

ANOVA – ANALÝZA ROZPTYLU

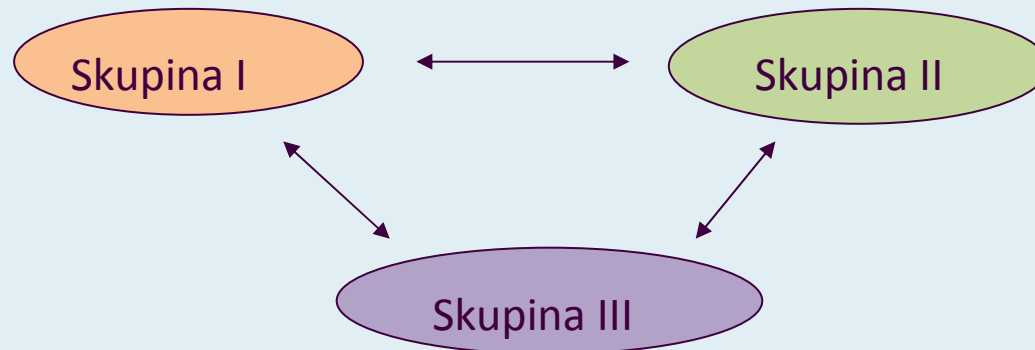


12. cvičení

Jednofaktorová ANOVA

- ❖ Test o shodě více než dvou středních hodnot
- ❖ Použití - příklady:
 - ❖ Porovnání výsledků přijímacího řízení u absolventů různých typů středních škol (gymnázium, SPŠ, SOU)
 - ❖ Srovnání obsahu dusíku u 5-ti příbuzných druhů rostlin
 - ❖ Srovnání platů podle bydliště respondentů (krajů)
 - ❖ Porovnání použité medikace na krevní tlak pacientů

Proč nepoužít řadu dvouvýběrových t -testů



- ❖ Porovnáváme-li k tříd (skupin), provádíme $\binom{k}{2} = \frac{k \cdot (k - 1)}{2}$ testů. V každém z nich je pravděpodobnost chyby prvního druhu α .
- ❖ Pravděpodobnost, že uděláme alespoň jednu chybu prvního druhu pak roste s počtem porovnávaných tříd.

Rostoucí chyba

- ❖ Pravděpodobnost chyby I. druhu při srovnávání typu „každý s každým“ - „Statistical fishing“.

	Hladina významnosti užívaná v t - testech		
Počet skupin (k)	0.10	0.05	0.01
2	0.10	0.05	0.01
3	0.27	0.14	0.03
5	0.65	0.40	0.10
10	0.99	0.90	0.36
20	1.00	1.00	0.85

- ❖ ANOVA zachovává výslednou hladinu významnosti α a rozumnou sílu testu.

Předpoklady testu ANOVA

- ❖ **Vyvážené třídění** – stejný rozsah jednotlivých výběrů, v praxi téměř nesplnitelný, čím vyváženější výběry jsou, tím věrohodnější jsou výsledky testu
- ❖ **Nezávislost výběrů** (pro porovnání $k > 2$ závislých výběrů používáme Friedmanův test)
- ❖ **Normalita rozdělení jednotlivých výběrů** (při výrazném porušení normality používáme Kruskal-Wallisův test)
- ❖ **Stejné rozptyly jednotlivých výběrů** = homoskedasticita (při porušení homoskedasticity používáme, stejně jako při porušení normality, Kruskal-Wallisův test)

Testy shody rozptylů - homoskedasticity

Předpoklad: $k > 2$ **nezávislých** výběrů z **normálního** rozdělení

Testujeme hypotézu $H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2$

Oproti alternativě $H_A : \overline{H_0}$ (alespoň jedna dvojice se liší)

Testy shody rozptylů (všechny součástí Statgraphicsu):

- ❖ Bartlettův test – citlivý na porušení předpokladu normality dat, v takovém případě potom volíme Levenův test
- ❖ Leveneův test – méně citlivý na porušení předpokladu normality, ale má menší sílu testu
- ❖ Hartleyův test
- ❖ Cochranův test

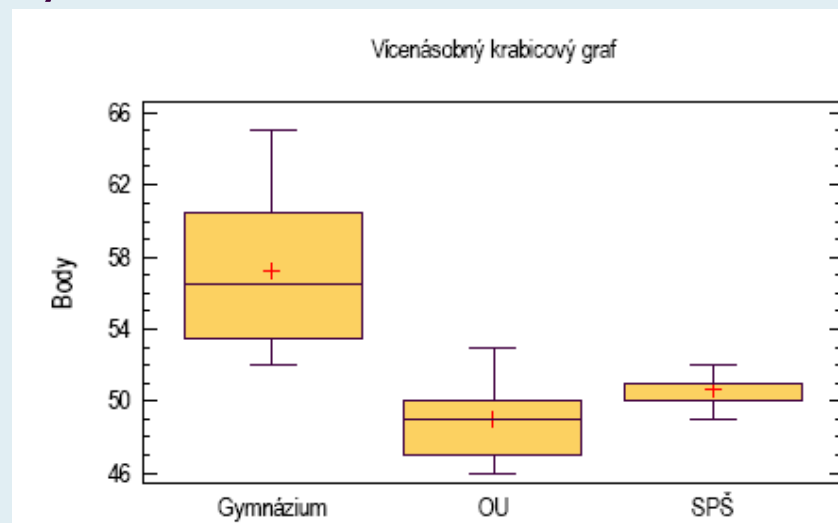
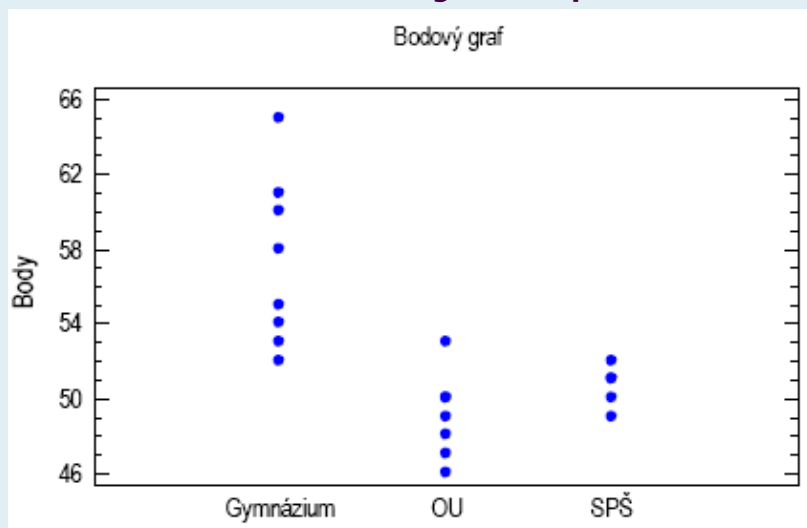
Příklad

Porovnejte úspěšnost absolventů gymnázii, SPŠ a odborných učilišť s maturitou (OU) u přijímací zkoušky z matematiky. Dosažené výsledky náhodně vybraných dvaceti studentů jsou uvedeny v následující tabulce.

Gymnázium	SPŠ	OU
55	52	47
54	50	53
58	51	49
61	51	50
52	49	46
60		48
53		50
65		

Příklad

Prvním krokem je explorační analýza dat:



	Gymnázium	OU	SPŠ
rozsah	8	7	5
průměr	57,3	49,0	50,6
výběrový rozptyl	20,5	5,3	1,3

Příčina rozdílných výsledků

❖ **Vliv sledovaného faktoru**

tj. rozdíly mezi kvalitou výuky na jednotlivých typech středních škol.

Projevuje se **rozdíly mezi třídami**.

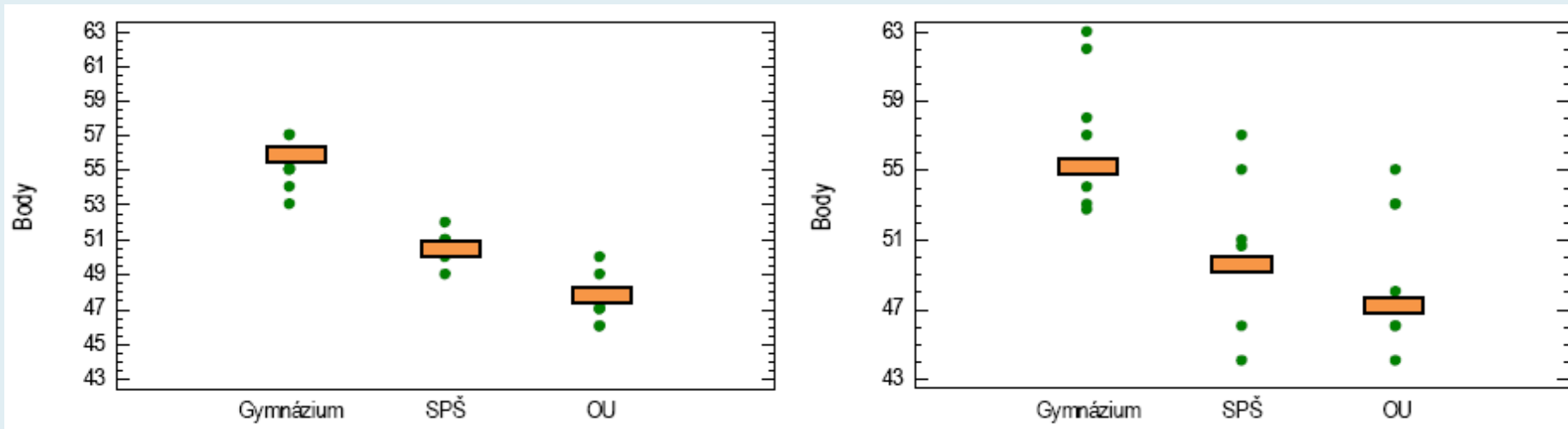
❖ **Reziduální (zbytkové) vlivy**

tj. rozdíly mezi školami v rámci tříd (není gymnázium jako gymnázium), rozdíly mezi pedagogy v rámci jedné školy, rozdíly mezi schopnostmi jednotlivých studentů, ...

Projevuje se **rozdíly uvnitř tříd**.

Analýza rozptylu

- ❖ Testuje jestli jsou průměry jednotlivých výběrů (tříd) rozdílné vlivem různých středních hodnot příslušných populací, nebo lze rozdíly mezi průměry přičíst na vrub náhodnému kolísání?



Srovnání datových souborů s nízkou a vysokou variabilitou uvnitř skupin

Jak tyto rozdíly kvantifikovat

- ❖ Rozdíly mezi třídami (vliv faktoru) kvantifikuje **mezitřídní variabilita** (součet čtverců mezi třídami) :

$$SS_B = \sum_{i=1}^k n_i (\bar{X}_i - \bar{X})^2$$

počet prvků v i-tém výběru

- ❖ Rozdíly uvnitř tříd (reziduální vlivy) kvantifikuje **vnitřní variabilita** (součet čtverců uvnitř tříd):

$$SS_W = \sum_{i=1}^k \sum_{j=1}^n (X_{ij} - \bar{X}_i)^2$$

Celková variabilita

- ❖ **Celková variabilita** (celkový součet čtverců) je definována jako **součet mezitřídní variability** a **vnitřní variability**.

$$SS_{TOTAL} = SS_W + SS_B \quad \Rightarrow \quad SS_{TOTAL} = \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2$$

Příklad – výpočet mezitřídní a vnitřní variability

	Skupina			
	Gymnázium 1	SPŠ 2	OU 3	
	55	52	47	
	54	50	53	
	58	51	49	
	61	51	50	
	52	49	46	
	60		48	
	53		50	
	65			
Rozsah	8	5	7	$n = 20$
Průměr $\bar{X}_i$	57,3	50,6	49,0	$\bar{\bar{X}} = \sum_{i=1}^k \sum_{j=1}^{n_i} \frac{X_{ij}}{n} = 52,7$
$(\bar{X}_i - \bar{\bar{X}})$	4,6	-2,1	-3,7	
$n_i(\bar{X}_i - \bar{\bar{X}})^2$	165,62	22,05	95,83	$\sum_{i=1}^k n_i(\bar{X}_i - \bar{\bar{X}})^2 = 283,5$

$$SS_B = \sum_{i=1}^k n_i (\bar{X}_i - \bar{\bar{X}})^2 = 8 \cdot (57,3 - 52,7)^2 + 5 \cdot (50,6 - 52,7)^2 + 5 \cdot (49 - 52,7)^2 = 283,5$$

$$SS_W = \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 = (55 - 57,3)^2 + (54 - 57,3)^2 + \dots + (52 - 50,6)^2 + \dots + (47 - 49)^2 + \dots + (50 - 49)^2 = 180,7$$

Java applet

<http://mi21.vsb.cz/modul/uvod-do-statistiky>

Všimněte si změn poměru mezitřídní a vnitřní variability při zachování průměrů a proměnném výběrovém rozptylu (DATA3 a DATA4).

Odhad společného rozptylu σ^2 za předpokladu platnosti H_0

- ❖ Odhad na základě mezitřídní variability (**rozptyl mezi třídami**, průměrný mezitřídní součet čtverců, vysvětlený rozptyl)

$$MS_B = \frac{SS_B}{DF_B} = \frac{\sum_{i=1}^k n_i (\bar{X}_i - \bar{X})^2}{k - 1}$$

počet tříd

- ❖ Odhad na základě vnitřní variability (**rozptyl uvnitř tříd**, průměrný součet čtverců uvnitř tříd, nevysvětlený rozptyl)

$$MS_W = \frac{SS_W}{DF_W} = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2}{N - k}$$

počet prvků ve výběrech (dohromady)

F-ratio (*F*-poměr)

Poměr dvou odhadů rozptylu (na základě výběrů z normálního rozdělení) má Fisher-Snedecorovo rozdělení.

$$F - ratio = \frac{MS_B}{MS_W} \rightarrow F(DF_B; DF_W)$$

Platí-li H_0 : MS_B je srovnatelné s MS_W , *F*-poměr se pohybuje kolem 1.

Platí-li H_A : MS_B je mnohem větší než MS_W , *F*-poměr je mnohem větší než 1.

ANOVA

Je možné, že výběry reprezentovány takto rozdílnými průměry pocházejí ze stejného rozdělení?

0. Předpoklady testu:

- ❖ nezávislé výběry
- ❖ normalita výběrů
- ❖ stejné rozptyly: $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2$ (homoskedasticita)

1. Formulace H_0 a H_A :

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$$

$$H_A : \overline{H_0}$$

ANOVA

2. Volba testové statistiky:

$$F - ratio = \frac{MS_B}{MS_W} \rightarrow F(DF_B(k - 1); DF_W(N - k))$$

3. Výpočet pozorované hodnoty testové statistiky

4. Výpočet p -value:

$$p - value = 1 - F(x_{OBS})$$

5. Formulace závěru

Tabulka ANOVA

Způsob prezentace výsledků ANOVY

Zdroj proměnlivosti	Součet čtverců	Stupně volnosti (DF)	Průměrný čtverec	Testová stat. F-poměr	P-value
Mezitřídní (faktor)	$SS_B = \sum_{i=1}^k n_i \cdot (\bar{X}_i - \bar{X})^2$	$k - 1$	$MS_B = \frac{SS_B}{k - 1}$		
Vnitřní (reziduální)	$SS_W = \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$	$N - k$	$MS_W = \frac{SS_W}{N - k}$	$F - ratio = \frac{MS_B}{MS_W}$	$1 - F(F - ratio)$
Celkový	$SS_{TOTAL} = \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2$	$N - 1$			

Příklad

0. Ověření předpokladů testu:

normalita každého výběru, nezávislost výběrů,
homoskedasticita

1. Formulace nulové a alternativní hypotézy:

$$H_0 : \mu_{GYM} = \mu_{SPŠ} = \mu_{OU}$$

$$H_A : \overline{H_0}$$

2. Volba testové statistiky, dopočítání neznámých:

$$F - ratio = \frac{MS_B}{MS_W} \rightarrow F(DF_B(k-1); DF_W(N-k))$$

$$MS_B = \frac{SS_B}{k-1} = \frac{283,5}{3-1} = 141,8 \quad MS_W = \frac{SS_W}{N-k} = \frac{180,7}{20-3} = 10,6$$

Příklad

3. Výpočet pozorované hodnoty testové statistiky:

$$F - ratio = x_{OBS} = \frac{141,8}{10,6} = 13,3 \rightarrow F(2;17)$$

4. Výpočet p -value:

$$p - value = 1 - F(13,3) = 0,0003$$

5. Formulace závěru:

$$p\text{-value} < 0,05$$

Na hladině významnosti 0,05 zamítáme nulovou hypotézu o shodě středních hodnot. Lze tedy tvrdit, že typ absolvované střední školy má vliv na výsledek přijímací zkoušky z matematiky.

Příklad

Tabulka ANOVA

Analysis of Variance					
Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Between groups	283,5	2	141,75	13,34	0,0003
Within groups	180,7	17	10,63		
Total (Corr.)	464,2	19			

Síla testu ANOVA

- ❖ Zvyšuje se se zvětšující se odchylkou od H_0 (to nelze ovlivnit)
- ❖ Zvyšuje se s počtem pozorování ve třídách
- ❖ Zvyšuje se s vyvážeností tříd
- ❖ Klesá s rostoucím počtem tříd

Post Hoc analýza

Metoda mnohonásobného porovnávání

- ❖ Vysoký F -poměr indikuje existenci významných změn mezi populačními výběrovými průměry a vede k **zamítnutí H_0** .
- ❖ V tomto případě **je nutné identifikovat, které z populací signalizují významnou odchylku průměru**.
- ❖ LSD metoda, Duncanův test, Tukeyův test pro významné rozdíly, Scheffého test a Bonferroniho test
- ❖ **POZOR!!!** Použijeme-li Post Hoc analýzu neoprávněně (v případě nezamítnutí H_0), můžeme získat informaci o **falešně významných** rozdílech mezi průměry.

1. Následující příklad je ukázkou klinické studie. Dvacet dva pacientů, kteří podstoupili operaci srdce, bylo náhodně rozděleno do tří skupin.

Skupina 1: Pacienti dostali 50 % oxidu dusného a 50 % kyslíkové směsi nepřetržitě po dobu 24 hodin;

Skupina 2: Pacienti dostali 50 % oxidu dusného a 50 % kyslíkové směsi pouze během operace;

Skupina 3: Pacienti nedostali žádný oxid dusný, ale dostali 35-50 % kyslíku po dobu 24 hodin.

Tabulka ukazuje koncentraci soli kyseliny listové v červených krvinkách ve všech třech skupinách po uplynutí 24 hodin ventilace.

Zjistěte, zda složení a způsob dané medikace má vliv na koncentraci soli kyseliny listové v červených krvinkách po uplynutí 24 hodin ventilace. Pro řešení ve Statgraphicsu použijte soubor Kys_listova.sf3.

2. Je třeba zjistit, zda se liší spotřeba automobilu při použití různých druhů benzínu. Zkouší se čtyři typy benzínu, jež se liší chemickým složením. Testovací jízdy se provádějí se 16 auty stejného modelu tak, že vždy čtyři auta použijí stejný benzín.

Výsledky měření spotřeby v l/100 km při jednotlivých jízdách jsou uloženy v datovém souboru [Spotreba.sf3](#). Rozhodněte pomoci testu, zda složení benzínu ovlivňuje jeho spotřebu ($\alpha = 0,05$).

Jak postupovat při nesplnění předpokladů?

❖ **Porušení homoskedasticity:**

Pokusíme se stabilizovat rozptyl pomocí transformací proměnných (není obsahem Statistiky I.). Pokud se nám rozptyl stabilizovat nepodaří, můžeme použít neparametrickou obdobu ANOVY – Kruskal – Wallisův test (vícevýběrový test o shodě mediánů).

❖ **Porušení normality:**

Pokud je splněna podmínka homoskedasticity, můžeme opět použít Kruskal – Wallisův test.

Kruskal – Wallisův test

1. Formulace H_0 a H_A :

$$H_0: x_{0,5_I} = x_{0,5_{II}} = \dots = x_{0,5_k}$$

H_A : neplatí H_0

2. Volba testové statistiky:

$$Q = \frac{12}{N \cdot (N + 1)} \cdot \sum_{i=1}^k \frac{T_i^2}{n_i} - 3 \cdot (N + 1) \rightarrow \chi_{k-1}^2$$

($N = \sum_{i=1}^k n_i$, T_i jsou součty pořadí pro jednotlivé výběry)

3. Výpočet pozorované testové statistiky

4. Výpočet p-value:

$$p - value = 1 - F(x_{OBS})$$

5. Formulace závěru

Příklad

Výběr			
I	II	III	IV
67	20	106	13
22	-13	127	49
10	11	13	97
55	5	79	85
94	38	37	46
-17	53	31	31
37	5	22	37
28		70	61
		76	10
		55	1
		91	
		25	

	Výběr			
	I	II	III	IV
	28	11	36	9,5
	12,5	2	37	23
	6,5	8	9,5	35
	25,5	4,5	31	32
	34	21	19	22
	1	24	16,5	16,5
	19	4,5	12,5	19
	15		29	27
			30	6,5
			25,5	3
			33	
			14	
Rozsah výběru n_i	8	7	12	10
Součty pořadí T_i	145	75	293	193,5
T_i^2	20022,3	5625,0	85849,0	37442,3
T_i^2/n_i	2502,8	803,6	7154,1	3744,2

$$Q = \frac{12}{N \cdot (N + 1)} \cdot \sum_{i=1}^k \frac{T_i^2}{n_i} - 3 \cdot (N + 1) = \frac{12}{37 \cdot (37 + 1)} \cdot (2502,8 + 803,6 + 7154,1 + 3744,2) - 3 \cdot (37 + 1) = 7,24$$

$$p - value = 1 - F(7,24) = 0,0645$$

3. Příklad pedagogického výzkumu: Zjistěte, zda používání elektronických stavebnic má pozitivní vliv na vytváření a rozvoj žákových vědomostí a dovedností. Pro ověření tohoto výzkumu byly získány údaje o bodovém hodnocení studentů SŠ při závěrečné zkoušce z Elektrotechniky. Studenti byli rozděleni do tří skupin – skupina A - zahrnovala studenty, kteří při výuce používali stavebnici ZEM Elektronik, skupina B – používala stavebnici pro technické práce a základy techniky pro 8. třídy, skupina C při výuce žádnou stavebnici nepoužívala. Dosažené výsledky jsou zaznamenány v následující tabulce. (pro řešení použijte Kruskal-Wallisův test).

A	6,4	6,8	7,2	8,3	8,4	9,1	9,4	9,7
B	2,5	3,7	4,9	5,4	5,9	8,1	8,2	
C	1,3	4,1	4,9	5,2	5,5	8,2		

Pro řešení ve Statgraphicsu použijte soubor Stavebnice.sf3

Test

1. Ke statistické úloze přiřadte vhodný test.

1. Ověřte, zda je průměrná výška dospělé populace v ČR větší než 170 cm (rozsah výběru je 120, byla ověřena normalita výběru).

a) Dvouvýběrový t test

b) Friedmanův test

c) Jednovýběrový t test

d) Jednovýběrový Wilcoxonův test

e) Test o parametru alternativního rozdělení

f) Test o rozptylu normálního rozdělení

g) Párový t test

1. Ke statistické úloze přiřadte vhodný test.

7. Bylo testováno 11 automobilů určité značky. Ověřte, zda se jejich pravé a levé přední pneumatiky ojíždějí srovnatelně. (Předpokládejte, že ojetí pneumatik [mm] má normální rozdělení.)

- | | |
|--|-----|
| a) Dvouvýběrový t test | = 6 |
| b) Friedmanův test | = 4 |
| c) Jednovýběrový t test | = 1 |
| d) Jednovýběrový Wilcoxonův test | = 2 |
| e) Test o parametru alternativního rozdělení | = 3 |
| f) Test o rozptylu normálního rozdělení | = 5 |
| g) Párový t test | = 7 |

2. Rozhodněte o pravdivosti následujících výroků:

- a) Při neparametrickém testu homogenity dvou binomických rozdělení nemusíme ověřovat žádné předpoklady o výběrech.
- b) Mannův-Whitneyův test se používá pro ověření shody úrovně ve dvou závislých výběrech.
- c) Každý test hypotézy $H_0: \mu_0 = \mu_1$, tj. hypotézy o shodě dvou středních hodnot je testem párovým.
- d) Jedním z předpokladů analýzy rozptylu je alespoň přibližná shoda rozptylů v jednotlivých skupinách.

2. Rozhodněte o pravdivosti následujících výroků:

-  a) Při neparametrickém testu homogenity dvou binomických rozdělení nemusíme ověřovat žádné předpoklady o výběrech.
-  b) Mannův-Whitneyův test se používá pro ověření shody úrovně ve dvou závislých výběrech.
-  c) Každý test hypotézy $H_0: \mu_0 = \mu_1$, tj. hypotézy o shodě dvou středních hodnot je testem párovým.
-  d) Jedním z předpokladů analýzy rozptylu je alespoň přibližná shoda rozptylů v jednotlivých skupinách.

2. Rozhodněte o pravdivosti následujících výroků:

- e) Reziduální rozptyl (v analýze rozptylu) lze určit jako aritmetický průměr rozptylů v jednotlivých skupinách.
- f) Post hoc analýza znamená, že stanovíme nejprve hypotézy H_0 , H_A , a „následně“ provedeme řešení.
- g) Kruskalův-Wallisův test se nazývá rovněž neparametrická ANOVA.

2. Rozhodněte o pravdivosti následujících výroků:

-  e) Reziduální rozptyl (v analýze rozptylu) lze určit jako aritmetický průměr rozptylů v jednotlivých skupinách.
-  f) Post hoc analýza znamená, že stanovíme nejprve hypotézy H_0 , H_A , a „následně“ provedeme řešení.
-  g) Kruskalův-Wallisův test se nazývá rovněž neparametrická ANOVA.

3. Doplňte:

- a) Test o shodě středních hodnot dvou populací může být oboustranný nebo
- b) Neparametrický test, při kterém srovnáváme úroveň dvou závislých (spárovaných) souborů se nazývá
- c) Parametrický test, při kterém srovnáváme střední hodnoty dvou souborů o stejném, avšak neznámém rozptylu se nazývá
- d) Hartleyův test homoskedasticity lze použít pouze v případě třídění.

3. Doplňte:

- a) Test o shodě středních hodnot dvou populací může být oboustranný nebo ..jednostranný.....
- b) Neparametrický test, při kterém srovnáváme úroveň dvou závislých (spárovaných) souborů se nazývá ..párový Wilcoxonův nebo znaménkový test.....
- c) Parametrický test, při kterém srovnáváme střední hodnoty dvou souborů o stejném, avšak neznámém rozptylu se nazývádvouvýběrový t test.....
- d) Hartleyův test homoskedasticity lze použít pouze v případě ..vyváženého..... třídění.