8. Kontingenční tabulky a χ^2 test

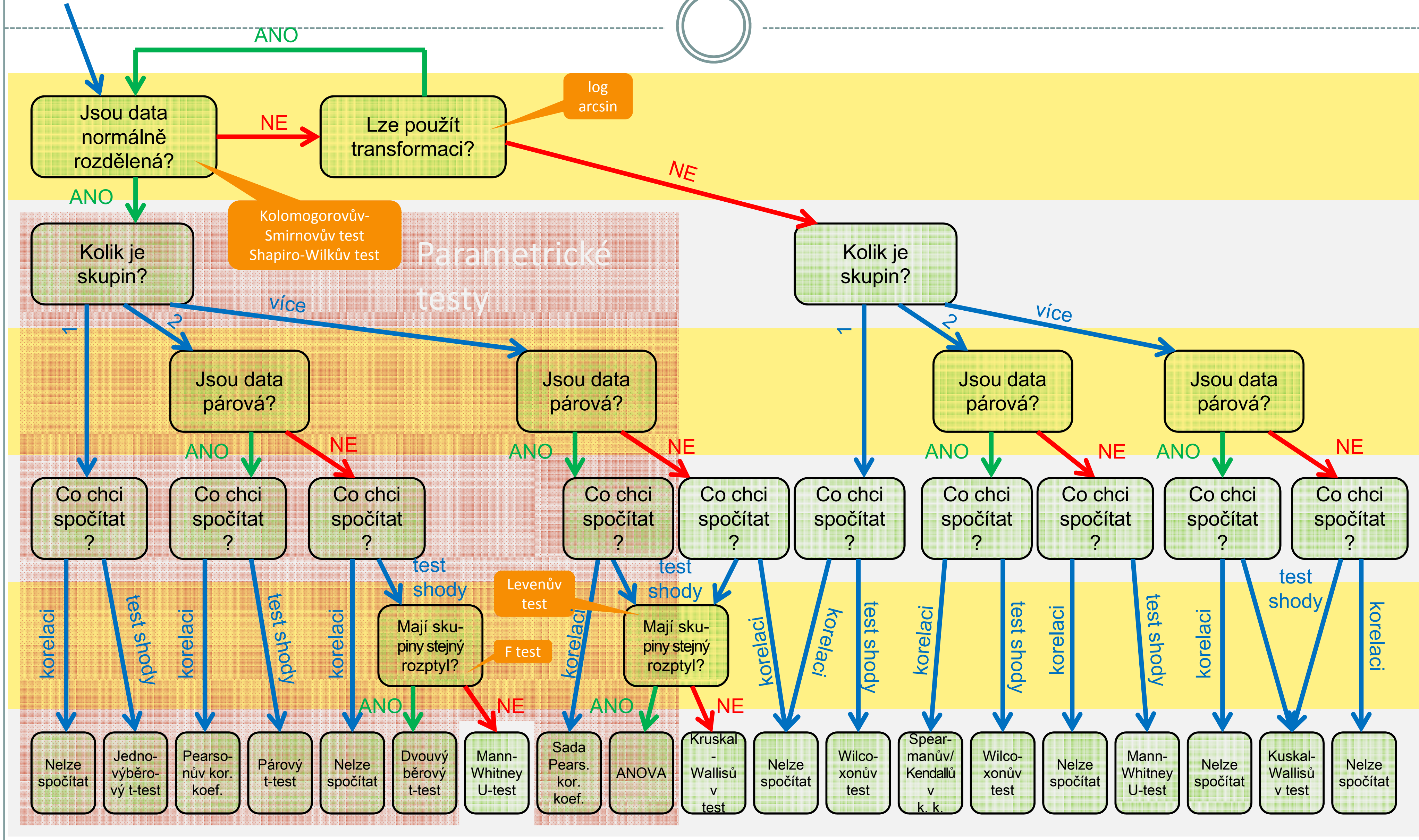
Znaménkový test a χ^2 test
Shrnutí statistických testů
Kontingenční tabulky

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

Shrnutí statistických testů

Typ srovnání	Nulová hypotéza	Parametrický test	Neparametrický test
1 skupina dat vs. etalon	**Střední hodnota je rovna hodnotě etalonu.**	jednovýběrový t-test	Wilcoxonův test; znaménkový test
2 skupiny dat nepárově	**Obě skupiny hodnot pochází ze stejného rozdělení.**	nepárový t-test	Mann-Whitneyův test
2 skupiny dat párově	**Zkoumaný efekt mezi páry hodnot je nulový.**	Párový t-test	Wilcoxonův test; znaménkový test
shoda rozdělení	**rozdělení dat ve skupině odpovídá teoretickému (vybranému) rozdělení.**	Shapiro-Wilkův test; Kolmogorovův-Smirnovův test; Lilieforsův test	χ2 test, test dobré shody
homoskedasticita (shoda rozptylů)	**rozptyl obou (všech) skupin je shodný.**	Levenův test	
více skupin nepárově	**Zkoumaný efekt mezi skupinami hodnot je nulový.**	ANOVA	Kruskal- Wallisův test
korelace	**Neexistuje (příčinná, důsledková) vazba mezi skupinami hodnot.**	Pearsonův koeficient	Spearmanův koeficient; Kendallův koeficient

Shrnutí statistických testů



Statistické testování – základní pojmy

- **Nulová hypotéza H_0**

  H_0: sledovaný efekt je nulový

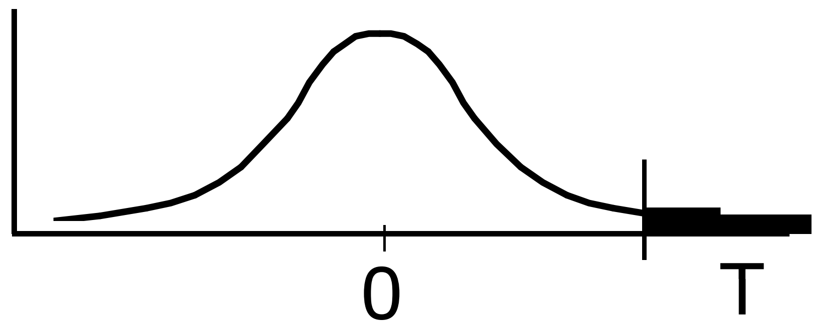
- **Alternativní hypotéza H_A**

  H_A: sledovaný efekt je různý mezi skupinami

- **Testová statistika**

$$\textbf{Testová statistika} = \frac{\textbf{Pozorovaná hodnota – Očekávaná hodnota}}{\textbf{Variabilita dat}} * \sqrt{\textbf{Velikost vzorku}}$$

- **Kritický obor testové statistiky**



Statistické testování odpovídá na otázku zda je pozorovaný rozdíl náhodný či nikoliv. K odpovědi na otázku je využit statistický model – testová statistika.

P-hodnota

Významnost hypotézy hodnotíme dle získané tzv. p-hodnoty, která vyjadřuje pravděpodobnost, s jakou číselné realizace výběru podporují H_0, je-li pravdivá.

P-hodnotu porovnáme s α (hladina významnosti, stanovujeme ji na **0,05**, tzn., že připouštíme 5 % chybu testu, tedy, že zamítneme H_0, ačkoliv ve skutečnosti platí).

P-hodnotu získáme při testování hypotéz ve statistickém softwaru.

- Je-li **p-hodnota ≤ α**, pak **H_0 zamítáme** na hladině významnosti α a **přijímáme H_A**
- Je-li **p-hodnota > α**, pak **H_0 nezamítáme** na hladině významnosti α

P-hodnota vyjadřuje pravděpodobnost za platnosti H_0, s níž bychom získali stejnou nebo extrémnější hodnotu testové statistiky.

Test dobré shody - základní teorie

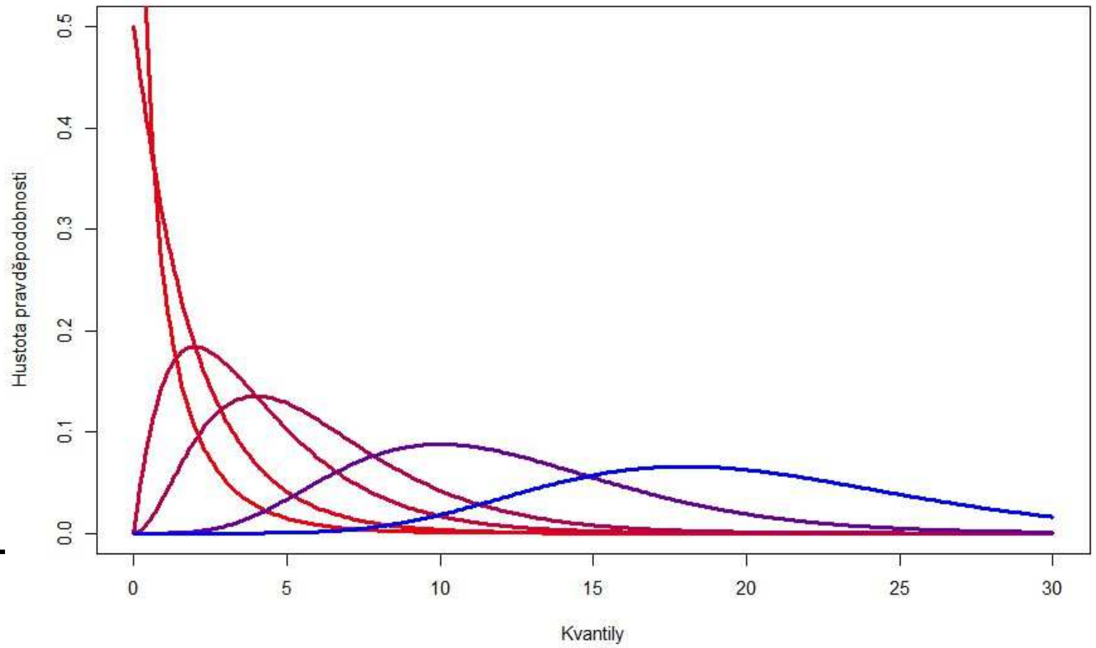
Testuje shodu reálné distribuce hodnot do n skupin s teoretickou distribucí.

Předpokladem je, že velikost rozdílu mezi očekávaným a skutečným počtem hodnot v každé skupině je náhodně rozdělená → multinomické rozdělení.

Součet druhých mocnin relativních rozdílů očekávaného a skutečného počtu hodnot má přibližně χ^2 rozdělení.

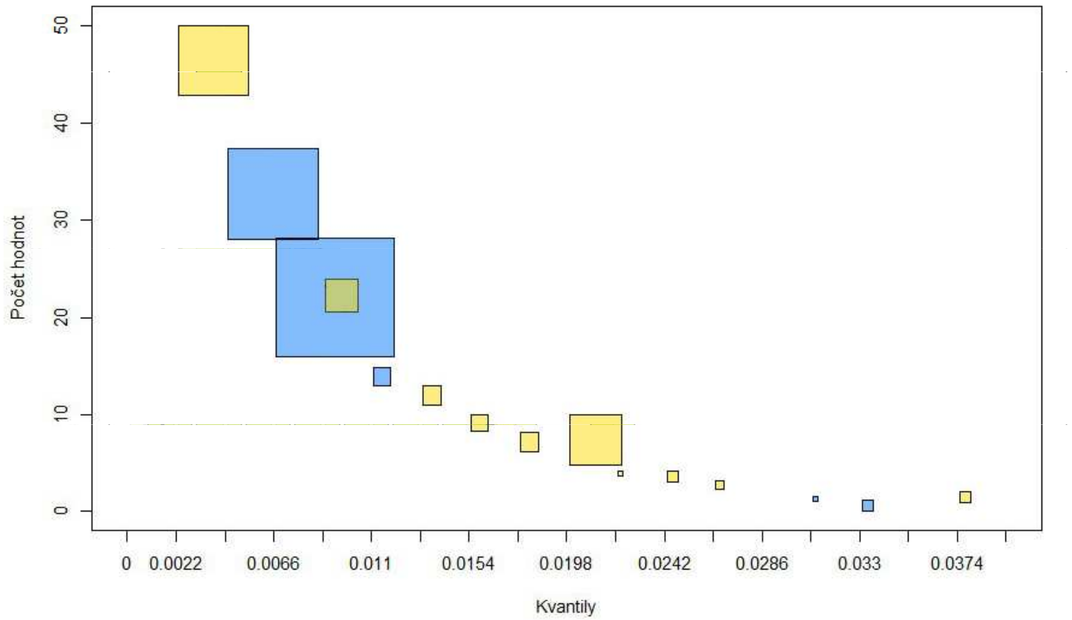
χ^2 rozdělení pro kladné hodnoty (suma čtverců) se liší podle počtu stupňů volnosti *k* (počtu skupin) - se zvyšujícím se *k* přechází v normální rozdělení.

$$\chi^2_{(s.v.)} = \sum \frac{\left[\textbf{pozorovaná četnost} - \textbf{očekávaná četnost}\right]^2}{\textbf{očekávaná četnost}}$$



Test dobré shody - základní teorie

$$\chi^2_{(s.v.)} = \underbrace{\frac{\left[\text{pozorovaná četnost} - \text{očekávaná četnost}\right]^2}{\text{očekávaná četnost}}}_{\text{1. jev}} + \underbrace{\frac{\left[\text{pozorovaná četnost} - \text{očekávaná četnost}\right]^2}{\text{očekávaná četnost}}}_{\text{2. jev}} + \ldots$$



Očekávané četnosti

V případě platnosti nulové hypotézy je poměr mezi buňkami jednoho sloupce v různých řádcích nezávislý na výběru tohoto sloupce.

V případě platnosti nulové hypotézy je poměr mezi buňkami jednoho řádku v různých sloupcích nezávislý na výběru tohoto řádku.

Pokud tyto poměry normalizujeme, získáváme tabulku očekávaných četností.

Řádkové a sloupcové součty se touto operací nemění.

Pozorované četnosti

	Ano	Ne	Σ
Ano	20	82	**102**
Ne	10	54	**64**
Σ	**30**	**136**	**166**

$$102 \times 30 / 166$$

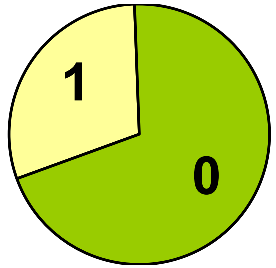
Očekávané četnosti

	Ano	Ne	Σ
Ano	18,4	83,6	**102**
Ne	11,6	52,4	**64**
Σ	**30**	**136**	**166**

Test dobré shody - základní teorie

Binomické jevy (1/0)

$$\chi^2_{(1)} = \underbrace{\frac{\left[\text{pozorovaná četnost} - \text{očekávaná četnost}\right]^2}{\text{očekávaná četnost}}}_{\text{I. jev 1}} + \underbrace{\frac{\left[\text{pozorovaná četnost} - \text{očekávaná četnost}\right]^2}{\text{očekávaná četnost}}}_{\text{II. jev 2}}$$



Příklad

✓ **10 000 lidí hází mincí** → **rub: 4 000 případů (R)**; **líc: 6 000 případů (L)**

? **Lze výsledek považovat za statisticky významně odlišný (nebo neodlišný) od očekávaného poměru R : L = 1 : 1 ?**

$$\chi^2_{(1)} = \frac{(4000 - 5000)^2}{5000} + \frac{(6000 - 5000)^2}{5000} = 400$$

Tabulková hodnota: $\chi^2_{(0,95)}(\nu = 1) = \underline{\underline{3,84}}$ $\qquad (0,95 = 1 - \alpha)$

Rozdíl je vysoce statisticky významný (p << 0,001)

Znaménkový test

Zjednodušení neparametrického párového Wilcoxonova testu.

Namísto velikosti rozdílů se počítá pouze jejich orientace (signum).

Případy, kde *sgn(d) = 0* se z analýzy vylučují.

Sečtou se kladné a záporné rozdíly a menší ze součtů je hledaná statistika *m*.

Statistika *m* se porovná s tabulkovou hodnotou pro danou hladinu pravděpodobnosti:

Počet párů *n*	Hladina významnosti (α) 0,01	0,05	0,10
5	-	-	0
6	-	0	0
7	-	0	0
8	0	0	1
9	0	1	1
10	0	1	1
11	0	1	2
12	1	2	2
13	1	2	3
14	1	2	3
15	2	3	3
16	2	3	4
17	2	4	4
18	3	4	5
19	3	4	5
20	3	5	5

Kontingenční tabulky
H0 :Nezávislost dvou jevů A a B

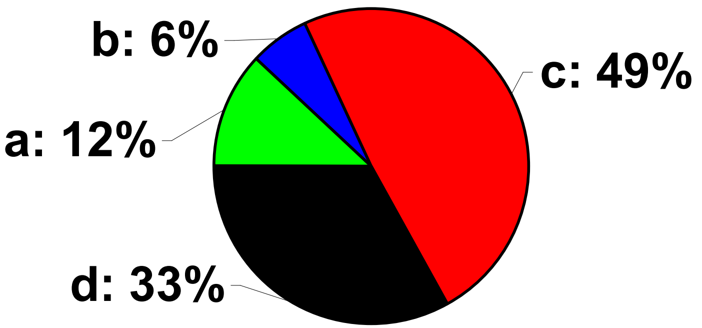
Kontingenční tabulka 2 x 2

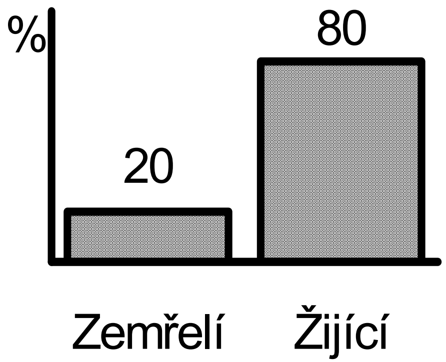
B ↓ \ A →	+	-	Σ
+	a	b	
-	c	d	
Σ			**suma sum**

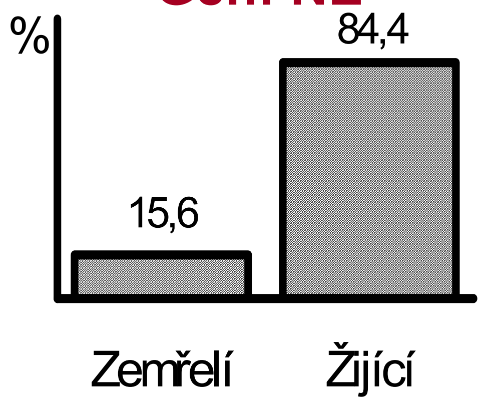
Kontingenční tabulky: příklad

gen \ †	Ano	Ne	Σ
Ano	20	82	102
Ne	10	54	64
Σ	30	136	166

Kontingenční tabulka v obrázku







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

Příklad – závislost pohlaví na onemocnění

	Zdraví	Nemocní	Celkem
Muži	50	50	100
Ženy	50	50	100
Celkem	100	100	200

	Zdraví	Nemocní	Celkem
Muži	45	55	100
Ženy	55	45	100
Celkem	100	100	200

	Zdraví	Nemocní	Celkem
Muži	35	65	100
Ženy	65	35	100
Celkem	100	100	200

	Zdraví	Nemocní	Celkem
Muži	5	95	100
Ženy	95	5	100
Celkem	100	100	200

Příklad – závislost pohlaví na onemocnění

Pozorované hodnoty

	Zdraví	Nemocní	Celkem
Muži	50	50	100
Ženy	50	50	100
Celkem	100	100	200

	Zdraví	Nemocní	Celkem
Muži	45	55	100
Ženy	55	45	100
Celkem	100	100	200

	Zdraví	Nemocní	Celkem
Muži	35	65	100
Ženy	65	35	100
Celkem	100	100	200

	Zdraví	Nemocní	Celkem
Muži	5	95	100
Ženy	95	5	100
Celkem	100	100	200

Očekávané hodnoty pro <u>všechny</u> tabulky vlevo

	Zdraví	Nemocní	Celkem
Muži	50	50	100
Ženy	50	50	100
Celkem	100	100	200

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

Příklad – závislost pohlaví na onemocnění

	Zdraví	Nemocní	Celkem
Muži	**50**	**50**	100
Ženy	**50**	**50**	100
Celkem	100	100	200

$X^2 = 0,0$
$p = 1,000$

	Zdraví	Nemocní	Celkem
Muži	**45**	**55**	100
Ženy	**55**	**45**	100
Celkem	100	100	200

$X^2 = 2,0$
$p = 0,157$

	Zdraví	Nemocní	Celkem
Muži	**35**	**65**	100
Ženy	**65**	**35**	100
Celkem	100	100	200

$X^2 = 18,0$
$p < \mathbf{0,0001}$

	Zdraví	Nemocní	Celkem
Muži	**5**	**95**	100
Ženy	**95**	**5**	100
Celkem	100	100	200

$X^2 = 162,0$
$p < \mathbf{0,0001}$

9. Analýza rozptylu a korelace

Parametrická analýza rozptylu

Post hoc testy

Kruskal-Wallisův test

Korelace

Lineární regrese

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

Anotace

- **t-test** slouží pro porovnání průměrů spojité proměnné ve dvou (diskrétních) skupinách.
- **Analýza rozptylu (ANOVA)** umožňuje totéž porovnání provést pro větší počet (diskrétních) skupin.
- **Korelační analýza** je využívána pro vyhodnocení míry vztahu dvou spojitých proměnných.
- **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.

Existují rovněž neparametrické varianty t-testu a ANOVy.

Analýza rozptylu - ANOVA

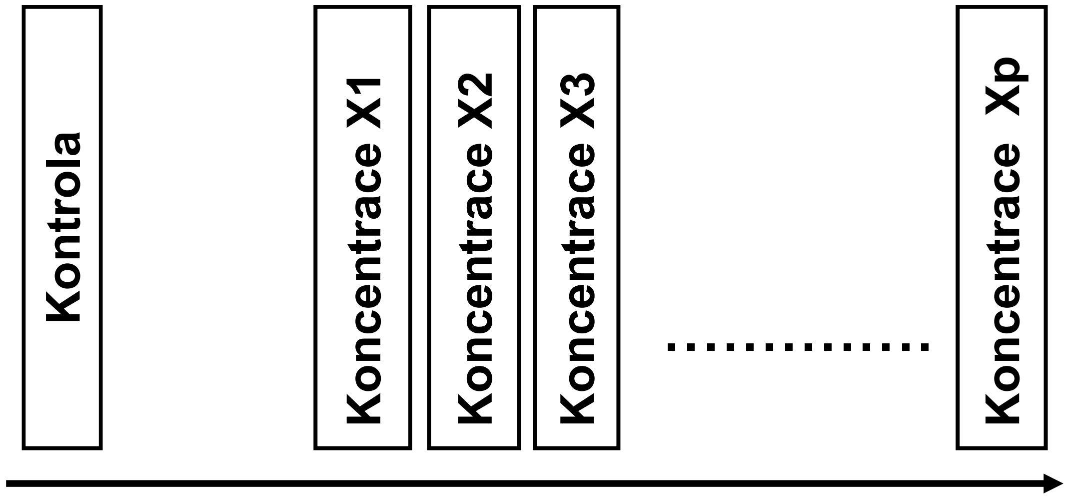
- Zobecnění dvouvýběrového t-testu
- ANOVA je základním nástrojem pro analýzu rozdílů mezi průměry v několika skupinách
- H_0: všechny střední hodnoty jsou stejné
  H_A: alespoň jedna dvojice středních hodnot se liší
- Předpoklady: normální rozložení ve skupinách, nezávislost skupin, shoda rozptylů (Levenův či Bartlettův test)
- Pokud H_0 zamítáme na hl. význ. α → nás zajímá, která dvojice středních hodnot se od sebe liší
  - metody mnohonásobného testování (tzv. post hoc testy), např. Schefféhо, Tukeyova metoda

Anotace

- 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 převažuje, 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

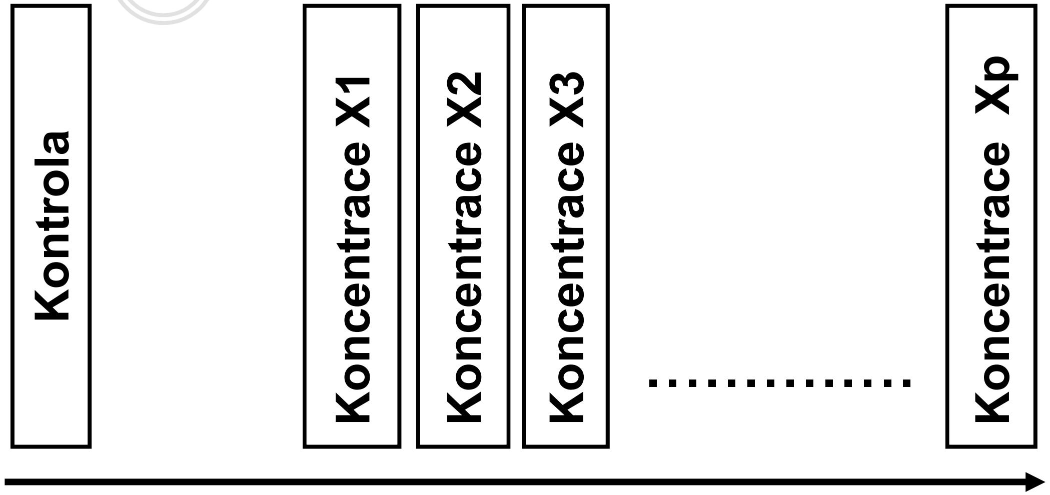


Rostoucí koncentrace testované látky / látek

- 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



Rostoucí koncentrace testované látky / látek

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**
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 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.

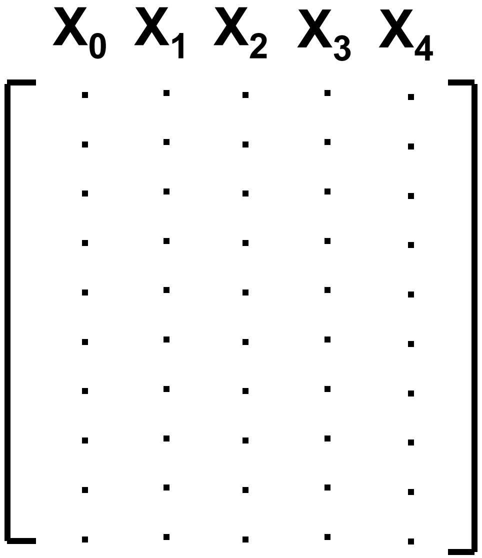
- **Nehomogenita rozptylu.** Velmi častý nedostatek experimentálních dat, často související s nenormalitou rozložení nebo s odlehlými hodnotami.

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

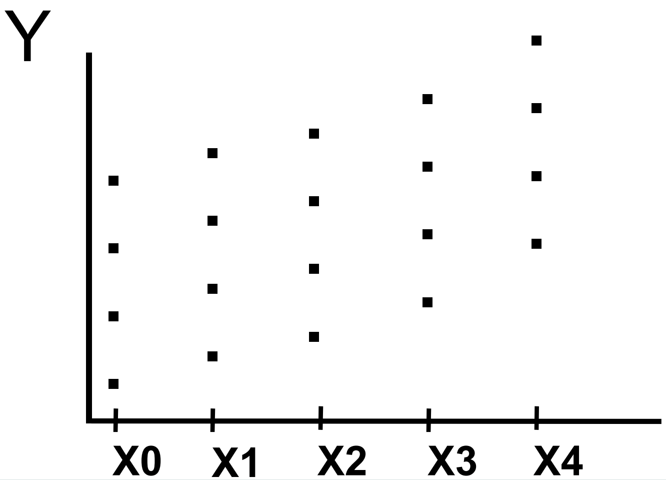
- **Neaditivita kombinovaného vlivu více pokusných zásahů.** Tuto situaci lze testovat jednak speciálními testy aditivity nebo přímo F testem kontrolujícím významnost vlivu interakce pokusných zásahů. Při významné interakci je nutné prozkoumat především její charakter ve vhodném experimentálním uspořádání.

Modely analýzy rozptylu

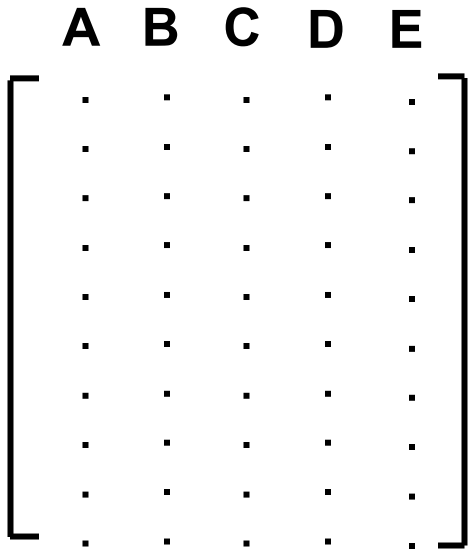
Model I. Pevný model



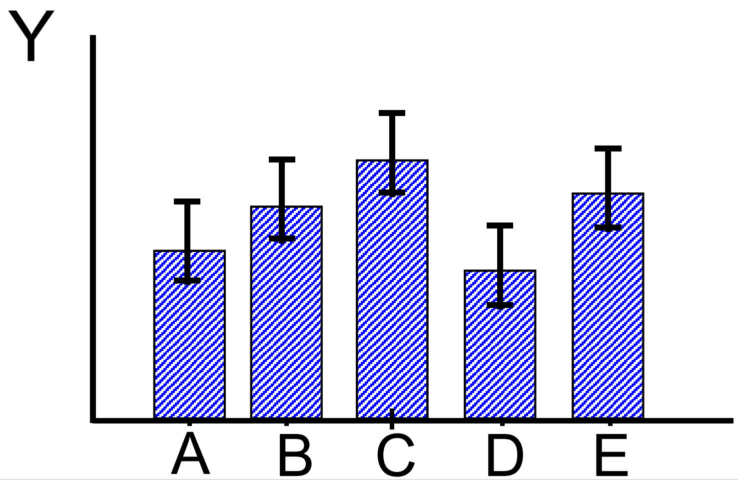
$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$



Model II. Náhodný model



$$y_{ij} = \mu + A_i + \varepsilon_{ij}$$



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

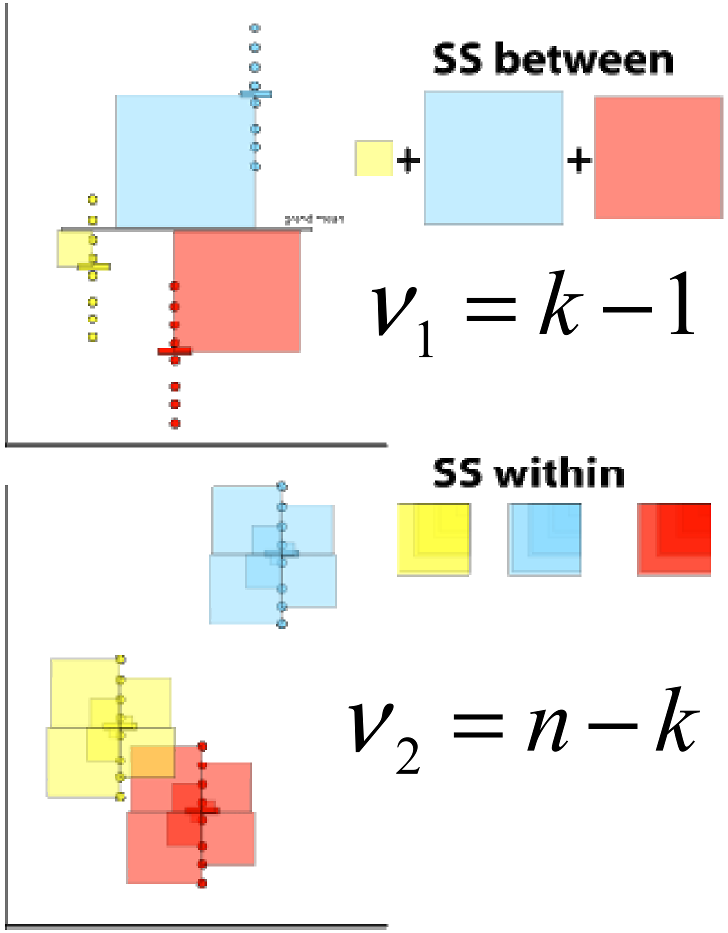
Rozptyl je počítán pro celkový průměr (tzv. grand mean) a průměry v jednotlivých skupinách dat

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)



$$\nu_1 = k - 1$$

$$\nu_2 = n - k$$

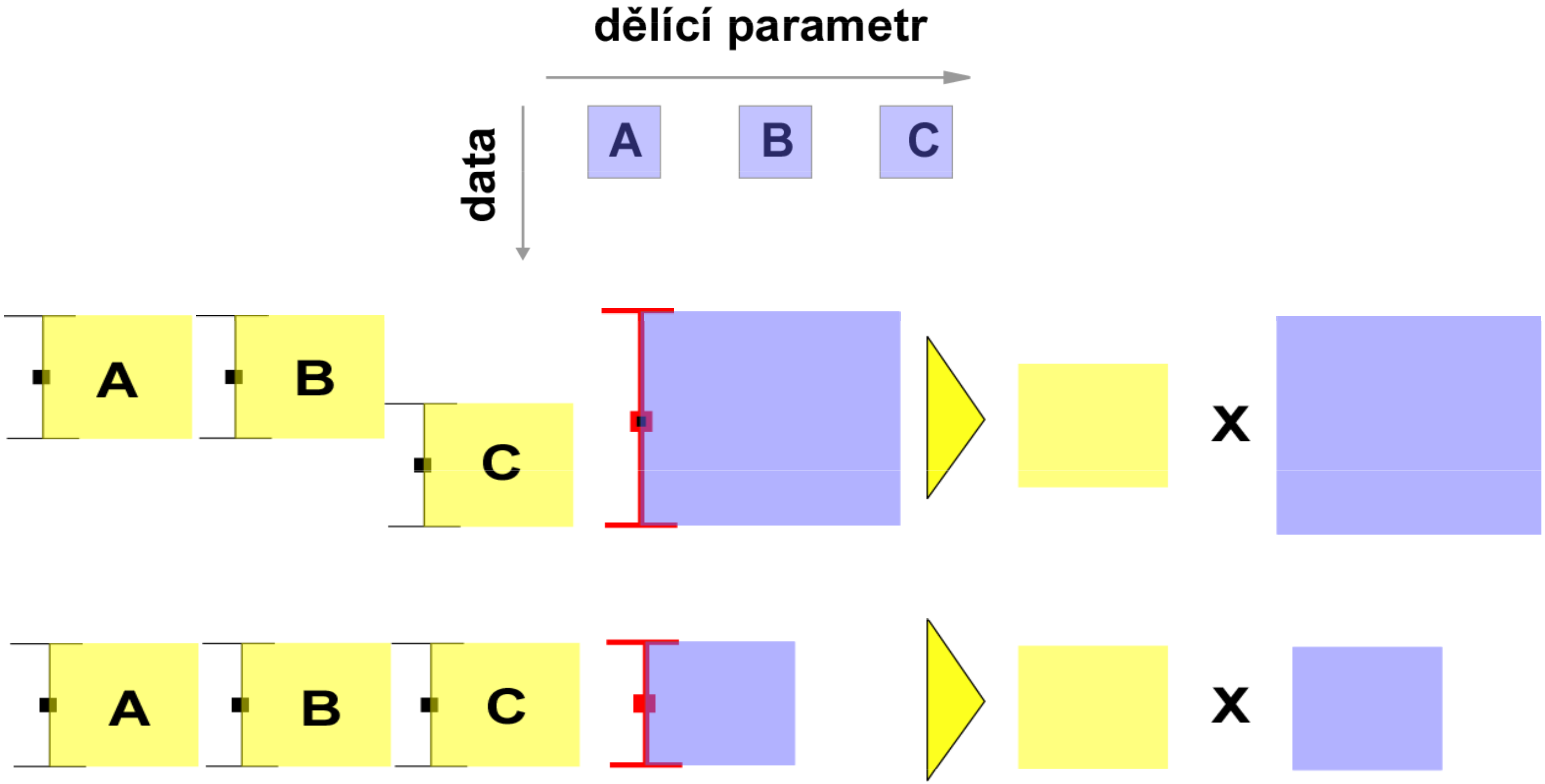
$$F = \frac{between\_groups}{within\_groups}$$

<u>Výsledný poměr (F) porovnáme s tabulkami F rozložení pro v_1 a v_2 stupňů volnosti</u>

SS=sum of squares

Jednoduchý ANOVA design

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

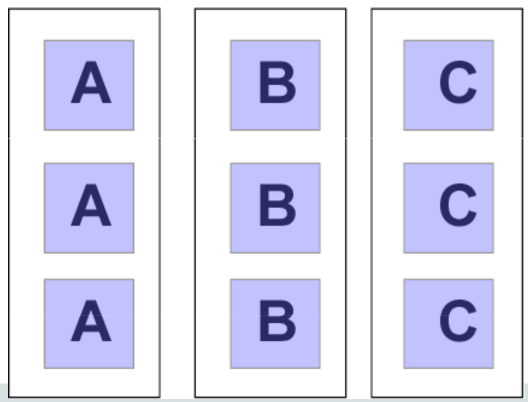


Nested ANOVA

- Rozdělení skupin na náhodné podskupiny (např. opakování experimentu), podskupiny jsou vždy v jedné skupině (ne kartézský součin).
- 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

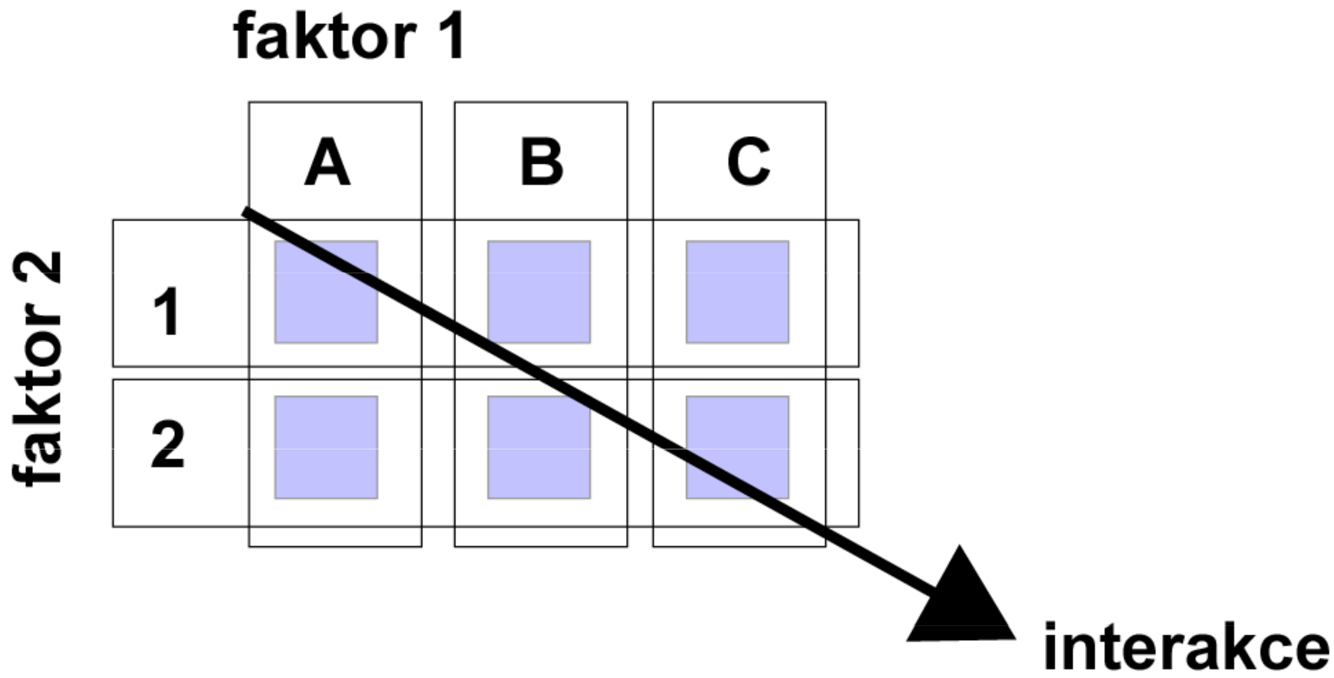


Two way ANOVA

Pro rozdělení do kategorií je zde více parametrů (možné jsou všechny varianty kartézského součinu).

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
<u>Tabulka ANOVA</u> - 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		

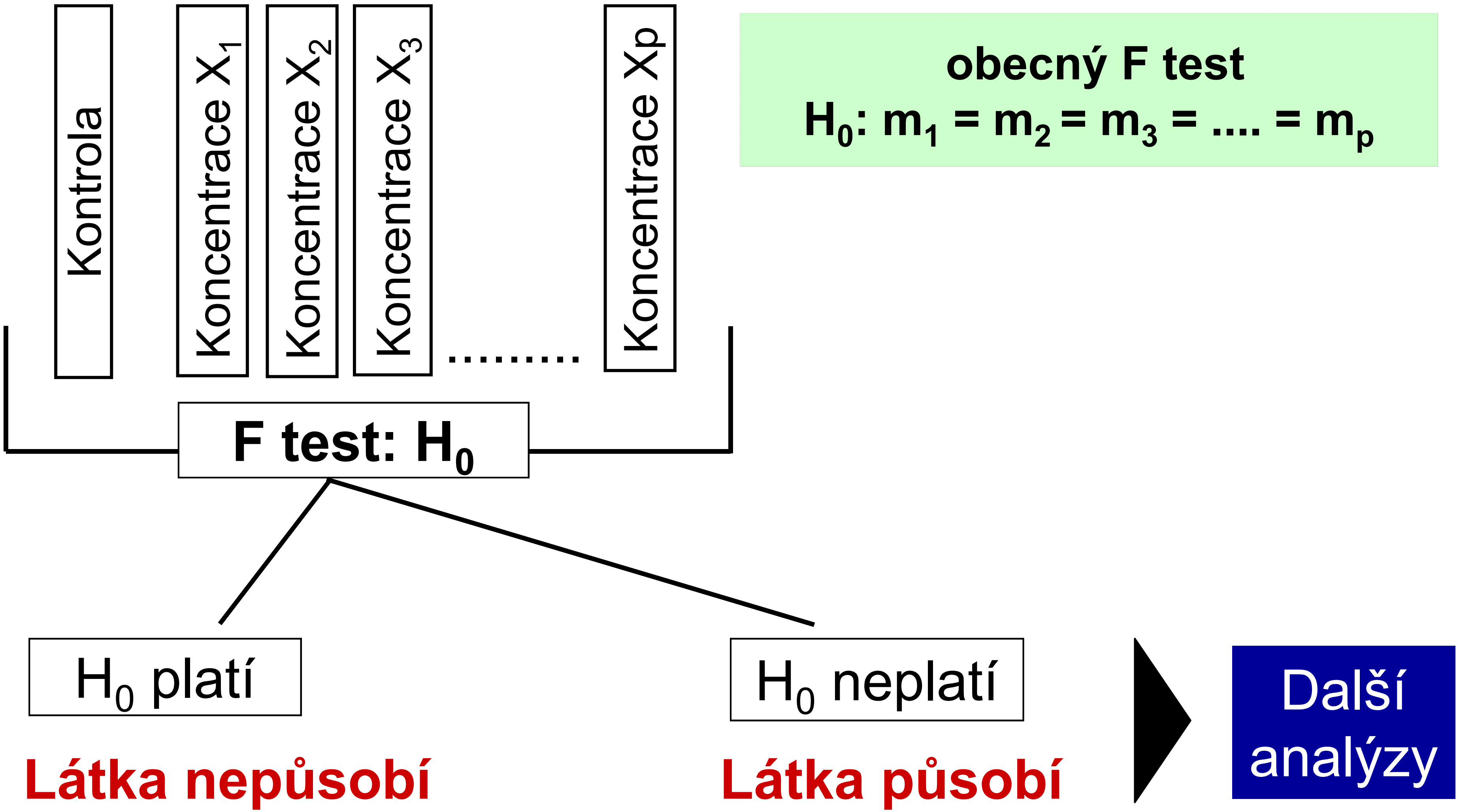
SS_B/SS_T → Kvantifikovaný podíl rozdílu mezi pokusnými zásahy na celkovém rozptylu

MS_B/MS_T → Statistická významnost rozdílu

Analýza rozptylu - obecný F test

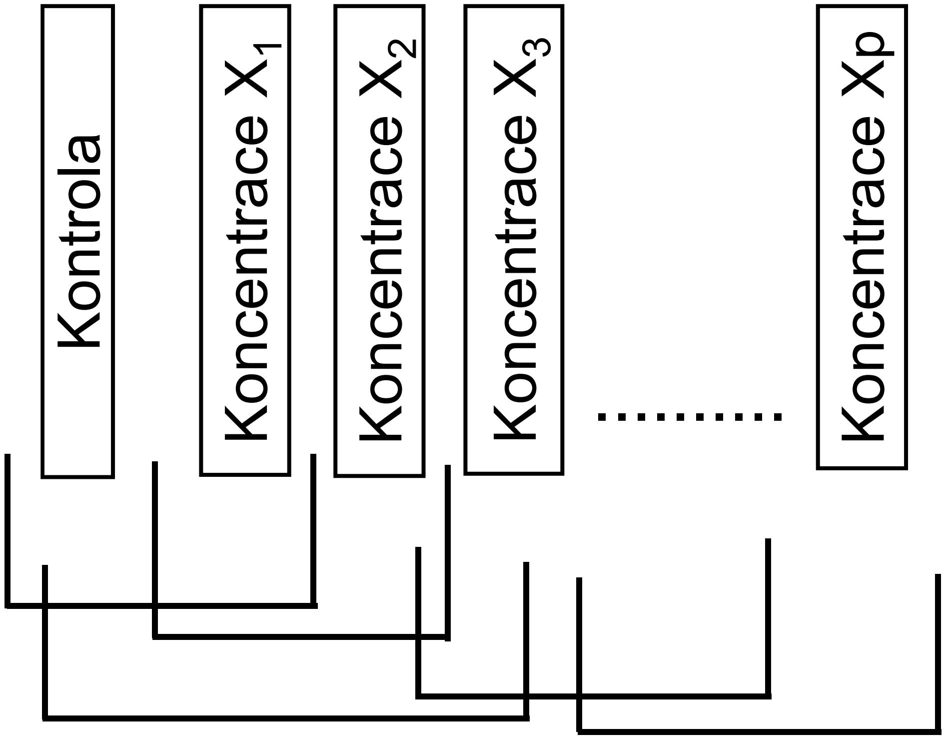
obecný F test

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



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

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

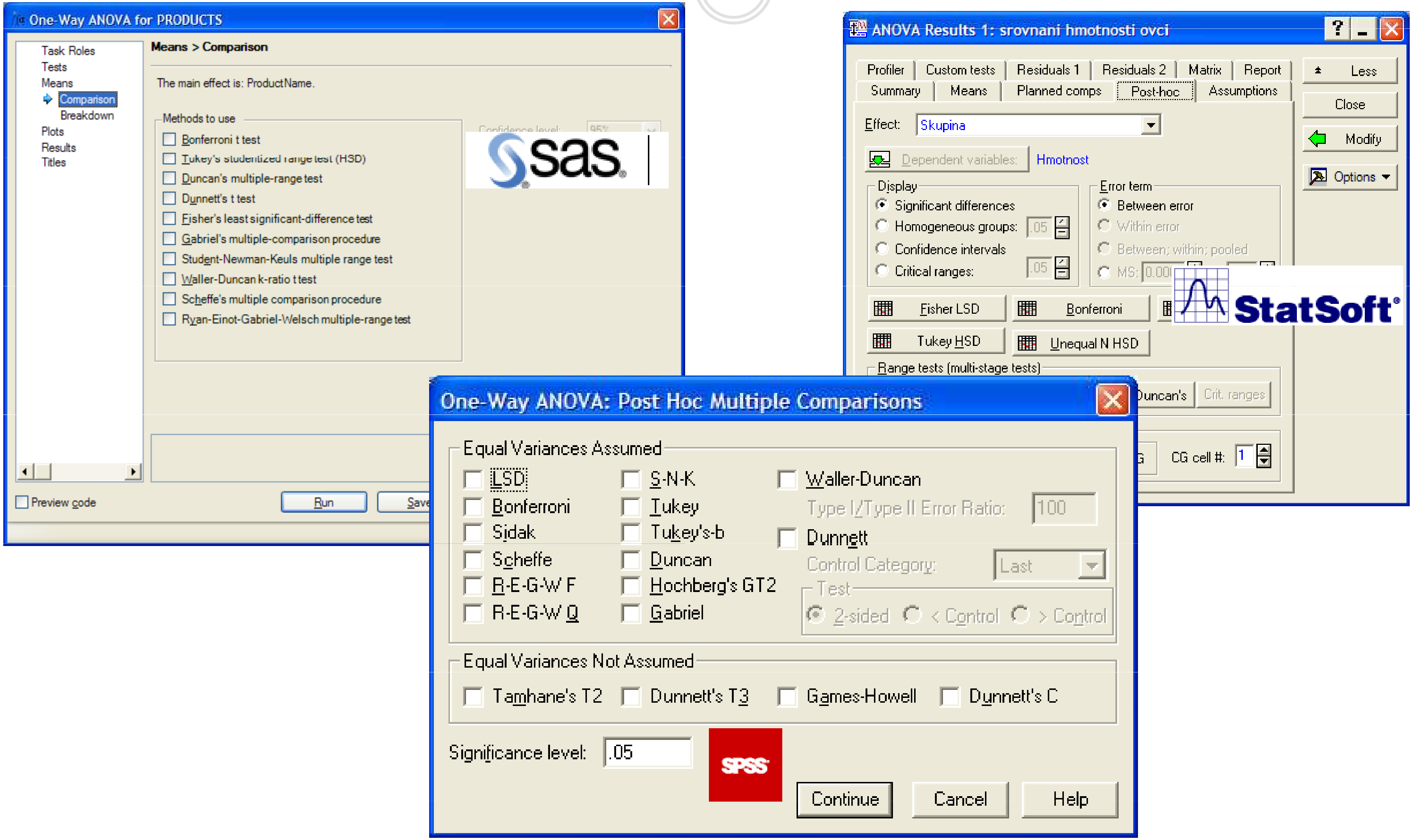
Testy pro jednoduché kontrasty

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

Testy nevhodné

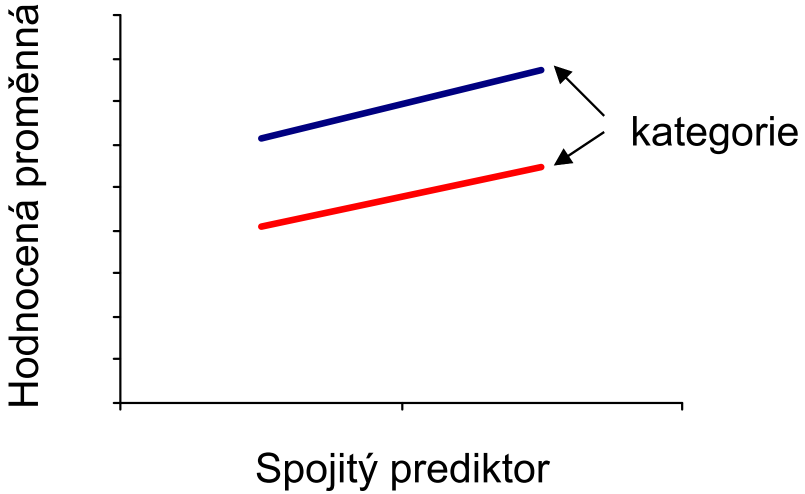
Duncan	**Student - Newmann-Keuls**	**Waller-Duncan k ratio**

Řada post-hoc testů v různých SW

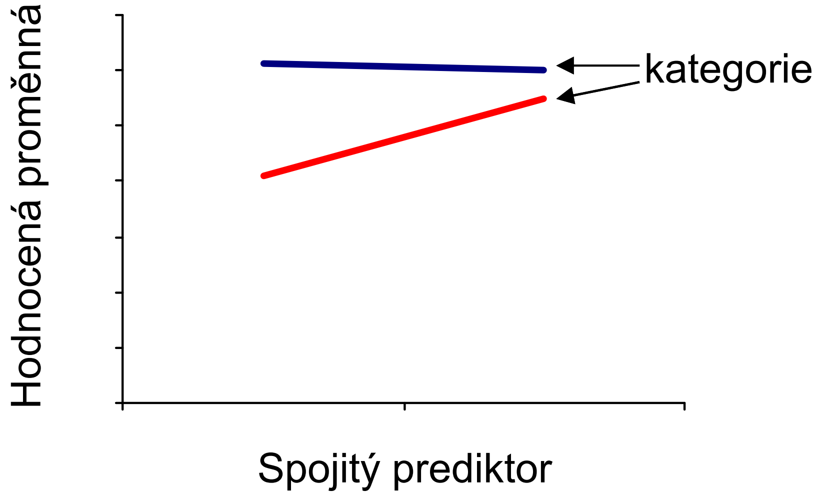


ANCOVA

- 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

Korelace a regrese

- 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

Korelace

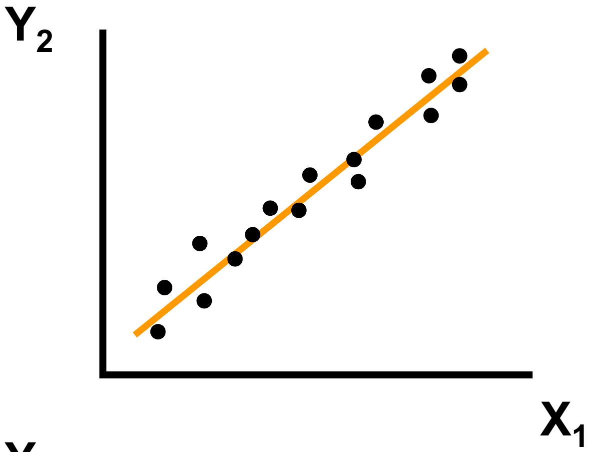
- K měření těsnosti lineárního vztahu 2 spojitých proměnných
  r = 0 → nekorelované
  r > 0 → kladně korelované
  r < 0 → záporně korelované
- H_0: proměnné X, Y jsou stochasticky nezávislé náhodné veličiny (r = 0)
  H_A: proměnné X, Y nejsou stochasticky nezávislé náhodné veličiny (r ≠ 0)
- Parametrický korelační koeficient:
  Pearsonův kor. koef. (dvourozměrné normální rozložení)
- Neparametrické korelační koeficienty:
  Spearmanův (pořadový) kor. koef., Kendallovo tau.

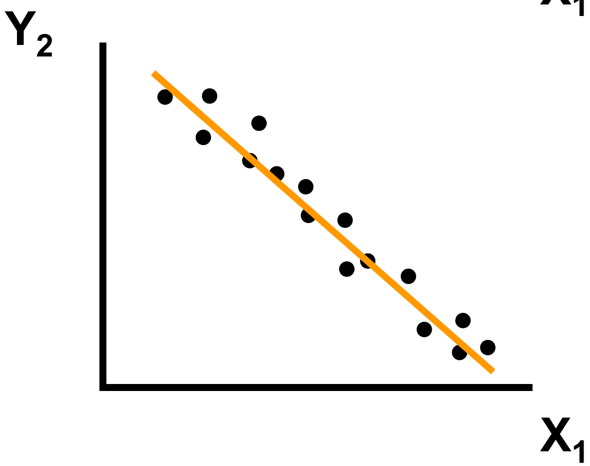
Jednoduchá lineární regrese

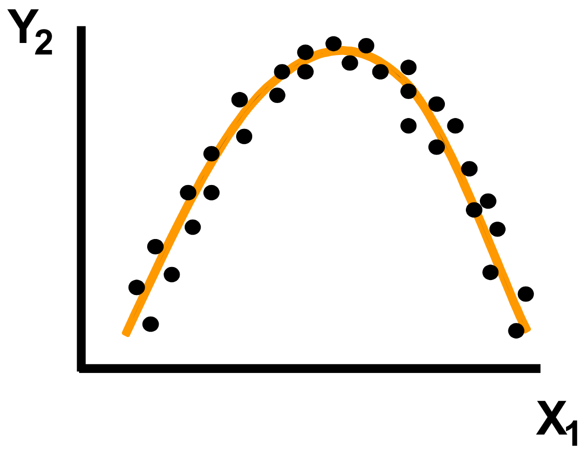
- V případě existence vzájemného vztahu (korelace) lze tento vztah podrobněji popsat.
- Cíl regresní analýzy: popsat závislost hodnot proměnné Y na hodnotách proměnné X.
- V případě lineární regrese je tento popis dán lineárním modelem tvaru $y = ax + b$.
- Existují i techniky nelineární regrese.
- Nemáme-li dostatek informací k teoretickému souboru, snažíme se odhadnout typ funkce pomocí dvourozměrného diagramu.

Základy korelační analýzy - I.

Korelace – vztah (závislost) dvou znaků (parametrů)







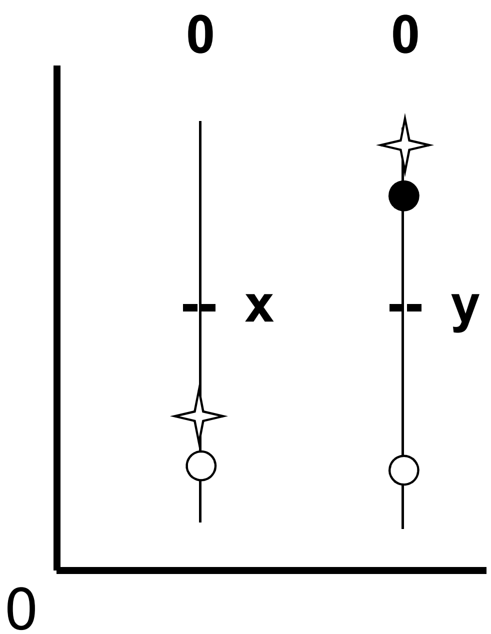
X_2 \ X_1	ANO	NE
ANO	a	b
NE	c	d

Základy korelační analýzy - II.

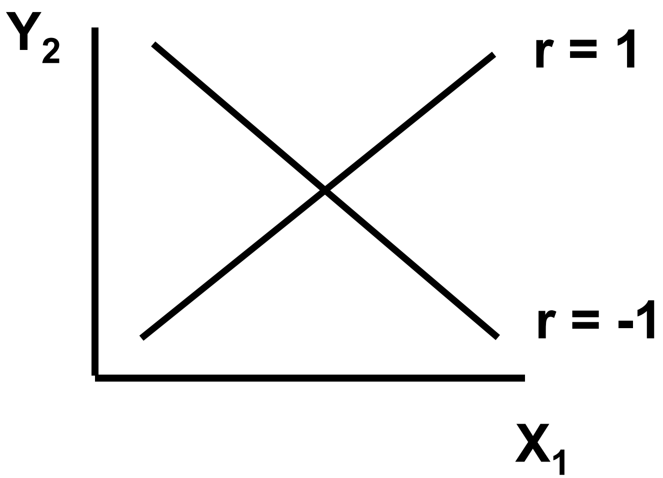
Parametrické míry korelace

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

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

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

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

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

II. $H_0 : \rho = \phi$

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

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

$P \leq 0{,}05$

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

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

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

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

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

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

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

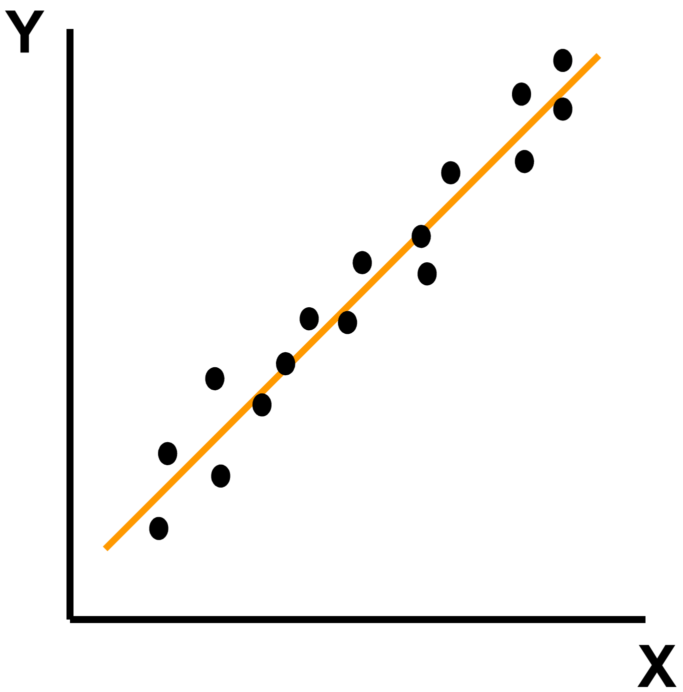
$$\textbf{tab}: \quad r_s(v = 6) = 0{,}89$$

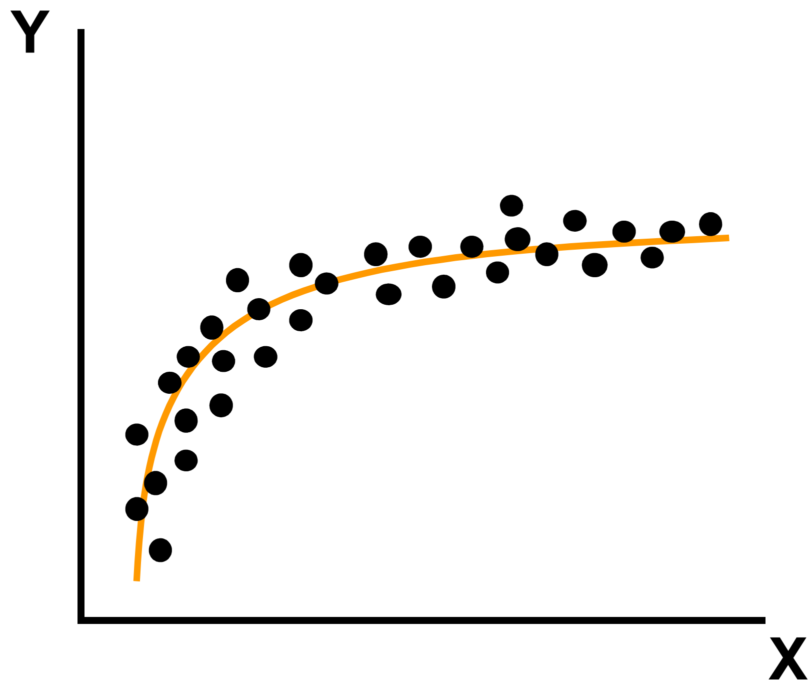
Pacient č.	1	2	3	4	5	6	7
Lékař 1	4	1	6	5	3	2	7
Lékař 2	4	2	5	6	1	3	7
d_I	0	-1	1	-1	2	-1	0

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

P = 0,358

Korelace v grafech I.





Vztahy velmi často implikují funkční vztah mezi Y a X.

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

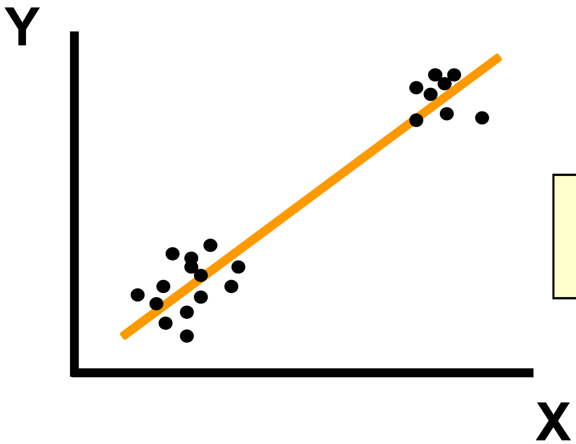
$$Y = a + b_1 \,.\, X_1 + b_2 \,.\, X_2 + b_3 \,.\, X_3$$

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

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

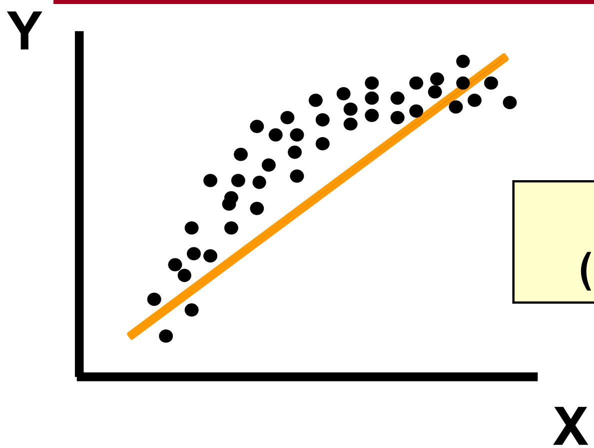
Korelace v grafech II.

Problém rozložení hodnot



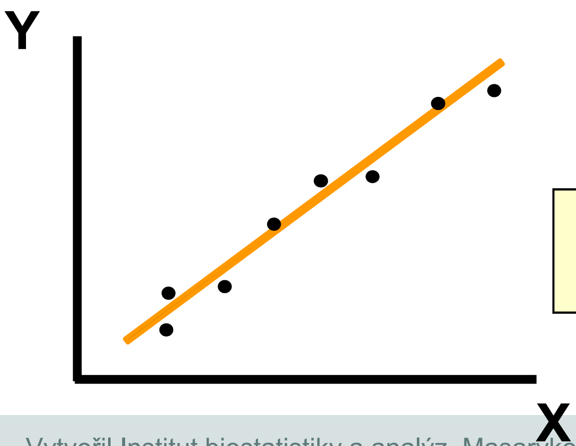
r = 0,981
(p < 0,001)

Problém typu modelu

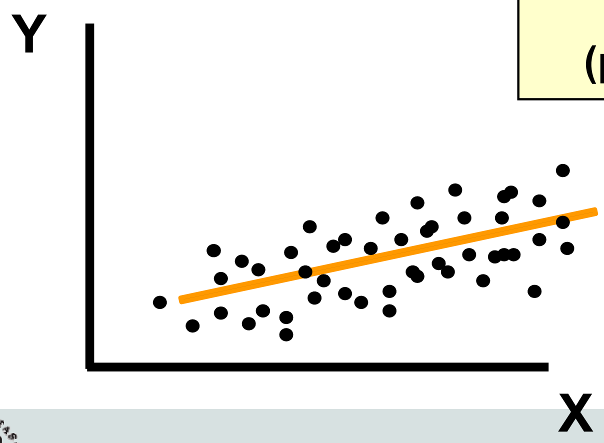


r = 0,761
(p < 0,032)

Problém velikosti vzorku



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



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