

Dispersioonanalüüs (ANOVA)

Dispersioonanalüüsi eesmärk on kontrollida gruppidevaheliste erinevuste statistilist olulisust. Erinevalt t-testist **võimaldab** ta **võrrelda enam kui kahte gruppi nende keskväärtuste põhjal**. Seega dispersioonanalüüsi kõige lihtsam mudel (One-Way ANOVA)¹ eeldab kahte tunnust: üht kategoriaalset tunnust, mis kirjeldab võrreldavad grupid ning üht arvtunnust, mille osas grupe võrrelda tahetakse.

Kui tegemist on kahe tunnusega, millest vähemalt üks on kategoriaalne, siis on sisult sama küsimust võimalik sõnastada nii erinevuse- kui seoseküsimusena (näiteks küsimus *Kas Tallinna, Tartu ja Pärnu õpilaste tulemused erinevad?* on ligilähedaselt samaväärne küsimusega *Kas õpilaste tulemused on seotud asukohaga?* või *Kas asukoht on õpilaste tulemust mõjutavaks faktoriks?*). Viimasest lähtuvalt nimetatakse dispersioonanalüüsis grupeerivat tunnust tihti ka **faktoriks** või faktortunnuseks.²

Dispersioonanalüüsi keerulisemate mudelite puhul võib lisanduda nii grupeerivaid kui ka uuritavaid tunnuseid. Kui grupeerivaid tunnuseid on enam kui üks (n uurijaid huvitab soo ja rahvuse koosmõju), siis räägitakse vastavalt kas kahe-faktorilisest, kolme-faktorilisest või jne dispersioonanalüüsist.³ Kui aga korraga soovitakse vaadelda mitut sõltuvat tunnust (n testi erinevate alamosade tulemusi või erinevaid sotsiaalseid parameetreid), siis räägitakse mitmemõõtmelisest dispersioonanalüüsist (MANOVA).⁴ Juhul kui võib oletada, et leidub selliseid tegureid, millel on sõltuva(te)le tunnus(t)ele märkimisväärne kõrvalmõju võib tekkida vajadus selle mõju statistiliseks kõrvaldamiseks mudelist, et mudel peegeldaks selgemini meid huvitava faktori mõju uuritava(te)le tunnus(t)ele. Sel juhul on tegemist kovariatsioonanalüüsiga (ANCOVA ja MANCOVA).⁵

Edaspidi peatume ainult dispersioonanalüüsi lihtsaimal mudelil.

Eeldused:

- * sõltuv tunnus peab olema arvtunnus (arvskaalaga samaväärseks võib pidada võrdsete vahemikega järjestusskaalat);
- * võrreldavad grupid peavad olema omavhel sõltumatud;
- * tulemuste hajuvused (dispersioonid) võrreldavates gruppides peavad olema sarnased;
- * tulemuste jaotus peaks olema ligilähedane normaaljaotusele.⁶

Hüpoteesid:

- $H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_n$ (alamkogumite keskväärtused ei erine)
 $H_1: \exists \mu_i \neq \mu_j$ (leidub vähemalt kaks alamkogumit, mille keskväärtused erinevad)

Olulisuse nivoo:

- $\alpha = \dots$ määrata, kui väike peab olema nullhüpoteesi kehimise tõenäosus, et võiks ta ümber lükata.

¹ ingl ANOVA = Analysis of Variance; One-Way ANOVA = ühefaktoriline dispersioonanalüüs

² Ka faktoranalüüsi puhul kasutatakse nimetust *faktortunnus*, mis ei ole aga samas tähenduses nagu dispersioonanalüüsi puhul.

³ Two-Way ANOVA, Three-Way ANOVA, ...

⁴ ingl MANOVA = Multiple Analysis of Variance

⁵ ANCOVA = Analysis on Covariance; MANCOVA = Multiple Analysis of Variance

⁶ selle eelduse mittetäidetust mõjutab analüüsi tulemust üldjuhul vähe.

Nullhüpoteesi kehtimise tõenäosuse hindamine:

Dispersioonanalüüs tegeleb niisiis keskväärtuste erinevuse hindamisega – oma nime on ta aga saanud tuntud hajuvuse näitaja järgi seetõttu, et keskväärtusi puudutava nullhüpoteesi kehtimise tõenäosust hinnatakse valmite hajuvuste kaudu. **Eeldusel, et nullhüpotees kehtib**, peaksid kõik võrreldavad grupid olema juhuslikud valimid ühest ja samast üldkogumist, seega peaks nende valmite hajuvuste põhjal saama hinnata selle üldkogumi hajuvust. Konkreetseks hajuvuse näitajaks, mida siin kasutatakse, on dispersioon. Üldkogumi hajuvust/dispersiooni saab aga hinnata kahel erineval viisil (vt ka joonis 1):

1. kuna juhusliku valimi dispersioon peaks olema ligilähedane üldkogumi dispersioonile, siis saame ühe hinnangu üldkogumi dispersioonile arvutades kõigi valimite dispersioonide (kaalutud) keskväärtuse.
Kuna sel viisil arvatud dispersiooni hinnang lähtub valimite sisemisest hajuvusest, siis võimegi rääkida, et oleme saanud **hinnangu üldkogumi dispersioonile valimite sisese dispersiooni põhjal**.⁷
2. teiselt poolt saame aga arutleda nii, et kui kõik valimid on juhuslikud valimid ühest üldkogumist, siis peaksid kõigi valimite jaotused sisalduma üldkogumi jaotuses ehk teisisõnu - kõik erinevatesse gruppidesse kuulvad objektid peaksid “ära mahtuma” üldkogumi jaotust iseloomustava jaotuskõvera alla. Selleks, et hinnata üldkogumi dispersiooni tuleks nüüd kõigepealt arvutada valimite keskväärtuste vaheline hajuvus (nii saame kätte keskväärtuse standardvea) ning seejärel, kasutades standardvea valemit, saame leida ka hinnangu üldkogumi dispersioonile. Kuna antud juhul lähtusime me gruppide vahelistest erinevustest, siis oleme nüüd saanud **hinnangu üldkogumi dispersioonile valimite vahelise dispersiooni põhjal**.

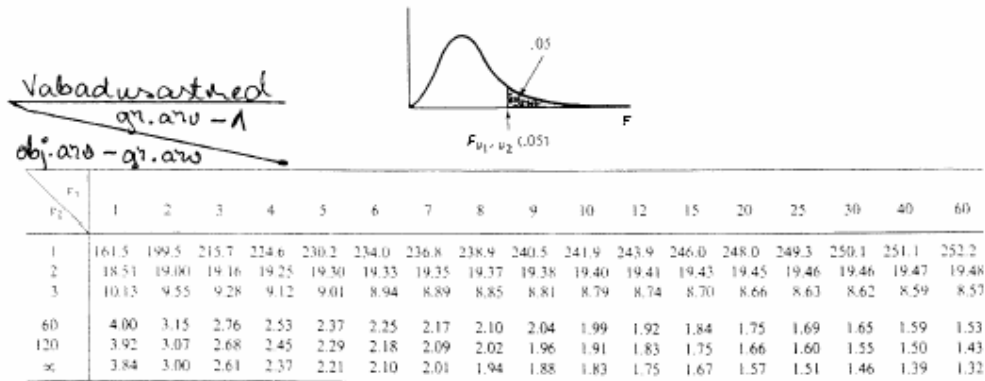
Saadud kaks hinnangut peaksid nullhüpoteesi kehtimise korral olema ligilähedased (erinevus võiks olla vaid statistilise vea piires). Siit edasi arutledes näeme, et kui need kaks hinnangut on aga väga erinevad, siis järelikult ei saa nullhüpotees kehtida ja me võime ta ümber lükata. Kahe arvnäitaja erinevuse hindamiseks võib kasutada lahutamistehet, samuti aga jagamistehet, mis annab meile suhtelise erinevuse. Viimast viisi dispersioonanalüüsi puhul kasutataksegi. Seega erinevuse suurust väljendab statistik F, mille saame:

$$F = \frac{s_{vv}}{s_{vs}} \quad \text{ehk} \quad \left(F = \frac{\text{hinnang üldkogumi dispersioonile valimite vahelise dispersiooni põhjal}}{\text{hinnang üldkogumi dispersioonile valimite sisese dispersiooni põhjal}} \right)$$

Kui erinevust ei ole (ja seega meie nullhüpoteesi kehtimise tõenäosus on suur), siis on F-i väärtus ligikaudu 1. Näeme, et mida suurem on valimite keskväärtuste vaheline erinevus, mis väljendub valimite vahelises dispersiooni hinnangus, seda suurem on ka erinevuse näitaja F. Nüüd tekib aga viimane tähtis küsimus: *Kui suur võib olla juhuslik viga kui nullhüpotees kehtib? ehk Kui suur peab olema erinevus, et me võiksime oma nullhüpoteesi ümber lükata?*

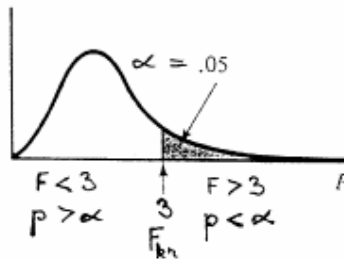
⁷ Siin peitub ka põhjus, miks dispersionanalüüsi eelduseks on, et valmite hajuvused peavad olema sarnased.

Selle hindamiseks peame konstrueerima teoreetilise jaotuse kõikvõimalikest erinevustest (F -idest), mis võivad tekkida juhuse tõttu kui meie eeldus e nullhüpoteesi kehtib. Selle jaotuse täpne kuju sõltub vabadusastmete arvust e antud juhul võrreldavate valimite arvust ning vaadeldavate objektide arvust. Jaotuse üldkuju on toodud joonisel 2.



Joonis 2. F -i kriitilise väärtuse saab kätte vastavatest tabelitest.

Näiteks kui võrreldavaid grupe on kolm ja vaadeldavaid objekte 200, siis on kriitiliseks väärtuseks $F_{kr} \approx 3$. Jooniselt 3 näeme, et kui meie tegelik erinevus (F) on kriitilisest väärtusest väiksem, siis on tõenäosus nullhüpoteesi kehtimiseks (p ehk Sig) suur ja me ei saa nullhüpoteesi ümber lükata; kui aga meie erinevus (F) on kriitilisest väärtusest suurem, siis on nullhüpoteesi kehtimise tõenäosus (p ehk Sig) väike ja seetõttu võime ta ümber lükata ning sel juhul oleme tõestanud valitud olulisuse nivool alternatiivse hüpoteesi kehtimise.



Joonis 3. F -i kriitiline ja tegelikud väärtused kolme grupi ja ≈ 200 objekti korral.

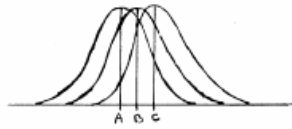
Kasutades arvutitarkvara nullhüpoteesi kehtimise hindamiseks, ei ole vajadust F -i kriitilise väärtuse otsimiseks tabelitest kuna arvutipaketid arvutavad meile täpse nullhüpoteesi kehtimise tõenäosuse (p ehk Sig).⁸ Seega peaksime otsuse langetamiseks nüüd võrdlema valitud olulisuse nivood (α) ning saadud olulisuse tõenäosust (p ehk Sig).

Otsus nullhüpoteesi ümberlukkamise kohta:

- $p > \alpha$ H_0 jääb kehtima
- $p \leq \alpha$ H_0 ümber lükata ja seega kehtib H_1

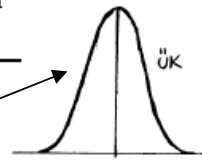
⁸ Käsitsi oleks selle arvutamine liiga tülikas ja aeganõudev.

N1. Kolm sarnast valimit



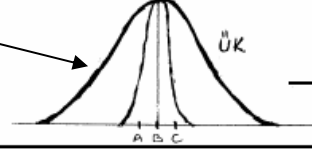
HINNANG ÜLDKOGUMI
DISPERSIOONILE
KAHEL VIISIL:

VALIMITE
SISESE DISP.
PÕHJAL



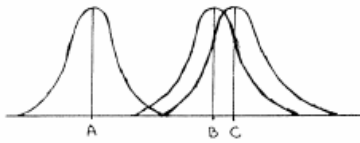
KAKS HINNANGUT
ERINEVAD VÄHE!

VALIMITE
VAHELISE
DISP.
PÕHJAL



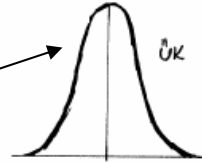
VALIMI
KESKVÄÄRTUSTE
JAOTUS

N2. Üks valim teisest erinev



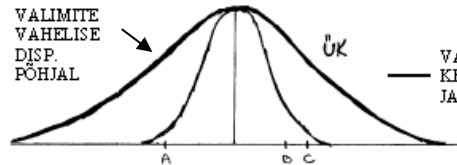
$$\text{st.viga} = \frac{\text{st.hälve}}{(\text{st.hälve}) \sqrt{N}}$$

VALIMITE
SISESE DISP.
PÕHJAL



KAKS HINNANGUT
ERINEVAD PALJU!

VALIMITE
VAHELISE
DISP.
PÕHJAL



VALIMI
KESKVÄÄRTUSTE
JAOTUS

SPSS'i kasutamine dispersioonanalüüsi tellimiseks ja tulemuste tõlgendamine:

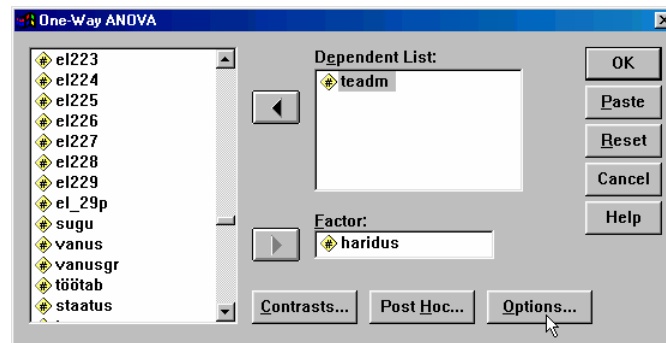
Näide: võrdleme erineva haridustasemega inimeste EL'i alaste teadmiste taset

H_0 : teadmiste tase alamkogumites ei erine

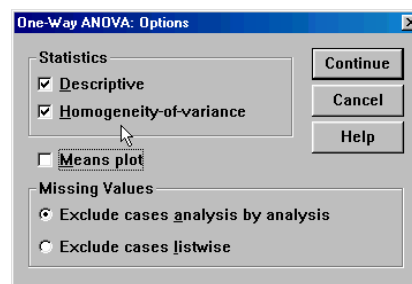
H_1 : leiduvad vähemalt kaks gruppi, mis esindavad erineva teadmiste tasemega üldkogumeid

Valime $\alpha = 0,05$

- Vali *Analyze/Compare Means/One-Way ANOVA...*
- Paiguta grupeeriv e sõltumatu tunnus väljale *Factor*:
- Paiguta uuritav e sõltuv tunnus väljale *Dependent List*:
(kui määrad enam kui ühe sõltuva tunnuse, siis viiakse iga tunnuse jaoks läbi eraldi analüüs)



- Kui soovid tellida kirjeldavaid arvnäitajaid võrreldavate gruppide kohta ning testi kontrollimaks gruppide hajuvuste erinevuse statistilist olulisust (vt ANOVA eeldusi!), siis vajuta nuppu *Options...* ning märgista vastavalt väljad *Descriptive* ja/või *Homogeneity-of-variance*

**Kirjeldavad arvnäitajad:****Descriptives**

Teadmiste indeks (max 3p)

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Alg- või põhiharidus	179	.9011	.9054	6.762E-02	.7677	1.0345	.00	3.00
Keskharidus	261	1.2522	.9084	5.618E-02	1.1416	1.3628	.00	3.00
Kutse- v. keskeriharidus	341	1.2269	.8469	4.587E-02	1.1366	1.3171	.00	3.00
Kõrgharidus	152	1.6182	.8841	7.168E-02	1.4766	1.7598	.00	3.00
Total	934	1.2352	.9059	2.965E-02	1.1770	1.2933	.00	3.00

Näeme, et valimis on võrreldavate gruppide keskmine teadmiste tase (*Mean*) erinev, teadmiste hajuvus (*Std. Deviation*) aga suhteliselt sarnane.

Levene'i test gruppide dispersioonide võrdlemiseks (ANOVA eeldus!):

Hüpoteesid:

$H_0: \sigma_1 = \sigma_2 = \sigma_3 = \dots = \sigma_n$ (alamkogumite dispersioonid ei erine)
 $H_1: \exists \sigma_i \neq \sigma_j$ (leidub vähemalt kaks alamkogumit, mille dispersioonid erinevad)

Olulisuse nivoo:

$\alpha = \dots$ tavaliselt 0,05 eelduse kontrollimiseks
 läbiviidavate testide puhul

Test of Homogeneity of Variances

statistik, mis väljendab erinevuse suurust		Teadmiste indeks (max 3p)				olulisuse tõenäosus p	
		Levene Statistic	df1	df2	Sig.		
		2.182	3	930	.089		

Kuna $p > \alpha$, siis jääme nullhüpoteesi juurde ning võime ANOVA eeldust hajuvuste sarnasuse kohta täidetuks pidada.

ANOVA põhitulem:

ANOVA

		Teadmiste indeks (max 3p)				
		Sum of Squares	df	Mean Square	F	Sig.
gruppide vaheline	Between Groups	42.429	3	14.143	18.172	.000
gruppide sisene	Within Groups	723.020	929	.778		
Total		765.448	932			

keskruut e üldkogumi dispersiooni hinnang

olulisuse tõenäosus p

Kuna saadud hinnangud üldkogumi dispersioonile erinevad märkimisväärselt ning ühtlasi

$p < \alpha$, siis võime H_0 ümber lükata ja tõestatuks lugeda sisuka hüpoteesi H_1 . Seega võime väita, et vähemalt kahel grupil võrreldavast neljast on teadmiste tase erinev statistiliselt olulisel määral.

Kirjeldavate arvnäitajate põhjal näeme, et kõige suurem erinevus on alghariduse ja kõrgharidusega inimeste teadmiste tasemetes, samas jääb keskharidusega inimeste teadmiste tase sinna vahele ning me ei oska praeguste tulemuste põhjal öelda, kas ka erinevused kesk- ja algharidusega inimeste ning kesk- ja kõrgharidusega inimeste tasemetes on statistiliselt olulised. Selle kindlakstegemiseks tuleks tellida lisatestit paariviisiliseks võrdluseks.

- Paariviisiliste võrdlustestide tellimiseks vajuta ANOVA põhiaknas nuppu *Post Hoc...* ning vali sobiv test.

Kastist *Equal Variances Assumed* leiad terve hulga teste, mis sobivad juhul kui gruppide hajuvusi võib sarnaseks pidada. Nende testide hüpoteesid on samad, mis t-testide puhul, kuid erinevus peitub selles, valitud olulisuse nivood arvestatakse

kogu testide komplekti kohta ja mitte iga üksiku testi jaoks eraldi (see vähendab juhusliku vea tõenäosust).⁹

Kastist *Equal Variances Not Assumed* leiad sobivad testid juhaks kui gruppide hajuvused on erinevad statistiliselt olulisel määral.

One-Way ANOVA: Post Hoc Multiple Comparisons

Equal Variances Assumed

☐ LSD ☐ S-N-K ☐ Waller-Duncan

☒ Bonferroni ☐ Tukey Type I/Type II Error Ratio: 100

☐ Sidak ☐ Tukey's-b

☐ Scheffe ☐ Duncan Control Category: Last

☐ R-E-G-W F ☐ Hochberg's GT2

☐ R-E-G-W Q ☐ Gabriel Test: 2-sided < Control > Control

Equal Variances Not Assumed

☐ Tamhane's T2 ☐ Dunnett's T3 ☐ Games-Howell ☐ Dunnett's C

Significance level: .05

Continue Cancel Help

Paariviisilised võrdlused:

Hüpoteesid:

$H_0: \mu_i = \mu_j$ (kahe võrreldava alamkogumi keskvaartused ei erine)

$H_1: \mu_i \neq \mu_j$ (kahe võrreldava alamkogumi keskvaartused erinevad)

Olulisuse nivoo:

$\alpha = \dots$ valida

keskväärtuste erinevus valimis

olulisuse tõenäosus p

Multiple Comparisons

Dependent Variable: Teadmiste indeks (max 3p)

Bonferroni

(I) Haridus	(J) Haridus	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Alg- või põhiharidus	Keskharidus	-.3511*	8.554E-02	.000	-.5772	-.1249
	Kutse- v. keskeriharidus	-.3257*	8.139E-02	.000	-.5409	-.1106
	Kõrgharidus	-.7171*	9.724E-02	.000	-.9742	-.4600
Keskharidus	Alg- või põhiharidus	.3511*	8.554E-02	.000	.1249	.5772
	Kutse- v. keskeriharidus	2.534E-02	7.252E-02	1.000	-.1664	.2171
	Kõrgharidus	-.3660*	8.996E-02	.000	-.6039	-.1282
Kutse- v. keskeriharidus	Alg- või põhiharidus	.3257*	8.139E-02	.000	.1106	.5409
	Keskharidus	-2.5337E-02	7.252E-02	1.000	-.2171	.1664
	Kõrgharidus	-.3914*	8.602E-02	.000	-.6188	-.1639
Kõrgharidus	Alg- või põhiharidus	.7171*	9.724E-02	.000	.4600	.9742
	Keskharidus	.3660*	8.996E-02	.000	.1282	.6039
	Kutse- v. keskeriharidus	.3914*	8.602E-02	.000	.1639	.6188

*. The mean difference is significant at the .05 level.

*. Erinevus on statistiliselt oluline nivool 0,05

Paariviisilistest võrdlustest näeme, et erinevus on statistiliselt oluline kõigi erinevate haridustasemete vahel, kusjuures erinevat tüüpi keskharidust omavate inimeste puhul teadmiste tase ei erine statistiliselt olulisel määral (erinevus sisuliselt puudub ka valimis).

⁹ Erand on siin LSD test, mille kasutamist võiks seetõttu vältida.