XI. Analýza rozptylu

Parametrická analýza rozptylu

Post hoc testy

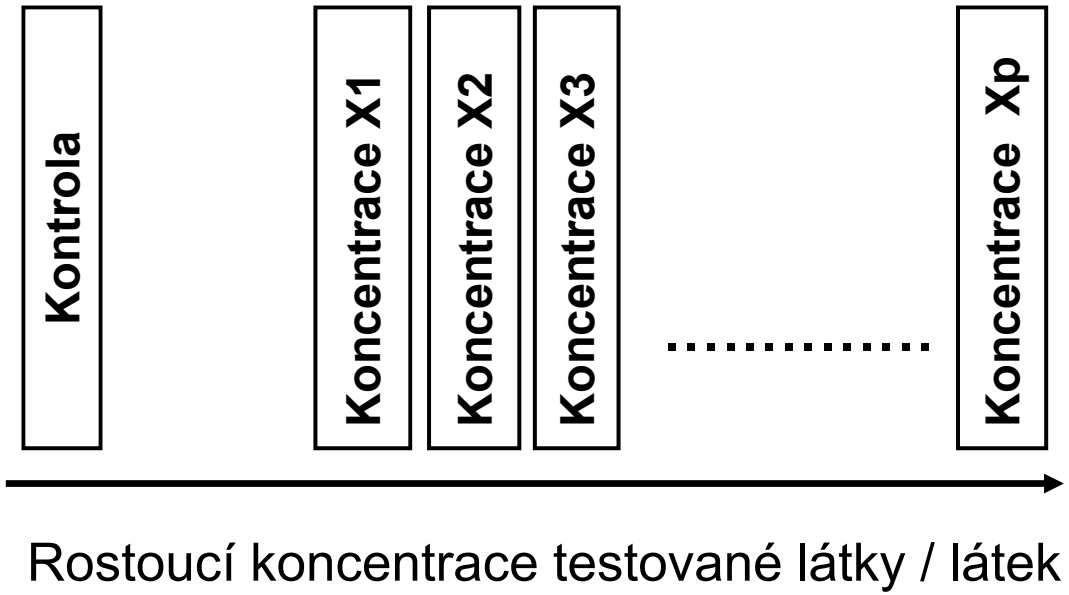
Vytvořil Institut biostatistiky a analýz, Masarykova univerzita
J. Jarkovský, L. Dušek

Anotace

- Analýza rozptylu je základním nástrojem pro analýzu rozdílů mezi průměry v několika skupinách pacientů.
- Základní myšlenka, na níž je ANOVA založena, je rozdělení celkové variability v datech (neznámé, dané pouze náhodným rozložením) na část systematickou (spjatou s kategoriemi pacientů, vysvětlená variabilita) a část náhodnou. Pokud systematická, tedy nenáhodná a vysvětlitelná část variability převažujeme, považujeme daný kategoriální faktor za významný pro vysvětlení variability dat.
- Analýza rozptylu vyhodnocuje pouze celkový vliv faktoru na variabilitu, v případě analýzy jednotlivých kategorií je třeba využít tzv. post-hoc testy

Analýza rozptylu - ANOVA

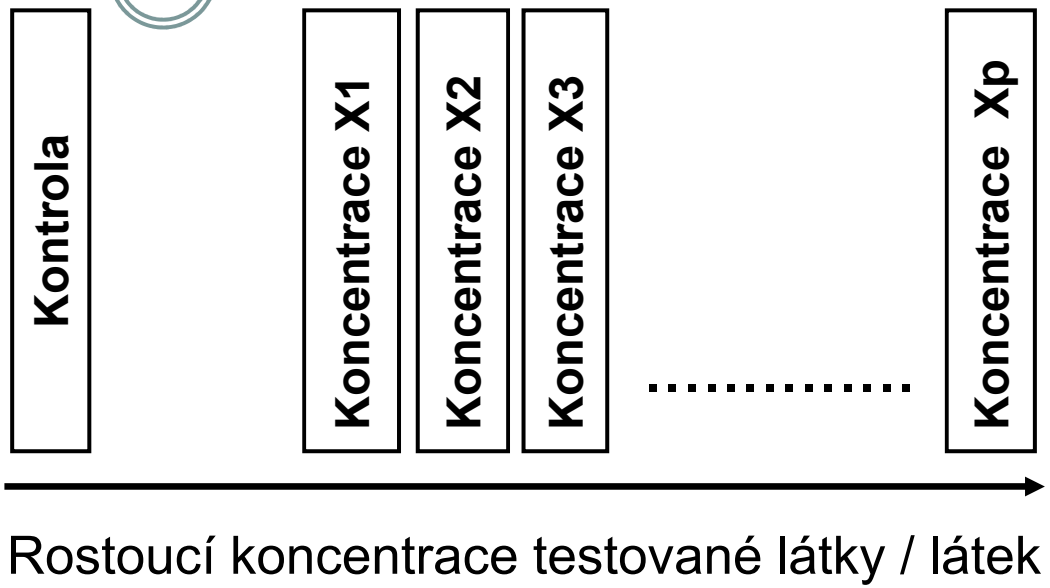
Základní technika sloužící k posouzení rozdílů mezi více úrovněmi pokusného zásahu



- Celkově významné změny v reakci biologického systému
- Vzájemné rozdíly účinku jednotlivých dávek
- Rozdíly účinku dávek od kontroly

Analýza rozptylu - ANOVA

Významné kroky analýzy, vedoucí k efektivnímu srovnání variant



- Splnění předpokladů analýzy
  Transformace dat
- Relevantnost kontroly
  (vliv vlastní aplikace látek)
- Vhodnost modelu ANOVA pro účely testu
- Vlastní srovnání variant
  Minimalizace chyb při ověřování hypotéz

Analýza rozptylu - ANOVA

SPLNĚNÍ PŘEDPOKLADŮ ANOVA JE NEZBYTNOU PODMÍNKOU POUŽITÍ TÉTO TECHNIKY

ANOVA
= parametrická analýza dat

1. **Předpoklad nezávislosti opakování experimentu**
2. **Homogenita rozptylu v rámci pokusných variant Bartlettov test**
3. **Normalita rozložení v rámci pokusných variant**

ALTERNATIVOU JSOU NEPARAMETRICKÉ METODY

Analýza rozptylu - ANOVA

Předpoklady analýzy rozptylu jsou nezbytné pro dosažení síly testu

• **Symetrické rozložení hodnot a normalita odchylek** od hodnoceného modelu ANOVA. Velkou část dat lze adekvátně normalizovat použitím logaritmické transformace. Předpoklad lognormální transformace může pochopitelně být teoreticky vyloučen u mnoha datových souborů obsahujících diskrétní parametry, kde je indikována vhodnost jiného typu transformace. U asymetricky rozložených a u diskrétních dat je nutné využít neparametrické alternativy analýzy rozptylu.

• **Homogenita rozptylu** je nutným předpokladem pro smysluplnost vzájemných srovnání pokusných variant. U testů toxicity by splnění tohoto předpokladu mělo být ověřováno (Bartlettův test), neboť vážné rozdíly (až řádové) v jednotkách testovaného parametru mohou nastat v důsledku inhibice dávkami látky. Nehomogenita rozptylu je často ve vztahu k nenormalitě (asymetrii) dat a lze ji odstranit vhodnou normalizující transformací.

• **Statistická nezávislost reziduí** vyhodnocovaného modelu ANOVA. Pokud odhad a posouzení korelačních vztahů mezi pokusnými variantami není přímo předmětem výzkumu, lze jejich vliv na vyhodnocení odstranit znáhodněním dat v rámci pokusných variant - tedy změnou pořadí v náhodné. Rozsah vlivu těchto autokorelačních vztahů musí být ovšem primárně omezen správností experimentálního uspořádání.

• **Aditivita** jako předpoklad týkající se složitějších experimentálních uspořádání. Exaktní otestování aditivity více pokusných faktorů je procedura poměrně náročná na experimentální design vyvážený co do počtu opakování. Je rovněž obtížné testovat interakci na nestandardních datech, neboť případná transformace může změnit charakter odchylek původních dat od hodnoceného modelu ANOVA.

Analýza rozptylu - ANOVA

Omezení aplikace ANOVA lze řešit

• **Chybějící data.** Vážným problémem jsou chybějící údaje o celé skupině kombinací testovaných látek, například u faktoriálních pokusů, kdy je znemožněno hodnocení experimentu jako celku.

• **Různé počty opakování** Jde o typický jev pro experimentální datové soubory. Při různých počtech opakování v experimentálních variantách jsou testy ANOVA citlivější na nenormalitu dat. Pokud jsou počty opakování zcela odlišné(až na řádové rozdíly), je nutno použít neparametrické techniky nebo analýzu rozptylu nevyvážených pokusů.

• **Odlehlé hodnoty.** Ojedinělé odlehlé hodnoty musí být před parametrickou analýzou rozptylu vyloučeny.

• **Nedostatek nezávislosti mezi rezidui modelu.** Jde o závažný nedostatek, zkreslující výsledek F-testu. Velmi často je tato skutečnost důsledkem špatného provedení nebo naplánování experimentu.

• **Nehomogenita rozptylu.** Velmi častý nedostatek experimentálních dat, často související s nenormalitou rozložení nebo s odlehlými hodnotami.

• **Nenormalita dat.** I v tomto případě lz situaci upravit vyloučením odlehlých hodnot nebo normalizující transformací.

• **Neaditivita kombinovaného vlivu více pokusných zásahů.** Tuto situaci lze testovat jednak speciálními testy aditivity nebo přímo F testem kontrolujícím významnost vlivu interakce pokusných zásahů. Při významné interakci je nutné prozkoumat především její charakter ve vhodném experimentálním uspořádání.

ANOVA – základní výpočet

- Základním principem ANOVY je porovnání rozptylu připadajícího na:
  - Rozdělení dat do skupin (tzv. effect, variance between groups)
  - Variabilitu objektů uvnitř skupin (tzv. error, variance within groups), předpokládá se, že jde o náhodnou variabilitu (=error)

1. Variabilita mezi skupinami

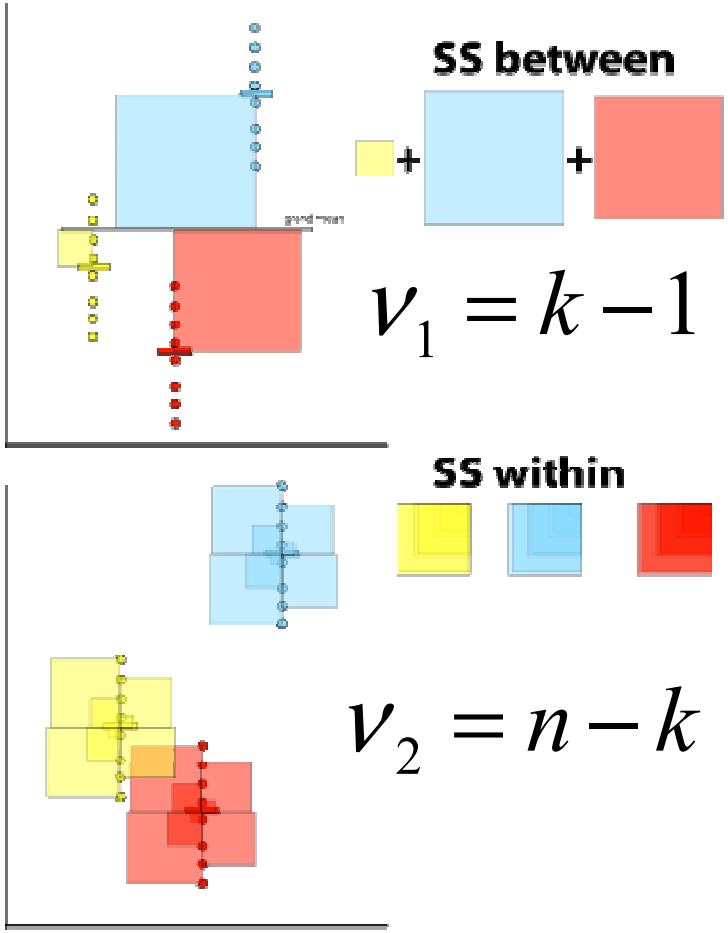
Rozptyl je počítán pro celkový průměr (tzv. grand mean) a průměry v jednotlivých skupinách dat

Stupně volnosti jsou odvozeny od počtu skupin (= počet skupin -1)

2. Variabilita uvnitř skupin

Rozptyl je počítán pro průměry jednotlivých skupin a objekty uvnitř příslušných, celková variabilita je pak sečtena pro všechny skupiny

Stupně volnosti jsou odvozeny od počtu hodnot (= počet hodnot - počet skupin)



$$\nu_1 = k - 1$$

$$\nu_2 = n - k$$

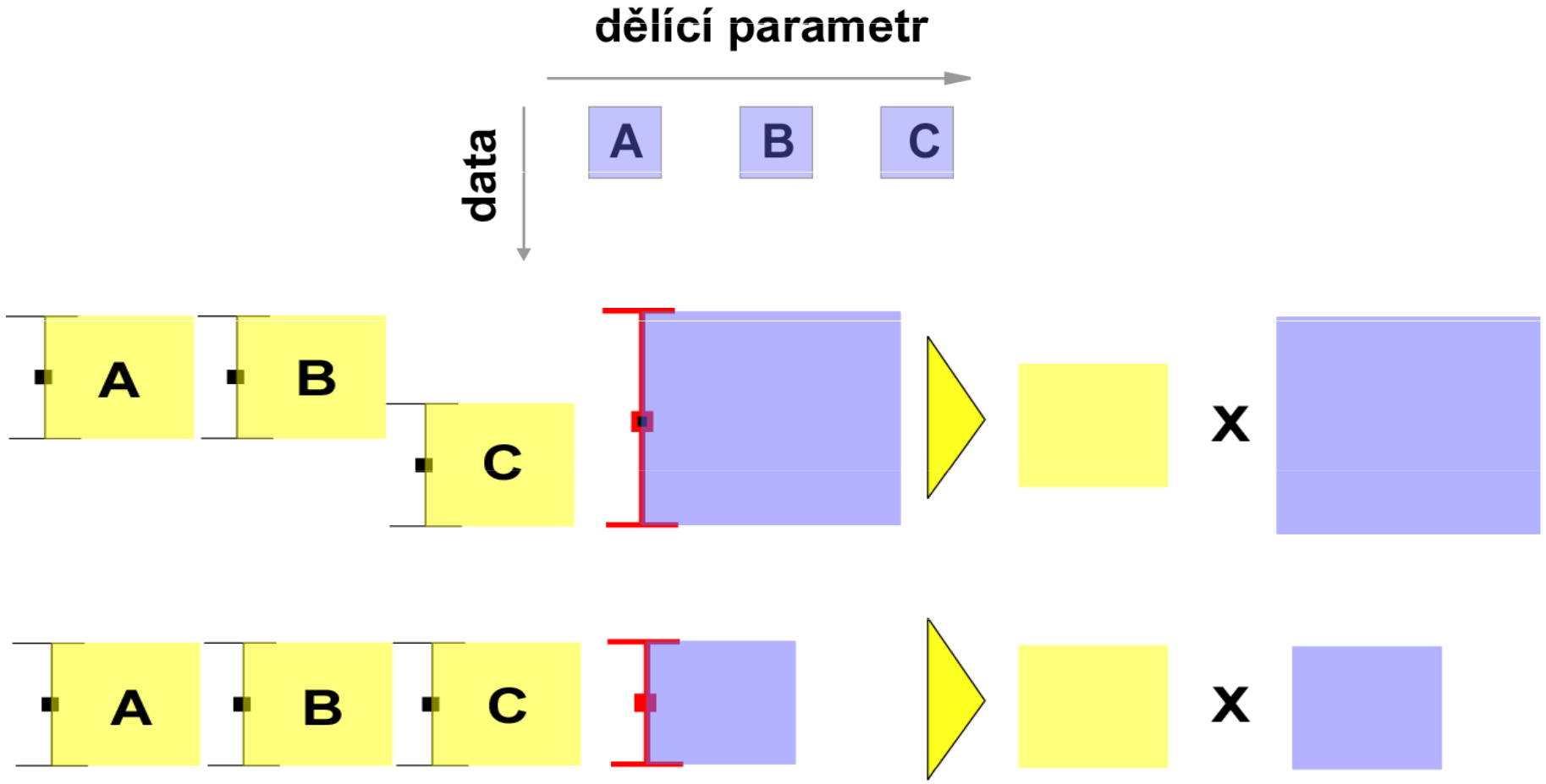
$$F = \frac{between\_groups}{within\_groups}$$

Výsledný poměr (F) porovnáme s tabulkami F rozložení pro v_1 a v_2 stupňů volnosti

SS=sum of squares

Jednoduchý ANOVA design

Nejjednodušším případem ANOVA designu je rozdělení na skupiny podle jednoho parametru.

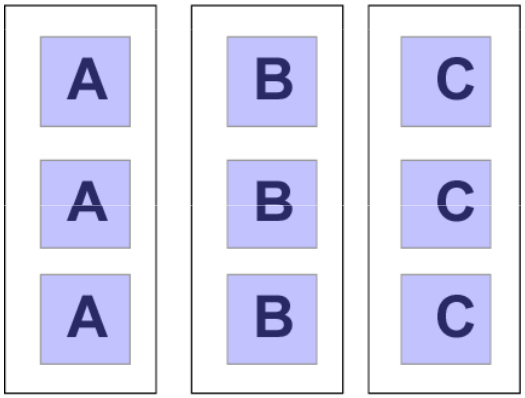


Nested ANOVA

- Rozdělení skupin na náhodné podskupiny (např. opakování experimentu)
- Cílem je zjistit, zda data v jedné skupině nejsou pouhou náhodou
- Nejprve je testována shoda podskupin v hlavních skupinách,
  - pokud jsou shodné, je vše v pořádku
  - pokud nejsou, stále lze zjišťovat, zda se variabilita uvnitř hlavních skupin liší od celkové variability

jednoduchá ANOVA

nested ANOVA

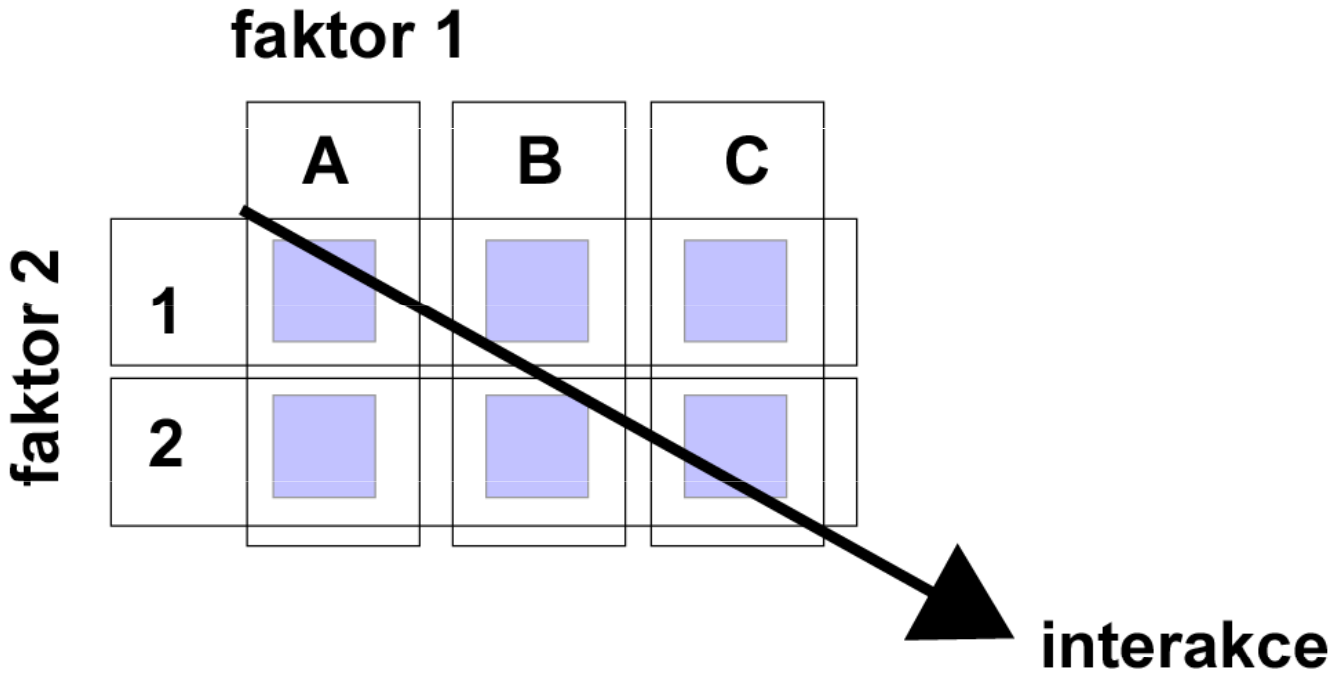


Two way ANOVA

Pro rozdělení do kategorií je zde více parametrů

Na rozdíl od nested ANOVY nejde o náhodná opakování experimentu, ale o řízené zásahy (např.vliv pH a koncentrace O_2)

Kromě vlivu hlavních faktorů se uplatňuje i jejich interakce



Modely analýzy rozptylu - základní výstup

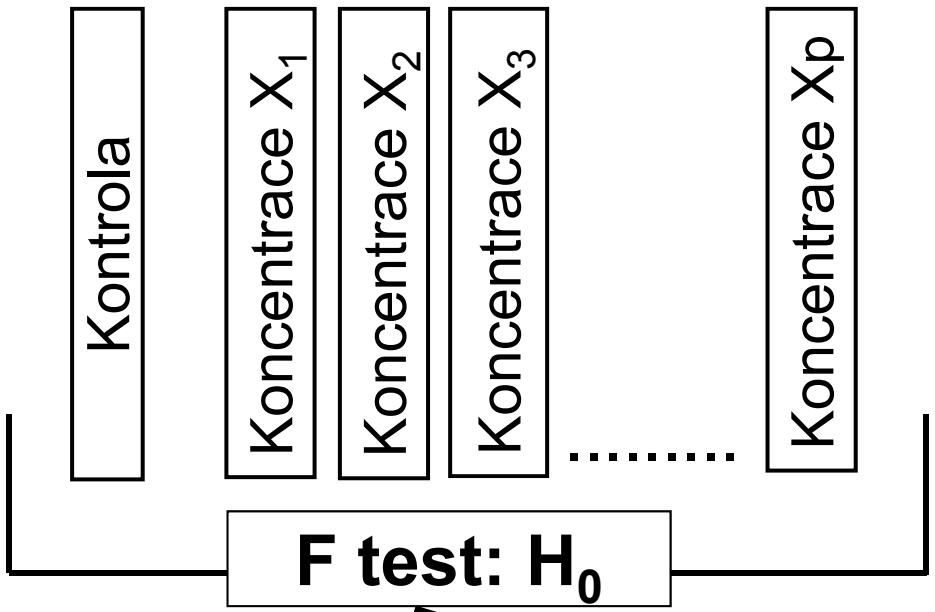
Základním výstupem analýzy rozptylu je
Tabulka ANOVA - frakcionace komponent rozptylu

Zdroj rozptylu	St. v.	SS	MS	F
Pok. zásah (mezi skupinami)	**a -1**	**SS_B**	**$SS_B/(a-1)$**	**MS_B/MS_E**
Uvnitř skupin	**N - a**	**SS_E**	**$SS_E/(N-a)$**	
Celkem	**N -1**	**SS_T**		

SS_B/SS_T → Kvantifikovaný podíl rozdílu mezi pokusnými zásahy na celkovém rozptylu

MS_B/MS_T → Statistická významnost rozdílu

Analýza rozptylu - obecný F test



obecný F test

H_0: $m_1 = m_2 = m_3 = ... = m_p$

H_0 platí

Látka nepůsobí

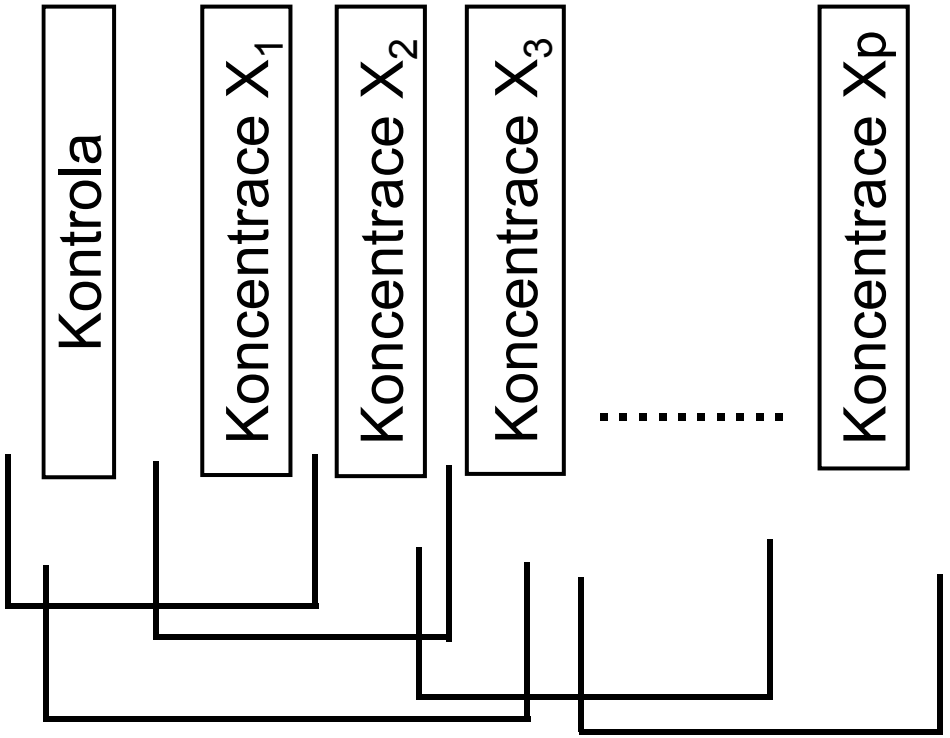
H_0 neplatí

Látka působí

Další analýzy

Analýza rozptylu - Testy kontrastů

ANOVA:H_0 zamítnuta
Testy kontrastů



Rozdíly v smysluplných kombinacích ?

Plánované

Neplánované

Pro srovnání variant s kontrolou

Testování kontrastů "Multiple range testy"

Parametrické

Neparametrické

Příklad: Anova - One way

Dávka rostlinného stimulátoru (0, 4, 8, 12 mg/l)
A = 4 ; n = 8

I. ANOVA

Bartlett's test: P = 0,9847
K-S test: P = 0,482 - 0,6525 pro jednotlivé kategorie

Source	D. f.	SS	MS	F
Between Groups	3	305,8	101,9	8,56
Within Groups	28	322,2	11,9	
Total (corr.)	31	638,0		

II. Multiple Range Test

NKS -test

Level	Average	Homogenous Groups		
0	34,8	x		
4	41,4		x	
12	41,8		x	
8	52,6			x

Příklad: Anova - One way

I. Zásah: 4 klinická stadia virové choroby (napadá kr. buňky)
Sledovaná veličina: aktivita enzymu v těchto krevních buňkách

$$H_o: \quad \mu_1 = \mu_2 = \mu_3 = \mu_4$$

n = 3
MODEL = ?

	I	II	III	IV
	22,8	16,4	11,2	14,2
	19,4	17,8	18,2	10,1
	12,5	19,1	15,8	12,8
Σ	65,7	53,3	45,2	37,1
průměr	21,9	17,8	15,1	12,4

II.

Source	D.f.	MS	F	P
Between groups	3	49,6	8,39	0,0075
Within groups	8	5,9		
Total (corr.)	11	-		

III. Komponenta rozptylu:

$$\sigma_A^2 \sim S_A^2 = \frac{MS_A - MS_e}{n} = \frac{49{,}6 - 5{,}9}{3} = 14{,}57$$

$$S_A^2 = 2{,}5 \cdot S_e^2$$

IV.

$$\rho_I \sim r_I = \frac{S_A^2}{S_A^2 + S_e^2} = 0{,}7142$$

Srovnání variant v testech

Srovnáváni variant po celkovém testu ANOVA

Mnoho existujících algoritmů není vhodných
pro konkrétní případ

Day and Quin
Ecological Monographs,1989

Test	Využití	Poznámka
Dunnett Williams	**Srovnání s kontrolou**	Ex. i modifikace pro různá n.
ANOVA testy (F)	**Orthogonální kontrasty**	Plánovaná srovnání
Ryan Q test	**Jednoduché kontrasty**	Vyhodnocen jako nejlepší test

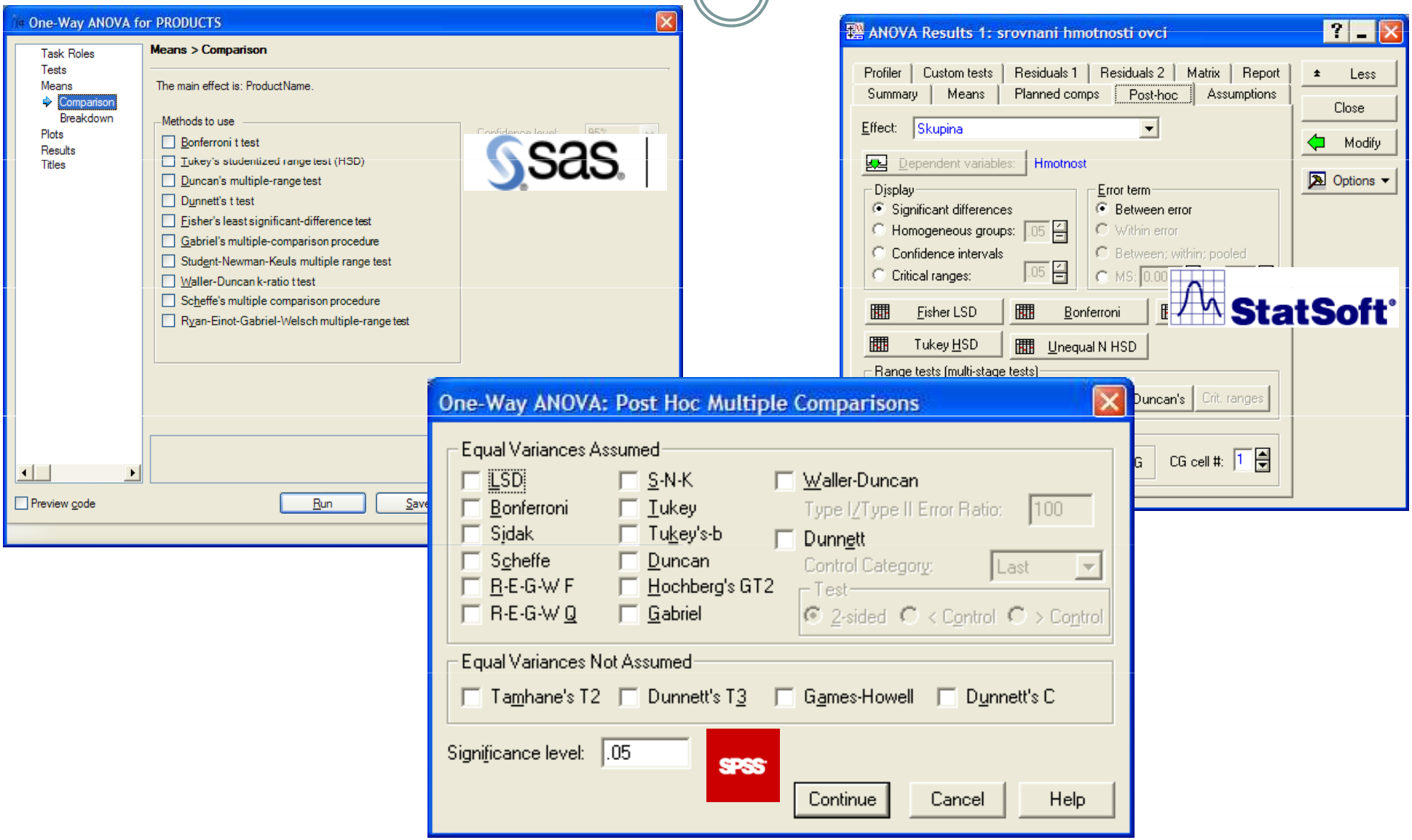
Testy pro jednoduché kontrasty

Scheffe	**Tukey**	**LSD**
Bonferroni	**Dunn-Sidák**	**Kramer**

Testy nevhodné

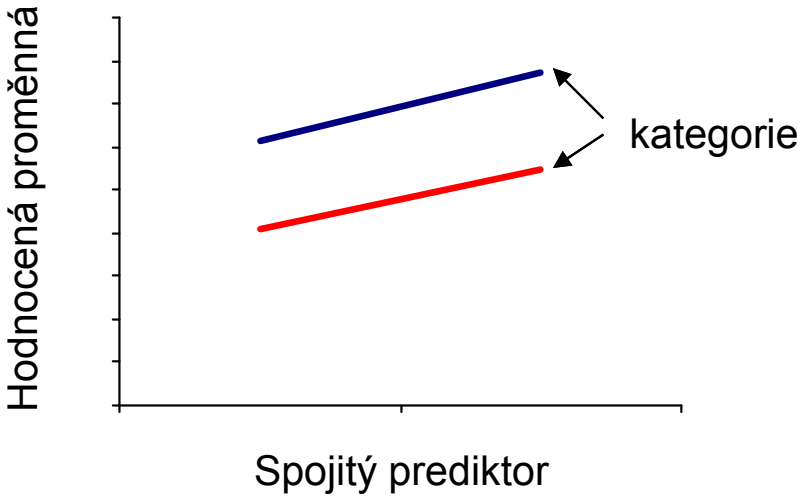
Duncan	**Student - Newmann-Keuls**	**Waller-Duncan k ratio**

Řada post-hoc testů v různých SW

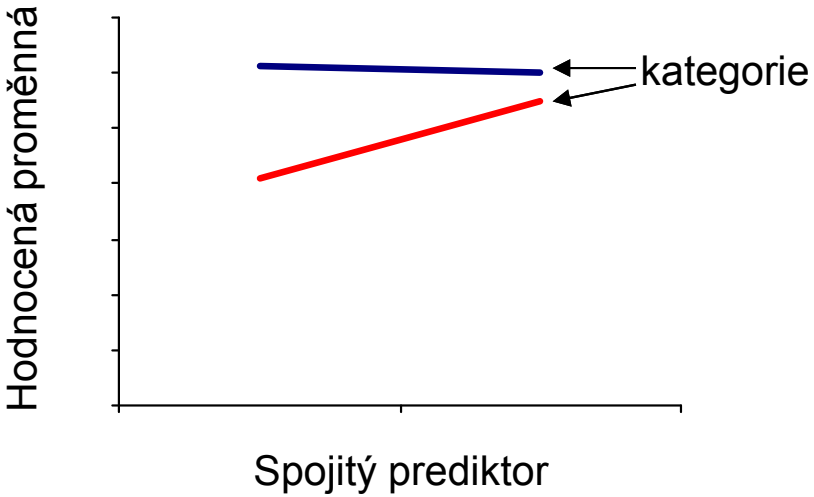


ANCOVA

- Rozšíření ANOVA
- Současná analýza kategoriálních a spojitých prediktorů
- Testování hypotézy paralelismu regresních vztahů



Kategorie pacientů (pokusný zásah) neovlivňuje vztah proměnných



Kategorie pacientů (pokusný zásah) ovlivňuje vztah proměnných

XI. Korelace

Parametrická a neparametrická korelace

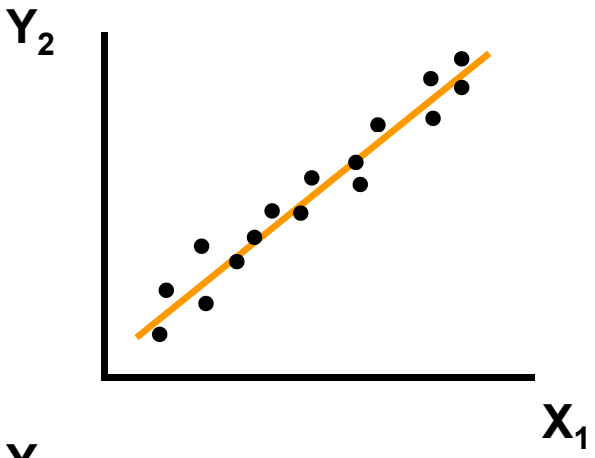
Vytvořil Institut biostatistiky a analýz, Masarykova univerzita
J. Jarkovský, L. Dušek

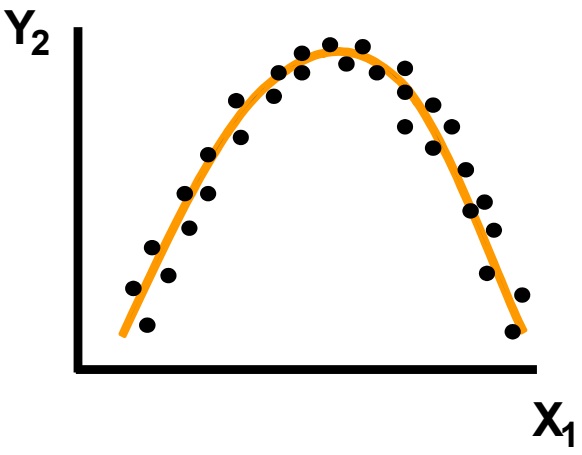
Anotace

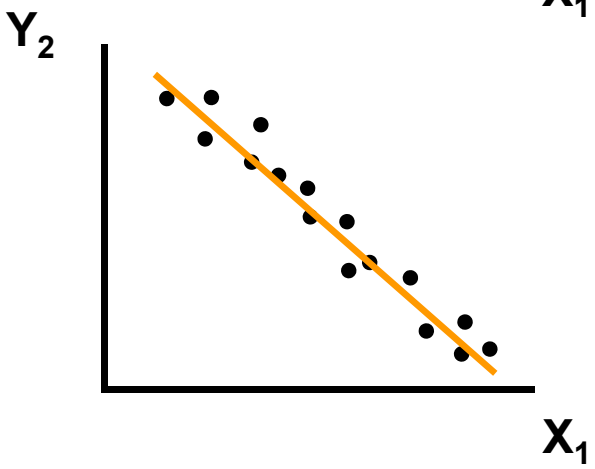
- Korelační analýza je využívána pro vyhodnocení míry vztahu dvou spojitých proměnných. Obdobně jako jiné statistické metody, i korelace mohou být parametrické nebo neparametrické
- Regresní analýza vytváří model vztahu dvou nebo více proměnných, tedy jakým způsobem jedna proměnná (vysvětlovaná) závisí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

Základy korelační analýzy - I.

Korelace - vztah (závislost) dvou znaků (parametrů)







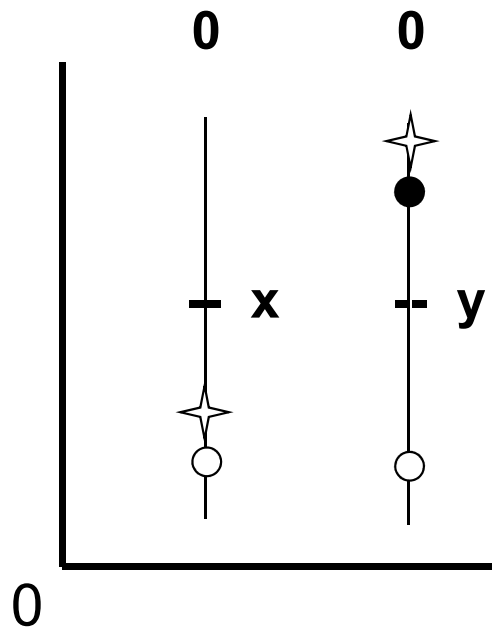
X_2 \ X_1	ANO	NE
ANO	a	b
NE	c	d

Základy korelační analýzy - II.

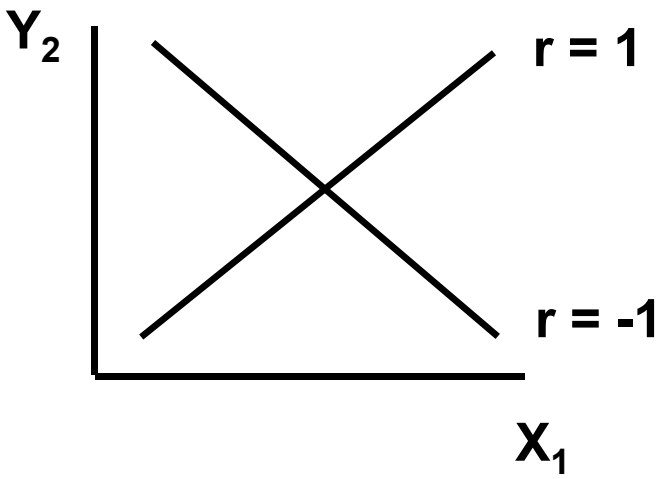
Parametrické míry korelace

Kovariance

$$Cov(x, y) = E(x_i - \bar{x}).(y_i - \bar{y})$$



Pearsonův koeficient korelace



Základy korelační analýzy - III.

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

tab : $r(v = 6) = 0,7076$

II. $H_0 : \rho = \phi$ $\qquad t = \left[\frac{r}{\sqrt{1 - r^2}}\right] \cdot \sqrt{n - 2}$ $\qquad v = n - 2$

$$t = \frac{0,7176}{0,6965} \cdot \sqrt{6} = 2,524$$

tab : $t_{0,975}^{(n-2)} = 2,447$

$P \leq 0,05$

Základy korelační analýzy - IV.
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

Základy korelační analýzy - V. Neparametrická korelace (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

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

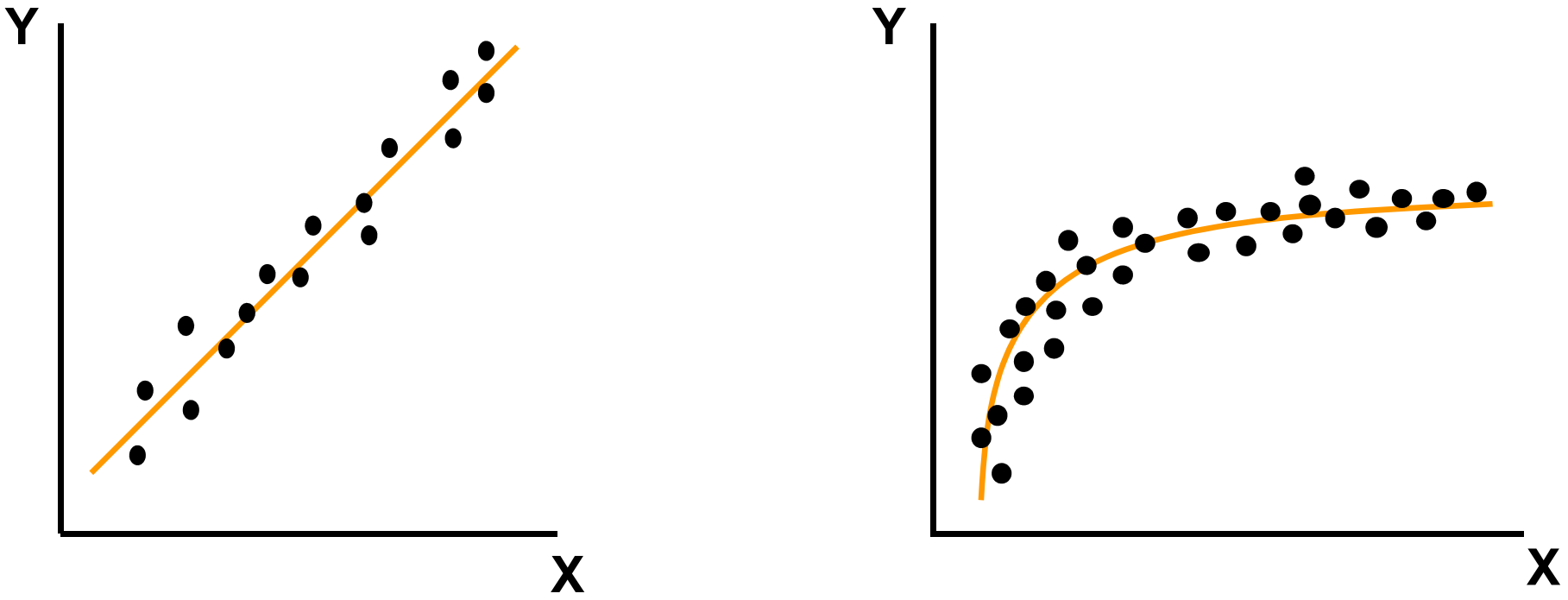
tab : $r_s(v = 6) = 0,89$

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

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

P = 0,358

Korelace v grafech I.



Vztahy velmi často implikují funkční vztah mezi Y a X.

$$Y = a + b \,.\, X$$

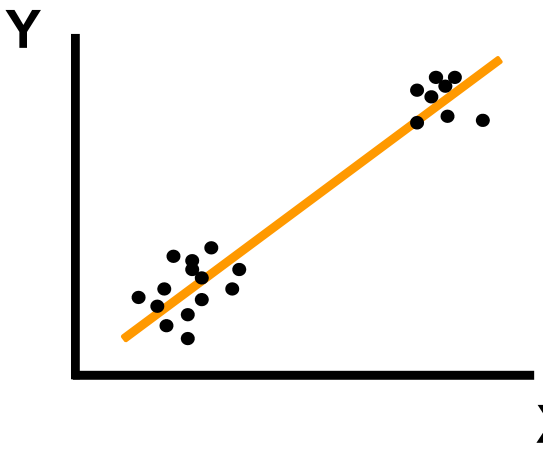
$$Y = a + b_1 \,.\, X_1 + b_2 \,.\, X_2 + b_3 \,.\, X_3$$

$$Y = a + b_1 \,.\, X_1 + b_2 \,.\, X_2$$

$$Y = a + b_1 \,.\, X_1 + b_2 \,.\, X_2 + b_3 \,.\, X_1 \,.\, X_2$$

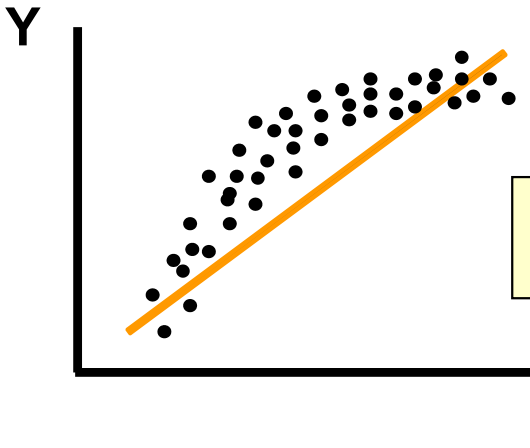
Korelace v grafech II.

Problém rozložení hodnot



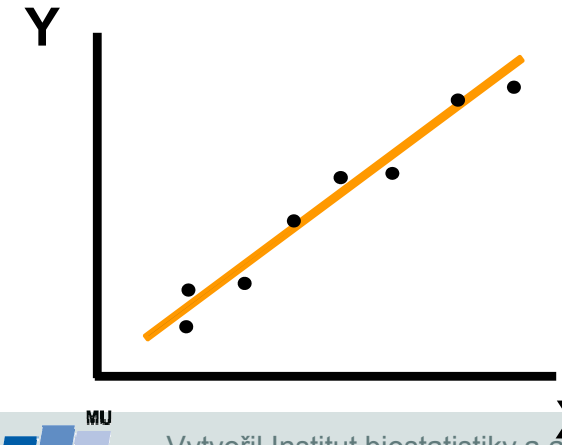
r = 0,981
(p < 0,001)

Problém typu modelu

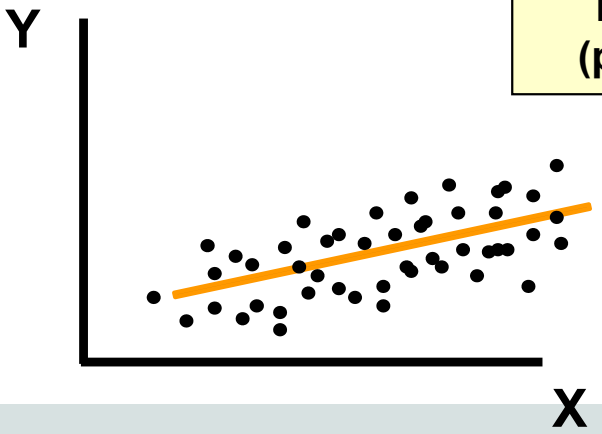


r = 0,761
(p < 0,032)

Problém velikosti vzorku



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



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