Přednáška 9

Poissonovo rozdělení

Popis rozložení a jeho využití

Anotace

- Poissonovo rozdělení se používá pro popis četnosti výskytu jevu na experimentální jednotku, příkladem může být počet mutací bakterií na Petriho misku nebo počet srdečních poruch na jednotku času

Poissonovo rozdělení

Celkový počet jevů v n nezávislých pokusech

E(x) = n p
D(x) = n p } E(x) = D(x)

$$P(r) = \frac{e^{-\mu} \cdot \mu^r}{r!} = e^{-\lambda} \cdot \frac{\lambda^r}{r!}$$

$\mu = \lambda =$ průměrný počet jevů z n pokusů

$$P(X = 0) = e^{-\mu}$$

$$P(X = 1) = e^{-\mu} \cdot \mu^2$$

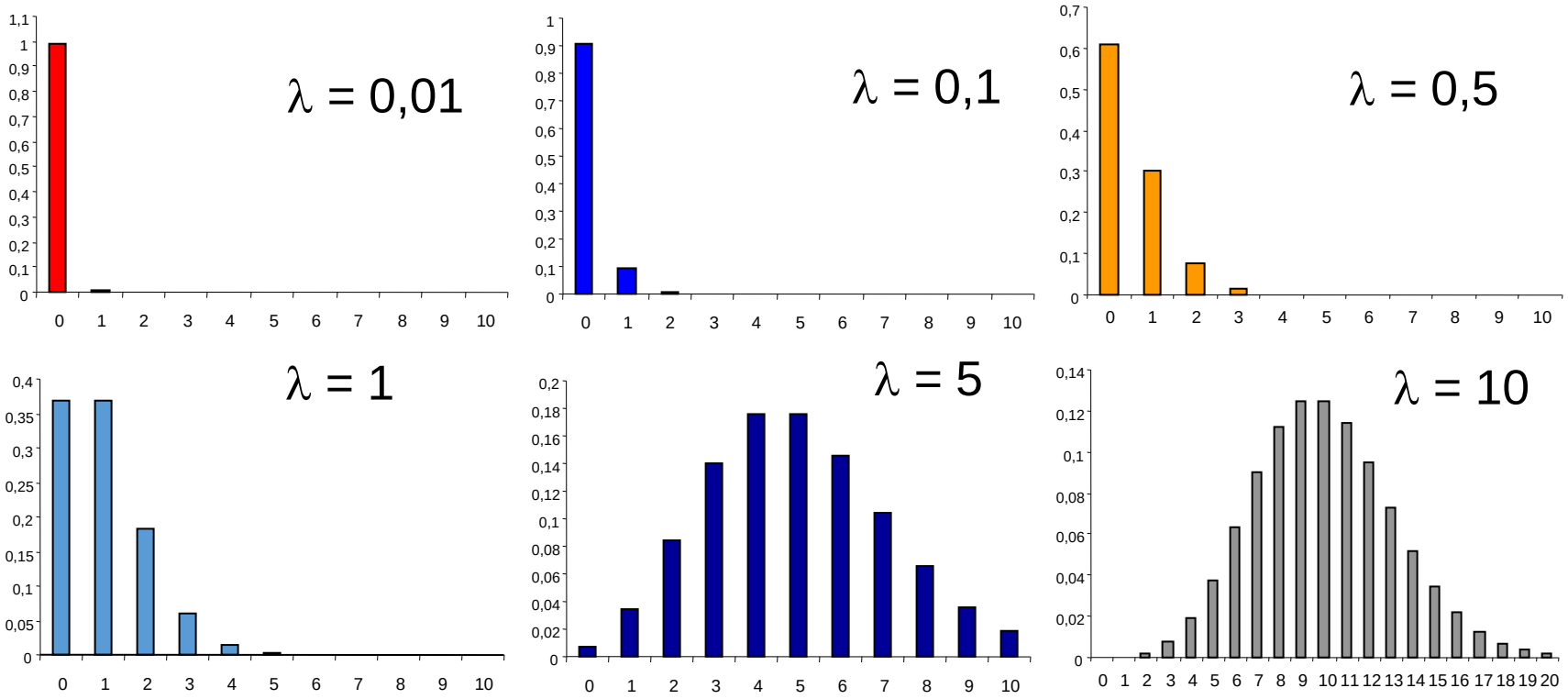
$$P(X = 2) = \frac{e^{-\mu} \cdot \mu^2}{2}$$

$$P(X = 3) = \frac{e^{-\mu} \cdot \mu^3}{(3)(2)}$$

$$P(X = 4) = \frac{e^{-\mu} \cdot \mu^4}{(4)(3)(2)}$$

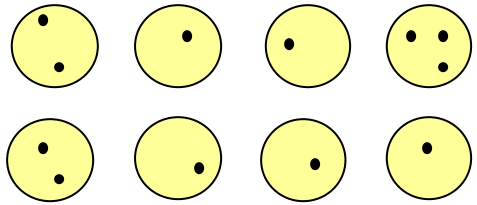
Poissonovo rozdělení jako model

$$\mathbf{P}(x = r) = e^{-\lambda} \cdot \frac{\lambda^r}{r!}$$



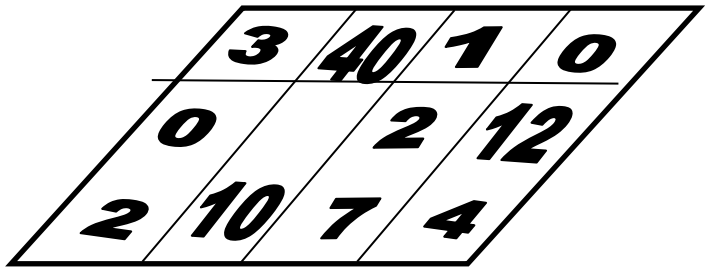
Poissonovo rozdělení v přírodě existuje

Mutace bakterií na inkubačních miskách



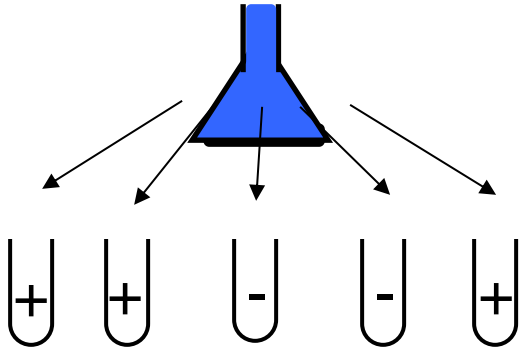
Výskyt jevu v prostoru

(počet žížal na určitou plochu pole)



Orientační stanovení jevu

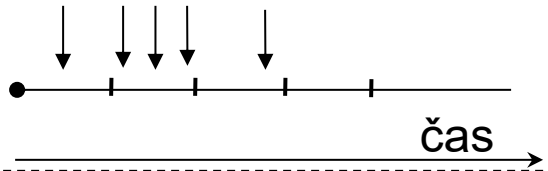
(při produkci plynu bakteriemi)



The most probable number technique

Výskyt jevu v čase

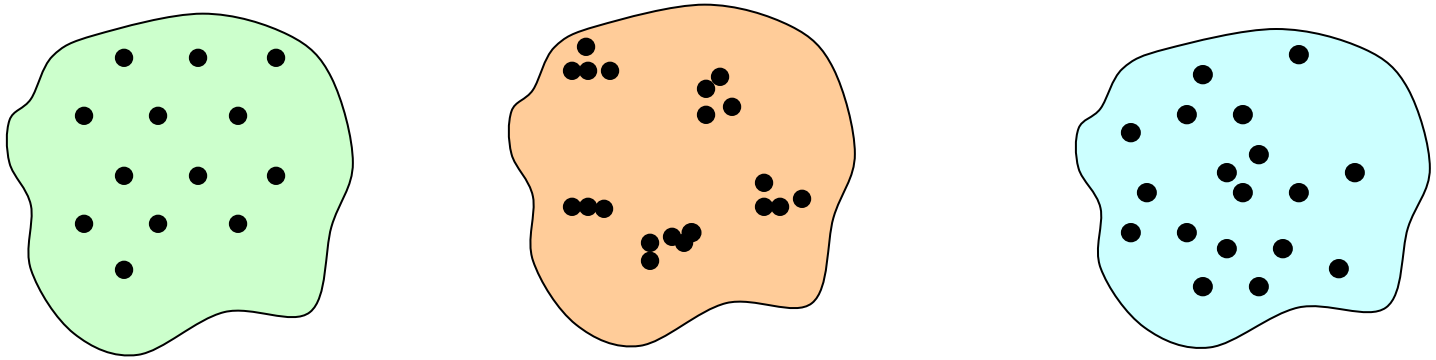
(srdeční arytmie v určitých časových intervalech)



Poissonovo rozdělení jako model pro náhodný výskyt jevů

Předpoklad: náhodná distribuce jevu mezi studovanými objekty
(příp. v čase, v prostoru).

$\sigma^2 < \mu$ $\sigma^2 > \mu$ $\sigma^2 = \mu$



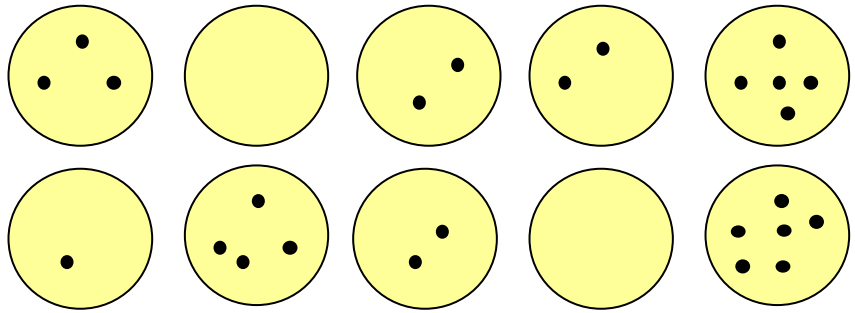
Uniform Clustered Random

Poisson

Pokud je λ spíše větší (~ 5 - 10), pak Poisson odpovídá spíše binomickému až normálnímu rozložení.

Formální prezentace Poissonova rozložení

Př: pokus.....10 000 bakterií na misce
n = 10 misek
Jev: mutace (r=25)
λ.....průměrný počet mutantů na jednu misku



$r = 25$

$\bar{x} \approx \lambda = 25/10 = 2{,}5$

95 % IS:

$$\bar{x} - Z_{1-\alpha/2} \cdot \sqrt{\frac{\bar{x}}{n}} \leq \lambda \leq \bar{x} + Z_{1-\alpha/2} \cdot \sqrt{\frac{\bar{x}}{n}}$$

$$2{,}5 - 1{,}96 \cdot \sqrt{0{,}25} \leq \lambda \leq 2{,}5 + 1{,}96 \cdot \sqrt{0{,}25}$$

$$1{,}52 \leq \lambda \leq 3{,}48$$

Poissonova náhodná proměnná

- Při měření počtu krvinek změněných určitou chorobou (relativně vzácné) je pozorován zředěný vzorek krve pod mikroskopem v komůrce rozdělené na stejně velká pole. Sledovaná veličina, udávající počet krvinek v i-tém poli může být považována za rozdělenou podle Poissonova rozložení:

- n = 169 = počet nezávislých pozorování proměnné
- r = 10 = počet pozorovaných krvinek

- Jaká je hodnota parametru ? Poissonova rozložení a jaká je jeho interpretace ?
- Jaký je interval 95% spolehlivosti pro parametr λ

- Pokud bychom sledovali celkový počet červených krvinek (opět v n = 169 nezávislých políčkách), bylo by i tuto proměnnou možno považovat za rozloženou podle Poissonova rozložení ? Uvažujte celkový počet pozorovaných krvinek jako 2013.

Výpočet intervalu spolehlivosti pro λ (bez aproximace na normální rozložení)

Spodní hranice IS

$$L_1 = \frac{\chi^2_{(f_1=2r)\,1-\alpha/2}}{2}$$

Horní hranice IS

$$L_2 = \frac{\chi^2_{(f_2=f_1+2)\,\alpha/2}}{2}$$

Poissonova náhodná proměnná

Konstantní zářič: n = 2608 časových intervalů (každý 7,5 s)
i: počet částic v intervalu (x)
s_i: pozorovaná četnost intervalů s i částicemi

$$\mathbf{P}(x=i)=\frac{\lambda^i\cdot e^{-\lambda}}{i!}\sim p_i$$

Poissonova proměnná:

* Výborný model pro experimenty, v nichž je během časového průběhu zjišťován počet výskytu určitého jevu

i	Počet intervalů s právě i zaznamenanými částicemi s_t	teoretické četnosti np_i	$\frac{(s_i-np_i)^2}{np_i}$
0	57	54,399	0,1244
1	203	210,523	0,2688
2	383	407,361	1,4568
3	525	525,496	0,0005
4	532	508,418	1,0938
5	408	393,515	0,5332
6	273	253,817	1,4498
7	139	140,325	0,0125
8	45	67,882	7,7132
9	27	29,189	0,1642
10	10	17,075 (= P{ξ ≥ 10})	0,0677
11	4		
12	2		
13	0		
	n = 2608	2608,00	12,8849

Poissonovo rozdělení: jednovýběrový test

$$P_{(r)} = \frac{(e^{-\lambda} \cdot \lambda^{r})}{r!}$$

Př: Počet hnízd křepelek na dané ploše

n = 8 000 "pod lokalit"
r = 28
} $\hat{p} = 0{,}0035$

Nechť je srovnávací soubor
(předchozí průzkum)

$$p_o = 0{,}0020$$

$$p_o \cdot 8\,000 = 16 = \mu = \lambda$$

$$H_o: \quad p \le p_o \sim \mu \le 16 \quad ?$$

1) Vzít data jako pocházející z populace:

$$P(r = 28) = \frac{e^{-16} \cdot 16^{28}}{28!} = 0{,}00192$$

2) $P(r \ge 28) = ?$
$[0{,}00411]$ } $< 0{,}05$ => H_o zamítnuta

r = 28 je příliš velké pro populaci s p_o

aby r = 28 bylo pravděpodobnější

$p > p_o$,

Analýza rozptylu

Parametrická analýza rozptylu

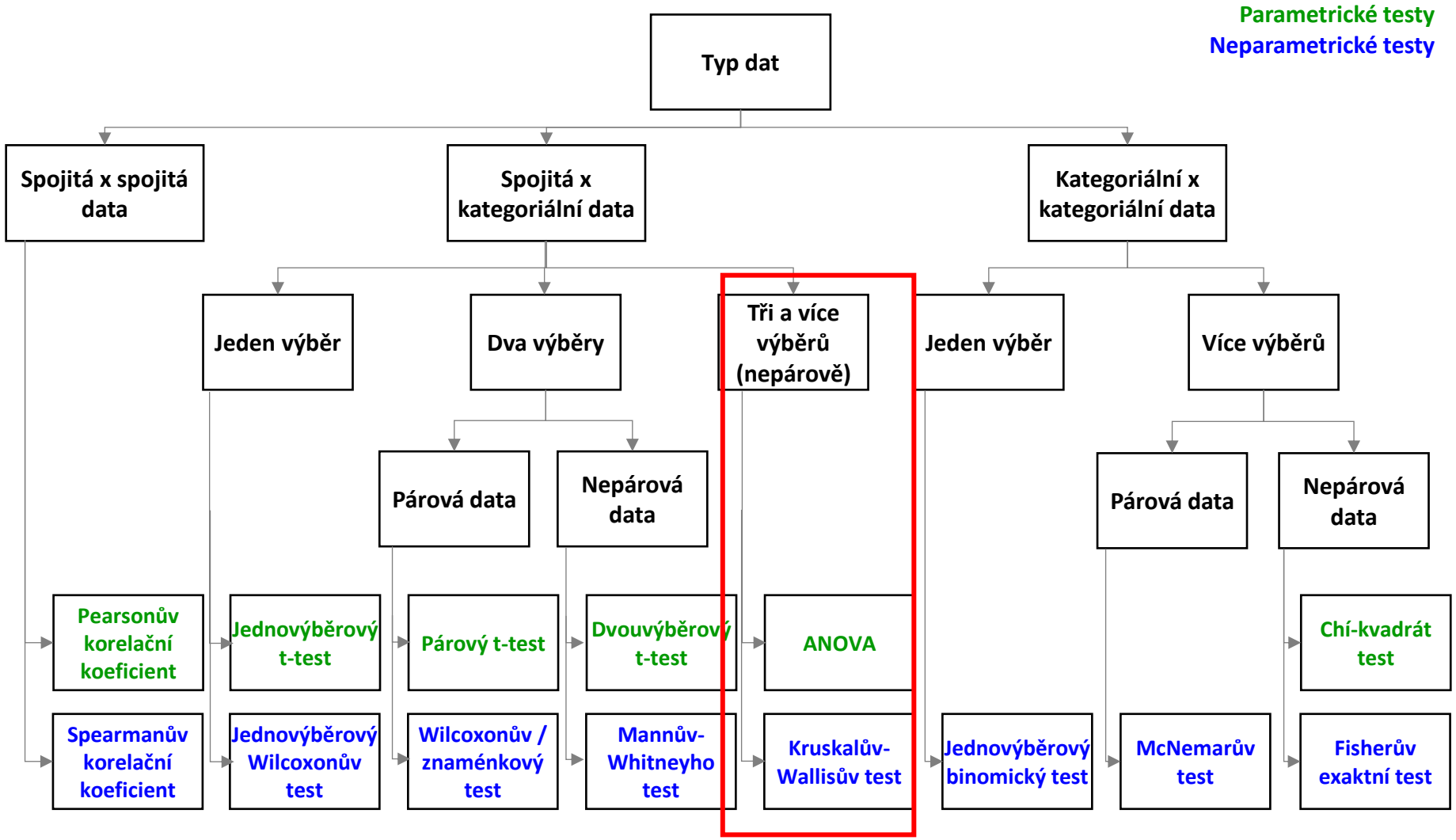
Post hoc testy

Kruskal-Wallis test

Anotace

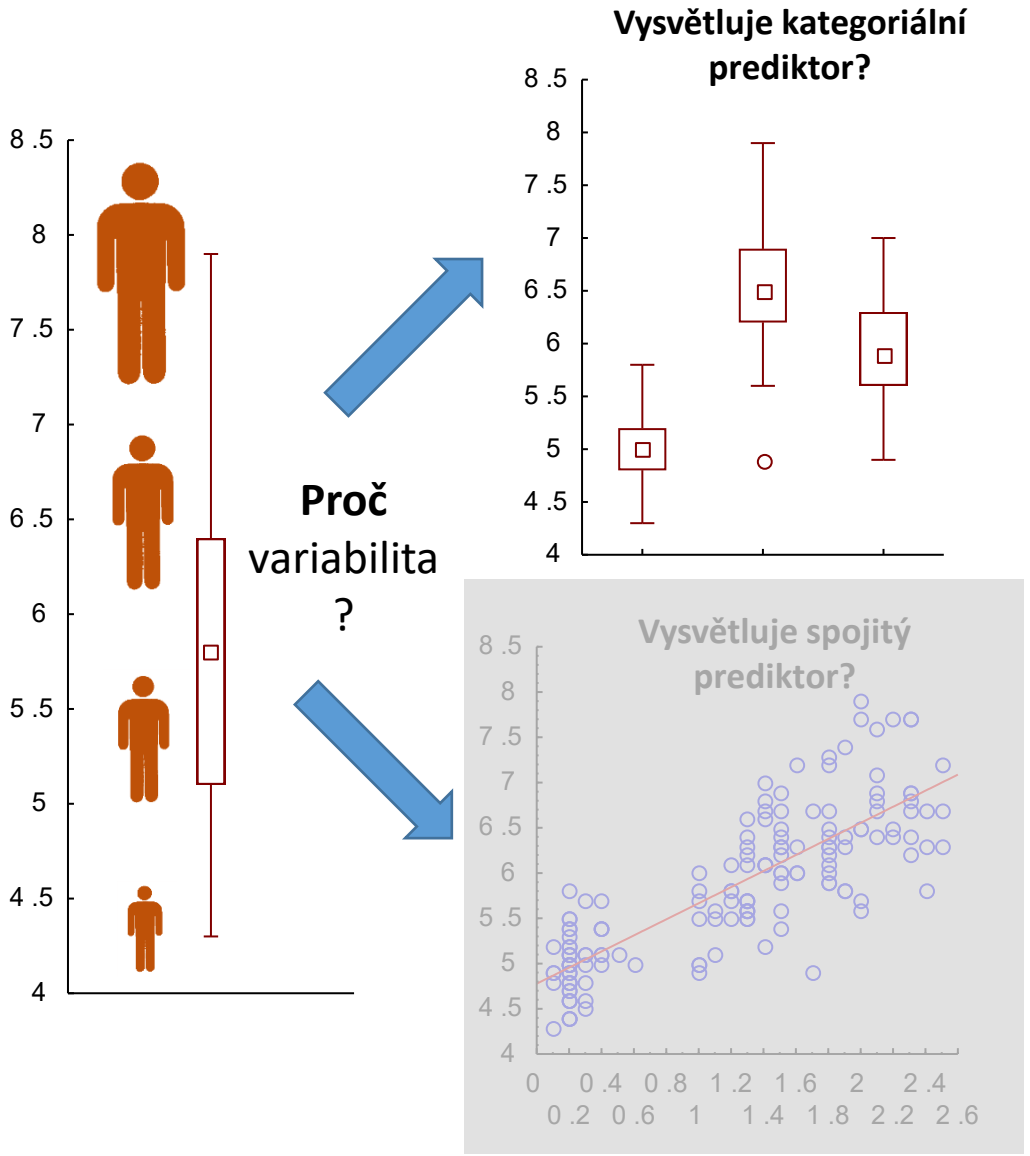
- Analýza rozptylu je základním nástrojem pro analýzu rozdílů mezi průměry v několika skupinách objektů.
- Základní myšlenka, na níž je ANOVA založena, je rozdělení celkové variability v datech (neznámé, dané pouze náhodným rozložením) na část systematickou (spjatou s kategoriemi pacientů, vysvětlená 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

Základní rozhodování o výběru statistických testů



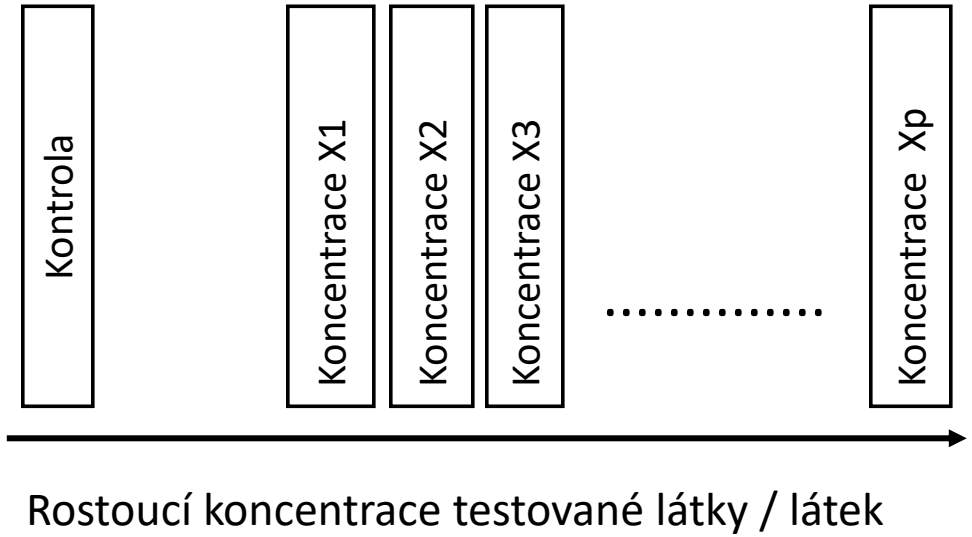
Cíl stochastického modelování

- 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



Analýza rozptylu - ANOVA

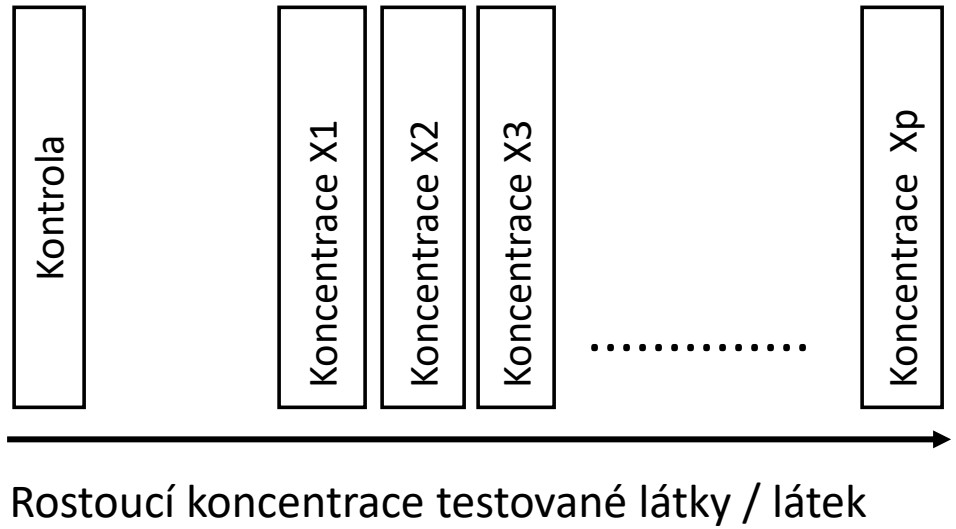
Základní technika sloužící k posouzení rozdílů mezi více úrovněmi pokusného zásahu



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

Analýza rozptylu - ANOVA

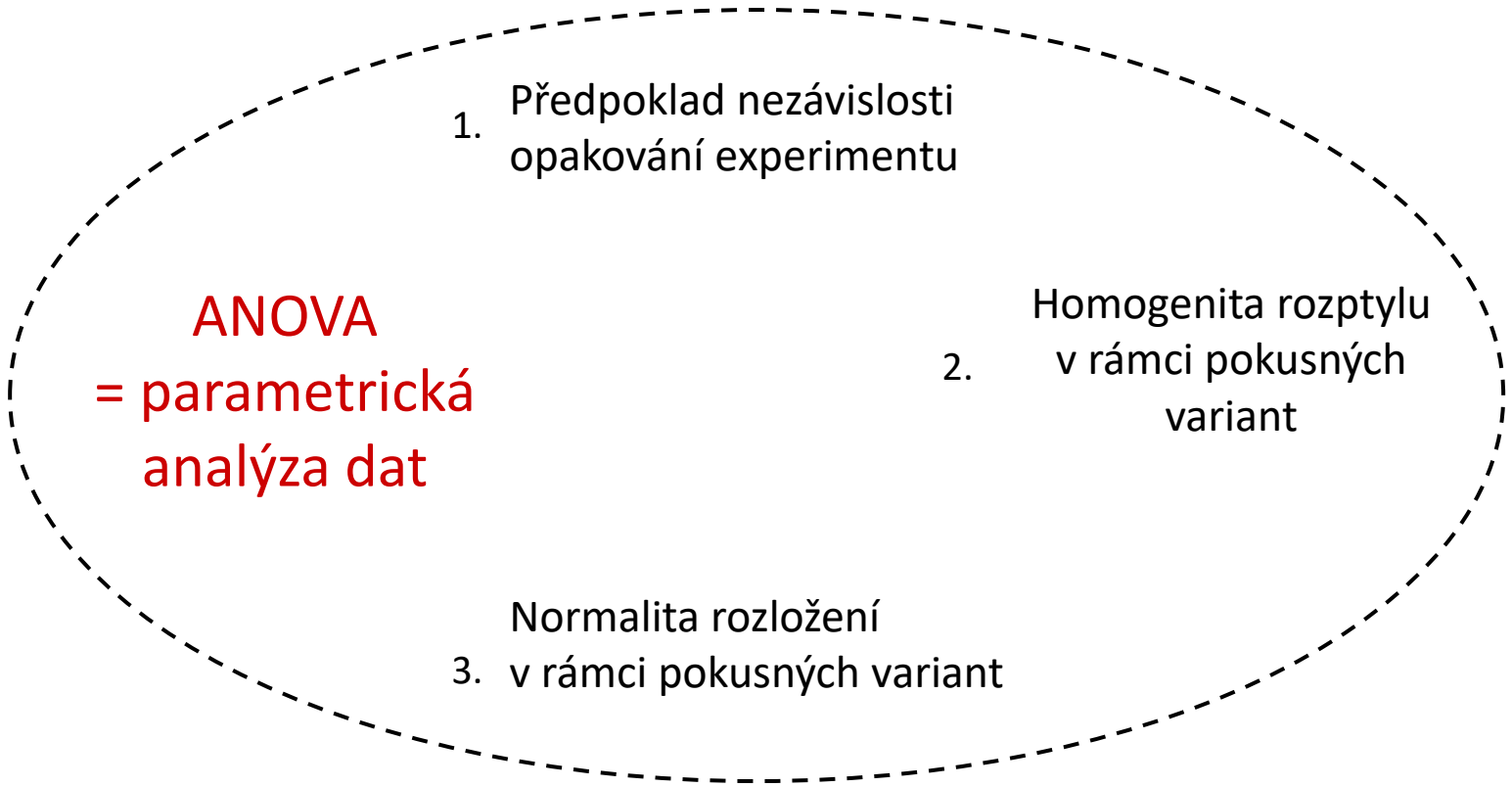
Významné kroky analýzy, vedoucí k efektivnímu srovnání variant



- 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



ALTERNATIVOU JSOU NEPARAMETRICKÉ METODY

ANOVA – předpoklady

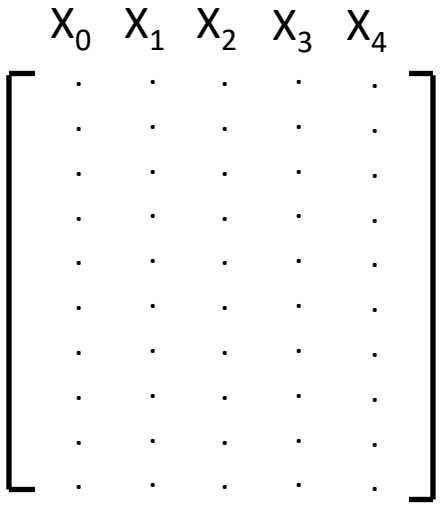
- Symetrické rozložení hodnot a normalita odchylek od hodnoceného modelu ANOVA. Velkou část dat lze adekvátně normalizovat použitím logaritmické transformace. Předpoklad lognormální transformace může pochopitelně být teoreticky vyloučen u mnoha datových souborů obsahujících diskrétní parametry, kde je 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.

Omezení aplikace ANOVA lze řešit

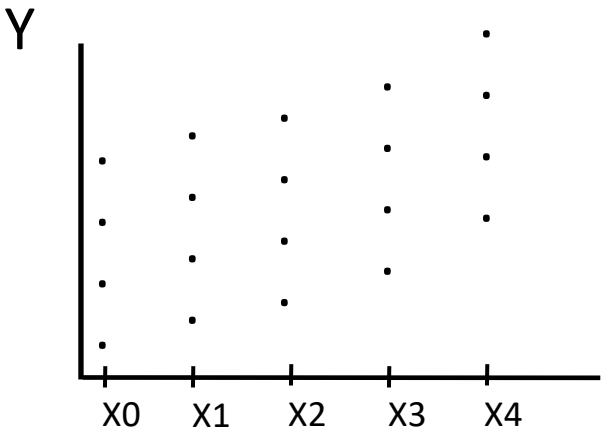
- **Chybějící data.** Vážným problémem jsou chybějící údaje o celé skupině kombinací testovaných látek, například u faktoriálních pokusů, kdy je znemožněno hodnocení experimentu jako celku.
- **Různé počty opakování.** Jde o typický jev pro experimentální datové soubory. Při různých počtech opakování 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.

Modely analýzy rozptylu

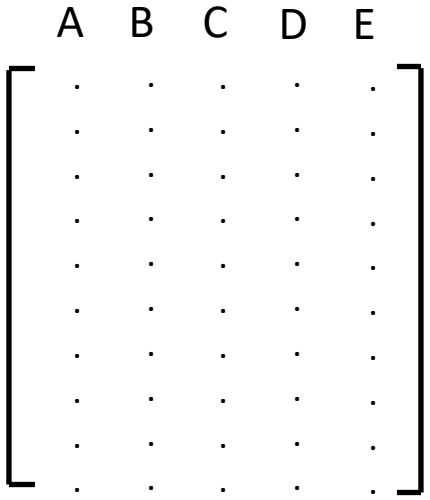
Model I. Pevný model



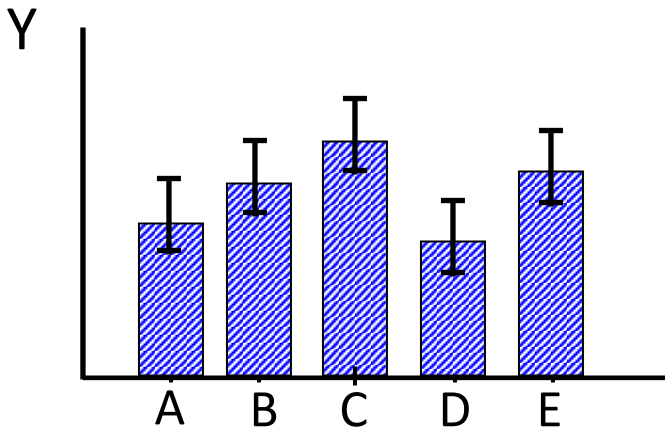
$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$



Model II. Náhodný model



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



Princip ANOVA

- 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

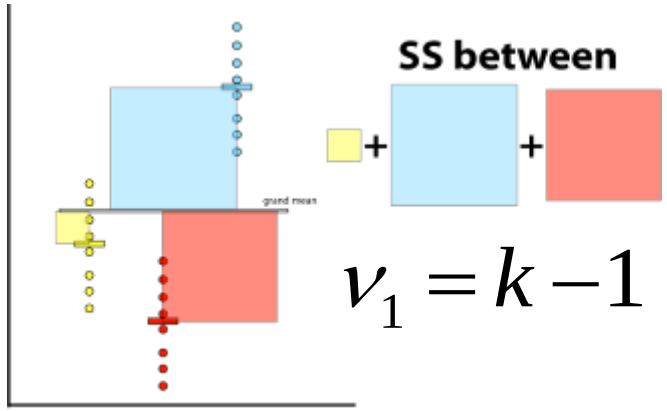
Rozptyl je počítán pro celkový průměr (tzv. grand mean) a průměry v jednotlivých skupinách dat

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)



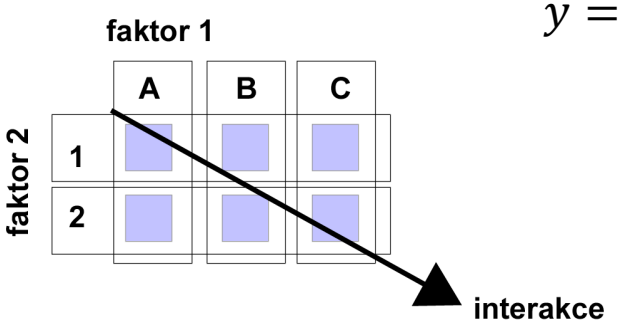
$$F = \frac{between\_groups}{within\_groups}$$

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

SS=sum of squares

Design modelu

- Design modelu znamená jaké proměnné a v jakých kombinacích budou vysvětlovat hodnocenou proměnnou
- Obecně je vhodné ať již expertně nebo jako výsledek předběžné analýzy vytvořit a ověřit hypotézy o vzájemných vztazích proměnných a podle těchto předběžných výsledků vytvářet finální model
- Tvorba designu modelu úzce souvisí s pojmy:
  - Analýza pouze hlavních efektů proměnných
  - Analýza interakcí mezi proměnnými a složitost interakcí
- Design modelu lze vyjádřit graficky nebo v rovnici nebo pomocí maticoveho zápisu

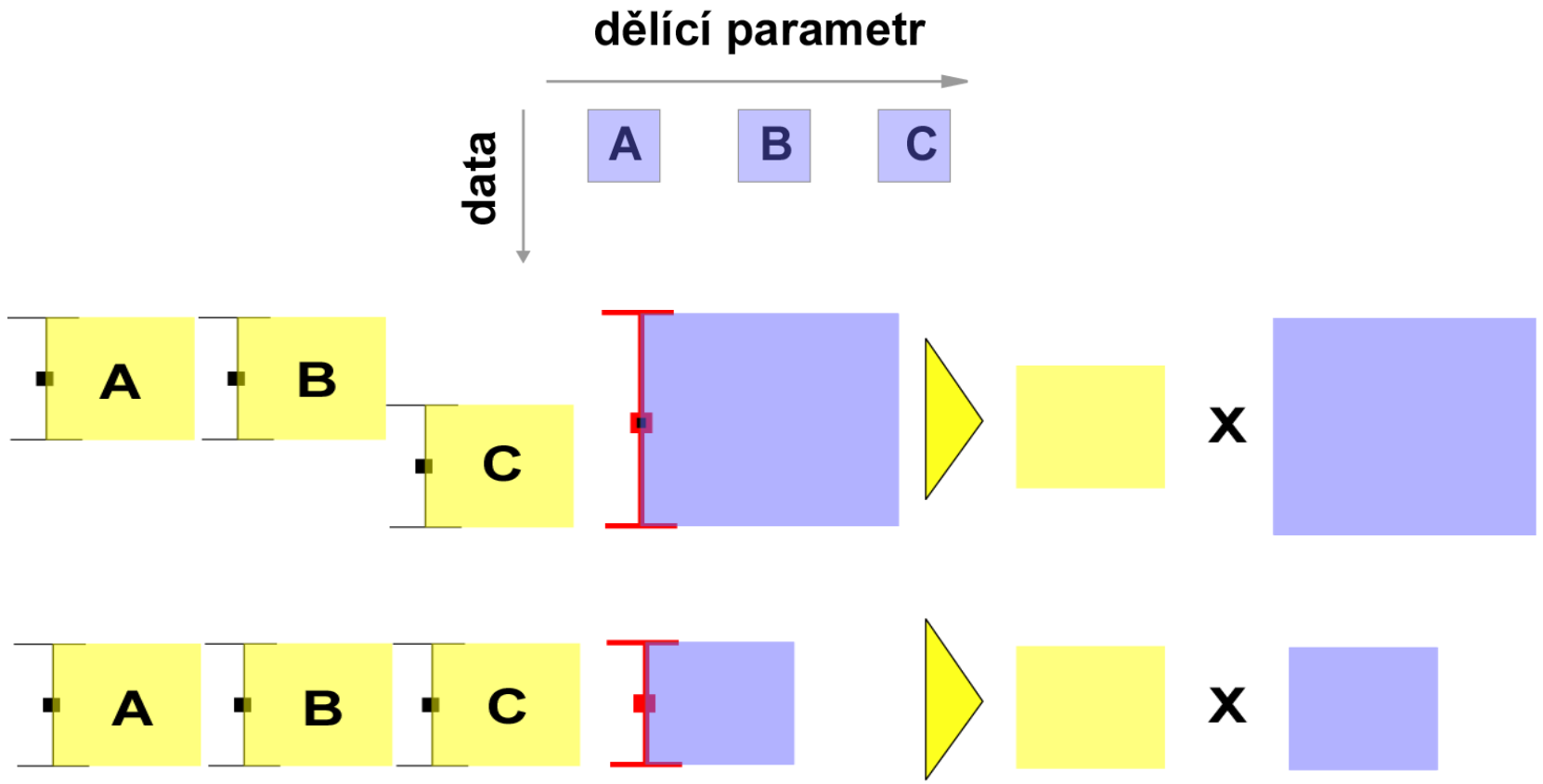


$$y = hmotnost * 1.5 + věk * 3.6 + hmotnost * věk * 1.8 + 9$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Jednoduchý ANOVA design

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

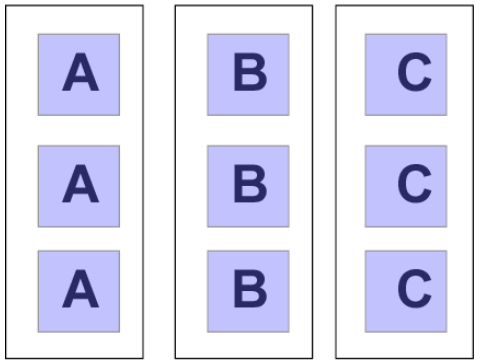


Nested ANOVA

- 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

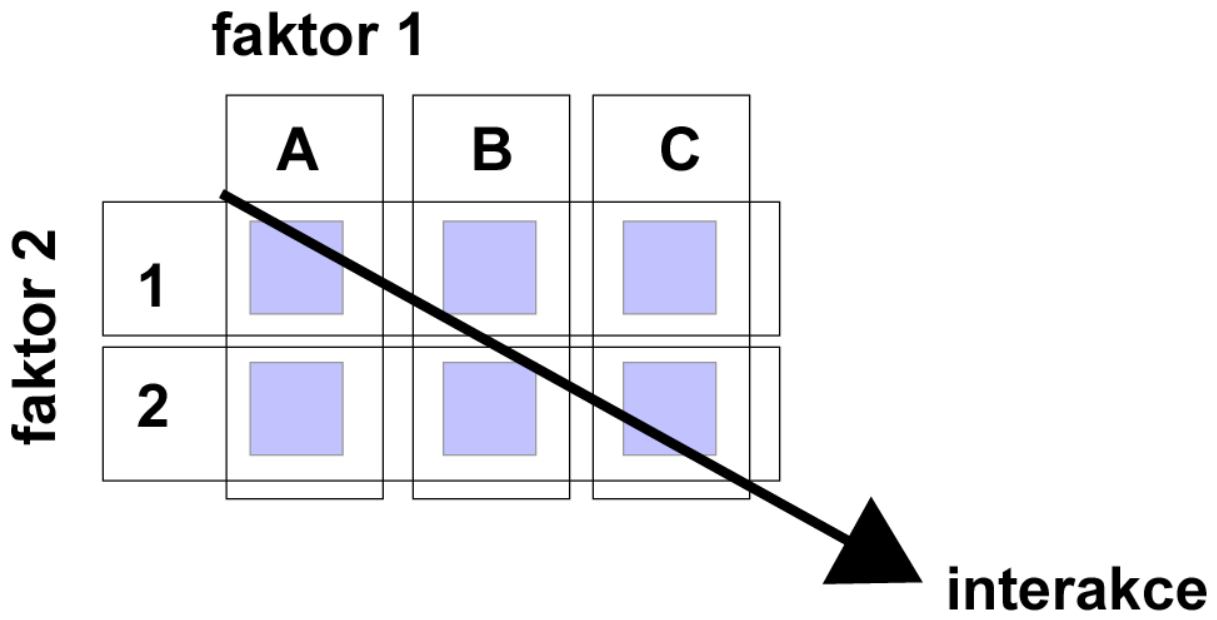
jednoduchá ANOVA

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 O2)
- Kromě vlivu hlavních faktorů se uplatňuje i jejich interakce



ANOVA – základní výstup

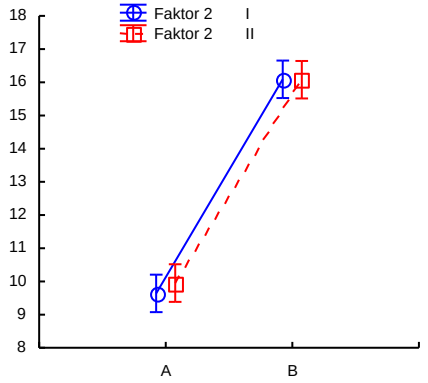
- Základním výstupem analýzy rozptylu je Tabulka ANOVA - frakcionace komponent rozptylu

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**		

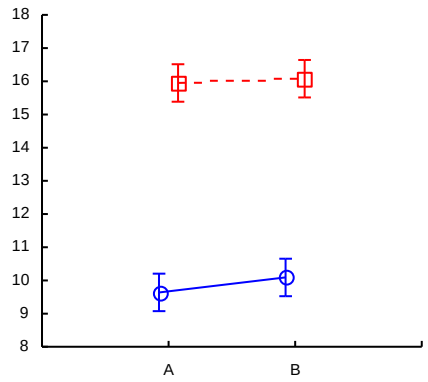
SS_B/SS_T → Kvantifikovaný podíl rozdílu mezi pokusnými zásahy na celkovém rozptylu

MS_B/MS_T → Statistická významnost rozdílu

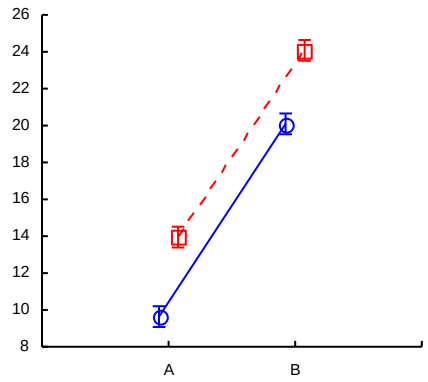
Hlavní efekty a interakce



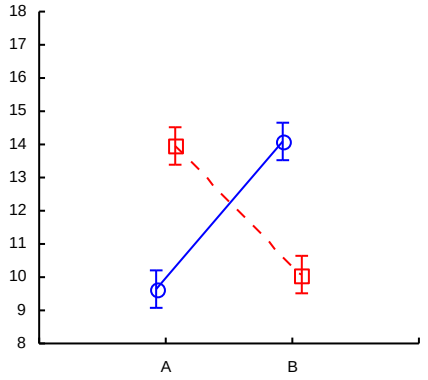
	SS	D.f.	MS	F	p
Intercept	33487	1	33487	8165.3	0.000
Faktor 1	**1978**	**1**	**1978**	**482.2**	**0.000**
Faktor 2	1	1	1	0.3	0.602
F1*F2	1	1	1	0.3	0.570
Error	804	196	4		



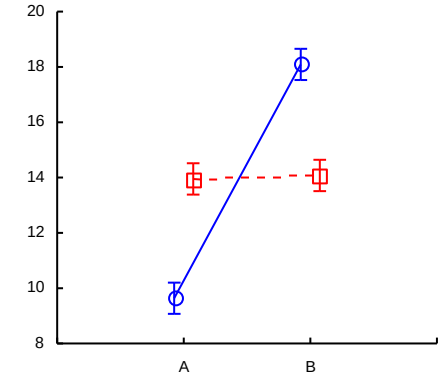
	SS	D.f.	MS	F	p
Intercept	33487	1	33487	8165.3	0.000
Faktor 1	4	1	4	1.0	0.314
Faktor 2	**1891**	**1**	**1891**	**461.1**	**0.000**
F1*F2	1	1	1	0.3	0.570
Error	804	196	4		



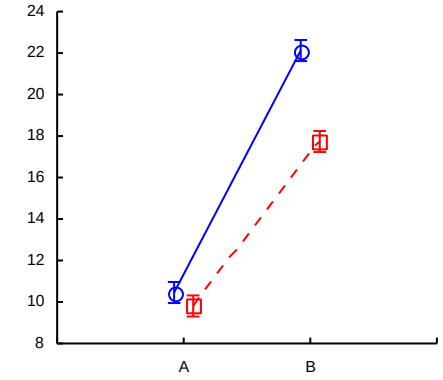
	SS	D.f.	MS	F	p
Intercept	57391	1	57391	13993	0.000
Faktor 1	5293	1	5293	1290.7	0.000
Faktor 2	861	1	861	209.9	0.000
F1*F2	1	1	1	0.3	0.570
Error	804	196	4		



	SS	D.f.	MS	F	p
Intercept	28511	1	28511	6952.0	0.000
Faktor 1	4	1	4	1.0	0.314
Faktor 2	1	1	1	0.3	0.602
F1*F2	**867**	**1**	**867**	**211.3**	**0.000**
Error	804	196	4		

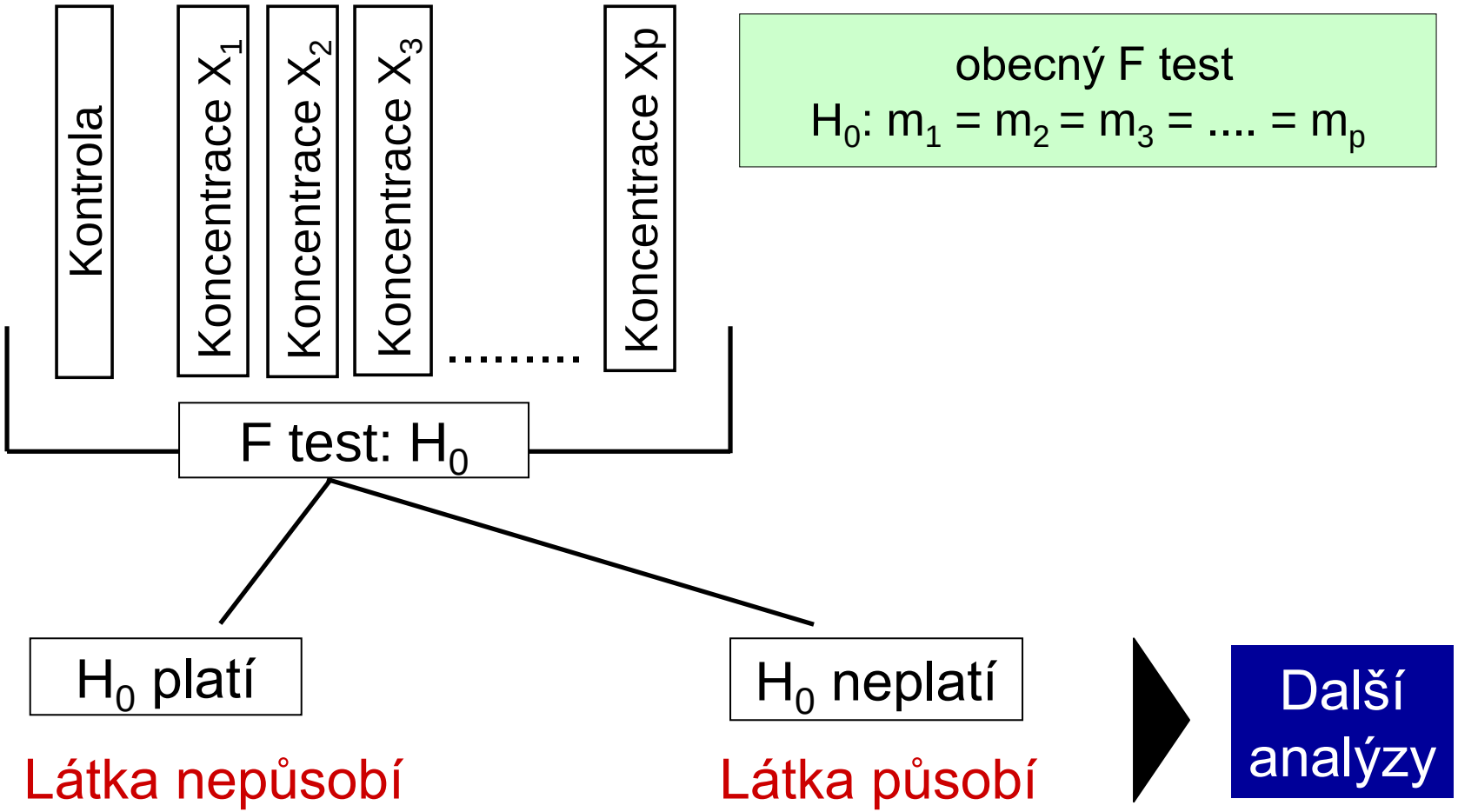


	SS	D.f.	MS	F	p
Intercept	38863	1	38863	9476.2	0.000
Faktor 1	**920**	**1**	**920**	**224.3**	**0.000**
Faktor 2	1	1	1	0.3	0.602
F1*F2	**867**	**1**	**867**	**211.3**	**0.000**
Error	804	196	4		



	SS	D.f.	MS	F	p
Intercept	45203	1	45203	13596	0.000
Faktor 1	**4799**	**1**	**4799**	**1443.4**	**0.000**
Faktor 2	**316**	**1**	**316**	**95.0**	**0.000**
F1*F2	**175**	**1**	**175**	**52.5**	**0.000**
Error	652	196	3		

Analýza rozptylu - obecný F test

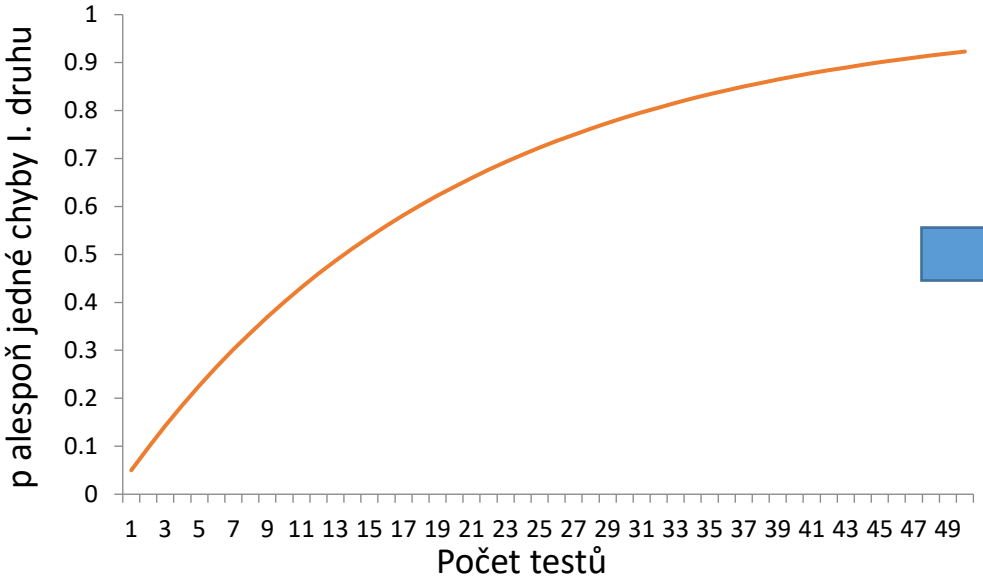


obecný F test

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

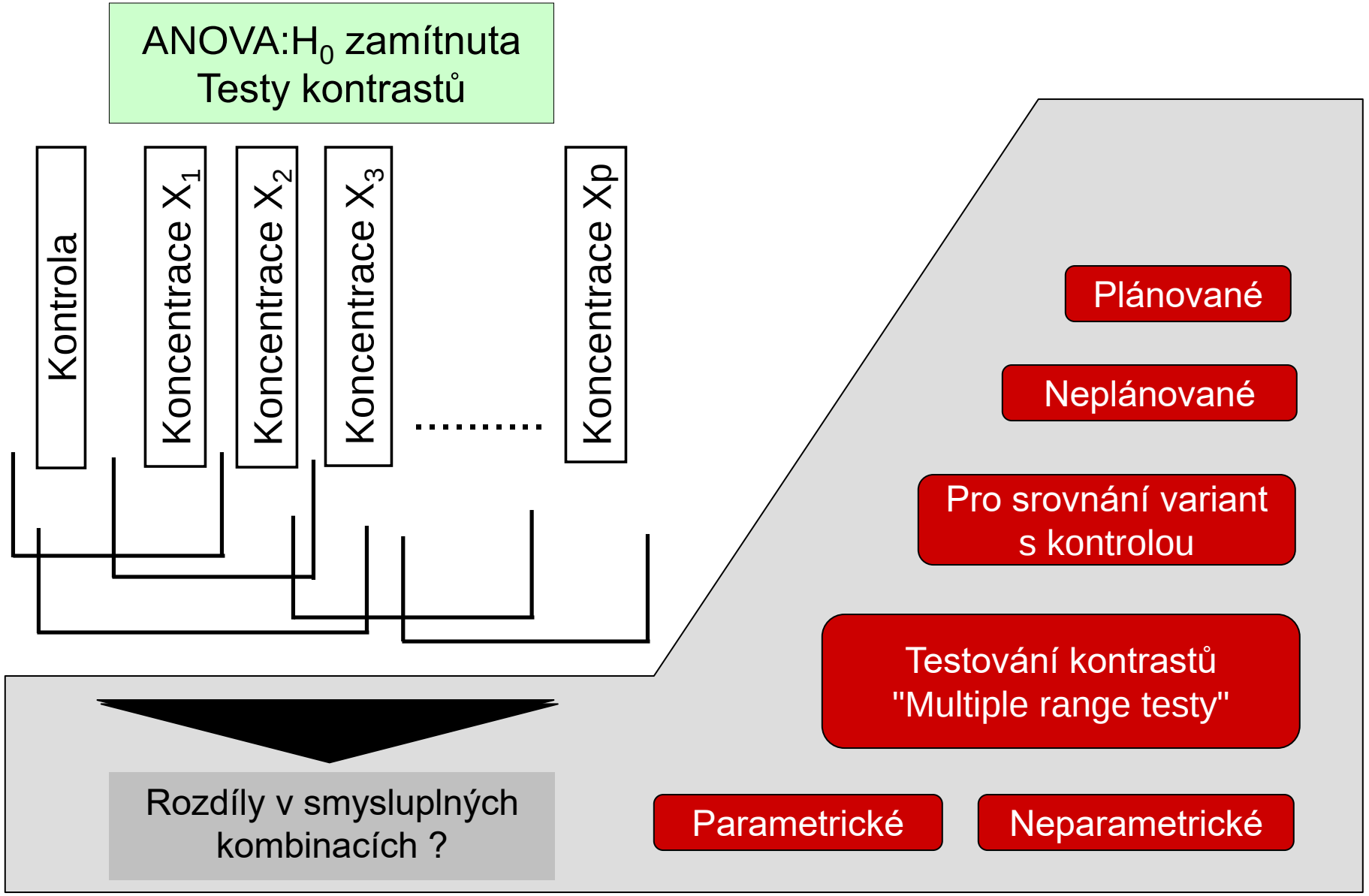
Testování dílčích hypotéz

- V řadě analýz je třeba pracovat se vzájemným testováním více skupin objektů stylem každý s každým
- Obecný postup analýzy je
  - Testování celkové významnosti – všechny skupiny navzájem (ENG: among groups)
  - Pokud je zjištěna celková významnost pokračuje testování analýzou již konkrétních kombinací dvojic skupin (ENG: between)
- Problémem je vliv mnohonásobného testování na statistickou významnost testů:
  - Každý jeden test má a=0.05 (chyba I. druhu)
  - Při mnohonásobném testování stoupá pravděpodobnost, že alespoň u jednoho testu dojde k chybnému zamítnutí nulové hypotézy (tedy k chybě I. druhu)

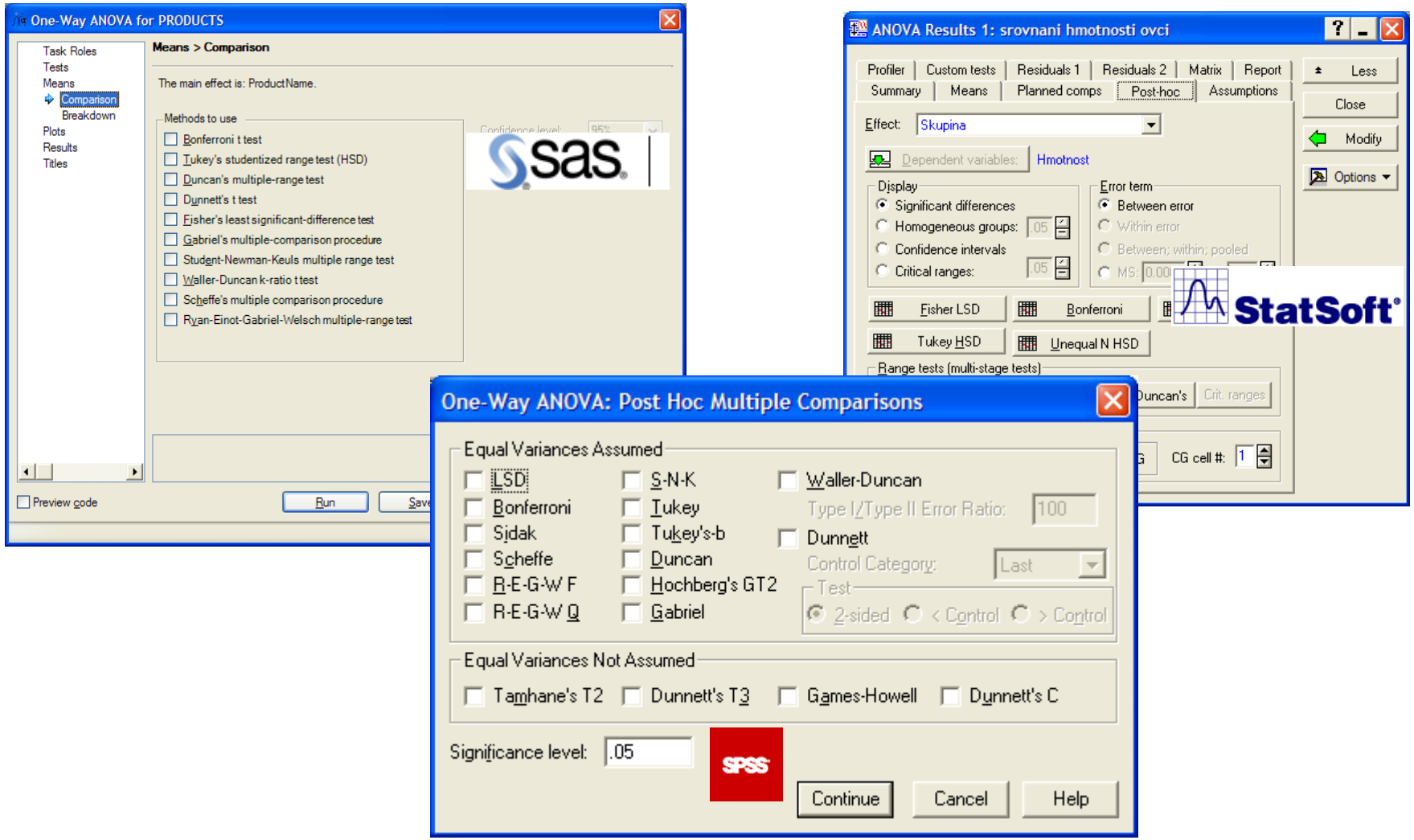


Řešením jsou různé procedury korigující hodnotu p (např. Bonferroniho korekce, FWR, FDR procedury apod.)

Analýza rozptylu - testy kontrastů



Řada různých post-hoc testů



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	p
Between	3	305.8	101.9	8.56	<0.001
Within	28	322.2	11.9		
Total	31	638			

II. Multiple Range Test (NKS –test)

Level	Average	Homogeneous groups		
0	34.8	x		
4	41.4		x	
12	41.8		x	
8	52.6			x