

Биостатистика и анализа података

Марија Цупарић

Дванаести час



- ★ Желимо да тестирамо једнакост више од две средње вредности
- ★ Имамо три или више популација или је популација подељена на три или више подпопулација
- ★ У питању је уопштење t —теста
- ★ Разматраћемо

једнофакторску дисперзиону анализу - уопштење независног t —теста

рандомизирани комплетни блок дизајн - уопштење спареног t —теста



- ★ Сам назив је повезан са тим да постоји само једна променљива која се директно проучава
- ★ Сваки елемент је класификован у тачно једну категорију те променљиве
- ★ Разликујемо две ситуације:
 - (1) Имамо једну популацију и интересује нас ефекат k различитих третмана. Случајан узорак обима n делимо на k подузорака обима $n_1, n_2, \dots, n_k$. Свака група добија различит третман. Тестирамо нулту хипотезу да су ефекти свих третмана једнаки, док је алтернативна да постоји бар нека разлика.
 - (2) Имамо k популација на којима проучавамо исто обележје. Извлаче се узорци обима $n_1, n_2, \dots, n_k$. Свакој групи даје се исти третман. Тестирамо нулту хипотезу да су ефекти третмана исти у свим популацијама, док је алтернативна хипотеза да постоји бар нека разлика.
- ★ У оба случаја тестирамо $H_0 : m_1 = m_2 = \dots = m_k$ против алтернативе $H_1 : m_i \neq m_j$ за неко i и j , где је m_i средња вредност за i -ти третман односно i -ту популацију



Пример

- (1) Разматра се једна популација биљака исте врсте. Испитује се утицај различитих метода ђубрења на просечну висину биљака. Третмани:

- T_1 стандардно ђубрење
- T_2 органско ђубрење
- T_3 минерално ђубрење
- T_4 комбиновано ђубрење

Случајан узорак од n биљака дели се у $k = 4$ група, при чему свака група добија другачији третман. Након одређеног периода мери се висина биљака.

$$H_0 : m_1 = m_2 = m_3 = m_4$$

H_1 : бар једна просечна висина се разликује

- (2) Разматрамо 4 различите популације биљака исте врсте које расту на различитим стаништима, а проучавамо исто обележје: висину биљака. Са сваког станишта узима се узорак обима n_1, n_2, n_3, n_4 . Све биљке се посматрају у истом периоду развоја и мере под истим условима.

$$H_0 : m_1 = m_2 = m_3 = m_4$$

H_1 : постоји разлика између бар два станишта

Користи се следећа нотација:

X_{ij} j -ти елемент у i -тој групи

n_i број елемената у i -тој групи

$T_{i\bullet} = \sum_{j=1}^{n_i} X_{ij}$ збир елемената у i -тој групи

$\bar{X}_{i\bullet} = \frac{T_{i\bullet}}{n_i}$ средња вредност елемената у i -тој групи

$T_{\bullet\bullet} = \sum_{i=1}^k \sum_{j=1}^{n_i} X_{ij} = \sum_{i=1}^k T_{i\bullet}$ збир свих елемената узорка

$\bar{X}_{\bullet\bullet} = \frac{T_{\bullet\bullet}}{n}$ средња вредност свих елемената узорка



Дефиниција: Модел једнофакторске дисперзионе анализе је

$$X_{ij} = m + (m_i - m) + (X_{ij} - m_i),$$

где је

- ★ m средња вредност свих популација (целе популације)
- ★ $m_i - m$ ефекат i -те групе (i -тог третмана)
- ★ $X_{ij} - m_i$ случајно (индивидуално) одступање у оквиру i -те групе (третмана)

Претпоставке модела:

- ★ k узорака из k група међусобно су независни
- ★ унутар сваке групе случајна променљива која се проучава има нормалну расподелу са средњом вредношћу m_i и истом дисперзијом σ^2 .



Тачкастим оцењивањем непознатих параметара модела добија се

$$X_{ij} - \bar{X}_{..} = (\bar{X}_{i.} - \bar{X}_{..}) + (X_{ij} - \bar{X}_{i.})$$

где је $\bar{X}_{i.}$ средња вредност узорка из i -те групе, док је $\bar{X}_{..}$ средња вредност свих елемената свих узорака.

Сабирањем квадрата ових једначина за свако X_{ij} добија се

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{..})^2 = \sum_{i=1}^k n_i (\bar{X}_{i.} - \bar{X}_{..})^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i.})^2$$

укупно одступање свих елемената узорка од заједничке средње вредности (укупна варијабилност целог узорка)

=

одступање средњих вредности група од заједничке средње вредности (укупно одступање међу групама) - то је оно што испитујемо да ли постоји

+

одступање елемената узорка од средње вредности своје групе (индивидуално одступање унутар група) - случајна грешка

$$SST = SSG + SSE$$



Дефиниција: Средње одступање по групама MSG и средње одступање унутар група MSE рачунају се као

$$MSG = \frac{SSG}{k-1}, \quad MSE = \frac{SSE}{n-k}.$$

Тест статистика је

$$F_0 = \frac{MSG}{MSE},$$

која има Фишерову расподелу с параметрима $\nu_1 = k-1$ и $\nu_2 = n-k$.
 p - вредност теста је површина Фишерове $F_{k-1, n-k}$ расподеле десно од вредности f_0 коју је тест статистика узела у узорку.

Рачунске формуле

$$SST = \sum_{i=1}^k \sum_{j=1}^{n_i} x_{ij}^2 - \frac{T_{\bullet\bullet}^2}{n}$$

$$SSG = \sum_{i=1}^k \frac{T_{i\bullet}^2}{n_i} - \frac{T_{\bullet\bullet}^2}{n}$$

$$SSE = SST - SSG$$



ANOVA табела

Извор варијације	Степени слободе	Сума квадрата	Средње одступање	F -статистика
Третман	$k - 1$	SSG	$MSG = \frac{SSG}{k - 1}$	$\frac{MSG}{MSE}$
Грешка	$n - k$	SSE	$MSE = \frac{SSE}{n - k}$	
Укупно	$n - 1$	SST		



Пример Социолог испитује утицај броја деце у породици на самосталност особе, на популацији бруцоша једног универзитета. Популација је подељена на четири групе, породице с једним, двоје, троје или више од троје деце. Узети су узорци обима 15, 15, 14 и 13 и сваком је дато да попуни анкетни лист на основу чијег је резултата процењена самосталност особе. Тестирамо нулту хипотезу $H_0 : m_1 = m_2 = m_3 = m_4$. Добијени подаци су:

Број деце у породици			
једно	двоје	троје	више од 3
59.1	61.2	73.4	73.1
84.4	71.0	69.3	95.7
76.0	46.6	64.9	91.1
59.5	54.0	48.7	49.7
60.1	66.6	67.7	94.9
73.4	56.6	72.5	65.8
64.1	70.5	68.8	75.8
69.4	72.8	79.9	77.2
56.4	58.5	77.7	86.2
67.1	48.7	79.2	61.1
97.6	63.3	56.7	83.1
58.5	74.8	60.1	95.6
70.7	53.1	69.8	83.8
51.8	69.9	58.2	
53.2	65.5		
$T_{1\bullet} = 1001.3$	$T_{2\bullet} = 933.1$	$T_{3\bullet} = 946.9$	$T_{4\bullet} = 1033.1$
$T_{\bullet\bullet} = 3914.4$			

$$\sum_{i=1}^4 \sum_{j=1}^{n_i} x_{ij}^2 = 59.1^2 + 84.4^2 + \dots + 83.8^2 = 277845.9$$

$$\frac{T_{..}^2}{n} = \frac{3914.4^2}{57} = 268816.27$$

$$SST = \sum_{i=1}^4 \sum_{j=1}^{n_i} x_{ij}^2 - \frac{T_{..}^2}{n} = 9029.63$$

$$\sum_{i=1}^4 \frac{T_{i.}^2}{n_i} = \frac{1001.3^2}{15} + \frac{933.1^2}{15} + \frac{946.9}{14} + \frac{1033.1}{13} = 271029.07$$

$$SSG = \sum_{i=1}^4 \frac{T_{i.}^2}{n_i} - \frac{T_{..}^2}{n} = 2212.80, \quad SSE = SST - SSG = 6816.83$$

$$MSG = \frac{SSG}{k-1} = 737.60, \quad MSE = \frac{SSE}{n-k} = \frac{6816.83}{57-4} = 128.62$$

$$f_0 = \frac{MSG}{MSE} = 5.73$$



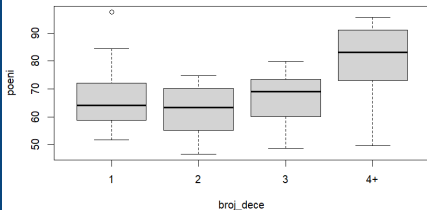
Извор варијације	Степени слободе	Сума квадрата	Средње одступање	F -статистика
Третман	3	2212.80	737.60	5.73
Грешка	54	6816.83	128.62	
Укупно	56	9029.63		

Користећи $F_{3,53}$ расподелу, добија се да је p -вредност теста 0.002 ($1 - pf(5.73, 3, 53)$), па закључујемо да постоји утицај броја деце у породици на самосталност особе.





```
samostalnost$broj_dece <- factor( samostalnost$broj_dece, levels =  
  c("jedno", "dvoje", "troje", "vise"), labels = c("1", "2", "3",  
  "4+") )  
plot(poeni~broj_dece,samostalnost)
```



```
jda.samostalnost<-aov(poeni~broj_dece,samostalnost)  
summary(jda.samostalnost)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
broj_dece	3	2213	737.6	5.735	0.00179 **
Residuals	53	6817	128.6		

 Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Једнофакторска дисперзиона анализа се користи да се упореде средње вредности k популација. Након анализе налазимо се у тачно једној од следеће две ситуације:

немамо доказа да одбацимо H_0 : Нисмо успели да нађемо никакву разлику међу k популација. Анализа се ту завршава.

одбацили смо H_0 : Постоји бар нека разлика међу средњим вредностима, али још увек немамо одговор међу којим групама постоји разлика. Анализа није завршена, потребно је накнадно упоређивање.

- ★ Постоји више метода накнадног упоређивања: овде ћемо размотрити Бонферонијеве T —тестове и Тјукијев тест поштених значајних разлика



- ★ Ако разматрамо k група, онда имамо $\binom{k}{2} = \frac{k(k-1)}{2}$ могућих парова средњих вредности које треба упоредити
- ★ Бонферонијев метод извршава $\frac{k(k-1)}{2}$ могућих тестова за тестирање $H_0 : m_i = m_j$ против $H_1 : m_i \neq m_j$ (двострани стандардни t —тестови)
- ★ Тест статистика за упоређивање m_i и m_j је

$$T_0 = \frac{\bar{X}_{i\bullet} - \bar{X}_{j\bullet}}{\sqrt{MSE(\frac{1}{n_i} + \frac{1}{n_j})}}$$

која, при H_0 , има Студентову расподелу са $n - k$ степени слободе

- ★ Пошто истовремено радимо $\frac{k(k-1)}{2}$ тестова, p —вредност сваког од њих мора бити мања од $\frac{2\alpha}{k(k-1)}$
- ★ Није увек обавезно обавити свих $\frac{k(k-1)}{2}$ тестова, већ истраживач може изабрати оне где сматра да ће открити разлику



Пример Проучава се нови полимер за уклањање токсичног отпада из воде. Експеримент је извршен на пет различитих температура. Резултати представљају проценат нечистоћа уклоњених третманом. Добијени су подаци:

температура				
I	II	III	IV	V
40	36	49	47	55
45	42	51	49	60
42	38	53	51	62
48	39	53	52	63
50	37	52	50	59
51	40	50	51	61

Прво размотримо *ANOVA* табелу са потребним вредностима

Узрок дисперзије	Број степени слободе	Сума квадрата	Средње одступање	Тест статистика
Третман	4	1458.13	364.53	47.78
Грешка	25	190.67	7.63	
Тотал	29	1648.80		

Како је p —вредност мала, одбацујемо нулту хипотезу о непостојању утицаја различитих температура.

Желимо сада, на нивоу $\alpha = 0.1$ да откријемо за које температуре је значајно различита количина избачених материја. Пошто имамо $\frac{5 \cdot 4}{2} = 10$ комбинација, треба да обавимо 10 тестова на нивоу 0.01. На пример, ако тестирамо $m_1 = m_2$ против $m_1 \neq m_2$ имамо да је

$$t_0 = \frac{|46 - 38.7|}{\sqrt{7.63(\frac{1}{6} + \frac{1}{6})}} = 4.58.$$

Користећи Студентову расподелу са $30 - 5 = 25$ степени слободе, добија се да је p -вредност мања од 0.01. Закључујемо да постоји разлика у количини одбачених материја при температурама I и II.

Слично се тестови обављају и у осталим случајевима.



```
temperatura_df <- data.frame( temperatura = factor(rep(c("I", "II",  
"III", "IV", "V"), each = 6)), vrednost = c( 40, 45, 42, 48, 50, 51,  
36, 42, 38, 39, 37, 40, 49, 51, 53, 53, 52, 50, 47, 49, 51, 52, 50, 51,  
55, 60, 62, 63, 59, 61 ) )  
pairwise.t.test(temperatura_df$vrednost, temperatura_df$temperatura)
```

Pairwise comparisons using t tests with pooled SD

data: temperatura_df\$vrednost and temperatura_df\$temperatura

	I	II	III	IV
II	0.00042	-	-	-
III	0.00780	2.1e-07	-	-
IV	0.03794	1.3e-06	0.41094	-
V	3.7e-08	6.7e-12	6.1e-05	8.8e-06

P value adjustment method: holm

Вредности у табели представљају p -вредности тестова, тако да одатле можемо да закључимо где постоје разлике у количини одбачених материја.

Овај тест је својеврсно побољшање Бонферијевих тестова. Тјукијева критика састоји се у томе да накнадна упоређивања нису међусобно независна и да се дисперзија оцењује на основу свих група, што мења расподелу тест статистике. Тест статистика за упоређивање m_i и m_j је иста као код Бонферијевог теста,

$$Q_0 = \frac{\bar{X}_{i\bullet} - \bar{X}_{j\bullet}}{\sqrt{MSE(\frac{1}{n_i} + \frac{1}{n_j})}}$$

а разлика је у томе што је сада њена расподела под нултом хипотезом такозвана Тјукијева расподела. Често се резултат теста приказује преко интервала поверења

$$\bar{X}_{i\bullet} - \bar{X}_{j\bullet} \pm q_{k,n-k,1-\alpha} \sqrt{MSE \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$$

и интервали који не садрже нулу представљају значајне разлике између одговарајућих група.



Пример Вратимо се на базу података са температурама и применимо Тјукијев тест.

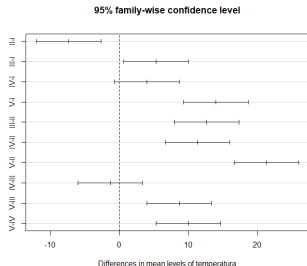


```
jda.temperatura<-aov(vrednost~temperatura, temperatura_df)
TukeyHSD(jda.temperatura)
plot(TukeyHSD(jda.temperatura))
```

Tukey multiple comparisons of means
95% family-wise confidence level

Fit: aov(formula = vrednost ~ temperatura, data = temperatura_df)

Stemperatura	diff	lwr	upr	p adj
II-I	-7.333333	-12.0159829	-2.650684	0.0009176
III-I	5.333333	0.6506838	10.015983	0.0199661
IV-I	4.000000	-0.6826496	8.682650	0.1206577
V-I	14.000000	9.3173504	18.682650	0.0000000
III-II	12.666667	7.9840171	17.349316	0.0000003
IV-II	11.333333	6.6506838	16.015983	0.0000018
V-II	21.333333	16.6506838	26.015983	0.0000000
IV-III	-1.333333	-6.0159829	3.349316	0.9167059
V-III	8.666667	3.9840171	13.349316	0.0001099
V-IV	10.000000	5.3173504	14.682650	0.0000135



Закључци су исти као и раније.

Пример Применимо Бонфереонијев тест на базу самосталност.



```
pairwise.t.test(samostalnost$poeni, samostalnost$broj_dece)
```

Pairwise comparisons using t tests with pooled SD

data: samostalnost\$poeni and samostalnost\$broj_dece

	1	2	3
2	0.6098	-	-
3	0.8350	0.6098	-
4+	0.0230	0.0011	0.0363

P value adjustment method: holm

Закључак је да постоји значајна разлика између породица с двоје и породица с више од четворо деце, док се за остала упоређивања не може рећи да постоје значајне разлике.

Сада применимо и Тјукијев тест.



```
TukeyHSD(jda.samostalnost)
```

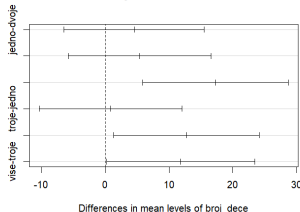
Tukey multiple comparisons of means
95% family-wise confidence level

Fit: aov(formula = poeni ~ broj_dece, data = samostalnost)

\$broj_dece

	diff	lwr	upr	p adj
jedno-dvoje	4.546667	-6.437592	15.53093	0.6923447
troje-dvoje	5.429048	-5.749638	16.60773	0.5744046
visе-dvoje	17.262564	5.863660	28.66147	0.0010458
troje-jedno	0.882381	-10.296304	12.06107	0.9967146
visе-jedno	12.715897	1.316994	24.11480	0.0231420
visе-troje	11.833516	0.247142	23.41989	0.0435942

95% family-wise confidence level



- ★ Поред утицаја различитих третмана на групе, може се јавити и неки други утицај (променљива која се не проучава директно)
- ★ Ова променљива може имати утицаја на закључивање у вези са стварном разликом у срединама услед третмана

Пример. Желимо да испитамо постоји ли разлика у четири врсте асфалта за асфалтирање аутопута. Променљива коју меримо је степен истрошености након годину дана. Међутим, поред квалитета асфалта на истрошеност утичу и други фактори, као што су фреквентност саобраћаја и временске прилике. Зато одређујемо три различита места (блокове) на којима постављамо четири различите врста асфалта (групе).

Дакле, на основу тог утицаја делимо узорак на блокове тако да у оквиру сваке групе по један елемент узорка припада сваком од блокова. Третмани су онда случајно додељени у оквиру блока.

Ово је заправо уопштење спареног t —теста

блок	група					
	1	2	3	...	k	
1	X_{11}	X_{21}	X_{31}	...	X_{k1}	$T_{\bullet 1}$
2	X_{12}	X_{22}	X_{32}	...	X_{k2}	$T_{\bullet 2}$
3	X_{13}	X_{23}	X_{33}	...	X_{k3}	$T_{\bullet 3}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
b	X_{1b}	X_{2b}	X_{3b}	...	X_{kb}	$T_{\bullet b}$
	$T_{1\bullet}$	$T_{2\bullet}$	$T_{3\bullet}$	...	$T_{k\bullet}$	$T_{\bullet\bullet}$

X_{ij} елемент у i -тој групи и j -том блоку

k број група (и број елемената у оквиру једног блока)

b број блокова (и број елемената у оквиру једне групе)

kb укупан број елемената

$$T_{i\bullet} = \sum_{j=1}^b X_{ij} \text{ збир елемената у } i\text{-тој групи}$$

$$T_{\bullet j} = \sum_{i=1}^k X_{ij} \text{ збир елемената у } j\text{-том блоку}$$

$$T_{\bullet\bullet} = \sum_{i=1}^k \sum_{j=1}^b X_{ij} \text{ збир свих елемената узорка}$$

Дефиниција: Модел је

$$X_{ij} = m + (m_i - m) + (m_{\bullet j} - m) + (X_{ij} - m_i - m_{\bullet j} + m),$$

где је

- ★ m средња вредност свих популација (целе популације)
- ★ $m_i - m$ ефекат i -те групе (i -тог третмана)
- ★ $m_{\bullet j}$ ефекат j -тог блока
- ★ $X_{ij} - m_i$ случајно (индивидуално) одступање у оквиру i -те групе (третмана)

Коришећем тачкастих оцена параметара модела добија се

$$SST = SSG + SSB + SSE,$$

укупно одступање
свих елемената
узорка од заједничке
средње вредности
(укупна
варијабилност целог
узорка)

=

одступање средњих
вредности група од
заједничке средње
вредности (укупно
одступање међу
групама)

+

одступање
под
утицајем
различитих
блокова

+

случајно
индивиду-
ално
одступање
грешка

Дефиниција: Средње одступање по групама MSG , средња одступања због блокова MSB и средње одступање унутар група MSE рачунају се као

$$MSG = \frac{SSG}{k-1}, \quad MSB = \frac{SSB}{b-1}, \quad MSE = \frac{SSE}{n-k}.$$

Тест статистика је

$$F_0 = \frac{MSG}{MSE},$$

која има Фишерову расподелу с параметрима $\nu_1 = k-1$ и $\nu_2 = (k-1)(b-1)$. p - вредност теста је површина Фишерове $F_{k-1, (k-1)(b-1)}$ расподеле десно од вредности f_0 коју је тест статистика узела у узорку.

Рачунске формуле

$$SST = \sum_{i=1}^k \sum_{j=1}^b X_{ij}^2 - \frac{T_{\bullet\bullet}^2}{kb}, \quad SSG = \sum_{i=1}^k \frac{T_{i\bullet}^2}{n_i} - \frac{T_{\bullet\bullet}^2}{kb}$$

$$SSB = \sum_{j=1}^b \frac{T_{\bullet j}^2}{k} - \frac{T_{\bullet\bullet}^2}{kb}, \quad SSE = SST - SSG - SSB$$



Блок дизајн треба користити ако је утицај блокова већи од случајног утицаја, тј. $MSB > MSE$, иначе је боље користити једнофакторску дисперзиону анализу.

ANOVA табела

Извор варијације	Степени слободe	Сума квадрата	Средње одступање	F-статистика
Третман	$k - 1$	SSG	$MSG = \frac{SSG}{k - 1}$	$\frac{MSG}{MSE}$
Блок	$b - 1$	SSB	$MSB = \frac{SSB}{b - 1}$	
Грешка	$(k - 1)(b - 1)$	SSE	$MSE = \frac{SSE}{(k - 1)(b - 1)}$	
Укупно	$kb - 1$	SST		

Пример (наставак). Желимо да упоредимо четири врсте асфалта на основу података из табеле.

блок	врста асфалта				
	1	2	3	4	
1	42.7	39.3	48.5	32.8	$T_{\bullet 1} = 163.3$
2	50.0	38.0	49.7	40.2	$T_{\bullet 2} = 177.9$
3	51.9	46.3	53.5	51.1	$T_{\bullet 3} = 202.8$
	$T_{1\bullet} = 144.6$	$T_{2\bullet} = 123.6$	$T_{3\bullet} = 151.7$	$T_{4\bullet} = 124.1$	$T_{\bullet\bullet} = 544.0$

Тестирамо хипотезу

$$H_0 : m_1 = m_2 = m_3 = m_4.$$

Одредимо потребне вредности

$$\sum_{j=1}^4 \sum_{i=1}^3 x_{ij}^2 = (42.7)^2 + (50.0)^2 + (51.9)^2 + \dots + (51.1)^2 = 25136.76$$

$$\frac{T_{\bullet\bullet}^2}{kb} = \frac{(544.0)^2}{12} = 24661.33$$

$$\sum_{i=1}^4 \frac{T_{i\bullet}^2}{b} = \frac{T_{1\bullet}^2}{3} + \frac{T_{2\bullet}^2}{3} + \frac{T_{3\bullet}^2}{3} + \frac{T_{4\bullet}^2}{3} = 24866.61$$

$$\sum_{j=1}^3 \frac{T_{\bullet j}^2}{k} = \frac{T_{\bullet 1}^2}{4} + \frac{T_{\bullet 2}^2}{4} + \frac{T_{\bullet 3}^2}{4} = 24860.79.$$

$$SSG = \sum_{i=1}^4 \frac{T_{i\bullet}^2}{b} - \frac{T_{\bullet\bullet}^2}{kb} = 205.28.$$

$$SSB = \sum_{j=1}^3 \frac{T_{\bullet j}^2}{k} - \frac{T_{\bullet\bullet}^2}{kb} = 199.46.$$

$$SST = \sum_{j=1}^3 \sum_{i=1}^4 x_{ij}^2 - \frac{T_{\bullet\bullet}^2}{kb} = 475.43.$$

$$SSE = SST - SSG - SSB = 70.69.$$

Извор варијације	Степени слободe	Сума квадрата	Средње одступање	F-статистика
Третман	3	205.28	68.43	5.81
Блок	2	199.46	99.73	
Грешка	6	70.69	11.78	
Укупно	11	475.43		

Како је $MSB > MSE$, блок дизајн је добар избор модела.

Тест статистика узима вредност $f_0 = 5.81$. p -вредност теста је 0.03. Закључујемо да постоји разлика међу различитим врстама асфалта.



```
asfalt <- data.frame(ostecenost=c(42.7,39.3,48.5,32.8,50.0, 38.0,
49.7, 40.2, 51.9,46.3,53.5,51.1),
  blok=c("1","1","1","1","2","2","2","2","3","3","3","3"),
  vrsta=c("1","2","3","4","1","2","3","4","1","2","3","4"))
bd.asfalt <- aov(ostecenost ~ vrsta + blok, asfalt)
summary(bd.asfalt)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
vrsta	3	205.3	68.42	5.807	0.0330 *
blok	2	199.4	99.73	8.463	0.0179 *
Residuals	6	70.7	11.78		

 Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

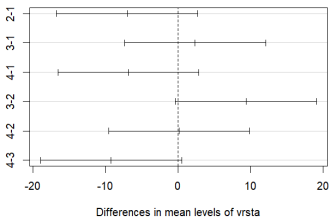


Као и у случају једнофакторске дисперзионе анализе, и код комплетног блок дизајна се може извезти Бонферонијев метод (са адекватно изабраним α) с тим што се у овом случају користи упарени t -тест. Међутим, овај метод је применљив само када је k мало (у супротном би α морало да буде сувише мало а самим тим тест би имао малу моћ).



```
plot(TukeyHSD(bd.asfalt, "vrsta"))
```

95% family-wise confidence level



Како је дисперзиона анализа повезана са линеарном регресијом?

Дисперзиона анализа је специјалан случај линеарног регресионог модела.

Ако имамо једну категоричку променљиву са k нивоа, у регресији се моделује помоћу индикаторских променљивих:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_{k-1} x_{i(k-1)} + \varepsilon_i.$$

Тестирање $H_0 : \beta_1 = \dots = \beta_{k-1} = 0$ је исто што и тестирање $H_0 : m_1 = \dots = m_k$ у дисперзионој анализи, па су и F тестови исти. Додавањем блокова добиће се вишеструка линеарна регресија.

Дисперзиона анализа се фокусира на разлагање дисперзије, док се регресија фокусира на прављење модела.





```
lm.samostalnost<-lm(poeni~broj_dece,samostalnost)
summary(lm.samostalnost)
```

Call:

```
lm(formula = poeni ~ broj_dece, data = samostalnost)
```

Residuals:

Min	1Q	Median	3Q	Max
-29.769	-7.653	1.093	7.693	30.847

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	62.207	2.928	21.244	< 2e-16 ***
broj_decejedno	4.547	4.141	1.098	0.277203
broj_decetroje	5.429	4.214	1.288	0.203274
broj_decevice	17.263	4.297	4.017	0.000187 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.34 on 53 degrees of freedom

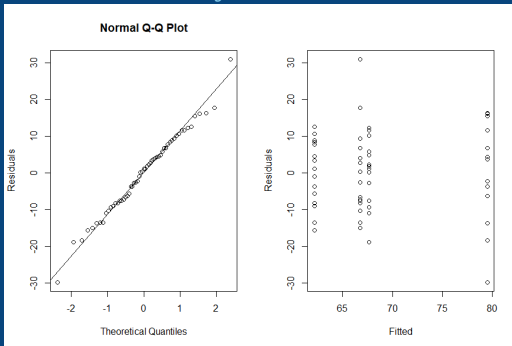
Multiple R-squared: 0.2451, Adjusted R-squared: 0.2023

F-statistic: 5.735 on 3 and 53 DF, p-value: 0.001787

Провера претпоставки



```
qqnorm(residuals(lm.samostalnost), ylab="Residuals")
qqline(residuals(lm.samostalnost))
plot(fitted(lm.samostalnost),
     residuals(jda.samostalnost), xlab="Fitted", ylab="Residuals")
```



Исти графици се добијају и ако се стави `jda.samostalnost` уместо `lm.samostalnost`.

- ★ Вишефакторска дисперзиона анализа, или факторски дизајн укључује више од једног фактора који имају две или више категорија. Приликом прављења дизајна треба обратити пажњу на комбинације категорија, такозвану интеракцију међу факторима.
- ★ Уравнотежени некомплетни блок дизајн се користи када нисмо у могућности да у сваком блоку имамо сваку групу. Тада формирамо дизајн пазећи да у сваком блоку имамо исти број група, као и да су групе равномерно заступљене.

