Vícerozměrné metody - cvičení

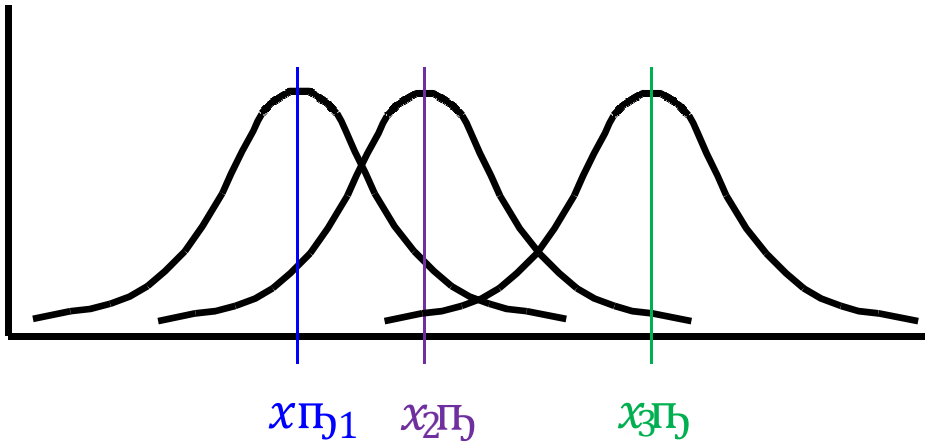
Mgr. Jan Zdražil

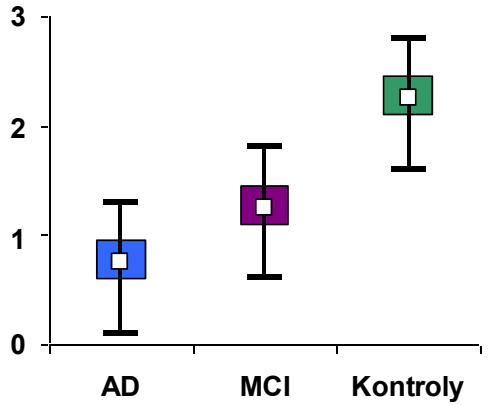
Cvičení 3

Analýza rozptylu pro vícerozměrná data

Analýza rozptylu (ANOVA) jednoduchého třídění

- **Srovnáváme tři a více skupin dat, které jsou na sobě nezávislé** (mezi objekty neexistuje vazba).
- Příklady: srovnání objemu hipokampu u pacientů s AD, pacientů s MCI a kontrol; srovnání kognitivního výkonu podle čtyř kategorií věku.

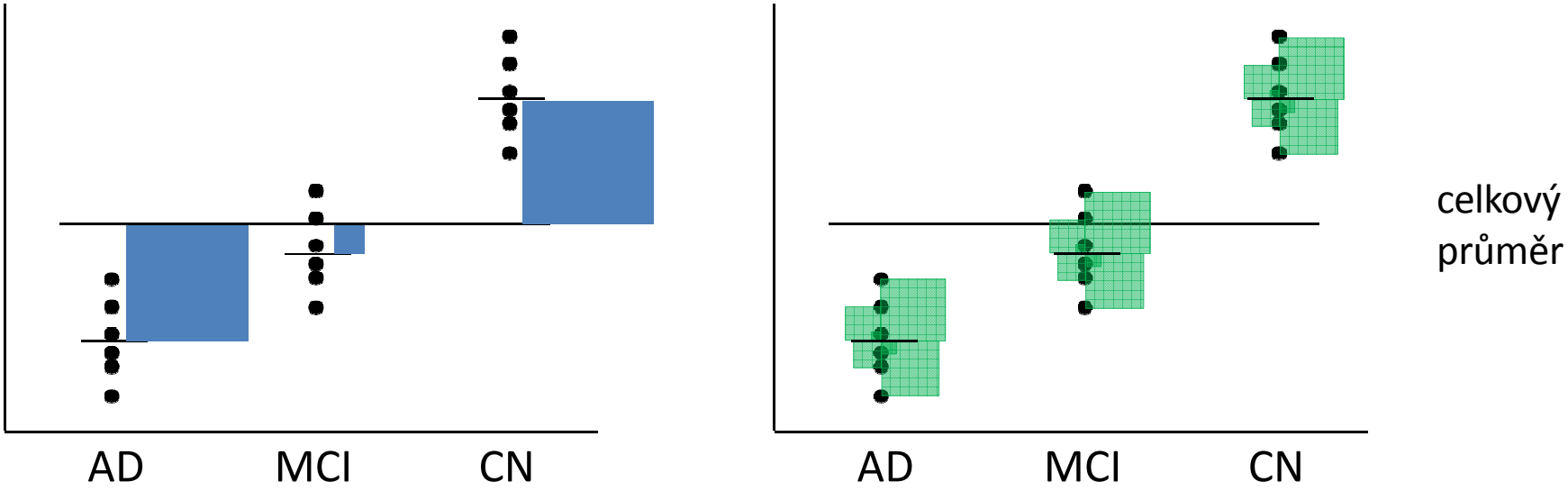




- Předpoklady: **normalita dat ve VŠECH skupinách**, **shodnost (homogenita) rozptylů VŠECH srovnávaných skupin**, nezávislost jednotlivých pozorování.
- Testová statistika: $F = \frac{S_A / df_A}{S_e / df_e}$

Analýza rozptylu (ANOVA) – princip

- Srovnání variability (rozptylu) mezi výběry s variabilitou uvnitř výběrů.



- Tabulka analýzy rozptylu jednoduchého třídění (One-Way ANOVA):

Variabilita	Součet čtverců	Počet stupňů volnosti	Průměrný čtverec	F statistika	p-hodnota
Mezi skupinami	S_A	$df_A = a - 1$	$MS_A = S_A/df_A$	$F = \frac{S_A / df_A}{S_e / df_e}$	p
Uvnitř skupin (reziduální var.)	S_e	$df_e = n - a$	$MS_e = S_e/df_e$		
Celkem	S_T	$df_T = n - 1$			

Analýza rozptylu jako lineární model

- Analýza rozptylu pro jednu vysvětlující proměnnou (jednoduché třídění) lze zapsat jako lineární model:

$$Y_{ij} = \mu_i + e_{ij} = \mu + \alpha_i + e_{ij}$$

Reziduum

Populační průměr

i-tý efekt faktoru A

- Nulovou hypotézu pak lze vyjádřit jako: $H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_k$

- **Rozšířením tohoto zápisu můžeme definovat další modely ANOVA**: více faktorů, hodnocení interakcí, opakovaná měření na jednom subjektu.

Analýza rozptylu pro vícerozměrná data

- podle počtu vysvětlovaných proměnných:
  - 1 vysvětlovaná proměnná – jednorozměrná analýza rozptylu (ANOVA)
  - 2 a více vysvětlovaných proměnných – vícerozměná analýza rozptylu (MANOVA)
- podle počtu faktorů:
  - 1 faktor – ANOVA jednoduchého třídění (jednofaktorová ANOVA)
  - 2 faktory – ANOVA dvojného třídění (dvoufaktorová ANOVA)
  - ...
- podle toho, zda se faktory ovlivňují či nikoliv:
  - faktory se mohou ovlivňovat – model s interakcí
  - faktory se neovlivňují – model bez interakce

Analýza rozptylu pro vícerozměrná data - příklady

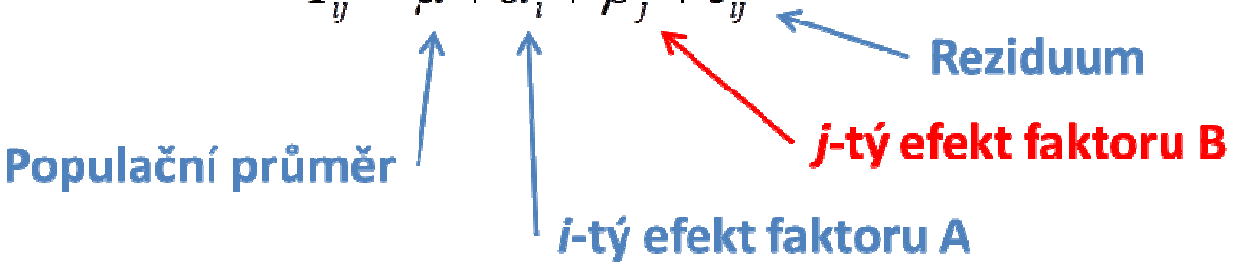
Počet proměnných: jednorozměrná x vícerozměrná analýza rozptylu
Počet faktorů: jednoduché x dvojné x trojné, ... třídění
Faktory se ovlivňují či neovlivňují: s interakcí x bez interakce

- zkoumáme dlouhodobý vliv třech typů léků na hodnoty systolického tlaku u stovky osob
- zkoumáme dlouhodobý vliv třech typů léků na hodnoty systolického tlaku u stovky osob, přičemž chceme zkoumat i vliv pohlaví, předpokládáme však, že ženy i muži reagují na jednotlivé léky obdobně (tzn. např. ženy s léky A a C budou mít nižší tlak než ženy s lékem B a muži s léky A a C budou mít také nižší tlak než muži s lékem B apod.)
- zkoumáme dlouhodobý vliv třech typů léků na hodnoty systolického tlaku u stovky osob, přičemž chceme zkoumat i vliv pohlaví, a předpokládáme, že ženy a muži budou reagovat na léky různě (tzn. např. ženy s léky A a C budou mít nižší tlak než ženy s lékem B, zatímco muži s léky A a B budou mít vyšší tlak než muži s lékem C apod.)
- zkoumáme dlouhodobý vliv třech typů léků na hodnoty systolického a diastolického tlaku u stovky osob
- zkoumáme dlouhodobý vliv třech typů léků a vliv pohlaví na hodnoty systolického a diastolického tlaku u stovky osob

Analýza rozptylu dvojného třídění (bez interakce)

- Uvažujeme dvě vysvětlující proměnné zároveň.
- Zápis modelu:

$$Y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$$



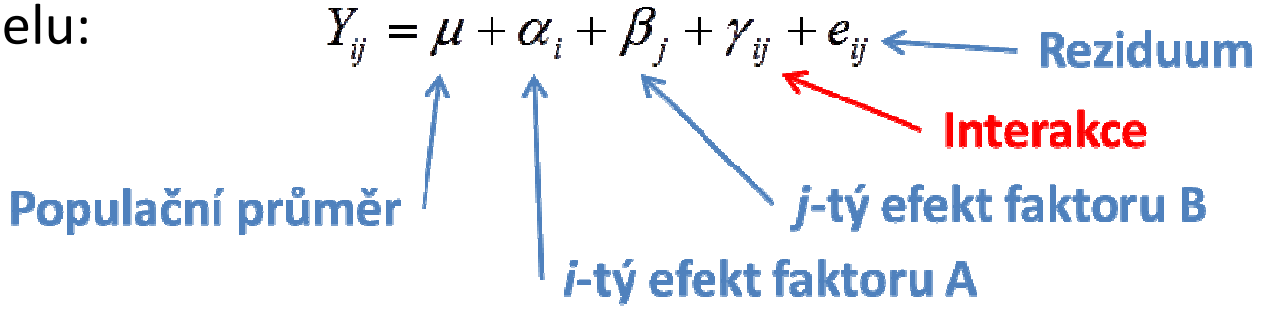
- Nulové hypotézy pak máme dvě: $H_{01}: \alpha_1 = \alpha_2 = \ldots = \alpha_k$, $H_{02}: \beta_1 = \beta_2 = \ldots = \beta_r$

Variabilita	Součet čtverců	Počet stupňů volnosti	Průměrný čtverec	*F* statistika	*p*-hodnota
Faktor A	S_A	$df_A = a - 1$	$MS_A = S_A / df_A$	F_A	p
Faktor B	S_B	$df_A = b - 1$	$MS_B = S_B / df_B$	F_B	p
Rezidua	S_e	$df_e = n - a - b + 1$	$MS_e = S_e / df_e$		
Celkem	S_T	$df_T = n - 1$			

Analýza rozptylu dvojného třídění s interakcí

- Uvažujeme dvě vysvětlující proměnné a zároveň i jejich společné působení.
- Zápis modelu:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ij}$$

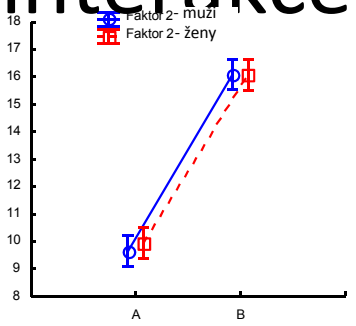


- Nulové hypotézy pak máme tři:

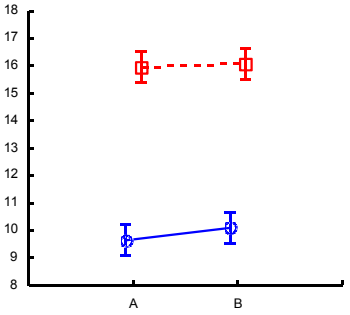
$$H_{01}: \gamma_{11} = \gamma_{12} = \ldots = \gamma_{kr} \qquad H_{02}: \alpha_1 = \alpha_2 = \ldots = \alpha_k \qquad H_{03}: \beta_1 = \beta_2 = \ldots = \beta_r$$

Variabilita	Součet čtverců	Počet stupňů volnosti	Průměrný čtverec	*F* statistika	*p*-hodnota
Faktor A	S_A	$df_A = a - 1$	$MS_A = S_A / df_A$	F_A	p
Faktor B	S_B	$df_A = b - 1$	$MS_B = S_B / df_B$	F_B	p
Interakce A×B	S_{AB}	$df_{AB} = (a-1)(b-1)$	$MS_{AB} = S_{AB} / df_{AB}$	F_{AB}	p
Rezidua	S_e	$df_e = n - ab$	$MS_e = S_e / df_e$		
Celkem	S_T	$df_T = n - 1$			

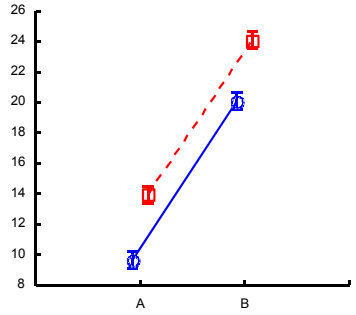
Hlavní efekty a interakce



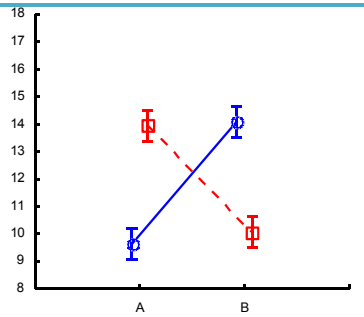
	SS	D.f.	MS	F	p
Faktor 1	**1978**	**1**	**1978**	**482.2**	**0.000**
Faktor 2	1	1	1	0.3	0.602
F1*F2	1	1	1	0.3	0.570
Error	804	196	4		



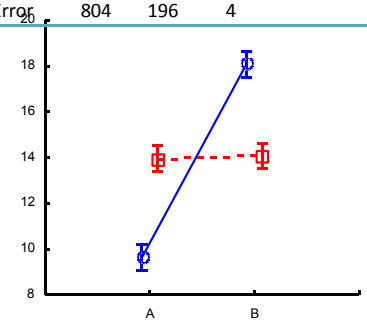
	SS	D.f.	MS	F	p
Faktor 1	4	1	4	1.0	0.314
Faktor 2	**1891**	**1**	**1891**	**461.1**	**0.000**
F1*F2	1	1	1	0.3	0.570
Error	804	196	4		



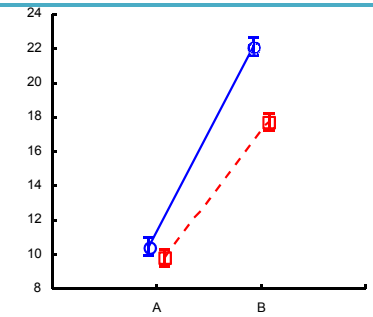
	SS	D.f.	MS	F	p
Faktor 1	5293	1	5293	1290.7	0.000
Faktor 2	861	1	861	209.9	0.000
F1*F2	1	1	1	0.3	0.570
Error	804	196	4		



	SS	D.f.	MS	F	p
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		

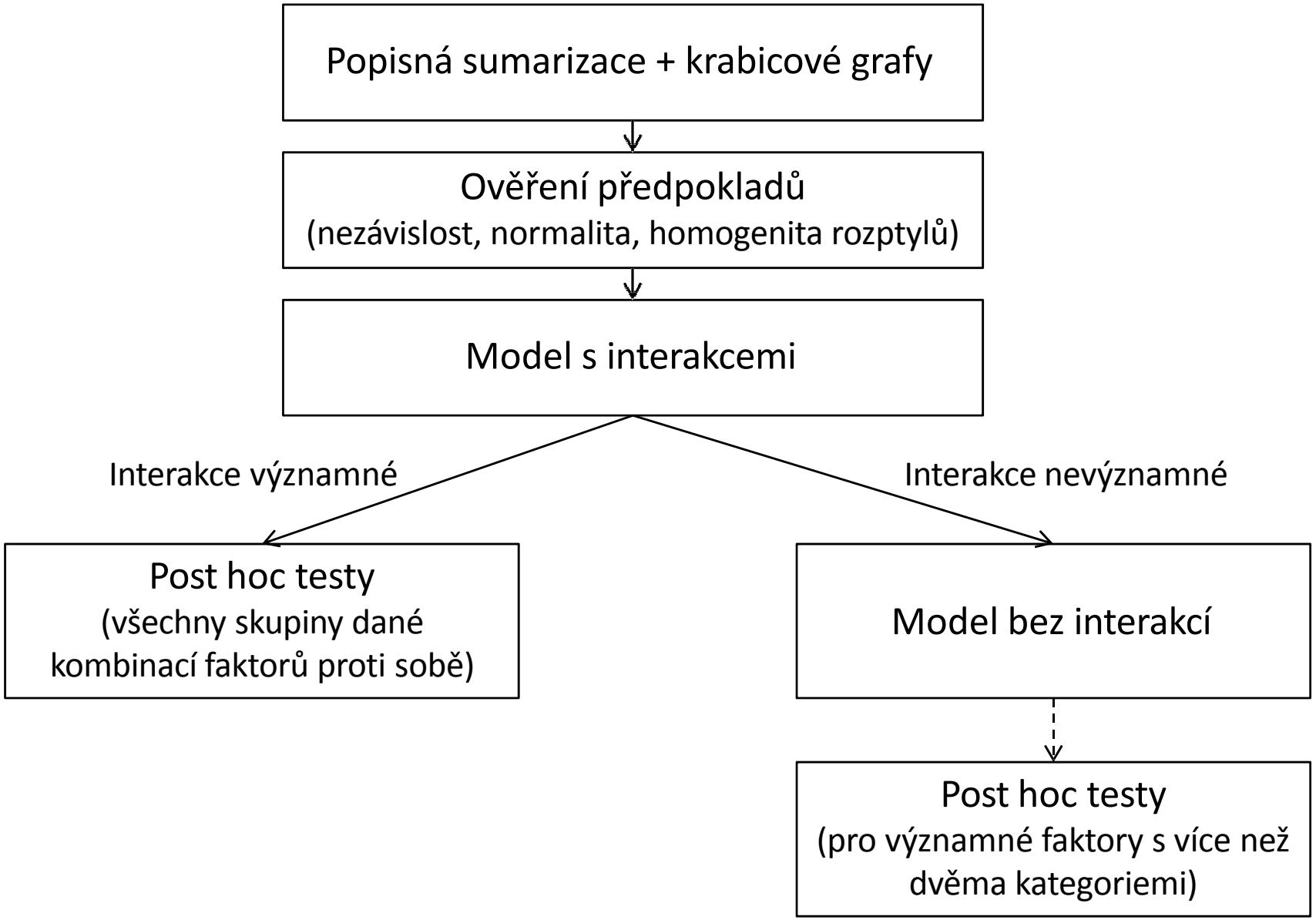


	SS	D.f.	MS	F	p
Faktor 1	**920**	**1**	**920**	**224.3**	**0.000**
Faktor 2	1	1	1	0.3	0.602
F1*F2	**867**	**1**	**867**	**211.3**	**0.000**
Error	804	196	4		



	SS	D.f.	MS	F	p
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 pro vícerozměrná data - postup



Úkol 1

Zjistěte, zda má vliv pohlaví a typ léku na počet nežádoucích účinků u pacientů s leukémií (neuvažujeme možnou interakci).

Pohlaví	Typ léku	Počet nežádoucích účinků
1	1	1
1	2	1
1	3	6
2	1	3
2	2	4
2	3	9

Variabilita	Součet čtverců	Počet stupňů volnosti	Průměrný čtverec	*F* statistika	*p*-hodnota
Faktor A	S_A	$df_A = a - 1$	$MS_A = S_A / df_A$	F_A	p
Faktor B	S_B	$df_A = b - 1$	$MS_B = S_B / df_B$	F_B	p
Rezidua	S_e	$df_e = n - a - b + 1$	$MS_e = S_e / df_e$		
Celkem	S_T	$df_T = n - 1$			

Úkol 2

Zjistěte, zda má vliv pohlaví a typ onemocnění na objem hipokampu.

Ukázka datového souboru:

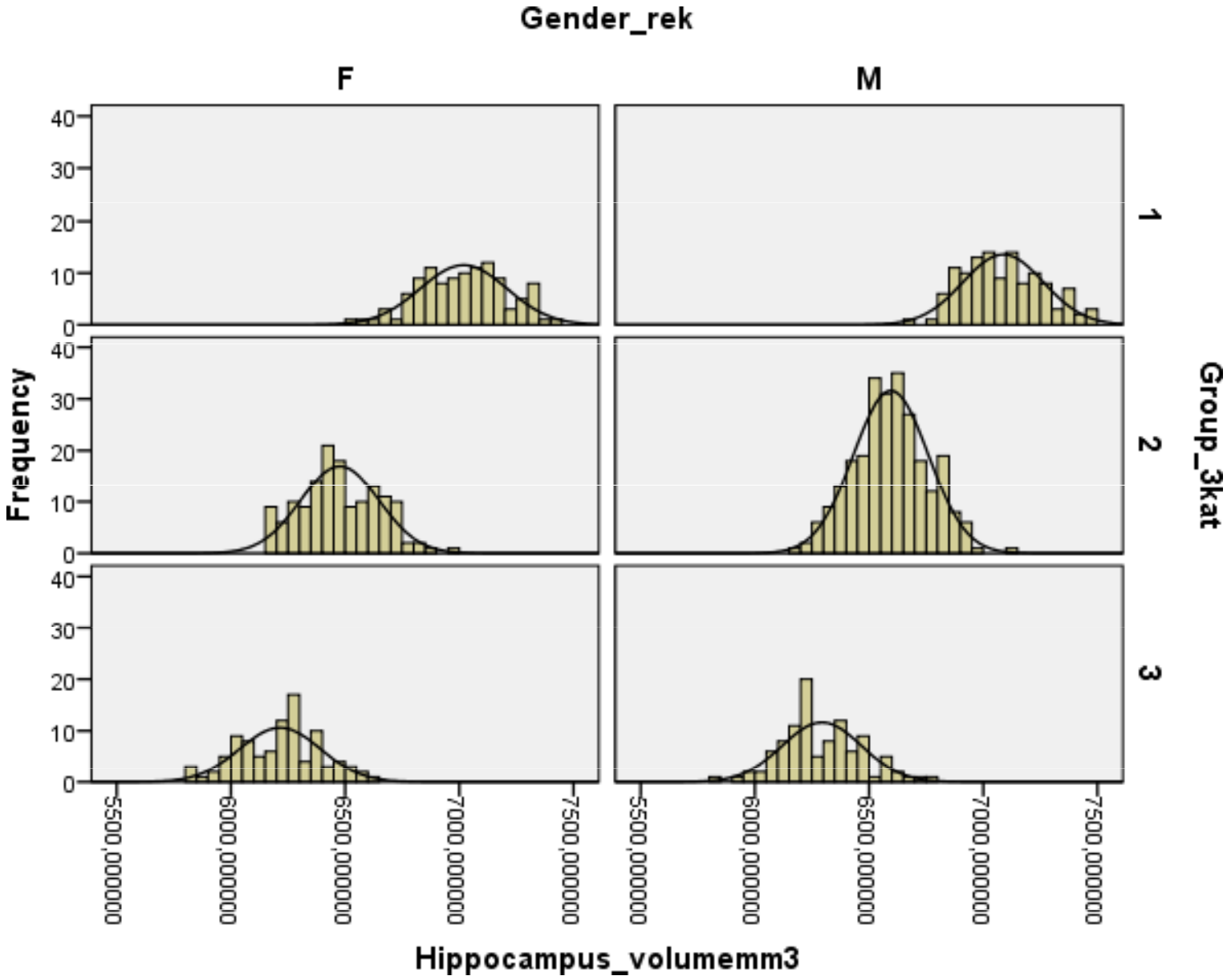
ID	Group_3kat	Gender_rek	Hippocampus_volume (mm3)
101	1	M	6996.1
102	1	F	7187.3
103	1	M	7030.2
331	2	M	6891.6
332	2	M	6332.9
334	2	F	6303.7
737	3	M	6170.8
739	3	F	5984.1
740	3	F	6052.4

Legenda k proměnné Group_3kat:
1...CN (kontroly)
2...MCI (mírná kognitivní porucha)
3...AD (Alzheimerova choroba)

Úkol 2 – popisná sumarizace dat

Skupina	Pohlaví	N	Průměr	SD	Medián	Minimum	Maximum
CN	F	110	7018.3	190.1	7036.1	6509.6	7430.1
	M	120	7087.3	176.0	7081.1	6674.4	7486.6
	Celkem	230	7054.3	185.7	7048.6	6509.6	7486.6
MCI	F	146	6476.7	171.8	6460.4	6155.1	6984.8
	M	260	6595.2	164.1	6589.5	6159.1	7125.6
	Celkem	406	6552.6	176.2	6555.0	6155.1	7125.6
AD	F	95	6215.0	178.8	6237.8	5805.2	6619.0
	M	102	6293.0	174.8	6250.8	5844.3	6756.9
	Celkem	197	6255.4	180.6	6248.0	5805.2	6756.9
Celkem	F	351	6575.6	364.8	6498.2	5805.2	7430.1
	M	482	6653.8	323.9	6610.0	5844.3	7486.6
	Celkem	833	6620.9	343.7	6580.9	5805.2	7486.6

Úkol 2 – ověření normality



Úkol 2 – homogenita rozptylů a nezávislost

Homogenita rozptylů:

Levene's Test of Equality of Error Variances[a,b]

		Levene Statistic	df1	df2	Sig.
Hippocampus_volume (mm3)	Based on Mean	.962	5	827	.440
	Based on Median	.852	5	827	.513
	Based on Median and with adjusted df	.852	5	815.047	.513
	Based on trimmed mean	.935	5	827	.457

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Dependent variable: Hippocampus_volume (mm3)

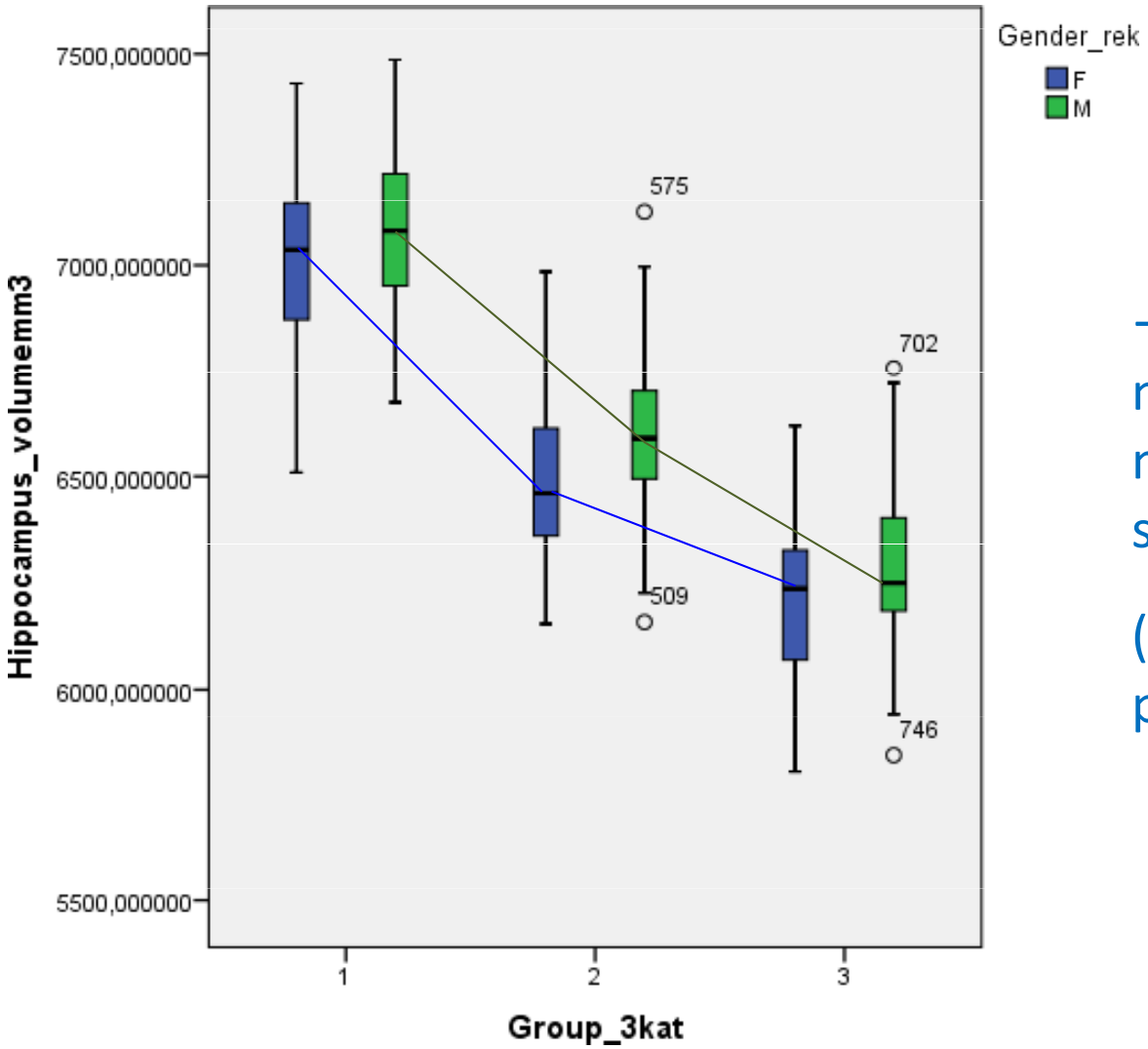
b. Design: Group_3kat + Gender_rek + Group_3kat * Gender_rek

p=0,440 > 0,05 → nezamítáme homogenitu rozptylů

Nezávislost:

Protože žádný subjekt nebyl současně ve více skupinách, nezávislost můžeme předpokládat.

Úkol 2 – krabicový graf



→ interakci sice očekávat nebudeme, přesto si ale model s interakcí raději spočítáme

(nejdřív ale musíme ověřit předpoklady)

Úkol 2 – model s interakcí

Tests of Between-Subjects Effects

Dependent Variable: Hippocampus_volumemm3

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Model	3,659E+10[a]	6	6098069036	201956,010	,000
Group_3kat	71984656,14	2	35992328,07	1191,995	,000
Gender_rek	1455184,169	1	1455184,169	48,193	,000
Group_3kat * Gender_rek	104654,379	2	52327,189	1,733	,177
Error	24971294,93	827	30195,036		
Total	36613385510	833			

a. R Squared = .999 (Adjusted R Squared = .999)

→ není statisticky významná interakce, proto spočítáme model bez interakce

Úkol 2 – model bez interakce

Tests of Between-Subjects Effects

Dependent Variable: Hippocampus_volumemm3

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Model	3,659E+10[a]	4	9147077390	302398,408	,000
Group_3kat	71962303,15	2	35981151,58	1189,521	,000
Gender_rek	1781192,205	1	1781192,205	58,885	,000
Error	25075949,31	829	30248,431		
Total	36613385510	833			

a. R Squared – .999 (Adjusted R Squared – .999)

→ statisticky významný vliv pohlaví i typu onemocnění na objem hipokampu

→ protože typ onemocnění má více než 2 kategorie, musíme provést post-hoc test, abychom zjistili, mezi kterými kategoriemi je statisticky významný rozdíl

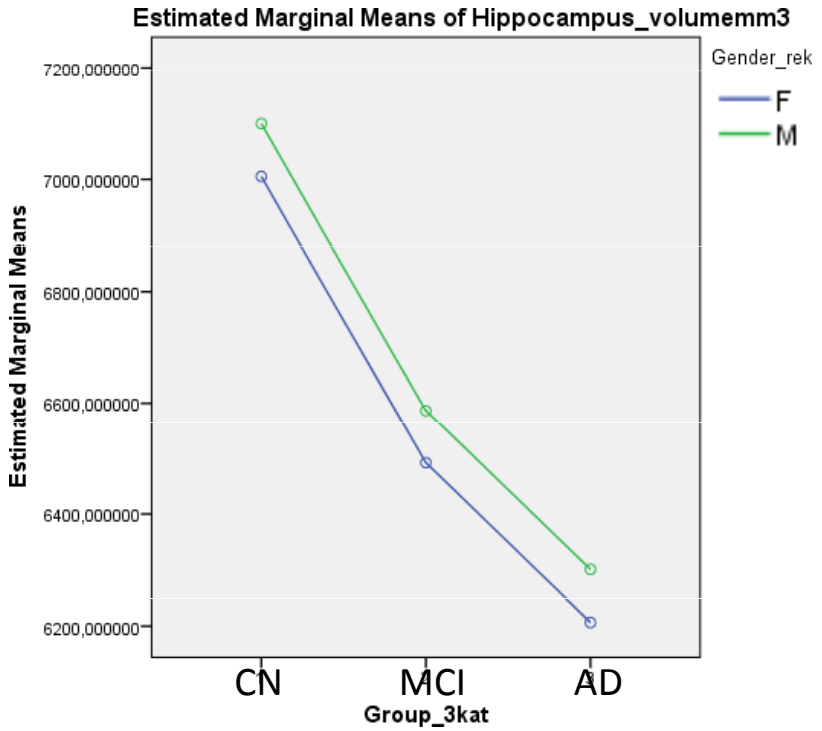
Úkol 2 – post-hoc testy a interpretace

Multiple Comparisons

Dependent Variable: Hippocampus_volume (mm3)

	(I) Group_3kat	(J) Group_3kat	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Scheffe	CN	MCI	501.721*	14.3533	.000	466.524	536.918
		AD	798.953*	16.8837	.000	757.551	840.355
	MCI	CN	-501.721*	14.3533	.000	-536.918	-466.524
		AD	297.232*	15.1013	.000	260.201	334.263
	AD	CN	-798.953*	16.8837	.000	-840.355	-757.551
		MCI	-297.232*	15.1013	.000	-334.263	-260.201

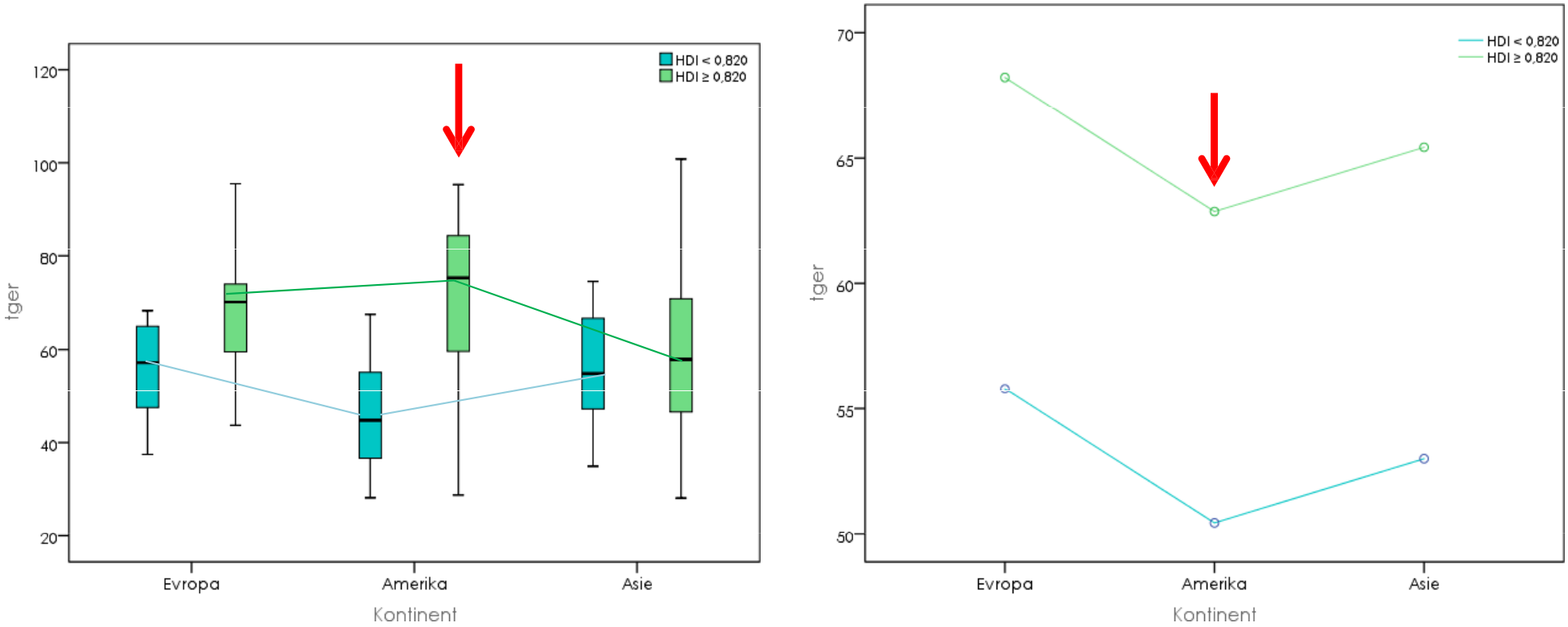
			Subset		
	Group_3kat	N	1	2	3
Tukey B[a,b,c]	AD	197	6255.382		
	MCI	406		6552.614	
	CN	230			7054.335



- statisticky významný vliv pohlaví i typu onemocnění na objem hipokampu, přičemž mezi pohlavím a typem onemocnění nenastává interakce
- u mužů statisticky významně vyšší objem hipokampu než u žen
- statisticky významný rozdíl v objemu hipokampu u všech 3 skupin subjektů podle typu onemocnění, přičemž u pacientů s AD je objem nejmenší a u CN největší

Upozornění I

Pozor, pokud mediány ukazují úplně jiný „trend“ než průměry!



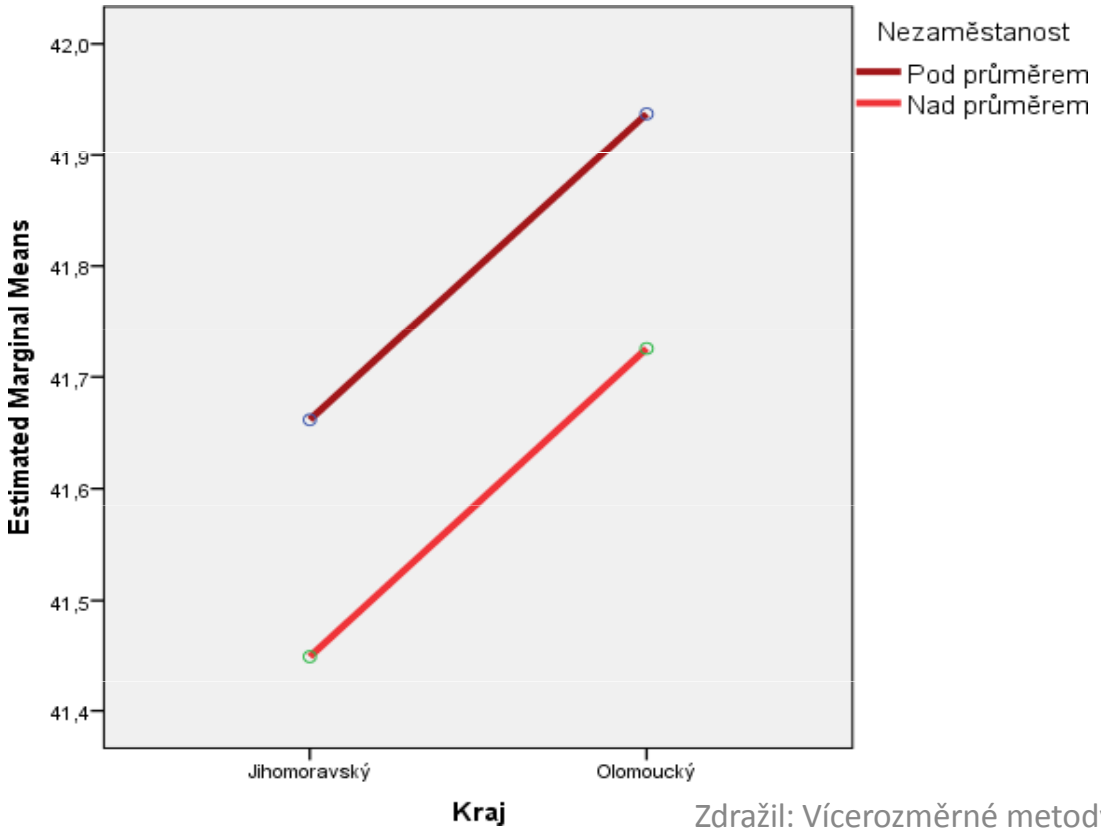
- znamená to, že tam zřejmě není splněn předpoklad normality
- pokud rozdíl není statisticky významný, není zpravidla potřeba to řešit
- pokud by ten rozdíl vyšel statisticky významně, je to problém!
- poznámka: je dobré mít měřítko na ose y stejné u obou grafů

Upozornění II

Pozor na interpretaci!

Na první pohled z grafu vypadá, že tam je vliv kraje i nezaměstnanosti, že to nevychází statisticky významně může být:

- malým počtem subjektů ve skupině
- ale i velikostí efektu! (tady efekty malé, průměry ve všech čtyřech skupinách se podle posledního grafu pohybují jen od cca 41,4 do 42!)



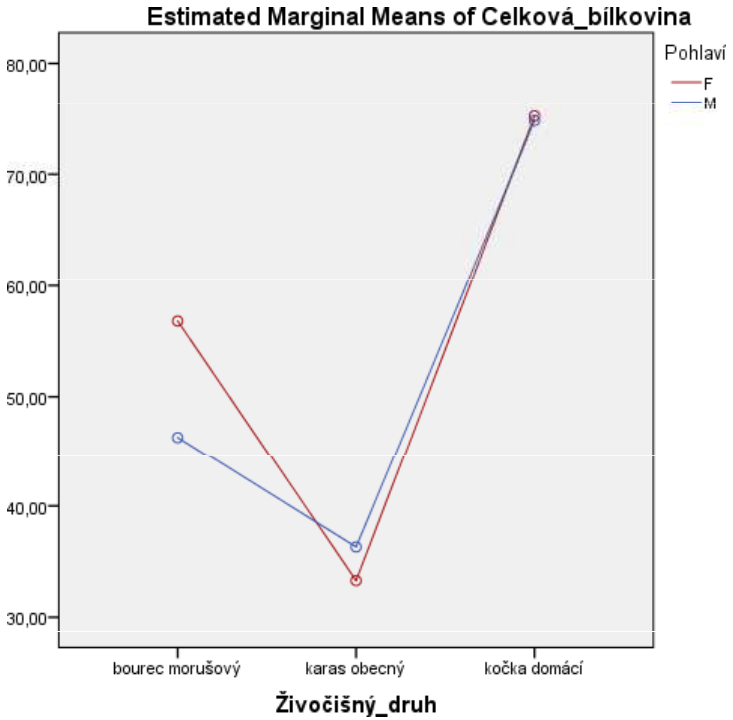
Doplnění – model s interakcemi

Tests of Between-Subjects Effects

Dependent Variable: Celková_bílkovina

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	152178,501[a]	5	30435,700	4942,124	,000
Intercept	1562235,885	1	1562235,885	253674,570	,000
Živočišný_druh	146815,301	2	73407,651	11919,874	,000
Pohlaví	931,626	1	931,626	151,277	,000
Živočišný_druh * Pohlaví	4431,573	2	2215,787	359,798	,000
Error	3288,599	534	6,158		
Total	1717702,985	540			
Corrected Total	155467,100	539			

a. R Squared = ,979 (Adjusted R Squared = ,979)



Unequal N HSD; variable Celková_bílkovina
Approximate Probabilities for Post Hoc Tests
Error: Between MS = 6.1584, df = 534.00

Cell No.	Živočišný_druh	Pohlaví	{1} 56.801	{2} 46.318	{3} 75.211	{4} 74.794	{5} 33.289	{6} 36.308
1	bourec morušový	F		0.000020	0.000020	0.000020	0.000020	0.000020
2	bourec morušový	M	0.000020		0.000020	0.000020	0.000020	0.000020
3	kočka domácí	F	0.000020	0.000020		0.870236	0.000020	0.000020
4	kočka domácí	M	0.000020	0.000020	0.870236		0.000020	0.000020
5	karas obecný	F	0.000020	0.000020	0.000020	0.000020		0.000020
6	karas obecný	M	0.000020	0.000020	0.000020	0.000020	0.000020	

Závěr:

- Nejvyšší koncentrace celkové bílkoviny zjištěny u kočky domácí a nejnižší u karase obecného.
- Vliv pohlaví různý u různých druhů. Největší vliv u bource morušového, přičemž F statisticky významně vyšší koncentrace než u M. Žádný vliv u kočky domácí. U karase obecného významně vyšší koncentrace u M než F.