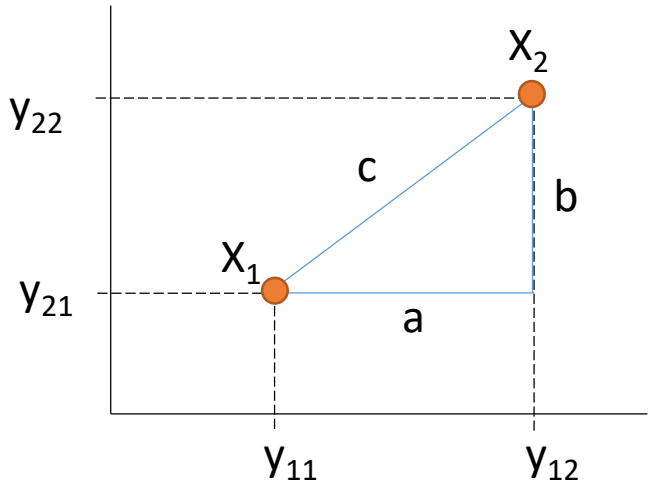


```
vmat3=matrix(c(1,1,1, 1,1,1, 1,1,1),3,3)
x3<-rmultnorm(1000,c(10,10, 10), vmat3, tol = 1e-10)
write.table(x3,"x3.txt")
```



Vícerozměrné hodnocení vychází z jednoduchých principů

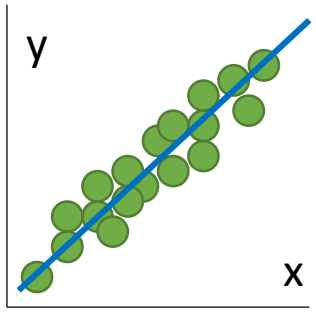
- Nejsnáze představitelným měřítkem vztahu dvou objektů ve vícerozměrném prostoru je jejich vzdálenost
- Nejjednodušším typem této vzdálenosti (bohužel s omezeným použitím na data společenstev) je Euklidovská vzdálenost vycházející z Pythagorovy věty



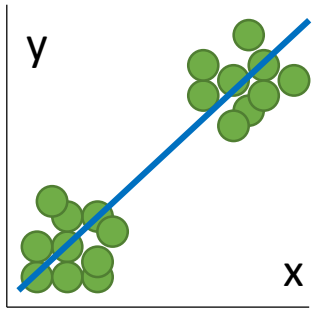
$$D_1(x_1, x_2) = \sqrt{\sum_{j=1}^{p} (y_{1j} - y_{2j})^2}$$

Korelace jako princip výpočtu vícerozměrných analýz

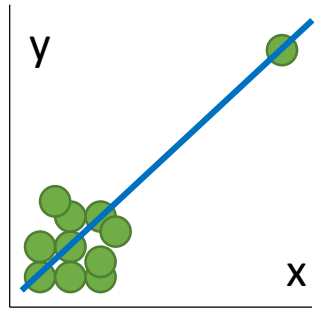
- Kovariance a Pearsonův korelační koeficient je základem analýzy hlavních komponent, faktorové analýzy jakož i dalších vícerozměrných analýz pracujících s lineární závislostí proměnných
- Předpokladem výpočtu kovariance a Pearsonova korelačního koeficientu je:
  - Normalita dat v obou dimenzích
  - Linearita vztahu proměnných
- Pro vícerozměrné analýzy je nejzávažnějším problémem přítomnost odlehlých hodnot



Lineární vztah – bezproblémové použití Pearsonova korelačního koeficientu



Korelace je dána dvěma skupinami hodnot – vede k identifikaci skupin objektů v datech



Korelace je dána odlehlou hodnotu – analýza popisuje pouze vliv odlehlé hodnoty

Analýza kontingenčních tabule jako princip výpočtu vícerozměrných analýz

- Abundance taxonů (nebo počet jakýchkoliv objektů) na lokalitách lze brát jako kontingenční tabulku a mírou vztahu mezi řádky (lokality) a sloupci (taxony) je velikost chi-kvadrátu

$$\chi^2_{(1)} = \frac{\left[\text{pozorovaná četnost} - \text{očekávaná četnost}\right]^2}{\text{očekávaná četnost}}$$

Počítáno pro každou buňku tabulky

	☠	☺
A	10	0
B	0	10

Pozorovaná tabulka

	☠	☺
A	5	5
B	5	5

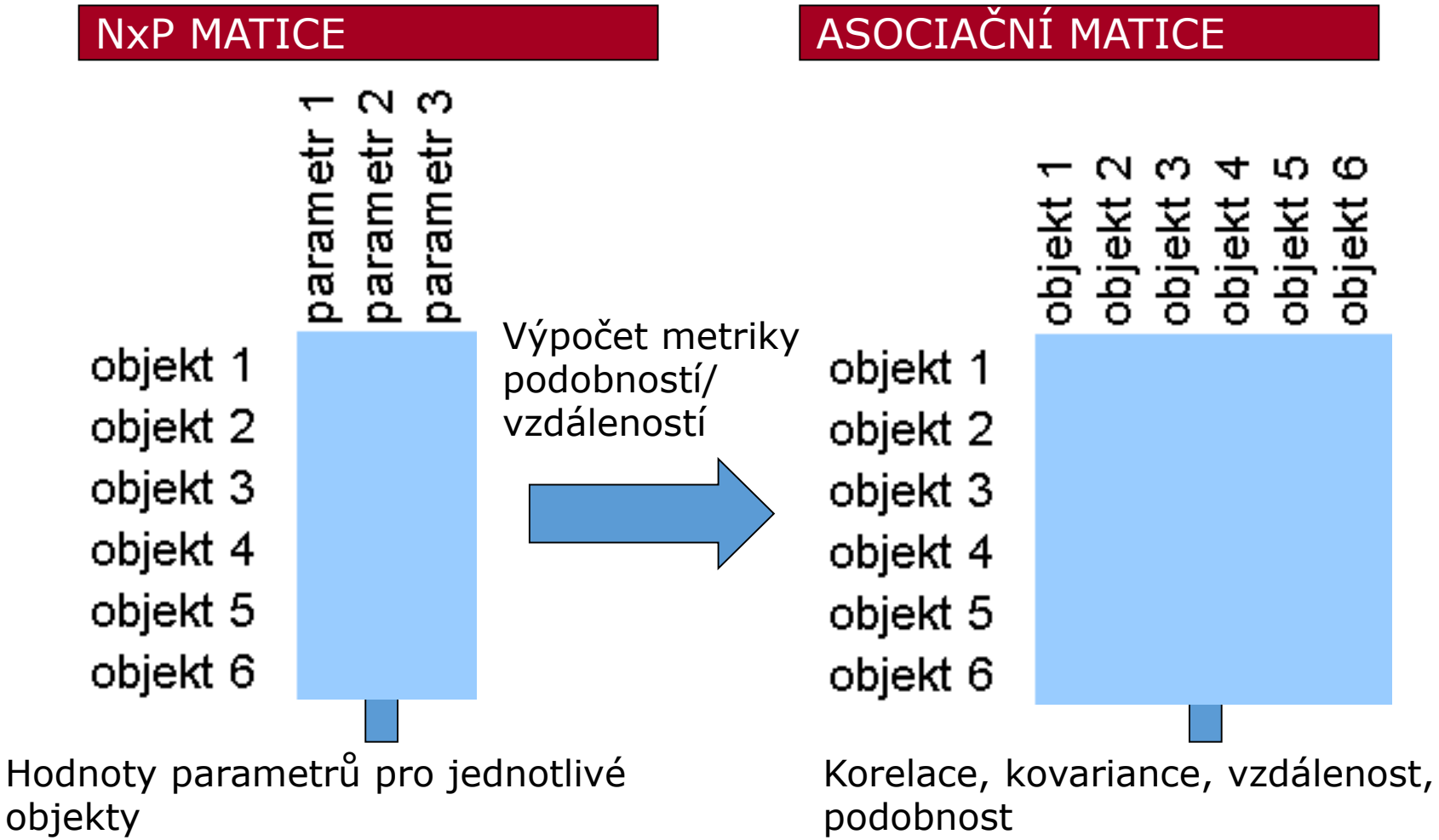
Očekávaná tabulka

Hodnota chi-kvadrátu definuje míru odchylky dané buňky (v našem kontextu vztahu taxon-lokalita) od situace, kdy mezi řádky a sloupci (taxon-lokalita) není žádný vztah

Pojmy vícerozměrných analýz

- Vícerozměrné metody: Název vícerozměrné vychází z typu vstupních dat, tato data jsou tvořena jednotlivými objekty a každý z nich je charakterizován svými parametry a každý z těchto parametrů můžeme považovat za jeden rozměr objektu.
- Maticová algebra: Základem práce s daty a výpočtů vícerozměrných metod je maticová algebra, matice tvoří jak vstupní, tak výstupní data a probíhají na nich výpočty.
- NxP matice: N objektů s p parametry pak vytváří tzv. NxP matici, která je prvním typem vstupu dat do vícerozměrných analýz.
- Asociační matice: Na základě těchto matic jsou počítány matice asociační na nichž pak probíhají další výpočty, jde o čtvercové matice obsahující informace o podobnosti nebo rozdílnosti (tzv. metriky) buď objektů (Q mode analýza) nebo parametrů (R mode analýza).Měřítko podobnosti se liší podle použité metody a typu dat, některé metody umožňují použití uživatelských metrik.

Vstupní matice vícerozměrných analýz



Hodnoty parametrů pro jednotlivé objekty

Korelace, kovariance, vzdálenost, podobnost

Základní typy vícerozměrných analýz

SHLUKOVÁ ANALÝZA

- vytváření shluků objektů na základě jejich podobnosti
- identifikace typů objektů

ORDINAČNÍ METODY

- zjednodušení vícerozměrného problému do menšího počtu rozměrů
- principem je tvorba nových rozměrů, které lépe vyčerpávají variabilitu dat

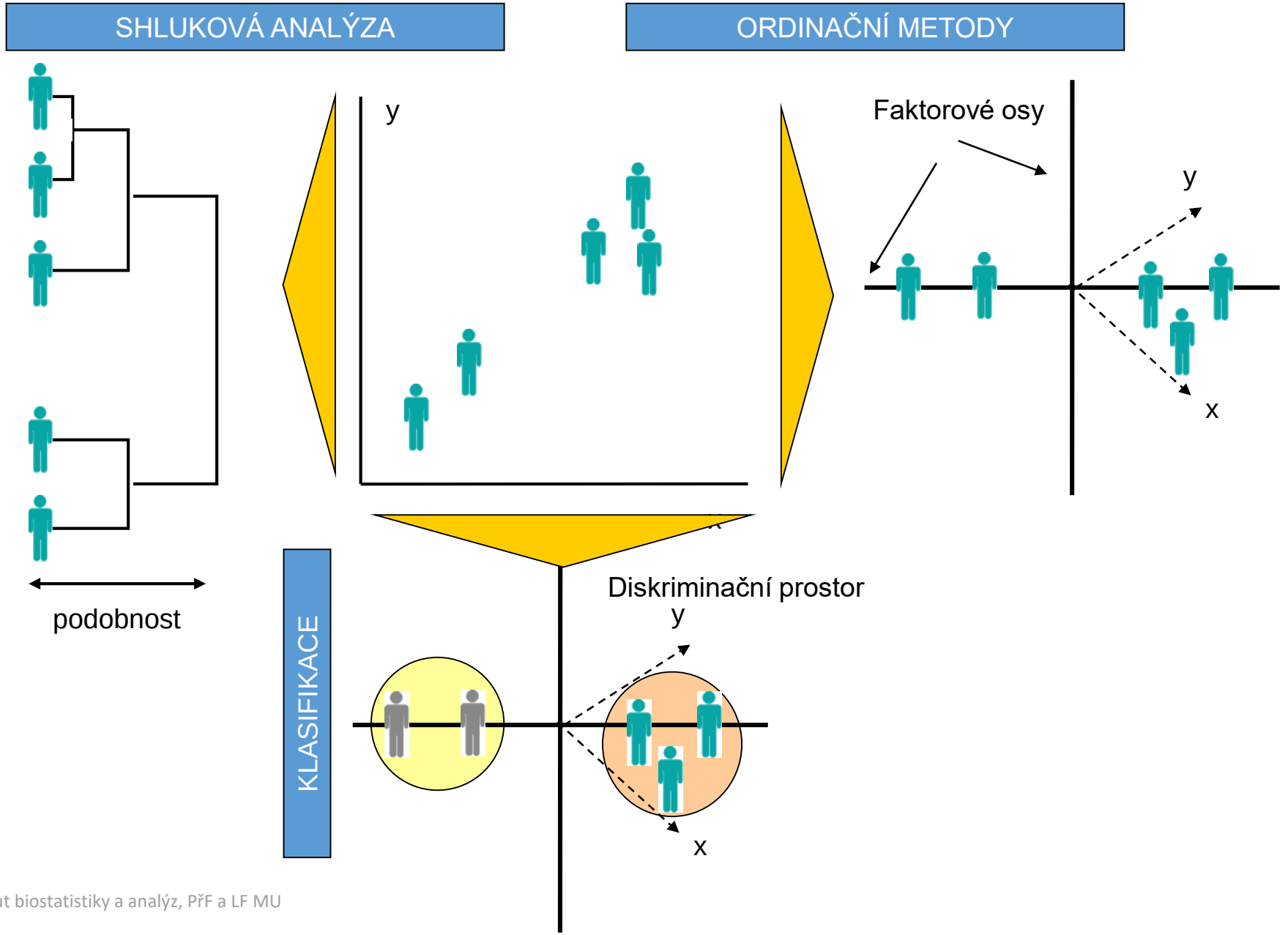
KLASIFIKACE

- Model zařazení neznámých pacientů do předem daných skupin
- Řada algoritmů

MODELOVÁNÍ

- Predikční modely s více prediktory
- Regresní metody i další typy algoritmů

Typy vícerozměrných analýz



Děkuji za pozornost, doufám jste si ze semestru něco odnesli ☺