

STATISTIKA ANALIZA VARIJANSE

Ispitujemo uticaj faktora na obeležje.

Ako se faktor koji utiče na obeležje – odnosno nezavisna promenljiva - mogu izraziti kvantitativno (kontinualne promenljive na intervalnoj skali) onda koristimo linearni regresioni model obrađen u prethodnom poglavlju.

Ako se faktor koji utiče na obeležje izražava kvalitativno (kategorijska promenljive na nominalnoj ili ordinalnoj skali), onda koristimo analizu varijanse.

- Da li postoji razlika između ocena u 3 ili više grupa za slušanje vežbi?
- Da li ima razlike u prosečnoj plati u različitim gradovima u Srbiji? (različitim zemljama)
- Da li im razlike u znanju matematike studenata FON, ETF, EKO, PMF
- Da li ima razlike u prodaji ako koristimo reklame na TV, Radio, Bilbordi..
- Itd...

Analiza varijanse ustvari ispituje da li postoji razlika između 3 ili više grupa vezano za neko obeležje (isto kao i t test za 2 grupe). To se manifestuje razlikom srednjih vrednosti. Ako postoji razlika srednjih vrednosti, znači ima razlike u grupama (postoji uticaj faktora). Ako su srednje vrednosti iste nema razlike u grupama (ne postoji uticaj faktora).

Jednakost srednjih vrednosti više skupova ispituje analizu varijanse. Jednakost sredina nam u stvari pokazuje da li postoji uticaj faktora na obeležje.

Suština analize varijanse je u razlaganju ukupnog varijabiliteta obeležja na sastavne komponente (izvore): varijabilitet koji nastaje pod uticajem kontrolisanih faktora i tzv. rezidualni varijabilitet koji nastaje pod uticajem ostalih (**nekontrolisanih**) faktora.

- Npr. – faktor grupa za slušanje vežbi, obeležje poeni na ispitu

GR1: $X_{11} \ X_{12} \ X_{13}$

GR2: $X_{21} \ X_{22} \ X_{23}$

GR3: $X_{31} \ X_{32} \ X_{33}$

Faktore čije delovanje ispituje označavamo velikim slovom A, B.. Modaliteti određenog faktora u analizi varijanse nazivaju se **tretmanima (nivoima)** i nivoe posmatranog faktora obeležavamo sa $A_1, A_2, \dots A_k$.

- Npr. – nivoi su različite grupe za slušanje vežbi.

Parametarska je metoda.

Jednofaktorska analiza varijanse

Istražuje uticaj jednog faktora A na varijabilitet posmatranog obeležja.

Faktor ima više kategorija - k nivoa (tretmana), $A_1, A_2, \dots, A_k$. Elementi statističkog skupa su podeljeni u grupe prema kategorijama. Prvu grupu čine elementi na koje je delovao prvi tretman faktora, drugu grupu čine oni elementi na koje je delovao drugi tretman, itd.

Kod svake grupe meri se vrednost obeležja X. Neka su to vrednosti $X_1, X_2, \dots$ koje odgovaraju grupama.

Ako faktor nema uticaja na obeležje, onda će svaka grupa imati iste karakteristike posmatranog obeležja. To znači da će karakteristike obeležja po grupama, $X_1, X_2, \dots$ biti iste (iste srednje vrednosti).

Ako faktor ima uticaja, onda će karakteristike biti različite

Ukoliko bi kontrolisani faktor značajno uticao na varijabilitet posmatranog obeležja on bi prouzrokovao varijacije između grupa, što bi se manifestovalo relativno velikim razlikama između srednjih vrednosti. Ostali nekontrolisani faktori mogu prouzrokovati varijacije unutar pojedinih grupa.

Analiza varijanse zasniva se na sledećoj ideji: što su veće varijacije između grupa u odnosu na varijacije unutar njih veći je uticaj kontrolisanog faktora u odnosu na nekontrolisani.

Osnovni cilj je da se proverí hipoteza $H_0 (m_1 = m_2 = \dots = m_k)$ a to je pretpostavka da ***faktor nema značajnog uticaja na posmatrano obeležje.***

Ukupan totalan varijabilitet obeležja X jednak je zbiru varijabiliteta nastalih pod dejstvom kontrolisanog faktora i varijabiliteta pod dejstvom rezidualnih (nekontrolisanih) faktora.

Ukupan varijabilitet – odstupanje opservacija svih uzoraka od njihove aritmetičke sredine:

$$q = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 = \sum_{i=1}^k (\bar{x}_i - \bar{x})^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 = \left[\sum_{i=1}^k \sum_{j=1}^n x_{ij}^2 - \frac{1}{n} \left(\sum_{i=1}^k \sum_{j=1}^n x_{ij} \right)^2 \right]$$

$$q_1 = \sum_{i=1}^k (\bar{x}_i - \bar{x})^2 = \left[\sum_{i=1}^k \frac{1}{n_i} \left(\sum_{j=1}^{n_i} x_{ij} \right)^2 - \frac{1}{n} \left(\sum_{i=1}^k \sum_{j=1}^n x_{ij} \right)^2 \right]$$

– zbir kvadrata odstupanja između grupa - faktorski varijabilitet između grupa, pod dejstvom kontrolisanih faktora

$$q_2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 \quad \boxed{= q - q_1}$$

- zbir kvadrata odstupanja unutar grupa - rezidualni varijabilitet unutar grupa, pod uticajem nekontrolisanih faktora

$q = q_1 + q_2$ - ukupan zbir kvadrata odstupanja

$S_1^2 = \frac{q_1}{k-1}$ - srednje kvadratno odstupanje između grupa – faktorska varijansa

$S_2^2 = \frac{q_2}{n-k}$ - srednje kvadratno odstupanje unutar grupa

$S^2 = \frac{q}{n-1}$ - ukupno srednje kvadratno odstupanje

	Zbir kvadrata odstupanja	Broj stepeni clobode	Srednje kvadratno odstupanje
Između grupa	q_1	$k-1$	$S_1^2 = \frac{q_1}{k-1}$
Unutar grupa	q_2	$n-k$	$S_2^2 = \frac{q_2}{n-k}$
Ukupan	q	$n-1$	$S^2 = \frac{q}{n-1}$

$$\tau = \frac{S_1^2}{S_2^2} : F_{(k-1), (n-k)}$$

$$P\{F > F_0\} = \alpha \quad - \text{desnostrana } c = (F_0; +\infty)$$

Dvofaktorska analiza varijanse

Posmatraju se dva faktora od kojih prvi ima r , a drugi s različitih tretmana. Na elemente statističkog skupa deluju oba faktora istovremeno.

Izvor varijacije	Zbir kvadrata odstupanja	Stepeni slobode	Ocena varijacija	F
Faktor A	S_A	$r-1$	V_A	$\frac{V_A}{V_R}$
Faktor B	S_B	$s-1$	V_B	$\frac{V_B}{V_R}$
Rezidual	S_R	$(r-1)(s-1)$	V_R	
Total	S_T	$rs-1$		

$$S_T = S_A + S_B + S_R$$

$$S_A = \frac{\sum_{i=1}^r S_i^2}{s} - \frac{S^2}{rs} \quad \text{- zbir kvadrata odstupanja između tretmana I faktora}$$

$$S_B = \frac{\sum_{j=1}^s S_j^2}{r} - \frac{S^2}{rs} \quad \text{- zbir kvadrata odstupanja između tretmana II faktora}$$

$$S_T = \sum_{i=1}^r \sum_{j=1}^s x_{ij}^2 - \frac{S^2}{rs} \quad \text{- totalni zbir kvadrata odstupanja}$$

$$S_R = S_T - (S_A + S_B) \quad \text{- rezidualni zbir kvadrata odstupanja}$$

$$S = \sum_i \sum_j x_{ij}$$

$$V_A = \frac{S_A}{r-1} \quad V_B = \frac{S_B}{s-1} \quad V_R = \frac{S_R}{(r-1)(s-1)}$$

VAŽNO: Faktori su uvek kategorijskog tipa.
Poredi se da li ima razlike u grupama.
Obeležje je uvek kontinualnog tipa
Unutar tablice su vrednosti posmatranog obeležja.

ZADACI:

1. Analizirajući tri vrste automobilskih guma, G1, G2 i G3 dobijeni su sledeći podaci o broju pređenih kilometara

G1	40	30	38	
G2	35	40	45	
G3	42	40	36	50

Gume su odabrane na slučajan način. Pretpostavka je da su uzete iz populacija sa normalnom raspodelom i sa istom varijansom. Da li se na osnovu ovih podataka može prihvatiti pretpostavka da je prosečan vek sve tri vrste guma isti?

Rešenje:

$$H_0 : m_1 = m_2 = m_3$$

H_1 : "Bar dve vrednosti matematičkih očekivanja se međusobno razlikuju"

				suma	
G1	40	30	38	108	
G2	35	40	45	120	
G3	42	40	36	50	168

Analiza varijanse se najčešće prikazuje tabelarno

	Zbir kvadrata odstupanja	Broj stepeni clobode	Srednje kvadratno odstupanje
Između grupa	q_1	$k-1$	$S_1^2 = \frac{q_1}{k-1}$
Unutar grupa	q_2	$n-k$	$S_2^2 = \frac{q_2}{n-k}$
Ukupan	q	$n-1$	$S^2 = \frac{q}{n-1}$

$$F = \frac{S_1^2}{S_2^2} : F_{(k-1), (n-k)}$$

$$k=3 \text{ (broj grupa)} \quad n_1=3 \quad n_2=3 \quad n_3=4 \quad n=10$$

$$\bar{x}_1 = \frac{40+30+38}{3} = 36 \quad \bar{x}_2 = \frac{35+40+45}{3} = 40 \quad \bar{x}_3 = \frac{42+40+36+50}{4} = 42$$

$$\sum_i \sum_j x_{ij} = 396 \quad \sum_i \sum_j x_{ij}^2 = 15954$$

$$q = \sum_i \sum_j x_{ij}^2 - \frac{1}{n} \left(\sum_i \sum_j x_{ij} \right)^2 = 15954 - \frac{1}{10} 396^2 = 15954 - \frac{1}{10} 156816 = 272.4$$

$$q_1 = \sum_i \frac{1}{n_i} \left(\sum_j x_{ij} \right)^2 - \frac{1}{n} \left(\sum_i \sum_j x_{ij} \right)^2 = \frac{1}{3} 108^2 + \frac{1}{3} 120^2 + \frac{1}{4} 168^2 - \frac{1}{10} 156816 = 15744 - 15681.6 = 62.4$$

$$q_2 = q - q_1 = 272.4 - 62.4 = 210$$

Dobijene rezultate uvrštavamo u tabelu analize varijanse:

	Zbir kvadrata odstupanja	Broj stepeni clobode	Srednje kvadratno odstupanje
Između grupa	$q_1=62.4$	2	$S_1^2 = 31.2$
Unutar grupa	$q_2=210$	7	$S_2^2 = 30$
Ukupan	$q=272.4$	9	$S^2 = 30.26$

$$\tau = \frac{S_1^2}{S_2^2} = \frac{31.2}{30} = 1.04$$

tab

$$P\{\tau > F_0\} = 0.05 \quad P\{\tau < F_0\} = 0.95 \quad \Rightarrow \quad F_0 = 4.74$$

$F_{2,7}$

$$c = (4.74; +\infty) \quad 1.04 \notin (4.74; +\infty) \Rightarrow F \notin c \Rightarrow H_0 - \text{tačna}$$

Prihvataemo H_0 i zaključujemo da je srednji vek svih analiziranih auto guma isti.

2. Prinos jedne žitarice zavisi od sorte i vremena setve. Radi provere nove sorte pšenice, izvršena su merenja prosečnih prinosa po hektaru i dobijeni su podaci.

Godišnje doba Sorta	Jesen	Zima	Proleće	Leto
1	80	86	89	69
2	75	79	92	72
3	70	78	86	66

Testirati razlike u prosečnim prinosima pojedinih sorti, kao i razlike u prosečnim prinosima za pojedina godišnja doba setve.

Rešenje:

Radi se analiza varijanse sa dva faktora. Napravićemo tabelu, tako što ćemo godišnja doba označiti sa A_i , $i=1,2,3,4$, a sorte pšenice sa B_j , $j=1,2,3$.

Godišnje doba Sorta	A_1	A_2	A_3	A_4
B_1	80	86	89	69
B_2	75	79	92	72
B_3	70	78	86	66

Formulisaćemo hipoteze koje se odnose na godišnja doba:

$$H_{01}: m_{A1} = m_{A2} = m_{A3} = m_{A4}$$

$$H_{11}: \text{"Efekat barem jednog nivoa faktora A se razlikuje od nule."}$$

Hipoteze koje se odnosi na sortu pšenice:

$$H_{02}: m_{B1} = m_{B2} = m_{B3}$$

$$H_{12}: \text{"Efekat barem jednog nivoa faktora B se razlikuje od nule."}$$

Za analizu varijanse pravimo tabelu oblika:

	Zbir kvadrata odstupanja	Broj stepeni slobode	Srednje kvadratno odstupanje
Između grupa	q_1	$k-1$	$S_1^2 = \frac{q_1}{k-1}$
Unutar grupa	q_2	$n-k$	$S_2^2 = \frac{q_2}{n-k}$
Ukupan	q	$n-1$	$S^2 = \frac{q}{n-1}$

Pri čemu je:

$$S_A = \frac{\sum_{i=1}^r S_i^2}{s} - \frac{S^2}{rs} \quad S_B = \frac{\sum_{j=1}^s S_j^2}{r} - \frac{S^2}{rs} \quad S = \sum_i \sum_j x_{ij}$$

$$S_T = \sum_{i=1}^r \sum_{j=1}^s x_{ij}^2 - \frac{S^2}{rs} \quad S_R = S_T - (S_A + S_B)$$

$$V_A = \frac{S_A}{r-1} \quad V_B = \frac{S_B}{s-1} \quad V_R = \frac{S_R}{(r-1)(s-1)}$$

	A ₁	A ₂	A ₃	A ₄	S _j	S _j ²
B ₁	80	86	89	69	324	104976
B ₂	75	79	92	72	318	101124
B ₃	70	78	86	66	300	90000
S _i	225	243	267	207	S=942	$\sum S_j^2 = 296100$
S _i ²	50625	59049	71289	42849	$\sum S_i^2 = 223812$	

r predstavlja broj nivoa A, tj. r=4

s predstavlja broj nivoa B, tj. s=3

$$S_A = \frac{223812}{3} - \frac{942^2}{4 \cdot 3} = 657 \quad S_B = \frac{296100}{4} - \frac{942^2}{4 \cdot 3} = 78$$

$$S_T = (80^2 + 75^2 + \dots + 66^2) - \frac{942^2}{4 \cdot 3} = 781 \quad S_R = 781 - (657 + 78) = 46$$

$$V_A = \frac{657}{3} = 219 \quad V_B = \frac{78}{2} = 39 \quad V_R = \frac{46}{6} = 7.67$$

Dobijene rezultate predstavimo tabelom

Izvor varijacije	Zbir kvadrata odstupanja	Stepeni slobode	Ocena varijacija	F
Faktor A	$S_A = 657$	$r-1=3$	$V_A = 219$	$V_A/V_R = 28.56$
Faktor B	$S_B = 78$	$s-1=2$	$V_B = 39$	$V_B/V_R = 5.08$
Rezidual	$S_R = 46$	$(r-1)(s-1)=6$	$V_R = 7.67$	
Total	$S_T = 781$	$rs-1=11$		

Prema tome, dobijene su vrednosti statistika

$\tau_1 = 28.56$ za testiranje H_{01} i $\tau_2 = 5.08$ za testiranje H_{02}

1) Određujemo kritičnu oblast za hipotezu koja se odnosi na godišnja doba

$$P\{\tau_1 > F_{01}\} = \alpha \quad \tau_1 : F_{(r-1), (r-1)(s-1)}$$

$$P\{\tau_1 > F_{01}\} = 0.05$$

tab

$$P\{\tau_{11} < F_{01}\} = 0.95 \Rightarrow F_{01} = 4.76 \quad c = (4.76; +\infty)$$

$F_{3,6}$

$$28.56 \in (4.76; +\infty)$$

Odbacujemo H_{01} i zaključujemo da faktor A (godišnja doba) sistematski utiče na varijabilitet posmatrane pojave (prinos po hektaru).

2) Određujemo kritičnu oblast za hipotezu koja se odnosi na sorte pšenice:

$$P\{\tau_2 > F_{02}\} = \alpha \quad \tau_2 : F_{(s-1), (r-1)(s-1)}$$

$$P\{\tau_2 > F_{02}\} = 0.05$$

tab

$$P\{\tau_2 < F_{02}\} = 0.95 \Rightarrow F_{02} = 5.14 \quad c = (5.14; +\infty)$$

$F_{2,6}$

$$5.08 \notin (5.14; +\infty)$$

Prihvataemo H_{02} i zaključujemo da faktor B (sorta pšenice) ne utiče sistematski na varijabilitet posmatrane pojave (prinos po hektaru).