

DISPERSIOONANALÜÜS

Dispersioonanalüüs (DA) on mõeldud tunnuse, mille väärtused eeldatavasti sõltuvad üheaegselt mõjuvatest sõltumatutest tunnustest, varieeruvuse analüüsiks.

Sõltuv tunnus	Y (intervall või suhteskaala)
Sõltumatu (faktor-)tunnus	X (suvaline skaala)

DA idee
Sõltuva tunnuse varieeruvus kirjeldatakse teda mõjutavate
faktortunnuste varieeruvuse kaudu.

$$S^2 = (\sum (Y_i - \bar{Y})^2) / (N-1), \text{ kus}$$

S² dispersioon,
My tunnuse Y aritmeetiline keskmine,
Y_i tunnuse Y väärtus I-ndal objektil,
N objektide arv.

$$SS = \sum (Y_i - \bar{Y})^2, \text{ kus}$$

SS tunnuse Y koguvarieeruvus.

Faktortunnus X omagu K väärtusklassi, st. $j=1,...,K$. Saame moodustada K sõltumatut valimit mahuga $N_1, N_2,..., N_K$

$$\mathbf{X} \begin{matrix} 1 & \mathbf{Y}_{11} & \mathbf{Y}_{12} & \dots & \mathbf{Y}_{1N1} \\ 2 & \mathbf{Y}_{21} & \mathbf{Y}_{22} & \dots & \mathbf{Y}_{2N2} \\ \dots & & & & \\ \mathbf{K} & \mathbf{Y}_{K1} & \mathbf{Y}_{K2} & \dots & \mathbf{Y}_{KNK}, \end{matrix}$$

$$\mathbf{N}_1 + \mathbf{N}_2 + \dots + \mathbf{N}_K = \mathbf{N}.$$

Igale X väärtusklassile (rühmale) $j = 1, 2, \dots, K$ arvutatakse

1) **tunnuse Y aritmeetiline keskmine**

$$M_{Yj} = (\sum_t Y_{jt})/N_j, \quad (M_{Y1}, M_{Y2}, \dots, M_{YK}), t=1, \dots, N_j,$$

2) **varieeruvus**

$$\underline{SS_j = \sum_t (Y_{jt} - M_{Yj})^2, (SS_1, SS_2, \dots, SS_K),}$$

$$M_Y = (j \sum_t \sum Y_{jt}) / (N_1 + N_2 + \dots + N_K)$$

$$SS = j \sum_t \sum (Y_{jt} - M_Y)^2, \quad j=1, \dots, K, t=1, \dots, N_j$$

$$SS = SS_X + SS_J, \quad \text{kus}$$

$SS_X = j \sum_t \sum (Y_{jt} - M_{Yj})^2$, so. sõltuva tunnuse Y varieeruvus faktortunnuse X tõttu (rühmade sisene),

$SS_J = j \sum N_j (M_{Yj} - M_Y)^2$, so. see osa Y varieeruvusest, mis on tingitud määramata faktorite poolt (jääk- e rühmade vaheline).

Kui 2 faktortunnust (A, B):

$$SS = SS_A + SS_B + SS_{AB} + SS_J$$

Kui 3 faktortunnust (A, B, C):

$$SS = SS_A + SS_B + SS_C + SS_{AB} + SS_{AC} + SS_{BC} + SS_{ABC} + SS_J$$

Näide

Tabel 1. Keskmise tarbitav alkoholikogus aastas inimese kohta (Y) (pinti) eri rahvusgruppides (X)

Jrk	Rahvus	N	M_{Yj}	<u>SD</u>	95% int.
1	Iirlased	119	24,25	25,60	19.61 ... 28,89
2	Itaallased	84	24,31	24,19	19,06 ... 29,56
3	Juudid	41	9,25	21,65	2,42 ... 16,08
4	Rootslased	74	16,56	26,75	10,37 ... 22,76
5	Inglased	90	21,87	21,56	17,36 ... 26,39

$$H_0: M_{Y1} = M_{Y2} = \dots = M_{YK}$$

	df	SS	M_{SS}	F	Olulisus F_{ARV}
Rühmade vaheline (SS_J)	4	9353,87	2338,47	3,96	0,0036
Rühmade sisene (SS_X)	403	237986,39	590,54		
Kokku (SS)	407	247340,26			

M_{Yj}	Jrk	Rahvus	1	2	3	4	5
9,25	1	Juudid					
16,56	2	Rootslased					
21,87	3	Inglased					
24,25	4	Iirlased	*				
24,31	5	Itaallased	*				

Ülddispersioon koosneb süstemaatilise ja veadispersiooni. Süstemaatiline dispersioon kujutab endast kogudispersiooni seda osa, mis on tingitud analüüsis osalevate tunnuste poolt. Veadispersioon on see osa kogudispersioonist, mis on põhjustatud rühmade erinevusest, so. juhuslike, määramata faktorite poolt.

Dispersioonanalüüs hõlmab süstemaatilise ning veadispersiooni jaotust ning analüüsi. Jaotusmeetodid erinevad sõltuvalt meie poolt rakendatavast konkreetsest eksperimentaalsest plaanist.

Rühmade keskmiste vahelisi erinevusi testitakse F-kriteeriumi abil, mis võimaldab võrrelda rühmade keskmiste vahelist dispersiooni rühmasisese dispersiooniga.

Dispersioonianalüüsi kasutamiseks peavad olema rahuldatud järgmised eeldused:

- 1) tulemusi mõõdetakse vähemalt intervallskaalal,
- 2) valimid koostatakse juhukorras,
- 3) valimid võetakse normaaljaotusega üldkogumitest,
- 4) vastavate üldkogumite dispersioonid on võrdsed.

ÜHESE DISPERSIOONANALÜÜSI KASUTAMINE KORRELEERITUD KESKMISTE KORRAL (sama valimi korduv mõõtmine või paarisvalimite eraldi mõõtmine)

Näide.

Atmosfääri saastatuse muutumise uurimiseks mõõdeti suitsu kontsentratsiooni erinevates kohtades üle terve linna. Mõõtmised teostati kolmel korral aastas ühe nädala vältel. Eesmärgiks on kindlaks määrata, kas suitsusisalduse tasemete keskmine kontsentratsioon nimetatud kolmel korral on muutunud.

Tabel. Atmosfääri saastumine: keskmine suitsu kontsentratsioon linna eri kohtades (mikrogrammi kuupmeetri kohta)

Asukohad	Suitsu kontsentratsioon		
	Oktoober (X1)	Veebruar (X2)	Mai (X3)
A	26	25	25
B	24	25	22
C	23	24	25
D	27	27	26
E	27	25	24
F	27	23	23
G	23	24	23
H	24	26	27

1) H_0 : suitsukontsentratsiooni keskmises tasemes eri ajaperioodidel olulisi erinevusi (muutusi) ei esine.

2) H_0 : suitsukontsentratsiooni keskmises tasemes eri mõõtmiskohtades olulisi erinevusi esine.

Antud näite puhul võib kogudispersiooni jagada järgnevalt:

SÜSTEEMAATILISED MÕJUD

1. Katsetevaheline dispersioon kujutab endast tulemuste gruppide keskmiste vahelist dispersiooni, mis on esile kutsutud sõltumatu tunnuse (X), antud juhul mõõtmise vahelise ajaperioodi poolt.

2. Kohtadevaheline dispersioon on dispersioon, mis on tingitud sellest, et samades kohtades teostati mõõtmisi kolmel erineval korral, mistõttu andmed on korreleeritud. Näiteks mõningad kohad omavad kõigil kolmel korral märkimisväärselt suuremaid tulemusi kui ülejäänud.

VEA MÕJUD

3. Jääkdispersioon on tingitud mõõtmiskoha muutusest kolmel erineval mõõtmiskorral. See mõõdab faktoreid, mis on põhjustatud mitte kohtade individuaalsete mõjude või kontrollimise aja poolt, vaid nimetatud tunnuste koosmõju poolt.

$$F_{\text{KATSETE VAHEL}} = \frac{\text{Katsetevaheline dispersioon}}{\text{Jääkdispersioon}}$$

$$F_{\text{KOHTADE VAHEL}} = \frac{\text{Kohtadevaheline dispersioon}}{\text{Jääkdispersioon}}$$

Tabel. Dispersioonanalüüsi tabel

	df	SS	M _{SS}	F	F _{KR} (0,05)
Katsetevaheline	2	2,34	1,17	0,62	3,72
Kohtadevaheline	7	27,30	3,89	2,07	2,78
Jääkdispersioon	14	26,33	1,88		
Kokku	23	55,97			

1) Võtame vastu H_0 ning järeldame, et antud perioodil ei esine suitsu keskmises kontsentratsioonis olulisi erinevusi.

2) Võtame vastu H_0 ning järeldame, et erinevates mõõtmiskohtades suitsu keskmises kontsentratsioonis olulised erinevused puuduvad, kui me vaatleme neid katsetest sõltumatult.

Kui katsete vahel oleks täheldatud olulisi erinevusi, tuleks rakendada Tukey-kriteeriumi.

FRIEDEMANNI JÄRJESTUSNUMBRITE KAHEPOOLSE DISPERSIOONANALÜÜSI KASUTAMINE

Näide

Kuuel juhtimiskonsultandil palutakse igaühel järjestada viis kandidaati sobivuse alusel tippjuhi kohale. Soovitakse selgitada, kas mingi ühe kandidaadi suhtes esineb konsultantide poolset olulist soosingut.

Tabel. Eelistused kandidaatide suhtes

Juhtimis- konsultandid	Kandi Daadi d				
	L	M	N	O	P
A	2	4	5	1	3
B	1	5	3	2	4
C	3	2	1	4	5
D	2	4	3	1	5
E	3	5	1	2	4
F	2	3	4	1	5

H_0 : kandidaatide järjestused oluliselt ei erine.

(Sisuliselt: konsultantide valikud pole suunatud ühele konkreetsele kandidaadile).

Antud probleemi puhul on sobiv kasutada Friedemanni järjestusnumbrite kahepoolset dispersioonanalüüsi. Oma näites vaatleme me üht ja sama subjektide gruppi (kuus juhtimiskonsultanti) iga järjestuse suhtes viiest (viis kandidaati). Friedeman'i kriteerium eeldab ordinaalseid andmeid:

$$X_r^2 = \frac{12 \cdot H}{n \cdot k \cdot (k+1)} - 3 \cdot n \cdot (k+1),$$

kus

n - tabeli ridade arv,

k - tabeli veergude arv,

V^2 – veeru (kandidaadi) järjestusnumbrite summa ruut,

H - järjestusnumbrite summade ruutude kogusumma ($\sum V^2$).

FRIEDEMANNI KRITEERIUMI (X^2_r) ARVUTAMINE

Tabel. Valijate eelistused kandidaatide suhtes (arvutamine)

Juhtimis- konsultandid	Kandi daadi				
	L	M	N	O	P
A	2	4	5	1	3
B	1	5	3	2	4
C	3	2	1	4	5
D	2	4	3	1	5
E	3	5	1	2	4
F	2	3	4	1	5
Σ	13	23	17	11	26

$$n=6, k=5$$

$$(V_1)^2=169, (V_2)^2=529, (V_3)^2=289, (V_4)^2=121, (V_5)^2=676,$$

$$H=(V_1)^2+(V_2)^2+(V_3)^2+(V_4)^2+(V_5)^2=1784.$$

$$X_r^2 = \frac{12 \cdot 1784}{6 \cdot 5 \cdot (5+1)} = 18 \cdot (5+1) = 10,9.$$

X^2_r OLULISUSE KONTROLLIMINE

5%-lisel olulisuse nivool (0,05) on X^2_r kriitiliseks väärtuseks 9,07. Meie poolt arvutatud väärtus on sellest suurem. Seetõttu võime H_0 hüljata ning järeldada, et juhtimiskonsultantide valikud on suunatud ühele konkreetsele kandidaadile (kandidaat "O").

Friedeman'i kriteeriumi efektiivsus parameetrilise F-kriteeriumi suhtes korreleeritud keskmiste korral väidetakse olevat $95,5 \cdot k/(k+1) \%$, kus k - tähistab tabeli veergude arvu.

KAHEPOOLNE DISPERSIOONANALÜÜS (erinevate gruppide ühekordne vaatlus)

Kahepoolset dispersioonanalüüsi kasutatakse kahe sõltumatu tunnuse toime hindamiseks sõltuvale tunnusele.

		Sõltumatu tunnus (Faktor B)	
		Immigrandid	Mitteimmigrandid
Sõltumatu tunnus A (Faktor A)	Kõrge stress	1. rühm	2. rühm
		9,1	8,2
		9,4	8,8
		9,6	8,7
		9,8	9,0
	Madal stress	3. rühm	4. rühm
		7,5	7,3
		7,1	7,2
		7,2	7,5
		7,1	7,0

Oletame, et tahame võrrelda kõrge ning madala stressiga immigrantlike ning mitteimmigrantlike (kultuuriline taust) gruppide psühholoogilist seisundit (ärevus, kartus) mõõdetuna sotsiaalse kohanemise skaalal. Peale skaala rakendamist suurele linlaste üldkogumile valitakse juhukorras välja võrdne arv inimesi, keda iseloomustab stressi- ning mittestressilaadne eluviis. Valim moodustatakse immigrantide ja migrantide areaalist ning viiakse läbi ärevuse mõõtmise test.

Kahepoolne dispersioonanalüüs lubab meil vastata reale küsimustele:

- Kas immigrandid erinevad oluliselt mitteimmigrantidest oma ärevustasemelt?
- Kas stress mõjutab ärevust?
- Kas kõrge stressiga immigrantide juures võib täheldada enam ärevustundeid kui madala stressiga immigrantide puhul?
- Jne.

Iga püstitatud küsimuse osas väidab nullhüpotees (H_0), et ei stressitase ega kultuuriline taust ega ka mingi nimetatud kahe faktori vaheline koosmõju ärevust ei mõjuta.

- 1) H_0 : immigrantide ja mitteimmigrantide (faktor B) ärevustulemuste aritmeetilised keskmised oluliselt ei erine.
- 2) H_0 : erinevatele stressiastmetele vastavate ärevustulemuste aritmeetilised keskmised oluliselt ei erine.
- 3) H_0 : stressi ja kultuuritausta koosmõju ärevustulemuste aritmeetilised keskmised oluliselt ei erine (st. stressi ning kultuurilise tausta ühismõjud ei avalda toimet ärevusele).

Kogudispersiooni võime jagada:

a) SÜSTEEMAATILISED MÕJUD

Rühmade vaheline dispersioon:

- 1) FAKTOR A - stressiastme mõju dispersioonile,
- 2) FAKTOR B - kultuuritausta mõju dispersioonile,
- 3) A ja B koosmõju - stressiastme ning kultuuritausta koosmõju dispersioonile.

b) VEAMÕJUD

Rühmasisene (jääk-) dispersioon on subjektide erinevustest rühmas ning kontrollimata faktoritest põhjustatud dispersioon.

Tabel. Dispersioonanalüüsi tabel

	df	SS	M_{SS}	F	F_{KR} (0,05)
Faktor A	1	13,51	13,51	190,23	4,75
Faktor B	1	0,60	0,60	8,47	4,75
Faktorite A ja B koostoime	1	0,68	0,68	9,60	4,75
Jääkdispersioon	12	0,85	0,07		
Kokku	15	15,64			

- 1) Hülgame H_0 ja järeldame, et migrantide ja mittemigrantide ärevustulemused erinevad oluliselt, so.

- immigrantide ärevustulemused on kõrgemad kui mitteimmigrantidel.
- 2) Hülgame H_0 ja järeldame, et kõrge stressiastmega subjektid omavad kõrgemaid ärevustulemusi kui madala stressiastmega subjektid.
 - 3) Hülgame H_0 ja järeldame, et stressiastme ning kultuuri ühistoime mõjutab oluliselt ärevustulemusi.

Me ei tea, milliste rühmade aritmeetilised keskmised on oluliselt erinevad. Teame vaid seda, et vähemalt kaks neist seda on. Siin ongi meil abiks Tukey kriteerium.

Tabel. Keskmiste erinevused

	Keskmine	M2	M3	M4
1. Rühm: Kõrge stressiga immigrantid	M1=9,48	*	*	*
2. Rühm: Kõrge stressiga mitteimmigrantid	M2=8,68		*	*
3. Rühm: Madala stressiga immigrantid	M3=7,23			
4. Rühm: Madala stressiga mitteimmigrantid	M4=7,25			

* - oluline nivool 0,01

Analüüsid keskmiste erinevuste tabelit, võime teha järgmised järeldused:

- a) kõrge stressiastmega immigrantid saavad kõrgemaid ärevustulemusi kui kõrge stressinäitajaga mitteimmigrantid.
- b) kõrge stressiastmega immigrantid saavad kõrgemaid ärevustulemusi kui madala stressiga immigrantid.
- c) kõrge stressinäitajaga mitteimmigrantid saavad kõrgemaid ärevustulemusi kui madala stressiga immigrantid.
- d) madala stressiastmega immigrantid ei saa oluliselt suuremaid ärevustulemusi kui madala stressiga mitteimmigrantid.