Analiza varijanse (ANOVA)

Analiza varijanse potpuno slučajnog rasporeda (prostog slučajnog rasporeda)

U istraživačkom radu često se testiraju ili ocenjuju više od dve sredine istog ili različitih osnovnih skupova. Statistički postupak kod ovakvih istraživanja poznat je pod nazivom analiza varijanse.
Analiza varijanse se sastoji u ispitivanju varijabiliteta aritmetičkih sredina iz više slučajno odabranih uzoraka, pri čemu se ukupan varijabilitet (ukupna varijansa) razdvaja na sastavne delove, odnosno na varijabilitet koji nastaje usled uticaja primenjenog tretmana i na slučajan varijabilitet.

U analizi varijanse polazimo od k uzoraka (tretmana) i izračunavamo njihove aritmetičke sredine.

Aritmetička sredina svakog od k uzoraka definisana je na sledeći način:

$$\overline{X}_{i\bullet} = \frac{\sum_{j=1}^{n_i} X_{ij}}{n_i}$$

n_i – broj jedinica u uzorku
X_{ij} – vrednost obeležja j-te jedinice i-tog tretmana

Opšta sredina svih jedinica (N) iz svih uzoraka definisana je sledećim izrazom:

$$\overline{X}_{\bullet\bullet} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}}{\sum_{i=1}^{k} n_i} \qquad N = \sum_{i=1}^{k} n_i$$

Ako su svi uzorci jednake veličine , odnosno sa jednakim brojem ponavljanja ukupan broj jedinica u analizi varijanse (N) može se iskazati na sledeći način:

$$n = n_1 = n_2 = ... = n_k$$
$$N = n \cdot k$$

Varijabilitet koji nastaje primenom odabranih *k* tretmana na *N* jedinica, u analizi varijanse iskazuje se na osnovu odstupanja svake individualne vrednosti obeležja od opšte sredine:

$$\left(X_{ij} - \overline{X}_{\bullet\bullet}\right) = \left(\overline{X}_{i\bullet} - \overline{X}_{\bullet\bullet}\right) + \left(X_{ij} - \overline{X}_{i\bullet}\right)$$

Ako se navedeni izraz kvadrira dobijaju se odgovarajuće sume kvadrata:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij}-\overline{X}_{\bullet\bullet}\right)^2=\sum_{i=1}^{k}\left(\overline{X}_{i\bullet}-\overline{X}_{\bullet\bullet}\right)^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij}-\overline{X}_{i\bullet}\right)^2$$

$$Q=Q_T+Q_P$$

$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij}-\overline{X}_{\bullet\bullet}\right)^2 \Rightarrow$ suma kvadrata totala Q

$\sum_{i=1}^{k}\left(\overline{X}_{i\bullet}-\overline{X}_{\bullet\bullet}\right)^2 \Rightarrow$ suma kvadrata tretmana Q_T (suma kvadrata između grupa; suma kvadrata objašnjene varijacije)

$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij}-\overline{X}_{i\bullet}\right)^2 \Rightarrow$ suma kvadrata pogreške Q_P (suma kvadrata unutar grupa; suma kvadrata neobjašnjene varijacije)

Na osnovu definisanih suma kvadrata proverava se polazna hipoteza u primeni metoda analize varijanse.

Polazna pretpostavka u analizi varijanse potpuno slučajnog rasporeda glasi:

$$H_0:\overline{X}_1=\overline{X}_2=\overline{X}_3=....=\overline{X}_k$$
$$H_1:\overline{X}_1\neq\overline{X}_2\neq\overline{X}_3\neq....\neq\overline{X}_k$$

Polazna hipoteza proverava se izvođenjem F testa. Za izvođenje ovog testa popunjava se tabela analize varijanse.

Izvori varijacije	**Stepeni slobode**	**Sume kvadrata**	**Sredine suma kvadrata (varijanse)**	**F- odnos**	**F - tablično**	
					0,05	**0,01**
Tretmani	**k-1**	Q_T	S_T^2	S_T^2/S_P^2	$r_{1=k-1}$	$r_{2=N-k}$
Pogreška	**N-k**	Q_P	S_P^2			
Total	**N-1**	**Q**				

Sume kvadrata se u praktičnom radu izračunavaju se primenom sledećih radnih formula:

$$Q=\sum_{i=1}^{k}\sum_{j=1}^{n_i}X_{ij}^2-C$$

C – je korektivni faktor $$C = \frac{\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} \mathrm{X}_{ij}\right)^2}{N}$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} \mathrm{X}_{ij} = T \qquad C = \frac{T^2}{N}$$

$$\sum_{j=1}^{n_i} \mathrm{X}_{ij} = T_i$$

$$n_1 = n_2 = ... = n_k$$

$$Q_T = \frac{\sum_{i=1}^{k}\left(\sum_{j=1}^{n_i} \mathrm{X}_{ij}\right)^2}{n} - C$$

$$Q_P = Q - Q_T$$

$$n_1 = n_2 = ... = n_k$$

$$Q_T = \frac{\sum_{i=1}^{k} T_i^2}{n} - C$$

$$n_1 \neq n_2 \neq ..... \neq n_k$$

$$Q_T = \sum_{i=1}^{k} \frac{\left(\sum_{j=1}^{n_i} \mathrm{X}_{ij}\right)^2}{n_i} - C$$

$$n_1 \neq n_2 \neq ..... \neq n_k$$

$$Q_T = \sum_{i=1}^{k} \frac{T_i^2}{n_i} - C$$

Sredine suma kvadrata izračunavaju se kao količnik suma kvadrata i odgovarajućih stepeni slobode.

$$S_T^2 = \frac{Q_T}{k-1} \qquad S_P^2 = \frac{Q_P}{N-k}$$

F odnos je količnik izračunatih varijansi i uvek je vrednost veća od nule.
F količnik upoređujemo sa kritičnim vrednostima iz tablica Fišerove distribucije.

$$F \langle F_{\alpha(r_1 < r_2)} \Rightarrow H_0$$

$$F^{**} \rangle F_{\alpha(r_1, r_2)} \Rightarrow H_1$$

$$F^{*} \rangle F_{0,05(r_1, r_2)} \Rightarrow H_1$$

Prihvatanje polazne hipoteze ukazuje da između primenjenih tretmana ne postoje statistički značajne razlike i time se analiza završava.
Ako se polazna hipoteza ne prihvati, već se utvrdi postojanje značajnih ili vrlo značajnih razlika između primenjenih tretmana, analiza varijanse se dalje nastavlja, da bi se utvrdilo između kojih tretmana postoje razlike. Pri tome je moguće izvesti više upoređenja između aritmetičkih sredina tretmana. Broj mogućih upoređenja određujemo na osnovu sledećeg izraza:

$\frac{k(k-1)}{2}$ - broj upoređenja sredina tretmana

Za testiranje razlika između sredina tretmana najčešće koristimo sledeće testove:

1. t – test
2. test najmanje značajne razlike – NZR test
3. višestruki test intervala – Dankanov test

t – test

Polazna hipoteza kod ovog testiranja glasi:

$$H_0 : \overline{X}_i = \overline{X}_j$$
$$H_1 : \overline{X}_i \neq \overline{X}_j$$

Za izvođenje ovog testa i proveru postavljene hipoteze utvrđuje se količnik t:

$$t = \frac{\overline{X}_i - \overline{X}_j}{S_{(\overline{X}_i - \overline{X}_j)}}$$

$\overline{X}_i, \overline{X}_j \Rightarrow$ aritmetičke sredine ispitivanih tretmana

$S_{(\overline{X}_i - \overline{X}_j)} \Rightarrow$ standardna greška razlike dve sredine

$$n_i = n_j \Rightarrow S_{(\overline{X}_i - \overline{X}_j)} = \sqrt{\frac{2 \cdot S_p^2}{n}}$$
$$n_i \neq n_j \Rightarrow S_{(\overline{X}_i - \overline{X}_j)} = \sqrt{S_p^2 \cdot \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

Izračunati količnik t, po apsolutnoj vrednosti upoređuje se sa kritičnim vrednostima iz tablica Studentove distribucije.

$|t| < t_{\alpha(N-k)} \Rightarrow H_0$

$|t| > t_{\alpha(N-k)} \Rightarrow H_1$

Test najmanje značajne razlike – NZR test

Za izvođenje NZR testa prvo se izračunavaju najmanje značajne razlike.

$$NZR_\alpha = t_{\alpha(N-k)} \cdot S_{(\overline{X}_i - \overline{X}_j)}$$

Kod ovog testa formira se pomoćna tabela gde se u prvoj koloni, aritmetičke sredine tretmana rangiraju prema veličini u vertikalnom nizu od maksimalne do minimalne vrednosti. U sledeće kolone se unose razlike sredina tretmana, koje su uvek pozitivne vrednosti. Razlike sredina tretmana upoređuju se sa izračunatim najmanje značajnim razlikama.
Na primer: da je u analizi varijanse primenjeno četiri tretmana A, B, C i D.
Najveću prosečnu vrednost ima tretman A, pa tretman B, zatim tretman C i najmanju vrednost sredine tretman D. Tabela za NZR test bi imala sledeći izgled:

Tretman		$\overline{X}_i$	$\overline{X}_i - \overline{X}_D$	$\overline{X}_i - \overline{X}_C$	$\overline{X}_i - \overline{X}_B$
	max				
A	**.**	$\overline{X}_A$	*	*	*
	.				
B	**.**	$\overline{X}_B$	*	*	
	.				
C	**.**	$\overline{X}_C$	*		
	.				
D	**min**	$\overline{X}_D$			

Višestruki test intervala – Dankanov test

Za izvođenje ovog testa izračunavamo standardnu grešku aritmetičke sredine:

$$S_{\overline{X}} = \sqrt{\frac{S_p^2}{n}}$$

$S_p^2 \Rightarrow$ varijansa pogreške iz tabele analize varijanse

Formiraju se zatim dve tabele za dva praga značajnosti (α = 0,05 i α= 0,01) sledećeg oblika:

Interval	**2**	**3**	**4**	**.....**	**k**

Kritična vrednost	
Najmanje značajni interval	

U tabeli u prvom redu upisuju se mogući intervali na osnovu broja posmatranih tretmana. Zatim se očitavaju kritične vrednosti iz tablica za višestruki test intervala za date pragove značajnosti α i stepene slobode pogreške iz tabele analize varijanse i to za svaki interval idući od 2, 3, 4,.....k, koje se upisuju u drugi red tabele. Očitane i upisane kritične vrednosti množe se sa izračunatom standardnom greškom aritmetičke sredine, a proizvod predstavlja vrednost najmenje značajnog intervala i njega upisujemo u treći red tabele.

Sa najmanje značajnim intervalima upoređujemo razlike aritmetičkih sredina tretmana.

Aritmetičke sredine tretmana rangiraju se u pomoćnoj tabeli u horizontalnom nizu od minimalne do maksimalne vrednosti.

Tretman	**D C B A** **min . . . . . max**

$\overline{X}_A - \overline{X}_D$ $\quad$ $\overline{X}_B - \overline{X}_D$ $\quad$ $\overline{X}_C - \overline{X}_D$

$\overline{X}_A - \overline{X}_C$ $\quad$ $\overline{X}_B - \overline{X}_C$

$\overline{X}_A - \overline{X}_B$

Primer 1. U ogledu sa tri sorte breskve (A, B, C) dobijeni su sledeći podaci o prinosu (vagona /ha) koji su dati u tabeli. Metodom analize varijanse ustanoviti da li postoji statistički značajna razlika između prosečnih prinosa posmatranih sorata. Utvrditi između kojih sorata postoje razlike u prosečnom prinosu.

Ponavljanja	**Sorte (tretmani)**		
n_i	**A**	**B**	**C**
1	**5**	**4**	**7**
2	**4**	**4**	**6**
3	**4**	**4**	**5**
4	**6**	**5**	**6**
5	**5**	**4**	**6**
6	**6**	**3**	**6**
T_i	**30**	**24**	**36**
$\overline{X}_i$	**5**	**4**	**6**

$k = 3$

$n = 6$

$N = 18$

$H_0 : \overline{X}_A = \overline{X}_B = \overline{X}_C$

$H_1 : \overline{X}_A \neq \overline{X}_B \neq \overline{X}_C$

Izvori varijacije	**Stepeni slobode**	**Sume kvadrata**	**Sredine suma kvadrata (varijanse)**	**F- odnos**	**F - tablično**	
					0,05	**0,01**
Tretmani	**k-1= 2**	**12**	**$S_T^2 = 6$**	**F = 11,32****	**3,68**	**6,36**
Pogreška	**N-k= 15**	**8**	**$S_P^2 = 0,53$**			
Total	**N-1= 17**	**20**				

$$Q = \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij}^2 - C = 5^2 + 4^2 + 7^2 + \dots\dots + 6^2 - C$$

$$C = \frac{\left(\sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij} \right)^2}{N} = \frac{90^2}{18} = 450$$

$$Q = \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij}^2 - C = 470 - 450 = 20$$

$$n_1 = n_2 = \dots = n_k$$

$$Q_T = \frac{\sum_{i=1}^{k} T_i^2}{n} - C = \frac{30^2 + 24^2 + 36^2}{6} - 450 = 12$$

$$Q_P = Q - Q_T = 20 - 12 = 8$$

$$S_T^2 = \frac{Q_T}{k-1} = \frac{12}{2} = 6 \qquad S_P^2 = \frac{Q_P}{N-k} = \frac{8}{15} = 0,53$$

$$F = \frac{S_T^2}{S_P^2} = \frac{6}{0,53} = 11,32$$

$$F_{0,05;2,15=3,68}$$

$$F_{0,01;2,15=6,36}$$

$$F^{**} \rangle F_{\alpha(r_1,r_2)} \Rightarrow H_1$$

t – test

$$H_0 : \overline{X}_i = \overline{X}_j$$
$$H_1 : \overline{X}_i \neq \overline{X}_j$$

$$\frac{k(k-1)}{2} = \frac{3 \cdot (3-1)}{2} = 3$$

$$t = \frac{\overline{X}_i - \overline{X}_j}{S_{(\overline{X}_i - \overline{X}_j)}}$$

$$n_i = n_j \Rightarrow S_{(\overline{X}_i - \overline{X}_j)} = \sqrt{\frac{2 \cdot S_p^2}{n}} = \sqrt{\frac{2 \cdot 0{,}53}{6}} = 0{,}42$$

$$|t| \langle t_{\alpha(N-k)} \Rightarrow H_0$$
$$|t| \rangle t_{\alpha(N-k)} \Rightarrow H_1$$

$$t_2 = \frac{\overline{X}_A - \overline{X}_C}{S_{(\overline{X}_i - \overline{X}_j)}} = \frac{5-6}{0{,}42} = |-2{,}38|^{*}$$

$$t_{0,05(15)} = 2{,}131$$
$$t_{0,01(15)} = 2{,}947$$

$$t_3 = \frac{\overline{X}_B - \overline{X}_C}{S_{(\overline{X}_i - \overline{X}_j)}} = \frac{4-6}{0{,}42} = |-4{,}76|^{**}$$

NZR test

$$H_0 : \overline{X}_i = \overline{X}_j$$
$$H_1 : \overline{X}_i \neq \overline{X}_j$$

$$NZR_\alpha = t_{\alpha(N-k)} \cdot S_{(\overline{X}_i - \overline{X}_j)}$$

$$NZR_{0,05} = 2{,}131 \cdot 0{,}42 = 0{,}895$$
$$NZR_{0,01} = 2{,}947 \cdot 0{,}42 = 1{,}238$$

Tretman		$\overline{X}_i - \overline{X}_B$	
C	**6**	**2****	**1***
A	**5**	**1***	
B	**4**		

$\overline{X}_C - \overline{X}_B = 6 - 4 = 2$
$\overline{X}_A - \overline{X}_B = 5 - 4 = 1$
$\overline{X}_C - \overline{X}_A = 6 - 5 = 1$

Višestruki test intervala – Dankanov test

$H_0 : \overline{X}_i = \overline{X}_j$
$H_1 : \overline{X}_i \neq \overline{X}_j$

$$S_{\overline{X}} = \sqrt{\frac{S_p^2}{n}} = \sqrt{\frac{0{,}53}{6}} = 0{,}297$$

α = 0,05

Interval	**2**	**3**
Kritična vrednost	**3,01**	**3,16**
Najmanje značajni interval	**0,894**	**0,939**

3,01· 0,297 = 0,894

3,16· 0,297 = 0,939

α = 0,01

Interval	**2**	**3**
Kritična vrednost	**4,17**	**4,37**
Najmanje značajni interval	**1,238**	**1,298**

4,17 · 0,297 = 1,238
4,37 · 0,297 = 1,298

Tretman	**B**	**A**	**C**
	4	**5**	**6**

$\overline{X}_C - \overline{X}_B = 6 - 4 = 2^{**} > 0{,}939; > 1{,}298$
$\overline{X}_C - \overline{X}_A = 6 - 5 = 1^{*} > 0{,}939$

$\overline{X}_A - \overline{X}_B = 5 - 4 = 1^{*} > 0{,}894; 1 < 1{,}238$