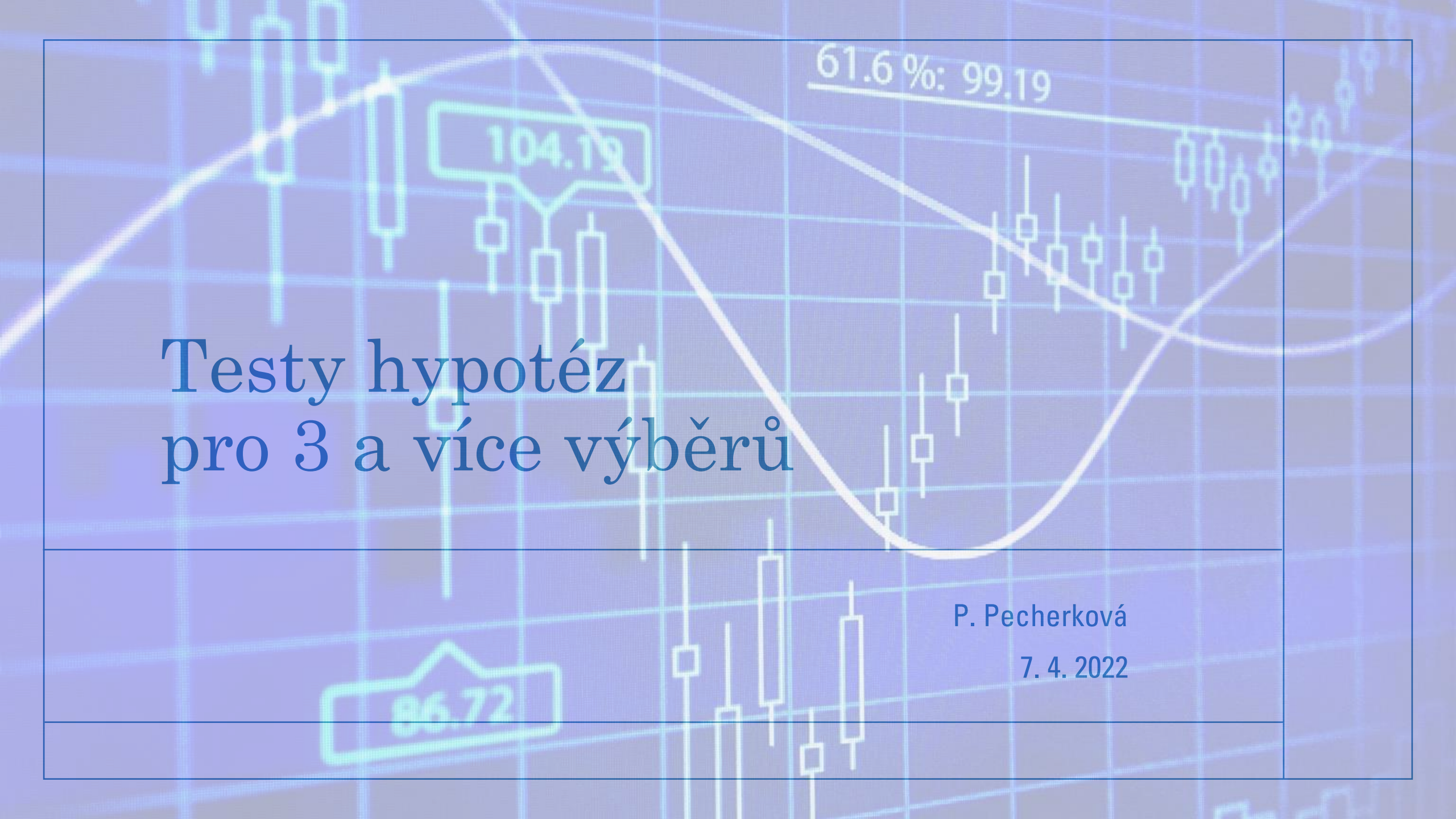




Testy hypotéz pro 3 a více výběrů

P. Pecherková

7. 4. 2022

TH: Počet výběrů

- TH rozlišuje pro určitý počet výběru:
 - 1 výběr: Někdo tvrdí, že parametr výběru č. 1 /rozdělení z výběru má hodnotu xy
 - Výrobce tvrdí, že průměrná doba výroby jeho výrobku je 2h.
 - Studenti tvrdí, že průměrná doba studia vysoké školy je vyšší než 5h/týdně.
 - 2 výběry: Někdo tvrdí, že parametr výběru č. 1 je stejný jako parametr výběru č. 2
 - Výrobce tvrdí, že vyrábí výrobky rychleji než jeho konkurent. (není zde dána žádná přesná hodnota)
 - Studenti tvrdí, že průměrná doba studia na vysoké škole je vyšší než na střední škole.
 - 3 a více výběrů: Někdo tvrdí, že parametr výběru č. 1 a parametr výběru č. 2 a parametr výběru 3.... jsou stejné
 - Výrobce tvrdí, že vyrábí výrobky stejně rychle jako jeho konkurenti. (není zde zadána žádná hodnota, ale naopak je víc konkurentů, tedy víc výběrů).
 - Studenti tvrdí, že průměrná doba jeho studia na vysoké škole je stejná jako na předcházejících školách. (výběry: mateřská škola, základní škola, střední škola, vysoká škola)
 - Určení rychlosti různými přístroji je stejně dobré (výběry: policejní radar, pevný radar, gps detekce...)

TH pro 3 a více výběrů

- Zkoumáme vztah mezi vysvětlovanými (y) a vysvětlujícími proměnnými (x).
- Vysvětlující proměnné označujeme jako faktory, které sledujeme.
- Příklady pro 1 faktor → nepárová/párová data:
 1. Cena vozidla (vysvětlovaná proměnná) závisí na stáří vozidla (faktor).
 2. Počet ujetých km je ovlivněn cenou benzínu.
 3. Velikost sklizně (vysvětlovaná proměnná) závisí na stáří stromu (faktor).
- Příklady 2 faktory → párová data:
 1. Cena vozidla (vysvětlovaná proměnná) závisí na stáří vozidla (faktor 1) nebo na značce vozidla (faktor2).
 2. Cena benzínu (vysvětlovaná proměnná) závisí na kurzu dolaru (faktor 1) nebo na prodejci (faktor 2)

TH: počet výběrů

Někdy se může stát, že se v průběhu zpracování dat změní situace a je nutno změnit test.
Ukážeme si to na příkladu: vývoj pohlaví

H0: Ověřte tvrzení, že pohlaví nemá vliv na průměrnou délku nákupu

- Druhy pohlaví: muž, žena
- H0: $\mu_1 = \mu_2$
- HA: $\mu_1 \neq \mu_2$

DŘÍVE

ttest 2 (normalita dat) / Mann-Whitney test

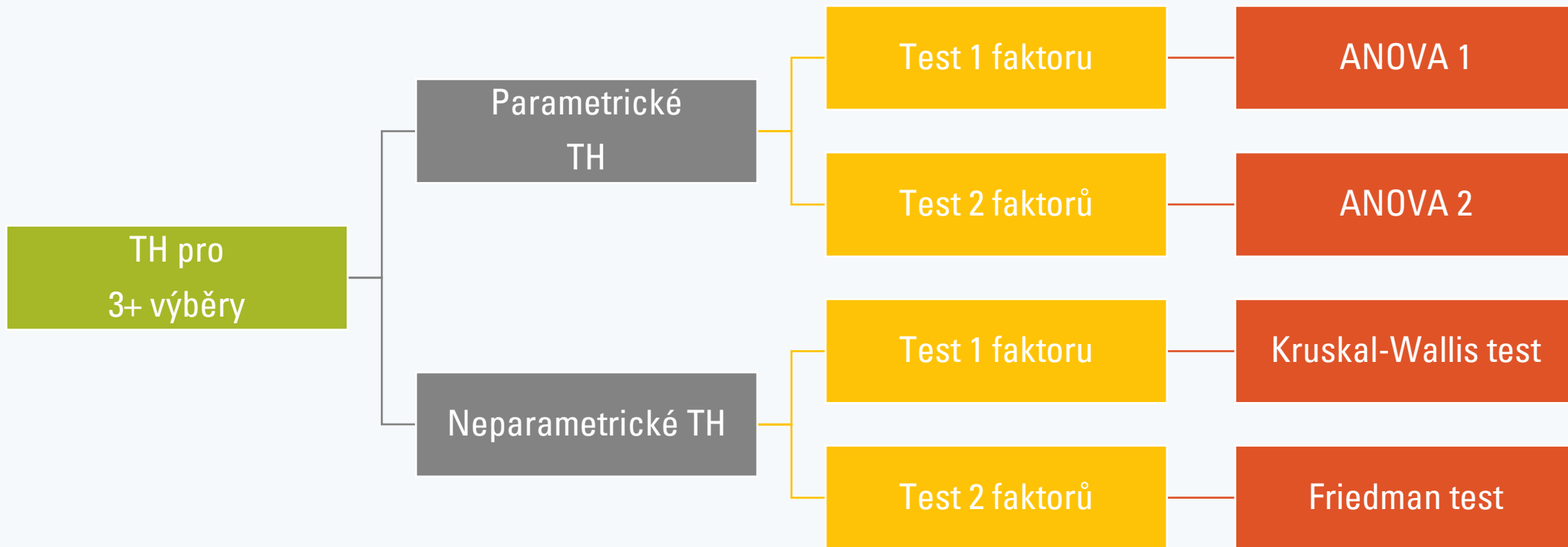
H0: Ověřte tvrzení, že pohlaví nemá vliv na průměrnou délku nákupu

- Druhy pohlaví: muž, žena, agender, bigender, Non-binary...
- H0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \dots$
- HA: $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \neq \dots$

DNES

ANOVA / Kruskal-Wallis test

TH pro 3 a více výběrů



Pre-analýza (obecně)

- Pro použití všech těchto testů:
 - Jedna proměnná je spojitá (vysvětlovaná) a ostatní jsou diskrétní (faktory)
- Pro užití ANOVA testu je nutné splnit 2 podmínky:
 1. Normalita dat (AD test, QQ plot, Shapiro test...)
 2. Rozptyly v jednotlivých skupinách jsou stejné:
 - Neočekává se absolutní shoda, ale zase otestujeme. Použijeme Bartlett test

Bartlett test

Na hladině významnosti 0,05 testujeme hypotézu, že ve skupinách mají stejný rozptyl. Tedy zda je shodný nevysvětlený rozptyl (viz další slide).

- $H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \dots$
- $H_A: \sigma_1^2 \neq \sigma_2^2 \neq \sigma_3^2 \neq \dots$

→ pro ANOVA test je nutné, abychom H_0 nezamítli.

ANOVA 1 (jednofaktorová ANOVA)

- Je založena na zjišťování rozdílů průměru mezi více skupinami (3+) pomocí testovacího kritéria F

- $F = \frac{\sigma_{MS}^2}{\sigma_{US}^2}$

- kde σ_{MS}^2 je rozptyl mezi skupinami (vysvětlený) a σ_{US}^2 je rozptyl uvnitř skupin (nevysvětlený).

- Co znamená rozptyl uvnitř skupin?

- Rozptyl v každé skupině může být různě velký – ale ne o moc (viz Bartlettův test).
- U rozptylu uvnitř skupiny nevíme proč tam je, je to náhodná chyba způsobena například chybou měření. Protože nevíme, proč tam je, říká se mu: nevysvětlitelný rozptyl.

- Co znamená rozptyl mezi skupinami?

- Spočítáme si průměr v každé skupině a spočítáme „společný“ průměr ($\bar{x}_{all}$).
- Spočítáme si rozptyl – každého datového bodu od tohoto „společného“ průměru, tedy $\frac{1}{n-1} \sum (x_i - \bar{x}_{all})$
- Čím jsou od sebe více vzdáleny jednotlivé výběry, tím větší tedy bude rozptyl vysvětlený

Hypotéza

- $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \dots$
- $H_A: \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \neq \dots$

Na několika benzínových stanicích v Evropě jsme sledovali cenu benzínu Natural 95 (rok 2020). Data jsme zapsali do tabulky. Otestujte tvrzení, že cena benzínu je v různých státech stejná.

Německo	33.82	34.72	35.14	35.58	35.93	34.76	33.50	34.10	33.85	
Bělorusko	17.99	17.85	18.41	18.41	17.23	18.06	17.93	18.72		
Itálie	39.85	40.88	39.46	39.57	40.79	42.48	39.95	41.20	40.55	42.02
ČR	30.83	29.61	29.65	29.81	28.53	31.16	30.49	29.84	31.12	29.27

Pre-analýza

1. Test normality dat

```
p_norm=[p_CR p_BY p_I p_DE]
```

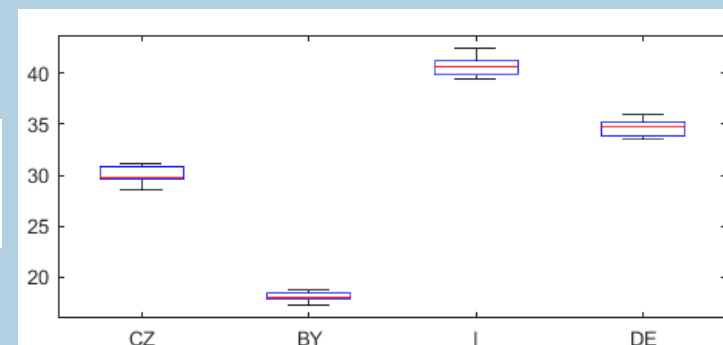
```
p_norm = 1×4
    0.5660    0.5965    0.6240    0.6987
```

Nezamítám hypotézu o normalitě dat.

2. Bartlettův test

```
Bartlett's statistic    4.23561
Degrees of freedom      3
p-value                 0.23712
```

Nezamítám hypotézu o shodě rozptylů.



Analýza

Test: ANOVA 1

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

$$H_A: \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$$

ANOVA Table					
Source	SS	df	MS	F	Prob>F
Groups	2391.07	3	797.024	1143.36	2.10502e-33
Error	23	33	0.697		
Total	2414.08	36			

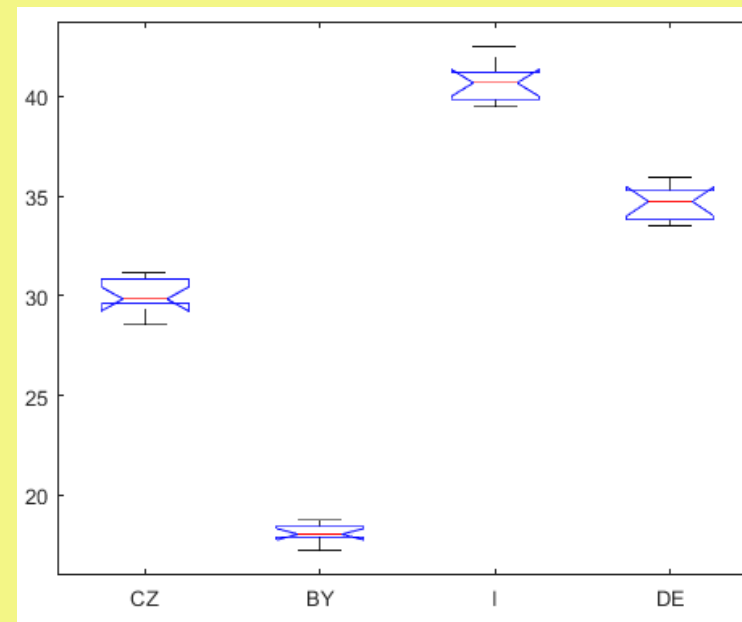
P-hodnota

Zamítáme H_0 o tom, že jsou ceny benzínu všude stejné.

Ale je důležité si říci, které výběry se liší od kterých a které se naopak vůbec neliší



Toto řeší post analýza



Toto není klasický krabicový diagram, zde:

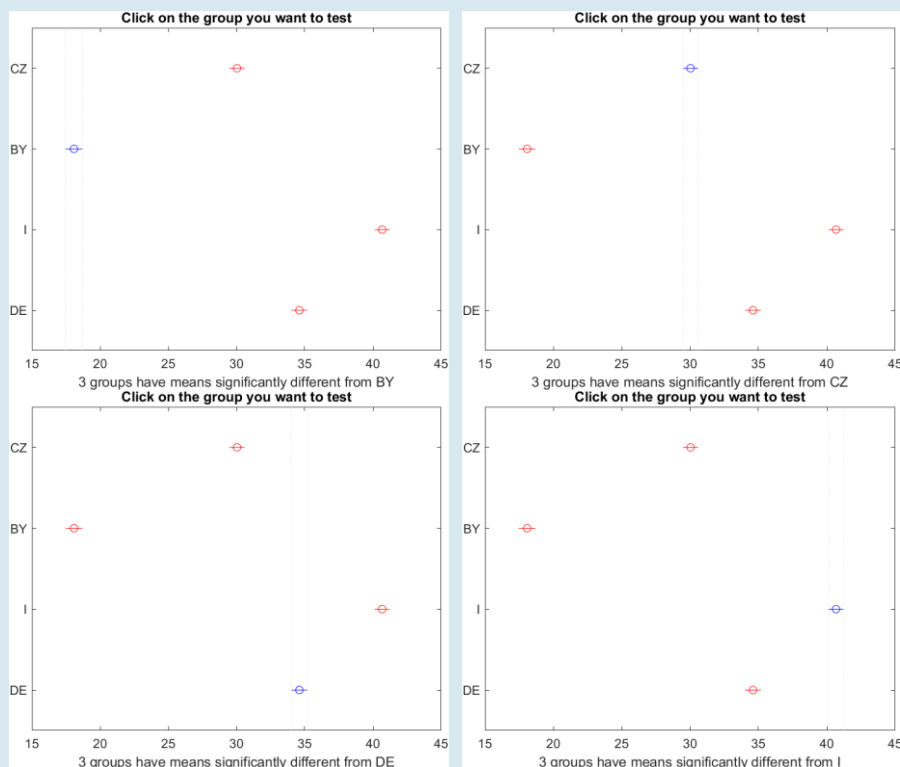
- Červená čára je střední hodnota (průměr).
- Modrý objekt je rozptyl

Post-analýza (obecně)

- Post-analýzu provádíme pouze tehdy, pokud zamítáme H_0 , tedy $p < 0,05$. Tímto testem zjistíme, které výběry případně skupiny výběrů se liší.
- Existuje několik druhů post-analýz, my si pro parametrický a neparametrický test ukážeme 1:
 1. Scheffého test, Tukey-Kramer test – pro ANOVA
 2. Bonferroni test, Tukey-Kramer test – pro Kruskal-Wallisův test / Friedmanův test
- Na cvičení budeme používat: Tukey-Kramer test pro oba druhy.

Post-analýza

- Protože jsme měli signifikantní rozdíl mezi skupinami (H_0 jsme zamítli), použili jsme Scheffého test jako post-analýzu pro ANOVA test.

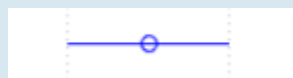


post_anova = 6x6

1.0000	2.0000	10.7867	11.9531	13.1196	0.0000
1.0000	3.0000	-11.7442	-10.6444	-9.5447	0.0000
1.0000	4.0000	-5.6997	-4.5698	-3.4400	0.0000
2.0000	3.0000	-23.7640	-22.5976	-21.4311	0.0000
2.0000	4.0000	-17.7178	-16.5230	-15.3281	0.0000
3.0000	4.0000	4.9447	6.0746	7.2045	0.0000

Výběr č. Výběr č. P-hodnota

V tabulce vidíme, že ani jeden výběr nemá střední hodnotu shodnou se střední hodnotou jiného výběru.



Vodorovné čáry v kroužku znamenají IS kolem střední hodnoty. Pokud se budou protínat, bude to znamenat, výběry mohou pocházet ze stejného souboru a tedy $p > \alpha$

ANOVA 2 (dvoufaktorová ANOVA)

- Dvoufaktorová anova je rozšířením jednofaktorové ANOVA, ale tentokrát se ptáme, zda je rozdíl mezi skupiny je dán rozdílem na základě faktoru 1 či faktoru 2.
- Předpoklad: párová data, normalita dat, stejné rozptyly ve skupinách (Bartlett test)

Tvrzení po řádcích (výši nákupu ovlivňuje den v týdnu)

- $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \dots$
- $H_A: \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \neq \dots$

Tvrzení po sloupcích (výši nákupu ovlivňuje věk nakupujícího)

- $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \dots$
- $H_A: \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \neq \dots$

Tvrdí se, že hodnota nákupu není ovlivněna ani pohlavím nakupujícího, ani dnem v týdnu . (Cena nákupu je spojitá náhodná veličina, den v týdnu diskrétní (Pondělí – Neděle) a totéž i pohlaví (dítě – dospělý – senior).

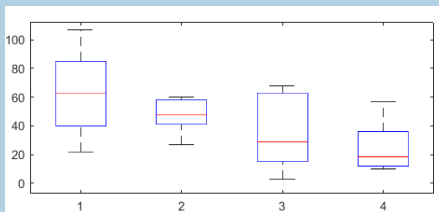
- Příklad:
- *Na 6 studentech jsme otestovali náročnost vyřešení rébusů. Každý student dostal 4 rébusy. Dobu jejich řešení (v minutách) jsme zapsali do tabulky. Na hladině významnosti 0,05 testujte tvrzení, že doba řešení rébusů nezáleží na rébusu ani na studentovi. Předpokládáme normalitu dat.*

	St. 1	St. 2	St. 3	St. 4	St. 5	St. 6
Rébus 1	85	107	61	22	40	65
Rébus 2	58	51	60	47	45	27
Rébus 3	15	30	68	63	28	3
Rébus 4	57	12	36	21	10	16

Pre-analýza

1. **Test normality dat:** ze zadání máme zadáno. Jinak bychom museli udělat test pro každý řádek a každý sloupec, tedy 10 testů normality. Každý by musel vyjít jako normální, abychom mohli použít ANOVA test.
2. **Bartlettův test:** musíme zjistit, zda jsou shodné rozptyly po sloupcích i po řádcích

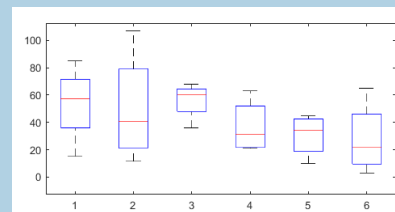
Po řádcích (4 rébusy)



Bartlett's statistic 4.01433
Degrees of freedom 3
p-value 0.25992

nezamítám

Po sloupcích (6 studentů)



Bartlett's statistic 4.40547
Degrees of freedom 5
p-value 0.49263

nezamítám

Analýza

Test: ANOVA2

Po řádcích (rébusy trvá vyřešit stejně dlouho) Po sloupcích (každý student řeší rébusy stejnou dobu)

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

$$H_A: \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$$

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$$

$$H_A: \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \neq \mu_5 \neq \mu_6$$

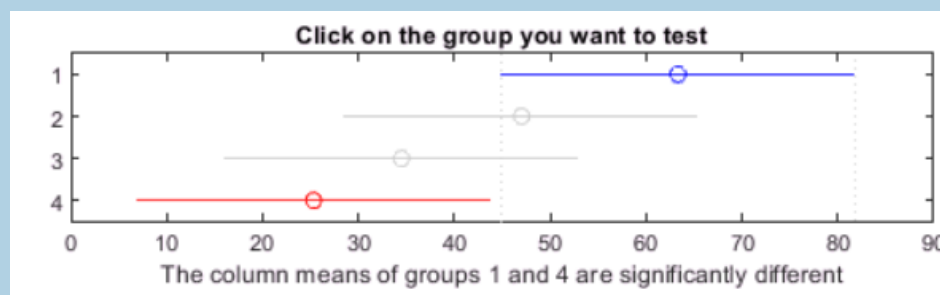
$p < \alpha \rightarrow$
hypotézu
zamítáme

ANOVA Table					
Source	SS	df	MS	F	Prob>F
Columns	4877.79	3	1625.93	3.32	0.0487
Rows	3042.21	5	608.44	1.24	0.3382
Error	7345.96	15	489.73		
Total	15265.96	23			

$p > \alpha \rightarrow$
hypotézu
nezamítáme

Post-analýza

Jen pro rozdíl v rébusech,
protože mezi studenty
rozdíl není



Statistický rozdíl je mezi
1. a 4. rébusem.

Kruskal-Wallisův test

- Neparametrická obdoba jednofaktorového testu ANOVA.
- Používáme v případě, kdy nemáme normalitu dat nebo stejný rozptyl dat → stačí, abychom zamítnuli tvrzení pro 1 výběr a musíme použít Kruskal-Wallisův test.

Hypotéza

- H_0 : všechna data (groups) pochází ze stejného rozdělení, tedy $\tilde{x}_{0,5}^{(1)} = \tilde{x}_{0,5}^{(2)} = \tilde{x}_{0,5}^{(3)} = \tilde{x}_{0,5}^{(4)} \dots$
- H_A : všechna data (groups) nepochází ze stejného rozdělení, tedy $\tilde{x}_{0,5}^{(1)} \neq \tilde{x}_{0,5}^{(2)} \neq \tilde{x}_{0,5}^{(3)} \neq \tilde{x}_{0,5}^{(4)} \dots$
- Dá se říci, že střední hodnota u všech výběrů je stejná
- Zamítnu-li, znamená to, že alespoň 2 výběry se liší

- Zkoumáme výsledky jednotných přijímacích zkoušek na víceletá gymnázia z okresů A, B, C a D. Zajímá nás, zda se výsledky okresů liší. Otestujte tvrzení, výsledky zkoušek jsou všude stejné.

A	79	77	98	98	72	100	74	96	67	98
B	89	79	60	94	46	69	82	91		
C	73	91	78	65	62	89	63	94	84	
D	63	64	72	58	60	86	61	62		

Pre-analýza

1. Test normality dat

```
[h,pkw1]=adtest(y1)
```

```
h = logical
     1
pkw1 = 0.0220
```

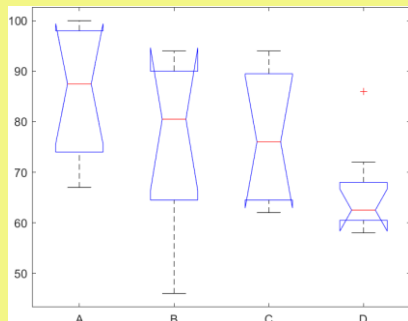
Hned prvním výběrem zamítám hypotézu, proto není nutné kontrolovat další a víme, že nemůžeme použít ANOVA 1.

Pokud si otestujeme i tak všechny výběry zjistíme, že zamítáme 1. a 4. výběr, naopak 2. a 3. nezamítáme.

Batlettův test dělat nemusíme.

Analýza

- H0: všechna data (groups) pochází ze stejného rozdělení, tedy $\tilde{x}_{0,5}^{(A)} = \tilde{x}_{0,5}^{(B)} = \tilde{x}_{0,5}^{(C)} = \tilde{x}_{0,5}^{(D)}$
- HA: všechna data (groups) nepochází ze stejného rozdělení, tedy $\tilde{x}_{0,5}^{(A)} \neq \tilde{x}_{0,5}^{(B)} \neq \tilde{x}_{0,5}^{(C)} \neq \tilde{x}_{0,5}^{(D)}$



Kruskal-Wallis ANOVA Table					
Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Groups	1076.54	3	358.847	10.27	0.0164
Error	2487.46	31	80.241		
Total	3564	34			

P-hodnota

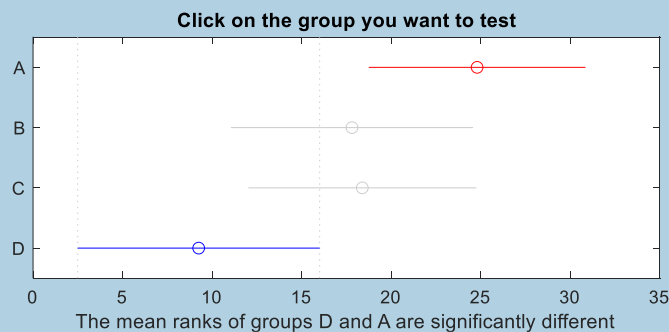
Zamítáme nulovou hypotézu, data nepochází ze stejného rozdělení.

Post-analýza

1.0000	2.0000	-5.8251	6.9875	19.8001	0.9012
1.0000	3.0000	-5.9998	6.4111	18.8220	1.0000
1.0000	4.0000	2.7374	15.5500	28.3626	0.0082
2.0000	3.0000	-13.7016	-0.5764	12.5488	1.0000
2.0000	4.0000	-4.9432	8.5625	22.0682	0.5664
3.0000	4.0000	-3.9863	9.1389	22.2641	0.3973

Výběr č. Výběr č.

P-hodnota



Statisticky významný rozdíl je mezi krajem A a D.

Friedmanův test

- Neparametrická obdoba dvoufaktorového testu ANOVA.
- Předpoklad: párová data
- Používáme v případě, kdy nemáme normalitu dat nebo stejný rozptyl dat → stačí, abychom zamítli tvrzení pro 1 výběr a musíme použít Friedmanův test
- U Friedmanova testu také sledujeme 2 faktory – tedy po sloupcích a po řádcích.

Tvrzení po řádcích

- H_0 : všechna data (groups) pochází ze stejného rozdělení, tedy $\tilde{x}_{0,5}^{(1)} = \tilde{x}_{0,5}^{(2)} = \tilde{x}_{0,5}^{(3)} = \tilde{x}_{0,5}^{(4)} \dots$
- H_A : všechna data (groups) nepochází ze stejného rozdělení, tedy $\tilde{x}_{0,5}^{(1)} \neq \tilde{x}_{0,5}^{(2)} \neq \tilde{x}_{0,5}^{(3)} \neq \tilde{x}_{0,5}^{(4)} \dots$

Tvrzení po sloupcích

- H_0 : všechna data (groups) pochází ze stejného rozdělení, tedy $\tilde{x}_{0,5}^{(1)} = \tilde{x}_{0,5}^{(2)} = \tilde{x}_{0,5}^{(3)} = \tilde{x}_{0,5}^{(4)} \dots$
- H_A : všechna data (groups) nepochází ze stejného rozdělení, tedy $\tilde{x}_{0,5}^{(1)} \neq \tilde{x}_{0,5}^{(2)} \neq \tilde{x}_{0,5}^{(3)} \neq \tilde{x}_{0,5}^{(4)} \dots$

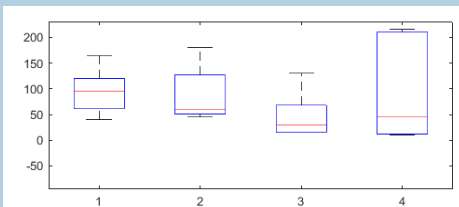
Na 6 myších jsme otestovali náročnost vyřešení bludiště. Každá myš musela proběhnout 4 bludiště. Dobu jejich cesty (ve vteřinách) jsme zapsali do tabulky. Na hladině významnosti 0,05 testujte tvrzení, že doba proběhnutí bludiště nezáleží na typu bludiště ani na myši. Předpokládáme normalitu dat.

	Myš 1	Myš 2	Myš 3	Myš 4	Myš 5	Myš 6
Bludiště 1	85	107	61	120	40	165
Bludiště 2	58	51	60	180	45	127
Bludiště 3	15	30	68	130	28	15
Bludiště 4	57	12	36	210	10	216

Pre-analýza

- 1. Test normality dat:** ze zadání máme zadáno. Jinak bychom museli udělat test pro každý řádek a každý sloupec, tedy 10 testů normality. Každý by musel vyjít jako normální, abychom mohli použít ANOVA test.
- 2. Bartlettův test:** musíme zjistit, zda jsou shodné rozptyly po sloupcích i po řádcích

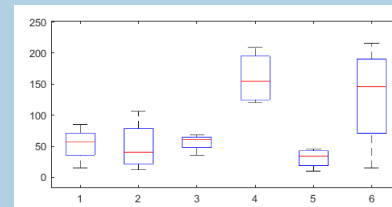
Po řádcích (4 bludiště)



Bartlett's statistic 4.22705
Degrees of freedom 3
p-value 0.23797

nezamítáme

Po sloupcích (6 myší)



Bartlett's statistic 11.4621
Degrees of freedom 5
p-value 0.043

ZAMÍTÁME

Analýza

Test: Friedman test

Po řádcích (bludiště jsou stejně těžká)

$$H_0: \tilde{x}_{0,5}^{(1)} = \tilde{x}_{0,5}^{(2)} = \tilde{x}_{0,5}^{(3)} = \tilde{x}_{0,5}^{(4)}$$

$$H_A: \tilde{x}_{0,5}^{(1)} \neq \tilde{x}_{0,5}^{(2)} \neq \tilde{x}_{0,5}^{(3)} \neq \tilde{x}_{0,5}^{(4)}$$

Friedman's ANOVA Table					
Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Columns	4.3333	3	1.44444	2.6	0.4575
Error	25.6667	15	1.71111		
Total	30	23			

nezamítáme

Test for column effects after row effects are removed

Po sloupcích (všechny myši běhají stejně rychle)

$$H_0: \tilde{x}_{0,5}^{(1)} = \tilde{x}_{0,5}^{(2)} = \tilde{x}_{0,5}^{(3)} = \tilde{x}_{0,5}^{(4)} = \tilde{x}_{0,5}^{(5)} = \tilde{x}_{0,5}^{(6)}$$

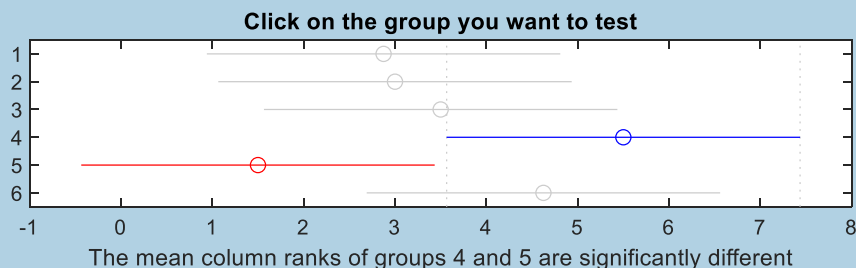
$$H_A: \tilde{x}_{0,5}^{(1)} \neq \tilde{x}_{0,5}^{(2)} \neq \tilde{x}_{0,5}^{(3)} \neq \tilde{x}_{0,5}^{(4)} \neq \tilde{x}_{0,5}^{(5)} \neq \tilde{x}_{0,5}^{(6)}$$

Friedman's ANOVA Table					
Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Columns	39.625	5	7.925	11.4	0.044
Error	29.875	15	1.99167		
Total	69.5	23			

zamítáme

Test for column effects after row effects are removed

Post-analýza



3.0000	5.0000	-1.8690	2.0000	5.8690	1.0000
3.0000	6.0000	-4.9940	-1.1250	2.7440	1.0000
4.0000	5.0000	0.1310	4.0000	7.8690	0.0361
4.0000	6.0000	-2.9940	0.8750	4.7440	1.0000
5.0000	6.0000	-6.9940	-3.1250	0.7440	0.2663

Statisticky významný rozdíl je mezi 4. a 5. myší.