E-kursuse "Kvantitatiivne andmeanalüüs (SPSS'iga)" materjalid

Aine maht 6 EAP

Anu Masso (Tartu Ülikool), 2011

TARTU ÜLIKOOL
Ajakirjanduse ja kommunikatsiooni instituut

Kvantitatiivne lähenemine. Esmane ühe- ja mitmemõõtmeline analüüs.

Kvantitatiivne andmeanalüüs

Anu Masso (PhD)

Kvantitatiivne lähenemine

- Kvantitatiivne lähenemine sotsiaalteadustes tekkis tänu arengutele erinevates teadusvaldkondades.
  - 16.saj. tõenäosusteooria; 17.saj. rahvastikustatistika; 19.saj. bioloogia, korrelatsiooni ja regressiooni, normaaljaotuse mõisted (Francis Galton); 19.saj. sotsiaalteadused, nähtuste omavaheline statistiline seos (nt Quetelet) jms.
- Sotsiaalteadustes kasutatakse mõistet andmeanalüüs: numbriliste andmete kogumine, korrastamine ja tõlgendamine.
  - Enamasti soovitakse valimit kasutades teha järeldusi populatsioonile; järelduste tegemisel kasutatakse matemaatilise statistika meetodeid.
  - Numbrid väljendavad teatud teoreetilise konstrukti või kontsepti väärtuseid või tasemeid; numbreid kasutatakse nähtuse tõlgendamisel.

Koolkonnad

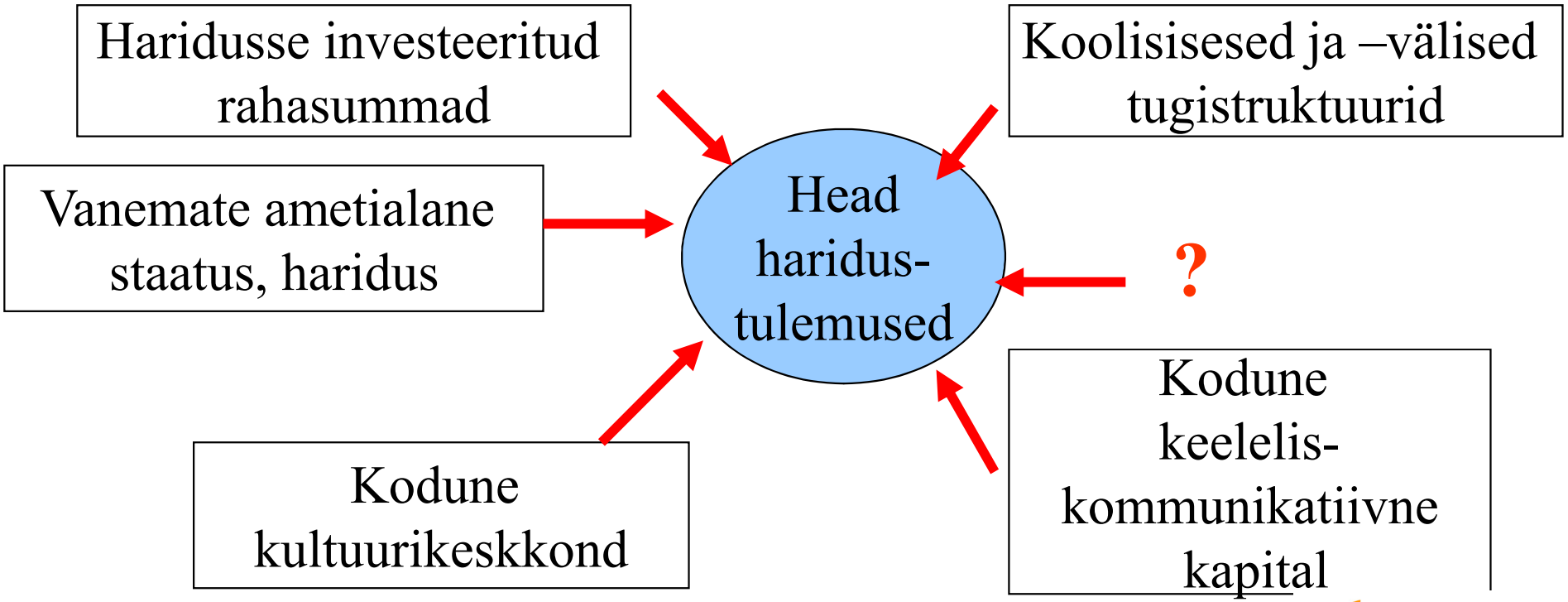
- Algsete positivistlike lähenemiste kõrvale on aegamööda tekkinud interpretatiivsed jt lähenemised.
  - Positivistliku lähenemise järgi peaks teaduslikku teooriat olema võimalik empiiriliselt ümber lükata või kinnitada (vt nt Karl Popper). Teaduslik teadmine on võimalik vaid läbi induktsiooni. Objektiivne lähenemine: *Kas on võimalik vastu võtta hüpotees, et nähtuste vahel on seos?*
  - Post-positivistliku epistemioloogia (nt Cook, Campbell, Lakatos) järgi esksiteerib maailm ka meie tajudest väljaspool, teadus peaks püüdlema selle mõistmise poole. Selleks tuleb nähtust mõõta erinevatel viisidel (triangulatsioon). Teadmine saadakse deduktsiooni ja induktsiooni kombinatsioonis.
  - Interpretatiivse lähenemise järgi võib kvantitatiivne lähenemine anda unikaalse panuse nähtuse uurimisel. Positivistlik lähenemine vähendab uurija ja tema tõlgenduste osa. Eesmärgiks pole vaid hüpoteeside testimine, vaid tõese mudeli vms leidmine (vt nt Taagepera). *Kuidas selgitada variatiivsust andmetes?*

Kvantitatiivsed uuringud Eestis

- TÜ ajakirjanduse ja kommunikatsiooni osakonnas: Eesti elanike esinduslik küsitlus “Mina.Maailm.Meedia” (2003, 2005, 2008).
- Eesti Sotsiaalteaduslik Andmearhiiv (sisaldab andmestikke, ankeete, jm mitte-elektroonilisi materjale); http://psych.ut.ee/esta (/ankeedid)
  - 1958-1965 embrüonaalne periood Nõukogude sotsioloogias, st esimesed regulaarsed meediauuringud (1965-66 Edasi lugejaskond).
  - 1965-1972 kuldajastu Nõukogude sotsioloogias; 1973-1975 repressioonide aeg, ideoloogiline puhastustöö; 1975-1980 lõpuni stagnaaeg.
  - 1990.a-tel rakenduslike uuringute kommertsialiseerumine, st turu- ja avaliku arvamuse uuringu firmade tekkimine (P.Vihalemm 2004).

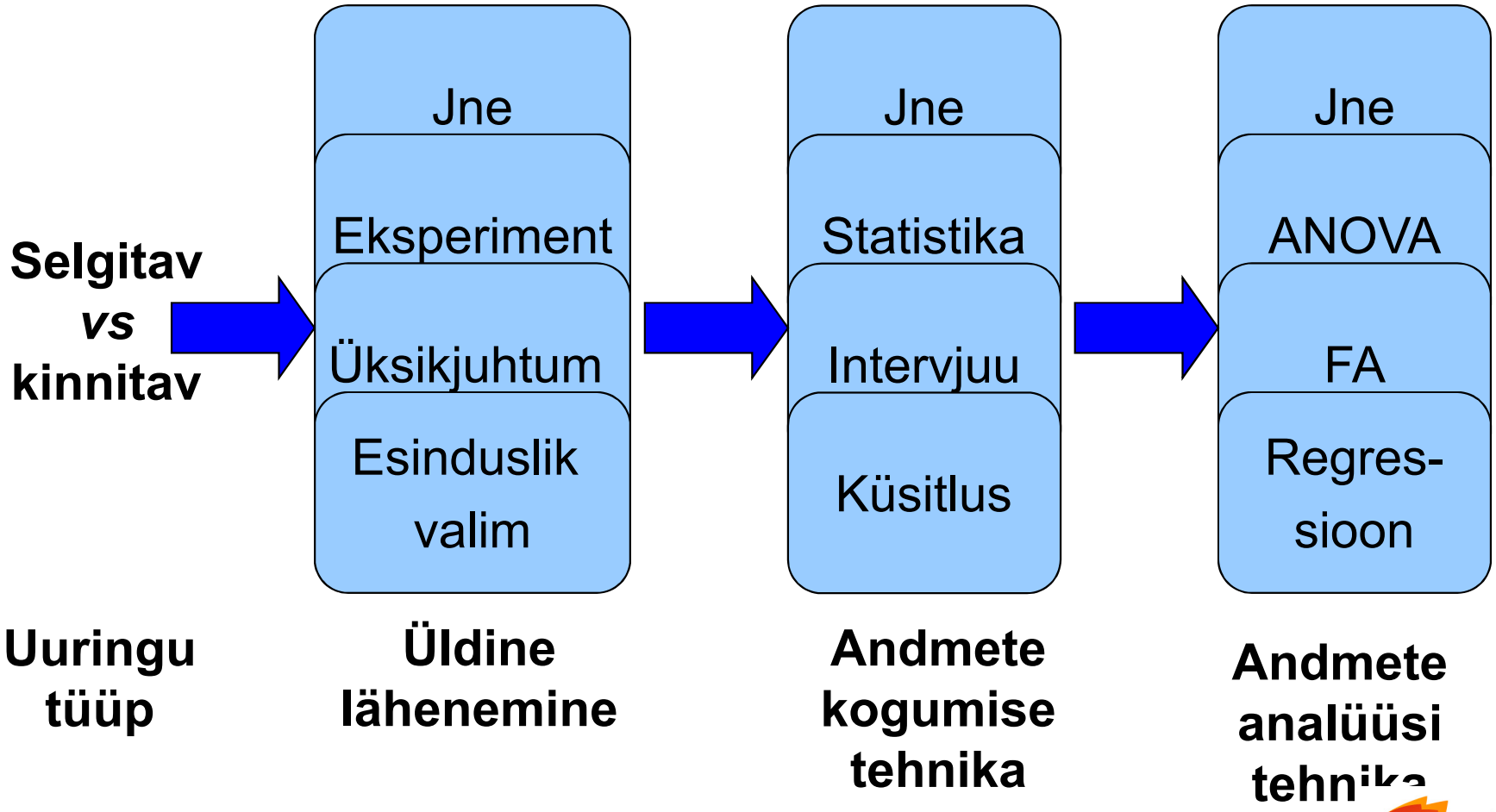
Kvantitatiivne probleem

- Kvantitatiivne uuring algab uurimisprobleemi, indikaatorite ja mõistete süsteemi määratlemisest.
  - *PISA uuringute järgi Soome haridussüsteem edukaim, st õpilastel parimad tulemused matemaatikas, ka üldine keskmine tase kõrge. Millest tuleneb haridussüsteemi edukus?*



Andmete kogumine

- Uuringutüübi ning üldise lähenemise valik pole otseselt seotud konkreetse andmete kogumise või analüüsi tehnika valikuga.



Andmestik I

- Kvantitatiivsed andmed saadud küsitluse, loendamise, vaatluse, mõõtmise vms teel.
- Analüüsimiseks tuleb andmed viia standardkujule (objekt-tunnus-tabel vorm). Moodustub andmestik, mis sisaldab küsimusi ehk tunnuseid *(nt sünniaasta)* uuritavate objektide (indiviidide) lõikes.
- Tunnustel on väärtused ehk kategooriad, mis näitavad tunnustel esinevaid omadusi (tavaliselt andmestikus märgitud numbritena, *nt 1-mees, 2-naine*).

Tabel: Objekt-tunnus-tabeli näide

	v1 ojektid	v2 sünniaasta	v3 surmaasta	v4 panus
Tunnused →	Francis	1822	1911	korrelatsioon
	Karl	1859	1936	sobitusaste
Objektid ↓	Ronald	1890	1962	olulisuse test

Andmestik II

- Longituuduuringute korral on ühe indiividi kohta andmestikus kaks rida, st sama indiviid on samadele küsimustele vastanud eri ajahetkedel.
- Stratifitseeritud juhuvalimi korral on uuritavad võetud juhuslikkuse alusel uuringu seisukohalt oluliste elanikkonnagruppide klastritest *(nt haridus, vanus, aga ka riigid, koolid vms)*

Tabel: Objekt-tunnus-tabeli näide

Tunnused →

Objektid ↓

V1 indiviidid	**V2 riigid**	**V3 vanus**	**V4 hinnang muutustele**
1	Eesti	15	Rõõmustavad
2	Eesti	42	Raske öelda
3	Leedu	30	Kurvastavad
4	Leedu	71	Pigem kurvastavad
5	Rootsi	59	Rõõmusta

Mõõteskaalad I

- Andmete kogumise viisi aluseks mõõteskaala määramine, st reeglid, mille kohaselt uurimisobjekti omadused seatakse vastavusse arvuliste vm formaliseeritud väärtustega.
  - Tunnused erinevad selle poolest, kui "hästi" neid on võimalik mõõta, st ku palju informatsiooni on mõõteskaala kaudu võimalik saada. *Nt vanus ja sissetulek on täpsemalt "mõõdetavad" kui küsimus – kas kultuur on üldiselt Teie elus olulisel kohal?*
  - Mõõtmisviga – erinevus uuritava nähtuse tegeliku väärtuse ning mõõdetud väärtuse vahel. Saadava info hulk ning mõõtmisvea suurus sõltub uuritava objekti mõõdetavusest, aga ka valitud mõõteskaalast.
  - Mõõtmisviga juhuslik *(st mõõtmisinstrumendi piiratud täpsus, vea vähendamiseks korrata uuringut)* või süstemaatiline *(nt küsimuse ebakorrektne sõnastus).*

Mõõteskaalad II

Mõõteskaala	Kirjeldus	Näide
Nominaalne ehk kategoriaalne tunnus	Teatud nähtuse omaduste nimekiri või loend	*Naine, mees* *Kristlus, budism, islam...*
Ordinaalne ehk järjestustunnus	Loogiliselt järjestatud skaalapunktid (hinnang, meeldivus)	*Täiesti nõus, üldiselt nõus, raske öelda, pigem ei ole nõus, olen täiesti vastu* *Jah, ei (poolt, vastu)*
Intervaall ehk kvantitatiivne ehk *skaala* tunnus (SPSS)	Võrdsete vahemikega skaalapunktid	*Vanus aastates* *Sissetulek kroonides*

- NB! Eristatakse ka suhteskaalat – iseloomulik on kokkuleppelise null-punkti olemasolu (*nt kaal ja temperatuur*); enamus analüüsitehnikaid ei erista seda intervallskaalast.

- Skaalast sõltub nähtuse uurimise täpsus ning konkreetse analüüsimeetodi valik.

Skaalade teisendamine

MIKS?

- Puuduvad väärtused [*missing values*], st mõõtmis- ja sisestusvead tuleb korrigeerida (*nt liidetakse mõne olemasoleva väärtusega või jäetakse analüüsist välja*).
- Võimalik suurendada tulemuste üldistatavust, keskenduda üksikule nähtusele (väärtusele); nt liita arvuliselt tasihoidlikult esindatud kategooriad. *Nt küsimusele "kui sageli Te jälgite CNN'i" vastas vaid 5% (73 indiviidi), et jälgib iga päev; statistiliste seoste analüüsimiseks vajalik grupi suurus 100 indiviidi.*

KUIDAS?

- Vähem "ranget" (*nt järjestusskaala*) skaalat võimalik ümber teisendada "rangemale" skaalale (*nt nominaalskaala*).
- Esialgse järjestusskaala (*nt jah, ei*) teisendamine arvskaalaks võimalik vaid läbi matemaatiliste transformatsioonide.
- Nominaaltunnuste korral peavad liidetavad väärtused sisuliselt kokku sobima, järjestustunnuse korral saab ühendada vaid naaberväärtuseid.

Näide I

Tunnus	Algne skaala	Teisendatud skaala
Arvutioskus	1- ei oska üldse 2- vähene 3- rahuldav 4- hea 5- väga hea	1- oskavad (5+4+3) 2- ei oska (1+2)
Kuivõrd on Eestis probleemiks riigi vaesus	1- kindlasti mitte 2- pigem mitte 3- ei tea, raske öelda 4- võib-olla ka seda 5- seda kindlasti 0- vastamata	1- vaesus probleemiks (5+4) 2- vaesus pole probleemiks (1+2) 3- ei tea, vastamata (3+0)

Enamasti “puuduvaid väärtuseid” ei analüüsita. Kui “puuduvate väärtuste” arv on suur (u 1/3 vastajatest), võib need liita loogiliselt kokkusobiva sisulise kategooriaga (nt “raske öelda”).

Näide II

Tunnus	Nominaal-skaala	Järjestus-skaala	Intervall-skaala
Vanus	1- alla 29-aastased 2- üle 30-aastased	1- 15-19 2- 20-29 3- 30-39 4- 40-49 5- 50-59 6- 60-74	Täpne vanus
Sugu	1- mees 2- naine		Naiselikkuse või mehelikkuse indeks
Poliitiku populaarsus	1- jah 2- ei	1.koht 2.koht 3.koht	Skaala –5 kuni +5

Järjestusskaala võimaldab analüüsida äärmuslikke juhtumeid, leida nähtuse selgitamise seisukohalt kriitilised piirid *(nt kuni 29-aastased on oma eluga oluliselt enam rahul).* Teisendamine nominaalskaalale võimaldab analüüsi fokusseerida *(nt rahulolevad pensionärid).*

Ülesanne

- Mis tüüpi skaaladega on tegemist – nominaalne, ordinaalne või kvantitatiivne?
- Kuidas oleks otstarbekas mõõteskaalat teisendada?

Kuivõrd Teid huvitab informatsioon Euroopa Liidu ja selle institutsioonide tegevuse kohta (Euroopa komisjon, Europarlament jt)?

Huvitab väga 4

Mõningal määral huvitab.....3

Huvitab vähe 2

Üldse ei huvita.....1

Ei oska öelda.....5

Analüüsi käik

- Andmete puhastamine
  - Tuleb kindlaks teha, et andmed on korrektselt sisestatud ja puuduvad väärtused on korrektselt defineeritud.
- Esmane ülevaade andmetest
  - Sirvida tunnuseid andmestikus, teha esialgne kirjeldav analüüs (keskväärtuste, protsentide, jooniste, tabelite vormis).
- Koond- ehk indekstunnuste loomine
  - Sama nähtust mõõtvate tunnuste koondamine üheks tunnuseks (nt liitmise teel).
- Seoste leidmine
  - Risttabelite, jooniste tegemine aitab leida nähtuse üldised mustrid ja seosed.
- Seoste analüüsimine
  - Seoste tugevuse analüüsimiseks arvutada seosekordajad. Seoste struktuuri uurimiseks analüüsida statistilisi mudeleid (nt regressioon).
  - Järelduste tegemiseks ning tulemuste korrektseks tõlgendamiseks tuleb arvutada seoste statistiline olulisus.

Andmeanalüüsi ülesanded

- Eesmärgiks on andmetes sisalduva *variatiivsuse* kirjeldamine (kirjeldav analüüs), variatiivsuse selgitamine (selgitav analüüs) või selle prognoosimine.
- Variatiivsus – tegelikkuses olemasolevad erinevused populatsiooni või valimi indiviidide hulgas.
  - Tunnuse varieeruvuse iseloomustamisel on esimeseks sammuks tunnuse empiirilise jaotuse (protsentjaotuse) koostamine.
  - Järjestus- ja intervallskaalal tunnuste korral tuuakse jaotuse kokkuvõtlikuks iseloomustamiseks sageli esile keskmine väärtus.
  - Erisugust tüüpi tunnuste korral tuleb kahe jaotuse ühisosa analüüsimiseks enamasti suurearvulised skaalapunktid teisendada ümber väiksemasse arvu skaaladesse.
  - Analüüs on kas ühemõõtmeline (tunnuste vaatlus ükshaaval) või kahemõõtmeline (mitme tunnuse koosanalüüs); uuriv (uue andmestiku korral, puuduvad eelteadmised) või kinnitav analüüs (andmete alusel püütakse kontrollida hüpoteese).

Ühemõõtmeline analüüs

- Eesmärgiks on uuritava nähtuse süstemaatiline kirjeldamine protsentjaotuste ja keskväärtuste kaudu, tabelite ja graafikute vormis.
  - Esmase kirjeldava analüüsi käigus vaadeldakse vaid üksiktunnuseid, jäetakse kõrvale tunnuste omavahelised seosed. Analüüsitehnikate valik sõltuvalt tunnuse skaalast.
  - Ühemõõtmelise analüüsi eesmärgiks võib olla esmane analüüs, *nt sisestusvigade leidmine andmestikus, edasiseks analüüsiks oluliste üldtendentside ja andmeteisenduste vajaduse väljaselgitamine.*
  - Ühemõõtmeline analüüsi eesmärgiks võib olla ka nö lõplik analüüs, *nt Interneti kasutajate protsendi väljaselgitamine, töötusmäära leidmine küsitluse teel lisaks ametkondlikule statistikale jms.*

Keskmised I

- Aritmeetiline keskmine võimaldab suurt hulka numbrilisi andmeid koondada ja välja tuua üldtendentse.
  - Arvutamine: väärtuste summa jagatud objektide arvuga. *Nt seitsme inimese keskmise vanuse arvutamiseks liidame vanused 19+22+ 32+45+51+65+74 ja jagame 7'ga = 44.*
  - Puuduseks tundlikkus äärmuslike väärtuste suhtes, kasutatakse eelkõige väikese *hajuvuse* korral keskväärtuse suhtes. *Nt keskmine vanus 44 ei ütle midagi selle kohta, kui palju on alla 20-aastaseid.*
  - Kasutatakse intervallskaala korral; järjestusskaala korral sobilik skaala loogilise keskpunkti olemasolu korral. *Nt kumb sõnapaar iseloomustab Teie tundeid Eesti riigi suhtes:*

	Väga hästi	Kesk-miselt	Vähesel määral	Ei seda ega teist	Vähesel määral	Kesk-miselt	Väga hästi	
Kiire	1	2	3	4	5	6	7	**Aeglane**

Keskmised II

- Järjestusskaalal tunnuste jaotuse kokkuvõtlikuks iseloomustamiseks tuuakse sageli esile punkt, millest väiksemate väärtuste osa on ½ - mediaan (järjestatud nimekirja keskel asuv punkt).
  - Mediaani kasutatakse juhtumitel, mil aritmeetilise keskmise kasutamine on ebasobiv tunnuse suure hajuvuse tõttu.
  - Mediaan jaotab kogumi vaadeldava tunnuse poolest kaheks võrdsagedaseks grupiks. ***Nt 7 inimese vanuse järjestamisel 19, 22, 32, 45, 51, 65, 74 on mediaan 45 aastat.*** Paaritu arvu väärtuste korral leitakse rea keskel asuvate kahe väärtuse aritmeetiline keskmine.
  - Mediaani kasutatakse tunnuse ümberkodeerimisel nominaalskaalale. *Nt kodeerimisel moodustuks 2 võrdse indiviidide arvuga kategooriat (nooremad kui 45 ja vanemad kui 45).*

Näide I

Tabel. Usaldus riiklike institutsioonide suhtes (aritmeetiline keskmine, skaala 1-ei usalda üldse, 5-usaldan täiesti)

	2003	2005	2008
Riigikogu	2,35	2,59	2,48
President	3,37	3,46	3,25
Politsei	2,84	3,01	3,41
Pangad	3,19	3,39	3,15
Kohtusüsteem	2,72	2,95	3,03
Kultuuritegelased	3,51	3,49	3,58
Eesti Televisioon	3,52	3,65	3,43
Eesti Raadio	3,50	3,51	3,43
Ajalehed	3,15	3,09	3,06
Internetiportaalid	2,76	2,65	2,74
Ettevõtjad	2,57	2,65	2,86

Analüüs aastate lõikes näitab, et kasvanud on usaldus politseisse, veidi vähenenud usaldus presidenti.

Analüüs indikaatoreite lõikes näitab, et kõigil aastatel on kõige enam usaldatud ETV'd, ER ja kultuuritegelasi. Kõige vähem usaldatakse Riigikogu.

Allikas: Uuring Mina.Maailm.Meedia 2003, 2005, 2008.a

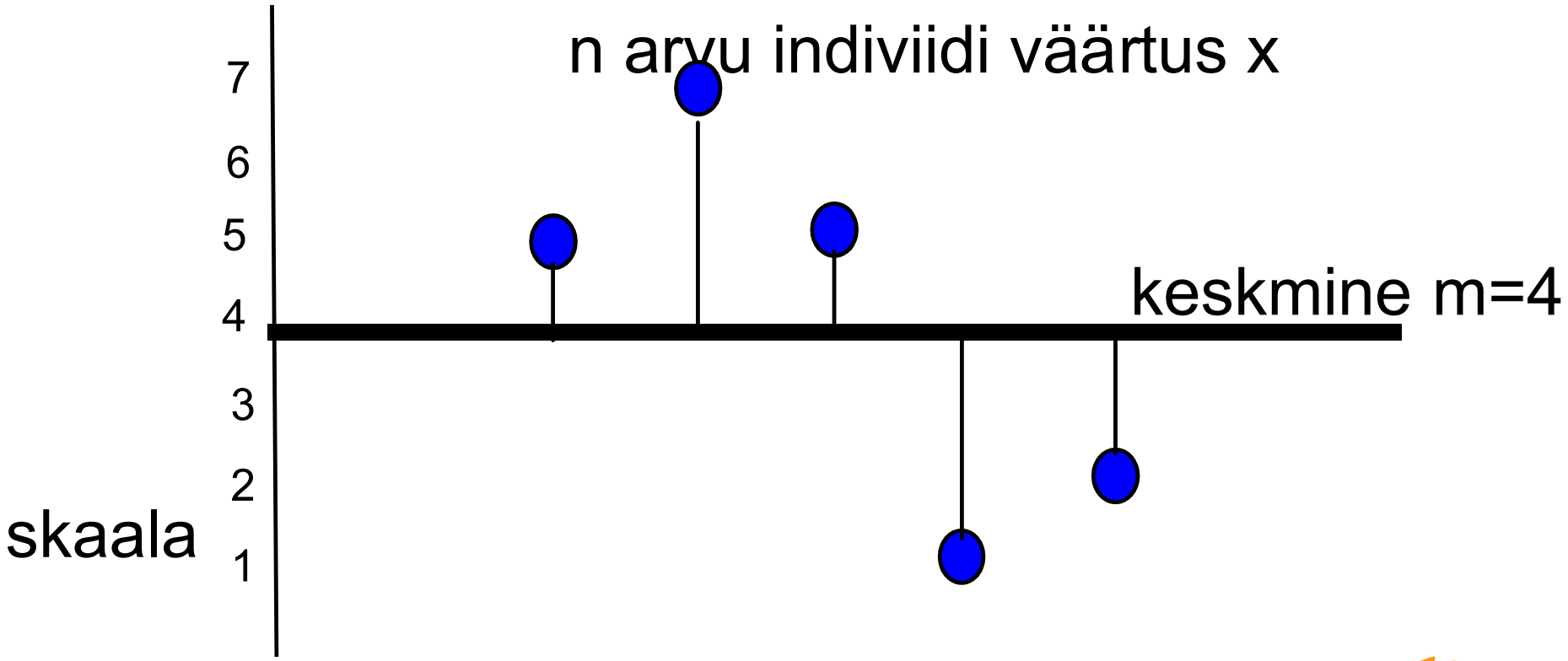
Hajuvusmõõdud

- Hajuvusmõõdud näitavad kõrvalekalde ulatust aritmeetilise keskmise suhtes; suure hajuvuse korral ei paikne üksikute indiviidide väärtused aritmeetilise keskmise lähedal.
  - Haar [*range*] on suurima ja väikseima väärtuse vahe, puuduseks – ei ütle midagi väikseima ja suurima väärtuse vahele jäävate väärtuste kohta.
  - Standardhälve ehk keskmine ruuthälve iseloomustab objektide paiknemist keskväärtuse suhtes, suur standardhälve näitab suure hulga indiviidide väärtuste erinevust keskväärtusest.
  - Dispersioon [*variance*] - standardhälbe ruut; väike dispersioon näitab suurt homogeenuse määra andmetes.
  - Standardviga [*standard error of mean*] ehk valimi standardviga näitab, kuivõib keskmine võib eri valimite lõikes erineda (võrreldakse mõõdetud keskväärtust arvutusliku hüpoteetilise väärtusega). Mida suurem on valimi grupp, seda väiksem standardviga.

Arvutamine

Standardhälbe arvutamiseks leitakse iga üksiku indiviidi väärtuse erinevus keskväärtusest, erinevuste ruudud liidetakse ning võetakse sellest ruutjuur.

$$\text{valimi st.hälve} = \sqrt{\frac{\sum (x-m)^2}{n-1}} = 2{,}45$$



Näide

Tabel. Usaldus riiklike institutsioonide suhtes (keskmine ja hajuvus, 1-ei usalda üldse, 5-usaldan täiesti)

	N	Haar	Kesk-mine	St.hälve	Disper-sioon	Keskmise st.viga
Kultuuri-tegelased	1456	4	3,58	0,83	0,69	0,02
ETV	1460	4	3,43	1,05	1,10	0,03
Politsei	1502	4	3,41	0,96	0,93	0,02
Kirik	1500	4	3,26	1,17	1,36	0,03
President	1501	4	3,25	1,25	1,55	0,03
Pangad	1499	4	3,15	0,98	0,96	0,03
Interneti-portaalid	1438	4	2,74	0,98	0,95	0,03
Riigikogu	1501	4	2,48	0,99	0,97	0,03

– Puudub “suure” või “väikese” hajuvuse piir. Olulisem samal skaalal mõõdetud tunnuste võrdlus omavahel. *Nt presidendi usalduse osas on vastused keskväärtuse suhtes kõige ebaühtlasemalt jaotunud.*

Allikas: Uuring Mina.Maailm.Meedia 2008.a

Sagedusjaotus I

- Tunnuse skaala üksikväärtuste (vastusevariantide) või väärtuste gruppide (liidetud väärtused) esinemissageduste rida vaadeldava andmekogumi alusel.
- Kasutatakse enamasti nominaal- (*nt sugu*) ja järjestusskaalal (*nt hinnang ühiskonna muutustele*) tunnuste analüüsimisel.
  - Absoluutne sagedus – indiviidide hulk absoluutarvudes iga üksikväärtuse korral.
  - Suhteline sagedus – absoluutse sageduse suhtarv indiviidide koguarvu (korrutades sajaga saame protsentjaotuse).
  - Kumulatiivne sagedus – antud väärtust mitteületava väärtuse osa kogumis, kasutatakse eelkõige suure arvu skaalapunktidega järjestustunnuste korral.

Sagedusjaotus II

Tabel. Suhtelise sageduse arvutamine

tunnuse väärtus	väärtuse sagedus	suhteline sagedus	suhteline sagegus %	kumulatiivne suhteline sagedus
a1	n1	n1/n	(n1/n) * 100	n1/n
a2	n2	n2/n	(n2/n * 100)	(n1 + n2) / n
a3	n3	n3/n	(n3/n *100)	(n1 + n2 + n3) / n
...	...	...	...	
summa	n	1	100	

-Suure valimi korral (n>100) võimaldab protsentjaotus anda parema ülevaate väärtuse esinemise sagedusest. Väikese valimi korral (n<100) eelistada suhtelist sagedust (*nt 2/3 nooremast vanusegrupist kasutab Internetti*).

-Kumulatiivse sageduse arvutamine aitab suure arvu väärtustega arvtunnuse korral leida piirid, mis on aluseks skaala jaotamisel "jämedamatesse" klassidesse (*nt 25-skaalaline indekstunnus teisendada 5-kategoorialiseks*).

Sagedusjaotus III

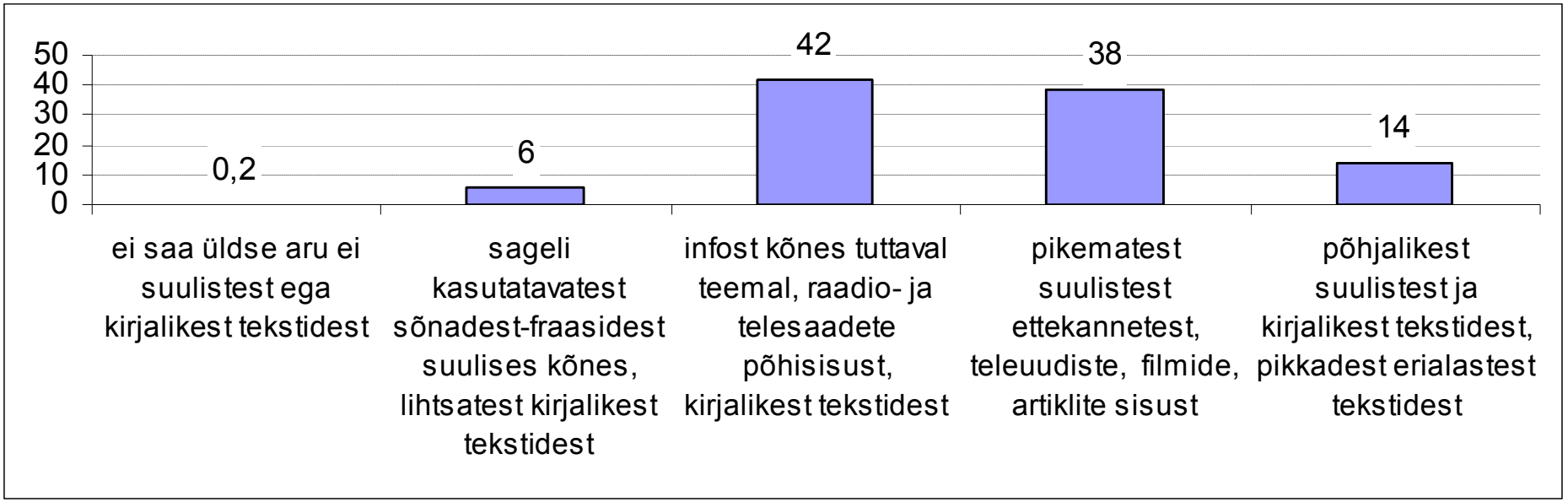
- Sagedusjaotuse keskse tendentsi väljatoomiseks kasutatakse moodi ehk kõige sagedamini esinevat väärtust.

  - Mood võib olla skaala üksikväärtus või teisendatud tunnuste korral väärtusklass *(nt arvtunnuse korral 44-aastased, nominaaltunnuse korral 20-29-aastased).*

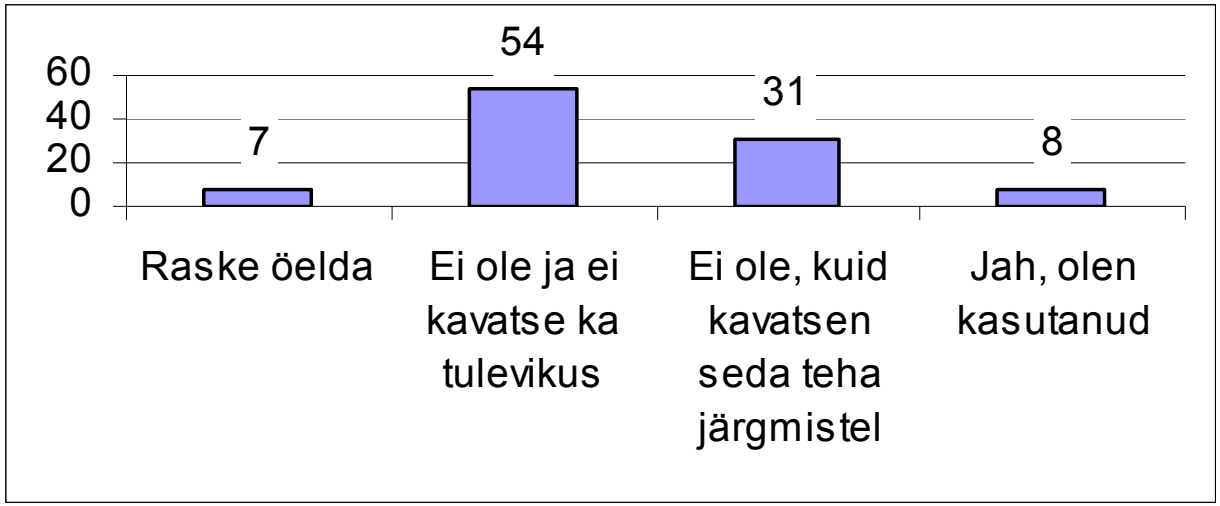
  - Moodi kasutamise puuduseks on tulemuste tõlgendamise raskused bi- ja multimodaalsete jaotuste korral (vastand unimodaalne) - mitme väärtuse võrdselt sage esinemine *(nt võrdne hulk inimesi väidavad, et meeldib / ei meeldi käia muuseumis).*

  - Jaotuse ilmekuse tõstmiset ja üldiste tendentside analüüsimist kergendab jaotuse elementide järjestamine sageduse alusel *(nt suuremast alustades, protsentide asemel võib kasutada järjekorranumbrit).*

Näide I



Joonis 1. Vene koolide abiturientide hinnang eesti keele oskuse tasemele
Allikas: Projekt "Vene laps venekeelse üldhariduskooli eestikeelses õppes" 2008.a.)



Joonis 2. Kas olete kasutanud e-valimiste võimalust?
Allikas: Uuring Mina.Maailm.Meedia 2008.a.

Bimodaalne jaotus ülemisel joonisel võib viidata ebakorrektsele ankeedile (kaks varianti vastajate jaoks raskelt eristatavad).

Alumisel joonisel on jaotus unimodaalne, st osalemine e-valimistel ja valmisolek selleks tagas

Näide II

Tabel. Majanduslanguse ohtlikkus pere olukorrale
(indeksi väärtuste sagedusjaotus)

	Sagedus	Protsent	Kumulatiivne sagedus
0	122	8	8
1	93	6	14
2	117	8	22
3	177	12	34
4	204	14	47
5	211	14	61
6	140	9	71
7	133	9	79
8	121	8	87
9	69	5	92
10	119	8	100
KOKKU	1507	100	

Allikas: Uuring Mina. Maailm. Meedia 2008

– Indeks arvutatud tunnuste alusel: *Kuivõrd ohtlikuks peate üldise majanduslangusega kaasnevat...toidu-kaupade kallinemist, bensiinihinna tõusu, eluasemekulude tõusu, töökoha kaotamise võimalust, raskusi laenude tagasi-maksmisel?*

– Kumulatiivne sagedus aitab otsustada edasiste skaala-teisenduste käiku.

Näide III

Tabel. Milliseid riike külastavad eestlased ja rootslased?

	eestlased	*rootslased*
Läti	**1** (+36)	**8** (-20)
Venemaa	**2** (+27)	**7** (-14)
Soome	**3** (+10)	**2** (+11)
Rootsi / Eesti	**4** (-1)	**6** (-9)
Saksamaa	**5** (-5)	**1** (+12)
Prantsusmaa	**6** (-20)	**3** (+10)
Inglismaa	**7** (-22)	**4** (+9)
USA	**8** (-24)	**5** (-4)
keskmine %	*24*	*29*

–Arvutatud on keskmine välismaal käimise protsent ning iga konkreetse maa külastamissageduse erinevus keskmisest (sulgudes)

– Küsimustikus samas blokis paiknevad üksikküsimused võib järjestada olulisuse alusel (järjekorranumbrid tabelis tumedas kirjas)

Allikas: Masso 2008.

Ülesanne

- **Otsida ankeedist järgmiste skaaladega tunnuseid:**
  - *Nominaalne skaala*
  - *Järjestus ehk ordinaalskaala*
  - *Arvuline ehk intervallskaala*

- **Millise ühemõõtmelise analüüsitehnika abil oleks otstarbekas nimetatud tunnuseid analüüsida?**
  - *Sagedusjaotus (absoluutarv, protsent)*
  - *Keskväärtused ja hajuvus (st.hälve)*

- **Kuidas tuleks tunnuste skaalasid teisendada, et analüüs sagedusjaotuste või keskväärtuste abil oleks korrektne?**

Mitmemõõtmeline analüüs

- Mitme tunnuse jaotuse samaaegne analüüs, eesmärgiks andmete kokkuvõtmine ning tunnustevaheliste seoste leidmine ja selgitamine.

  MILLINE MEETOD? Valik sõltub andmete iseloomust, uurimisprobleemist ja analüüsi eesmärkidest.

  - Esmase analüüsi eesmärgiks on variatiivsuse kirjeldamine (*nt risttabeli abil uuritakse, millisest soost, haridusega, sissetulekuga inimesed kasutavad enam Internetti*).
  - Seoste leidmine, selgitamine (*nt regressioonanalüüsi abil võimalik omavahel võrrelda erinevate tegurite olulisust Interneti kasutamise selgitamisel*).
  - Varjatud struktuuride leidmine, komplekssuse vähendamine andmetes (*nt klasteranalüüs võimaldab leida Interneti kasutajate tüpoloogia*).

Risttabel I

- Mitte-arvuliste ehk kategoriaalsete andmete mitmemõõtmelise analüüsi esimeseks sammuks on risttabelite (kahe tunnuse ühisjaotus) tegemine.
  - Tabeli veergudes ja ridades on tunnused, veeru ja rea ristumiskohal näidatakse tunnuste väärtuste koosesinemise sagedus.
  - Lihtsaim 2x2 risttabel (kahemõõtmeline analüüs), keerukam nt kolme tunnuse risttabel (kolmemõõtmeline analüüs).

		A		
		A1	A2	KOKKU
B	B1	a	b	a+b
	B2	c	d	c+d
	KOKKU	a+c	b+d	n=a+b+c+d

Risttabel II

- Absoluutarvuna andmete esitamisel lisatakse üldjaotused ehk tunnuste üksikute väärtuste summa ridade ja veergude lõikes *(näite tabelistes veerg KOKKU).*
- Protsentidena (veeru-, rea- või koguprotsendina) – protsentide kasutamine aitab tunnuste vahelise seose määramisel (juhul kui seos üldse esineb).

Tabel 1. Absoluutarvud

	mees	naine	KOKKU
täistööaeg	377	131	508
osaline tööaeg	10	288	298
ei tööta	12	40	52
KOKKU	399	459	858

Tabel 2. Veeruprotsent

	mees	naine
täistööaeg	94	29
osaline tööaeg	3	63
ei tööta	3	9
KOKKU	100	100

Tabel 3. Reaprotsent

	mees	naine	KOKKU
täistööaeg	74	26	100
osaline tööaeg	3	97	100
ei tööta	23	77	100

Tabel 4. Koguprotsent

	mees	naine	KOKKU
täistööaeg	44	15	
osaline tööaeg	1	34	
ei tööta	1	5	
KOKKU			100%

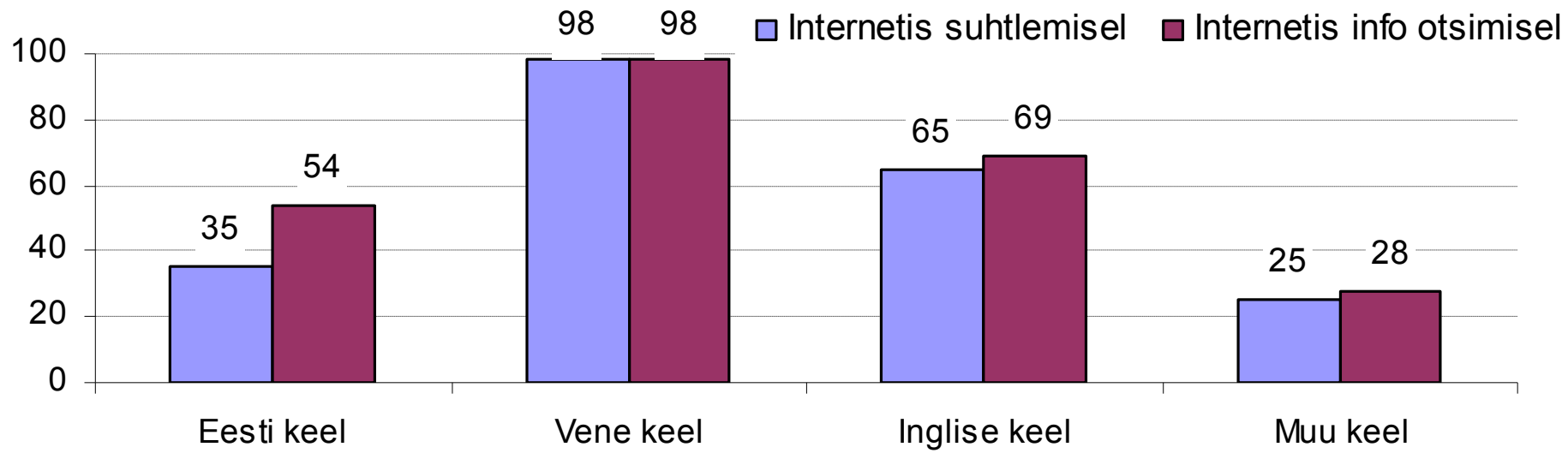
Allikas : Aronsson, Åke (1999). SPSS. En introduktion till basmodulen. Lund: Studentlitteratur.

Tõlgendamine I

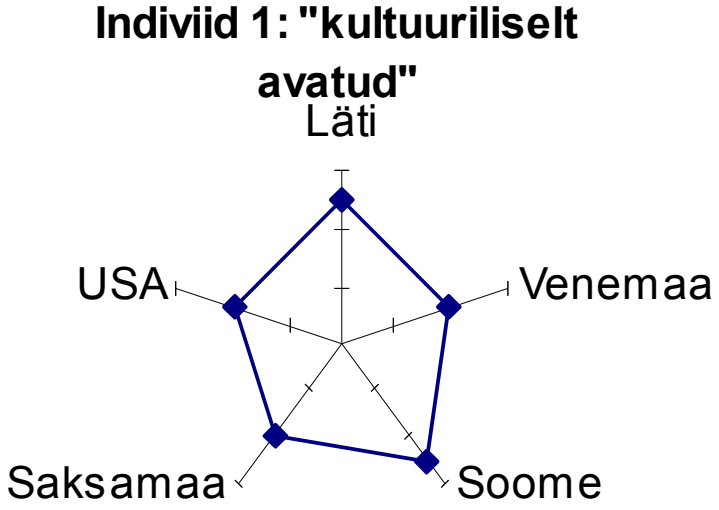
- Kas esineb andmetes teatud tendents? Kas tunnuste vahel on seos?
  - Kui erinevus protsentides (nt veeruprotsentide võrdlemisel ridade lõikes) >15, võib piisavalt suure valimi korral öelda (N>200), et tunnuste vahel on seos.
  - Mida suurem on erinevus protsentides, seda väiksem on tõenäosus, et erinevus on tingitud juhusest. Kui erinevus on 0, pole tunnuste vahel seost.

Veeruprotsent	**Reaprotsent**	**Koguprotsent**
Kas mehed töötavad täisajaga enam kui naised? *Kui suur osa meestest töötavad osalise tööajaga?*	*Kui suur osa osalise ajaga töötajatest on mehed?* *Kui palju osalise ajaga töötajatest on naised?*	*Kui palju kõikidest uuritutest on osalise ajaga töötajad ja naised?* *Kui palju kõikidest uuritutest on täisajaga töötavad mehed?*
Võrreldakse ühte veergu teise veeruga samas reas.	Võrreldakse ühte rida teise reaga samas veerus.	Kõiki lahtreid, tabeliruute võimalik üksteise

Näide I



Joonis: Võõrkeelte kasutamine Internetis (tulpdiagramm, *Allikas: HTM, vene koolide abiturientide küsitlus 2008*)



- Kahe tunnuse korral eelistada lihtsamaid graafikuid (nt tulp-, joon, hajuvusdiagramm).

- Enama kui kahe tunnuse korral kasutada ikoongraafikuid (võimaldab väikese indiviidide arvu korral esile tuua indiviidide tüpoloogiaid)

Joonis: Kultuurilise avatuse tüpoloogia
(tähediagramm, *Allikas: Mina.Maailm.Meedia 2003a.*)

Näide II

% within Vanus kolmene

		Millal te viimati Internetti kasutasite			
		Pole kunagi arvutit kasutanud	Viimase poole aasta jooksu või harvem	Vähemalt viimasel kuul	KOKKU
Vanus kolmene	15-29	,5%	1,2%	**98,3%**	100,0%
	30-54	,9%	4,3%	**94,8%**	100,0%
	55-74	**4,0%**	**13,3%**	82,7%	100,0%
KOKKU		1,2%	4,5%	94,3%	100,0%

% within Millal te viimati Internetti kasutasite

		Millal te viimati Internetti kasutasite			
		Pole kunagi arvutit kasutanud	Viimase poole aasta jooksu või harvem	Vähemalt viimasel kuul	KOKKU
Vanus kolmene	15-29	14,3%	9,8%	**38,8%**	37,2%
	30-54	35,7%	45,1%	**47,7%**	47,5%
	55-74	**50,0%**	45,1%	13,4%	15,3%
KOKKU		100,0%	100,0%	100,0%	100,0%

Tabel: Interneti kasutamine vanuse lõikes (rea ja veeru protsendid, *Allikas: Mina.Maailm.Meedia 2008*)

- Reaprotsent (üleval) näitab, et harva Internetti kasutajatest on enamus üle 55-aastased.

- Veeruprotsent (all) näitab, et alla 29-aastastest on enamus kasutanud Internetti sageli ehk viimase kuu jooksul.

Näide III

		sõltumatu tunnus x1	sõltumatu tunnus x2	*erinevus:*
sõltuv tunnus	y1	50	15	**15-50=-35**
	y2	35	40	**40-35=5**
	y3	15	45	**45-15=30**
		100	100	

		sõltumatu tunnus x1	sõltumatu tunnus x2	*erinevus:*
sõltuv tunnus	y1	20	40	**40-20=20**
	y2	20	40	**40-20=20**
	y3	60	20	**20-60=-40**
		100	100	

		sõltumatu tunnus x1	sõltumatu tunnus x2	*erinevus:*
sõltuv tunnus	y1	45	15	**15-45=-30**
	y2	10	70	**70-10=60**
	y3	45	15	**15-45=30**
		100	100	

- Tabelites on veeruprotsente võrreldud ridade lõikes. Keskse tendentsi määratlemiseks leitakse kõik suurimad protsendid konkreetse rea lõikes.

- Sõltuvalt suuremate protsentide paiknemisest tabelis (erinevus peab olema vähemalt 15%) on seos tunnuste vahel lineaarne või kõverjooneline.

Ülesanne

- **Lugeda artiklit indeksite moodustamise teemal ja vastata järgmistele küsimustele:**
  - *Mis on indeks ehk koondtunnus?*
  - *Mis on indekstunnuste kasutamise puudused (võrreldes üksikute algtunnustega)?*
  - *Mis võivad olla indekstunnuste eelised?*
  - *Tooge näiteid indeksitest, millega olete kokku puutunud (nt millest kuulnud, lugenud vms)?*

Analüüs indeksitega

- Indeksid ehk koondtunnused – spetsiaalse metoodika alusel leitud näitajad, mille alusel üldistatakse teatud nähtuse iseloomu või selle arengut.
  - Indeks esindab mitut tunnust, st arvutuslikult ühendatakse mitmed analüüsiühikud (st ankeedi küsimused).
  - Võimaldab analüüsida komplekssеid sotsiaalseid nähtuseid, mida on raske üksiktunnustega mõõta.
  - Koondab ja üldistab andmestikku (*Nt ajakirjanduse ja kommunikatsiooni osakonna uurimuse "'Mina.Maailm.Meedia" andmestikes on olnud ligi 800 üksiktunnust, mille põhjal on moodustatud üle 100 üldise koondtunnuse*).

Näide

Tabel: Sotsiaalse võrgustikukapitali indeks rahvarühmades (% rahvarühmast, *p<.01)

		Madal	Alla keskmise	Kesk-mine	Üle kesk-mise	Kõrge
Sugu	Mees	8	13	52	18	8
	Naine	7	11	51	21	10
Ankeedi keel	Eesti	8	12	53	18	9
	Vene	6	12	50	22	10
Haridus*	Põhiharidus	10	12	55	17	5
	Keskharidus	8	13	52	20	8
	Kõrgharidus	5	10	46	23	16

Indeks loodi järgmiste algtunnuste alusel: (1) *kelle puhul järgnevas nimekirjas Te võite öelda, et tunnete nendega mingit ühtekuuluvustunnet?* (2) isiklikud kontaktid eri maades, (3) *Kuivõrd sageli pöörduvad kaaslased, tuttavad, pere liikmed Teie poole nõu ja arvamuse küsimiseks?*

Allikas: Pruulmann-Vengerfeldt 2004

Indeksite moodustamine I

1. Leitakse indikaatorid (tunnused), mis kirjeldavad antud nähtust, mõõdetakse nende numbrilised väärtused, määratletakse nähtuse seisukohalt olulised skaala punktid
   - Algtunnused peavad olema mõõdetud sarnasel skaalal, tunnused peavad sisuliselt kokku sobima.
   - Algtunnustele leitakse ühine nimetus ehk ühismõõdustaja (*nt majanduslik kapital*).
2. Algtunnused ühendatakse
   - Intervallskaalal tunnused liidetakse – summaindeks (*nt leitakse poliitikute usalduse indeks, liites kokku hinnangud kolmele poliitikule skaalal* –5 kuni +5).
   - Nominaalskaalal tunnuste väärtustele omistatakse kindel punktide arv ning loendatakse punktid kokku - loendusindeks (*nt poliitikute usalduse küsimuse korral arvestada 'üldse mitte'=0, 'mõnevõrra'=1, 'täiesti'=2 jne*).

Indeksite moodustamine II

3. Koontunnuse skaalad kodeeritakse ümber (lühendatakse) järgides tunnuste algse jaotuse loogikat.

   - Tavaliselt kodeeritakse skaala ümber 5-palliseks, kus 1.skaalapunkt - tunnuse puudumine või väga vähene esinemine, 2.skaalapunkt – vähene või alla keskmise, 3.skaalapunkt – keskmine, 4.skaalapunkt – suur või üle keskmise, 5.skaalapunkt – väga suur esinemine

   - Skaalade lühendamisel järgida algse jaotuse kuju. Nt sümmeetrilise jaotuse korral 1. ja 5. skaalapunktis 10-15% vastajatest, 2. ja 4. skaalapunktis 15-20% vastajaid, 3.skaalapunktis 30-40% vastajaid.

   - Analüüsis vaadeldakse enamasti vaid ülemist kolmandikku, st vastajaid, kes koondtunnuse skaalal on kogunud keskmisest rohkem punkte (nt suured poliitikahuvilised).

Allikas: Lauristin, Vihalemm 2004

Indeksite omadused

- Vähendavad tunnuste ulatust, st pole võimalik kirjeldada üksikuid tunnuseid.
  - Uuritava nähtuse lühendatud, abstraktne kujutis, mistõttu indekseid kombineeritakse ja näitlikustatakse alati üksiktunnuste analüüsiga.
- Mõõdetud tunnuste arvutuslik konstrukt, mistõttu moodustab ise uue tunnuse intervallskaalal.
- Peab sisaldama vaid ühte dimensiooni uuritavast nähtusest, vastuoluliste tunnuste liitmisel võib tulemuseks olla moonutatud pilt nähtusest.

Näide I

Tabel. Eesti keele valdamise ja mõistmise indeks

Indeksi algväärtused	Lühendatud indeks 5-ne skaala*	Lühendatud indeks 3-ne skaala
0 (3%)	0-2 (15%): puudub või väga madal	0-4 (38%): madal
1 (4%)	3-4 (23%): alla keskmise	5-6 (36%): keskmine
2 (8%)	5-6 (36%): keskmine	7-9 (26%): kõrge
3 (11%)	7-8 (19%): üle keskmise	
4 (12%)	9 (7%): väga suur	
5 (15%)		
6 (21%)		
7 (13%)		
8 (6%)		
9 (7%)		

Indeksi arvutamisel on kokku liidetud 3 algtunnust: (1) *Milliseid võõrkeeli Te üldiselt oskate (eesti keel)? (2) Hinnake täpsemalt oma eesti keele mõistmise taset! Saan eesti keeles aru...* (3) *Hinnake võimet end eesti keeles väljendada! Suudan eesti keeles...*

Allikas: Masso 2009

Näide II

Tabel. Eesti keele valdamise ja mõistmise sotsiaaldemograafiline iseloomustus (%)

		Madal	Keskmine	Kõrge
Sugu	Mees	41	42	17
	Naine	31	50	18
Kodakondsus	Eesti	31	49	20
	Muu	50	43	7
Kooli asukoht	Tallinna ja Lõuna-Eesti	35	42	23
	Ida-Eesti	36	52	12
Eesti keeles aine õppimine gümnaasiumis	Jah	35	45	20
	Ei	36	55	10
Osatavate võõrkeelte arv	0 või 1 keel	71	24	5
	2 keelt	40	49	11
	3 ja enam keelt	14	53	33

Allikas: Masso 2009

Iseseisvaks lugemiseks

- Tooding, L.-M. (2007). Andmed ja andmeanalüüsi käik. Rmt: *Andmete analüüs ja tõlgendamine sotsiaalteadustes*, Tartu: Tartu Ülikooli Kirjastus, lk 13-36.
- Taagepera, R. (2008). Why social sciences are not scientific enough. Rmt: *Making social sciences more scientific*, Oxford: Oxford University Press, lk 3-13.
- Tooding, L.-M. (2007). Tunnuse jaotus ja seda kokku võtvad parameetrid. Rmt: *Andmete analüüs ja tõlgendamine sotsiaalteadustes*, Tartu: Tartu Ülikooli Kirjastus, lk 39-77.

TARTU ÜLIKOOL
Ajakirjanduse ja kommunikatsiooni instituut

Järeldamine statistiliste hüpoteeside kaudu. Seosekordajad.

Kvantitatiivne andmeanalüüs

Anu Masso (PhD)

Tunnustevaheline seos I

- Andmeanalüüsi eesmärgiks on varieeruvuse kirjeldamine ja varieeruvuse põhjuste selgitamine.
  - Varieeruvus – kuivõrd konkreetse nähtuse omadused erinevad teatud valimi alagruppides (üldmõiste *vs* hajuvust kirjeldav mõiste).
  - Seos – *sõltuvus* kahe mõõdetud nähtuse vahel, mil ühe sündmuse esinemine muudab tõenäoliseks ka teise sündmsue esinemise. Seos *vs* lineaarne seos [ingl.k. *correlation*], seos *vs* põhjuslik seos [ingl.k. *causal relationship*], .
- Metodoloogiline komplekssus (kvalitatiivsete ja kvantitatiivsete meetodite triangulatsiooni).
  - Massimeedia ja sotsiaalsete muutuste seos (Katz 1981): uudised kui sotsiaalse kogemuse organiseerija (Alexander), massikommunikatsioon kui arvamuste vormija (Noelle-Neumann), massimeedia kui õpetaja (MacCormack), TV-uudised mitte-refleksiivse sotsiaalse teadvuse allikas (Tuchm

Tunnustevaheline seos II

- Tunnustevahelise seose olemasolu korral aitab ühe tunnuse jaotus selgitada teise tunnuse jaotust.
  - Seoste analüüsimine toimub tunnuspaaride (nt risttabel) või komplekssemate seosemudelite vormis (nt regressioon).
  - Seose olemasolu hindamiseks risttabelites tuleb 2x2 tabelites leida suuremad protsendid ridade lõikes (*nt kui arvutatud veeru protsent*) ning analüüsida "üldist tendentsi".
  - Mida suurem on risttabelis erinevus protsentides, seda väiksem on tõenäosus, et erinevus on tingitud juhusest. Kui erinevus on 0, pole andmete vahel seost.
  - Suuremate tabelite korral tuleb lisaks protsentjaotusele arvutada üks või mitu sobivat seosekordajat, et väita seose olemasolu või selle tugevust üldpopulatsioonis.

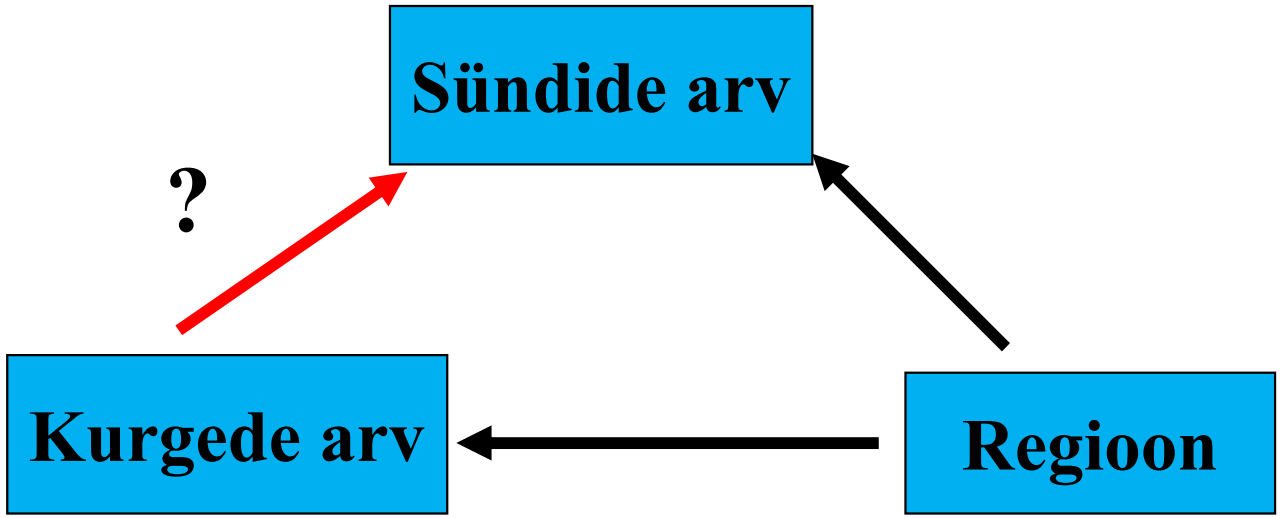
Tunnustevaheline seos III

- Enamasti huvitavad uurijat tunnustevaheliste seoste olemasolu. Ka seose puudumine võib uurimisprobleemi seisukohalt olla oluline tulemus.
  - Orinaalskaalal tunnuste korral on võimalik rääkida seose suunast (*nt positiivne seos: suurema sissetulekuga kõrgem hinnang poliitikule, negatiivne seos: suurema sissetulekuga madalam hinnang*).
  - Mida suurem on risttabelis erinevus protsentides (või mida suurem on erinevus kahe valimi grupi keskväärtustes), seda tugevam seos on.
  - Põhjuslik seos eeldab ühe tunnuse ajalist järgnevust teisele.

“Kolmanda tunnuse probleem” I

Kas seos esineb alati või vaid teatud tingimustel?
Kas on olemas kolmas faktor, mis kutsub esile seost kahe tunnuse vahel?

Kahe tunnuse vahelise seose täpsemaks analüüsimiseks vaadata huvi all oleva 2 tunnuse ja kolmanda tunnuse seost.



"Kolmanda tunnuse probleem" II

- Mis võib algse seosega juhtuda kolmanda tunnuse kaasamisel analüüsi?
- Tõeline, tegelik seos
  - Kahe tunnuse vahel seos jääb alles. St seos pole põhjustatud kolmandast tunnusest (*nt sõltumata riigist töötavad naised enam osalise tööajaga ja mehed enam täistööajaga*).
- Näiline seos
  - Kahe tunnuse vahel seos kaob mõlema kolmanda tunnuse kategooria korral. St seos on tingitud kolmandast tunnusest (*nt Eestis töötavad naised enam täistööajaga, Rootsis mehed osalise ajaga*).
- Tinglik seos
  - Seos kaob ühe kolmanda tunnuse kategooria korral. St seos tunnuste vahel kehtib vaid ühe kolmanda tunnuse väärtuse korral (*nt naised töötavad enam osalise tööajaga vaid Rootsis, Eestis naised ja mehed võrdselt*).

Näide I

Tabel 1. Tööaeg meeste ja naiste seas Rootsis

	mees	naine	KOKKU
täistööaeg	37	13	50
osaline tööaeg	9	35	44
ei tööta	2	4	6
KOKKU	48	52	100

Seos olemas.

Tabel 2. Tööaeg meeste ja naiste seas Eestis

	mees	naine	KOKKU
täistööaeg	33	35	68
osaline tööaeg	11	12	23
ei tööta	4	5	9
KOKKU	48	52	100

Seos puudub.

- **Kolmanda tunnuse "kontrolli all" hoidmine:** kahe tunnuse seose analüüsimine kolmanda tunnuse lõikes. Lihtsaim võimalus – kolmemõõtmeline risttabel .

Allikas: Hüpoteetilised andmed.

Näide II

Tabel. Interneti kasutamise sagedus vanuse lõikes (%).

		Millal Te viimati internetti kasutasite?			
		harvem kui pool aastat tagasi	vähemalt viimase poole aasta jooksul	vähemalt viimasel nädalal	KOKKU
Vanus kolmene	15-29	11,6%	12,7%	**39,6%**	37,2%
	30-54	39,5%	**56,4%**	47,3%	47,4%
	55-74	**48,8%**	30,9%	13,1%	15,3%
KOKKU		100,0%	100,0%	100,0%	100,0%

- Tabelist ilmneb seos Interneti kasutamise ja vanuse vahel. Seose olemasolust annab tunnistust suuremate protsentide paiknemine mööda nö tabeli diagonaali (märgitud tumedalt).
- Seose olemasolu võimalik järeldada, kui veeru protsentide erinevus ridade lõikes on suurem kui 15 protsendiühikut.

Allikas: Mina.Maailm.Meedia 2008.

Näide III

Tabel. Interneti kasutamise seos vanusega sugude lõikes (%).

Sugu			Millal Te viimati internetti kasutasite? harvem kui pool aastat tagasi	vähemalt viimase poole aasta jooksul	vähemalt viimasel nädalal	KOKKU
Mees	Vanus	15-29	5,3%	8,7%	**42,3%**	39,4%
		30-54	**68,4%**	**65,2%**	45,0%	46,8%
		55-74	**26,3%**	**26,1%**	12,7%	13,8%
	KOKKU		100,0%	100,0%	100,0%	100,0%
Naine	Vanus	15-29	16,7%	15,2%	**37,4%**	35,4%
		30-54	16,7%	**48,5%**	**49,1%**	47,8%
		55-74	**66,7%**	36,4%	13,5%	16,8%
	KOKKU		100,0%	100,0%	100,0%	100,0%

- Tabelist ilmneb seos Interneti kasutamise ja vanuse vahel nii meeste kui ka naiste lõikes. St tegemist on tõelise seosega Interneti kasutamise ja vanuse vahel, mis pole selgitatav kolmandast tunnusest ehk soolisest kuuluvusest.

Allikas: Mina.Maailm.Meedia 2008.

Näide IV

Tabel. Interneti kasutamise seos vanusega sissetuleku lõikes (%).

Sissetulek ühe pereliikme kohta kuus			Millal Te viimati internetti kasutasite? harvem kui pool aastat tagasi	vähemalt viimase poole aasta jooksul	vähemalt viimasel nädalal	KOKKU
kuni 4000kr	Vanus	15-29	16,7%	10,7%	**41,4%**	38,0%
		30-54	44,4%	**64,3%**	52,0%	52,5%
		55-74	**38,9%**	25,0%	6,6%	9,5%
	KOKKU		100,0%	100,0%	100,0%	100,0%
4001-6000 kr	Vanus	15-29			**36,9%**	33,2%
		30-54	31,6%	**50,0%**	**45,2%**	44,5%
		55-74	**68,4%**	50,0%	17,9%	22,3%
	KOKKU		100,0%	100,0%	100,0%	100,0%
üle 6001 kr	Vanus	15-29	**33,3%**	7,7%	**38,8%**	37,8%
		30-54	**50,0%**	**53,8%**	**45,9%**	46,2%
		55-74	16,7%	**38,5%**	15,2%	16,0%
	KOKKU		100,0%	100,0%	100,0%	100,0%

- Seos Interneti kasutamise ja vanuse vahel ilmneb kahes madalamas sissetulekugrupis. Seos on tinglik, st madalama sissetulekuga gruppides on noored aktiivsemad Interneti kasutajad, suurimas sissetulekugrupis on igapäevaseid kasutajaid võrdselt eri vanu[illegible] gruppides

Allikas: Mina.Maailm.Meedia 2008.

Ülesanne

- **Eesti järgmine suur eesmärk: rahvuse ja kultuuri säilitamine *(Allikas: Epl, 7.02.11)***

  Uuringufirma Klaster küsis Eesti Päevalehe tellimusel jaanuari lõpus ligi 500 inimeselt, mis peaks olema Eesti siht nüüd, kui oleme juba saavutanud mitu suurt eesmärki: kuulume NATO-sse ja Euroopa Liitu ning oleme võtnud kasutusele euro. Selgus, et inimesed on üsna konservatiivsed. Üle poole vastanutest (54,1 protsenti) arvas, et meie suur väljakutse on, nagu põhiseaduse avalausetes kirjas, eesti rahvuse ja kultuuri säilitamine. Sellist tahet väljendasid nii rikkad kui ka vaesed, nii kõrgharidusega kui ka madalama haridustasemega inimesed. Teistsugune vaatenurk oli vaid neil küsitluses osalenutel, kes vastasid ka, et soovivad järgmise Eesti peaministrina näha Keskerakonna esimeest Edgar Savisaart.

Ülesanne

- **Küsimused:**
  - ***Leidke tunnused (nähtused), mille vahelist seost on analüüsitud?***
  - ***Millised kolmandad tunnused võivad aidata kirjeldatud seoseid selgitada?***
  - ***Milles võib seisneda mainitud uuringu viga?***

Statistiline tõenäosus I

- Seose statistilise olemasolu väitmiseks tuleb hinnata tõenäosust, kuivõrd valimi põhjal tehtud oletused kehtivad ka üldpopulatsioonis, kust konkreetne valim on võetud.
  - Populatsiooni ehk üldkogumi moodustab uurimisülesande sisuga piiritletud kõigi uurimisobjektide kogum. Valim on kindlal viisil eraldatud osa üldkogumist, mida analüüsis kasutatakse üldkogumi asemel.
  - Statistilise käsitluse jaoks on oluline, et valiku põhiprintsiibiks oleks juhuslikkus, mis tagab kõigile populatsiooni indiviididele võrdsed võimalused valimisse sattuda.
  - Valimi enese kohta kehtivad kõik järeldused täpselt, üldkogumi kohta aga teatava veavõimalusega. Vea suurust iseloomustatakse võimaliku eksimise tõenäosusega.

Allikas: Tooding 2007.

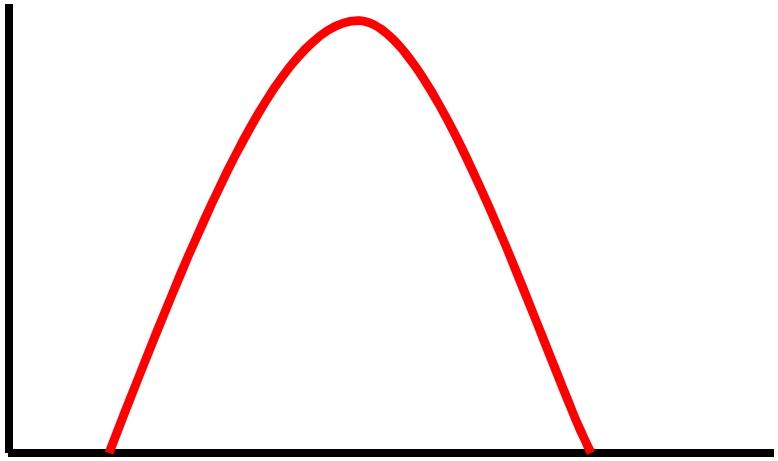
Statistiline tõenäosus II

- Sündmuse tõenäosus on arv 0 ja 1 vahel, kus väärtus 0 tähistab sündmuse võimatust (ebatõenäoline) ja väärtus 1 sündmuse kindlat esinemist (tõenäoline).
- Praktikas kasutatakse statistilise seose olemasolu hindamiseks olulisuse nivood ehk vea ülempiiri, mis näitab kui suur statistiline viga on lubatud, kui väidame seose olemasolu.
  - Hüpoteesipaari kontrollimisel tehakse järeldus nii, et esimest liiki vea tõenäosus ei ületaks olulisuse nivood. Traditsioonilised olulisusnivoo väärtused on 0,05; 0,01, harvem ka 0,10 (st seos kehtib 95, 99 või 90 juhul 100st).
  - Mida väiksem on olulisusnivoo, seda tõsikindlam on tulemus, kuid ühtlasi õnnestub sel juhul alternatiivhüpoteesi suhteliselt raskemini vastu võtta.

Allikas: Tooding 2007.

Normaaljaotus

- Normaaljaotus on tuntuim klassikalistest tõenäosusjaotustest.
  - Normaaljaotus on ühetipuline keskväärtuse suhtes sümmeetriline jaotus.
  - Mida suurem on standardhälve, seda väiksema järsakusastmega on kõver. 95,5% väärtustest paikneb kahe standardhälbe ulatuses keskväärtusest. 68,3% väärtustest paikneb ühe standardhälbe kaugusel.



Allikas: Tooding 2007.

- Normaaljaotusest räägitakse vaid arvuliste ehk intervallskaalal tunnuste korral, st sellest sõltuvalt toimub konkreetsete statistiliste testide või analüüsitehnikate valik.

Statistiline hüpotees

- Statistiline hüpotees - oletus üldkogumi jaotuse kohta tervikuna või jaotuse mõne parameetri kohta; oletust kontrollitakse valimi põhjal.
  - Null-hüpotees - H_0, üldkogumi vastamine teatud standardile, kus puuduvad erinevused ja seosed (st teatud kooskõla).
  - Alternatiivhüpotees - H_1, ehk sisukas hüpotees, mida uurija soovib tõestada (tavaliselt mingi erinevuse või seose olemasolu).
- Hüpoteeside testimine - hüpoteesi paikapidavuse kontrollimine teatud eeskirja (testi, kriteeriumi) alusel.
  - Vead hüpoteeside testimisel: I liiki viga, kus võetakse vastu H_1, kuid õige on H_0. II liiki viga, kus jäädakse H_0 juurde, ehkki õige on H_1.

Allikas: Tooding 2007.

Seosekordajad

- Arvuline näitaja (nn “indeks”), mis kvantifitseerib seose olemasolu või tugevuse kahe tunnuse vahel.
  - **Seosekordaja** [*association coefficient*]- mõistet kasutatakse kvalitatiivsete tunnuste (st nominaal- või ordinaalskaala) korral.
  - **Korrelatsioonikordaja** [*correlation coefficient*]- mõistet kasutatakse kvantitatiivsete tunnuste (st intervall- või suhteskaala) korral.
  - **Statistiline test** - mõistet kasutatakse kvantitatiivsete (st intervall- või suhteskaala) tunnuste keskväärtuste võrdlemisel eri valimigruppides.
  - NB! Spearman’i korrelatsioonikordaja on seosekordaja, mida kasutatakse ordinaalskaalal tunnuste vahelise seose hindamiseks.

Seosekordaja valimine

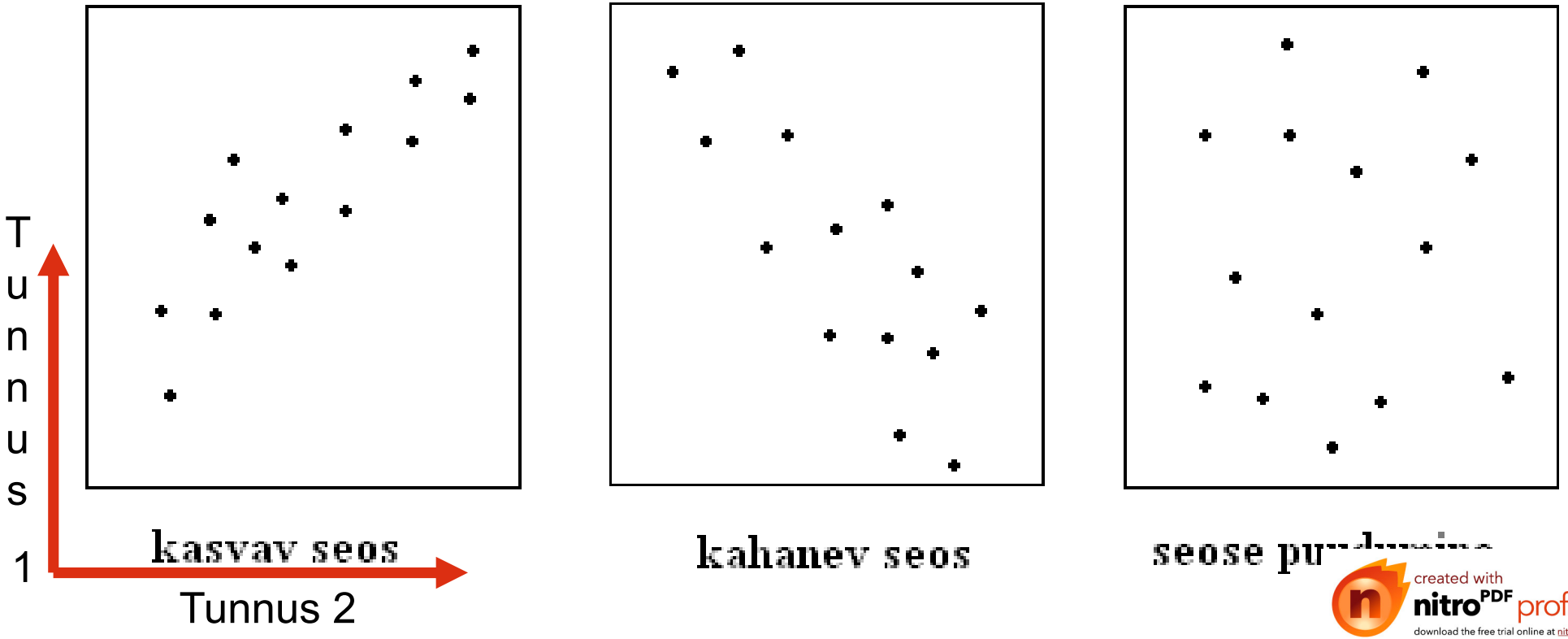
- Skaala tüüp: seosekordajad on arvutatavad konkreetset tunnuse skaalat arvestades.
- Eesmärk: seose olemasolu, tugevus või selle suund?
- Tõlgendamine: enamus seosekordajate väärtuseid varieerub 0 (seose puudumine) ja 1 (täielik seos) vahel. Raske on tõlgendada seose tugevust ja suunda Hii-ruudu korral.
- Marginaaljaotus: osa seosekordajaid (nt Lambda) on tundlikud ebaühtlaste marginaaljaotuste suhtes (st kogusummad risttabelis) (*nt kordaja võib olla 0 ka seose olemasolu korral*).
  - Arvutada mitu seosekordajat: nii on ühte kordajat kasutades võimalik hinnata seose olemasolu (nt Hii-ruut), teist kasutades hinnata seose tugevust (Lambda, Crameri V).

Korrelatsioonikordaja

- Korrelatsioonikordajat kasutatakse seose hindamiseks ordinaal- või intervallskaalal tunnuste vahel. Iseloomustab seose tugevus ja seose suund.
- Seose tugevust (rangust) hinnatakse kordaja suuruse alusel.
  - Kordaja vähim võimalik väärtus on –1 ja suurim 1, st täielik lineaarne sõltuvus. Keskmise tugevusega seoseks peetakse tavaliselt kordaja väärtust vahemikus $0,3<r<0,7$. Tugeva seosega on tegemist alates kordaja väärtusest $r>0,7$.
- Seose suunda hinnatakse kordaja märgi alusel (+ või -).
  - Kordaja positiivne väärtus $\Rightarrow$ samasuunaline seos, st mõlema muutuja väärtused üheaegselt kasvavad/kahanevad. Kordaja negatiivne väärtus $\Rightarrow$ vastassuunaline seos, st ühe muutuja kasvades teine kahaneb.

Hajuvusdiagramm

- Kahe tunnuse väärtuspaaride kandmine kahemõõtmelisele koordinaatteljestikule, võimaldab ligikaudselt hinnata seose olemasolu ja selle kuju.
  - Annab ülevaate äärmuslike indiviidide esinemisest. Ei võimalda täpselt määratleda seose iseloomu. NB! Võimalik kasutada eelkõige intervallskaalal, aga ka ordinaalskaalal tunnuste korral.



Pearson'i ja Spearman'i korrelatsioonikordajad

- Pearson'i kordaja – arvtunnuste korral.

$$r = \frac{1}{n}\sum_{i=1}^{n} \frac{x_i - m_x}{s_x} \times \frac{y_i - m_y}{s_y}$$

n on valimi maht, x_i on esimese tunnuse väärtus indiviidil järjekorranumbriga i ja y_i teise tunnuse väärtus indiviidil järjekorranumbriga i. Suurused s_x ja s_y on vastavalt esimese ja teise tunnuse standardhälbed, m on keskväärtused.

- Spearman'i astakkorrelatsiooni kordaja - ordinaalskaalal tunnuste korral.
  - Iga indiviidi jaoks leitakse nende astakute (järjekorranumbrite) vahe, mille indiviid saaks kogumit vaadeldavate tunnuste põhjal variatsioonreaks järjestades.
  - Kordaja põhineb summaarsel astakute erinevusel, mis on standardiseeritud nii, et kordaja väärtused on arvude –1 ja 1 vahel.

Näide

Tabel. Meedia usalduse seosed valdkonniti (Spearman'i korrelatsioon)

	Eesti Televisioon	Eesti Raadio	Ajalehed	Interneti-portaalid
Eesti Televisioon	1,000	,656**	,477**	,164**
Eesti Raadio	,656**	1,000	,516**	,254**
Ajalehed	,477**	,516**	1,000	,279**
Internetiportaalid	,164**	,254**	,279**	1,000

**. Correlation is significant at the 0.01 level (2-tailed).

- Mitme tunnuse korral nimetatakse paarikaupa korrelatsioonikordajate tabelit korrelatsioonimaatriksiks. Maatriksi diagonaalelementide väärtused võrduvad 1'ga, näidates, et tunnus on alati täielikus sõltuvuses iseendaga.
- Usaldus Eesti Raadiosse ja Eesti Televisiooni on seotud, st mida rohkem usaldatakse ühte, seda rohkem usaldatakse ka teist (r=0,656). Korrelatsioonikordajat väljendades protsendina (0,656*0,656*100=43) näeme, et siiski vaid alla poole ühe tunnuse variatiivsusest on selgitatav teise tunnuse variatiivsuse poolt (kolmanda tunnuse mõju?).

Allikas: uurimus Mina.Maailm.Meedia (2002)

T-test I

- Kasutatakse kahe valimi grupi keskväärtuste statistiliselt olulise erinevuse hindamiseks.
  - Sõltumatute valimite T-test, nt kui on konkreetne mõõtmistingimus ning erinevaid subjekte uuritakse selle tingimuse seisukohalt *(nt usalduse keskväärtus naiste ja meeste lõikes).*
  - Sõltuvate valimite T-test, nt kui on kaks mõõtmistingimust ning samad subjektid osalesid mõlemas mõõtmistingimuses *(nt longituuduurimuses vastatakse küsimusele eri ajahetkedel).*

$$t = \frac{\text{tegelik erinevus valimi keskmiste vahel} - \text{teoreetiline erinevus valimi keskmiste vahel}}{\text{kahe valimi keskmiste erinevuse standardviga}}$$

T-test II

- Kui erinevus valimi gruppide vahel on suurem kui teoreetiline jaotus standardvea alusel, siis on T-testi väärtus suur ja olulisustõenäosus väike.
- Mida suurem on tegelik keskmiste erinevus valimi gruppide vahel, seda kindlamalt võib väita keskmiste erinevust ka populatsioonis (*nt keskmiste erinevus esineb populatsioonis, pole tingitud juhuslikest sündmustest*).
- Suur T-testi ja väike olulisustõenäosuse väärtus näitavad statistilist seost vaid juhul, kui on tegemist juhusliku valimiga.

Näide

Tabel: Usaldus riiklike institutsioonide suhtes rahvuste lõikes
(keskmine,T-test, 1-ei usalda üldse, 5-usaldan täiesti)

	Ankeedi keel:	N	Keskmine	St.hälve	T-test	p
Kirik	Eesti keel	1034	3,07	1,13	-9,595	0,000
	Vene keel	466	3,68	1,15		
Valitsus	Eesti keel	1031	2,87	0,96	12,601	0,000
	Vene keel	465	2,18	1,06		
President	Eesti keel	1034	3,72	0,99	26,153	0,000
	Vene keel	467	2,21	1,13		
Eesti Televisioon	Eesti keel	994	3,79	0,85	21,833	0,000
	Vene keel	465	2,67	1,02		
Internetiportaalid	Eesti keel	984	2,77	0,95	1,643	0,101
	Vene keel	454	2,68	1,03		
Kultuuritegelased	Eesti keel	993	3,58	0,79	0,116	0,005
	Vene keel	463	3,57	0,91		

T-testi suur väärtus ja $p<.01$ või $p<.05 \Rightarrow H_1$ (keskväärtused on statistiliselt oluliselt erinevad). Internetiportaalide keskmine usaldus on eestlaste ja venelaste hulgas sarnane. Kui venelased usaldavad statistiliselt oluliselt enam kirikut, siis eestlased valitsust, presidenti ja ETV'd.

Allikas: Mina.Maailm.Meedia (2008)

F-test

- F-test – normaaljaotusega üldkogumite dispersioonide erinevuste võrdlemiseks.

$$F = \frac{\text{gruppide vaheline dispersioon}}{\text{gruppide sisene dispersioon}}$$

- Tõlgendamine:
  - F on alati positiivne väärtus. F suur väärtus ja $p<.05 \Rightarrow H_1$ (üldkogumite dispersioonid on erinevad).
  - F-testi kasutatakse dispersioonanalüüsis (ANOVA), regressioonanalüüsis vms.

Dispersioonanalüüs (ANOVA)

- Dispersioonanalüüs ehk ANOVA *(analysis of variance)* – mitme valimi grupi keskmiste võrdlemiseks.
  - Mõõdetakse teatavate tunnuste (ka sõltumatud tunnused või faktortunnused) mõju uuritava tunnuse (ka sõltuv tunnus või argumenttunnus) keskväärtusele;
  - Nn faktortunnused on kindla arvu väärtusega arvulised või (loogilise keskpunktiga) järjestusskaalal tunnused.
  - I liiki ehk fikseeritud mudel (analüüsitakse kõiki tasemeid). II liiki ehk juhuslik mudel (analüüsitakse teatud tasemeid).

Näide

Tabel: Dispersioonanalüüs: usaldus institutsioonidesse haridusgruppide lõikes

		N	Keskmine	St.hälve	F	p
Eesti riik	Alghariidus	383	3,28	1,04	1,418	0,243
	Põhiharidus	756	3,23	1,07		
	Kõrgharidus	367	3,34	1,04		
Valitsus	Alghariidus	383	2,66	1,06	0,906	0,404
	Põhiharidus	756	2,60	1,05		
	Kõrgharidus	367	2,69	1,09		
Riigikogu	Alghariidus	383	2,54	1,01	5,690	0,003
	Põhiharidus	756	2,39	1,01		
	Kõrgharidus	367	2,57	0,94		
President	Alghariidus	383	3,25	1,21	0,047	0,954
	Põhiharidus	756	3,23	1,24		
	Kõrgharidus	367	3,25	1,35		
Politsei	Alghariidus	383	3,36	1,02	1,892	0,151
	Põhiharidus	756	3,38	1,00		
	Kõrgharidus	367	3,49	0,90		
Pangad	Alghariidus	383	3,10	1,09	8,317	0,000
	Põhiharidus	756	3,06	0,99		
	Kõrgharidus	367	3,32	0,91		

Allikas: Mina.Maailm.Meedia (2008)

Hii-ruut

- Hii-ruut (X^2) on seosekordaja, mis sobib seose olemasolu hindamiseks nominaal-, ordinaal-, või arvskaalal tunnuste korral.
  - Kõige enam leiab Hii-ruut kasutust nominaalskaalal või väikese kategooriate arvuga ordinaalskaalal tunnuste korral.
  - Hii-ruut ei võimalda midagi öelda seose tugevuse ega selle suuna kohta (erinevalt korrelatsioonikordajast).
- Hii-ruudu vähimaks väärtuseks on 0 (tunnuste vahelise seose puudumine); maksimaalne väärtus sõltub valimi suurusest ja sagedustabeli väljade arvust.
  - Seose olemasolu hindamiseks tuleb lisaks Hii-ruudu väärtusele jälgida kordaja statistilise olulisuse tõenäosust.

Hii-ruudu arvutamine I

- Võrreldakse erinevust reaalse ja hüpoteetilise jaotuse vahel. Kui erinevus on väike, on Hii-ruudu väärtus väike; kui erinevus on suur, on kordaja väärtus suur.
  - Reaalne jaotus on kahe tunnuse eri väärtuste koosesinemise sagedus risttabelis selliselt, nagu see uuritava valimi korral avaldub.
  - Hüpoteetiline jaotus on arvutuslik kahe tunnuse koosesinemise sagedus rissttabelis eelduse korral, et tunnused on omavahel sõltumatud.

Tabel 1. Reaalne jaotus

		X		
		x1	x2	KOKKU
Y	y1	a	b	a+b
	y2	c	d	c+d
KOKKU		a+c	b+d	n

Tabel 2. Hüpoteetiline jaotus

		X		
		x1	x2	KOKKU
Y	y1	(a+c)(a+b)/n	(a+b)(b+d)/n	a+b
	y2	(a+c)(c+d)/n	(b+d)(c+d)/n	c+d
KOKKU		a+c	b+d	n

Hii-ruudu arvutamine II

Tabel 1. Reaalne jaotus: meeste ja naiste tööaeg

	mees	naine	KOKKU
täistööaeg	37	13	R1=50
osaline tööaeg	9	35	R2=44
ei tööta	2	4	R3=6
KOKKU	V1=48	V2=52	N=100

Tabel 2. Hüpoteetiline jaotus: meeste ja naiste tööaeg

	mees	naine	KOKKU
täistööaeg	R1*V1/N=24	R1*V2/N=26	24+26=50
osaline tööaeg	R2*V1/N=21	R2*V2/N=23	21+23=44
ei tööta	R3*V1/N=3	R3*V2/N=3	3+3=6
KOKKU	24+21+3=48	26+23+3=52	N=100

Hüpoteetilise ehk teoreetilise jaotuse arvutamiseks tuleb konkreetses tabeli väljas korrutada vastav rea- ja veerusumma (absoluutarvudes) ning jagada tulemus kogu indiviidide arvuga, kes antud küsimusele vastasid (*siin 100*).

Hii-ruudu arvutamine III

Tabel 3. Reaalse ja hüpoteetilise jaotuse vahe

	mees	naine	KOKKU
täistööaeg	37-24=13	13-26=-13	50-50=0
osaline tööaeg	9-21=-12	35-23=12	44-44=0
ei tööta	2-3=-1	4-3=1	6-6=0
KOKKU	48-48=0	52-52=0	100-100=0

Tabel 4. Teisendatud ruuthälbed (hälbe ruut/hüpoteetiline jaotus)

	mees	naine	KOKKU
täistööaeg	13*13/24=7	-13*-13/26=7	
osaline tööaeg	-12*-12/21=7	12*12/23=6	
ei tööta	-1*-1/3=0,3	1*1/3=0,3	
KOKKU	7+7+0,3=14,3	7+6+0,3=13,3	Hii-ruut: 14+13=27

Saadud Hii-ruudu väärtuse olulisust kontrollitakse Hii-ruudu kriitiliste väärtuste alusel, mida võimalik leida nt erinevatest statistika õpikutest. Statistikaprogrammid (nt SPSS) arvutavad nii Hii-ruudu väärtuse kui ka kordaja olulisuse tõenäosuse.

Tõlgendamine I

- Statistilise otsustuse loogika seose hindamisel Hii-ruudu abil:
  - Null-hüpotees: kaks (või enam) tunnust on sõltumatud, st erinevus reaalse ja hüpoteetilise jaotuse vahel on 0 (Hii-ruudu väärtus väike), juhul kui $p \geq 0,05$ (või $p \geq 0,01$);
  - Alternatiivhüpotees: kaks tunnust on omavahel sõltuvad, st erinevus reaalse ja hüpoteetilise jaotuse vahel on erinev nullist ning tulemus pole tingitud juhusest (Hii-ruudu väärtus suur), juhul kui $p \leq 0,05$ (või $p \geq 0,01$).

Tõlgendamine II

- Testib vaid seose olemasolu tunnuste vahel, kuid ei ütle midagi seose tugevuse ega selle suuna kohta.
- Analüüsi tegemisel jälgida, et tabeli väljade ehk vabadusastmete (df) arv poleks liialt suur.
  - Df=(k-1)(r-1), kus k-veergude arv tabelis, r=ridade arv tabelis. Suure arvu tabeli väljade korral ei pruugi Hii-ruut seost näidata ($\Rightarrow$ vajadus tunnuseid ümber kodeerida);
  - Hii-ruutu ei tohiks kasutada, kui hüpoteetilise jaotuse korral on 20% tabeli väljades absoluutarvarv <5 või kui mõnes tabeli väljas on hüpoteetilise jaotuse sagedus <1.
- Kuna Hii-ruudu maksimaalne väärtus sõltub tabeli väljadest ja valimist, ei pruugi kaks Hii-ruudu väärtust olla omavahel võrreldavad.

Näide

Tabel: Interneti kasutamise seos vanusega (% ja Hii-ruut)

		Millal Te viimati internetti kasutasite?			
		harvem kui pool aastat tagasi	vähemalt viimase poole aasta jooksul	vähemalt viimasel nädalal	KOKKU
Vanus kolmene	15-29	11,6%	12,7%	**39,6%**	37,2%
	30-54	39,5%	**56,4%**	47,3%	47,4%
	55-74	**48,8%**	30,9%	13,1%	15,3%
KOKKU		100,0%	100,0%	100,0%	100,0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	62,911[a]	4	,000
Likelihood Ratio	55,283	4	,000
Linear-by-Linear Association	52,966	1	,000
N of Valid Cases	1134		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 6,60.

- Protsentjaotustest ilmneb seos tunnuste vahel.
- Hii-ruudu väärtus on statistiliselt oluline nivool p<.001, st protsentjaotuste erinevus on statistiliselt oluline ning seos Interneti kasutamise ja vanuse vahel kehtib ka järeldusi tehes üldpopulatsioonile

Allikas: Mina.Maailm.Meedia (2008)

Teisi seosekordajaid I

Seosekordaja	Skaala	Mida testib?	Tabeli suurus	Sisu
Hii-ruut, X^2 [*Chi-square*]	Nominaal, ordinaal, intervall	Seose olemasolu / puudumist	m x n	Tabeli väljades ei või hüpoteetiline jaotus olla <1 või rohkem kui 20% väljadest <5. Ei ütle midagi seose tugevuse ega suuna kohta.
Fii, Φ [*Phi*]	Nominaal	Seose olemasolu ja tugevust	2 x 2	Tabelites >2 x 2 võib Φ olla >1. Saadud Hii-ruudu normeerimisel.
Cramer'i V	Nominaal	Seose olemasolu ja tugevust	m x n	2 x 2 tabelis V= Φ. Saadud Hii-ruudu normeerimisel. Maksimaalne väärtus=1.
Kontingentsus-kordaja, C	Nominaal	Seose olemasolu ja tugevust	m x n	C maksimaalne väärtus on 0,87. Saadud Hii-ruudu normeerimisel.
Lambda, λ	Nominaal	Proportsionaalset vea vähenemist, sümmetriline või asümmeetriline seos.	m x n	Väärtus 0 ja 1 vahel. Võib olla 0 ka sõltuvuse korral, kui marginaalide jaotused äärmised ebaühtlased. Tuleb eelnevalt määratleda sõltuv ja sõltumatu tunnus.
Määramatuse koefitsient, U [*Uncertainty coefficient*]	Nominaal	Seose suunda	m x n	Testib määramatuse vähenemist sõltuvas tunnuses, mis on tingitud sõltumatu tunnuse teadmisest. Võtab arvesse marginaalide jaotuste erinevuse.

Teisi seosekordajaid II

Seosekordaja	Skaala	Mida testib?	Tabeli suurus	Sisu
Gamma	Ordinaal	Seose olemasolu.	m x n n x n	Sarnane Lambdaga. Väärtus +1: suuremad protsendid on koondunud tabelis diagonaalis vasakult paremale, -1: protsentjaotused on koondunud diagonaalis paremalt vasakule.
Somers'i D	Ordinaal	Seose olemasolu.	m x n n x n	Sarnane Gamma'ga. 1: väärtused asetsevad piki tabeli diagonaali, 0-tunnused on sõltumatud.
Kendall'i Tau	Ordinaal	Seose olemasolu	m x n n x n	Võrdleb tabelis tunnuste paaride kooskõla. Varieerub –1 ja 1 vahel.
Spearman'i Roo	Ordinaal	Seose olemasolu		Astakkorrelatsiooni kordaja, 0 ja 1 vahel (1=täiuslik seos).

NB! Kui üks tunnus nominaalskaalal ja teine ordinaalskaalal, siis kasutada ordinaalskaala seosekordajat. Kui üks tunnus nominaalskaalal ja teine intervallskaalal, siis kasutada spetsiifilisi seosekordajaid (nt Eeta, Kappa jms).

Erinevatel skaaladel tunnuste seose analüüsimisel alati võimalik kasutada Hii-ruutu.

Näide

Tabel. Eluga rahulolu Eestis ja Euroopas sotsiaaldemograafiline iseloomustus (% ja Cramer'i V)

		Euroopa elanike rahulolu eluga (%)			Eesti elanike rahulolu eluga (%)			Seose tugevus (Cramer'i V)*	
		Madal	Kesk-mine	Kõrge	Madal	Kesk-mine	Kõrge	Euroopa	Eesti
Etniline vähemus	Jah	9	7	4	42	35	19	0,076	0,209
	Ei	92	93	96	58	65	81		
Sugu	Mees	40	43	51	40	42	47	0,097	0,064
	Naine	61	57	49	60	58	53		
Tasustatava töö olemasolu	Jah	33	45	74	38	50	75	0,352	0,301
	Ei	67	56	26	62	50	25		
Perekonnaseis	Abielus	48	50	54	41	44	45	0,102	0,124
	Vallaline	21	27	29	25	29	38		
	Muu	31	23	17	33	27	16		

- Objektiivsete faktorite (sissetulek) kõrval on eluga rahulolu kujunemisel olulised subjektiivsed faktorid (enda määratlemine enamus- või vähemusrahvusena). Etniline tegur on eluga rahulolu määramisel Eestis olulisem kui Euroopas keskmiselt.

Allikas: Euroopa Sotsiaaluuring (2006)

Ülesanne

- **Lugeda artiklit ja vastata järgmistele küsimustele:**
  - *Milliseid seosekordajaid on kasutatud?*
  - *Millisel skaalal on tunnused, mille seoseid on nende seosekordajatega arvutatud?*
  - *Mis tüüpi seostega on tegemist (tegelik, näiline, tinglik)?*
  - *Leidke näiteid tugevatest ja nõrkadest seostest?*

Iseseisvaks lugemiseks

- **Tooding**, L-M. (2007). Ptk 4. Järeldamine statistiliste hüpoteeside kaudu. Rmt: *Andmete analüüs ja tõlgendamine sotsiaalteadustes*, Tartu: Tartu Ülikooli Kirjastus, lk 140-154.
- **Tooding**, L-M. (2007). Ptk 5.1. Seos jaotuste vahel. Rmt: *Andmete analüüs ja tõlgendamine sotsiaalteadustes*, Tartu: Tartu Ülikooli Kirjastus, lk 196-227.
- **Field**, F. (2000). Chapter 6: Comparing two means. Rmt: *Discovering statistics using SPSS for Windows: advanced techniques for the beginner*, London etc: Sage, lk 206-242.
- **Masso**, A. (2009). Etnilised erisused rahulolu hinnangutes: Eesti eripära Euroopa kontekstis. M. Lauristin (toim.) Rmt. *Eesti inimarengu aruanne*.

TARTU ÜLIKOOL
Ajakirjanduse ja kommunikatsiooni instituut

Lineaarne ja mittelineaarne regressioon

Kvantitatiivne andmeanalüüs

Anu Masso (PhD)

Kirjeldav *vs* ennustav / selgitav lähenemine

- Kirjeldav ja selgitav lähenemine
  - Segastrateegiad parimad: statistiline kirjeldav analüüs (vastused piiratud küsimustele), kontseptuaalsed mudelid (võimaldavad küsimuste-ringi laiendada).
- Statistiline sõltuvus
  - Kval.hüpoteesid → Andmekogumine → Seoste testimine → Kvant.mudelid (kvant.hüpoteesid) → Täiendav andmekogumine → Seoste testimine → Mudeli täpsustamine → Testimine →…
- Sotsiaalsete seaduspärasuste leidmine
  - Oluline leida seoste kontseptuaalsed selgitused – võimalike analüüsitavate tunnuste arvu suurendamine. Teooria – paljude üksteisega seotud seaduspärasuste kombinatsioon; seaduspärasused on omakorda kontseptuaalselt põhjendatud ning empiiriliselt testitud.

Allikas: Taagepera 2008

Ülesanne

- **Lugeda teksti (Taagepera 2008, lk 187-198) ja vastata järgnevatele küsimustele:**
  - ***Mis on kirjeldava ja selgitava/ennustava lähenemise erinevused?***
  - ***Mis on statistiline sõltuvus ja millised võiksid olla vahendid selle analüüsimiseks?***
  - ***Kuidas leida sotsiaalseid seaduspärasusi (vrd gravitatsiooniseadus)?***

Statistiline sõltuvus I

- Statistilise sõltuvuse korral on tegemist seose eriliigiga, kus eeldatakse, et ühe tunnuse väärtuste varieeruvus sõltub teise tunnuse väärtustest.
  - Eristatakse statistiliselt sõltumatuid (*nt sots.dem.*) ja sõltuvaid *(nt Interneti kasutamine)* tunnuseid.
  - Otsustus, milline tunnus millisest sõltub, vajab põhjalikku argumentatsiooni, nii teoreetilist kui empiirilist *(nt varasemad uuringud, täiendav kvalitatiivne analüüs).*
- Sõltuvust analüüsitakse enamasti kompleksete statistiliste mudelite abil.
  - Mudelid võimaldavad analüüsida samaaegselt mitme sõltumatu tunnuse mõju sõltuvatele tunnustele või nende mõju olulisust omavahel võrrelda.

Statistiline sõltuvus II

- Sõltuvat ja sõltumatut tunnust eristavate analüüsimeetodite kasutamine toimub arvestades tunnuste skaala tüüpi.

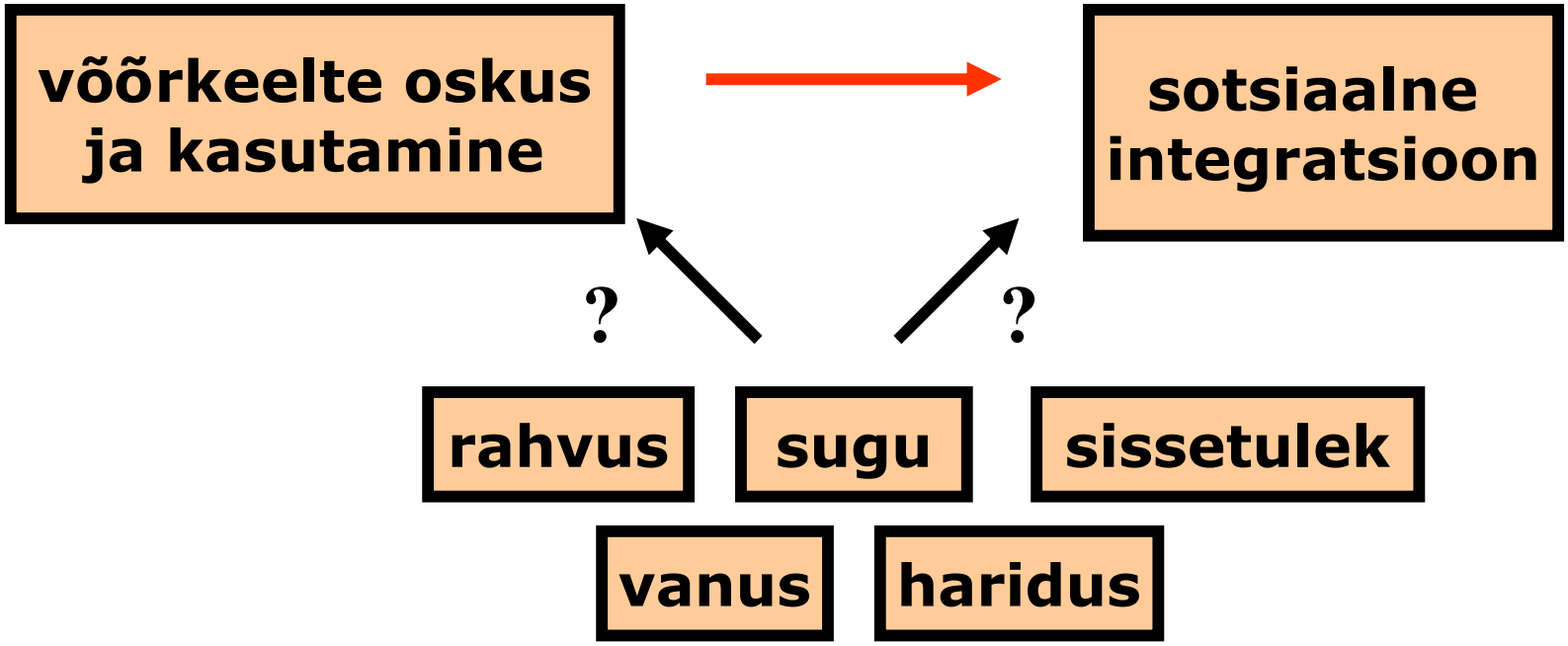
Meetod	Sõltuv tunnus	Sõltumatu tunnus	Näide
Dispersioon-analüüs	Intervall, ordinaalne	Nominaalne, ordinaalne	*Kuivõrd sõltub riiklike institutsioonide usaldus vanusest?*
Lineaarne regressioon	Intervall	Intervall, dihhotoomne	*Kuivõrd sõltub IQ skoori tulemus pea ümbermõõdust JA pikkusest?*
Multinoomne logistiline regressioon	Dihhotoomne, nominaalne	Dihhotoomne, nominaalne	*Kuivõrd sõltub võõrkeelte oskus sotsiaalsest, majanduslikust JA kultuurilisest kapitalist?*

Regressioonanalüüs

- Regressioon – analüüsitehnika, mis võimaldab analüüsida seost kahe või enama tunnuse vahel, eesmärgiga *ennustada* ühe tunnuse väärtust teiste tunnuste väärtuste alusel.
  - Nt koolis/ülikoolis käidud aastate arvu abil saame regressiooniga ennustada võimalikku sissetulekut; või reklaamide arv võib aidata ennustada toote müügiedu.
  - Regressioon võimaldab leida mudeli, mis analüüsib nähtust tervikuna, st samaaegselt mitme sõltumatu tunnuse mõju (nt müügiedu ilmneb vaid siis, kui reklaam on kaupluse aknal või kesklinna kauplustes).
  - Erinevalt Hii-ruut-testist ja korrelatsioonanalüüsist tegemist nn suunatud seosega, st regressioonanalüüs näitab, milline tunnustest loob varieeruvuse teises tunnuses.

Näide

- **Hüpotees:** võõrkeelte oskus ja kasutamine on kultuuriline kapital, mis suurendab sotsiaalset integratsiooni (*nt toimetulek ühiskondlike muutustega, majanduslik ettevõtlikkus, kuulumine organisatsioonidesse, identiteet*).



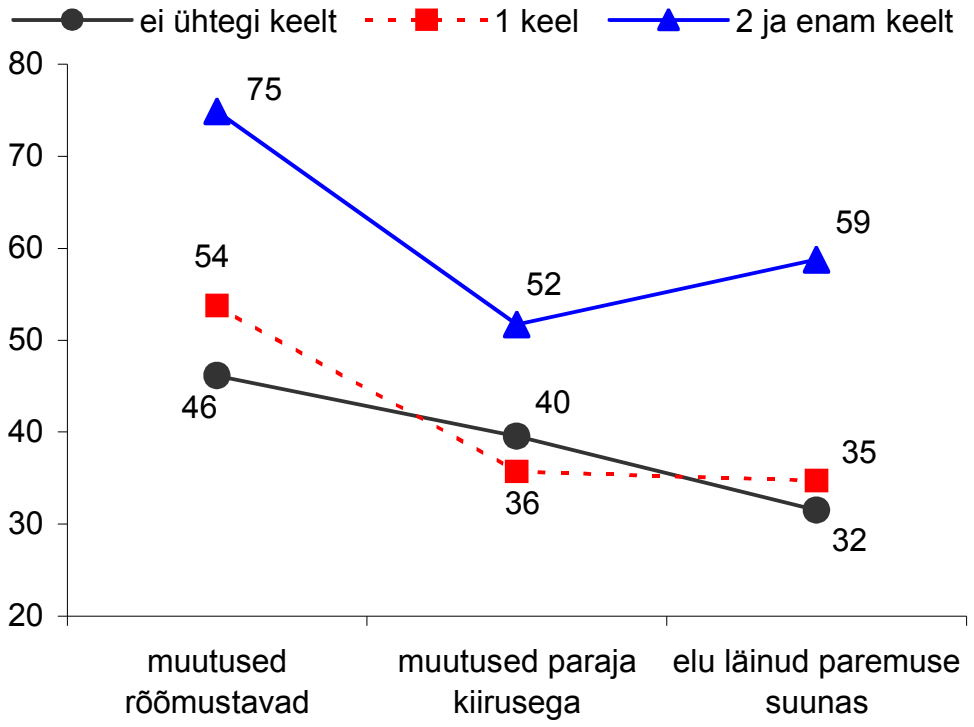
- **Küsimus:** Kas võõrkeelte oskus on ainuke sotsiaalset integratsiooni kirjeldav tunnus? Kuivõrd on seos seletatav teiste taustatunnuste kaudu? Kui suur on võõrkeelte oskuse ja kasutamise "puhas mõju", st tunnuse mõju üksinda sotsiaalsele integratsioonile.

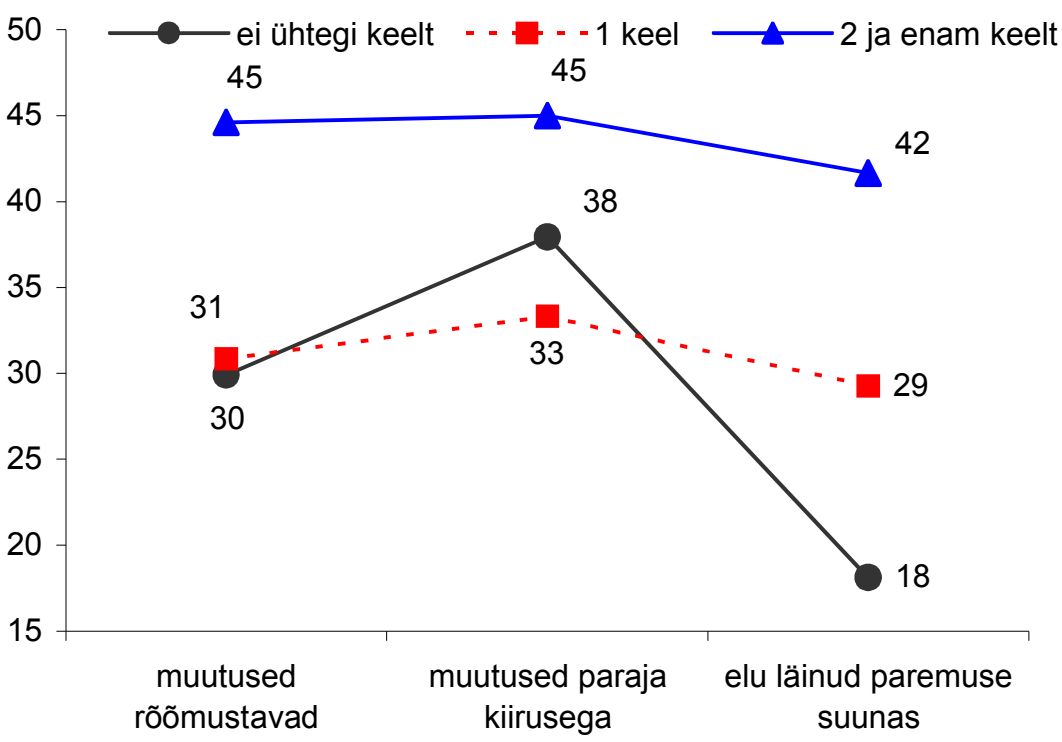
Seoste analüüs

- Latentse tunnuse mõju selgitamiseks ja varjatud tunnuste mõju “kontrolli all hoidmiseks” on järgmised võimalused:
- Seoseanalüüs risttabeli ja seosekordajate alusel
  - Võrreldakse seost kahemõõtmelises ja kolmemõõtmelises risttabelis. Analüüsitakse, kas olemasolev kahe tunnuse seos jääb alles (tõeline, tegelik seos) või kaob ära (näiline seos).
  - **NB!** Piiranguks on tunnuste arv, st kahe tunnuse seose analüüsimisel on võimalik korraga vaid 1 taustatunnust “kontrolli” all hoida.
- Regressioonanalüüs
  - Regressioonanalüüs võimaldab arvutada mudeli, kuhu on võimalik samaaegselt kaasata kõik tunnused, mille mõju konkreetsele tunnusele soovitakse analüüsida.

Näide I

Tulemus (risttabel): kahe ja enama võõrkeele oskus ja kasutamine suurendab sotsiaalset integratsiooni, st ühiskonnas toimunud muutusi tunnetatakse positiivsemalt.





Joonis 1. Ühiskondlike muutuste tunnetamine võõrkeelte lõikes eestlastel (vasakul) ja venelastel (paremal)

Allikas: Uurimus Mina.Maailm.Meedia 2002.

Näide II

Tabel 1. Ühiskondlike muutuste tunnetamise prognoosimine sotsiaal-demograafiliste tunnuste ja võõrkeelte alusel (multinominaalse logistilise regressiooni tulemused, + seose olemasolu nivool p<=.01)

	Muutused on olnud rõõmustavad / kurvastavad	Muutused on olnud liiga kiired / paraja kiirusega	Teie ja Teie pere elu on liikunud paremuse / halvemuse suunas	KOKKU seoste arv reas
Keele oskus ja kasutamise ulatus	+			1
Ankeedi keel	+	+	+	3
Vanus	+	+	+	3
Sugu				0
Haridus	+	+		2
Sissetulek	+	+	+	3
KOKKU seoste arv veerus	5	4	3	12

- **Tulemus (regressioon):** Keelteoskus oluline vaid muutuste rõõmustavana-kurvastavana tunnetamisel, seos muutuste kiiruse ja isikliku elu muutuse tunnetamisega (mis ilmnes eelnevalt jooniselt) on tegelikult tingitud ankeedi keele, vanuse ja sissetuleku mõj

Ülesanne

Ülesanne 1:

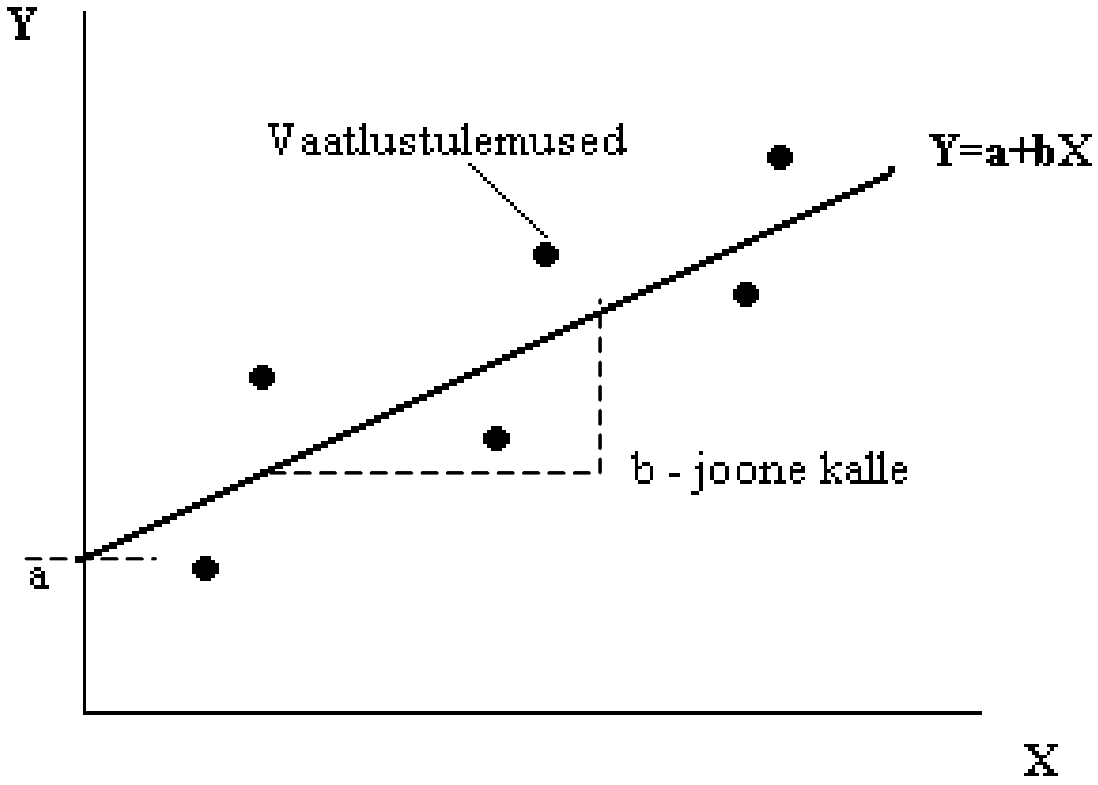
- ***Kas tunnuste vahel esineb seos ning milline on seose suund (negatiivne või positiivne)?***

Ülesanne 2:

- ***Milliste seostega on tabelites tegemist, kas näilise või tõelise seosega?***
- ***Kas seos tunnuste haridus ja toitumisharjumused on selgitatav tunnuste sissetulek ja perekonnaseis kaudu?***

Lineaarne regressioon

- Lineaarne kahe arvtunnuse regressioonanalüüs – kui suur osa sõltuva tunnuse varieeruvusest on tingitud sõltumatu tunnuse varieeruvusest.



Regressioon on kirjeldatav sirgjoonega, mis väljendub võrrandiga:

Y=a+bX+ε,

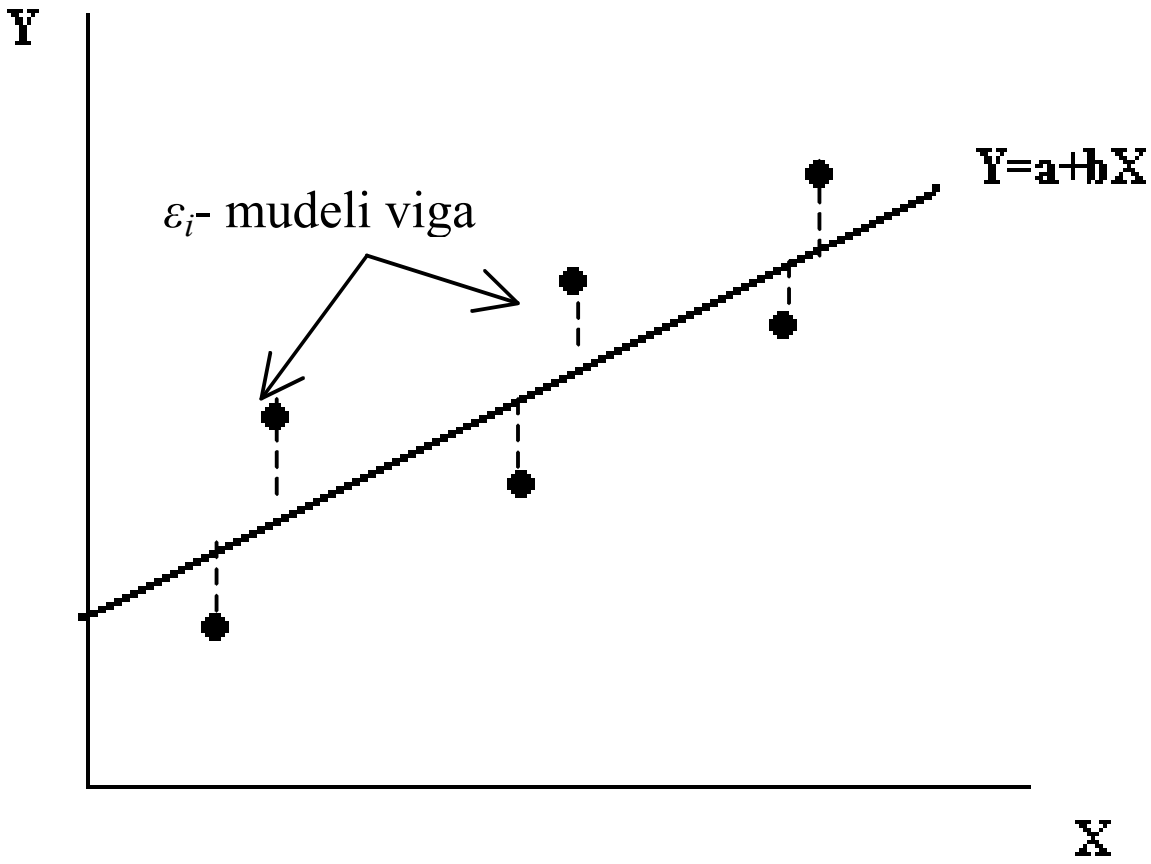
a on joone lõikumispunkt y-teljega,

b on joone kalle,

ε on mudeli viga.

Joonis 2. Regressioon, joone lõikumispunkt, kalle

Vähimruutude meetod



Joonis 3. Regressioon, mudeli vead

Mudeli viga ehk jääk [*residual*] ε:

kuivõrd konkreetne vaatlustulemus hälbib regressioonijoonest

- Eesmärgiks on konstrueerida sirge, mis kõige paremini iseloomustaks konkreetsete vaatlustulemuste vahelist seost. Leitakse sirge, mille puhul on vigade ruutude summa minimaalne (vähimruutude meetod).

Regressioonikordaja

- Lihtsa ehk paarisregressiooni korral huvitab meid kahe tunnuse vaheline seos.
  - $Y=\beta_0 + \beta_1 X_i + \varepsilon_{i,,}$ kus Y on sõltuv tunnus, X_i on i'nda indiviidi sõltumatu tunnuse väärtus, β_1 on joone kalle, β_0 on joone lõikumispunkt, $\varepsilon_{i,}$ on i'nda indiviidi väärtuse erinevus regressiooni joone poolt ennustatavast väärtusest.
- Tulemuste analüüsimisel ja tõlgendamisel huvitab meid suurus, mitme ühiku võrra muutub sõltuva tunnuse väärtus siis, kui sõltumatu tunnuse väärtus muutub ühe ühiku võrra.
- Seda seost iseloomustab regressioonikordaja ehk **β** (beeta) väärtus (tulemuste tõlgendamisel on oluline jälgida nii kordaja väärtust kui ka selle olulisust).

Mudeli sobivus

- Regressioonmudeli statistilist sobivust (st kuivõrd valimi põhjal leitud mudel kehtib järeldusi tehes populatsioonile) hindab F-test.

$$F = \frac{MS_M}{MS_R}$$

-*MSm* on mudelis olevate tunnuste keskmiste ruudud, *MSr* on jääkide keskmiste ruudud.

- F-test mõõdab, kui hästi sõltuva tunnuse ennustamist parandab sõltumatu tunnuse väärtuse teadmine võrreldes olukorraga, kui meil pole sõltumatu tunnuse väärtus teada.

• Statistiliselt sobivaks loetakse mudelit, mille F-testi väärtus on suur ja olulisuse tõenäosuse väärtus väike (<.01).

Mitmene regressioon

- Kahe või enama sõltumatu tunnuse mõju ühele sõltuvale tunnusele. Analüüsitavad tunnused on kõik <u>arvtunnused.</u>
  - $Y=\beta_0 + \beta_1X_1 + \beta_2X_2 +... + \beta_nX_n + \varepsilon_i$, kus Y on sõltuv tunnus, β_1 on esimese sõltumatu tunnuse koefitsient (X_1), β_2 on teise sõltumatu tunnuse koefitsient (X_2), ε_i, on i'nda indiviidi väärtuse erinevus regressiooni joone poolt ennustatavast väärtusest (*Nt sissetuleku ennustamine koolis käidud aastate, vanuse ja tööl käidud aastate alusel*).
  - Tihti uurija ei tea, millised või kui palju sõltumatuid muutujaid annavad rahuldava mudeli. On terve rida meetodeid, kuidas sõltumatuid tunnuseid mudelisse lisatakse – kas kõik korraga, järk-järgult vms.

Näide I

- **<u>Hüpotees:</u>** uudistesaate populaarsus pole tingitud mitte niivõrd saatejuhi poplaarsusest, vaid uudistesaatele eelneva saate populaarsusest (st nn “jäänukefekt”, vaatajad jäävad TV ette).
  - <u>Sõltuv tunnus:</u> uudistesaate populaarsus.
  - <u>Sõltumatu tunnus:</u> saatejuhi populaarsus, uudistesaatele eelneva saate populaarsus.
- **Tulemus (regressioon):** Regressioonanalüüsi tulemusena ei leidnud kinnitust püstitatud hüpotees.
  - St uudistesaate populaarsust ennustab uudistesaate juhi populaarsus (*st saatejuhi populaarsuse suurenedes 1 skaala palli võrra suureneb uudistesaate populaarsus 0,7 palli*). Uudistesaate populaarsus pole aga mõjutatud uudistesaatele eelneva saate populaarsusest.

Allikas: Hüpoteetilised andmed.

Näide II

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,753[a]	,568	,536	1,14292

a. Predictors: (Constant), uudistesaate juhi populaarsus, eelneva saate reiting

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	46,311	2	23,155	17,727	,000[a]
	Residual	35,269	27	1,306		
	Total	81,580	29			

a. Predictors: (Constant), uudistesaate juhi populaarsus, eelneva saate reiting

b. Dependent Variable: uudistesaate reiting

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	,924	,718		1,286	,209
	eelneva saate reiting	,066	,201	,070	,329	,744
	uudistesaate juhi populaarsus	,792	,242	,696	3,272	,003

a. Dependent Variable: uudistesaate reiting

R-ruut: determinatsiooni-kordaja, näitab mudeli headust, varieerub 0-1, suure väärtuse korral kirjeldab mudel andmeid hästi.

F-test: keskmiste ruutude määr, suur F-i väärtus ja väike olulisus **(Sig)** näitavad, et tulemused pole saadud juhuslikult.

T-test: suur T väärtus ja väike olulisustõenäosus – regress.kordaja 0-st erinev, ilmneb seos.

Beta: näitab mitme ühiku võrra suureneb sõltuv tunnus, kui sõltumatu tunnuse väärtus suureneb 1 ühiku võrra.

Näide III

Tabel: Regressioonanalüüsi tulemuste esitamine

(sõltuv tunnus: kihtikuuluvushinnang)

	Standardiseeritud regressiooni-kordaja β	t	Olulisus
Konstant		5,91	0,000
Rahaline jõukus	0,22	6,61	0,000
Konsumerism	0,16	5,28	0,000
Kogemus ja kontaktid läänemaailmas	0,09	3,28	0,001
Optimistlik eluhoiak	0,10	3,61	0,000
Arvuti kasutamise mitmekülgsus	0,07	2,46	0,014
Eesi-sisene liikuvus	0,09	3,29	0,001
Perekonnaseis	-0,07	-2,84	0,005
Sugu	0,07	2,71	0,007
Kodakondsus	0,06	2,33	0,020

Allikas: Kalmus, Veronika; Lauristin, Marju; Pruulmann-Vengerfedt, Pille (2004). Eesti elavik 21.sajandi algul. Ülevaade uurimuse Mina. Maailm. Meedia. tulemustest, Tartu: Tartu Ülikooli Kirjastus.

Mittelineaarne regressioon

- Võimaldab analüüsida ühe või enama sõltumatu tunnuse mõju ühele sõltuvale tunnusele juhul, kui analüüsitavad tunnused on kategoriaalsed (ordinaal- või nominaalskaalal).
  - Lineaarses regressioonis on huvi all, kuivõrd varieeruvus sõltumatus tunnuses aitab ennustada varieeruvust sõltuvas tunnuses.
  - Mittelineaarses regressioonis on huvi all, kuivõrd kuulumine sõltumatu tunnuse kindlasse kategooriasse aitab ennustada kuulumist sõltuva tunnuse konkreetsesse kategooriasse.
  - Nt kolme ja enama võõrkeele oskajad on võrreldes 2 ja vähema võõrkeelte oskajatega hinnanud enda aktiivsust lisateenistuse hankimisel kõrgemaks.

Logistiline regressioon

- Logistilist regressiooni kasutatakse olukordades, kus soovitakse subjekte klassifitseerida teatud sõltumatute tunnuste kogumiku alusel.
  - Kõige laiema kasutusvõimalusega on multinoomne logistiline regressioon. Sõltuv tunnus võib erinevalt teistest meetoditest olla kategoriaalne (vt nt binaarne logistiline regressioon), sõltumatu tunnus arvuline või kategoriaalne
  - Logistiline regressioon kasutab suhtelise tõenäosuse mõistet, st uuritakse sündmuse toimumise võimalust sündmuse mittetoimumise seisukohalt. Kui toimumine on tõepärasem kui mittetoimumine, siis on see suhe arvust 1 suurem, vastupidises olukorras aga väiksem.
  - Suhteliste tõenäosuste suhet mõõdetakse regressioonikordajaga eksponentastmel. Logaritmitud regressioonikordajad (st mitte eksponentastmel) on intuitiivselt raskemini mõistetavad.

Allikas: Tooding 2007.

Mudeli sobivus

- Logilistilise regressiooni korral hinnatakse mudeli sobivust Pseudo R-ruudu (Nagelkerke jms) ja Hii-ruudu kordajate abil.
  - Kui mudeli headust hindava Hii-ruudu väärtus on statistiliselt oluline ($p<0.01$) ja Nagelkerke väärtus suur (lähedane 1-le), võib leitud mudelit pidada statistiliselt sobivaks.
  - Parima mudeli leidmiseks tuleb erinevate tunnustega mudeleid võrrelda omavahel. Statistiliselt sobivaimaks loetakse mudelit, kus Nagelkerke kordaja väärtus on suurim.
- Üksikute sõltumatute tunnuste olulisust sõltuva tunnuse ennustamisel näitavad logaritmitud Beeta-kordajad.
  - Kordajate statistilist olulisust hinnatakse Wald'i kordaja abil. Võimaldab hinnata seda, millisel sõltumatutest tunnustest on kõige suurem mõju sõltuvale tunnusele.

Tõlgendamine

- Puudub numbriline nullpunkt, mille alusel sõltumatute tunnuste abil ennustada sõltuvat tunnust. Võetakse kokkuleppeline kategooria, millele vastandatakse teisi kategooriaid.
  - Võrreldav taustakategooria peaks olema selge sisuga ning teistest kategooriatest hästi eristatav. Tehniliselt on taustakategooriaks kõige lihtsam võtta kas skaala esimest või viimast kategooriat (*nt võrreldakse noorimat ja keskmist vanusegruppi kõige vanemaga*).
  - Hindamaks tõenäosust, et sõltuva tunnuse väärtus on võrdne selle kõrgeima/viimase väärtusega, võrreldakse mudelis logit-kordajaid või kordajaid eksponentastmel (=Beeta-kordaja tavalises lineaarses regressioonis).

Allikas: Tooding 2007

Näide I

- **Hüpotees:** suurema võõrkeelte oskusega indiviidid on majandusliku kapitali poolest rikkamad.
  - Eeldatakse, et kommunikatiivne kapital (võõrkeelte oskus) on muudetav teisteks kapitalideks, sarnaselt sotsiaalse ja kultuurilise kapitaliga.
- Multinoomse logistilise regressiooni abil analüüsiti järgmiseid tunnuseid:
  - sõltuv tunnus majanduslik kapital (koondindeks tunnustest: kuivõrd perel jätkub raha erinevateks väljaminekuteks, nt riiete ostmiseks, õppimiseks jne)
  - sõltumatu tunnus võõrkeelte oskuse ja kasutamise ulatus, ankeedi keel, vanus, sugu, haridus, sissetulek.

Allikas: Uuring Mina.Maailm.Meedia 2002.

Näide II

Model Fitting Information

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	1611,011			
Final	1113,529	497,482	44	,000

Pseudo R-Square

Cox and Snell	,483
Nagelkerke	,506
McFadden	,214

Chi-square: mudel kirjeldab andmeid hästi, kuna mudeli sobivust hindav Hii-ruudu olulisus p<0,01.

Nagelkerke: mudel kirjeldab täiuslikult ennustatavat tunnust juhul, kui kordaja väärtus=1. Mudeli sobivuse hindamiseks võrrelda erinevate tunnustega mudeleid omavahel ning valida parima kirjeldusjõuga mudel (suurima Nagelkerke kordaja väärtusega).

Näide III

Likelihood Ratio Tests

Effect	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	1734,204[a]	,000	0	
KEELEGR	1755,310	21,106	12	,049
V674	1757,911[b]	23,707	4	,000
VANUS	1887,725[b]	153,520	20	,000
V593	1773,244	39,040	4	,000
HARIDUS	1793,357[b]	59,153	8	,000
SISSETUL	1920,768	186,564	16	,000

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

a. This reduced model is equivalent to the final model because omitting the effect does not increase the degrees of freedom.

b. There is possibly a quasi-complete separation in the data. Either the maximum likelihood estimates do not exist or some parameter estimates are infinite.

Likelihood ratio: sõltuvat tunnust "majanduslik kapital" aitavad kirjeldada ankeedi keel, vanus, sugu, haridus, sissetulek ($p<0,01$). Võõrkeelte oskus ja kasutamise ulatus pole oluline majandusliku kapitali ennustamisel ($p>0,01$).

Näide IV

Tabel. Huvi jälgida uudiseid eri riikides toimuva kohta sots.dem tunnuste lõikes (kõrge huvi)

		Läti (1)	Soome	Venemaa	Inglismaa
Sots.dem.	Sündinud riigis (2)	0.335*	0.528***	-1.200**	
tunnused	Mehed (3)	0.205*	0.226*	0.484***	
	Vanus: 57-74 (4)	0.459***			-0.737***
	Vanus: 42-56	0.799***			-0.477***
	Vanus: 29-41	0.551***		0.339**	
	Haridus: kõrg (5)	1.272***	1.163***	1.282***	1.255***
	Haridus: kesk	0.296**	0.367**	0.456***	0.370**
	Sissetulek: üle 4001 (6)		1.06***		0.387**
	Sissetulek: 2001-4000		0.584***	0.299**	0.611***
	Sissetulek: 1501-2000		0.297*		0.334*
	Chi-square	211***	326***	254***	268***
	R-Square (Nagelkerke)	0.19	0.282	0.226	0.237

- Logaritmitud regressiooni kordajad: * p<0.1; ** p<0.05; *** p<0.01; (1) võrdluskategooriaks “madal huvi riigi suhtes” (=0, tabelist jäetud välja, (2) võrdlus: pole sündinud riigis, (3) võrdlus: naised, (4) võrdlus: 15-28, (5) võrdlus: põhiharidus, (6) võrdlus: alla 1500 kr.

Allikas: Uuring Mina.Maailm.Meedia 2002.

Näide V

Tabel: Lemmiktegevuste prognoosimine 1979 ja 1997.aastal

Tegevus		*B*	*Exp (B)*	*p*
Kodu, pere	Naine*	0,540	1,716	***
	1997 **	0,396	1,486	***
Teater, kontsert jms	Naine	1,599	4,946	***
	1997	-0,681	0,506	***
Sõbrad	Naine	0,410	1,507	***
	1997	0,571	1,770	***
Raadio, TV ms	Naine	-0,071	0,931	
	1997	-0,938	0,391	***
Töö	Naine	0,065	1,068	
	1997	0,265	1,304	***
Erialane enesetäiendamine	Naine	0,298	1,348	***
	1997	0,397	1,487	***
Sport	Naine	-0,618	0,539	***
	1997	0,337	1,400	***
Peod	Naine	0,315	1,371	***
	1997	1,103	3,014	***

* taustaks mehed
** taustaks 1979
*** p<0,01
** p<0,05
* p<0,1

Allikas: Paalandi, V. (2001)

Näide VI

Tabel: Elutegevuste mõju prognoosimisel sotsiaal-demograafiliste tunnuste võrdlus

Sõltumatu muutuja	töö	erialane enesetäiendamine	pere, kodu	sport	poliitiline tegevus
Sugu: naine	0	++	+++	---	---
Rahvus: mitte-eestlane	--	--	---	-	0
Perekonnaseis: mitte abielus	+	0	--	0	0
Kooliskäidud aastaid: kuni 11 *	0	---	+++	0	0
Koolikäidud aastaid: 12-15*	0	--	+++	0	0
Elukoha tüüp: Tallinn**	0	0	---	+	0
Elukoha tüüp: muu linn**	0	++	-	0	0

* taustaks: 16 ja rohkem aastat
** taustaks maa
+++', '---': p<0,01
++', '--': p<0,05
+', '-': p<0,1

Allikas: Paalandi, V. (2001)

Ülesanne

- **Lugeda artiklit ja vastata järgnevatele küsimustele:**
  - ***Milliseid meetodeid kasutatakse andmete analüüsimisel?***
  - ***Milliste seoseid (milliste tunnuste vahel) analüüsib esitatud mudel? Mida võib mudeli põhjal sisuliselt järeldada?***
  - ***Milliseid statistilisi (seose)kordajaid on andmete esitamisel kasutatud?***

Iseseisvaks lugemiseks

- Taagepera, R. (2008). From descriptive to predictive approaches. Rmt: *Making social sciences more scientific*, Oxford: Oxford University Press, lk 187-198.
- Tooding, L-M. (2007). Ptk 6. Seosemudelid. Rmt: *Andmete analüüs ja tõlgendamine sotsiaalteadustes*, Tartu: Tartu Ülikooli Kirjastus, lk 254-310.
- Field, F. (2000). Chapter 4. Regression. Rmt: *Discovering statistics using SPSS for Windows: advanced techniques for the beginner*, London etc: Sage. Lk 103-162.

TARTU ÜLIKOOL
Ajakirjanduse ja kommunikatsiooni instituut

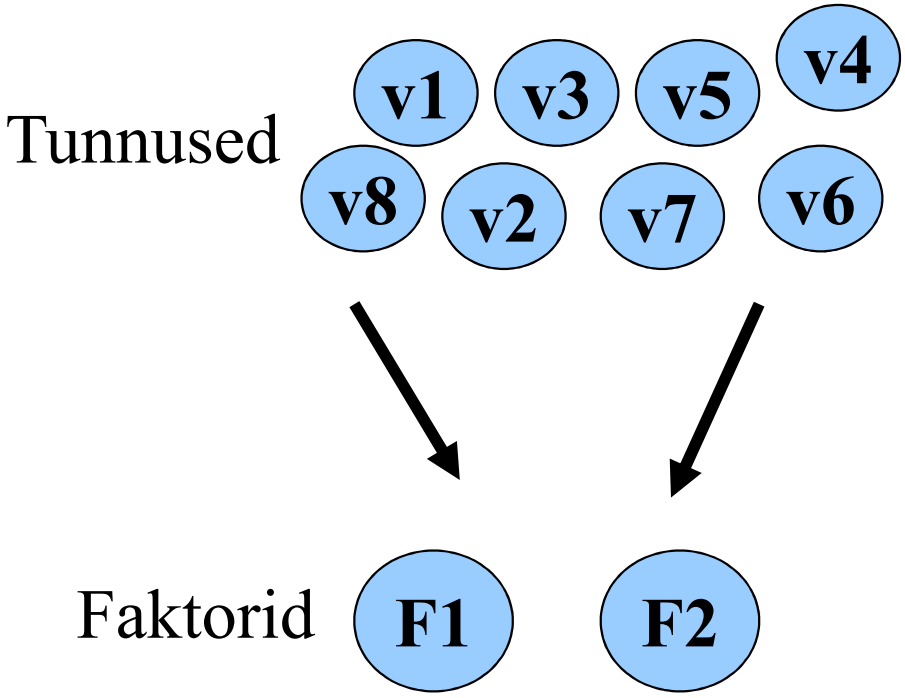
Faktor- ja klasteranalüüs. Analüüs indeksitega

Kvantitatiivne andmeanalüüs

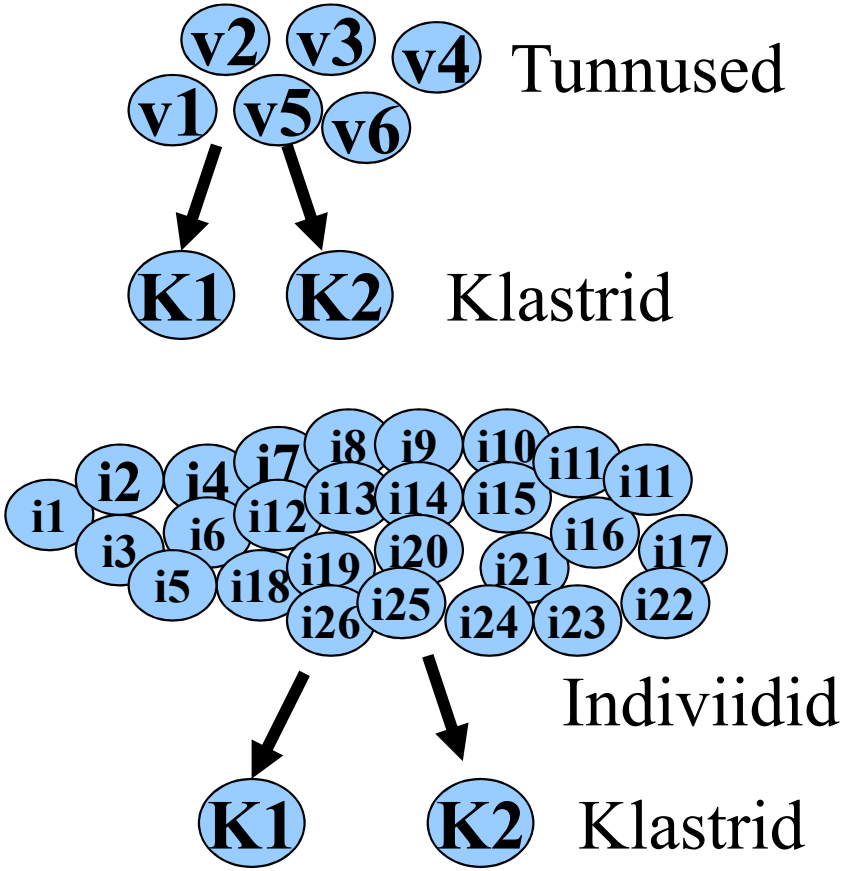
Anu Masso (PhD)

Andmete rühmitamine

- Mitmemõõtmelised analüüsitehnikad, mille eesmärgiks on tunnuste või indiviidide grupeerimine teatud varjatud dimensioonide või eelnevalt määratud tunnuste alusel.



FAKTORANALÜÜS

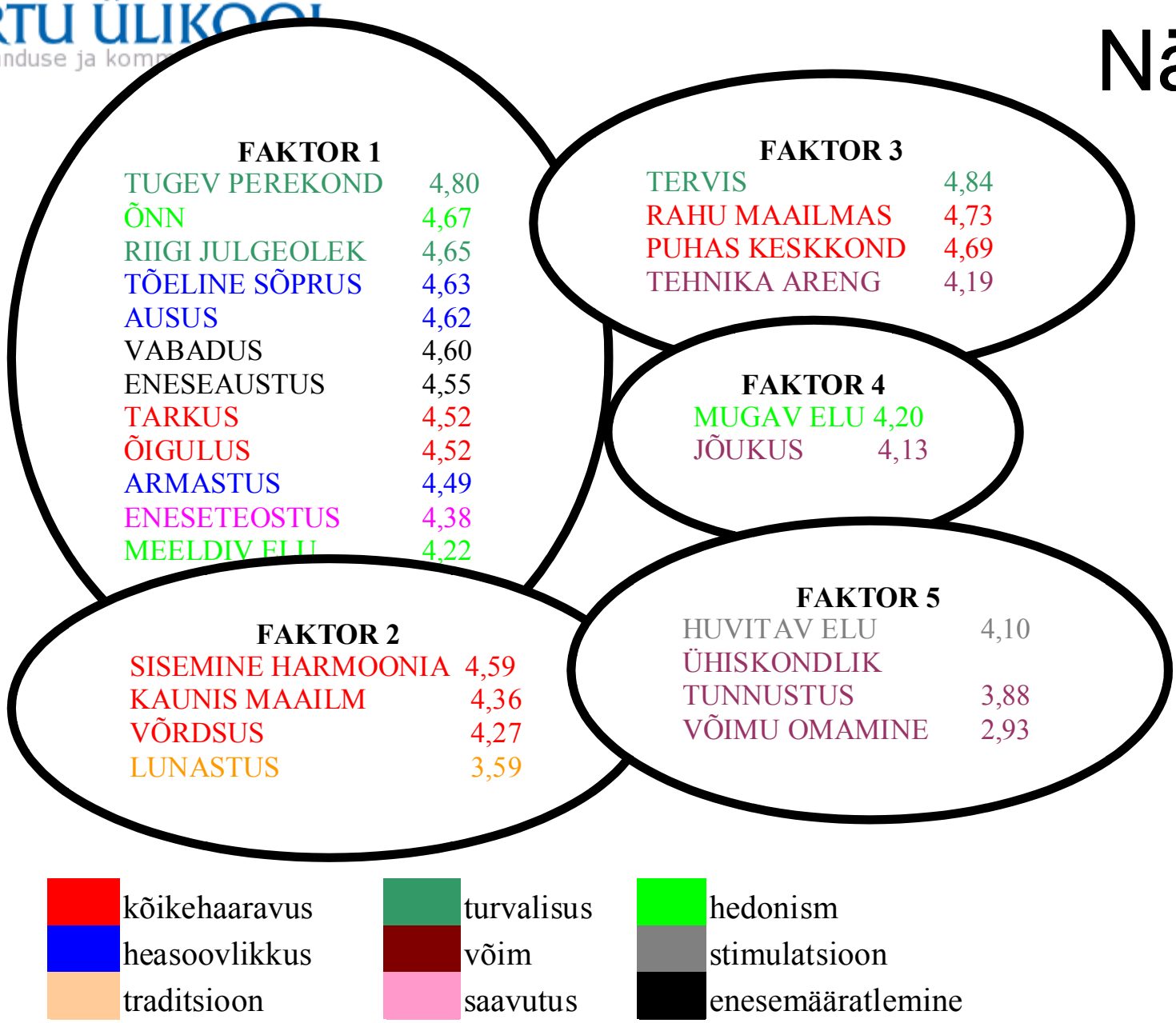


KLASTERAN

Faktoranalüüs

- Mitmemõõtmeline analüüsitehnika, kus suurt arvu vaadeldavaid tunnuseid kirjeldatakse väikese arvu uute tunnuste abi, mida nimetatakse faktoriteks.
  - Tehnika on suunatud tunnuste struktuurile, st andmestikus ei eristata sõltuvat ega sõltumatut tunnust. Eesmärgiks on andmestiku koondamine ja tulemuste üldistamine (sarnaselt indeksite moodustamisele).
  - Faktoranalüüs kvantitatiivse analüüsitehnikana loob tunnuste tüpoloogiad tuginedes tunnuste vahelistele kovariatsioonidele jms matemaatilistele operatsioonidele.
  - Faktoranalüüsi eelduseks on suure hulga intervall- või ordinaalskaalal mõõdetud tunnuste vahel korrelatsioonide olemasolu.

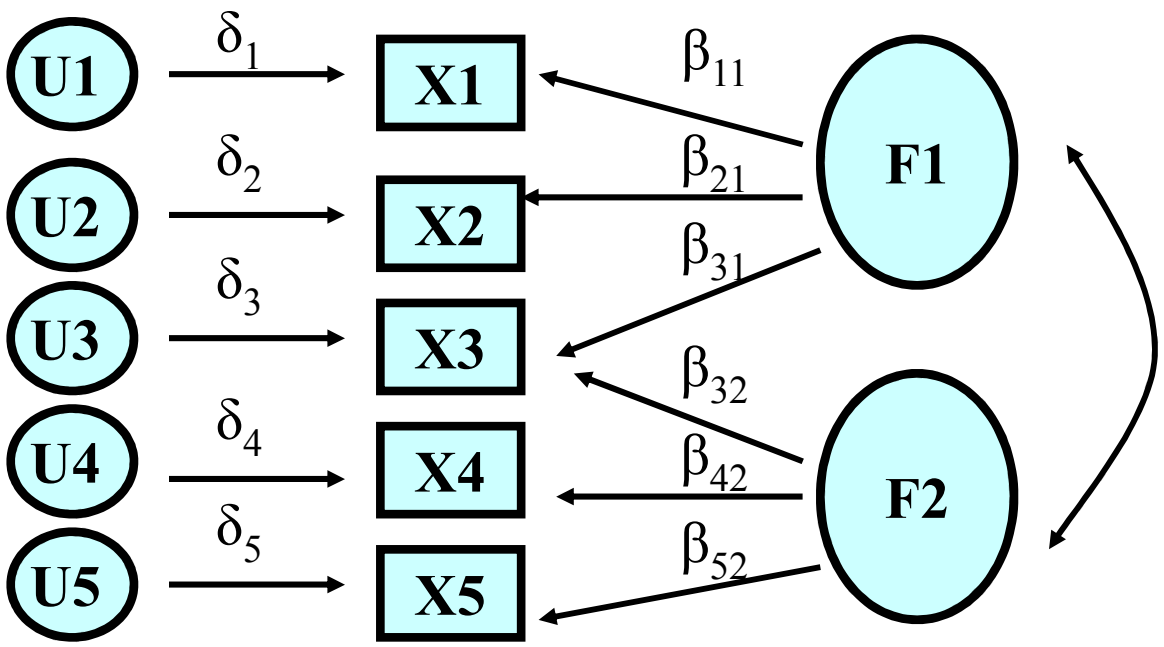
Näide



Joonis: 25 väärtustunnuse struktuur (erinevate värvidega märgitud Schwarz'i väärtusdimensioonid, numbrid näitavad keskväärtust)

Allikas: Masso, A., Vihalemm, T. 2003

Arvutamine I



X – vaadeldavad tunnused;

F – faktorid või latentsed tunnused, st otseselt mitte mõõdetavad tunnused, kirjeldavad algtunnuseid;

U – tunnuse X "unikaalne" faktor, mida iseloomustab jääk (vrdl viga lineaarses regressioonis);

B – faktorlaadung, mis näitab, kui suur muutus F'is mõjutab X'i;

δ – näitab, kui suur on unikaalse faktori U mõju X'ile.

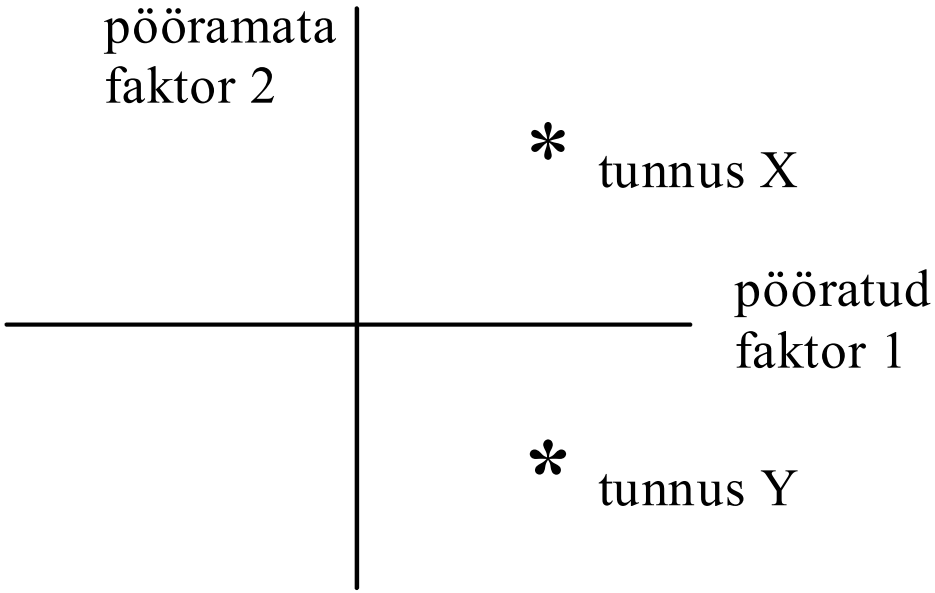
Arvutamine II

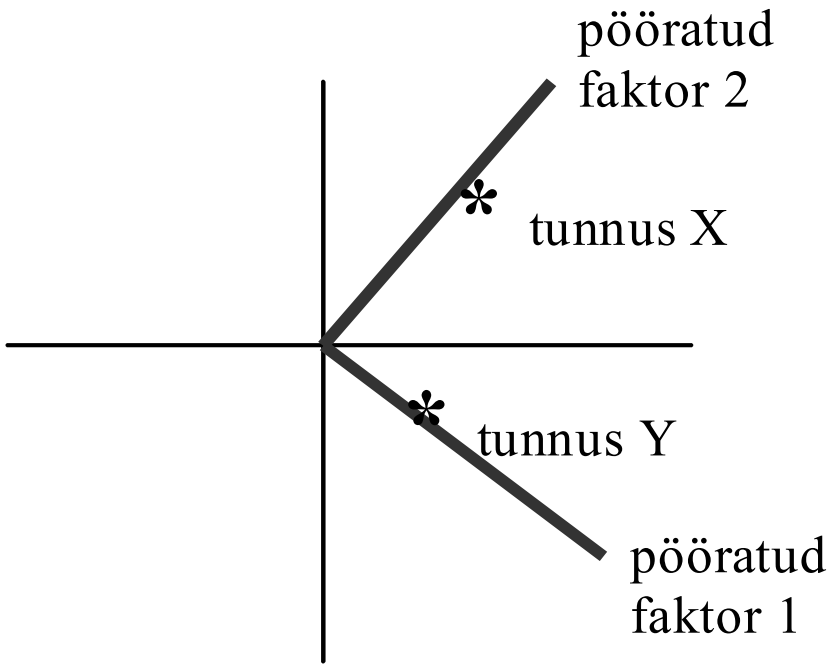
Lin.kombinatsioonid	**Faktormaatriks**		**Kommunaliteet**
	F1	**F2**	h_i^2
$X_1=b_{11}F_1+b_{12}F_2+d_1U_1$	b_{11}	b_{12}	$\sum b_{1j}^2 = h_1^2$
$X_2=b_{21}F_1+b_{22}F_2+d_2U_2$	b_{21}	b_{22}	$\sum b_{2j}^2 = h_2^2$
$X_3=b_{31}F_1+b_{32}F_2+d_3U_3$	b_{31}	b_{32}	$\sum b_{3j}^2 = h_3^2$
$X_4=b_{41}F_1+b_{42}F_2+d_4U_4$	b_{41}	b_{42}	$\sum b_{4j}^2 = h_4^2$
$X_5=b_{51}F_1+b_{52}F_2+d_5U_5$	b_{51}	b_{52}	$\sum b_{5j}^2 = h_5^2$
Omaväärtus:	$\sum b_{i1}^2$	$\sum b_{i2}^2$	

Kommunaliteet on konkreetse tunnuse faktorlaadungite ruutude summa, mis näitab unikaalsete faktorite kirjeldusjõudu.

Omaväärtus on faktorlaadungite ruutude summa konkreetse faktori korral, näitab konkreetse faktori kirjeldusjõudu.

Telgede pööramine





Teljed joonisel tähistavad kahte faktorit, st mida lähemal teljele asub tunnus, seda kõrgem on faktorlaadung konkreetses faktoris.

Pööramise eesmärgiks on optimeerida tulemust selliselt, et leida selgemat faktorstruktuuri, st tunnused saavad ühe faktori juures kõrge, teise juures madala faktorlaadungi.

FA meetodid I

- **Peakomponentide meetod**
  - Eesmärgiks on leida väike arv komponente, mis selgitaksid võimalikult palju andmetes olevast koguvarieeruvusest.
  - Kasutatakse esmases analüüsis tunnuste üldise struktuuri uurimiseks. Komponendid on kui lineaarsed kombinatsioonid vaadeldud tunnuste vahel.
- **Faktoranalüüsi meetod**
  - Eesmärgiks on leida faktorid, mis selgitaksid võimalikult palju ühisest varieeruvusest.
  - Faktoranalüüsi kasutatakse, et selgitada teatud kompleksse fenomeni taga olevaid varjatud faktoreid.

FA meetodid II

- Kinnitav faktoranalüüs [*confirmatory factor analysis*]
  - Faktorite arvu soovitakse piirata konkreetsele arvule, otsustus tehakse enne analüüsi (tuginedes empiirilistele või teoreetilistele teadmistele).
  - Iseloomustab ortogonaalne telgede pööramine, mille tulemusena faktorid on omavahel nõrgalt korreleeritud (nt *varimax*). Konkreetne tunnus on tugevamalt seotud ühe faktoriga (=selgem faktorstruktuur).
- Uuriv faktoranalüüs [*explorative factor analysis*]
  - Puuduvad eelnevad piirangud, selged hüpoteesid andmete struktuuri kohta.
  - Iseloomustab mitte-ortogonaalne telgede pööramine (st saadakse omavahel nõrgalt korreleeritud faktorid, mistõttu üks tunnus võib samaaegselt kuuluda mitmesse faktorisse) või jäetakse teljed pööramata.

Faktorite arv I

Faktorite arvu määramiseks, tunnuste üldise struktuuri leidmiseks sobib peakomponentide meetod.

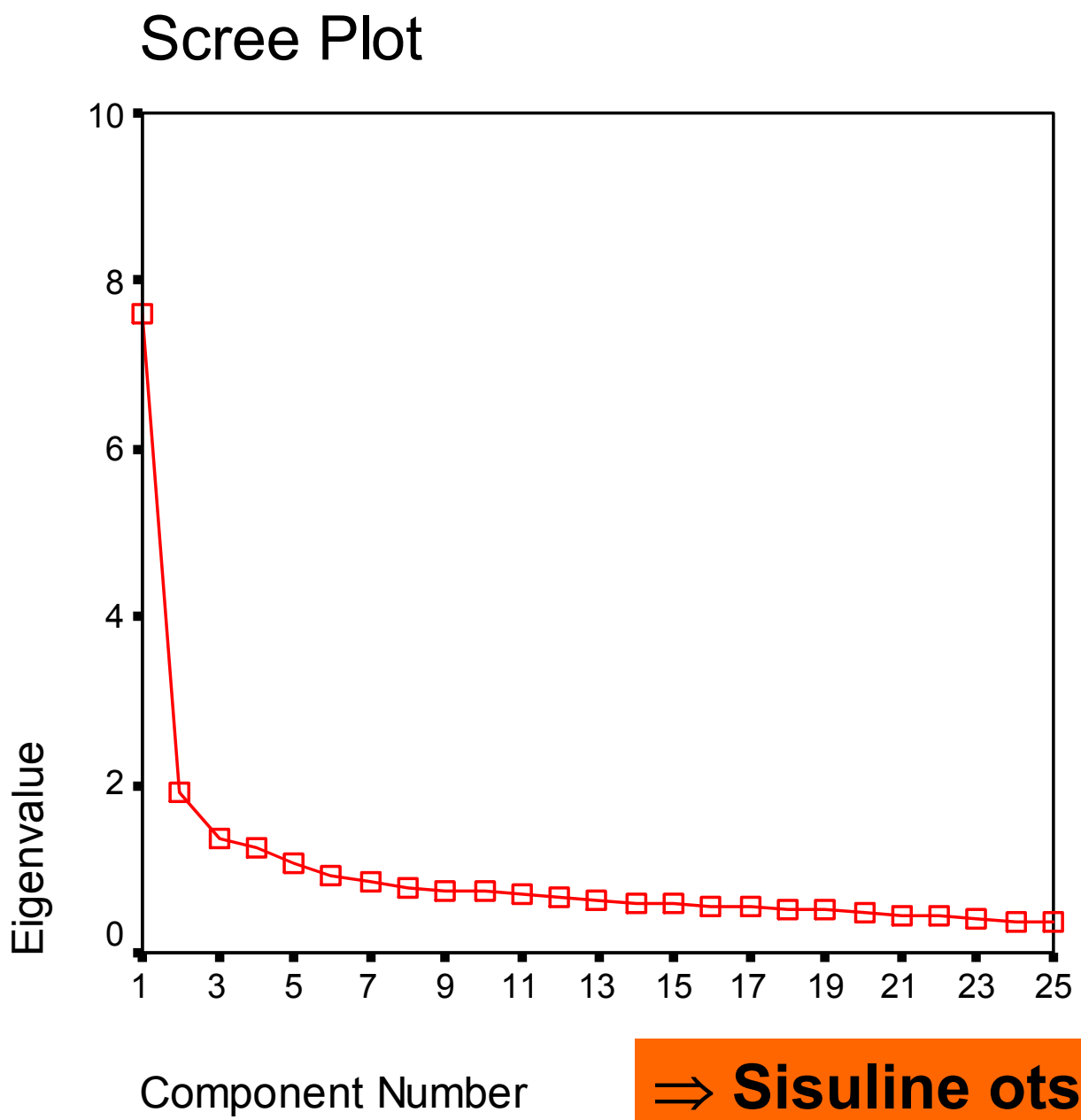
Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	7,602	30,406	30,406	7,602	30,406	30,406
2	1,905	7,618	38,024	1,905	7,618	38,024
3	1,357	5,427	43,451	1,357	5,427	43,451
4	1,261	5,045	48,497	1,261	5,045	48,497
5	1,058	4,232	52,728	1,058	4,232	52,728
6	,934	3,738	56,466			
7	,847	3,387	59,853			
8	,783	3,132	62,985			
9	,748	2,991	65,976			
10	,720	2,879	68,855			
11	,680	2,721	71,577			
12	,655	2,622	74,198			
13	,634	2,536	76,735			
14	,598	2,393	79,128			
15	,593	2,372	81,500			
16	,566	2,265	83,765			
17	,541	2,164	85,929			
18	,521	2,082	88,011			
19	,508	2,031	90,042			
20	,465	1,859	91,902			
21	,452	1,807	93,708			
22	,445	1,779	95,487			
23	,398	1,591	97,078			
24	,377	1,508	98,586			
25	,354	1,414	100,000			

Extraction Method: Principal Component Analysis.

Kaiseri kriteerium: kasutatakse faktorite arvu määramiseks. Selle järgi tuleb analüüsida vaid neid faktoreid, mille omaväärtus on suurem kui 1; Kriteeriumi puuduseks see, et võidakse analüüsida liiga palju faktoreid.

Faktorite arv II



Catell'i testi: teine võimalus faktorite arvu määramiseks. Selle järgi tuleb analüüsida omaväärtuste joonist ning leida punkt, kus joonis enam järsult ei lange.

Testi puuduseks see, et võidakse analüüsida liiga vähe faktoreid.

⇒ **Sisuline otsustus olulisem!**

Tulemuste tõlgendamine

- Antud faktoriga seotuks loetakse tunnused, mille faktorlaadung antud faktoris on >=0,3.
  - Suurima faktorlaadungiga tunnused antud faktoris määravad faktori iseloomu ning on aluseks faktori nimetamisel; selliseid tunnuseid nimetatakse faktori tuum- ehk põhitunnusteks.
  - Ühe faktoriga võib olla seotud mitu tunnust, st faktorlaadung >3 on rohkemas kui ühes faktoris; selliseid tunnuseid nimetatakse faktori lisa- ehk mitmedimensionaalseteks tunnusteks.
  - Faktorite selgema struktuuri leidmiseks võrreldakse tunnuste struktuuri nt erinevates sots.-dem. gruppides; tuumtunnusteks loetakse erinevates gruppides “stabiilsed” tunnused. Mitmedimensionaalsed tunnused võib analüüsist välja jätta.

Rotated Factor Matrix[a]

	Factor				
	1 FAKTOR Isiksuslik ja sotsiaalne tasakaal	**2 FAKTOR Hingeline tasakaal**	**3 FAKTOR Keskkond ja füüsiline heaolu**	**4 FAKTOR Materiaalne kindlustatus ja heaolu**	**5 FAKTOR Enesekehtestamine**
ÕNN	**,623**	,132	,180	*,380*	,002
ENESEAUSTUS	**,564**	,267	,168	,130	,247
ARMASTUS	**,539**	,027	,108	,118	,200
AUSUS	**,535**	*,411*	,269	-,088	-,004
MEELDIV ELU	**,489**	,058	-,006	*,396*	,263
TARKUS	**,478**	*,318*	,211	-,029	,209
TUGEV PEREKOND	**,474**	,229	*,330*	,166	-,019
ENESETEOSTUS	**,461**	,097	,118	,065	*,435*
TÕELINE SÕPRUS	**,441**	,273	,173	,037	,141
ÕIGLUS	**,432**	*,348*	,155	,060	,194
RIIGI JULGEOLEK	**,392**	*,327*	,282	,109	-,013
VABADUS	**,362**	,262	,204	,223	,137
LUNASTUS	,083	**,619**	,059	,034	,148
KAUNIS MAAILM	,207	**,526**	*,345*	,080	,095
VÕRDSUS	,175	**,418**	,091	,192	-,006
SISEMINE HARMOONIA	*,341*	**,415**	,223	,122	,037
TERVIS	,228	,067	**,690**	,161	,080
PUHAS KESKKOND	,213	,172	**,680**	,007	,045
RAHU MAAILMAS	,254	*,347*	**,428**	,155	-,110
TEHNIKA ARENG	,052	,203	**,382**	,199	*,309*
MUGAV ELU	,073	,160	,183	**,649**	,188
JÕUKUS	,136	,068	,068	**,532**	,269
VÕIMU OMAMINE	,090	-,026	-,019	,206	**,544**
ÜHISKONDLIK TUNNUSTUS	,220	*,378*	,007	,174	**,456**
HUVITAV ELU	,223	,154	,153	,274	**,364**

Extraction Method: Principal Axis Factoring.
Rotation Method: Varimax with Kaiser Normalization.

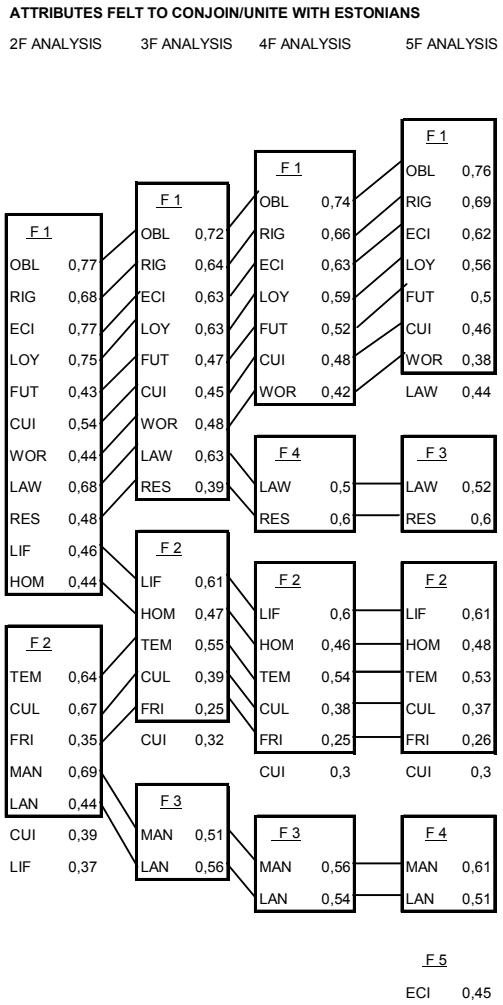
a. Rotation converged in 9 iterations.

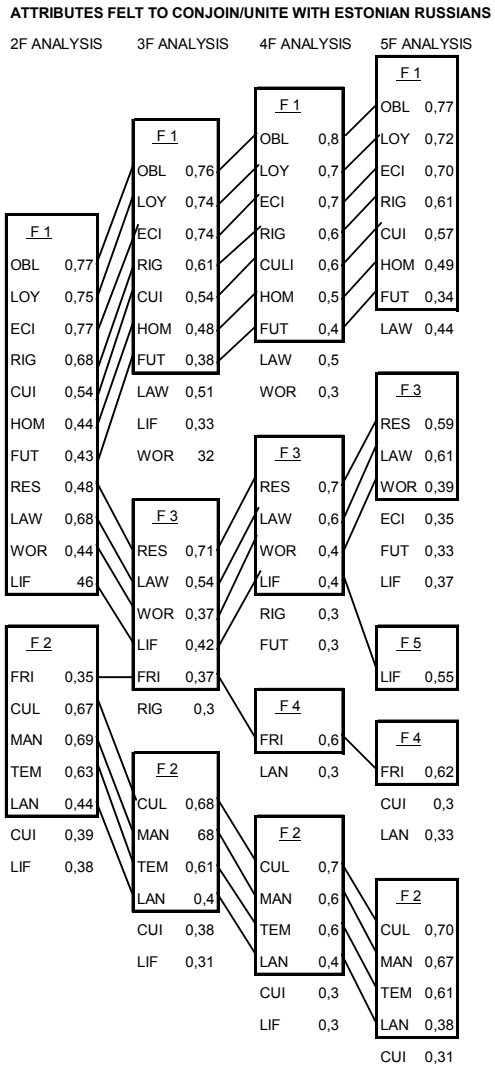
Näide I

Tuum-tunnused on märgitud siniselt ja rasvaselt.

Lisa-tunnused on märgitud siniselt ja kaldkirjas.

Näide II





LNG – common language
LIF – lifestyle
TEM –similar temperament
HOM – homeland
CUL – similar culture
CUI – cultural interests

RES – common place of residence
ECI – economic interests
WOR – daily worries
LAW – state, laws
LOY – loyalty towards the state

RIG – economic and political rights
OBL – economic and political obligations
FRI – friends, acquaintances
FUT – common future, aims
MAN – manners

Faktorgraafiline meetod FA tulemuste esitamiseks:

* Eristatakse faktorite tuum- ja lisatunnuseid.

* Näidatakse struktuuri muutust 2-, 3-, ja n-faktorilise lahendi korral.

* Võrreldakse faktorstruktuuri eri valimi alagruppides.

Allikas: T. Vihalemm, A. Masso 2002

Ülesanne

- Lugeda teksti (Vihalemm, Kalmus 2008) ja vastata järgmistele küsimustele:
  - *Millist andmete rühmitamise meetodit on kasutatud?*
  - *Millise meetodiga on analüüsitud seoseid leitud rühmade ja taustatunnuste vahel?*
  - Mis on iseloomulik transitsioonikultuurile (tuginedes analüüsitud Eesti näitele)?

Klasteranalüüs

- Mitmemõõtmeline analüüsitehnika, mida kasutatakse erinevate tunnuste või indiviidide gruppide leidmiseks andmestikus. Võimaldab andmestikku koondada ja üldistada.
  - Klasteranalüüs leiab struktuurid andmetes, võimaldab uurida klastritesse kuuluvate indiviidide omadusi, selgitamata põhjuseid, *miks* sellised struktuurid esinevad.
  - Klasteranalüüs loob tunnuste või indiviidide tüpoloogiad tuginedes indiviidide või tunnuste vahelistele kaugustele (nt Eukleidilistele kaugustele).
  - Tüpoloogiad kas teoreetilised (tüüpi moodustavad tunnused on eelnevalt teada) või empiirilised (aluseks on vastajate jagunemine tüüpi moodustavate tunnuste alusel).

Näide

Tabel: Elulaadi tüpoloogia algse 22 tunnuse alusel.

Mitme-külg-selt aktiiv-ne	Tööle orient., kultuuri-lembeli-ne	Hasart-ne, meele-lahu-tuslik	Tehnili-selt harras-tuslik	Uue meedia, seltsielu kesk-ne	Kodu-keskne tradit-siooniline	Kirja-sõna-keskne, tradit-siooni-line	Pas-siivne
9%	15%	5%	14%	13%	15%	13%	17%
Kino, teater, näitus, klubid, matkad, jms.	Teater, näitus, töö-ja ärireisid, konveren ts jms.	Kino, kohvik, klubi, reisid jms.	Töö-ja ärireisid, matkad, looduses jms.	Kino, klubid, külas jms.	Kirik, ilukirjandus , aiatööd, käsitöö jms.	Kirik, ilukirjand us jms.	*Ei tegele praktilis elt millega gi.*

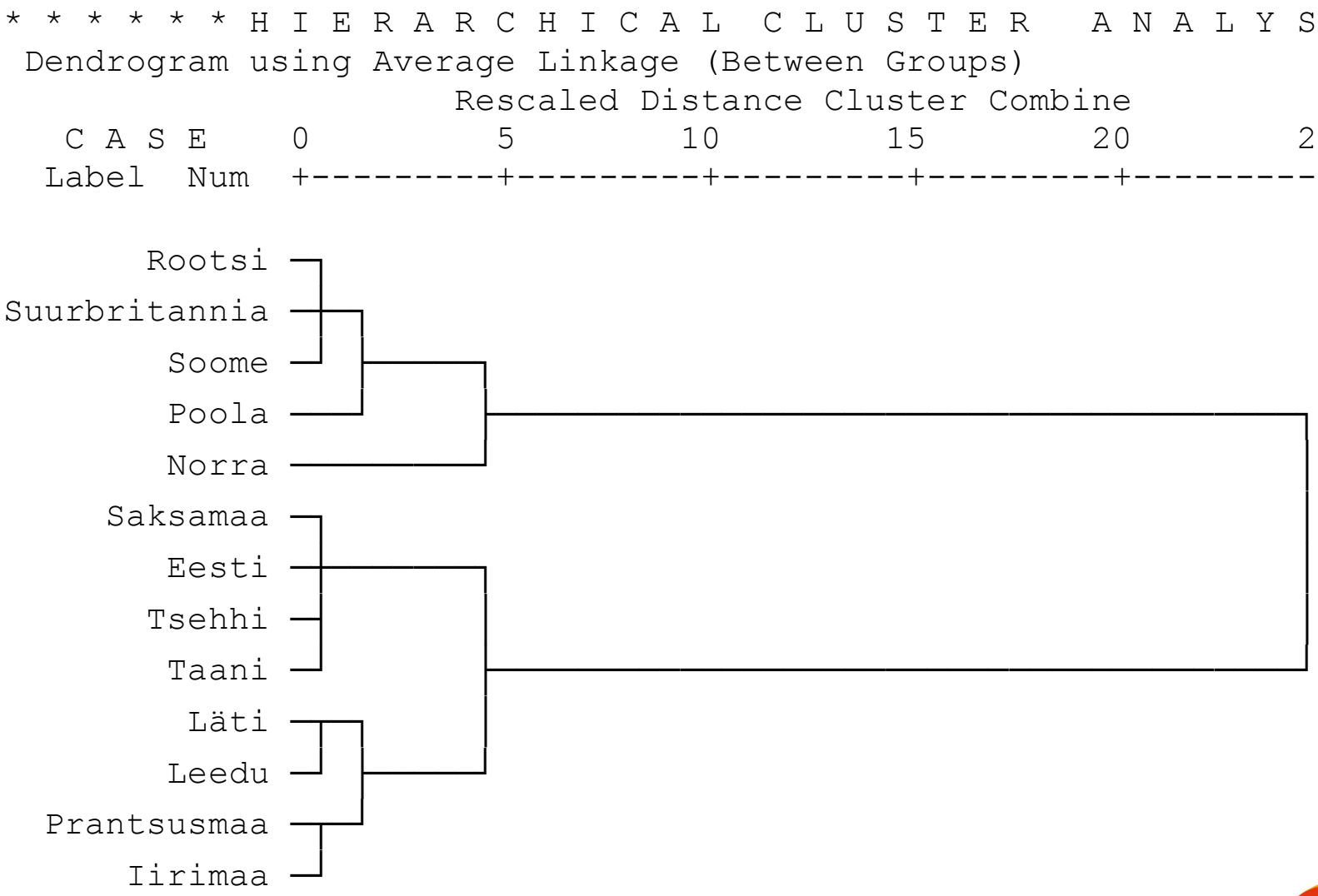
Allikas: Nigul, A. (2004). Elulaad Eestis, rmt. V. Kalmus, M. Lauristin, P. Pruulmann-Vengerfeldt (toim) Eesti elavik 21.sajandi algul, Tartu:

KA meetodid

- Hierarhiline klasteranalüüs
  - Klastrid kombineeritakse süstemaatiliselt tuginedes kaugustele tunnuste vahel. Tulemuseks teatud homogeensed grupid.
  - Võimalik grupeerida nii indiviide kui ka tunnuseid. Tunnuste grupeerimisel sarnane faktoranalüüsiga. Indiviidide grupeerimisel on sobiv väikese vaatluste arvu korral (st N<200).
- K-keskmiste meetod
  - Leiab andmetes teatud algoritmide alusel suhteliselt homogeensed indiviidide grupid. Eelnevalt on vajalik määrata soovitav klastrite arv.
  - Kasutatakse suure vaatluste arvu korral (st N>200), nn “kiire klasterdamismeetod”.

Hierarhiline KA

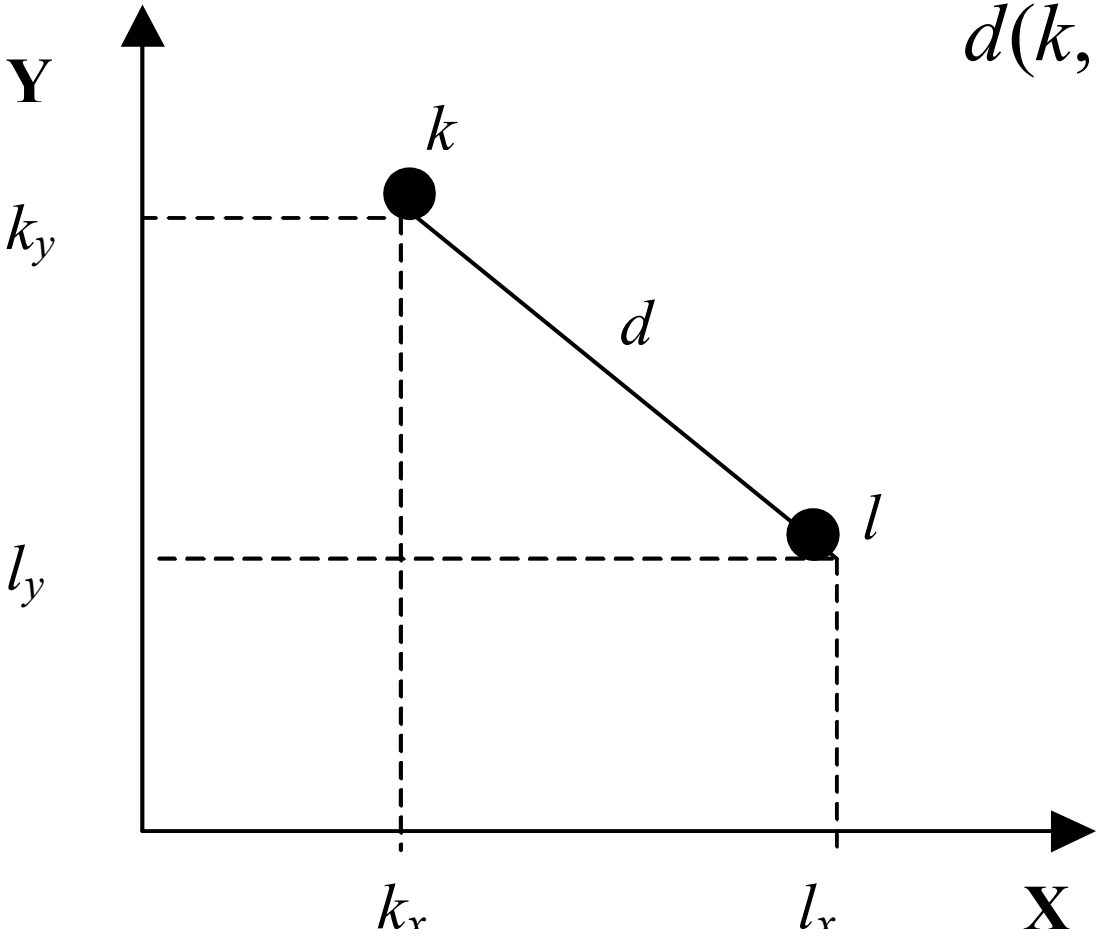
Hierarhilise KA tulemus: Kuivõrd on riike võimalik grupeerida järgmiste andmete alusel: *kodune Interneti ühendus, kulutused IKT'le, regulaarsed Interneti kasutajad jne.*



Allikas: Eurostat 2007.

Arvutamine I

Uurimisühikute vaheliste kauguste arvutamisel kõige sagedamini kasutatavaks meetodiks Eukleidiliste kauguste meetodit, st geomeetriliste distantside loomine multidimensionaalses ruumis.



$$d(k,l) = \sqrt{(k_x - k_y)^2 + (l_x - l_y)^2}$$

Eukleidiline kaugus on ruutjuur erinevuste ruutude summast konkreetsete kahe tunnuse koordinaatide vahel. Eukleidilisi kauguseid sobiv arvutada vaid juhul, kui tunnused on mõõdetud samal skaalal.

Arvutamine II

Olgu meil viis uuritavat ühikut, mida soovitakse klassifitseerida. Kauguseid analüüsitavate ühikute vahel iseloomustab järgmine maatriks:

$$\mathbf{D_1}= \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{pmatrix} 0 & 2 & 6 & 10 & 9 \\ 2 & 0 & 5 & 9 & 8 \\ 6 & 5 & 0 & 4 & 5 \\ 10 & 9 & 4 & 0 & 3 \\ 9 & 8 & 5 & 3 & 0 \end{pmatrix} \end{array}$$

Esimese grupi moodustavad kõige väiksemate kauguste alusel A ja B. Väikseim kaugus selle grupi ning kolme ülejäänud uuritava ühiku C, D ja E on järgmine:

$$D_{(AB)C}=\min\{d_{AC}, d_{BC}\}=d_{BC}=5$$
$$D_{(AB)D}=\min\{d_{AB}, d_{BD}\}=d_{BD}=9$$
$$D_{(AB)E}=\min\{d_{AE}, d_{BC}\}=d_{BE}=8$$

Allikas: Aronsson, Åke (1999). En introduktion till basmodulen, Lu
Studentlitteratur.

Arvutamine III

Selle tulemusena saame järgmise uue maatriksi:

$$\mathbf{D_2}= \begin{array}{c} \\ (AB) \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{cccc} (AB) & C & D & E \end{array} \\ \begin{pmatrix} 0 & 5 & 9 & 8 \\ 5 & 0 & 4 & 5 \\ 9 & 4 & 0 & 3 \\ 8 & 5 & 3 & 0 \end{pmatrix} \end{array}$$

Väikseimaks kauguseks maatriksis on DE (3), mis moodustavad seega teise grupi.

$\mathbf{D}_{(AB)C}=5$

$\mathbf{D}_{(AB)(DE)}=\min\{d_{AD}, d_{AE}, d_{BD}, d_{BE}\}=d_{BE}=8$

$\mathbf{D}_{(DE)C}=\min\{d_{CD}, d_{CE}\}=d_{CD}=4$

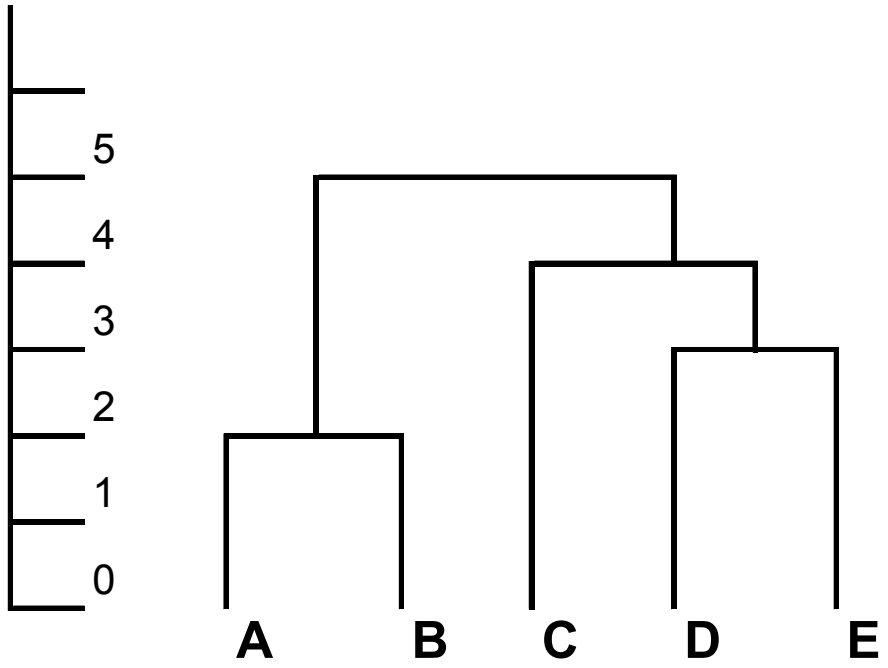
Selle tulemused saame järgmise maatriksi:

$$\mathbf{D_3}= \begin{array}{c} \\ (AB) \\ C \\ (DE) \end{array} \begin{array}{c} \begin{array}{ccc} (AB) & C & (DE) \end{array} \\ \begin{pmatrix} 0 & 5 & 8 \\ 5 & 0 & 5 \\ 8 & 4 & 0 \end{pmatrix} \end{array}$$

Arvutamine IV

Väikseim kaugus on $d_{(DE)C}$. Seetõttu liitub indiviid C grupiga, mis sisaldab ühikuid D ja E.

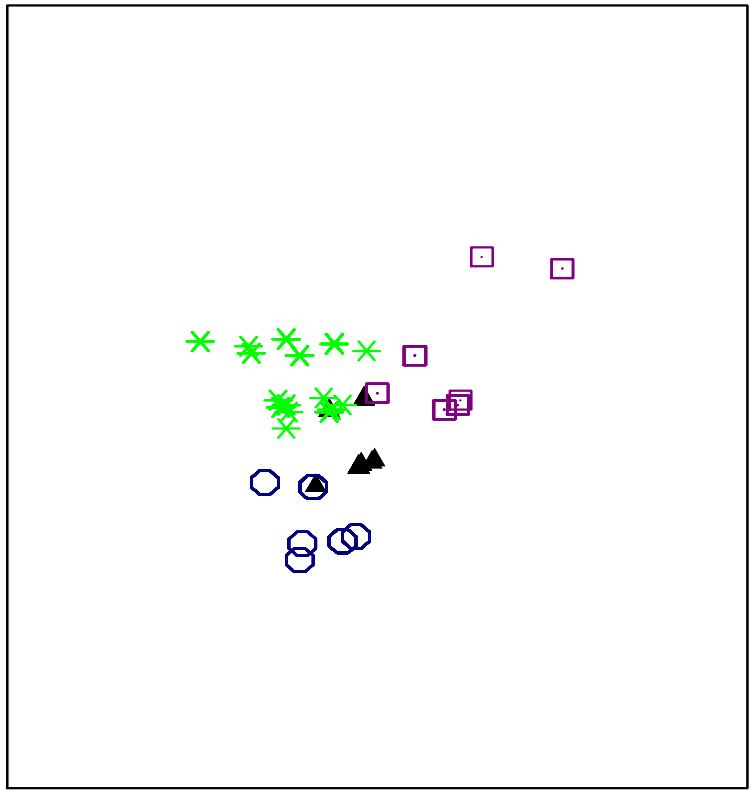
Lõpuks ühinevad kaks gruppi, st viimasel sammul sisaldab grupp kõiki analüüsitavaid ühikuid. Erinevate tunnuste ühinemist kirjeldab järgmine dentrogramm:



Vertikaalselt on esitatud erinevate gruppide/klastrite omavahelised kaugused, horisontaalselt on esitatud analüüsitavad ühikud

K-keskmiste meetod

K-keskmiste meetodi tulemus: Identiteedi grupid Eesti venekeelse elanikkonna hulgas järgmiste identifitseerivate kategooriate alusel*: Eesti kodanik, Eesti ühiskonna liige, Eesti vene kogukonna liige, endise NL kodanik, Venemaa kodanik. Baltikumi elanik. vastamata.*



Allikas: Vihalemm, T., Masso, A. (2003). Identity Dynamics of Russian-Speakers of Estonia in the Transition Period, Journal of Baltic St 92-116.

Arvutamine

- Tunnuste vaheliste distantside arvutamiseks kasutatakse enamasti Eukleidilisi distantse.
  - Esimesel sammul arvutatakse esialgsed klastrite keskpunktid. Igal järgneval sammul [ingl.k *iteration*] grupeeritakse vaatlused tuginedes Eukleidilistele kaugustele ning lähima naabri meetodile, st ühte klastrisse paigutatakse analüüsiühikud selliselt, et nendevahelised kaugused oleksid võimalikult väikesed.
  - Protsessi jätkatakse seni, kuni klastrite keskmised ei muutu rohkem kui on konkreetne piirväärtus või kui on saavutatud iteratsioonide piir.
  - Uurija peab eelnevalt määrama soovitava klastrite arvu.

Näide I

1. Elulaadi tüpoloogia leidmiseks viidi läbi faktoranalüüs peatelgede meetodil telgede *varimax* pööramisega. Leiti 8 tegevuse orientatsiooni.
   - *Mitmekülgselt aktiivne, tööle orienteeritud kultuurilembene, hasartne, tehniliselt harrastuslik, uue meedia ja seltsielu keskne, kodukeskne traditsiooniline, passiivne.*
2. Klasteranalüüsi K-keskmiste meetodil viidi läbi rühmitamine, leidmaks indiviidide gruppe.
   - Kõige paremini tõlgendatavaks osutus juba faktoranalüüsis välja tulnud kaheksa klastriga lahend.
   - Faktortunnuste alusel tehtav klasteranalüüs suurendab rühmitamise usaldusväärsust ning lihtsustab tulemuste tõlgendamist (erinevalt algtunnuste kasutamisele klasteranalüüsis).

Allikas: Nigul, A. 2004

Näide II

Tabel. Internetikasutajate tüüpide sotsiaaldemograafiline koosseis (% tüübist, p<.01)

	Mitmekülg-ne interaktiivne võrgusolija	Mitmekülg-ne infokasutaja	Suhtleja	Eraeluliste teenuste kasutaja	Virtuaalses avalik-kuses osaleja	Vähe-kasutaja
Mehed	64	43	46	36	62	48
Naised	37	57	54	64	38	52
15-19 a	19	6	49	2	28	9
20-29 a	48	25	26	38	25	12
30-44 a	25	42	16	38	28	38
45-54 a	56	20	7	16	12	25
55-64 a	3	7	2	5	7	9
65-74 a	0	0	0	1	0	8
Eestlased	72	77	61	76	76	70
Venelased	28	23	39	24	24	30
Koguvalim	14	16	18	14	18	20

Allikas: Pruulmann-Vengerfeldt, P. 2004

FA ja KA kasutamine

	FAKTORANALÜÜS	KLASTERANALÜÜS
Eesmärk	* Suure arvu tunnuste "kokkusurumine" väiksemaks arvuks. * Varjatud dimensioonide leidmine (tunnuste struktuuri analüüsimine).	1. Suure arvu tunnuste struktuuri analüüsimine. 2. Indiviidide gruppide leidmine määratud tunnuste alusel.
Tunnuste skaalad	* Intervallskaala (arvuline); * Järjestusskaala; * "Järjestatavad" dihhotoomiad (nt *1-meeldib, 0-ei meeldi*).	* Intervallskaala (arvuline) või järjestustunnus K-keskmiste meetodi korral; * Arvulised või dihhotoomsed (järjestatavad) hierarhilise klasteranalüüsi korral; * Eri skaalal järjestustunnused tuleb eelnevalt standardiseerida.

Kriitika

- Faktoranalüüsi kriitika
  - Eeltingimuseks korrelatsioonanalüüsi kasutamine, mis eeldab intervallskaalat, normaaljaotust jms.
  - "Esimese faktori probleem" – esimene faktor kaldub olema suurima kirjeldusjõuga (sisaldab suurimat tunnuste arvu).
  - Erinevad FA meetodid võivad anda erinevaid tulemusi. Kui ka leitakse selge faktorstruktuur, võib tihti olla raske neile tähendusliku nime andmine.
- Klasteranalüüsi kriitika
  - Probleemiks skaala piirang, st tunnused peavad olema mõõdetud samades skaalaühikutes (erinevatel skaaladel mõõdetud tunnused tuleb eelnevalt standardiseerida).
  - Meetodi valimine võib mõjutada struktuuri andmetes, saadud klastrid võivad olla petlikud.

    NB: lahendusena tuleks kasutada paralleelselt mitmeid erinevaid klaster- ja faktoranalüüsi meetodeid.

Ülesanne

- Lugeda teksti (Nigul 2004) ja vastata järgmistele küsimustele:
  - *Milliseid andmete rühmitamise meetodeid on kasutatud?*
  - *Milliseid kriteeriume on kasutatud rühmade arvu määramisel?*
  - Kumb lähenemine – Bourdieu või Reimer'i – sobib paremini antud andmete tõlgendamisel?

Iseseisvaks lugemiseks

- Field, A. (2000). Factor analysis. Rmt: *Discovering Statistics Using SPSS for Windows: Advanced Techniques for Beginners,* London: Sage, lk 423-470

- Everitt, B. S., Dunn, G. (1997). Chapter 6: Cluster analysis. Rmt: *Applied multivariate data analysis*, London etc: Arnold, lk 99-126.

- Nigul, A. (2004). Elulaad Eestis. V.Kalmus, M.Lauristin, P.Pruulmann-Vengerfeldt (toim) Rmt: *Eesti elavik 21.sajandi algul. Ülevaade uurimuse Mina. Maailm. Meedia tulemustest*, Tartu: Tartu Ülikooli Kirjastus, lk 83-95.

PRAKTIKUMI ÜLESANNE NR 1
Kvantitatiivne andmeanalüüs (SPSS'i abil)

Avada andmestik I PRAKT_Meema2008.sav (Allikas: uuring *mina.Maailm.Meedia*) ja vastata esitatud küsimustele kirjalikult.

Küsimus 1. Esimeseks analüüsi etapiks on ühemõõtmelise kirjeldava analüüsi tegemine. Arvutada sagedusjaotused ja keskmised tunnuste lõikes, mis kirjeldavad väärtushinnanguid (k53-k77). *Analyse⇒Descriptive Statistics⇒Frequencies/Descriptives.*

Küsimus	Vastus
1.1. Mida väärtustavad Eesti elanikud kõige enam / kõige vähem?	
1.2. Kumb viis andmete analüüsimiseks on antud juhul sobivam – sagedusjaotused või keskmised? Põhjendada!	
1.3. Kuidas võiks väärtuse tunnuseid ümber kodeerida? Põhjendada!	

Küsimus 2. Üheks mitmemõõtmelise analüüsi võimaluseks on indeksite arvutamine. Arvutada indeks algtunnuste k53, k54, k60, k62, k66 abil. Selleks defineerida esmalt vanade tunnuste väärtused ümber järgnevalt: 5(väga tähtis)= 2, 4(ehk tähtis)=1 ning kõik ülejäänud väärtused=0. Liita saadud uued tunnused kokku ehk arvutada summaindeks. *Transform⇒Recode into Different Variables; Transform⇒Compute Variable*

Küsimus	Vastus
2.1. Mis võiks olla saadud indekstunnuse nimi ehk algtunnuste ühismõõdustaja?	
2.2. Kuidas on Eesti elanikud jaotunud arvutatud väärtusindeksi lõikes (arvutada sagedused)?	
2.3. Kuidas võiks saadud indekstunnust ümber kodeerida (kirjeldada, kodeerida indekstunnus, milline on sagedusjaotus)?	

Küsimus 3. Enamkasutatavaks mitmemõõtmelise analüüsi vormiks on risttabelid. Analüüsida Arvutatud indekstunnuse variatiivsust vanuse, soo ja ankeedi keele lõikes.
Analyse⇒Descriptive Statistics⇒Crosstabs.

Küsimus	Vastus
3.1. Millistes valimi gruppides on indekstunnuse väärtused kõrgemad / madalamad?	
3.2. Kuidas selgitada sellist väärtushinnangute variatiivsust?	

PRAKTIKUMI ÜLESANNE NR 1I
Kvantitatiivne andmeanalüüs (SPSS'i abil)

Avada andmestik II_PRAKT_Meema2008.sav ja vastata esitatud küsimustele kirjalikult (*Andmeallikas: uuring Mina.Maailm.Meedia 2008*).

Küsimus 1. Esimeseks etapiks seoste analüüsimisel on risttabelite tegemine. Uurimisküsimus: millistes vanusgruppides on ENSV-nostalgia suurim?
Esmalt defineerida andmestikus kõikidele küsimustele vastamata jätnud indiviidid puuduvateks väärtusteks. *Variable view⇒Missing values.*
Edasi teha kahemõõtmeline risttabel tunnustega K14-K17 (nt veerutunnus) ning *vanusuus* (nt reatunnus). Arvutada protsendid ja seosekordajad Hii-ruut, Cramer'i V. *Analyse⇒Descriptive Statistics⇒Crosstabs.*

Küsimus	Vastus
1.1. Millistes vanusegruppides on ENSV nostalgia suurim / väikseim (vt protsent)?	
1.2. Kas seos tunnuste vahel on statistiliselt oluline nivool p≤0.01 (vt seosekordajate tabel)?	
1.3. Millise tunnuse korral on seos vanusega kõige tugevam / kõige nõrgem (vt Cramer'i V)?	
1.4. Mida tulemustest võib sisuliselt järeldada? Kuidas tulemusi tõlgendada?	

Küsimus 2. Teha kolmemõõtmeline risttabel, lisades olemasolevale veerutunnusele (valida üks tunnus tunnustest k14-k17) ning reatunnusele (vanusuus) ka kolmanda mõõtme ehk tunnuse ankeedi keel (k672). Arvutada protsendid ja seosekordaja Hii-ruut. *Analyse⇒Descriptive Statistics⇒Crosstabs.*

Küsimus	Vastus
2.1. Mis juhtub seosega kolmanda tunnuse lisamisel (seos jääb alles või kaob)?	
2.2. Kas seos vanuse ja ENSV nostalgia vahel on tõeline, näiline või tinglik?	
2.3. Mida järeldada analüüsist sisuliselt? Ehk kas kolmas tunnus ankeedi keel	

aitab seda seost selgitada?	

Küsimus 3. Sageli kasutatavaks mitmemõõtmelise analüüsi vormiks on keskväärtuste võrdlemine kahes valimi grupis. Uurimisküsimus: kuivõrd usaldus institutsioonidesse erineb ankeedi keele lõikes?
Küsimusele vastamiseks teha sõltumatute valimite T-test tunnuste k79-k98 ning tunnuse sugu lõikes (k651). *Analyse⇒Compare Means⇒Independent-Samples T-test*

Küsimus	Vastus
3.1. Milliseid institutsioone usaldavad enam mehed / milliseid naised? (vt keskväärtused)?	
3.2. Millised keskväärtused on statistiliselt oluliselt erinevad nivool p≤0.01 (vt olulisustõenäosus)?	
3.3. Mida tulemuste põhjal võib sisuliselt järeldada?	

Küsimus 4. Seoste analüüsimiseks intervall- ja ordinaalskaalal tunnuste korral kasutatakse korrelatsioonikordajat. Uurimisküsimus: kuidas on omavahel seotud üksikute riiklike institutsioonide usaldus? Teha korrelatsioonanalüüs tunnuste k79-k98 vahel. *Analyse⇒Correlate⇒Bivariate*

Küsimus	Vastus
4.1. Millist korrelatsioonikordajat tuleks siin kasutada (Pearson, Spearman)? Põhjendada!	
4.2. Milliste tunnuste vahel on seos tugevaim / nõrgim?	
4.3. Kas korrelatsioonid on statistiliselt olulised nivool p≤0.01?	
4.4. Mida tulemuste põhjal sisuliselt järeldada? Kuidas tulemusi tõlgendada?	

PRAKTIKUMI ÜLESANNE NR 1II
Kvantitatiivne andmeanalüüs (SPSS'i abil)

Avada andmestikud III_PRAKT_ylesanne.sav ja vastata esitatud küsimustele kirjalikult (A*ndmeallikas: uuring Mina.Maailm.Meedia 2008*).

Ülesanne 1.1: Vastata järgmisele hüpoteesile (andmestik III PRAKT_ylesanne1.sav): Lääne meediakanalite jälgimist (*index084*) ennustab võõrkeelte kasutamise ulatus (*index061*).
Analüüsi esimeseks sammuks on seoste analüüsimine korrelatsioonanalüüsi abil. Teha korrelatsioonanalüüs tunnuste index084 ja index061 vahel (analüüsi võib lisada ka teisi sisuliselt põhjendatud unnuseid, nt *index012, index016, index061, index078, index079, index080, index081, index082, index106, vanus*). Analüüsiks: *Analyse⇒Correlatite⇒Bivariate*

Küsimus	Vastus
1.1. Millist korrelatsioonikordajat kasutada ja miks (Pearson, Spearman)?	
1.2. Mida näitab korrelatsioonanalüüs sisuliselt?	
1.3. Milliseid tunnuseid võiks regressioonanalüüsis kasutada sõltumatute muutujatena?	

Ülesanne 1.2: Teha tunnuste regressioonanalüüs, kus sõltuvaks tunnuseks on Lääne meediakanalite jälgimine (*index084*) ja sõltumatuteks tunnusteks on *index012, index016, index061, index078, index079, index080, index081, index082, index106, vanus*; sõltuvate tunnuste valik siin vaba, oluline sisuline põhjendatus; enne küsimustele vastamist võib teha regressioonanalüüsi erinevate sõltumatute tunnustega). *Analyse⇒Regression⇒Linear.*

Küsimus	Vastus
1.1. Millised tunnused valisite sõltumatuteks tunnusteks? Palun põhjendada!	
1.2. Mida võib öelda mudeli headuse (statistilise sobivuse) kohta? Milliste statistiliste suuruste alusel seda otsustada?	
1.3. Mida võib järeldada regressioonikordajate alusel? Kas algselt püstitatud hüpotees leidis kinnitust? Palun põhjendada!	

Ülesanne 2.1. Vastata järgmisele hüpoteesile: indiviidide majanduslik kapital on selgitatav võõrkeelte oskuse kaudu. Esmalt uurida risttabeli ning Hii-ruut testi abil tunnuse indk110 seoseid taustatunnustega *indk061*, vanus2, sugu, haridus (analüüsi võib lisada ka teisi sisuliselt põhjendatud tunnuseid). *Analyse⇒Descriptive⇒Crosstabs*.

Küsimus	Vastus
1.1. Milliste tunnuste vahel ilmnesid statistiliselt olulised seosed?	
1.2. Mida seoseanalüüsi põhjal võib sisuliselt järeldada?	
1.3. Milliseid tunnuseid võiks regressioonanalüüsis kasutada sõltumatute muutujatena?	

Ülesanne 2.2. Hüpoteesile vastamiseks teha multinoomne logistiline regressioonanalüüs, kus sõltuvaks tunnuseks on indk110 ja sõltumatuteks tunnusteks: *indk061*, vanus2, sugu, haridus . *Analyse⇒Regression⇒Multinomial Logistic*.

Küsimus	Vastus
1.2. Mida järeldada mudeli statistilise headuse / sobivuse kohta kohta? Milliste statistiliste suuruste alusel seda otsustada?	
1.2. Mida võib öelda mudelisse valitud tunnuste kirjeldusjõu kohta?	
1.3. Mida võib järeldada regressioonikordajate alusel? Kas algselt püstitatud hüpotees leidis kinnitust? Palun põhjendada!	

PRAKTIKUMI ÜLESANNE NR IV
Kvantitatiivne andmeanalüüs (SPSS'i abil)

Avada andmestik IV PRAKT_Meema2008.sav ja vastata esitatud küsimustele kirjalikult (kättesaadav moodle's aine kodulehel, Allikas: Mina.Maailm.Meedia 2008).

Küsimus 1. Uurida elulaadi tunnuste struktuuri ja leida sobiv elulaadi tüpoloogia. Selleks teha esmalt elulaadi tunnuste korrelatsioonanalüüs (v207-240).
Analyse⇒Analyse⇒Correlate⇒Bivariate.

Küsimus	Vastus
1.1. Millist korrelatsioonikordajat kasutada ja miks (Pearson või Spearman)?	
1.2. Mida näitab korrelatsioonikordaja siin sisuliselt (nt milliste tegevuste vahel on korrelatsioonid tugevad / nõrgad?)	
1.3. Kas korrelatsioonide alusel on alust oletada tunnuste grupeerumist)?	

Küsimus 2: Järgmiseks analüüsi etapiks on teha elulaadi tunnuste faktoranalüüs. Proovida teha analüüsi erinevate faktorite arvuga. *Analyse⇒Data Reduction⇒Factor, Principal axis(Method), Varimax(Rotation).*

Küsimus	Vastus
2.1. Millise faktorite arvuga struktuur on statistiliselt sobivaim (kui suure osa tunnuste koguvarieeruvusest need faktorid kirjeldavad)?	
2.2. Mitme faktoriga lahendus on sisuliselt sobivaim (st pakub parimat tõlgendusvõimalust)?	
2.3. Kas esineb mitmedimensionaalseid tunnuseid (faktorlaadung >0,3 mitmes faktoris)?	
2.4. Kuidas võiks faktoreid nimetada (faktorisse kuuluvate üksiktunnuste alusel)?	

Küsimus 3. Lisaks tunnuste struktuuri analüüsile on võimalik faktortunnuseid analüüsida teiste taustatunnuste lõikes. Selleks tuleks faktortunnused salvestada uute tunnustena: *Analyse⇒Data Reduction⇒Factor; Scores⇒Save as variables.* Kuna salvestatud faktortunnused on arvulised, tuleks analüüsida faktortunnuste keskväärtuste erinevust valimi alagruppides, nt T-testi abil: *Analyse⇒Compare Means⇒Independent T-test* (grupeerivaks tunnuseks valida *k672*).

Küsimus	Vastus
3.1. Milliste faktortunnuste korral on erinevused ankeedi keele lõikes statistiliselt oluliselt erinevad (vt T-testi olulisustõenäosus, p≤.01)?	
3.2. Millised elulaadi orientatsioonid on iseloomulikud eri rahvusrühmade hulgas (vt keskväärtused)?	

Küsimus 4: Faktortunnuste alusel klasteranalüüsi tegemine võimaldab leida individuaalsed elulaadi grupid ning neid taustatunnuste lõikes analüüsida. Klasteranalüüsi tegemiseks faktortunnuste baasil: *Analyse⇒Classify⇒K-means*. Proovida teha analüüsi erinevate klastrite arvu korral. Valida sisuliselt ja statistiliselt sobivaim klasterlahend.

Küsimus	Vastus
4.1. Milline on üksikute tunnuste kirjeldusjõud klasterkuuluvuse määramisel (vt *ANOVA* tabel)?	
4.2. Mis on saadud klastrite sisu, st milliste klastrite ehk indiviidide gruppidega on tegemist (vt *final cluster centers*)?	
4.3. Kui suured on saadud klastrid arvuliselt? Kas võiks mõne arvuliselt väikese klastri analüüsist välja jätta või mõned väikesed, sisuliselt kokkusobivad klastrid kokku liita?	