Přednáška 8 – Testy hypotéz více veličin

Minulý týden jsme se seznámili s testy hypotéz pro dva výběry, kde se testy, podobně jako i pro jeden výběr, dělily na parametrické s předpokladem normality a neparametrické bez předpokladu normality, a navíc bylo potřeba zvolit mezi testy pro párové a nepárové výběry. Dnes se soustředíme na testy pro více výběrů.

Testy hypotéz více výběrů

V případě vícevýběrových těstů přidáme navíc ještě další předpoklad – jednofaktorové a dvoufaktorové testy. Faktor je určitá podmínka, na základě které zkoumáme shodu výběrů (správně by se mělo říct spíš shodu souborů, ze kterých výběry pochází). Například, porovnáváme plat zaměstnanců v závislosti na vzdělání (bereme v úvahu, například, tři možnosti: bez maturity, maturita, vysoká škola). Tady je vzdělání jeden faktor, v závislosti na kterém porovnáváme platy. Dále nás může zajímat, jak se liší platy v závislosti na vzdělání a na pohlaví. V tomto případě máme dva faktory - vzdělání a pohlaví.

V následující tabulce jsou testy pro více výběrů, které budeme používat (je také dostupná na webu na odkazu Jak zvolit test hypotéz). U každého testu jsou uvedeny specifické předpoklady jeho použití a nulové hypotézy.

Jednofaktorové	
Parametrické	Neparametrické
ANOVA – **anova_1** Předpoklady: $N(\mu, \sigma^2)$, $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$, nepárové výběry $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$	Kruskal-Wallisův test **kruskal_test** Předpoklady: bez $N(\mu, \sigma^2)$, nepárové výběry $H_0 : \tilde{X}_{0,5(1)} = \tilde{X}_{0,5(2)} = \ldots = \tilde{X}_{0,5(k)}$

Dvoufaktorové	
Parametrické	Neparametrické
ANOVA – **anova_2** Předpoklady: $N(\mu, \sigma^2)$, $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$ $H_0 : \mu_1 = \ldots = \mu_n$ (v řádcích) $H_0 : \mu_1 = \ldots = \mu_k$ (ve sloupcích)	Friedmanův test **friedman_test** Předpoklady: bez $N(\mu, \sigma^2)$ párové výběry $H_0 : \tilde{X}_{0,5(1)} = \ldots = \tilde{X}_{0,5(k)}$

Pre-analýza: Ověření předpokladu stejných rozptylů pro ANOVA:
Bartlettův test – **bartlett_test** $H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$

Post-analýza: v případě zamítnutí nulové hypotézy	
po zamítnutí ANOVA	po zamítnutí Kruskal-Wallisova nebo Friedmanova testu
Scheffého test – **scheffe_test** $H_0 : \mu_1 = \mu_2$ nebo $\mu_1 = \mu_3$ nebo $\ldots$ $\mu_1 = \mu_k$ nebo $\mu_2 = \mu_3$ nebo $\ldots$ (ve dvojicích)	Bonferroniho test $H_0 : \tilde{X}_{0,5(1)} = \tilde{X}_{0,5(2)}$ nebo $\tilde{X}_{0,5(1)} = \tilde{X}_{0,5(3)}$ nebo $\ldots$ $\tilde{X}_{0,5(1)} = \tilde{X}_{0,5(k)}$ nebo $\tilde{X}_{0,5(2)} = \tilde{X}_{0,5(3)}$ nebo $\ldots$ (ve dvojicích)

Podíváme se na každý z testů podrobněji.

Jednofaktorová ANOVA

Jednofaktorový test Anova (angl. analysis of variance) je určen pro testování shody více výběrů na základě určitého faktoru. Tento test použijeme za předpokladu:

- normality všech výběrů,
- stejných rozptylů výběrů.

Pokud data nesplňují jeden z těchto předpokladů, nemůžeme použít test Anova a musíme zvolit jeho neparametrickou alternativu (Kruskal-Wallisův test). Výběry nemusí být párové.

Příklad: Zajímá nás, zda se liší průměrný denní počet kroků mezi různými věkovými kategoriemi – mládeží, dospělými a seniory. Data z fitness náramků, mobilních telefonů a chytrých hodinek jsme si zapsali do tabulky:

osoba \ věk	mládež	dospělí	senioři
1	7652	6114	3468
2	7483	11026	4579
3	9806	8941	4668
4	5854	10412	4030
5	6448	8693	2283
6	5835	11212	4914
7	4300	9895	3263
8	6624	6626	6242
9	7815	7714	
10		9163	
11		9344	
12		9166	

Řešení: Tady máme 3 nepárové výběry počtů kroků mládeže, dospělých a seniorů a jeden faktor – věk. Abychom věděli, zda můžeme použít test Anova, musíme zaprvé ověřit předpoklad normality všech tří výběrů. Pokud jsme nezamítli nulovou hypotézu, že data pochází z normálního rozdělení pro všechny 3 výběry, můžeme pokračovat dál.

Dále pro použití testu Anova je nezbytné ověřit předpoklad stejných rozptylů výběrů. K tomu je určen Bartlettův test (**bartlett_test**). Nulová hypotéza Bartlettova testu tvrdí, že rozptyly všech výběrů jsou stejné (toto tvrzení se nikdy nemění):

$$H_0: \quad \sigma_1^2 = \sigma_2^2 = \ldots = {\sigma_k}^2, \quad \text{tj., v našem případě} \quad \sigma_1^2 = \sigma_2^2 = \sigma_3^2.$$

Alternativní hypotéza říká: $H_A:$ alespoň jeden rozptyl se liší, tj., nejsou stejné.

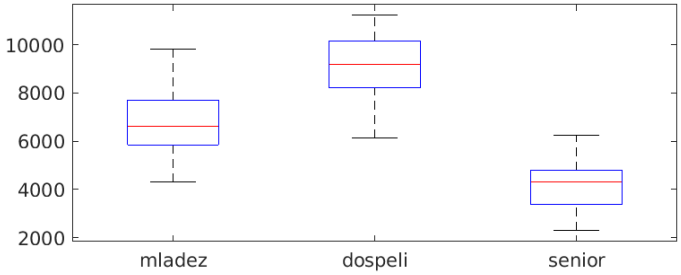
Tady je k počet výběrů. Bartlettův test je pouze pravostranný, ale směr testu tady nevolíme. Používá následující statistiku:

$$T = \frac{(N-k)\ln(S_p^2) - \sum_{i=1}^{k}(n_i-1)\ln(s_i^2)}{1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\frac{1}{n_i-1} - \frac{1}{N-k}\right)} \sim \chi^2\text{-rozdělení}, \quad \text{kde } k \text{ – počet výběrů, } n_i \text{ – rozsah } i\text{-ho výběru,}$$

$$N = \sum_{i=1}^{k} n_i, \qquad S_p^2 = \frac{1}{N-k}\sum_{i=1}^{k}(n_i-1)s_i^2 \quad \text{je souhrnný rozptyl,} \quad s_i^2 - \text{výběrový rozptyl } i\text{-ho výběru.}$$

Pro náš příklad se p-hodnota Bartlettova testu rovná 0.7259, což znamená, že nezamítáme nulovou hypotézu o shodě rozptylů všech výběrů. Naše data splňují nutné předpoklady (normalitu+rozptyly), tedy můžeme použít test Anova. Pozor, pokud bychom zamítli Bartlettův test i s ověřeným předpokladem normality dat, nemůžeme použít Anova a musíme zvolit jeho neparametrickou alternativu!

Při testování shody rozptylů je vhodné vykreslit krabicové diagramy výběrů, kde můžeme případné rozdíly v datech graficky znázornit:



Na obrázku vidíme, že se rozpětí výběrů moc neliší.

Po ověření nutných předpokladů můžeme použít jednofaktorový test Anova (**anova_1**). Nulová hypotéza říká, že střední hodnoty výběrů jsou stejné (toto tvrzení nikdy neměníme):

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k, \quad \text{tj., pro náš příklad} \quad \mu_1 = \mu_2 = \mu_3,$$

slovy: mládež, dospělí a senioři ujdou denně v průměru stejný počet kroků, tj., věk (jeden faktor) nemá vliv na počet kroků.

Alternativní hypotéza říká: H_A : alespoň jedna střední hodnota se liší, tj., neujdou stejné množství kroků a věk má vliv.

Anova je pouze pravostranný test, ale směr testu tady nevolíme. Princip výpočtu statistiky je následující. Označíme věkové kategorie V_1, V_2 a V_3 a spočteme průměry ze sloupců pro každý věk $\bar{V}_1$, $\bar{V}_2$, $\bar{V}_3$. Zprůměrujeme tyto průměry ze sloupců a dostaneme průměr průměrů z celé tabulky

$$\bar{V} = \frac{1}{k}\sum_{i=1}^{k} \bar{V}_i.$$

Dále spočteme rozptyl mezi sloupci, tj., mezi různými věkovými kategoriemi, který po algebraických úpravách je

$$s_M^2 = \sum_{i=1}^{k} n_i(\bar{V}_i - \bar{V})^2, \quad \text{kde} \quad n_i \text{ – rozsah } i\text{-ho výběru.}$$

Tento rozptyl je tzv. vysvětlený rozptyl, protože ho můžeme vysvětlit rozdílem průměrů ze sloupců $\bar{V}_i$. Dále spočteme rozptyl dat z celé tabulky, ktery tady bude

$$s_T^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(V_{j,i} - \bar{V}_i)^2, \quad \text{kde} \quad V_{j,i} \text{ – počet kroků } j\text{-té osoby } i\text{-ho výběru.}$$

Tento rozptyl je tzv. nevysvětlený – nemůžeme ho vysvětlit, protože nevíme, proč se hodnoty ve výběrech liší. Statistika testu Anova je podíl vysvětleného a nevysvětleného rozptylu:

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

Pro náš příklad p-hodnota=0.0000007, což znamená, že zamítáme nulovou hypotézu, že věk nemá vliv na denní počet kroků.

Pokud jsme zamítli nulovou hypotézu Anova, znamená to, že všechny střední hodnoty nejsou stejné. To je vícevýběrový test, tj., nemůžeme hned říct, která ze středních hodnot (věkových kategorií) se lišila (zkuste to porovnat s dvouvýběrovým). K tomu, abychom zjistili, který výběr se lišil, nám poslouží testy post-analýzy, například Scheffého test.

Scheffého test (**scheffe_test**) porovnává střední hodnoty výběrů ve dvojicích podobně jako **t_test**, ale jeho nulová hypotéza se skládá z variant

$$H_0 : \mu_1 = \mu_2 \text{ nebo } \mu_1 = \mu_3 \text{ nebo } \ldots \mu_1 = \mu_k \text{ nebo } \mu_2 = \mu_3 \text{ nebo } \ldots.$$

Výsledkem použití Scheffého testu je tabulka, kde v horní části nad hlavní diagonálou jedničkami jsou označeny dvojice, které dávaly odlišné střední hodnoty. Pro náš příklad to je tabulka 3x3:

$$\begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

kde 1 znamenají, že se lišily střední hodnoty dvojic mládež – dospělí, mládež – senior a dospělí – senior, což celkem vysvětluje, proč byla p-hodnota tak malá.

Kruskal-Wallisův test

Kruskal-Wallisův test (**kruskal_test**) je neparametrická alternativa jednofaktorového testu Anova. To znamená, že ho použijeme v případě více výběrů, pokud alespoň jeden z výběrů nepochází z normálního rozdělení nebo pokud nejsou rozptyly výběrů stejné. Výběry nemusí být párové.

Příklad: Zkoumáme výsledky jednotných přijímacích zkoušek na víceletá gymnázia z okresů A, B, C a D. Zajímá nás, zda se výsledky okresů liší.

uchazeč \ okres	A	B	C	D
1	79	89	73	91
2	77	79	91	64
3	98	60	76	72
4	67	94	65	58
5	98	46	62	60
6	98	69	89	86
7	72	82	63	61
8	100	91	94	62
9	74		84	
10	96			

Řešení: Tady máme 4 nepárové výběry a jeden faktor – okres. Otestovali jsme data na normalitu a hned u prvního výběru jsme museli zamítnout předpoklad normality. Dále už nemusíme testovat předpoklady – můžeme použít pouze neparametrický Kruskal-Wallisův test*.

Nulová hypotéza Kruskal-Wallisova testu zní: rozdělení, ze kterých pochází výběry, jsou stejná, případně: jejich mediány jsou stejné:

$$H_0 : \tilde{X}_{0,5(1)} = \tilde{X}_{0,5(2)} = \ldots = \tilde{X}_{0,5(k)}, \text{ tj., pro náš příklad } \tilde{X}_{0,5(A)} = \tilde{X}_{0,5(B)} = \tilde{X}_{0,5(C)} = \tilde{X}_{0,5(D)},$$

slovně: výsledky přijímacích zkoušek jsou stejné, tj., nezáleží, z jakého okresu je uchazeč.

Alternativní hypotéza říká: H_A : výsledky alespoň jednoho okresu se liší, tj., okres má vliv.

Test je pouze pravostranný. Používá statistiku s využitím pořadí:

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

- kde k – počet výběrů, n_i – rozsah i-ho výběru,
- r_{ij} – pořadí i-té hodnoty z j-ho výběru,
- $N = \sum_{i=1}^{k} n_i$, $\bar{r} = \frac{1}{2}(N+1)$,
- $\bar{r}_i$ – průměr pořadí v j-tém výběru.

Pro náš příklad nám vyšla p-hodnota=0.061, což znamená, že velmi těsně nezamítáme nulovou hypotézu, že okres nemá vliv na výsledky uchazeče. Je nutné si ale poznamenat, že kdybychom zamítli nulovou hypotézu, podobně jako v případě Anova, můžeme využít post-analýzy a zjistit, ze kterého okresu se výsledky lišily. Za tímto účelem se doporučuje použití Bonferroniho testu, což je alternativa Scheffého testu pro neparametrické vícevýběrové testy. Test je také založen na zkoumání rozdílů ve dvojicích výběrů (viz tabulka Jak zvolit test hypotéz).

*Kdybychom měli 2 nepárové výběry, jaký test by šlo použít? Mann-Whitneyův test

Dvoufaktorová ANOVA

Dvoufaktorový test Anova (**anova_2**) patří k parametrickým testům, který použijeme v případě, že máme tři a více výběrů a 2 faktory, v závislosti na kterých posuzujeme shodu výběrů, a data přitom splňují následující předpoklady:

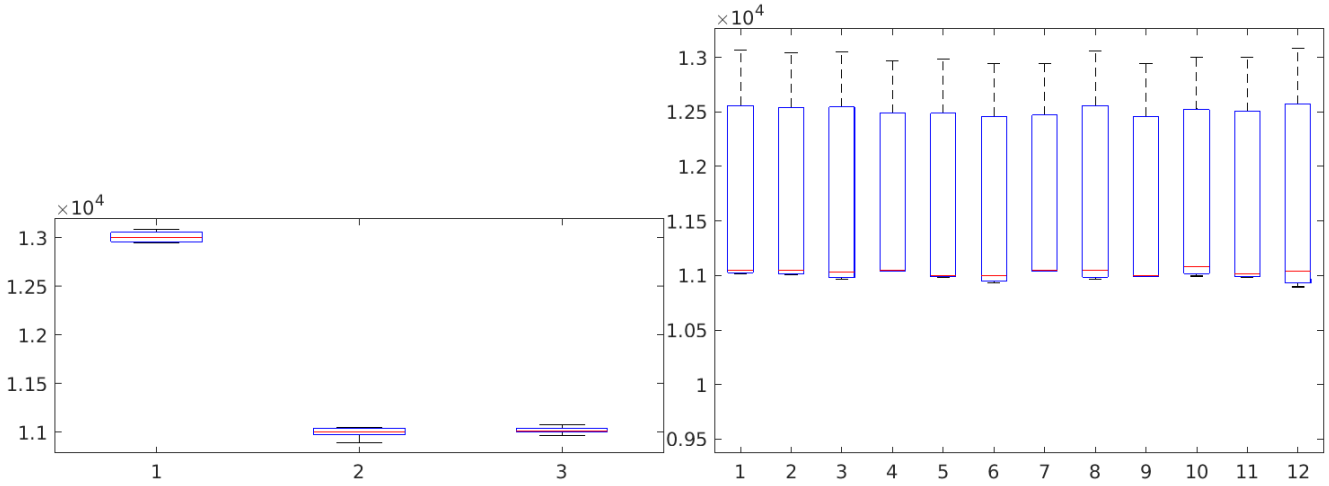
- normální rozdělení dat v řádcích a ve sloupcích tabulky s daty,
- stejné rozptyly dat v řádcích a ve sloupcích.

To znamená, že předpoklady zůstanou stejné, jako v případě **anova_1**, ale musíme je splnit v obou směrech tabulky s daty. Z toho důvodu by měly být počty dat stejné pro všechny výběry. Pokud data nesplňují alespoň jeden z předpokladů, nemůžeme použít dvoufaktorový test Anova a musíme zvolit jeho neparametrickou alternativu Friedmanův test.

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 (posuzuje se úroveň příjmů, LTV, atd.). Data máme v tabulce. Předpokládáme normální rozdělení hodnot jak ve sloupcích, tak i v řádcích.

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

Řešení: Tady máme 3 výběry se stejným počtem dat a 2 faktory – banka a žadatel. Testujeme, zda má banka vliv na výši splátky a zda má žadatel vliv. Předpokládá se normální rozdělení dat v obou směrech tabulky, takže nemusíme testovat data na normalitu. Musíme ale ověřit předpoklad stejných rozptylů pomoci Bartlettova testu. Vychází nám p-hodnota ve sloupcích 0.3292 a p-hodnota v řádcích 1, tj., nezamítáme obě nulové hypotézy, že rozptyly ve sloupcích a v řádcích jsou stejné. Můžeme se přesvědčit, že rozptyly jsou shodné také graficky:



Na obrázku vlevo vidíme krabicové diagramy 3 bank, vpravo – 12 žadatelů. V obou případech je vidět, že se rozpětí moc neliší. Dokážete říct, zda se liší střední hodnoty?

Data splňují všechny předpoklady pro dvoufaktorový test Anova (**anova_2**), tedy ho můžeme použít. Dvoufaktorová Anova funguje tak, jako kdybychom prováděli dvakrát jednofaktorový test Anova, ale zvlášť ve sloupcích (pro banky) a v řádcích (pro žadatele). Tím pádem, má test dvě nulové hypotézy – ve sloupcích a v řádcích:

$$H_0 : \mu_E = \mu_F = \mu_G, \quad \text{střední hodnoty jsou stejné, tj., banka nemá vliv na výši splátky (ve sloupcích).}$$

$$H_A : \quad \text{alespoň jedna střední hodnota se liší, tj., alespoň jedna banka má vliv.}$$

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_{12}, \quad \text{střední hodnoty jsou stejné, tj., žadatel nemá vliv na výši splátky (v řádcích.)}$$

$$H_A : \quad \text{alespoň jedna střední hodnota se liší, tj., alespoň jeden žadatel má vliv.}$$

Test používá dvě statistiky:

$$F_{\text{sloupec-banka}} = \frac{\text{vysvětlený rozptyl}}{\text{nevysvětlený rozptyl}}, \qquad F_{\text{řádek-žadatel}} = \frac{\text{vysvětlený rozptyl}}{\text{nevysvětlený rozptyl}}, \quad \text{obě} \sim \text{Fisherovo rozdělení.}$$

Tím pádem, máme dvě p-hodnoty – ve sloupcích a v řádcích:

$$\text{p-hodnota}_{\text{sloupec-banka}} = 9.898D - 32, \qquad \text{p-hodnota}_{\text{řádek-žadatel}} = 0.6725,$$

což znamená, že zamítáme nulovou hypotézu ve sloupcích, že banka nemá vliv, a nezamítáme nulovou hypotézu v řádcích, že žadatel nemá vliv. To bylo vidět i na krabicových diagramech, že se střední hodnoty bank lišily, ale střední hodnoty všech žadatelů byly shodné. Závěr: je jedno, jaký jsme žadatel, důležité je, do jaké banky půjdeme:-)

Testy post-analýzy se používají podobně jako pro jednofaktorový test Anova.

Poznámka: Dále existuje ještě tzv. dvoufaktorová Anova s opakováním. Kdybychom ji použili na náš příklad, znamenalo by to, že jako faktory máme banku, žadatele, a každý žadatel si navíc přijde několikrát.

V tomto případě testujeme shodu středních hodnot
- ↗ ve sloupcích (v závislosti na bance),
- → v řádcích (v závislosti na žadateli),
- ↘ nezávislost řádků a sloupců.

Proto jsou k tomu tři nulové hypotézy. První dvě jsou totožné s dvoufaktorovým testem Anova. Třetí nulová hypotéza tvrdí, že sloupce a řádky jsou nezávislé. Jako výsledek máme tedy tři p-hodnoty.

Friedmanův test

Friedmanův test (**friedman_test**) je neparametrický dvoufaktorový test, který je určen pro ordinální data (tj., dokážeme data uspořádat) a párové výběry bez předpokladu normality. Specifika Friedmanova testu spočívá v tom, že testuje shodu výběrů s ohledem na jeden faktor ze dvou, a vliv druhého faktoru se snaží eliminovat. Nežádoucí faktor je tvořen pomocí nezávislých bloků, tj., data v blocích nemají vazbu na jiné bloky.

Příklad: 12 náhodně vybraných zákazníků hodnotí kvalitu 5 e-shopů od 0 (nejhorší) do 10 (nejlepší). Ptáme se, zda jsou hodnocení e-shopů stejná. Nepředpokládáme normalitu dat. Máme tabulku hodnocení:

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

Řešení: Každý zákazník hodnotí každý e-shop (tj., máme párové výběry) a je jinak přísný. Nemusíme testovat data na normalitu, což znamená, že použijeme neparametrický test. Máme tady 2 faktory – zákazník (hodnotitel) a e-shop. Každý zákazník tvoří nezávislý blok hodnocení, hodnotí tedy nezávisle na jiných zákaznících. Potřebujeme testovat shodu výběrů v závislosti na e-shopu, čili odpovídáme na otázku, zda e-shop ovlivňuje hodnocení. Ale děláme to nezávisle na zákazníkovi – to je ten faktor, který nás nezajímá.

Nulová hypotéza Friedmanova testu zní: rozdělení, ze kterých pochází výběry, jsou stejná, případně: jejich mediány jsou stejné:

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

slovně: všechny e-shopy jsou hodnoceny stejně.

Alternativní hypotéza říká: H_A : alespoň jeden e-shop je hodnocen jinak.

Statistika testu se počítá podle následujícího postupu. Hodnotu v každém řádku nahradíme jejím pořadím a sečteme pořadí ve sloupcích:

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

Pomocí pořadí tento test odstraní subjektivní přísnost zákazníků – tj., hodnocení je nezávislé na zákazníkovi. Statistika je:

$$T = 3n(k+1)\sum_{i=1}^{k} R_i \sim \chi 2\text{-rozdělení.}$$

Testy post-analýzy se používají podobně jako pro Kruskal-Wallisův test.