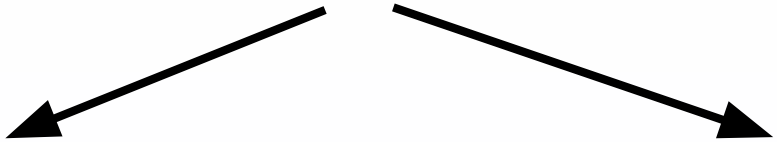
Přednáška 8 – Testy hypotéz více výběrů

- parametrické testy s předpokladem normality
- neparametrické testy bez předpokladu normality



Jednofaktorové testy

Příklad:

- plat v závislosti na vzdělání (bez maturity, maturita, VŠ)
- jeden faktor – vzdělání

Dvoufaktorové testy

Příklad:

- plat v závislosti na vzdělání a zaměstnancích
- dva faktory – vzdělání a zaměstnanci

- Faktor – podmínka, na základě které zkoumáme shodu výběrů
- Tabulka testů hypotéz:
  http://staff.utia.cas.cz/uglickich/pdfka/volbaTH.pdf

Jednofaktorový test ANOVA (angl. analysis of variance)

Použijeme za předpokladu:

- normalita všech výběrů
- stejné rozptyly výběrů

- Data nesplňují jeden z předpokladů – zvolíme neparametrickou alternativu
- Výběry nemusí být párové

Příklad: průměrný denní počet kroků mezi různými věkovými kategoriemi

- věk – faktor

mládež	dospělí	senioři
7652	6114	3468
7483	11026	4579
9806	8941	4668
5854	10412	4030
6448	8693	2283
5835	11212	4914
4300	9895	3263
6624	6626	6242
7815	7714	
	9163	
	9344	
	9166	

Postup:

1. Test normality všech výběrů
2. Pre-analýza – Bartlettův test

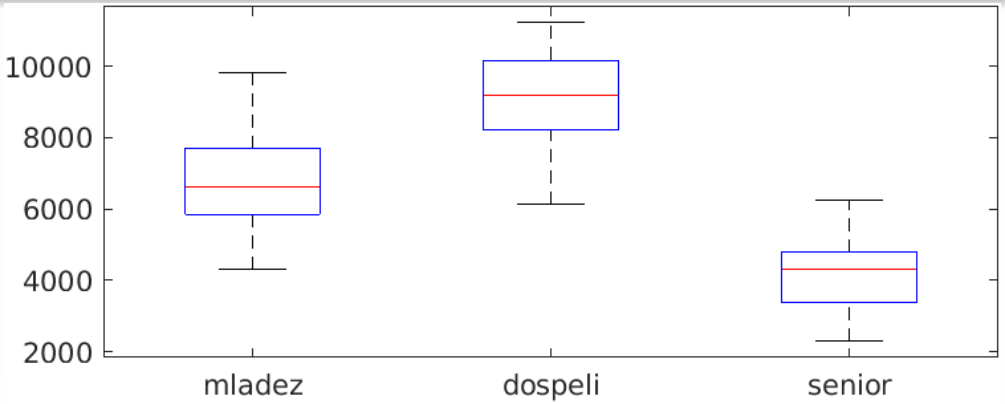
$H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k{}^2$ – rozptyly jsou stejné

H_A : alespoň jeden rozptyl se liší

$$T = \frac{(N-k)\ln(S_p^2) - \sum_{i=1}^{k}(n_i-1)\ln(s_i^2)}{1+\frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\frac{1}{n_i-1}-\frac{1}{N-k}\right)} \sim \chi^2$$

$$N = \sum_{i=1}^{k} n_i, \quad S_p^2 = \frac{1}{N-k}\sum_{i=1}^{k}(n_i-1)s_i^2$$

- Pouze pravostranný
- p-hodnota $= 0.7259$



Postup:

3. Po ověření předpokladů – Anova

- Pouze pravostranný

Nulová hypotéza: střední hodnoty jsou stejné

$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ – mládež, dospělí a senioři ujdou denně v průměru stejný počet kroků, tj., věk nemá vliv na počet kroků

Alternativní hypotéza:

H_A : alespoň jedna střední hodnota se liší, tj., věk má vliv

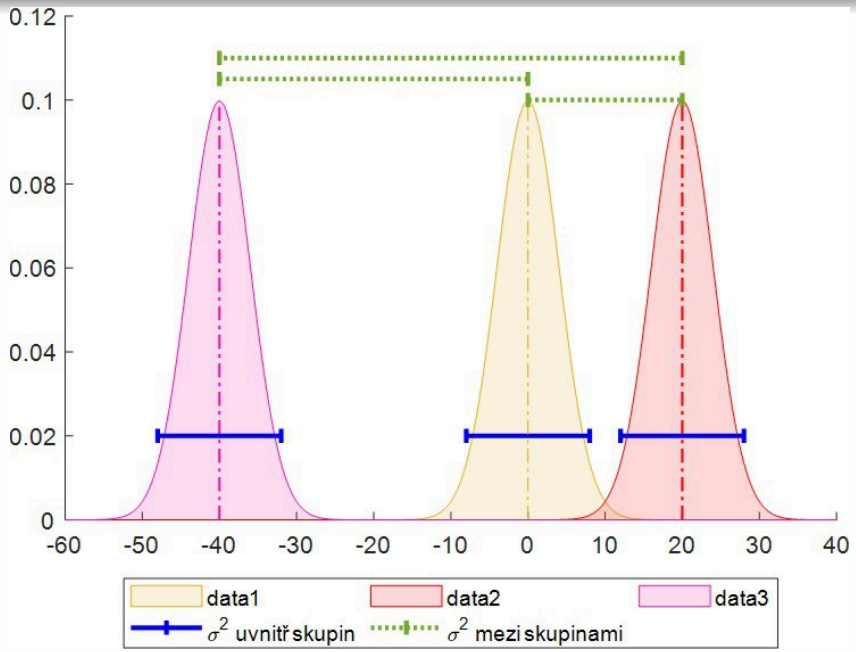
- V_1, V_2, V_3 – věkové kategorie
- $\bar{V} = \frac{1}{k}\sum_{i=1}^{k} \bar{V}_i$ – průměr průměrů z celé tabulky

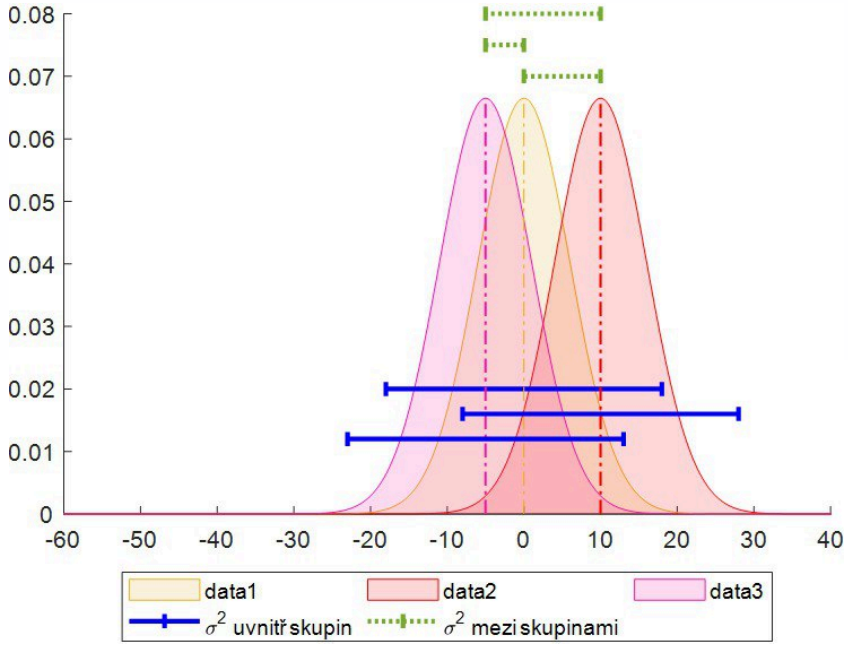
$$F = \frac{s_M^2}{s_T^2} = \frac{\text{vysvětlený rozptyl}}{\text{nevysvětlený rozptyl}} \sim \text{Fisherovo rozdělení}$$

- $s_M^2 = \sum_{i=1}^{k} n_i(\bar{V}_i - \bar{V})^2$ – rozptyl mezi sloupci – vysvětlený
- $s_T^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(V_{j,i} - \bar{V}_i)^2$ – rozptyl uvnitř sloupců (z celé tabulky) – nevysvětlený

p-hodnota=0.0000007

Příklady F = vysvětlený/nevysvětlený rozptyl





Příklad 1 – zamítneme H_0

$$\frac{\textbf{vysvětlený}}{\text{nevysvětlený}} \rightarrow \text{vyšší F} \rightarrow \text{nižší } p_h$$

Příklad 2 – nezamítneme H_0

$$\frac{\textbf{vysvětlený}}{\text{nevysvětlený}} \rightarrow \text{nižší F} \rightarrow \text{vyšší } p_h$$

4. Post-analýza – v případě zamítnutí nulové hypotézy

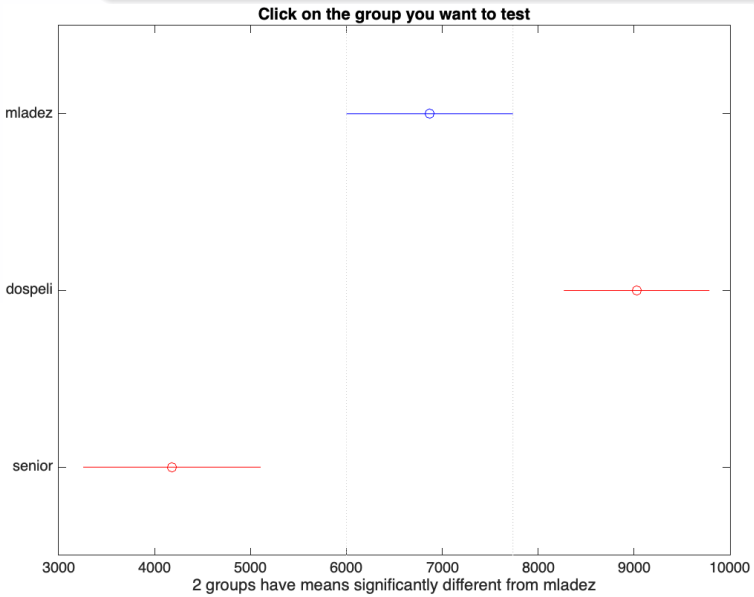
Post-analýza – porovnává střední hodnoty výběrů ve dvojicích podobně jako t-test

Nulová hypotéza:

$H_0 : \mu_1 = \mu_2$ nebo $\mu_1 = \mu_3$ nebo $\ldots$ $\mu_1 = \mu_k$ nebo $\mu_2 = \mu_3$ nebo $\ldots$

Poznámka:

- různé testy post-analýzy
- pro parametrické (Scheffého), pro neparametrické (Bonferroniho)
- na cvičení – Tukey-Kramer test (vhodný pro oba případy)




```
post_anova = 3×6
10^3  ×
    0.0010    0.0020   -3.7833   -2.1569   -0.5306    0.0000
    0.0010    0.0030    0.8955    2.6877    4.4798    0.0000
    0.0020    0.0030    3.1612    4.8446    6.5280    0.0000
```

- poslední sloupec – p-hodnoty
- lišily se střední hodnoty všech dvojic

Kruskal-Wallisův test – jednofaktorový neparametrický test

- alespoň jeden výběr nemá normalitu / rozptyly výběrů nejsou stejné
- výběry nemusí být párové

Příklad: výsledky jednotných přijímacích zkoušek na víceletá gymnázia z okresů A, B, C a D

A	B	C	D
79	89	73	91
77	79	91	64
98	60	76	72
67	94	65	58
98	46	62	60
98	69	89	86
72	82	63	61
100	91	94	62
74		84	
96			

- okres – faktor
- výběr A nemá předpoklad normality – pouze KW

Nulová hypotéza: rozdělení (mediány) jsou stejné

$H_0 : \tilde{X}_{0,5(1)} = \tilde{X}_{0,5(2)} = \ldots = \tilde{X}_{0,5(k)}$

výsledky jsou stejné, tj., na okresu nezáleží

Alternativní hypotéza:

H_A : výsledky alespoň jednoho okresu se liší,

tj., okres má vliv

$$T = (N-1)\frac{\sum_{i=1}^{k} n_i(\bar{r}_i - \bar{r})^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(r_{ij} - \bar{r})^2}$$

- Pouze pravostranný
- p-hodnota=0.061

- Post-analýza v případě zamítnutí – Bonferroniho test, Tukey-Kramer

Dvoufaktorový test ANOVA – dva faktory

Předpoklady:

- normální rozdělení dat v řádcích a sloupcích tabulky s daty
- stejné rozptyly dat v řádcích a sloupcích
- stejný počet dat

Příklad: Porovnáváme předpokládanou výši hypoteční splátky u bank E, F, G na byt 2+kk od 12 žadatelů. Zajímá nás, zda na výši splátky má vliv banka, kde požádáme o hypotéku a také žadatel. Předpokládáme normální rozdělení hodnot jak ve sloupcích, tak i v řádcích.

- 2 faktory – banka, žadatel

žadatel \ banka	E	F	G
1	13063	11046	11018
2	13038	11046	11006
3	13049	11028	10964
4	12967	11038	11049
5	12980	10983	11002
6	12945	10931	10999
7	12943	11046	11037
8	13058	10962	11051
9	12944	10991	11000
10	13001	10994	11081
11	12997	11018	10983
12	13083	10895	11036

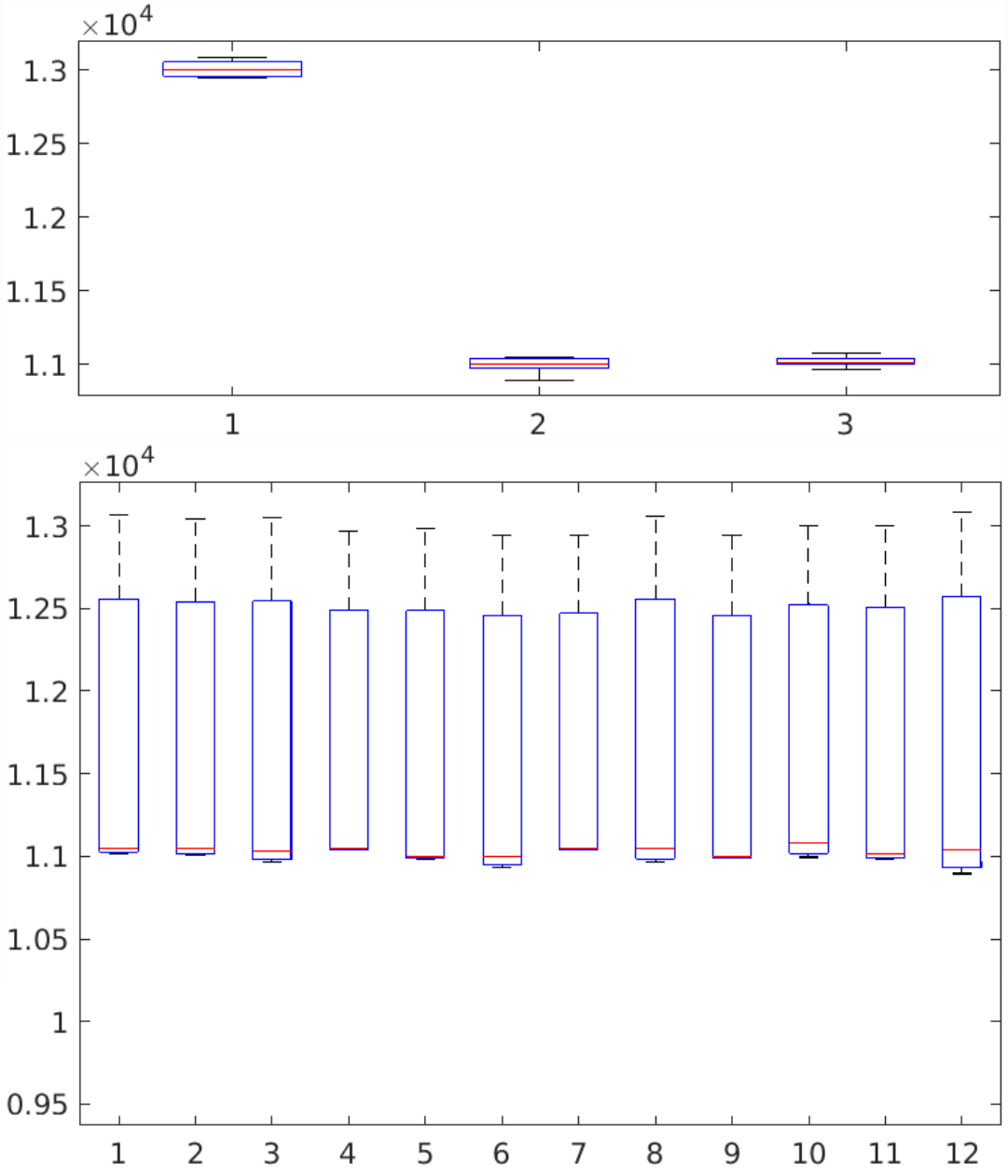
Ověření předpokladu stejných rozptylů – Bartlettův test:

- pro banky ve sloupcích:

p-hodnota $= 0.3292$

- pro žadatele v řádcích:

p-hodnota $= 1$



- Dvoufaktorová Anova – jednofaktorová Anova zvlášť ve sloupcích (pro banky) a řádcích (žadatele)
- dvě nulové hypotézy – ve sloupcích a v řádcích

Nulová hypotéza: střední hodnoty ve sloupcích jsou stejné

$H_0 : \mu_E = \mu_F = \mu_G$ – banka nemá vliv na výši splátky

Alternativní hypotéza: alespoň jedna střední hodnota se liší

H_A : alespoň jedna banka má vliv

Nulová hypotéza: střední hodnoty v řádcích jsou stejné

$H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_{12}$ – žadatel nemá vliv na výši splátky

Alternativní hypotéza: alespoň jedna střední hodnota se liší

H_A : alespoň jeden žadatel má vliv

$$F_{\text{s-b}} = \frac{\text{vysvětlený rozptyl}}{\text{nevysvětlený rozptyl}}$$

- p-hodnota$_{\text{s-b}} = 9.898e-32$
- Zamítáme H_0 ve sloupcích

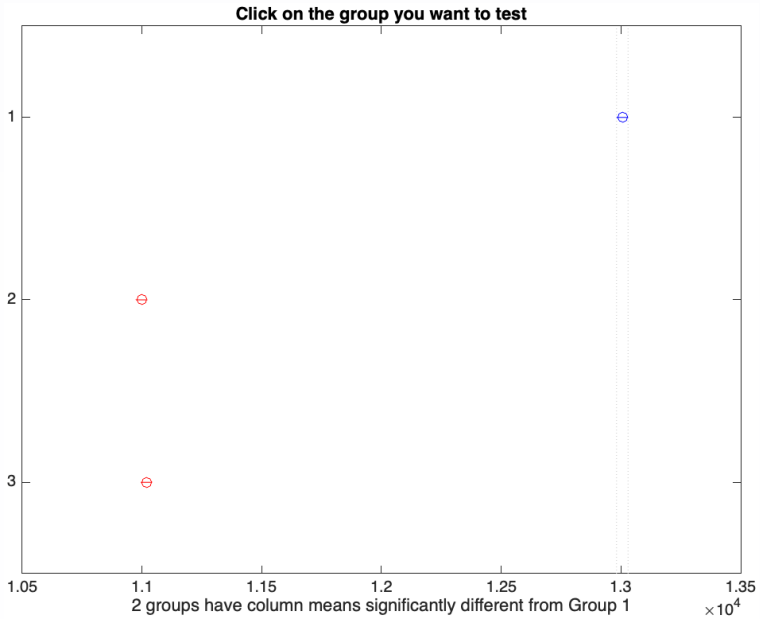
$$F_{\text{r-z}} = \frac{\text{vysvětlený rozptyl}}{\text{nevysvětlený rozptyl}}$$

- p-hodnota$_{\text{r-z}} = 0.6725$
- Nezamítáme H_0 v řádcích

Závěr: Záleží na bance

Všimněte si středních hodnot na krabicových diagramech

Testy post-analýzy – pro sloupce



post_anova2 = *3×6*

10^3 ×

0.0010	0.0020	1.9593	2.0075	2.0557	0
0.0010	0.0030	1.9387	1.9868	2.0350	0
0.0020	0.0030	−0.0688	−0.0207	0.0275	0.0005

- lišila banka E

Dvoufaktorový test Anova s opakováním

- banka, žadatel, každý žadatel si navíc přijde několikrát
- testuje se shoda středních hodnot
  - ↗ ve sloupcích (v závislosti na bance)
  - → v řádcích (v závislosti na žadateli)
  - ↘ nezávislost řádků a sloupců
- tři nulové hypotézy:
  - první dvě – totožné s dvoufaktorovým testem Anova
  - třetí H_0: sloupce a řádky jsou nezávislé
- tři p-hodnoty

Friedmanův test – neparametrický dvoufaktorový test

- párové výběry bez předpokladu normality
- dva faktory

Příklad:

12 náhodně vybraných zákazníků hodnotí kvalitu 5 e-shopů od 0 (nejhorší) do 10 (nejlepší). Ptáme se, zda jsou hodnocení e-shopů stejná a zda každý zákazník hodnotí e-shopy stejně. Nepředpokládáme normalitu dat.

- dvě H_0:
  – v řádcích, ve sloupcích
- dvě p-hodnoty:
  – v řádcích, ve sloupcích

- párové výběry
- 2 faktory – zákazník, e-shop

zákazník \ e-shop	A	B	C	D	E
1	7	3	5	7	2
2	1	2	8	8	2
...	...	...	...	...	...
12	...	...	...	...	...

H_0 a H_A

Nulová hypotéza: rozdělení (mediány) ve sloupcích jsou stejná

$H_0 : \tilde{X}_{0,5(A)} = \tilde{X}_{0,5(B)} = \ldots = \tilde{X}_{0,5(E)}$

e-shop nemá na hodnocení vliv

Alternativní hypotéza:

H_A : alespoň jeden e-shop je hodnocen jinak

Nulová hypotéza: rozdělení (mediány) v řádcích jsou stejná

$H_0 : \tilde{X}_{0,5(1)} = \tilde{X}_{0,5(2)} = \ldots = \tilde{X}_{0,5(12)}$

zákazník nemá na hodnocení vliv

Alternativní hypotéza:

H_A : alespoň jeden zákazník hodnotí jinak

Výpočet statistiky

- s využitím pořadí v obou směrech:

zákazník \ e-shop	A	B	C	D	E
1	7(4)	3(2)	5(3)	7(4)	2(1)
2	1(1)	2(2)	8(3)	8(3)	2(2)
...	...	...	...	...	...
	$R_1 = 5$	$R_2 = 4$	$R_3 = 6$	$R_4 = 7$	$R_5 = 3$

- $T = 3n(k+1)\sum_{i=1}^{k} R_i \sim \chi 2$-rozdělení
- p-hodnota ve sloupcích (e-shop), p-hodnota v řádcích (zákazník)
- Testy post-analýzy – podobně jako pro Kruskal-Wallisův test