TARTU ÜLIKOOL
EUROOPA KOLLEDŽ

B e S t
e-õppe programm

Euroopa Liit
Euroopa Sotsiaalfond

Eesti tuleviku heaks

TARTU ÜLIKOOL

STATISTILISE ANALÜÜSI TEOSTAMINE EXCELI JA SPSSI ABIL

Kerly Krillo

Tartu Ülikool, sotsiaalteaduslike rakendusuuringute keskus

Tööturu ja tööpoliitika programmi juht

kerly.krillo@ut.ee

26.02.2010

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

I ALUSTUSEKS

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Tunnikontroll ☺

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Tunnikontroll ☺

- Milline on Sinu lõputöö teema?
- Millised on Sinu ootused ainekursusele (milliseid meetodeid loodad õppida)?
- Kas oled varem SPSSiga kokku puutunud?
- Kas kasutad MS Exceli versiooni 2007?
- Mida tähendab korrelatsioon?
- Mida mõõdab standardhälve?
- Millal eelistada moodi aritmeetilisele keskmisele?

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Ainekursuse eesmärk

Kui tudeng, kes on kursusele registreerunud, oskab mai lõpuks iseseisvalt kasutada Excelit ja SPSSi lihtsama **kvantitatiivse** statistilise andmetöötluse tegemiseks…

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Ainekursuse eesmärk

Kui tudeng, kes on kursusele registreerunud, oskab mai lõpuks iseseisvalt kasutada Excelit ja SPSSi lihtsama **kvantitatiivse** statistilise andmetöötluse tegemiseks…

… siis olen mina oma ülesande täitnud ☺

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Ainekursuse eesmärk

Kui tudeng, kes on kursusele registreerunud, oskab mai lõpuks iseseisvalt kasutada Excelit ja SPSSi lihtsama statistilise andmetöötluse tegemiseks…

… siis olen mina oma ülesande täitnud ☺

Eesmärgiks on **ÕPPIMINE**

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Ainekursuse eesmärk

Kui tudeng, kes on kursusele registreerunud, oskab mai lõpuks iseseisvalt kasutada Excelit ja SPSSi lihtsama statistilise andmetöötluse tegemiseks…

… siis olen mina oma ülesande täitnud ☺

Eesmärgiks on **ÕPPIMINE KOOS TÖÖTAMISE** kaudu ning ma loodan, et see saab olema meeldiv mõlemale poolele

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Ainekursuse eesmärk

Teisisõnu, ainekursuse läbinu

- teab lihtsamate statistiliste näitajate (aritmeetiline keskmine, mediaan, mood, standardhälve, korrelatsioon jne) sisu
- on võimeline valima olenevalt uurimisprobleemist analüüsi teostamiseks sobivad statistilised karakteristikud
- oskab iseseisvalt statistilist analüüsi teostada ning tulemusi tõlgendada
- teab olulisemaid statistiliste andmete andmebaase (Eesti Statistikaameti, Eurostati, Maailmapanga, Rahvusvahelise Valuutafondi jne andmebaasid)

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Ainekursuse ülesehitus

26. veebruar kell 10.15-11.45	**Sissejuhatus ainesse**. Statistika põhimõisted ja -kontseptsioonid
26. veebruar kell 12.00-13.30	**Statistilise analüüsi teostamine Excelis**: - kategooriliste tunnuste teisendamine - statistiliste funktsioonide kasutamine - jooniste tegemine
27. veebruar kell 9.30-11.00	**Statistilise analüüsi teostamine Excelis:** - Töötamine suurte andmemassiividega
27. veebruar kell 11.15-12.45	**Statistilise analüüsi teostamine Excelis:** - Töövahend PivotTable
23. aprill kell 10.15-11.45	**Statistilise analüüsi teostamine Excelis:** - Töövahend Data Analysis
23. aprill kell 12.00-13.30	**Andmebaasid**: -mikro- ja makroandmebaasid - Eesti Statistikaameti andmebaas - Eurostati andmebaas (EL-27 + veel mõnede riikide andmed) - Maailmapanga, IMFi jt andmebaasid

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Ainekursuse ülesehitus

24. aprill kell 9.30-11.00	**Statistilise analüüsi teostamine SPSSis**
24. aprill kell 11.15-12.45	**Statistilise analüüsi teostamine SPSSis**
22. mai kell 9.30-11.00	**Statistilise analüüsi teostamine SPSSis**
22. mai kell 11.15-12.45	**Statistilise analüüsi teostamine SPSSis**

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Ainekursuse hinde kujunemine

- 4 kodutööd a max 10 punkti
- arvestus max 60 punkti

- praktikumitööd
- **Loeng/praktikumist puudumisel tuleb praktikum järele vastata (st lahendada praktikumiülesanne iseseisvalt, tõlgendada tulemusi)**

Virtuaalne kohtumispaik: Moodle

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Mõtteharjutus

- Mida loodad ainekursuse käigus õppida (konkreetsed meetodid)?

Uurimisküsimuse püstitamine

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Analüüsi ettevalmistamine

Autor: Kerly Krillo

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Analüüsi ettevalmistamine

Uurimisküsimuse püstitamine

Andmete olemasolu väljaselgitamine

Autor: Kerly Krillo

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Analüüsi ettevalmistamine

Uurimisküsimuse püstitamine

Andmete olemasolu väljaselgitamine

Andmed on olemas

Uuringu teostamise kavandamine

Autor: Kerly Krillo

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Analüüsi ettevalmistamine

Uurimisküsimuse püstitamine

Andmete olemasolu väljaselgitamine

Vajalikke andmeid ei ole olemas

Andmed on olemas

Uuringu teostamise kavandamine

Autor: Kerly Krillo

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Analüüsi ettevalmistamine

Uurimisküsimuse püstitamine

Andmete olemasolu väljaselgitamine

Vajalikke andmeid ei ole olemas

Kuidas jätkata?

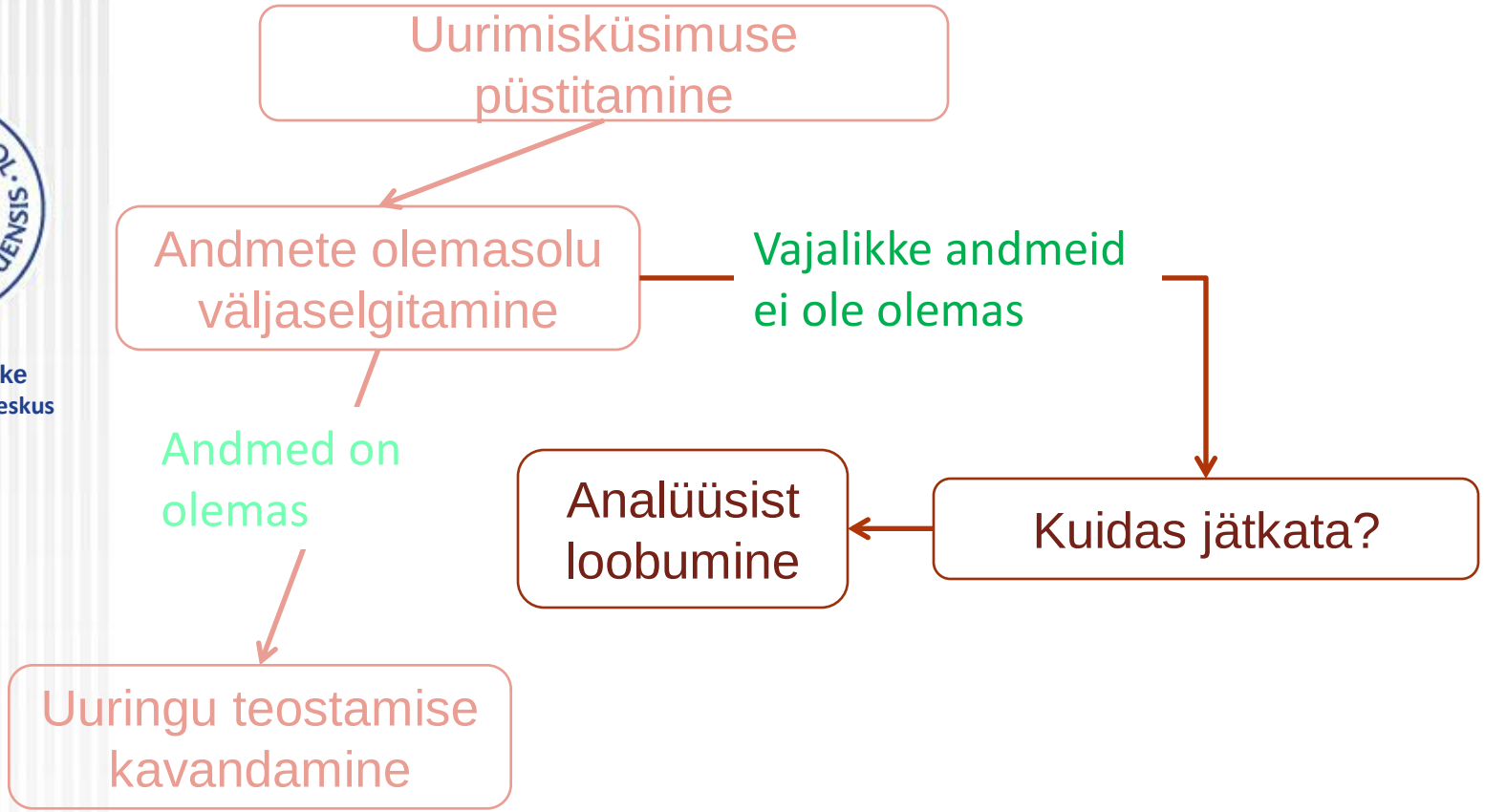
Andmed on olemas

Uuringu teostamise kavandamine

Autor: Kerly Krillo

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

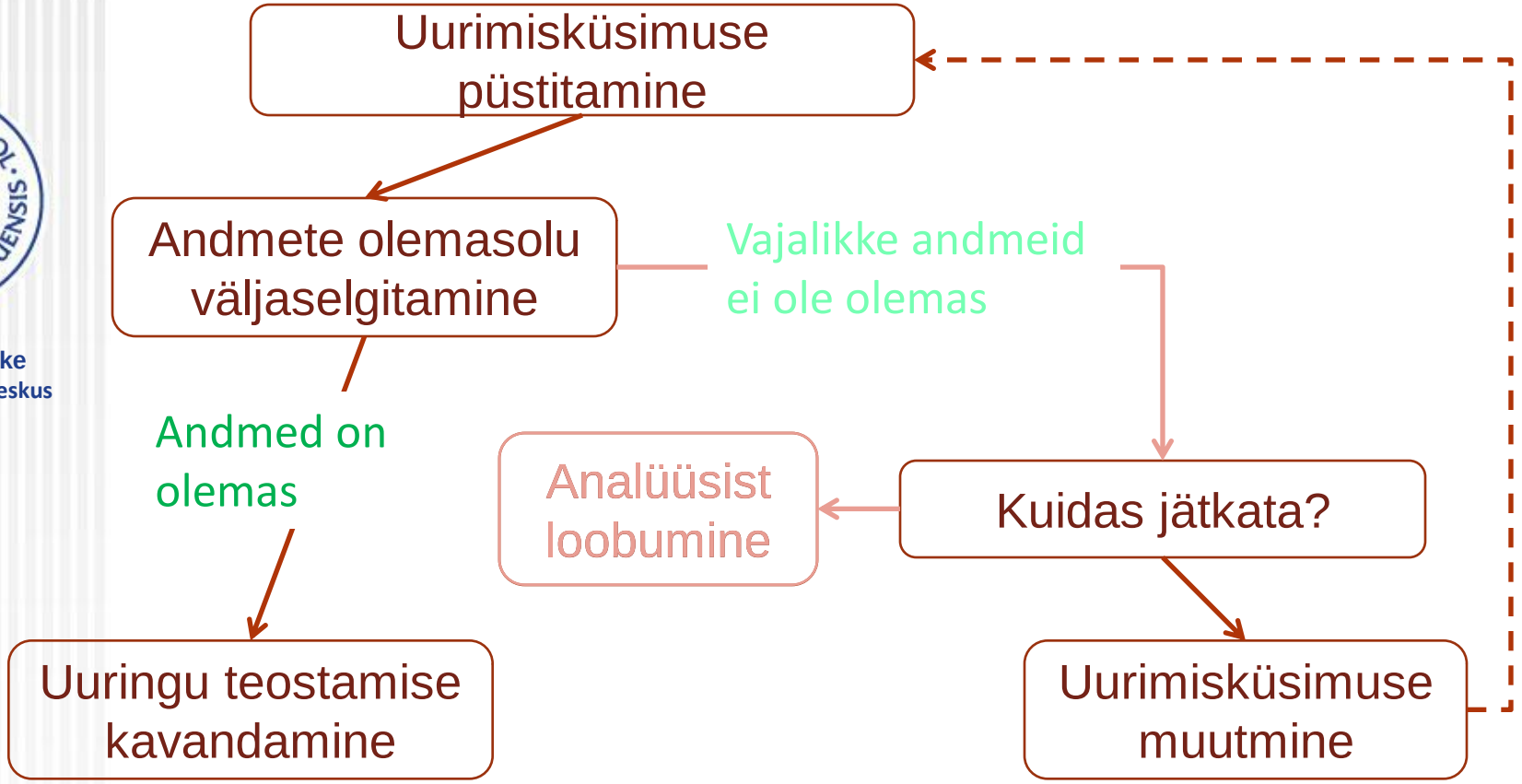
Analüüsi ettevalmistamine



Autor: Kerly Krillo

Sotsiaalteaduslike rakendusuuringute keskus [RAKE]

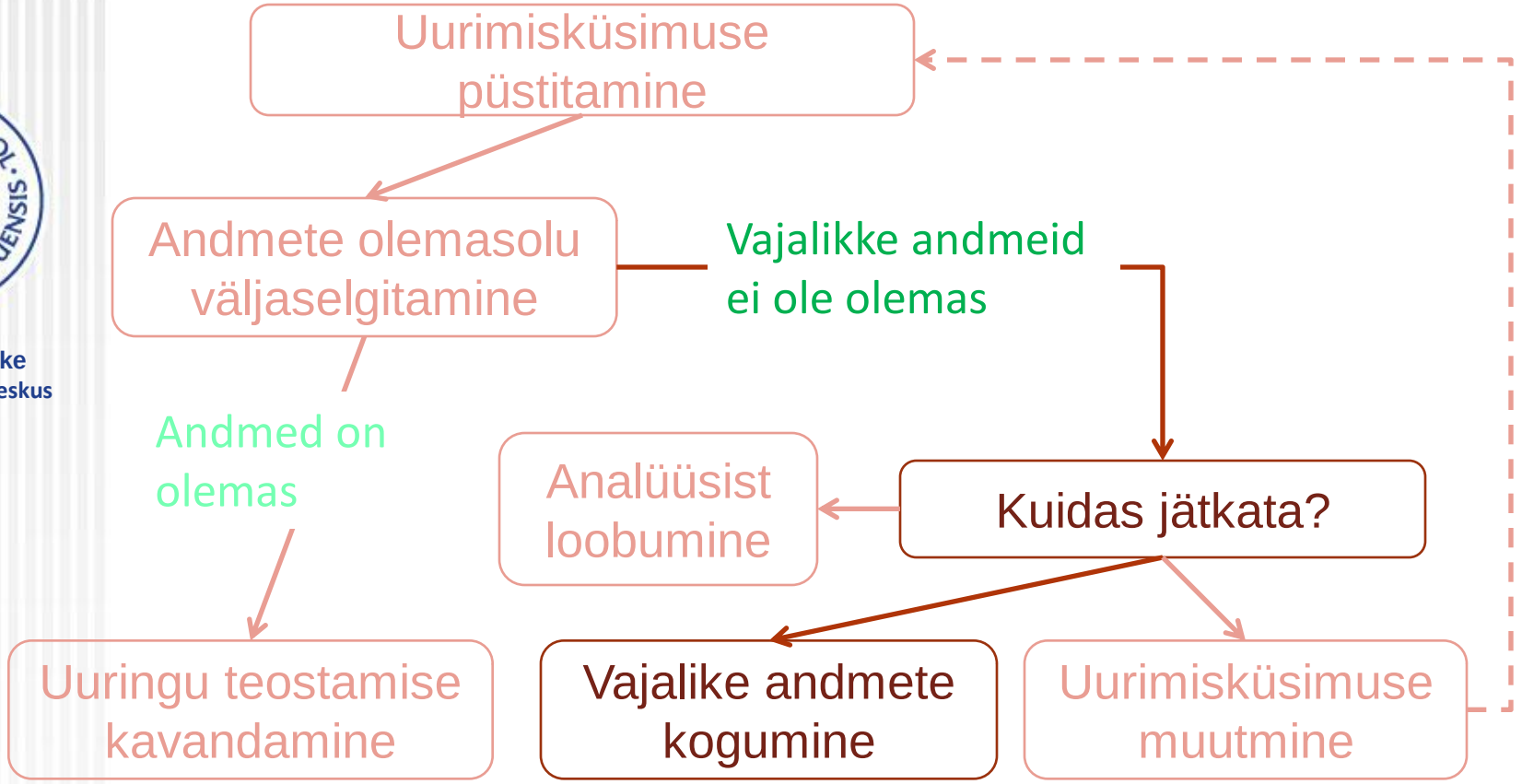
Analüüsi ettevalmistamine



Autor: Kerly Krillo

Sotsiaalteaduslike rakendusuuringute keskus [RAKE]

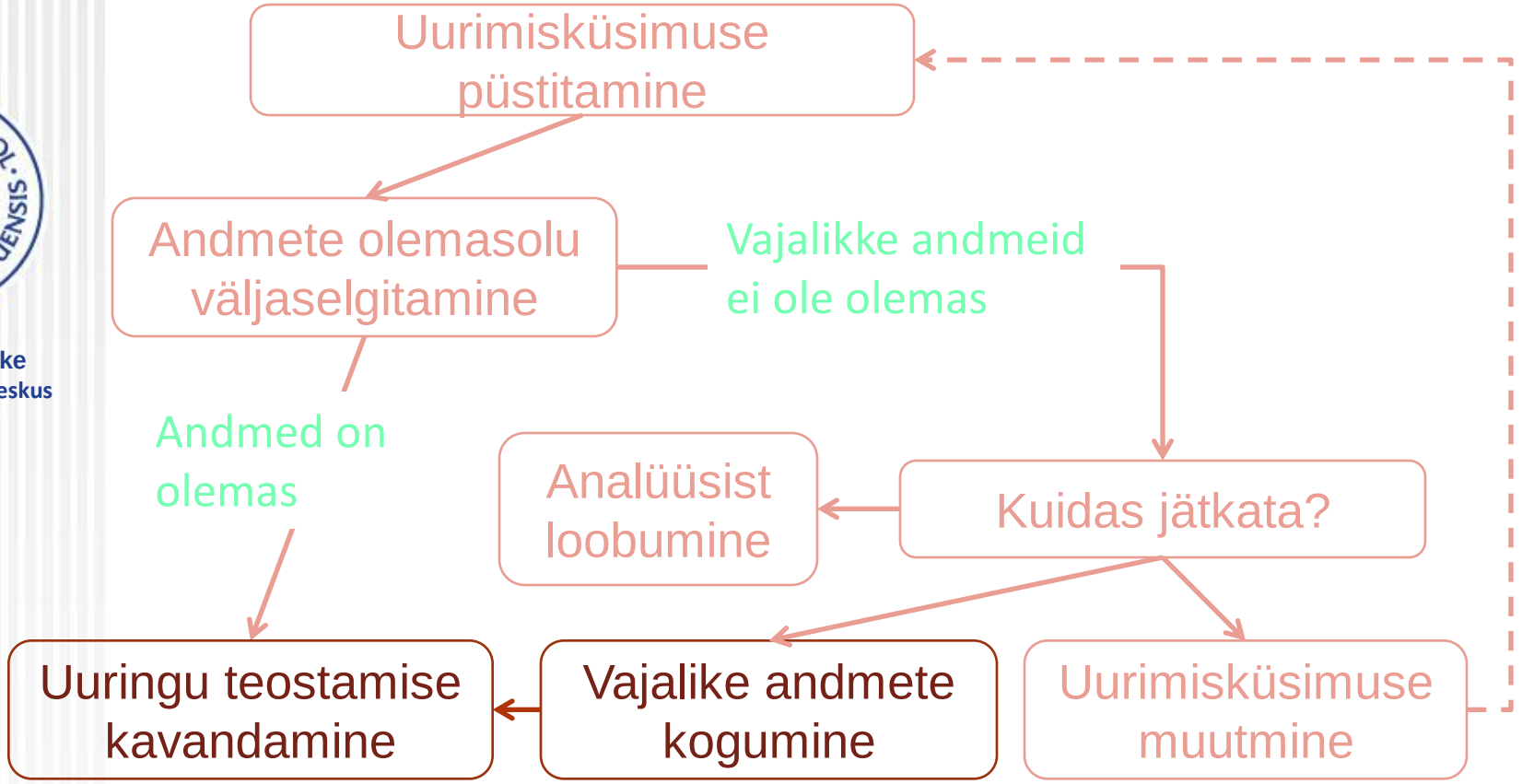
Analüüsi ettevalmistamine



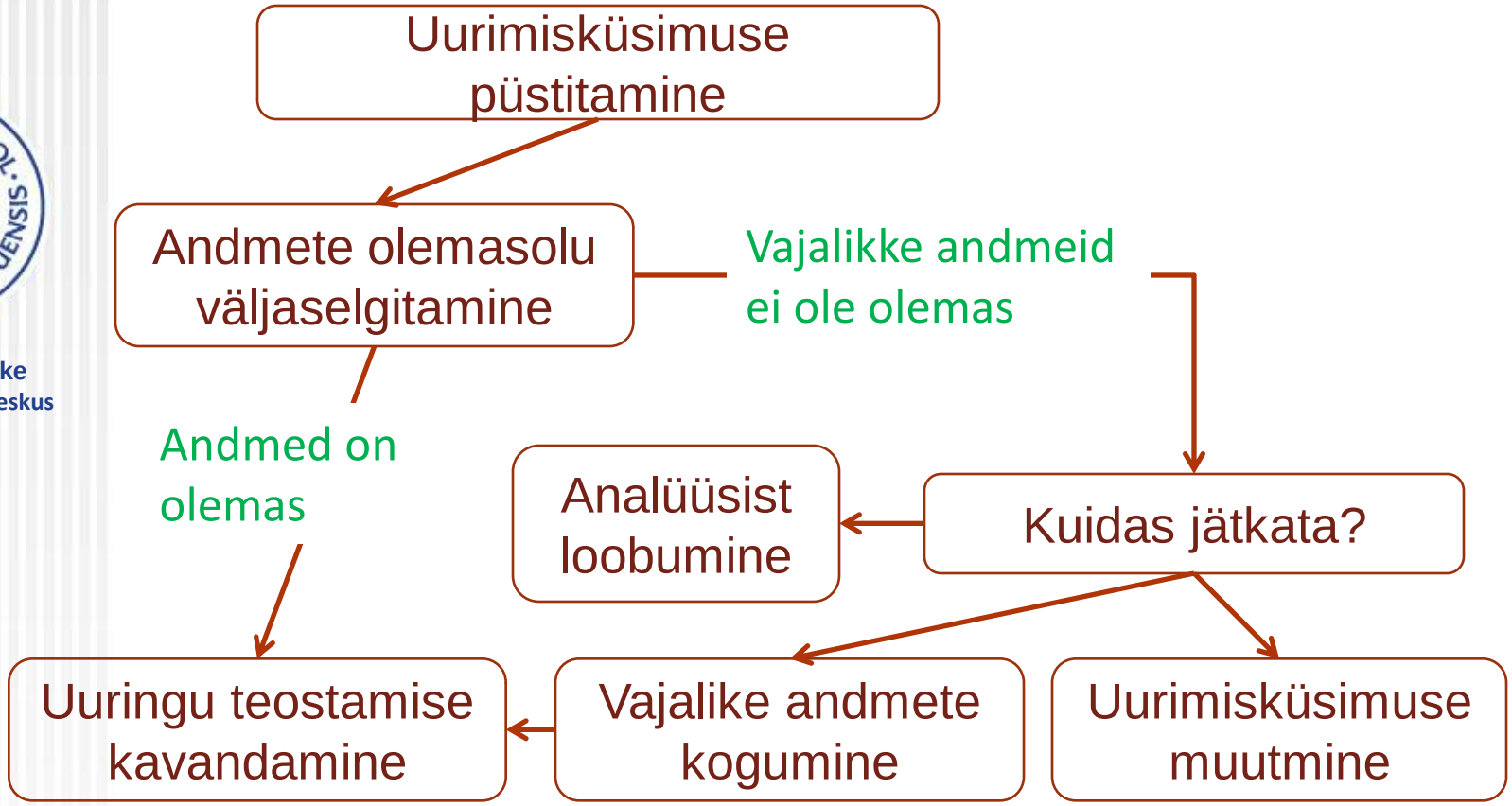
Autor: Kerly Krillo

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Analüüsi ettevalmistamine



Autor: Kerly Krillo



Analüüsi ettevalmistamine

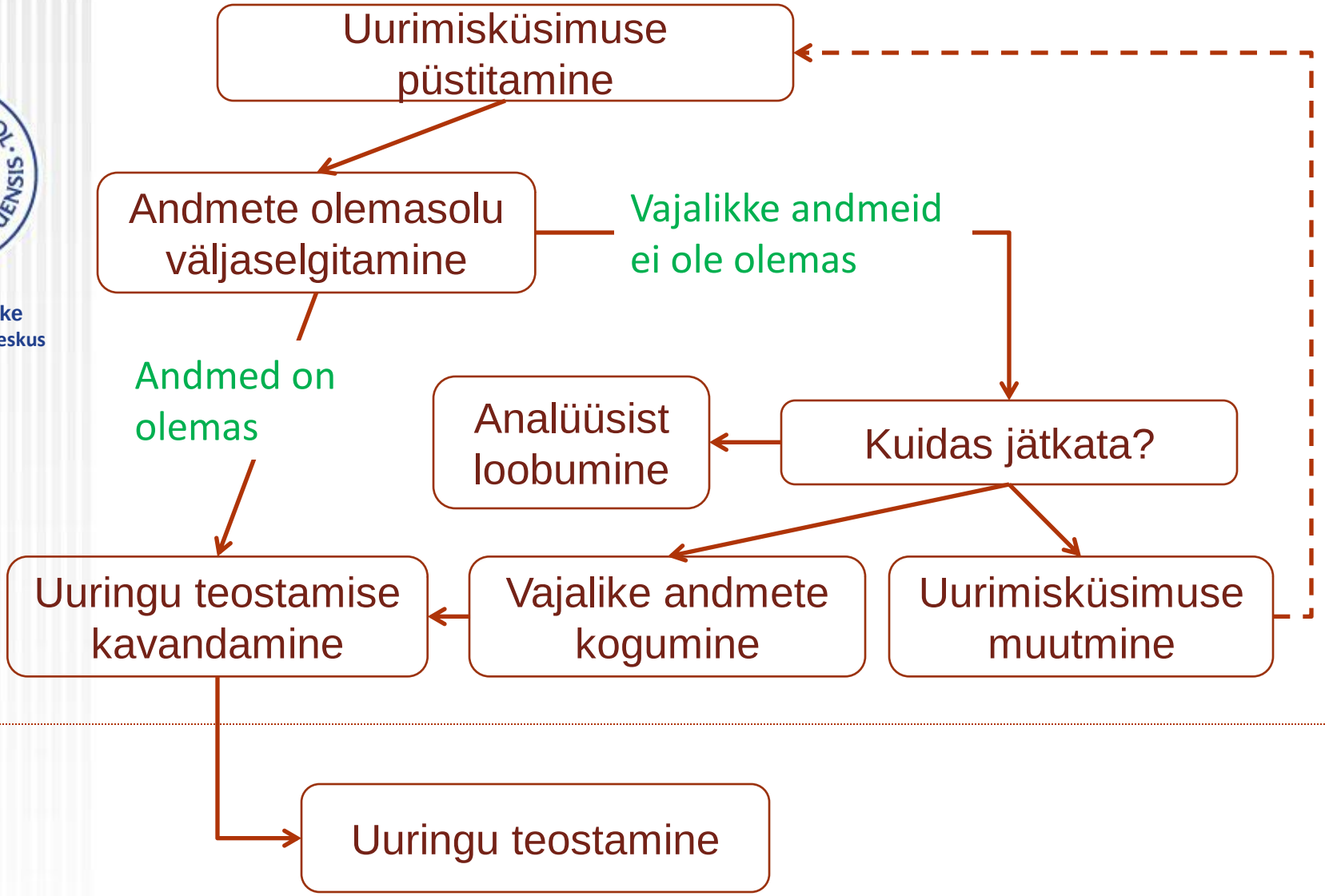
Analüüsi teostamine

Sotsiaalteaduslike rakendusuuringute keskus [RAKE]

Autor: Kerly Krillo

Sotsiaalteaduslike rakendusuuringute keskus [RAKE]

Analüüsi ettevalmistamine



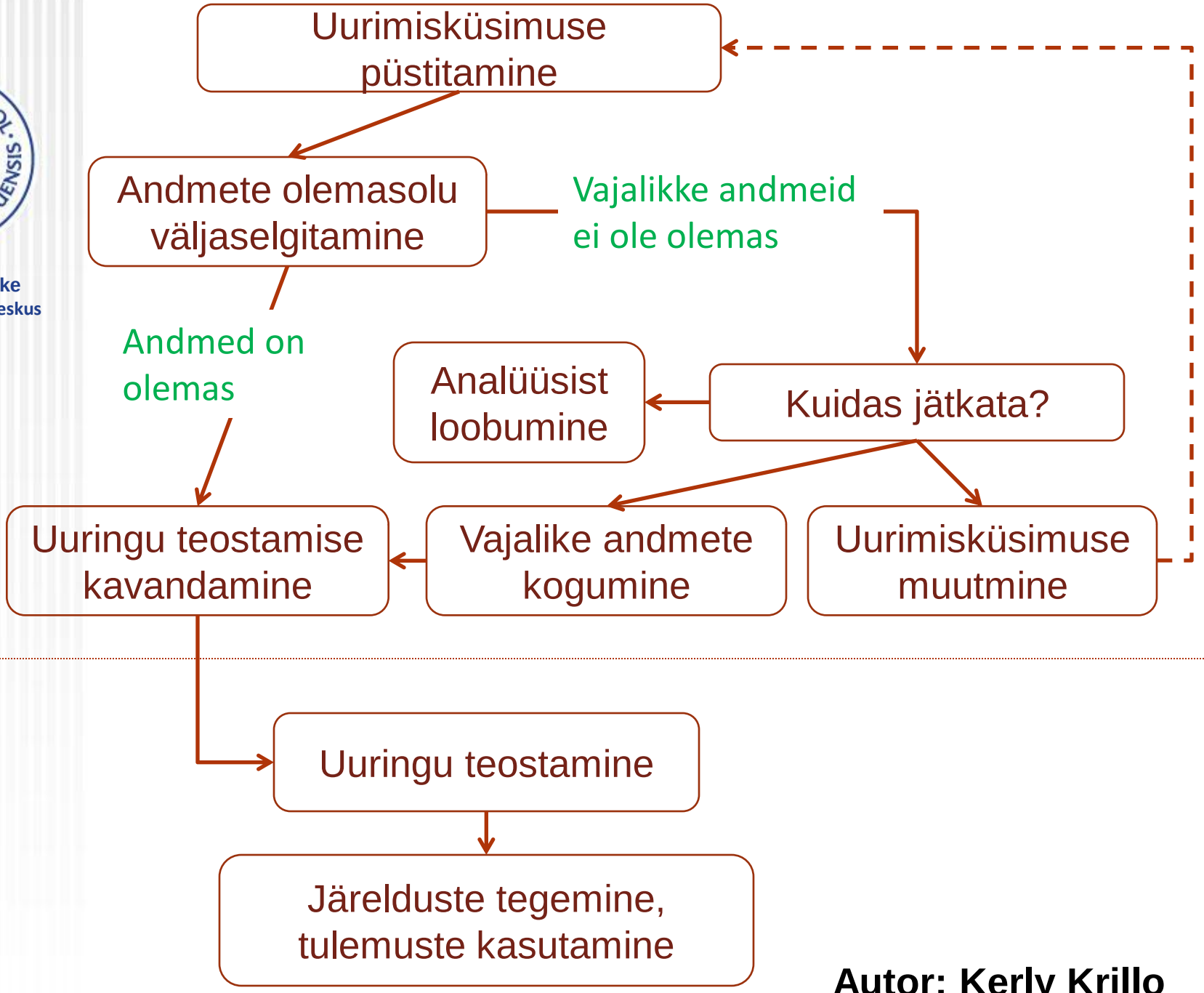
Analüüsi teostamine

Autor: Kerly Krillo

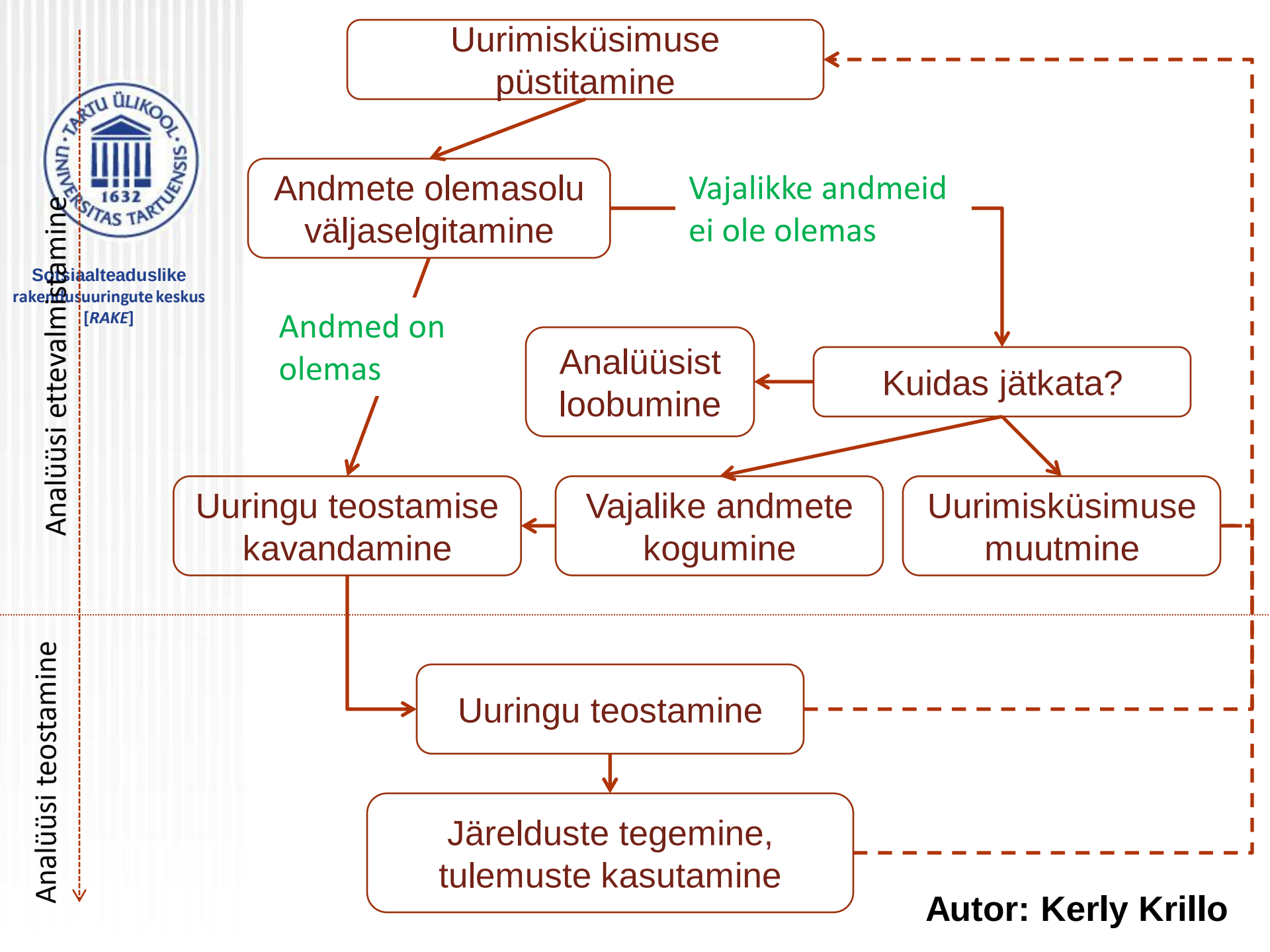
Sotsiaalteaduslike rakendusuuringute keskus [RAKE]

Analüüsi ettevalmistamine

Analüüsi teostamine



Autor: Kerly Krillo


Uurimisküsimuse püstitamine
Andmete olemasolu väljaselgitamine
Vajalikke andmeid ei ole olemas
Andmed on olemas
Analüüsist loobumine
Kuidas jätkata?
Uuringu teostamise kavandamine
Vajalike andmete kogumine
Uurimisküsimuse muutmine
Uuringu teostamine
Järelduste tegemine, tulemuste kasutamine
Analüüsi ettevalmistamine
Analüüsi teostamine
Sotsiaalteaduslike rakendusuuringute keskus [RAKE]
Autor: Kerly Krillo

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Uuringu teostamine

- Andmete korrastamine (vajadusel kodeerimine, rühmitamine jms)
- Andmete kontrollimine

✓ erindid

✓ sisestus-, loogika- jms vead

✓ lüngad – millest tingitud (kas juhuslik või mitte)

- Andmete valiidsus ja reliaablus

→hinnang andmete kvaliteedile → edasine andmete analüüsi tase

NB! Taust määrab andmete töötlemise ja interpreteerimise, tulemused tuleb siduda kogu uurimistsükliga!

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

II ANDMETE ANALÜÜS

26.02.2010

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Andmete korraldamise viise

- Ristlõikeandmestik – staatiline (M X N)
- Kordusmõõtmiste andmestik – üht ja sama tunnust mõõdetakse korduvalt

a) mõõdetavad tunnused samad, aga vastajad erinevad (nt Eesti tööjõu-uuringud)

b) indiviidid ja küsimused suuresti samad – longituud(ne)uuring

- Aegrida –sama tunnuse mõõtmine teatud ajavahemiku järel, tüüpiliselt palju mõõtmiskordi

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

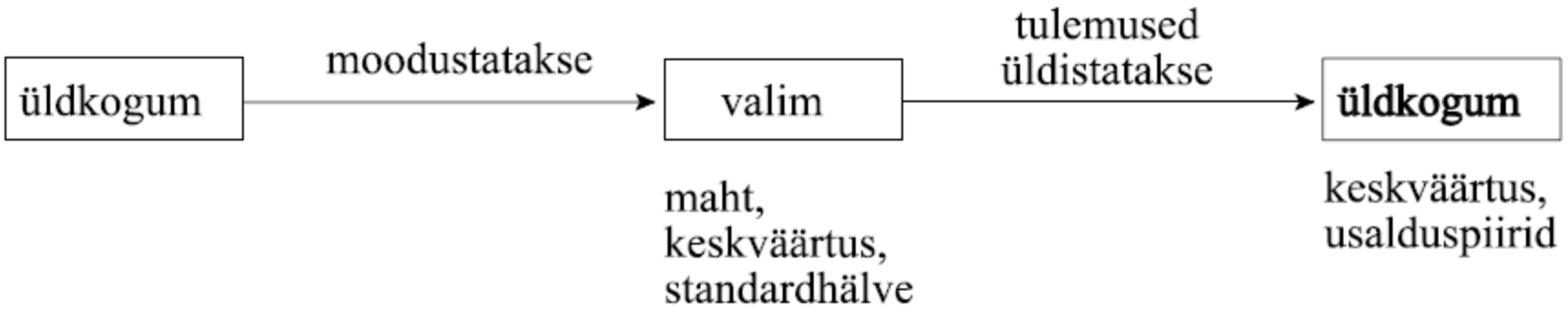
Andmete saamisviisid

- Esmased andmed – kogutakse uurija poolt/tema poolt määratletud eesmärkidel
- Teisesed andmed – on juba kellegi teise poolt kogutud teistel eesmärkidel (nt Statistikaamet, mõni teine era- või riigiettevõte)

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Andmete saamisviisid

- Üldkogum – kõikne vaatlus, uuritakse kõiki üldkogumi elemente (nt rahvaloendus)
- Valim – küsitletakse vaid osa üldkogumist



NB! Valimi korral on aktuaalne **kaalumine**, tagamaks saadud tulemuste üldistatavus

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Andmete mõõteskaalad – Stevensi tüpoloogia

- Nominaalskaala – vähe väärtusi, diskreetsed, pole loogiliselt järjestatavad (nt sugu, rahvus, värvus)

Ei saa teha arvutusi! Saab loendada ja leida sagedusi

- Ordinaal- ehk järjestusskaala – vähe väärtusi, diskreetsed, on loogiliselt järjestatavad (nt eelistused)

Sageli ei ole intervallid skaalajaotuse vahel sisuliselt ühepikkused (nt väga halb, halb, hea, väga hea; hinde "4" saanu ei pruugi olla teadmiste tasemelt täpselt kaks korda parem kui hinde "2" saanu) → võib teha tehteid, mis ei muuda tunnuse väärtuste järjekorda; aga nt aritmeetiline keskmine ei kanna sisukat infot

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Andmete mõõteskaalad – Stevensi tüpoloogia

- Arvuline skaala (intervallskaala) - palju väärtusi (nt vanus, pulsi sagedus, töötajate arv)

a) diskreetne – variandid selgelt eristunud (nt täisarvud)

b) pidev – iga kahe mõõtmistulemuse vahele on võimalik asetada veel kolmas

c) erijuht: binaarne ehk dihhotoomne (kaheväärtuseline)

- vahemikskaala – nullpunkti asukoht on kokkuleppeline (nt Celsiuse skaala, aeg).

Võib leida vahesid, aga mitte suhteid

- Suhteskaala – nullpunkt fikseeritud (nt kaal, pikkus)

NB! Arvulise tunnuse võib teisendada nominaal- või järjestustunnuseks, aga vastupidine teisendus ei ole võimalik!

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Andmete kodeerimine

- **Statistilise analüüsi teostamiseks on vaja andmed kodeerida!**

- Puuduvad väärtused – kood valitakse nii, et see eristuks selgelt teistest tunnuse väärtustest (nt 99, 999, ., “ “ jne)

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Keskmised

- **Mood** - tunnuse väärtuste hulgas kõige sagedamini esinev väärtus.

Moodi omadusi

1) saab kasutada nii nominaalskaala, järjestikskaala kui ka intervallskaala korral. NB! Juhul, kui arvulisel tunnusel on palju väärtusi (ja tavaliselt on), on sageli otstarbekas andmed enne moodi leidmist intervallidesse grupeerida.
2) teatud juhtudel mood puudub (st kõik tunnuse väärtused esinevad sama arv kordi).
3) teatud juhtudel on tunnusel mitu moodi (st on mitu ühesuguse sagedusega väärtust). Sel juhul on tegemist **multimodaalse kogumiga.**

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Keskmised

- **Mediaan –** variatsioonrea keskmine liige, st mediaanist mõlemale poole jääb 50% elementide koguarvust. Teisisõnu, mediaan jaotab **järjestatud** statistilise rea kaheks.

Mediaani omadusi:

1. paarituarvulise elementide arvuga rea korral on mediaan järjestatud rea keskmine liige
2. paarisarvulise elementide arvuga rea korral leitakse mediaan kahe keskmise liikme aritmeetilise keskmisena
3. Võib kasutada järjestus- ja arvtunnuse korral

Eelis võrreldes aritmeetilise keskmisega: **ei ole tundlik ekstremaalsete väärtuste suhtes!**

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Keskmised

- **Aritmeetiline keskmine**

Aritmeetilise keskmise omadusi:

1. korrektne on kasutada arvulise tunnuse korral, enamasti ei ole õige intervallskaalal mõõdetud tunnuse korral, kindlasti pole õige kasutada nominaaltunnuse korral
2. võimaldab võrrelda elementide näitaja väärtusi aritmeetilise keskmisega;
3. võimaldab arvutada teisi statistilisi näitajaid;
4. sõltub igast üksikust elemendist ja on seetõttu **on tundlik ekstremaalsete väärtuste suhtes**.

Edasiarendus: kaalutud keskmine (võtab arvesse andmerea eripärasid)

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Veel kirjeldavaid statistikuid

- Kvantiilid – jaotavad statistilise rea võrdseteks osadeks

Kvantiili nimetus	Mitu kvantiili on	Mitmeks osaks jaotavad	Märkused
mediaan	1	2	Mõlemale poole jääb 50% rea liikmetest
kvartiilid	3	4	Igas neljandikus on 25% rea liikmetest.
detsiilid	9	10	Igas kümnendikus on 10% rea liikmetest.
tsentiilid e. **protsentiilid** e. **pertsentiilid**	99	100	Igas sajandikus on 1% rea liikmetest.

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Absoluutsed variatsiooninäitarvud

- **Variatsiooniamplituud -** rea kõige suurema ja kõige väiksema liikme arvväärtuste vahe
- **Dispersioon ehk keskmine ruuthälve (variance) -** ruuthälvete aritmeetiline keskmine

  NB! Dispersiooni mõõtühikuks on mõõdetava tunnuse dimensiooni ruut, mis raskendab tõlgendamist. Näiteks aastates mõõdetud vanuse dispersiooni ühikuks on aasta ruut.
- **Standardhälve (*standard deviation)* ehk ruutkeskmine hälve -** dispersiooni ruutjuur

  Standardhälve mõõtühikud on samad, mis aritmeetilisel keskmisel ja üksikutel väärtustel. Näiteks kui vanus on aastates, siis ka vanuse standardhälve on aastates.

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Absoluutsed variatsiooninäitarvud

- NB! **Absoluutsete variatsioonnäitarvude abil ei saa võrrelda**
- eri ühikutes mõõdetavate suuruste varieerumist;
- väga erineva nivoo ümber toimuvaid kõikumisi.

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Suhtelised variatsiooninäitarvud

- **Variatsioonikoefitsient -** standardhälbe ja aritmeetilise keskmise jagatis

NB! **Absoluutsed ja suhtelised variatsioonnäitarvud on informatiivsed enamasti vaid arvuliste tunnuste korral**.

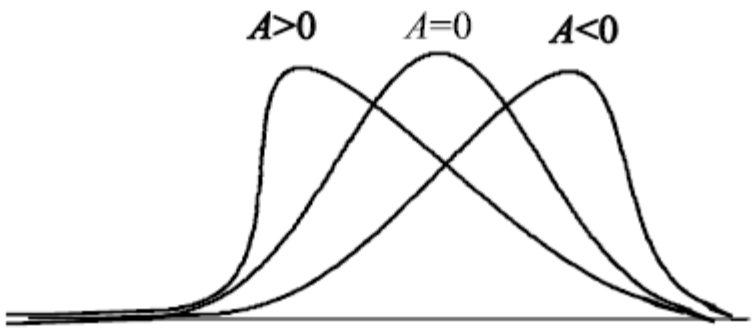
Järjestusskaala korral võib varieeruvuse hindamiseks kasutada näiteks kvantiilide vahesid.

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Jaotuse kuju iseloomustavad karakteristikud

- **Asümmeetriakordaja** – iseloomustab jaotuse sümmeetriat:

1. sümmeetrilise jaotuse korral *A=0.*
2. paremale (arvtelje positiivses suunas) väljavenitatud jaotuse korral on asümmeetriakordaja positiivne, vasakule väljavenitatud (negatiivses suunas) jaotuse korral negatiivne

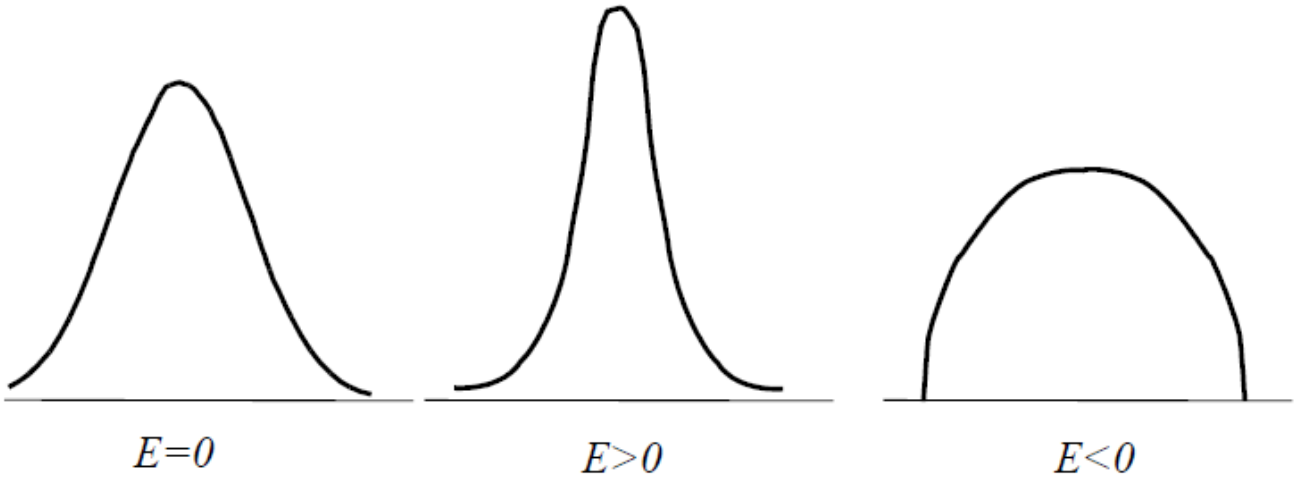


Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Jaotuse kuju iseloomustavad karakteristikud

- **Ekstsess** (*kurtosis*)– iseloomustab jaotuse püstakust:

1. Normaaljaotuse korral on E=0.
2. Kui püstakus on suurem, on jaotus kitsam. Väikese püstakuse korral “sabad” kaovad



Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

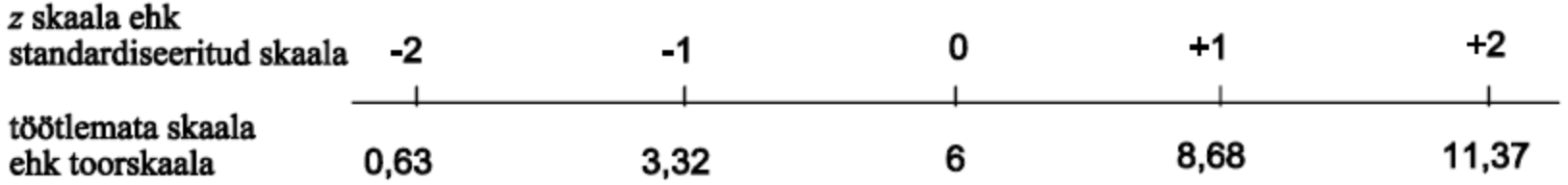
Jaotuse kuju iseloomustavad karakteristikud

NB! Asümmeetriakordajat ja ekstsessi on mõtet leida vaid suurte valimite korral (N = 30 või 50)

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Standardiseerimine

- **z-väärtus (z-score) näitab,** mitmekordse standardhälbe kaugusel keskväärtusest asub uuritava objekti väärtus.
- Standardiseeritud skaalal on keskväärtus alati 0 ja standardhälve 1.
- Eelis kasutamisel: tagab, et kõik analüüsis kasutatavad muutujad nö mängivad tulemuses ühesugust rolli



Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

III JOONISTE TEGEMINE EXCELIS

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Jooniste tüübid Excelis

- tulpdiagrammid
- joondiagrammid
- sektordiagrammid
- lintdiagrammid
- kihtdiagrammid
- xy-diagrammid (punktdiagrammid)
- börsidiagrammid
- pinddiagrammid
- rõngasdiagrammid
- mulldiagrammid
- radiaaldiagrammid

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Tulpdiagrammid

- sobivad mingi perioodi jooksul andmetes toimunud muutuste näitamiseks või üksuste võrdluse illustreerimiseks
- kategooriad on tavaliselt paigutatud horisontaalteljele ning väärtused vertikaalteljele

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Tulpdiagrammide tüübid

1) kobartulpdiagramm ja ruumiline kobartulpdiagramm - võrdlevad väärtusi kategooriate lõikes
2) virntulpdiagramm ja ruumiline virntulpdiagramm - kuvatakse üksikute elementide seos tervikuga, võrreldes eri kategooriate kõigi väärtuste osakaalu kogusummas
3) 100% virntulpdiagramm ja ruumiline 100% virntulpdiagramm - võrdlevad eri kategooriate kõigi väärtuste protsentuaalset osakaalu kogusummas
4) ruumiline tulpdiagramm - kasutatakse andmete võrdlemiseks korraga nii kategooriate kui ka sarjade lõikes
5) silinder-, koonus- ja püramiiddiagrammid

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Joondiagrammid

- kuvatakse ajaliselt järjestikused andmed ühisel skaalal, seega sobivad joondiagrammid hästi näiteks andmete trendi näitamiseks võrdsete ajavahemike tagant
- kategooriaandmed on jaotatud ühtlaselt horisontaalteljele ning väärtuste andmed ühtlaselt vertikaalteljele
- rohkem kui kümne arvsildi puhul tuleks kasutada punktdiagrammi

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Joondiagrammide tüübid

- joondiagramm ja tähistega joondiagramm - sobivad trendide kuvamiseks ajaliselt või järjestatud kategooriate kaupa, eriti kui andmepunkte on palju ja nende esitamise järjestus on oluline
- virnjoondiagramm ja tähistega virnjoondiagramm - saab kasutada iga väärtuse osakaalu trendi kuvamiseks ajaliselt või järjestatud kategooriate kaupa
- 100% virnjoondiagramm ja 100% tähistega virnjoondiagramm
- ruumiline joondiagramm

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Sektordiagramm

- kuvatakse ühe andmesarja elementide maht kõigi elementide kogusumma suhtes. Sektordiagrammil kuvatakse andmepunktid protsendina tervikust
- sektordiagrammi kasutatakse enamasti järgmistel juhtudel:
  - diagrammile paigutatakse ainult üks andmesari
  - ükski diagrammile paigutatavatest väärtustest pole negatiivne
  - diagrammile paigutatavate väärtuste hulgas pole peaaegu ühtegi nullväärtust
  - teil on maksimaalselt seitse kategooriat
  - kategooriad esitatakse sektordiagrammi osadena.

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Sektordiagrammi tüübid

- sektordiagramm ja ruumiline sektordiagramm
- sektordiagrammil sektordiagrammist või lintdiagrammil sektordiagrammist
- irdsektordiagramm ja ruumiline irdsektordiagramm

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Lintdiagrammid

- sobivad üksikute elementide võrdluste illustreerimiseks

Tüübid:

- kobarlintdiagramm ja ruumiline kobarlintdiagramm - võrdlevad väärtusi kategooriate lõikes
- virnlintdiagramm ja ruumiline virnlintdiagramm - kuvatakse üksikute elementide seos tervikuga
- 100% virnlintdiagramm ja ruumiline 100% virnlintdiagramm - võrdleb eri kategooriate kõigi väärtuste protsentuaalset osakaalu kogusummas
- horisontaalsed silinder-, koonus- ja püramiiddiagrammid

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

XY-diagrammid (punktdiagrammid)

- kuvatakse mitme andmesarja arvväärtuste seosed või kantakse diagrammile kaks arvude rühma ühe x- ja y-koordinaatide sarjana
- Horisontaalteljel (x-teljel) kuvatakse üks komplekt arvandmeid ja vertikaalteljel (y-teljel) teine. Need väärtused kombineeritakse andmepunktideks ja kuvatakse ebaühtlaste intervallide või kobaratena. Punktdiagramme kasutatakse tavaliselt arvandmete kuvamiseks ja võrdlemiseks

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

XY-diagrammide tüübid

- tähistega punktdiagramm
- sujuvjoontega punktdiagramm ning tähiste ja sujuvjoontega punktdiagramm
- sirgjoontega punktdiagramm ning sirgjoonte ja tähistega punktdiagramm

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Kihtdiagrammid

- rõhutavad aja jooksul toimunud muutuste suurusjärku ning neid saab kasutada tähelepanu juhtimiseks kogusummade trendile
- näitab osade seost tervikuga

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Teisi jooniste tüüpe Excelis

- börsidiagrammid
- pinddiagrammid
- rõngasdiagrammid
- mulldiagrammid
- radiaaldiagrammid

Nende kohta saate iseseisvalt rohkem lugeda konspektist

Tänan tähelepanu eest!

TARTU ÜLIKOOL
1632

STATISTILISE ANALÜÜSI TEOSTAMINE EXCELI JA SPSSI ABIL

Kerly Krillo

Tartu Ülikool, sotsiaalteaduslike rakendusuuringute keskus

Tööturu ja tööpoliitika programmi juht

kerly.krillo@ut.ee

27.02.2010

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

I LIIGENDTABELID

(inglise keeles PivotTable)

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Loeng/praktikumi eesmärk

Pärast selle kursuse lõpetamist

- mõistab tudeng PivotTable'i funktsiooni kasulikkust
- Oskab tudeng muuta PivotTable-aruande loomise abil andmeid arusaadavamaks

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Sissejuhatuseks – kuna liigendtabeleid kasutada?

PivotTable-aruanded võimaldavad mõne sekundiga töölehest uute vaadete loomist

- Väga kasulik abivahend, kui andmeid palju ja neist on keeruline ülevaadet saada
- Kui on kasulik saada kiiresti ülevaade andmete erinevatest dimensioonidest.
- PivotTable-aruande koostamine tähendab sisuliselt teabe osade üksteisega sobitamist

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Liigendtabelite võlu ja valu

- Andmeid on kõige parem korraldada, kui uurija teab, mida tal on vaja teada ☺

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Pole õigeid ja valesid lahendusi ☺

- Liigendtabelite tegemisel pole mõtet muretseda aruande valesti paigutamise pärast!
- PivotTable-aruande koostamise mõte ongi selles, et aruande koostaja saab välju ühest kohast teise liigutada, et näha, kuidas üks või teine paigutus välja näeb.
- Soovi korral on võimalik vaid paari klikiga teave ümber tõsta. Sellist teisaldamist nimetatakse paigutuse muutmiseks ning see on tööprotsessi loomulik osa.

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Nõuded lähteandmetele

- esimene rida peab sisaldama pealkirja igale veerule

  NB! Viisard kasutab neid veerupealkirju **väljade** (termin andmerühmade kohta) nimedena. neid välju on võimalik PivotTable-aruande paigutusalale lohistada ja kukutada
- aruandes kasutatavate andmete hulgas ei tohi olla tühje ridu ega veerge
- iga veerg peab sisaldama ainult ühte sorti andmeid (nt teksti või numbrilisi väärtusi)
- excel teeb pivottable-aruandes automaatselt vahekokkuvõtted ja üldkokkuvõtted. kui lähteandmed sisaldavad automaatseid vahekokkuvõtteid ning üldkokkuvõtteid, mis on tehtud menüü **andmed** käsu **vahesummad** abil, eemaldage need kokkuvõtted enne aruande koostamist sellesama käsu abil

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Liigendtabelite tegemine

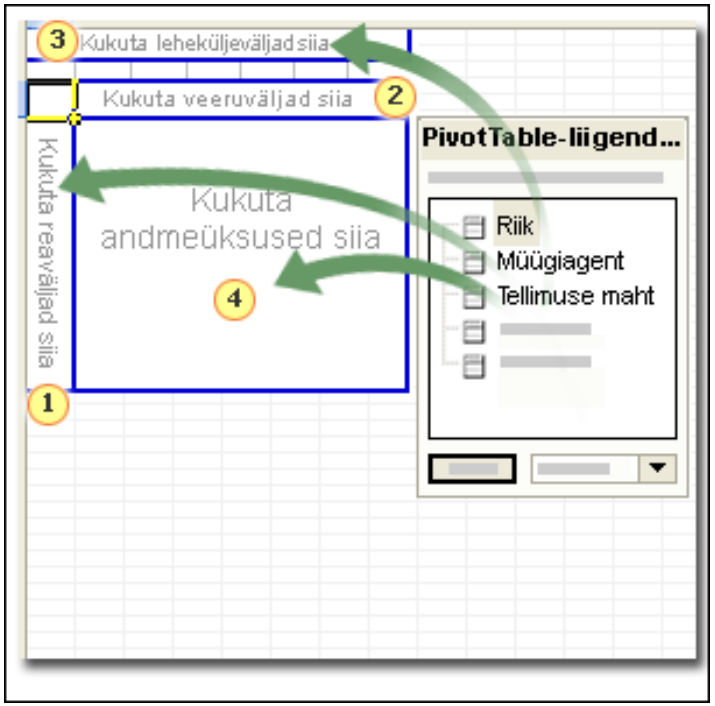
Viisardi (so abivahend) abil kuvatakse uus tööleht, kus on kõik vajalik PivotTable-aruande koostamiseks:

- **PivotTable-liigendtabeli väljaloend** (st "kastike", kust lohistatakse välju) ning
- paigutusala.

Uurijaülesandeks on valitud väljad loendist ühele neljast kukutusalast (reaväljade, veeruväljade, andmeüksuste või leheküljeväljade ala) lohistada.

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Liigendtabel – milline info kuhu?



1. Reaväljade alal kuvatakse andmed vertikaalselt, nii et ühel real on üks üksus

2. Veeruväljade alal kuvatakse andmed horisontaalselt, nii et ühes veerus on üks üksus

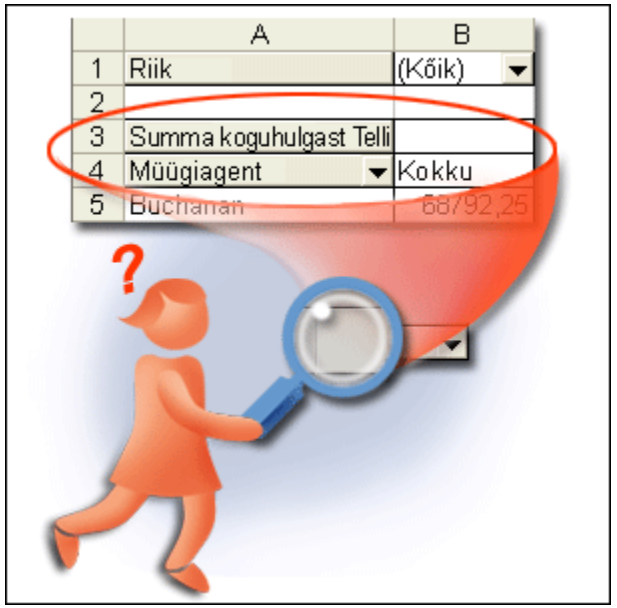
3. Leheküljeväljade alal kuvatakse andmeid lehekülgedena, rühmitades või eraldades niimoodi sinna lisatud andmeüksusi

4. Andmeüksuste ala on see, kus kuvatakse ja summeeritakse numbrilisi andmeid

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Hallid kastid, valged lahtrid

Iga PivotTable-aruande hall kast
Sisaldab mõne välja nime.
Excel asetab väljanimed automaatselt kastidesse, mis muudab need kergemini nähtavaks.
NB! Nimetus saab hõlpsalt muuta:
Selleks tuleb uus nimi sisestada ja
Vajutada klahvi “enter”



Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Mida saab veel teha?

- andmete sorteerimine
- andmete värskendamine - aruandesse muutuse tegemine võtab ainult mõne sekundi. Selleks tuleb klõpsata tööriistariba **PivotTable** nuppu **Värskenda andmed** ja aruannet värskendatakse uue muutusega
- Andmeid saab loendada, summeerida, leida miinimumi, maksimumi, keskmist…

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Harjutamine teeb meistriks!

Ja nüüd asume tööle praktilise näitega…

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

II SPSS

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Siin me pikalt sissejuhatust ei tee,…

- … vaid asume kohe asja kallale ja teeme SPSSiga tutvust

Tänan tähelepanu eest!

TARTU ÜLIKOOL

Töötamine suurte andmehulkadega

Janika Alloja

Tartu Ülikool, sotsiaalteaduslike rakendusuuringute keskus

Tööturu ja tööpoliitika programmi analüütik

Janika.Alloja@ut.ee

Milleks selline teema vajalik?

- Definitsioon *siin*: **palju ridu** ja/või palju veerge

- Suurte andmehulkade erisused
  - Raske saada ülevaadet
  - Arvutuste tegemine ajakulukas

- Kuidas analüüsida?
  - Andmete grupeerimine ja summeerimine
  - “Laiemalt kitsamale” analüüs
  - Uute muutujate loomine

- NB! Meetodid võivad töötada aeglaselt!
- NB! Baseerub Office 2007-l

- Meetode võib kasutada ka väiksemamahuliste andmebaaside korral

Nipid andmeanalüüsil

- Rida või veergude “külmutamine”

  View → Window → Freeze

- Veeru päis

  Merge Home → Alignment

  Wrap

- Ridade või veergude peitmine

  View → Window → Hide/Unhide

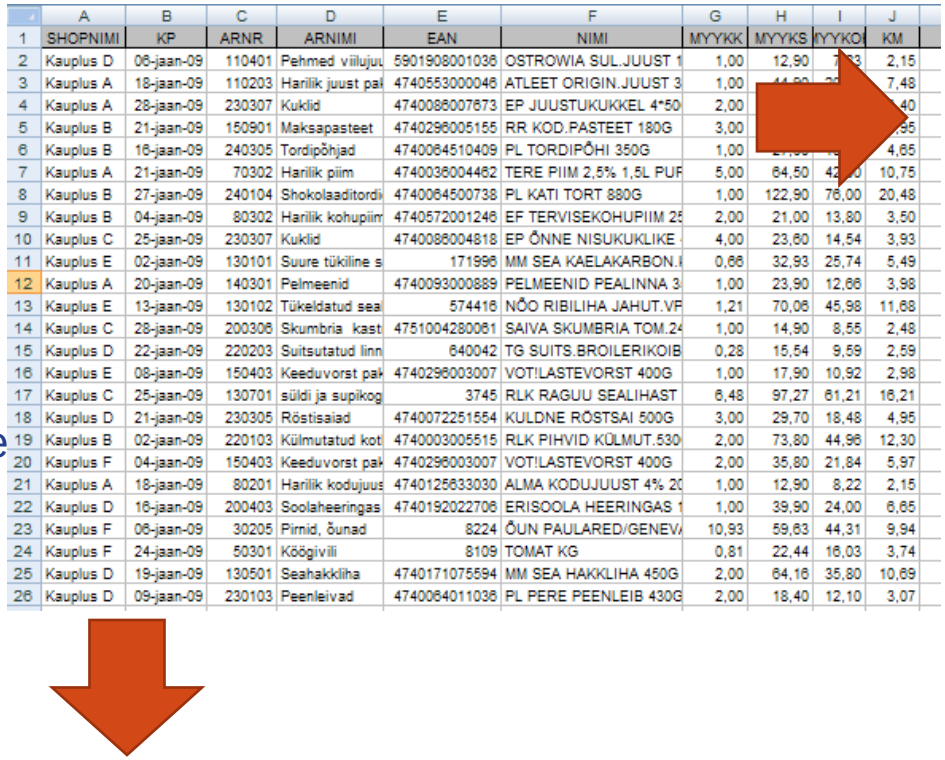
- Ebavajaliku info kustutamine

  crtl+shift+↓

- Valemi kopeerimine

  vasakul on kõigis ridades andmed

  vasakul on tühje lahtreid: shift ja crtl+D



	A	B	C	D	E	F	G	H	I	J
1	SHOPNIMI	KP	ARNR	ARNIMI	EAN	NIMI	MYYKK	MYYKS	MYYKO	KM
2	Kauplus D	06-jaan-09	110401	Pehmed viilujuu	5901908001036	OSTROWIA SUL.JUUST 1	1,00	12,90	7,53	2,15
3	Kauplus A	18-jaan-09	110203	Harilik juust pal	4740553000046	ATLEET ORIGIN.JUUST 3	1,00	44,90	29,…	7,48
4	Kauplus A	28-jaan-09	230307	Kuklid	4740086007673	EP JUUSTUKUKKEL 4*50	2,00	…	…	2,40
5	Kauplus B	21-jaan-09	150901	Maksapasteet	4740296005155	RR KOD.PASTEET 180G	3,00	…	…	2,95
6	Kauplus B	16-jaan-09	240305	Tordipõhjad	4740064510409	PL TORDIPÕHI 350G	1,00	27,90	18,…	4,65
7	Kauplus A	21-jaan-09	70302	Harilik piim	4740036004462	TERE PIIM 2,5% 1,5L PUF	5,00	64,50	42,00	10,75
8	Kauplus B	27-jaan-09	240104	Shokolaaditordi	4740064500738	PL KATI TORT 880G	1,00	122,90	76,00	20,48
9	Kauplus B	04-jaan-09	80302	Harilik kohupiim	4740572001246	EF TERVISEKOHUPIIM 25	2,00	21,00	13,80	3,50
10	Kauplus C	25-jaan-09	230307	Kuklid	4740086004818	EP ÕNNE NISUKUKLIKE	4,00	23,60	14,54	3,93
11	Kauplus E	02-jaan-09	130101	Suure tükiline s	171996	MM SEA KAELAKARBON.I	0,66	32,93	25,74	5,49
12	Kauplus A	20-jaan-09	140301	Pelmeenid	4740093000889	PELMEENID PEALINNA 3	1,00	23,90	12,66	3,98
13	Kauplus E	13-jaan-09	130102	Tükeldatud sea	574416	NÕO RIBILIHA JAHUT.VF	1,21	70,06	45,98	11,68
14	Kauplus C	28-jaan-09	200306	Skumbria kast	4751004280061	SAIVA SKUMBRIA TOM.24	1,00	14,90	8,55	2,48
15	Kauplus D	22-jaan-09	220203	Suitsutatud linn	640042	TG SUITS.BROILERIKOIB	0,28	15,54	9,59	2,59
16	Kauplus E	08-jaan-09	150403	Keeduvorst pak	4740296003007	VOT!LASTEVORST 400G	1,00	17,90	10,92	2,98
17	Kauplus C	25-jaan-09	130701	süldi ja supikog	3745	RLK RAGUU SEALIHAST	6,48	97,27	61,21	16,21
18	Kauplus D	21-jaan-09	230305	Röstisaiad	4740072251554	KULDNE RÖSTSAI 500G	3,00	29,70	18,48	4,95
19	Kauplus B	02-jaan-09	220103	Külmutatud kot	4740003005515	RLK PIHVID KÜLMUT.530	2,00	73,80	44,96	12,30
20	Kauplus F	04-jaan-09	150403	Keeduvorst pak	4740296003007	VOT!LASTEVORST 400G	2,00	35,80	21,84	5,97
21	Kauplus A	18-jaan-09	80201	Harilik kodujuus	4740125633030	ALMA KODUJUUST 4% 20	1,00	12,90	8,22	2,15
22	Kauplus D	16-jaan-09	200403	Soolaheeringas	4740192022706	ERISOOLA HEERINGAS 1	1,00	39,90	24,00	6,65
23	Kauplus F	06-jaan-09	30205	Pirnid, õunad	8224	ÕUN PAULARED/GENEVA	10,93	59,63	44,31	9,94
24	Kauplus F	24-jaan-09	50301	Köögivili	8109	TOMAT KG	0,81	22,44	16,03	3,74
25	Kauplus D	19-jaan-09	130501	Seahakkliha	4740171075594	MM SEA HAKKLIHA 450G	2,00	64,16	35,80	10,69
26	Kauplus D	09-jaan-09	230103	Peenleivad	4740064011036	PL PERE PEENLEIB 430G	2,00	18,40	12,10	3,07

Kui on vaja leida vaid 1 näitaja

- Tingimuslik loendamine

  COUNTIF(range,criteria)

  COUNTIFS(range1,criteria1,range2,criteria2…)

- Tingimuslik summeerimine

  SUMIF(range,criteria,sum_range)

  SUMIFS(sum_range,criteria_range1,criteria1, criteria_range2,criteria2…)

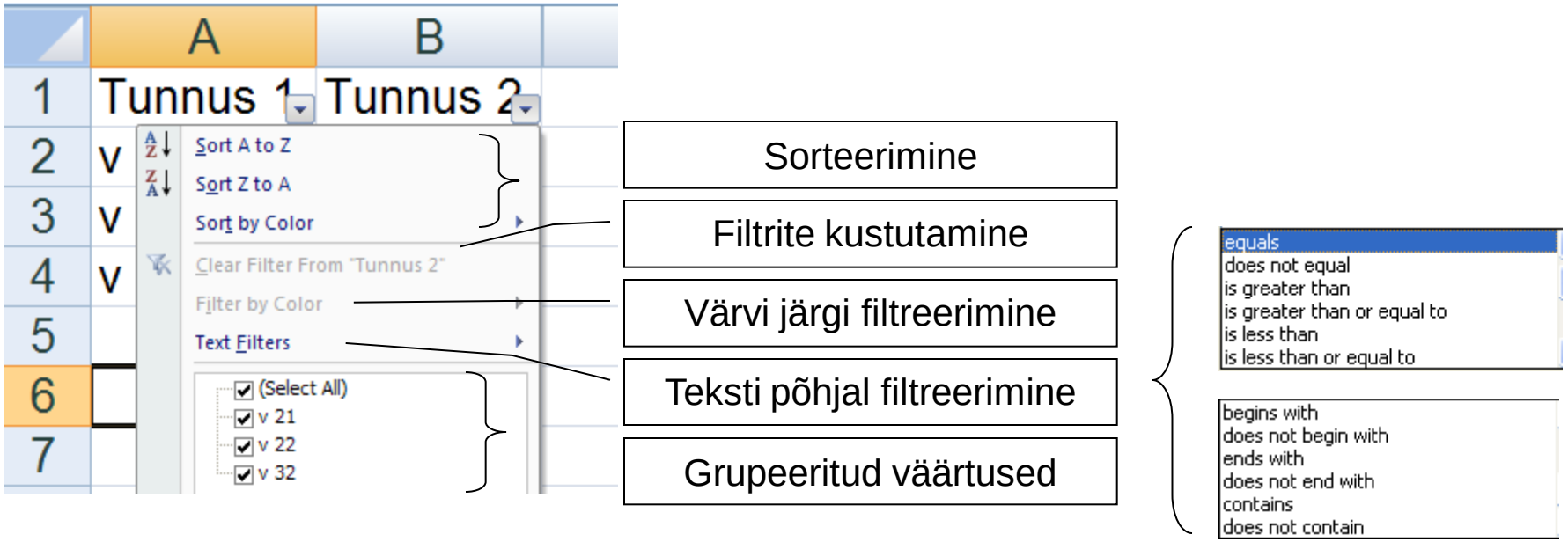
- Kaalutud summeerimine

  SUMPRODUCT(array1,array2,array3, ...)

	A	B	F	G	N
1	SHOPNIMI	KP	NIMI	MYYKK	Hind
2	Kauplus A	19-jaan-09	EP JUUBELISAI 350G	2,00	13,30
3	Kauplus A	03-jaan-09	MH SINGIPTSA KÜLMUT.200G	1,00	16,90
4	Kauplus A	07-jaan-09	RLK FRIKADELLID 350G	2,00	26,90
5	Kauplus A	10-jaan-09	KARTUL KG	4,17	3,90
6	Kauplus A	15-jaan-09	VÄIKE TOM VAN-ROS 125ML	11,00	5,12
7	Kauplus A	29-jaan-09	RAKVERE VERIKÄKK 440G	3,00	17,90
8	Kauplus A	07-jaan-09	TERE PEREJ.MUSTIKA 400G	1,00	13,50
9	Kauplus A	15-jaan-09	SLPT MUHU P/S VORST 320G	1,00	26,90
10	Kauplus A	18-jaan-09	VICI KÖÖGIV.BURGER 300G	2,00	21,90
11	Kauplus A	23-jaan-09	NÕO T/S VORST LOSSI 105G	2,00	22,90
12	Kauplus A	09-jaan-09	JUUST HOLLANDI LEIB 150G	5,00	18,75
13	Kauplus A	09-jaan-09	ALMA JOGURT MAASIKA 1KG	4,00	12,50
14	Kauplus A	14-jaan-09	VICI KRABIPULGAD 300G	1,00	23,90
15	Kauplus A	11-jaan-09	RM BROILER KILES KG	1,76	36,67
16	Kauplus A	11-jaan-09	EP KODUP.RUKKILEIB 800G	8,00	18,90
17	Kauplus A	03-jaan-09	TERE MARJA HAPS 1KG	2,00	19,60
18	Kauplus A	08-jaan-09	VICI KRABIPULGAD 300G	1,00	23,90
19	Kauplus A	29-jaan-09	TERE ROSINATORU 300G	4,00	13,50

Andmete grupeerimine - Filter

- Home → Editing→ Sort&Filter
- Data → Sort&Filter



Andmete grupeerimine ja summeerimine - Subtotal

- Data→Outline→Subtotal

Vaade 1: Grand Total

Vaade 2: Grupid kokku

Vaade 3: Kõik andmed

	A	B	C	D	E
1	ARNIMI	SHOPNIMI	EAN	NIMI	MYYKK
2	Dieetjogurt	Kauplus A	4740125522105	ALMA A+B JOG.MAASIKA 1KG	1,00
3	Dieetjogurt	Kauplus A	4740125538076	A JOG.MUSTIKA-JÕHV.150G	1,00
4	Dieetjogurt	Kauplus E	4740125522105	ALMA A+B JOG.MAASIKA 1KG	1,00
5	Dieetjogurt	Kauplus F	4740125522105	ALMA A+B JOG.MAASIKA 1KG	1,00
6	**Dieetjogurt Total**				4,00
7	Harilikud kreemid	Kauplus B	4740113034382	F KOHUP.KR MAASIKA 150G	2,00
8	Harilikud kreemid	Kauplus C	4740113034511	F KOHUP.KR VIRSIKU 150G	1,00
9	Harilikud kreemid	Kauplus F	4740113034511	F KOHUP.KR VIRSIKU 150G	2,00
10	**Harilikud kreemid Total**				5,00
11	Harilikud pudingud	Kauplus B	4740036001126	T VANILLIPUDING 125G	1,00
12	Harilikud pudingud	Kauplus D	4740036005728	TERE KARAM.PUDING 250G	2,00
13	Harilikud pudingud	Kauplus F	4740036001126	T VANILLIPUDING 125G	1,00
14	**Harilikud pudingud Total**				4,00
15	**Grand Total**				13,00

Funktsioon SUBTOTAL

Andmete summeerimine – Subtotal (funktsioon SUBTOTAL)

- *SUBTOTAL(function_num, ref1, ref2, ...)*

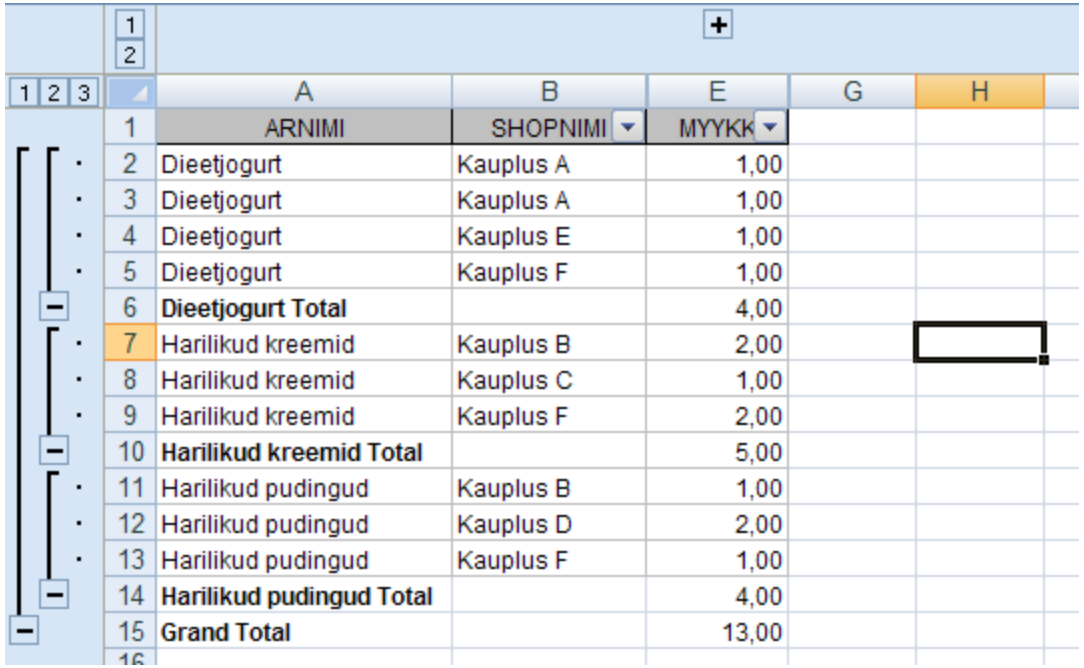
Function_ number	Function
1	AVERAGE
2	COUNT
3	COUNTA
4	MAX
5	MIN
6	PRODUCT
7	STDEV
8	STDEVP
9	SUM
10	VAR
11	VARP

Andmete grupeerimine - Group

- Data→Outline→Group

Veergude grupeerimine
Group

Ridade grupeerimine
Subtotal või Group



	A	B	E	G	H
1	ARNIMI	SHOPNIMI	MYYKK		
2	Dieetjogurt	Kauplus A	1,00		
3	Dieetjogurt	Kauplus A	1,00		
4	Dieetjogurt	Kauplus E	1,00		
5	Dieetjogurt	Kauplus F	1,00		
6	**Dieetjogurt Total**		4,00		
7	Harilikud kreemid	Kauplus B	2,00		
8	Harilikud kreemid	Kauplus C	1,00		
9	Harilikud kreemid	Kauplus F	2,00		
10	**Harilikud kreemid Total**		5,00		
11	Harilikud pudingud	Kauplus B	1,00		
12	Harilikud pudingud	Kauplus D	2,00		
13	Harilikud pudingud	Kauplus F	1,00		
14	**Harilikud pudingud Total**		4,00		
15	**Grand Total**		13,00		

Uute muutujate loomine

- Lihtsad arvutused

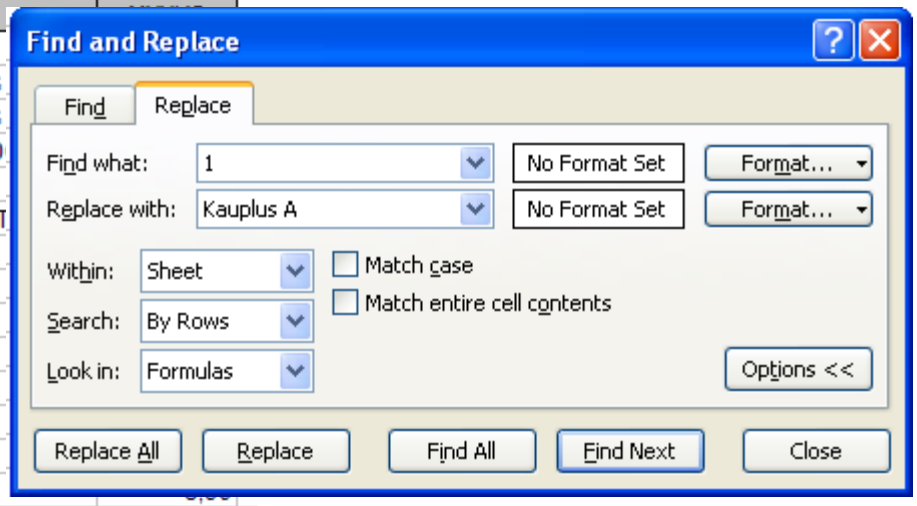
Näide : kokku

Kulutused leibkonnaliikme kohta = Leibkonna kulutused

Leibkonnaliikmete arv

- Asendamine Replace (Home→ Editing) /tinglikult/

	A	B	C
1	Kaupluse kood	NIMI	
2	1	A JOG.MAIUS KIRSI 200G	
3	2	A JOG.MUSTIKA-JÕHV.150G	
4	2	A JOG.MUSTIKA-JÕHV.150G	
5	2	A KOH.KREEM MAASIKA 150	
6	2	BANAAN CHIQUITA KG	
7	1	BONO KOH.DESSERT MUST	
8	1	EP HEA SAI 300G VIIL	
9	1	EP JUUSTUKUKKEL 4*50G	
10	1	EP KANEELISAIAKE 3*55G	
11	2	EP MOSKVA SAIAKE 3*45G	
12	2	EP MOSKVA SAIAKE 3*45G	
13	1	EP MUST VORMILEIB 280G	
14	2	EP MUST VORMILEIB 280G	
15	1	EP RUKKITASKU 4TK 270G	



Uute muutujate loomine

- Funktsioon IF *IF(logical_test,value_if_true,value_if_false)*

	A	B	C
1	NIMI	MYYKS	Müügihinnang
2	APELSIN NAVELINA KG	271,07	müügihitt
3	BAUER PORGAND MINI 400G	14,90	=IF(B3>=200;"müügihitt";"tavaline")
4	BONO KOH.DESSERT MUSTIKA	12,30	tavaline
5	BR.KOIB KÜLMUTATUD 1,5KG	59,90	tavaline
6	EP KODUP.RUKKILEIB 800G	325,50	müügihitt
7	EP MUST VORMILEIB 280G	20,70	tavaline

Tunnuseid 1-3

- Funktsioon VLOOKUP *VLOOKUP(lookup_value,table_array,col_index_num,range_lookup)*

	A	B	C	D	E	F	G
1	Kaupluse kood	NIMI	MYYKS	Kaupluse nimi		*Kaupluse kood*	*Kaupluse nimi*
2	1	A JOG.MAIUS KIRSI 200G	8,92	=VLOOKUP(A2;F:G;2;FALSE)		*1*	*Kauplus A*
3	4	A JOG.MUSTIKA-JÕHV.150G	5,90	Kauplus D		*2*	*Kauplus B*
4	7	A JOG.MUSTIKA-JÕHV.150G	5,90	Kauplus G		*3*	*Kauplus C*
5	2	A KOH.KREEM MAASIKA 150G	9,10	Kauplus B		*4*	*Kauplus D*
6	9	BANAAN CHIQUITA KG	5,92	Kauplus I		*5*	*Kauplus E*
7	7	BONO KOH.DESSERT MUSTIKA	4,10	Kauplus G		*6*	*Kauplus F*
8	7	EP HEA SAI 300G VIIL	4,90	Kauplus G		*7*	*Kauplus G*
9	9	EP JUUSTUKUKKEL 4*50G	6,90	Kauplus I		*8*	*Kauplus H*
10	1	EP KANEELISAIAKE 3*55G	8,50	Kauplus A		*9*	*Kauplus I*
11	2	EP MOSKVA SAIAKE 3*45G	8,50	Kauplus B			
12	2	EP MOSKVA SAIAKE 3*45G	8,50	Kauplus B			
13	3	EP MUST VORMILEIB 280G	6,90	Kauplus C			
14	5	EP MUST VORMILEIB 280G	6,90	Kauplus E			

Tunnuseid üle 4

Andmete analüüs - PivotTable

PivotTable (Insert → Tables) võimaldab:

1. Andmete grupeerimine ja summeerimine

Sum of MYYKS	SHOPNI						
KP	Kauplus A	Kauplus B	Kauplus C	Kauplus D	Kauplus E	Kauplus F	Grand Total
1.01.2009	4 046	3 705	2 046	1 711	1 305	3 205	16 018
2.01.2009	12 576	15 123	5 937	4 381	4 548	7 250	49 815
3.01.2009	9 154	11 585	4 658	3 500	3 964	7 638	40 500
4.01.2009	20 372	21 047	9 067	7 545	4 646	10 320	72 997
5.01.2009	15 739	28 018	11 596	5 112	5 289	6 622	72 376
6.01.2009	21 834	21 239	8 543	6 147	5 023	9 832	72 618
7.01.2009	17 240	15 619	7 738	5 977	4 769	11 189	62 531

Lihtne risttabel

Sum of MYYKS		SHOPNIMI		
ARNIMI	NIMI	Kauplus A	Kauplus B	Kauplus C
⊞Arbuus,melon		48	311	
⊞Banaan		3 891	4 142	1 788
⊟Dessertkohupiimakreemid	KOH.KR.MUSTIKAKISELLIGA	47	74	84
	KOH.KR.RABAR.KISSELLIGA	28	28	37
	KOH.KR.VAARIKAKISSELLIGA	19	65	
	TERE KOH.KR APRIK.150G	133	195	65
	TERE KOH.KR MAASIKA 150G	123	279	
	TERE KOH.KR. MURELI	57	112	93
Dessertkohupiimakreemid Total		406	752	279
⊞Dieetjogurt		325	294	193
⊞Dieetkohupiimakreemid		78	190	114
⊞Eksootilised puuviljad		810	1 692	8
⊞Feta juustud		92	135	45
⊞Fileed,palad paneeritud		173	19	19
⊞Friikartul		1 731	855	440
⊞Frikadellid		2 929	2 794	915

"Laiemalt kitsamale" risttabel

Andmete analüüs - PivotTable

2. Erinevuste, muutuste, struktuuri arvutamine

3		Data KP			
4		Sum of MYYKS		Sum of MYYKS2	
5	ARNIMI	15.01.2009	31.01.2009	15.01.2009	31.01.2009
6	Banaan	871	291		-66,6%
7	Dessertkohupiimakreemid	187	66		-65,0%
8	Dieetjogurt	62	12		-81,5%
9	Eksootilised puuviljad	187	69		-63,0%
10	Friikartul	349	116		-66,7%
11	Frikadellid	423	276		-34,8%
12	Hallitusjuustud	89	17		-80,9%
13	Hapupiim	47	16		-65,6%
14	Harilik hapukoor	2 497	1 103		-55,8%
15	Harilik juust pakitud	2 511	1 043		-58,4%
16	Harilik juust viilu	1 296	679		-47,6%
17	Harilik kodujuust	803	202		-74,9%

% muutus

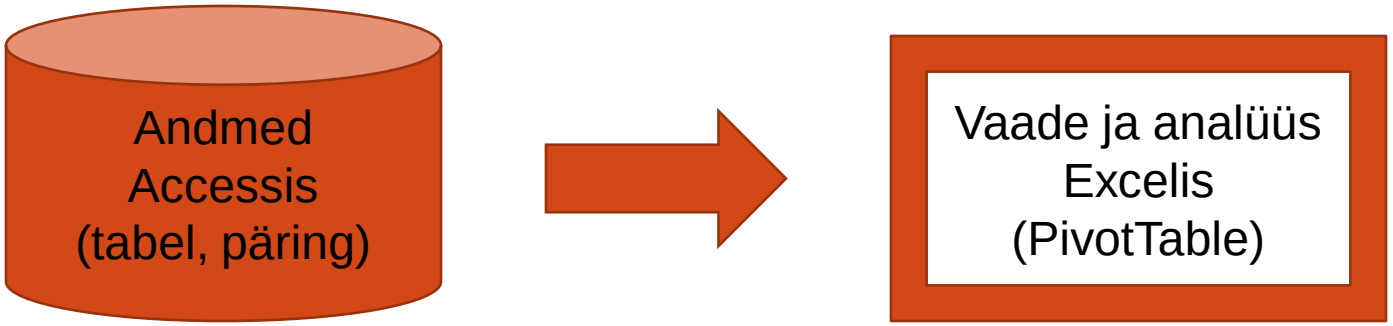
3. Uute muutujate arvutamine

3		Data		
4	ARNIMI	Sum of MYYKS	Sum of MYYKK	Sum of Hind
5	Arbuus,melon	640,95	27,56	23,26
6	Banaan	16 033,36	935,14	17,15
7	Dessertkohupiimakreemid	3 833,61	408,00	9,40
8	Dieetjogurt	1 664,80	118,00	14,11
9	Eksootilised puuviljad	3 317,74	132,24	25,09
10	Friikartul	5 374,58	295,00	18,22
11	Frikadellid	10 479,12	394,00	26,60
12	Galetid	30,80	2,00	15,40
13	Grill-linnuliha	68,90	1,00	68,90

Uus muutuja

Kui andmeid on väga palju …

- Loo andmebaas Accessi ja seo PivotTable kaudu Exceliga (Data→ Get External Data)



Kokkuvõtteks

		Plussid	Miinused
ANALÜÜS	Filter	Kiire Saab grupeerida	Ei saa summeerida Pole paindlik
	Subtotal	Saab grupeerida ja summeerida	Pole paindlik
	Pivot	Väga paindlik	?
UUTE MUUTUJATE LOOMINE	Filter	Numbriliste ja tekstiliste tunnuste asendamine	Ainult asendamine
	Replace	Numbriliste ja tekstiliste tunnuste asendamine Case-sensitive	Ainult asendamine
	IF	Numbriliste ja tekstiliste tunnuste asendamine Saab kasutada >,< Mitu tunnust ja muutujat	Mitme tunnuse ja muutuja korral läheb kiiresti keeruliseks
	VLOOKUP	Numbriliste ja tekstiliste tunnuste asendamine Võib olla väga palju tunnuseid	Ei saa kasutada >,< Ainult 1 muutuja baasil
	Pivot	Väga kiire	Uusi muutujad saab luua ainult arvutamise teel

Teadmiste kinnistamine

- Praktikumiülesanne
- Kodune mõtteharjutus

Tänan tähelepanu eest!

TARTU ÜLIKOOL
1632

STATISTILISE ANALÜÜSI TEOSTAMINE EXCELI JA SPSSI ABIL

Kerly Krillo

Tartu Ülikool, sotsiaalteaduslike rakendusuuringute keskus

Tööturu ja tööpoliitika programmi juht

kerly.krillo@ut.ee

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Esmalt paar jooksvat küsimust

Kokkusaamine mais – nihutaks selle 22. mailt **15. maile**?

Kodutööde tagasiside

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Tänase praktikumi teema:

I statistiline andmeanalüüs SPSSiga

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Loeng/praktikumi eesmärk

Pärast selle praktikumi läbimist

- Oskab tudeng iseseisvalt teostada lihtsamat statistilist analüüsi SPSSi abil ja saadud tulemusi sisukalt tõlgendada

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

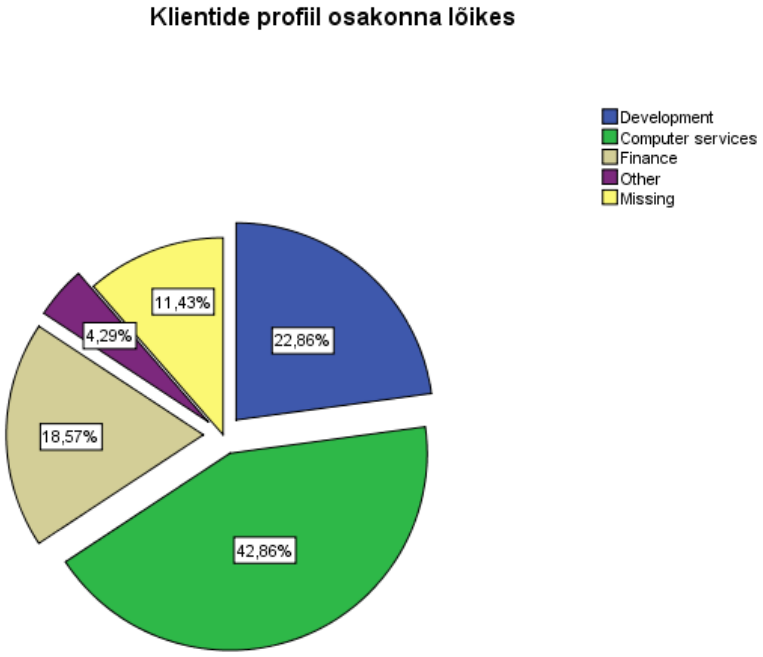
Sagedused (Frequencies)

- Nn **esimene pilguheit** andmetele (muuhulgas annavad ülevaate, kas andmeid on piisavalt, et usaldusväärset analüüsi teostada, kui palju on puuduvaid väärtusi, kas on sisestamisvigu jne)
- Kasutatakse peamiselt kategooriliste (st nominaal- ja järjestus-) tunnuste puhul ning võimaldab hõlpsalt saada ülevaade, millised on muutuja "tüüpilised" väärtused (st millised väärtused esinevad sagedamini, millised harvemini), millises vahemikus muutuja väärtused varieeruvad jne

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Sagedused (Frequencies) - nominaaltunnus

- **Analyze → Descriptive Statistics → Frequencies**
- Avame andmefaili **“contacts.sav”**



Footnote

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Sagedused (Frequencies) - nominaaltunnus

Osakaal nendest, kellel on muutuja väärtus olemas (st pole "Missing")

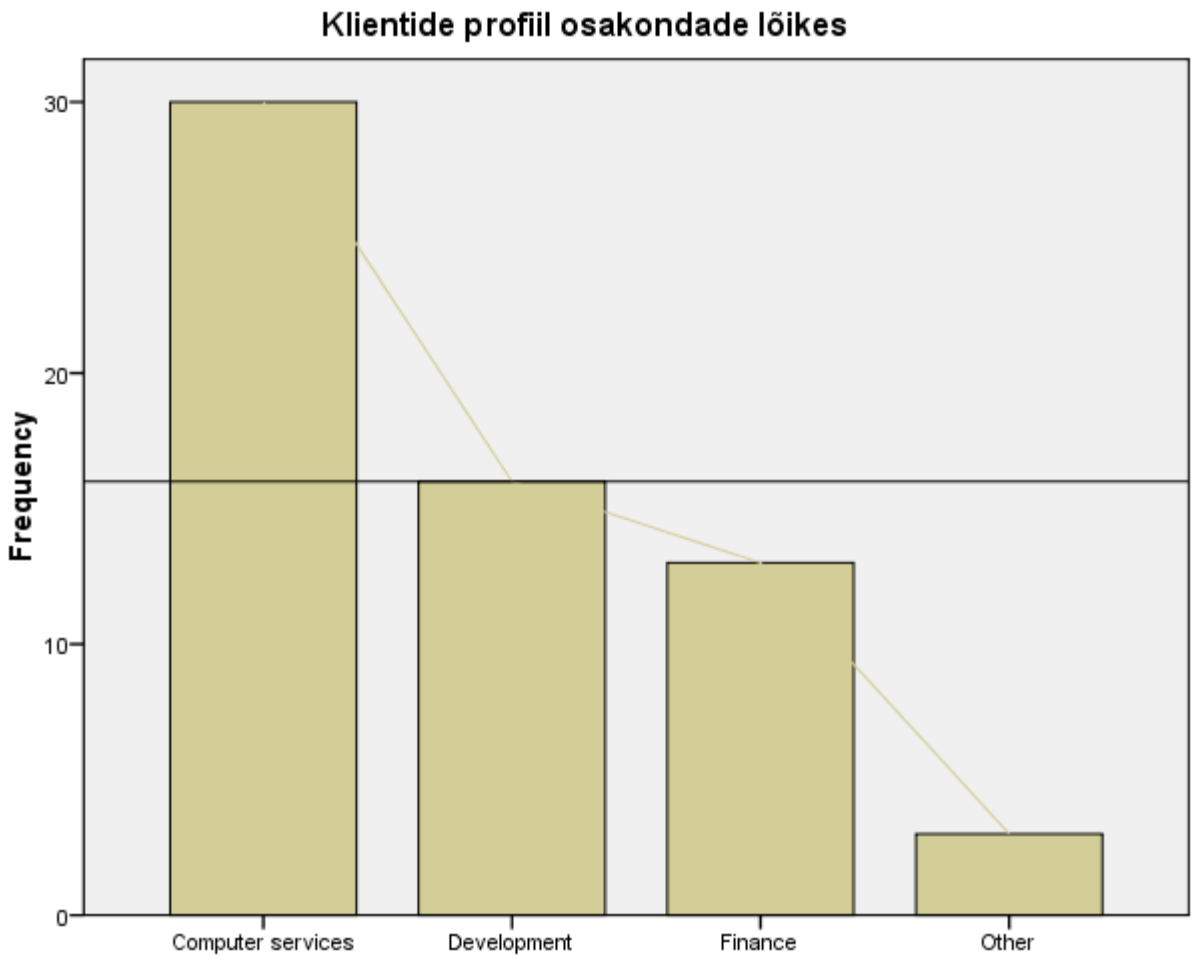
Osakaal kõikidest

Department

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Development	16	22,9	25,8	25,8
	Computer services	30	42,9	48,4	74,2
	Finance	13	18,6	21,0	95,2
	Other	3	4,3	4,8	100,0
	Total	62	88,6	100,0	
Missing	Don't know	8	11,4		
Total		70	100,0		

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Sagedused (Frequencies) - nominaaltunnus



Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Sagedused (Frequencies) - järjestustunnus

- Järjestustunnuste korral on kumulatiivset osakaalu kajastaval veerul suurem sisu kui nominaalsete andmete korral

62% meie kontaktidest kuuluvad vähemasti kõrgemasse Juhtkonda (senior manager)

Company rank

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Pres/CEO/CFO	6	8,6	10,2	10,2
	VP	13	18,6	22,0	32,2
	Sr. manager	18	25,7	30,5	62,7
	Jr. manager	11	15,7	18,6	81,4
	Employee	11	15,7	18,6	100,0
	Total	59	84,3	100,0	
Missing	Don't know	11	15,7		
Total		70	100,0		

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Sagedused (Frequencies) – pidev tunnus

- Ei ole mõtet “tellida” sagedustabelit, vaid pigem kirjeldavate statistikute tabel

Statistics

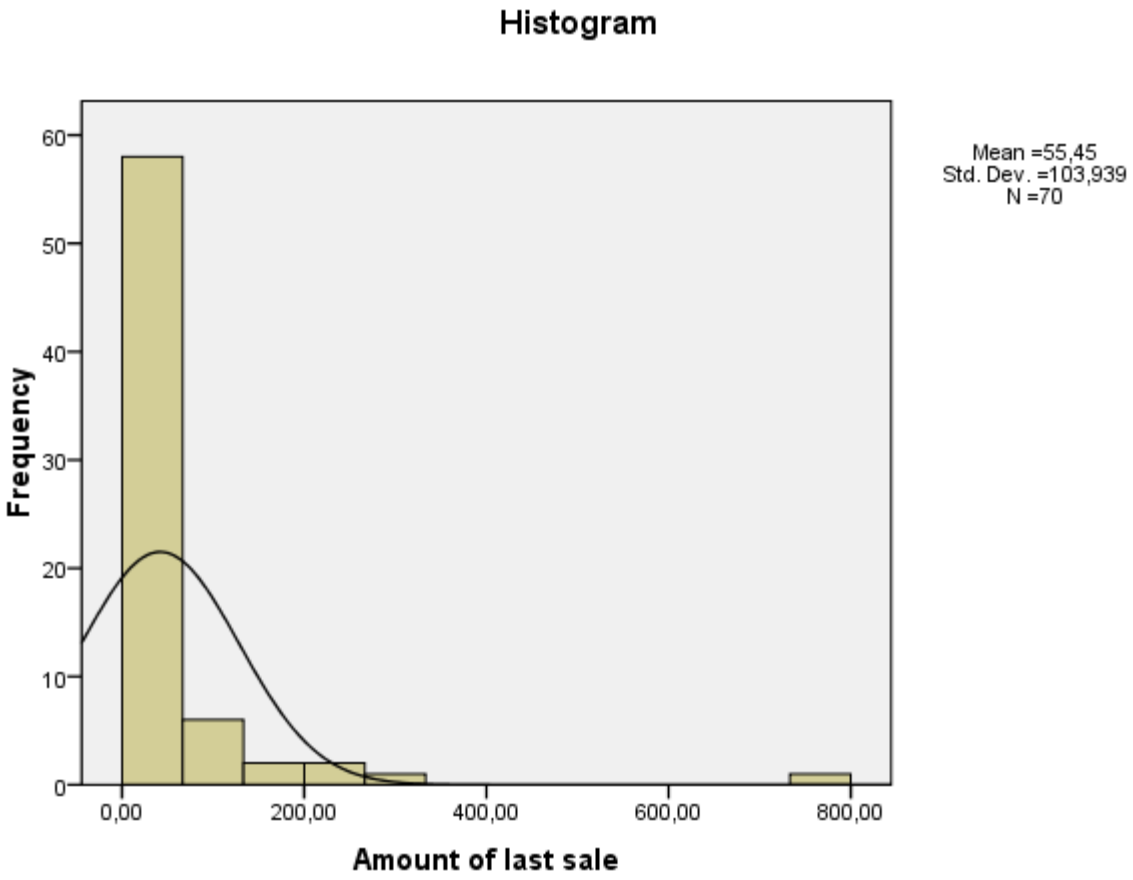
Amount of last sale

N	Valid	70
	Missing	0
Mean		55,4500
Median		24,0000
Std. Deviation		103,93940
Skewness		5,325
Std. Error of Skewness		,287
Kurtosis		34,292
Std. Error of Kurtosis		,566
Minimum		6,00
Maximum		776,50
Percentiles	25	12,0000
	50	24,0000
	75	52,8750

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Sagedused (Frequencies) – pidev tunnus

- Histogramm



Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Sagedused (Frequencies) – andmete transformeerimine

- Kui pideva muutuja korral ei ole andmed kaugeltki normaaljaotusega, on paljude statistiliste protseduuride tulemused ebausaldusväärsed
- Seda probleemi aitab teatud juhul lahendada muutujate transformeerimine, mis viib teisendatud muutuja jaotuse normaaljaotusele lähedasemaks
- Tüüpilisim transformatsioon ehk teisendus on logaritmimine

Transform → Compute Variable (loome muutuja ln (sale)

Sagedused (Frequencies) – andmete transformeerimine

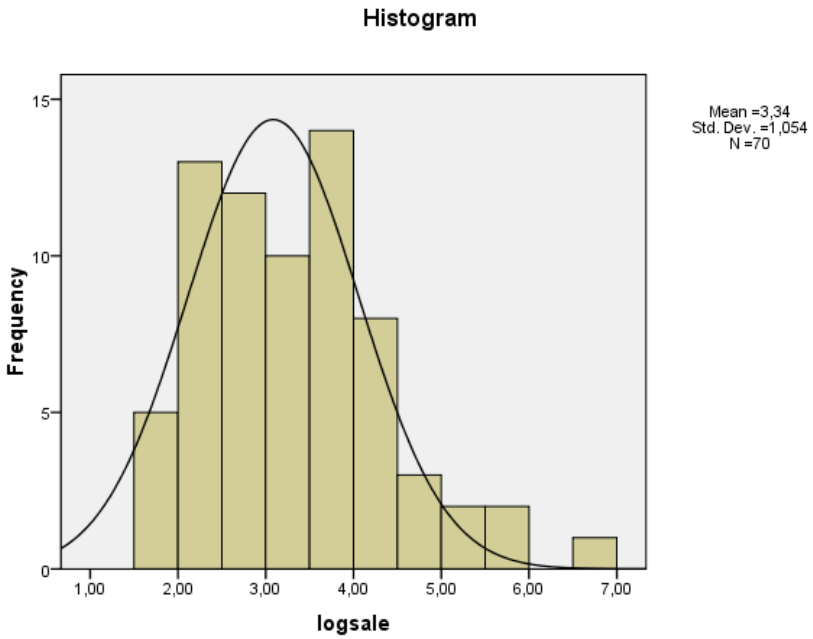
Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

- Tulemused

Statistics

logsale

N	Valid	70
	Missing	0
Mean		3,3373
Median		3,1772
Std. Deviation		1,05361
Skewness		,721
Std. Error of Skewness		,287
Kurtosis		,367
Std. Error of Kurtosis		,566
Minimum		1,79
Maximum		6,65
Percentiles	25	2,4849
	50	3,1772
	75	3,9679



Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Explore

- **Analyze → Descriptive Statistics → Explore**

NB! Statistikute tabelit saab liigendada! Selleks tuleb teha tabelil topeltklikk, valida **Pivot → Pivoting Trays…**

Descriptives

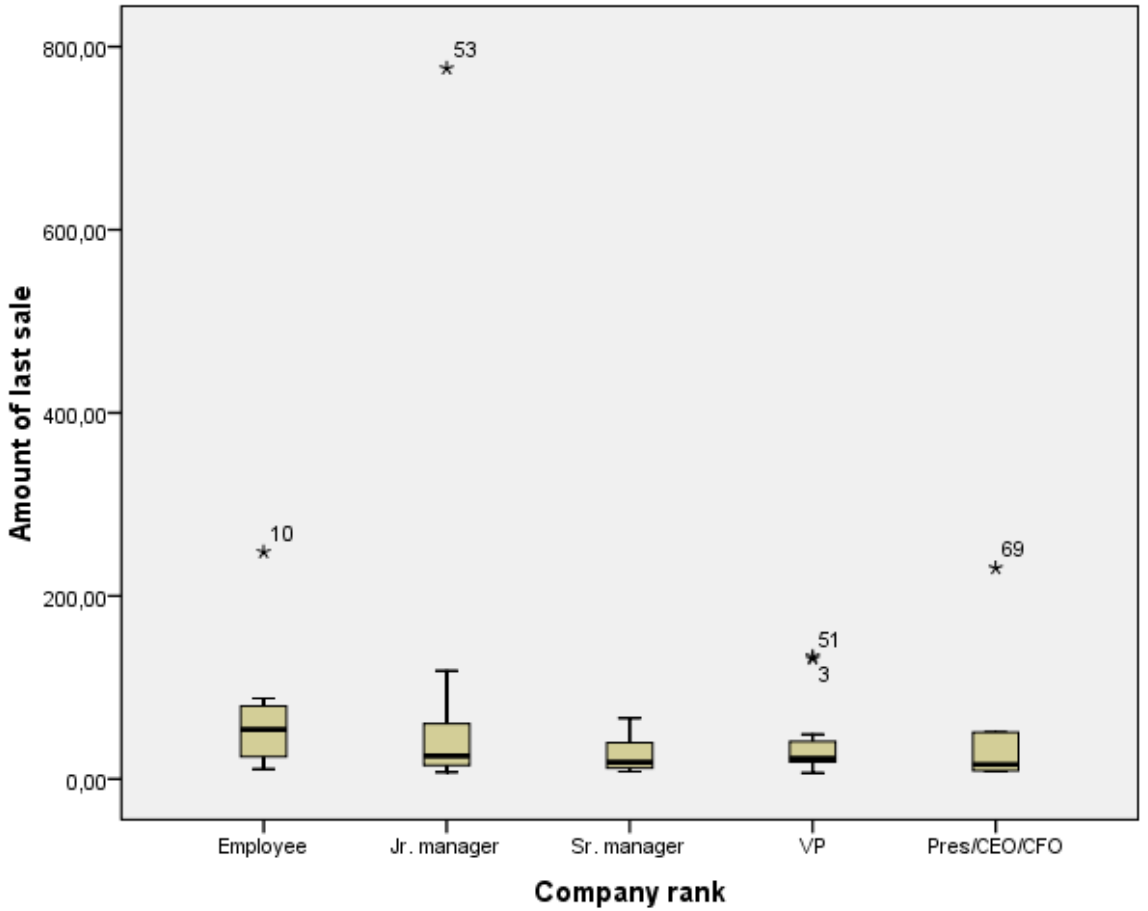
Statistics= Median

	Company rank	Statistic
Amount of last sale	Employee	54,0000
	Jr. manager	25,0000
	Sr. manager	18,2500
	VP	22,5000
	Pres/CEO/CFO	15,7500

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Explore

- **Karpdiagramm –** hea vahend eri kategooriate võrdlemiseks



Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Explore

- Võimalik on leida ka ekstreemsed väärtused, testida normaaljaotust jne.

*Kui Sig. < olulisuse tõenäosus (tavaliselt 0.05), siis **ei ole** tegu normaaljaotusega*

Tests of Normality

	Company rank	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
Amount of last sale	Employee	,288	11	,011	,732	11	,001
	Jr. manager	,384	11	,000	,464	11	,000
	Sr. manager	,218	18	,023	,862	18	,013
	VP	,283	13	,005	,716	13	,001
	Pres/CEO/CFO	,352	6	,020	,630	6	,001

a. Lilliefors Significance Correction

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Explore

- **Tüvi-ja-leht (*Stem-and-leaf*) diagramm**

Loe eelistest võrreldes histogrammi ja sagedustabeliga:

http://www.purplemath.com/modules/stemleaf.htm

Amount of last sale Stem-and-Leaf Plot
for
rank= Sr. manager

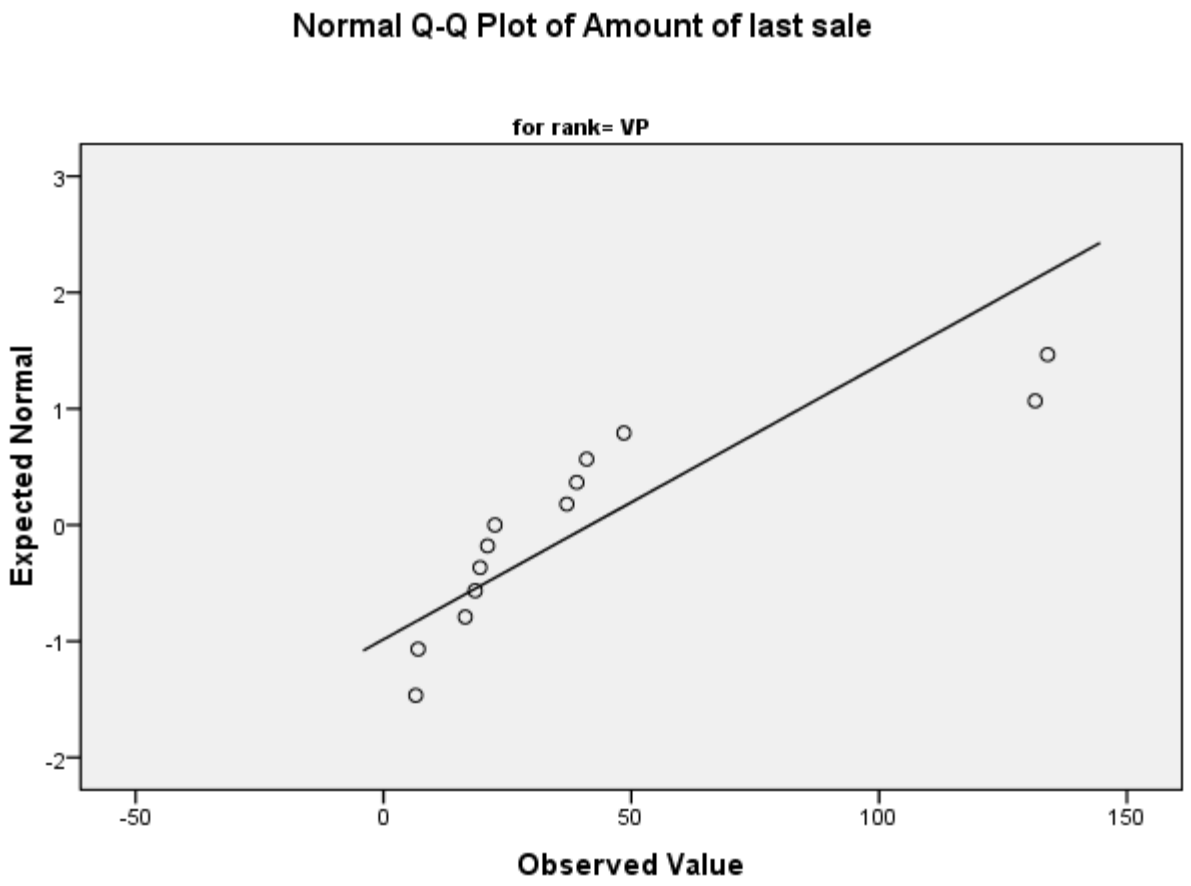
Frequency	Stem & Leaf
4,00	0 . 8899
5,00	1 . 22445
3,00	2 . 139
2,00	3 . 59
2,00	4 . 89
1,00	5 . 8
1,00	6 . 6

Stem width: 10,00
Each leaf: 1 case(s)

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Explore

- Q-Q joonis – näitab kõrvalekaldeid normaaljaotusest



Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Risttabelid

- Kasutatakse kahe kategoorilise (st nominaal- või järjestus-) tunnuse vaheliste seoste analüüsimiseks
- SPSS-is on võimalik kontrollida sõltumatust ka statistiliste testidega

- **Analyze → Descriptive Statistics → Crosstabs**
- Avame andmefaili **“satisf.sav”**

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Risttabelid

… kahe muutuja vaheliste seoste analüüs

Store * Service satisfaction Crosstabulation

Count

		Service satisfaction					
		Strongly Negative	Somewhat Negative	Neutral	Somewhat Positive	Strongly Positive	Total
Store	Store 1	25	20	38	30	33	146
	Store 2	26	30	34	27	19	136
	Store 3	15	20	41	33	29	138
	Store 4	27	35	44	22	34	162
Total		93	105	157	112	115	582

… üksi ei võimalda siiski teha järeldusi, kas erinevused on “tõelised” või üksnes juhuslikud

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Risttabelid

… küll aga saab selliseid järeldusi teha, tuginedes hii-ruut testile

Kui Sig.< olulisuse tõenäosus (tavaliselt 0,05),
Siis on muutujate vahel mingi statistiliselt oluline seos

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	16,293[a]	12	,178
Likelihood Ratio	17,012	12	,149
Linear-by-Linear Association	,084	1	,772
N of Valid Cases	582		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 21,73.

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Risttabeleid

… saab teha ka erinevates lõigetes (*layer variable*)

Store * Service satisfaction * Contact with employee Crosstabulation

Count

Contact with employee			Service satisfaction					Total
			Strongly Negative	Somewhat Negative	Neutral	Somewhat Positive	Strongly Positive	
No	Store	Store 1	16	9	18	17	19	79
		Store 2	2	15	16	13	12	58
		Store 3	9	14	23	22	14	82
		Store 4	17	14	19	10	10	70
	Total		44	52	76	62	55	289
Yes	Store	Store 1	9	11	20	13	14	67
		Store 2	24	15	18	14	7	78
		Store 3	6	6	18	11	15	56
		Store 4	10	21	25	12	24	92
	Total		49	53	81	50	60	293

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Risttabeleid

… sel juhul on hii-ruut testi tulemused sootuks teised

Chi-Square Tests

Contact with employee		Value	df	Asymp. Sig. (2-sided)
No	Pearson Chi-Square	20,898[a]	12	,052
	Likelihood Ratio	22,937	12	,028
	Linear-by-Linear Association	3,514	1	,061
	N of Valid Cases	289		
Yes	Pearson Chi-Square	25,726[b]	12	,012
	Likelihood Ratio	25,777	12	,012
	Linear-by-Linear Association	1,993	1	,158
	N of Valid Cases	293		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 8,83.

b. 0 cells (,0%) have expected count less than 5. The minimum expected count is 9,37.

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Nominaaltunnuste vahelised seosed

- Crameri V – vt

http://planetmath.org/encyclopedia/CramersV.html

- Fii – vt

http://changingminds.org/explanations/research/analysis/phi.htm

Directional Measures

Contact with employee				Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
No	Nominal by Nominal	Lambda	Symmetric	,036	,030	1,178	,239
			Store Dependent	,068	,044	1,498	,134
			Service satisfaction Dependent	,005	,028	,164	,869
		Goodman and Kruskal tau	Store Dependent	,023	,009		,067[c]
			Service satisfaction Dependent	,016	,006		,112[c]
		Uncertainty Coefficient	Symmetric	,027	,010	2,604	,028[d]
			Store Dependent	,029	,011	2,604	,028[d]
			Service satisfaction Dependent	,025	,010	2,604	,028[d]

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on chi-square approximation

d. Likelihood ratio chi-square probability.

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Järjestustunnuste vahelised seosed

- *2) Muutujate vaheline seos on võrdlemisi nõrk*

1) Muutujate vahel on statistiliselt oluline seos

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Kendall's tau-b	,107	,033	3,267	,001
	Kendall's tau-c	,102	,031	3,267	,001
	Gamma	,140	,043	3,267	,001
N of Valid Cases		582			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Directional Measures

			Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Somers' d	Symmetric	,107	,033	3,267	,001
		Shopping frequency Dependent	,104	,032	3,267	,001
		Overall satisfaction Dependent	,110	,034	3,267	,001

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Risttabelid – sündmuse toimumise suhteline risk (relative risk of an event)

- Suhteline risk on sündmuste toimumise tõenäosuste suhe

tõenäosus, et ajalehe tellija ei vasta/ tõenäosus, et ajalehe mittetellija ei vasta

tõenäosus, et ajalehe tellija vastab/ tõenäosus, et ajalehe mittetellija vastab

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for Newspaper subscription (Yes / No)	1,774	1,511	2,082
For cohort Response = Yes	1,668	1,445	1,924
For cohort Response = No	,940	,924	,957
N of Valid Cases	6400		

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Risttabelid – sündmuse toimumise suhteline risk (relative risk of an event)

- Sündmuse šansside suhe (*odds ratio*) – tõenäosus, et sündmus toimub/tõenäosus, et sündmus ei toimu

tõenäosus, et ajalehe tellija vastab =
= 13.7% / 86.3% = 0.158

tõenäosus, et ajalehe mittetellija vastab =
= 8.2% / 91.8% = 0.089

***šansside suhe**=*
= 0.158% / 0.089% = 1.775 = 1.668 / 0.94

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for Newspaper subscription (Yes / No)	1,774	1,511	2,082
For cohort Response = Yes	1,668	1,445	1,924
For cohort Response = No	,940	,924	,957
N of Valid Cases	6400		

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Risttabelid – sündmuse toimumise suhteline risk (relative risk of an event)

- šansside suhet saab kasutada suhtelise riski lähendina juhul, kui on täidetud mõlemad alljärgnevad tingimused:

1) sündmuse toimumise tõenäosus on madal (<0.1)
2) tegu on juhtumiuuringuga (*case study*)

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Risttabelid – sündmuse toimumise suhteline risk (relative risk of an event)

- šansside suhte homogeensuse test – kontrollib, kas eri gruppides ilmnevad erinevused on statistiliselt oluliselt erinevad 1-st
- Breslow-Day ja Tarone'i statistikud testivad šansside suhte homogeensust üle kontrollmuutja (*layer variable*) gruppide

Sig. > 0.05, seega šansside suhted on homogeensed

Tests of Homogeneity of the Odds Ratio

	Chi-Squared	df	Asymp. Sig. (2-sided)
Breslow-Day	4,030	3	,258
Tarone's	4,026	3	,259

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Risttabelid – sündmuse toimumise suhteline risk (relative risk of an event)

- Cochrani ja Mantel Haenszeli statistikud testivad, kas risttabeli rea- ja veerumuutujad on sõltumatud, kui arvesse on võetud kontrollmuutuja mõju

Sig. < 0.05, seega seos on oluline

Tests of Conditional Independence

	Chi-Squared	df	Asymp. Sig. (2-sided)
Cochran's	68,916	1	,000
Mantel-Haenszel	68,178	1	,000

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Kirjeldavad statistikud (Descriptives)

- Võimaldab
- võrrelda ligikaudu normaaljaotusega jaotunud muutujaid
- leida muutujate lõikes ebaharilikke objekte

Analyze → Descriptive Statistics → Descriptives

Kasutame andmefaili "**telco.sav**"

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Long distance last month	1000	,90	99,95	11,7231	10,36349
Toll free last month	1000	,00	173,00	13,2740	16,90212
Equipment last month	1000	,00	77,70	14,2198	19,06854
Calling card last month	1000	,00	109,25	13,7810	14,08450
Wireless last month	1000	,00	111,95	11,5839	19,71943
Valid N (listwise)	1000				

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Kirjeldavad statistikud (Descriptives)

- Pärast nulliliste väärtuste eemaldamist:

Kõige kasumlikumad

Descriptive Statistics

	N	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Long distance last month	1000	11,7231	10,36349	2,966	,077	14,052	,155
Toll free last month	475	27,9453	13,82910	3,465	,112	26,735	,224
Equipment last month	386	36,8389	10,39568	,756	,124	,641	,248
Calling card last month	678	20,3260	12,62916	2,150	,094	7,572	,187
Wireless last month	296	39,1348	15,32916	1,359	,142	3,079	,282
Valid N (listwise)	131						

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Kirjeldavad statistikud (Descriptives)

- Ebaharilike objektide leidmine – z-skoor

NB! Z-skoori kasutamise eelduseks on, et muutuja peab olema ligikaudselt normaaljaotusega!

Üheks võimaluseks on kasutada logaritmilist transformatsiooni

Descriptive Statistics

	N	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Log-long distance	1000	2,1821	,73455	,166	,077	-,001	,155
Log-toll free	475	3,2397	,41381	,304	,112	1,107	,224
Log-equipment	386	3,5681	,27756	,037	,124	-,344	,248
Log-calling card	678	2,8542	,55729	,081	,094	,109	,187
Log-wireless	296	3,5983	,36729	,200	,142	-,168	,282
Log-income	1000	3,9572	,80375	,701	,077	,669	,155
Valid N (listwise)	131						

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Kirjeldavad statistikud (Descriptives)

- Graafiliseks illustreerimiseks võib kasutada karpdiagrammi (*box-plot*)

Descriptive Statistics

	N	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Log-long distance	1000	2,1821	,73455	,166	,077	-,001	,155
Log-toll free	475	3,2397	,41381	,304	,112	1,107	,224
Log-equipment	386	3,5681	,27756	,037	,124	-,344	,248
Log-calling card	678	2,8542	,55729	,081	,094	,109	,187
Log-wireless	296	3,5983	,36729	,200	,142	-,168	,282
Log-income	1000	3,9572	,80375	,701	,077	,669	,155
Valid N (listwise)	131						

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Keskmised (Means)

- Sobib nii arvuliste andmete kirjeldamiseks kui analüüsimiseks

Analyze → Compare Means→ Means

Kasutame andmefaili "**hourlywagedata.sav**"

Report

Hourly Salary

Years Experience	Mean	N	Std. Deviation
5 or less	18,0416	221	3,86667
6-10	18,9169	460	3,77816
11-15	19,6616	752	3,90528
16-20	20,2876	729	3,82786
21-35	21,2594	539	4,08669
36 or more	21,6342	210	3,61826
Total	20,0159	2911	4,00309

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Keskmised (Means)

- Ühefaktoriline dispersioonanalüüs (*one-way ANOVA*)

Analyze → Compare Means→ Means

Kasutame andmefaili "**smokers.sav**"

Report

Age when first smoked a cigarette

# Cigarettes smoked per day past 30 days	Mean	N	Std. Deviation
1 to 5 cigarettes each day	15,81	1119	4,452
6 to 15 cigarettes (about 1/2 pack) each	15,89	1594	4,820
16 to 25 cigarettes (about 1 pack) each	15,63	1604	5,450
26 to 35 cigarettes (about 1 1/2 pk) eac	14,18	622	4,066
35 or more cigarettes (about 2 packs) ea	14,45	461	4,376
Total	15,48	5400	4,866

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Keskmised (Means)

- Ühefaktoriline dispersioonanalüüs

Vanuse ja suitsetamise taseme vahel on lineaarne seos

Vanuse ja suitsetamise taseme vahel on mittelineaarne seos

ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
Age when first smoked a cigarette * # Cigarettes smoked per day past 30 days	Between Groups	(Combined)	1974,095	4	493,524	21,158	,000
		Linearity	1321,500	1	1321,500	56,655	,000
		Deviation from Linearity	652,595	3	217,532	9,326	,000
	Within Groups		125841,118	5395	23,326		
	Total		127815,213	5399			

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Keskmised (Means)

- …samas on seos nõrk

Measures of Association

	R	R Squared	Eta	Eta Squared
Age when first smoked a cigarette * # Cigarettes smoked per day past 30 days	-,102	,010	,124	,015

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

T-testid

- Ühe valimi t-test (*One-Sample T-Test*)
- Paarisvalimi t-test (*Paired-Samples T-Test*)
- Sõltumatute valimite t-test (*Independent Samples T-Test*)

Ühe valimi t-test

- Kontrollib, kas erinevus valimi keskmise ja etteantud suuruse vahel on statistiliselt oluline
- SPSSis saab ette anda usaldusnivoo ning hõlpsalt kuvada iga testitava muutuja kirjeldavad statistikud
- Eeldus: muutuja on ligikaudu normaaljaotusega

Analyze → Compare Means → One-Sample T-Test

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Ühe valimi t-test

One-Sample Test

# Cigarettes smoked per day past 30 days		Test Value = 14					
		t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
						Lower	Upper
1 to 5 cigarettes each day	Age when first smoked a cigarette	13,595	1118	,000	1,810	1,55	2,07
6 to 15 cigarettes (about 1/2 pack) each	Age when first smoked a cigarette	15,692	1593	,000	1,894	1,66	2,13
16 to 25 cigarettes (about 1 pack) each	Age when first smoked a cigarette	11,972	1603	,000	1,629	1,36	1,90
26 to 35 cigarettes (about 1 1/2 pk) eac	Age when first smoked a cigarette	1,081	621	,280	,176	-,14	,50
35 or more cigarettes (about 2 packs) ea	Age when first smoked a cigarette	2,228	460	,026	,454	,05	,85

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Paarisvalimi t-test

- Kasutatakse, et testida hüpoteesi, et kahe muutuja vahel ei ole statistiliselt olulisi erinevusi (nn ***pre-post*** analüüs)

SPSSis on lisaks veel võimalik kuvada

- Iga testitava muutuja korral kirjeldavad statistikud
- Iga paari vahelise Pearsoni korrelatsioonikoefitsiendi

Analyze → Compare Means → Paired-Samples T-Test

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Paarisvalimi t-test

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Triglyceride	138,44	16	29,040	7,260
	Final triglyceride	124,38	16	29,412	7,353
Pair 2	Weight	198,38	16	33,472	8,368
	Final weight	190,31	16	33,508	8,377

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Triglyceride & Final triglyceride	16	-,286	,283
Pair 2	Weight & Final weight	16	,996	,000

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	Triglyceride - Final triglyceride	14,063	46,875	11,719	-10,915	39,040	1,200	15	,249
Pair 2	Weight - Final weight	8,063	2,886	,722	6,525	9,600	11,175	15	,000

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Sõltumatute valimite t-test

kasutatakse kahe valimikeskmise erinevuste statistilise olulisuse kontrollimiseks

Analyze → Compare Means → Independent-Samples T-Test

Group Statistics

	Type of mail insert received	N	Mean	Std. Deviation	Std. Error Mean
$ spent during promotional period	Standard	250	1566,3890	346,67305	21,92553
	New Promotion	250	1637,5000	356,70317	22,55989

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Sõltumatute valimite t-test

kasutatakse kahe valimikeskmise erinevuste statistilise olulisuse kontrollimiseks

Kahel grupil on muutuja varieeruvus sama

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
$ spent during promotional period	Equal variances assumed	1,190	,276	-2,260	498	,024	-71,11095	31,45914	-132,91995	-9,30196
	Equal variances not assumed			-2,260	497,595	,024	-71,11095	31,45914	-132,92007	-9,30183

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Ühefaktoriline dispersioonanalüüs
(one-way ANOVA)

Kasutatakse kontrollimaks hüpoteesi, et kahes või enamas grupis on muutuja keskväärtused sarnased

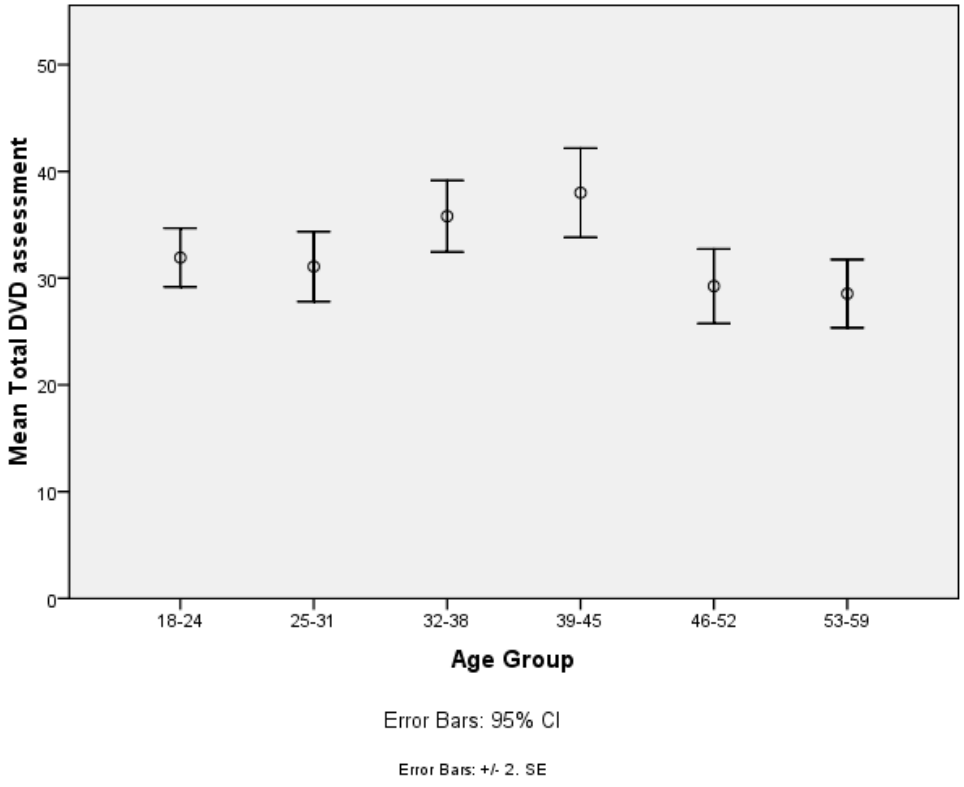
Analyze → Compare Means → One-Way ANOVA

Kasutame andmefaili "**dvdplayer.sav**"

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Ühefaktoriline dispersioonanalüüs)

Esmalt analüüsime graafiliselt, kas muutuja varieeruvus on gruppides sarnane



Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Ühefaktoriline dispersioonanalüüs

Descriptives

Total DVD assessment

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
18-24	13	31,92	4,958	1,375	28,93	34,92	26	39
25-31	12	31,08	5,664	1,635	27,48	34,68	24	40
32-38	10	35,80	5,308	1,679	32,00	39,60	30	44
39-45	10	38,00	6,600	2,087	33,28	42,72	28	47
46-52	12	29,25	6,047	1,746	25,41	33,09	20	41
53-59	11	28,55	5,298	1,598	24,99	32,10	18	37
Total	68	32,22	6,359	,771	30,68	33,76	18	47

Test of Homogeneity of Variances

Total DVD assessment

Levene Statistic	df1	df2	Sig.
,574	5	62	,720

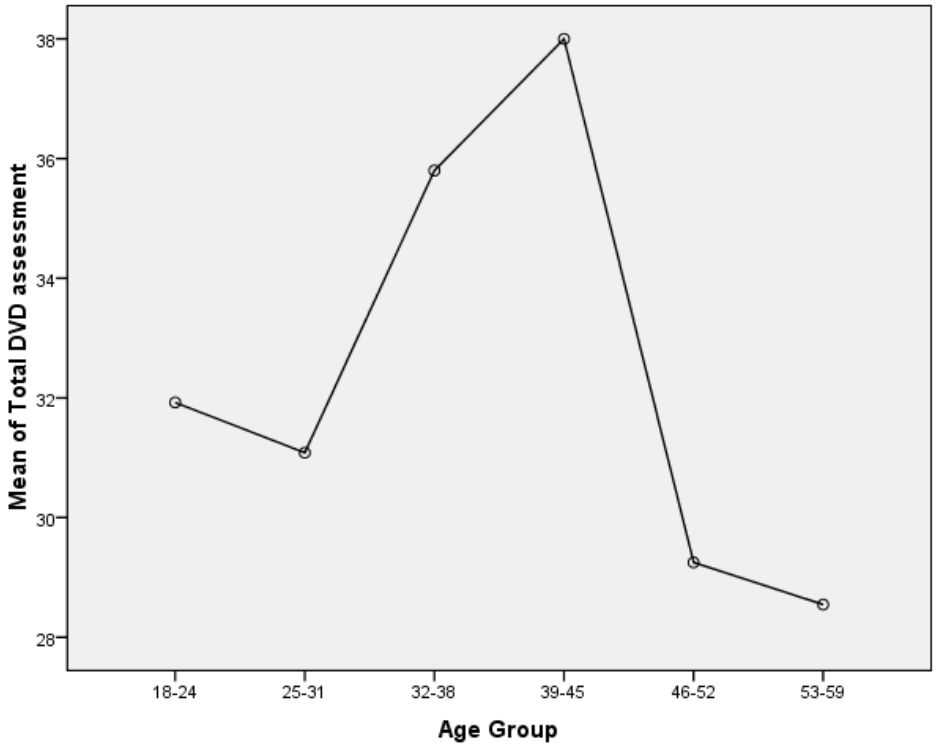
Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Ühefaktoriline dispersioonanalüüs

ANOVA

Total DVD assessment

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	733,274	5	146,655	4,601	,001
Within Groups	1976,417	62	31,878		
Total	2709,691	67			



Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Ühefaktoriline dispersioonanalüüs

Etteantud gruppide võrdlemine: **contrasts**

Contrast Coefficients

Contrast	Age Group					
	18-24	25-31	32-38	39-45	46-52	53-59
1	0	0	-1	1	0	0
2	,5	,5	0	0	-,5	-,5

Contrast Tests

		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
Total DVD assessment	Assume equal variances	1	2,20	2,525	,871	62	,387
		2	2,61	1,633	1,596	62	,116
	Does not assume equal variances	1	2,20	2,678	,821	17,209	,423
		2	2,61	1,594	1,635	42,280	,110

Iga grupi võrdlus iga grupiga: **post hoc**

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Korrelatsioonanalüüs

Võimaldab analüüsida muutujate vahelise seose olemasolu, tugevust ja suunda

Pidevate muutujate korral kasutatakse enamasti Pearsoni korrelatsioonikoefitsienti, mis mõõdab **lineaarse** seose tugevust

Kasutame andmefaili **“car_sales.sav”**

Analyze → Correlate → Bivariate

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Korrelatsioonanalüüs

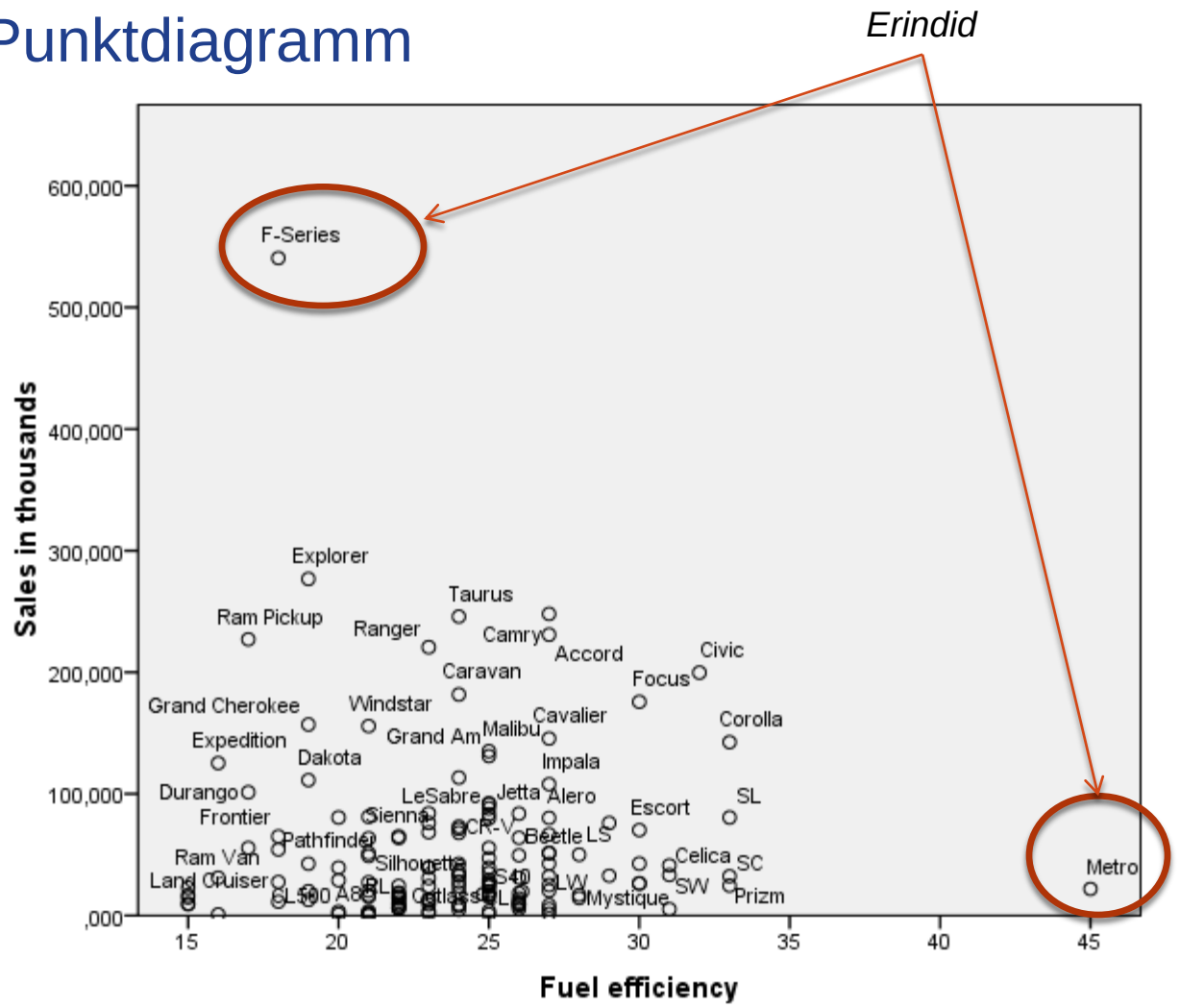
Correlations

		Fuel efficiency	Sales in thousands
Fuel efficiency	Pearson Correlation	1	-,017
	Sig. (2-tailed)		,837
	N	154	154
Sales in thousands	Pearson Correlation	-,017	1
	Sig. (2-tailed)	,837	
	N	154	157

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Korrelatsioonanalüüs

Punktdiagramm

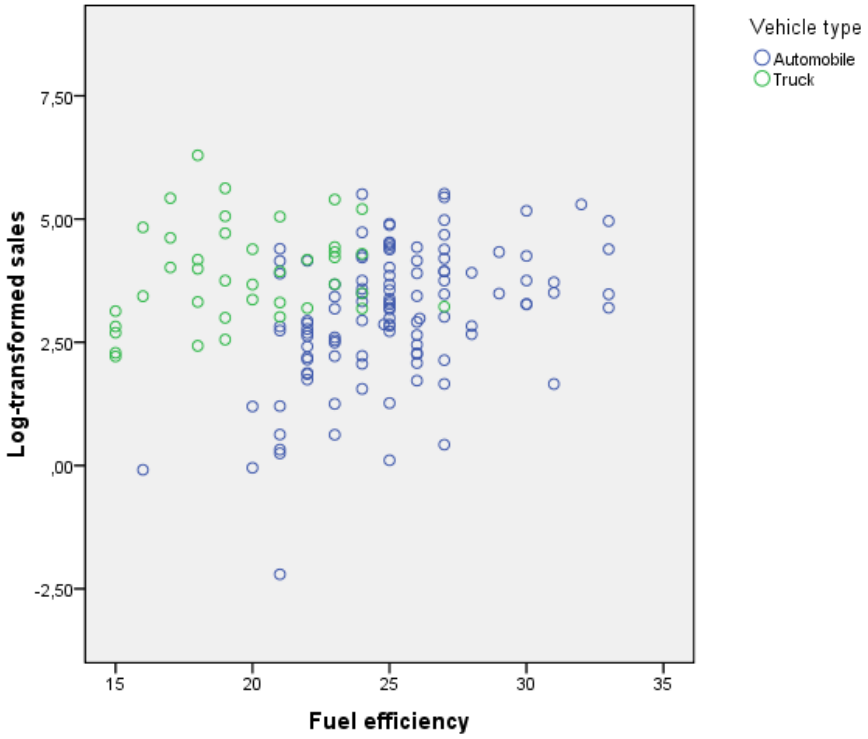


Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Korrelatsioonanalüüs

Correlations

		Fuel efficiency	Log-transformed sales
Fuel efficiency	Pearson Correlation	1	,136
	Sig. (2-tailed)		,093
	N	153	153
Log-transformed sales	Pearson Correlation	,136	1
	Sig. (2-tailed)	,093	
	N	153	156



Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Korrelatsioonanalüüs

Correlations

Vehicle type			Fuel efficiency	Log-transformed sales
Automobile	Fuel efficiency	Pearson Correlation	1	,451**
		Sig. (2-tailed)		,000
		N	113	113
	Log-transformed sales	Pearson Correlation	,451**	1
		Sig. (2-tailed)	,000	
		N	113	115
Truck	Fuel efficiency	Pearson Correlation	1	,203
		Sig. (2-tailed)		,210
		N	40	40
	Log-transformed sales	Pearson Correlation	,203	1
		Sig. (2-tailed)	,210	
		N	40	41

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

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Korrelatsioonanalüüs

Correlations

	Vehicle type			Fuel efficiency	Log-transformed sales	Sales in thousands
Spearman's rho	Automobile	Fuel efficiency	Correlation Coefficient	1,000	,425**	,425**
			Sig. (2-tailed)	.	,000	,000
			N	113	113	113
		Log-transformed sales	Correlation Coefficient	,425**	1,000	1,000**
			Sig. (2-tailed)	,000	.	.
			N	113	115	115
		Sales in thousands	Correlation Coefficient	,425**	1,000**	1,000
			Sig. (2-tailed)	,000	.	.
			N	113	115	115
	Truck	Fuel efficiency	Correlation Coefficient	1,000	,237	,237
			Sig. (2-tailed)	.	,141	,141
			N	40	40	40
		Log-transformed sales	Correlation Coefficient	,237	1,000	1,000**
			Sig. (2-tailed)	,141	.	.
			N	40	41	41
		Sales in thousands	Correlation Coefficient	,237	1,000**	1,000
			Sig. (2-tailed)	,141	.	.
			N	40	41	41

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

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Osakorrelatsioonanalüüs

Kirjeldab kahe muutuja vahelist korrelatsiooni, võttes arvesse ühe või enama muutuja mõju

Kõik analüüsis kasutatavad muutujad peaksid olema **arvulised** *(scale)*

Kasutame andmefaili **“health_funding.sav”**

Analyze → Correlate → Partial

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Osakorrelatsioonanalüüs

Correlations

Control Variables			Health care funding (amount per 100)	Reported diseases (rate per 10,000)	Visits to health care providers (rate per 10,000)
-none-[a]	Health care funding (amount per 100)	Correlation	1,000	,737	,964
		Significance (2-tailed)	.	,000	,000
		df	0	48	48
	Reported diseases (rate per 10,000)	Correlation	,737	1,000	,762
		Significance (2-tailed)	,000	.	,000
		df	48	0	48
	Visits to health care providers (rate per 10,000)	Correlation	,964	,762	1,000
		Significance (2-tailed)	,000	,000	.
		df	48	48	0
Visits to health care providers (rate per 10,000)	Health care funding (amount per 100)	Correlation	1,000	,013	
		Significance (2-tailed)	.	,928	
		df	0	47	
	Reported diseases (rate per 10,000)	Correlation	,013	1,000	
		Significance (2-tailed)	,928	.	
		df	47	0	

a. Cells contain zero-order (Pearson) correlations.

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Lineaarne regressioonanalüüs

Kasutatakse, analüüsimaks sõltuva muutuja lineaarset seost sõltumatu(te) muutuja(te)ga.

Eeldused:

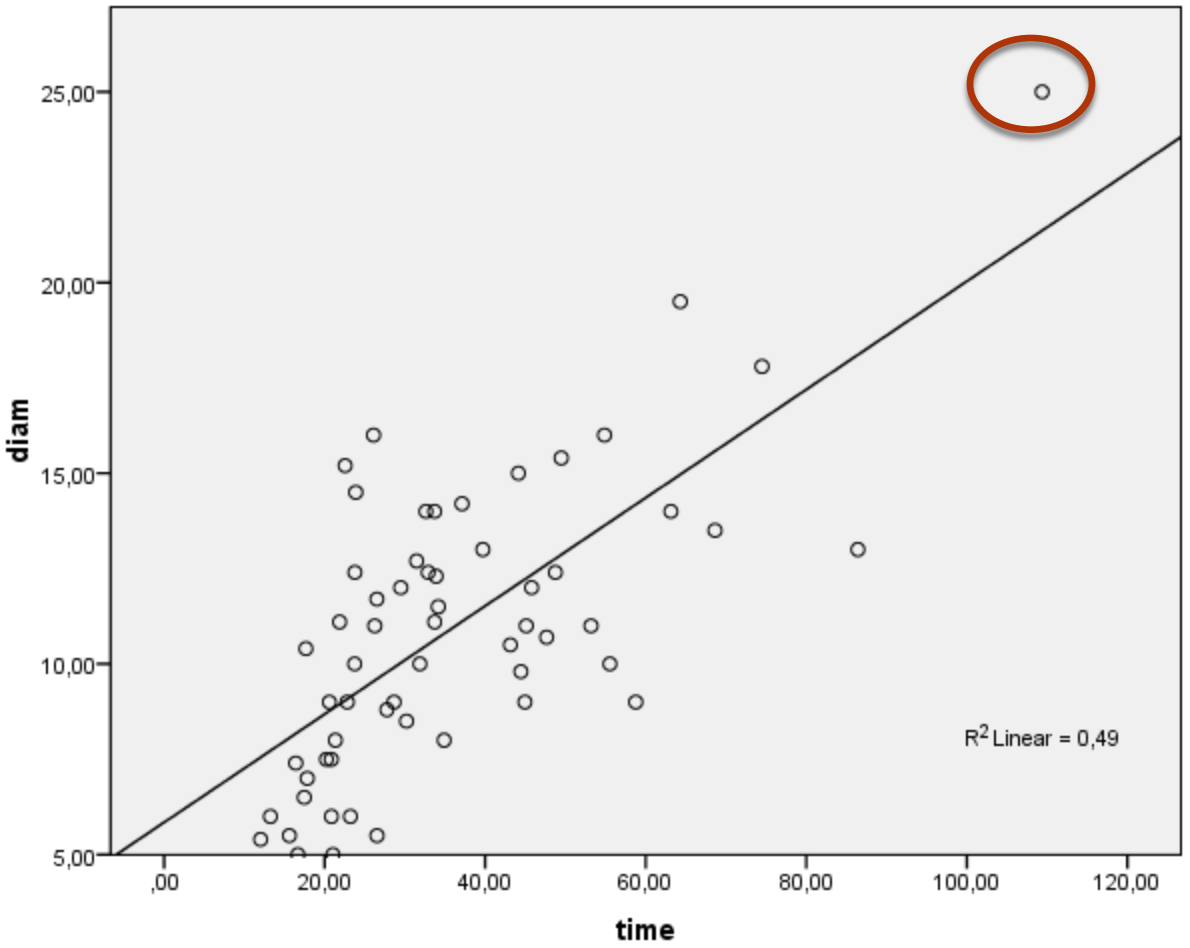
- vealiige on normaaljaotusega, keskväärtusega 0
- vealiikme varieeruvus on konstantne ja ei sõltu mudeli sõltumatutest muutujatest

Kasutame andmefaili **“polishing.sav”**

Analyze → Regression → Linear

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Lineaarne regressioonanalüüs



Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Lineaarne regressioonanalüüs

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-1,955	5,402		-,362	,719
	diam	3,457	,467	,700	7,407	,000

a. Dependent Variable: time

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	10287,173	1	10287,173	54,865	,000[a]
	Residual	10687,511	57	187,500		
	Total	20974,684	58			

a. Predictors: (Constant), diam

b. Dependent Variable: time

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Lineaarne regressioonanalüüs

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,700[a]	,490	,482	13,69307

a. Predictors: (Constant), diam

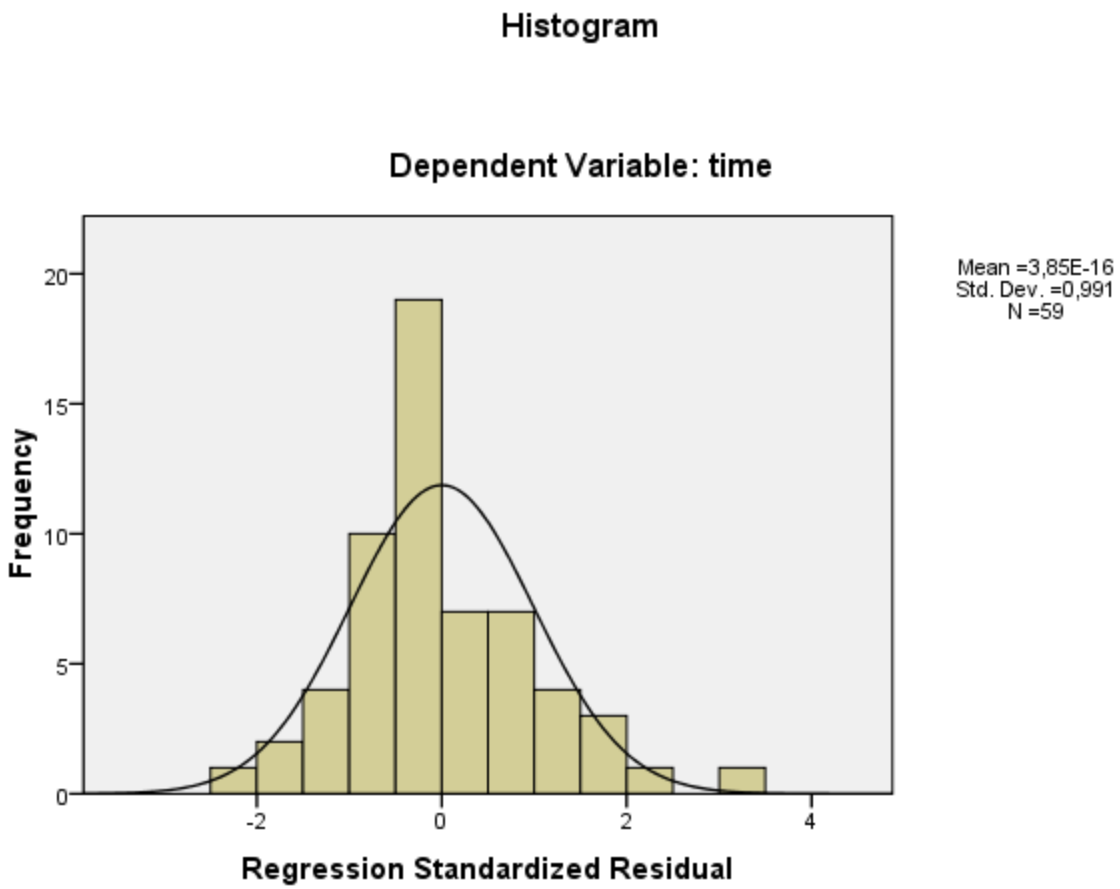
b. Dependent Variable: time

Descriptive Statistics

	Mean	Std. Deviation	N
time	35,8171	19,01664	59
diam	10,9271	3,85276	59

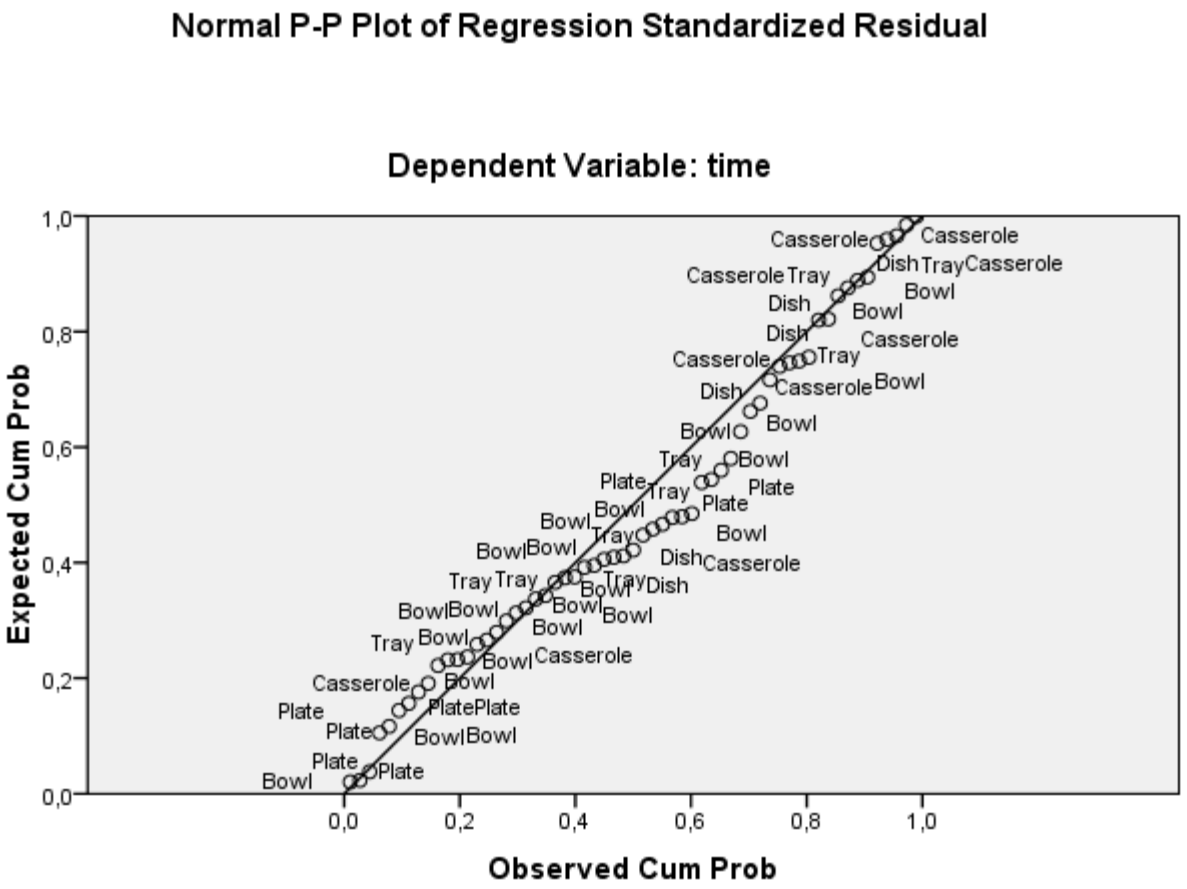
Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Lineaarne regressioonanalüüs



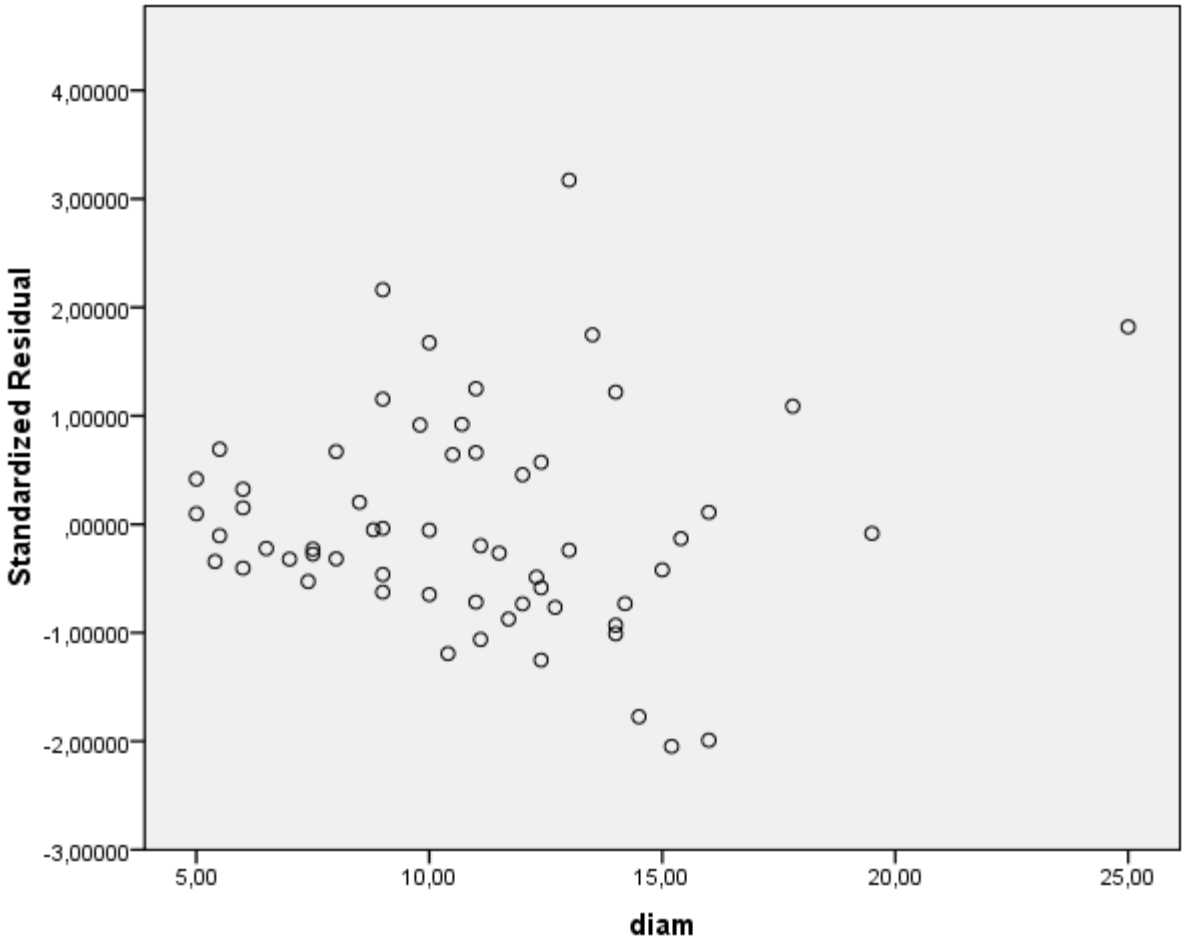
Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Lineaarne regressioonanalüüs



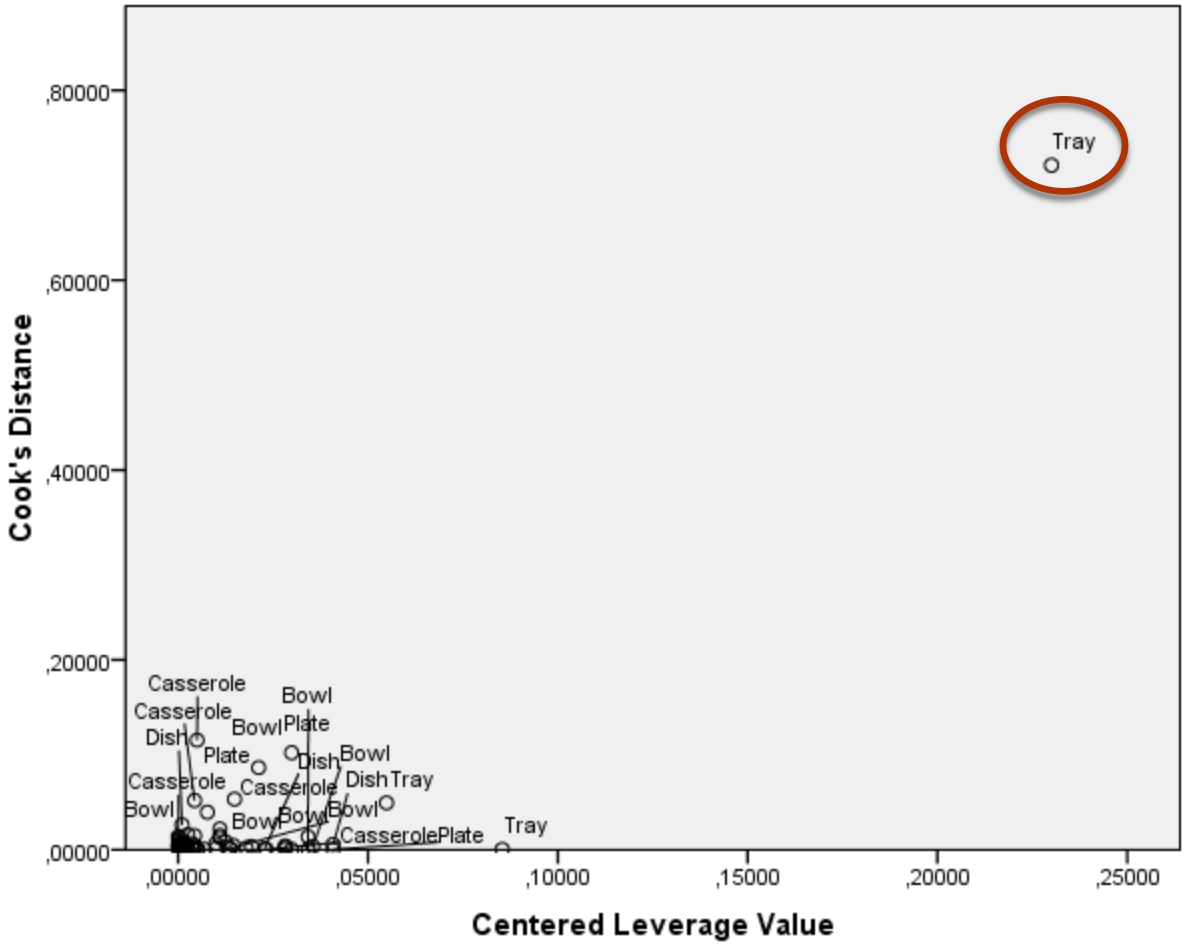
Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Lineaarne regressioonanalüüs



Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Lineaarne regressioonanalüüs



Tänan tähelepanu eest!

TARTU ÜLIKOOL
1632

STATISTILISE ANALÜÜSI TEOSTAMINE EXCELI JA SPSSI ABIL

Kerly Krillo

Tartu Ülikool, sotsiaalteaduslike rakendusuuringute keskus

Tööturu ja tööpoliitika programmi juht

kerly.krillo@ut.ee

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Tänase praktikumi teemad

Kodutööde tagasiside

Regressioonanalüüs

Klasteranalüüs

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Loeng/praktikumi eesmärk

Pärast selle praktikumi läbimist

- oskab tudeng iseseisvalt teostada regressioon- ja klasteranalüüsi SPSSi abil ja saadud tulemusi sisukalt tõlgendada

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

I Kodutööde tagasiside

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Töö suurte andmemassiividega

1. Millises summas on müüdud harilikku piima? Kasuta funktsiooni SUMIF ja sea tingimus segmendi nime järgi. Kodutöösse esita vastus ja lahenduskäik funktsioonina.

   Siin probleeme ei esinenud, kõik vastasid õigesti.

   Vastus=SUMIF(D:D;"Harilik piim";H:H) ning 133 316

2. Millises summas on kaupluses D müüdud õunu? Kasuta Filtrit. Kodutöösse esita vastus.

   Esmalt tuli kasutada kahte filtrit: filtreerida välja kauplus D (kaupluse veerg) ja tekstifiltrit kasutades filtreerida toote nimes sõna „õun“ (c*ontains* „õun“). Siis tähelepanelikumad ehk märkasid, et nimekirja sattus ka teisi tooteid, mille nimes õun (ÕUNA-MANDLIKOOK, NÕO LÕUNA SARDELL jne) ja välistasid need tooted (valisid õige segmendi – Pirnid, õunad või välistasid ebasobivad tooted).

   Vastus: 1951

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Töö suurte andmemassiividega

3. Loo uus muutuja kaalukaupade kohta (=tooted, mis on kahe- kuni kuuekohalise EANiga), nimeta „Kauba tüüp“ (tunnused kaalukaup ja tavakaup). Kasuta funktsiooni IF. Kodutöösse esita lahenduskäik funktsioonina.

Siin võis kasutada ja ka kasutati erinevaid teid, nt

=IF(E2<1000000;"kaalukaup";"tavakaup")

=IF(E2>999999;"tavakaup";"kaalukaup")

=IF(LEN(E2)<7;"kaalukaup";"tavakaup")

Kes ei tea, siis LEN() annab teksti pikkuse

=IF(E2<=703002;"kaalukaup";"tavakaup")

Järjestati EANi järgi ja leiti viimane kuuekohaline kood

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Töö suurte andmemassiividega

4. Kasutades Subtotalit, grupeeri andmed muutuja „Kauba tüüp“ lõikes, summeerides müügikoguse ja müügisumma. Kodutöösse esita vaade, mis näitab kaalukaupade/tavakaupade summeeritud andmeid (vaade nr 2).

Müügikogus	Müük summas	kauba tüüp
20480,31	465466,67	kaalukaup Total
117503,47	1435328,33	tavakaup Total
137983,78	1900795,00	Grand Total

Paar inimest ütles, et neil Subtotal uue tunnusega ei tööta. Võib olla, et jäeti valem välja kopeerimata (IF), ei oska täpsemalt kommenteerida. Siiski oleks soovinud sellistest probleemidest kuulda enne kodutöö esitamist.

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Töö suurte andmemassiividega

5. Kasuta PivotTabelit ja esita müügisumma kaupluste ja muutuja „Kauba tüüp“ lõikes. Müügisumma esita protsendina kogukäibest, st mitu % moodustavad kogukäibest kaalukaubad ja mitu % tavakaubad kaupluste lõikes.

Valida tuli „% of row“

Sum of Müük summas	kauba tüüp		
Kaupluse nimi	kaalukaup	tavakaup	Grand Total
Kauplus A	27,82%	72,18%	100,00%
Kauplus B	24,22%	75,78%	100,00%
Kauplus C	23,78%	76,22%	100,00%
Kauplus D	25,20%	74,80%	100,00%
Kauplus E	24,26%	75,74%	100,00%
Kauplus F	18,84%	81,16%	100,00%
Grand Total	24,49%	75,51%	100,00%

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Töö suurte andmemassiividega

6. BOONUSPUNKTIÜLESANNE. Loo uus muutuja „Tooterühm“, mille loomiseks vajaliku vastavustabeli leiad lehelt Vastavustabel2. Kasuta funktsiooni VLOOKUP. Kasuta PivotTabelit ja esita müügisumma kaupluste ja tooterühmade lõikes. Kodutöösse esita sama tabel vaid tooterühmade Jäätis, Jogurt ja Juustud kohta.

Kõik need, kes selle ülesande ette võtsid, tegid õigesti!

Sum of Müük summas	tooterühm			
Kaupluse nimi	Jogurt	Juustud	Jäätis	Grand Total
Kauplus A	11841,68	30935,42	6568,11	49345,21
Kauplus B	13340,47	30048,32	6003,89	49392,69
Kauplus C	5241,84	13655,34	3007,71	21904,89
Kauplus D	2143,17	6683,84	5116,34	13943,35
Kauplus E	4371,04	10242,41	4999,26	19612,70
Kauplus F	7697,93	16641,33	7004,98	31344,24
Grand Total	44636,13	108206,66	32700,29	185543,08

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

T-testid

- **Ühe valimi t-test** (*One-Sample T-Test*) - kontrollib, kas erinevus valimi keskmise ja etteantud suuruse vahel on statistiliselt oluline

- **Paarisvalimi t-test** (*Paired-Samples T-Test*) - kasutatakse, et testida hüpoteesi, et kahe muutuja vahel ei ole statistiliselt olulisi erinevusi (nn ***pre-post*** analüüs)

- **Sõltumatute valimite t-test** (*Independent Samples T-Test*) - kasutatakse kahe valimi keskmiste erinevuste statistilise olulisuse kontrollimiseks

Analyze → Compare Means

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Ülesanne 3

Kontrollige, kas valgenahaliste keskmine kooliskäidud aastate arv on 12 (ehk teisisõnu, kas „keskmisel“ valgenahalisel ameeriklasel on keskharidus).

Selleks tuleb esmalt analüüsi kaasata vaid need vaatlused, kus muutuja „race“ väärtuseks on 1 – „white“. Seda saate teha järgmiselt:

1. „Data“ → „Select Cases“ → “If condition is satisfied“ → race=1 või
2. „Data“ → „Split File“→ “Compare Groups“

Seejärel **„Analyze“ → „Compare Means“ → „One-Sample T-Test“**

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Ülesanne 3

One-Sample Statistics

Race of Respondent		N	Mean	Std. Deviation	Std. Error Mean
White	Highest Year of School Completed	1262	13,06	2,955	,083
Black	Highest Year of School Completed	199	11,89	2,677	,190
Other	Highest Year of School Completed	49	12,47	4,001	,572

NB!!!

One-Sample Test

		Test Value = 12					
						95% Confidence Interval of the Difference	
Race of Respondent		t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
White	Highest Year of School Completed	12,699	1261	,000	1,056	,89	1,22
Black	Highest Year of School Completed	-,556	198	,579	-,106	-,48	,27
Other	Highest Year of School Completed	,821	48	,416	,469	-,68	1,62

**NB! KUI OLETE VAJALIKUD ANDMETABELID GENEREERINUD,
ÄRGE UNUSTAGE „SELECT CASES“/”SPLIT FILE” MAHA VÕTTA!!!**

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Ülesanne 4

Analüüsige, kas erinevused valge- ja mustanahaliste keskmistes kooliskäidud aastate arvus on statistiliselt olulised. Selleks

- **püstitage null- ja alternatiivne hüpotees (ehk H_0 ja H_1);**

  H_0 : valge- ja mustanahaliste keskmine koolikäidud aastate arv on sama

  H_1 : H_0 ei kehti

- **Tehke sõltumatute valimite t-test**

  „Analyze“ → „Compare Means“ → “Independent Samples t-test“ (võrreldavad grupid on „1“ – valgenahalised ja „2“ – mustanahalised).

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Ülesanne 4

Mida järeldate Levene'i testi põhjal (sh püstitage ka selle testiga kontrollitavad hüpoteesid)?

H_0 : valge- ja mustanahaliste koolikäidud aastate arvu varieeruvus on sama

H_1 : H_0 ei kehti

Mida järeldate t-testi tulemuste põhjal?

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Highest Year of School Completed	Equal variances assumed	9,153	,003	5,219	1459	,000	1,162	,223	,725	1,598
	Equal variances not assumed			5,607	279,767	,000	1,162	,207	,754	1,570

H_1 H_1

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Ülesanne 5

Analüüsige, kes eri ametipositsioonidel töötajatel („occcat80“) on keskmine vanus „age“ ja kooliskäidud aastate arv erinev. Selleks andke esmalt lühiülevaade, kasutades kirjeldavaid statistikuid. Seejärel teostage dispersioonanalüüs, sh

- **püstitage uuritav hüpoteeside paar**

  H_0 : eri ametipositsioonil olevate töötajate keskmine vanus on sama

  H_1 : H_0 ei kehti

  H_0 : eri ametipositsioonil olevate töötajate keskmine kooliskäidud aastate arv on sama

  H_1 : H_0 ei kehti

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Ülesanne 5

Tehke dispersioonanalüüs, „Analyze“ → „Compare Means“ → “One-Way ANOVA“ →

Dependent List: „age“ ja „educ“

NB! Dispersioonanalüüsis on sõltuv tunnus alati pidev!

Factor: „occcat80“.

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
Age of Respondent	Between Groups	1130,038	5	226,008	,736	,597
	Within Groups	433018,087	1410	307,105		
	Total	434148,124	1415			
Highest Year of School Completed	Between Groups	3894,709	5	778,942	129,110	,000
	Within Groups	8500,737	1409	6,033		
	Total	12395,446	1414			

H_0

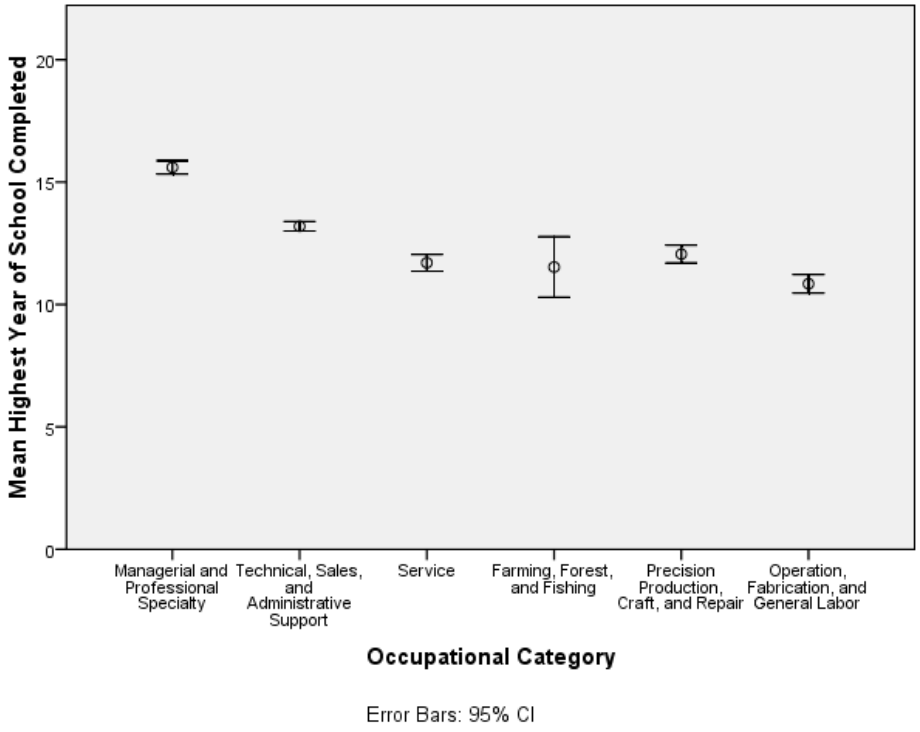
H_1

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Ülesanne 5

Analüüsige, millistes ametipositsioonide kategooriates on erinevused kooliskäidud aastate arvus sarnased, millistes erinevad. Selleks tehke esmalt nö endale pildi saamiseks joonis, nt

„Graphs“ → „Chart Builder“ → „Bar“ → „Simple Error Bar“



Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Ülesanne 5

Kasutades “One-Way ANOVA“ võimalust „Contrasts“, analüüsige, kas:

- erinevused juhtide (muutuja „occcat80“ kategooria „1“) ja lihttööliste (kategooria „6“) keskmistes kooliskäidud aastate arvus on statistiliselt olulised
- erinevused teenindustöötajate (muutuja „occcat80“ kategooria „3“) ja põllumajandustöötajate (kategooria „4“) keskmistes kooliskäidud aastate arvus on statistiliselt olulised.

Contrast Coefficients

Contrast	Occupational Category					
	Managerial and Professional Specialty	Technical, Sales, and Administrative Support	Service	Farming, Forest, and Fishing	Precision Production, Craft, and Repair	Operation, Fabrication, and General Labor
1	1	0	0	0	0	-1
2	0	0	1	-1	0	0

Sotsiaalteaduslike
rakendusuuringu
[*RAKE*]

Ülesanne 5

Contrast Tests

		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
Highest Year of School Completed	Assume equal variances	1	4,76	,213	22,367	1409	,000
		2	,18	,444	,398	1409	,691
	Does not assume equal variances	1	4,76	,235	20,223	429,032	,000
		2	,18	,632	,279	40,974	,781

H_1

H_0

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Tänase praktikumi teema:

I Regressioonanalüüs SPSSiga

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Regressioonanalüüs

Regressioonanalüüs on ökonomeetrilises analüüsis kasutatav põhimeetod, mis võimaldab kvantitatiivselt mõõta majandusnähtuste vahelisi seoseid.

Regressioonanalüüsi kvantitatiivseks tulemuseks on **hinnatud regressioonimudel** (**regressioonivõrrand**), mis kirjeldab statistilist seost sõltuva muutuja (endogeense muutuja) ning sõltumatu(te) (eksogeense muutuja) vahel.

Kui meil on ainult üks eksogeenne muutuja:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{u}_i$$

Lineaarse mudeli korral näitab parameetri hinnang $\hat{\beta}_1$, kui palju suureneb/väheneb sõltuva muutuja väärtus, kui sõltumatu muutuja väärtus kasvab ühe ühiku võrra

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Regressioonanalüüs

Kui meil on mitu sõltumatut muutujat, siis on tegu mitmese regressioonimudeliga

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1,i} + \hat{\beta}_2 X_{2,i} + \ldots + \hat{\beta}_k X_{k,i} + \hat{u}_i$$

Parameeter *näitab, kui palju muutub muutuja Y keskväärtus, kui sõltumatu muutuja muutub ühiku võrra ning teiste sõltumatute muutujate väärtused jäävad samaks (järgitud on ceteris paribus printsiip).*

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Regressioonanalüüs

NB! Klassikalise regressioonimudeli eeldused:

1. Juhuslike vigade *tinglikud keskväärtused (sõltumatu muutuja X fikseeritud väärtuse korral) on võrdsed nulliga iga i korral:*

$$E(u_i) = E(u|X_i) = 0 \text{ iga } i \text{ korral}$$

2. Juhuslike vigade tinglikud dispersioonid on konstantsed ja ei sõltu sõltumatutest ehk eksogeensetest muutujatest (homoskedastiivsus)

$$\text{var}(u_i|X_i) = E[u_i - E(u_i)]^2 = \sigma^2 = const$$

Kui nõue pole täidetud, on tegu **heteroskedastiivsus**ega

3. Juhuslikud vead ei korreleeru omavahel, s.t. nende kovariatsioon on null

$$\text{cov}(u_i, u_j) == 0 \text{ kui } i \neq j$$

Kui juhuslikud vead korreleeruvad omavahel, siis esineb mudelis **autokorrelatsioon**

4. Juhuslikud vead ei korreleeru sõltumatute muutujatega

$$cov\ (u\ ,\ X_{ij}) = 0 \qquad \text{iga } i \text{ ja } j \text{ korral, } 1 \leq j \leq k$$

5. Juhuslikud vead on normaaljaotusega

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Regressioonanalüüs

Kui klassikalise regressioonimudeli eeldused 1-4 on täidetud, saab anda hinnanguid vähimruutude meetodil leitud hinnangute keskväärtuste ja dispersioonide kohta. Tegemist on **punkthinnangute** omaduste uurimisega. Kui soovitakse leida ka usalduspiire ning testida hüpoteese hinnangute kohta, siis on oluline teha eeldusi juhuslike vigade jaotuse kohta. Seega viienda, normaaljaotuse nõude täidetus on eelkõige oluline vähimruutude meetodil leitud hinnangute omaduste analüüsimiseks.

BLUE – parim (**b**est) lineaarne (**l**inear) nihketa (**u**nbiased) hinnang (**e***stimator*)

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Regressioonanalüüs SPSSis

Täna tegeleme **lineaarse** regressioonanalüüsiga!

Andmebaas: **car_sales.sav**

Analyze → Regression → Linear

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	130,300	10	13,030	13,305	,000[a]
	Residual	138,082	141	,979		
	Total	268,383	151			

a. Predictors: (Constant), Fuel efficiency, Length, Price in thousands, Vehicle type, Width, Engine size, Fuel capacity, Wheelbase, Curb weight, Horsepower

b. Dependent Variable: Log-transformed sales

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Regressioonanalüüs SPSSis

Multikollineaarsus

Andmebaas: **car_sales.sav**

Analyze → Regression → Linear

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	130,300	10	13,030	13,305	,000[a]
	Residual	138,082	141	,979		
	Total	268,383	151			

a. Predictors: (Constant), Fuel efficiency, Length, Price in thousands, Vehicle type, Width, Engine size, Fuel capacity, Wheelbase, Curb weight, Horsepower

b. Dependent Variable: Log-transformed sales

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Regressioonanalüüs SPSSis

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,697[a]	,486	,449	,98960

a. Predictors: (Constant), Fuel efficiency, Length, Price in thousands, Vehicle type, Width, Engine size, Fuel capacity, Wheelbase, Curb weight, Horsepower

NB! Multikollineaarsuse oht!!!

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	-3,017	2,741		-1,101	,273					
	Vehicle type	,883	,331	,293	2,670	,008	,274	,219	,161	,304	3,293
	Price in thousands	-,046	,013	-,502	-3,596	,000	-,552	-,290	-,217	,187	5,337
	Engine size	,356	,190	,281	1,871	,063	-,135	,156	,113	,162	6,159
	Horsepower	-,002	,004	-,092	-,509	,611	-,389	-,043	-,031	,112	8,896
	Wheelbase	,042	,023	,241	1,785	,076	,292	,149	,108	,200	4,997
	Width	-,028	,042	-,073	-,676	,500	,037	-,057	-,041	,313	3,193
	Length	,015	,014	,148	1,032	,304	,215	,087	,062	,178	5,605
	Curb weight	,156	,350	,075	,447	,655	-,041	,038	,027	,131	7,644
	Fuel capacity	-,057	,047	-,167	-1,203	,231	-,016	-,101	-,073	,189	5,303
	Fuel efficiency	,081	,040	,262	2,023	,045	,121	,168	,122	,217	4,604

a. Dependent Variable: Log-transformed sales

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Regressioonanalüüs SPSSis

NB! Multikollineaarsuse oht!!!

Model	Dimension	Eigenvalue	Condition Index
1	1	9,920	1,000
	2	,733	3,678
	3	,259	6,193
	4	,050	14,051
	5	,019	22,589
	6	,008	35,942
	7	,005	44,275
	8	,003	58,480
	9	,002	76,175
	10	,001	130,747
	11	,000	148,267

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Multikollineaarsus

Vähimruutude meetodi kasutamisel klassikalise regressioonimudeli parameetrite hindamisel on oluliseks eelduseks, et **mudeli sõltumatud muutujad ei ole omavahel otseses lineaarses seoses**. ***Kui sõltumatute muutujate vahel on otsene lineaarne seos, siis vähimruutude meetodit regressioonimudeli parameetrite hindamisel kasutada ei saa.***

Kui mudeli sõltuvad muutujad on omavahel tugevas korrelatiivses seoses, on mudelis tegemist **multikollineaarsusega** *(multicollinearity).* Sel juhul on võimalik küll vähimruutude meetodit parameetrite hindamiseks kasutada, kuid probleeme võib tulla mudeli parameetrite sisulisel tõlgendamisel. Võimalikd viited multikoolineaarsusele:

- regressioonimudel on statistiliselt oluline, kuid kõik parameetrid või enamus neist ei ole statistiliselt olulised.

- statistiliselt oluline ning statistiliselt oluliste parameetritega mudel ei ole kooskõlas majandusteoreetiliste seisukohtadega ega kasutada olevate andmetega.

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Samm-sammuline regressioonanalüüs

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions		
				(Constant)	Price in thousands	Wheelbase
1	1	1,885	1,000	,06	,06	
	2	,115	4,051	,94	,94	
2	1	2,847	1,000	,00	,02	,00
	2	,150	4,351	,01	,97	,01
	3	,003	33,412	,99	,00	,99

a. Dependent Variable: Log-transformed sales

Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Samm-sammuline regressioonanalüüs

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,552[a]	,304	,300	1,11553
2	,655[b]	,430	,422	1,01357

a. Predictors: (Constant), Zscore: Price in thousands

b. Predictors: (Constant), Zscore: Price in thousands, Zscore: Wheelbase

c. Dependent Variable: Log-transformed sales

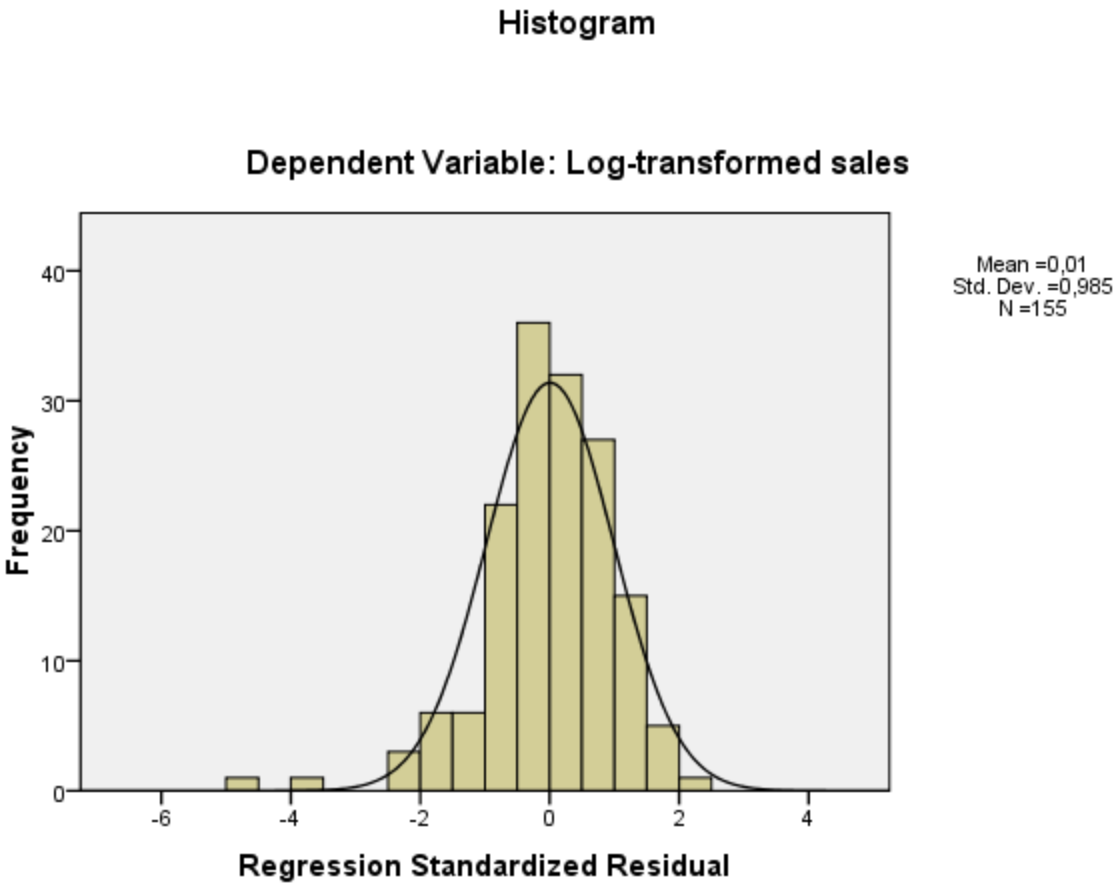
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	4,684	,194		24,090	,000		
	Price in thousands	-,051	,006	-,552	-8,104	,000	1,000	1,000
2	(Constant)	-1,822	1,151		-1,583	,116		
	Price in thousands	-,055	,006	-,590	-9,487	,000	,988	1,012
	Wheelbase	,061	,011	,356	5,718	,000	,988	1,012

a. Dependent Variable: Log-transformed sales

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Samm-sammuline regressioonanalüüs



Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

III Klasteranalüüs

Andmete standardiseerimine. Näide

- Kolme indiviidi A, B ja C andmed nende toote X ostmise tõenäosuse ning toote X telereklaamide vaatamisele kulutatud aja kohta.

Objekt	Ostmise tõenäosus	Reklaamide vaatamisaeg	
		Minutites	Sekundites
A	60	3.0	180
B	65	3.5	210
C	63	4.0	240

Andmete standardiseerimine. Näide

- Järjestus, kui aeg mõõdetud minutites

Objektide paar	Lihtne Euclideani distants		Ruutu tõstetud ehk absoluutne Euclideani distants		City-block distants	
	Väärtus	Astak	Väärtus	Astak	Väärtus	Astak
A-B	5.025	3	25.25	3	5.5	3
A-C	3.162	2	10.00	2	4.0	2
B-C	2.062	1	4.25	1	2.5	1

Andmete standardiseerimine. Näide

- Järjestus, kui aeg mõõdetud sekundites

Objektide paar	Lihtne Euclideani distants		Ruutu tõstetud ehk absoluutne Euclideani distants		City-block distants	
	Väärtus	Astak	Väärtus	Astak	Väärtus	Astak
A-B	30.41	2	925	2	35	2
A-C	60.07	3	3609	3	63	3
B-C	30.06	1	904	1	32	1

Andmete standardiseerimine. Näide

- Kasutades andmete standardiseerimist

Objektide paar	Standardiseeritud väärtused		Lihtne Euclideani distants		Ruutu tõstetud ehk absoluutne Euclideani distants		City-block distants	
	Ostmise tn	Min/sek vaatamisaeg	Väärtus	Astak	Väärtus	Astak	Väärtus	Astak
A-B	-1.06	-1.0	2.22	2	4.95	2	2.99	2
A-C	0.93	0.0	2.33	3	5.42	3	3.19	3
B-C	0.13	1.0	1.28	1	1.63	1	1.79	1

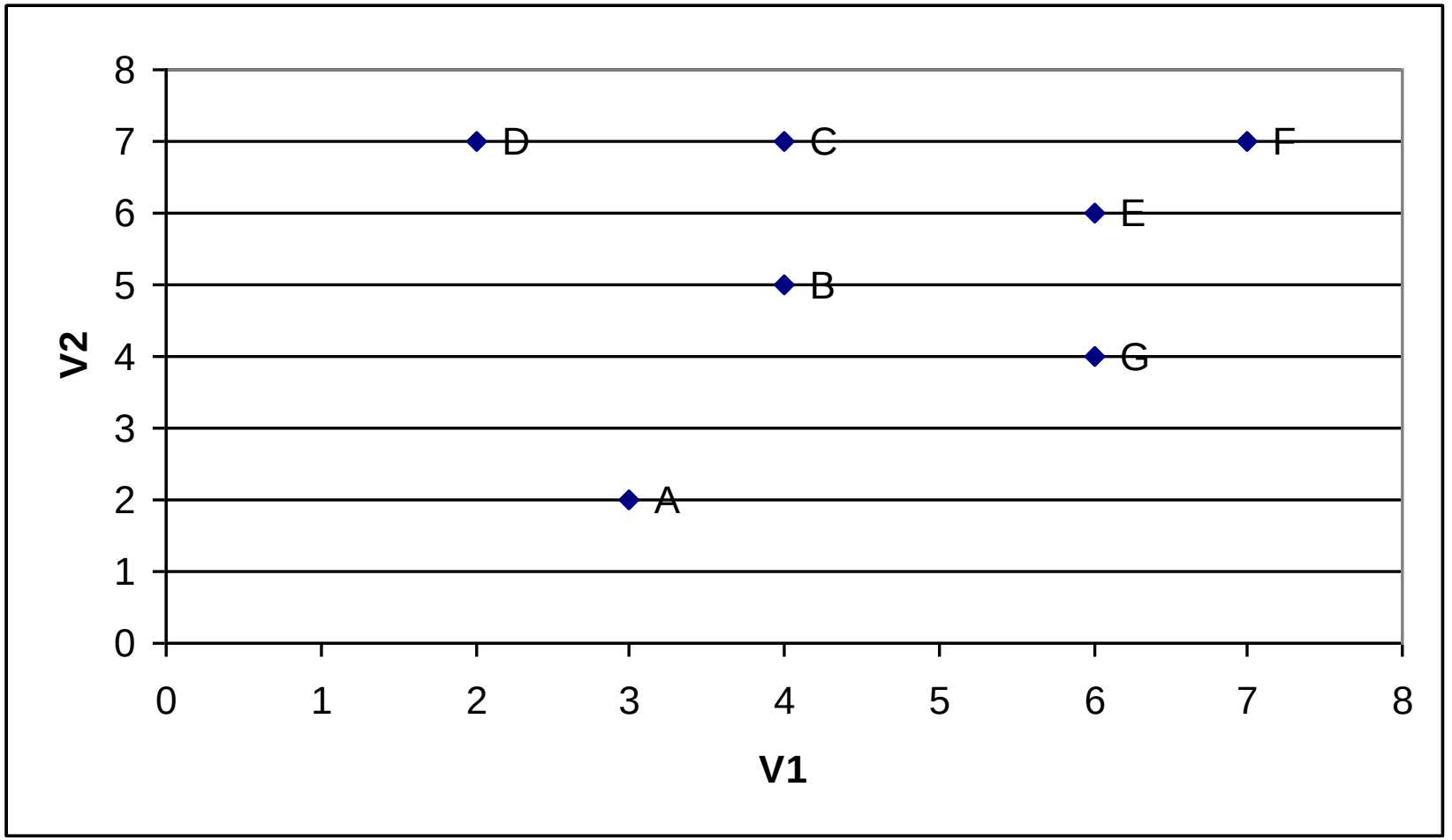
Näide

- Kaupluse omanik soovib analüüsida klientide lojaalsust kauplusele ning tootele (skaala 0-10)
- Pilootvalim: 7 inimest

Andmed

	vastaja						
klastermuutuja	A	B	C	D	E	F	G
V_1	3	4	4	2	6	7	6
V_2	2	5	7	7	6	7	4

Andmeid kirjeldav punktdiagramm

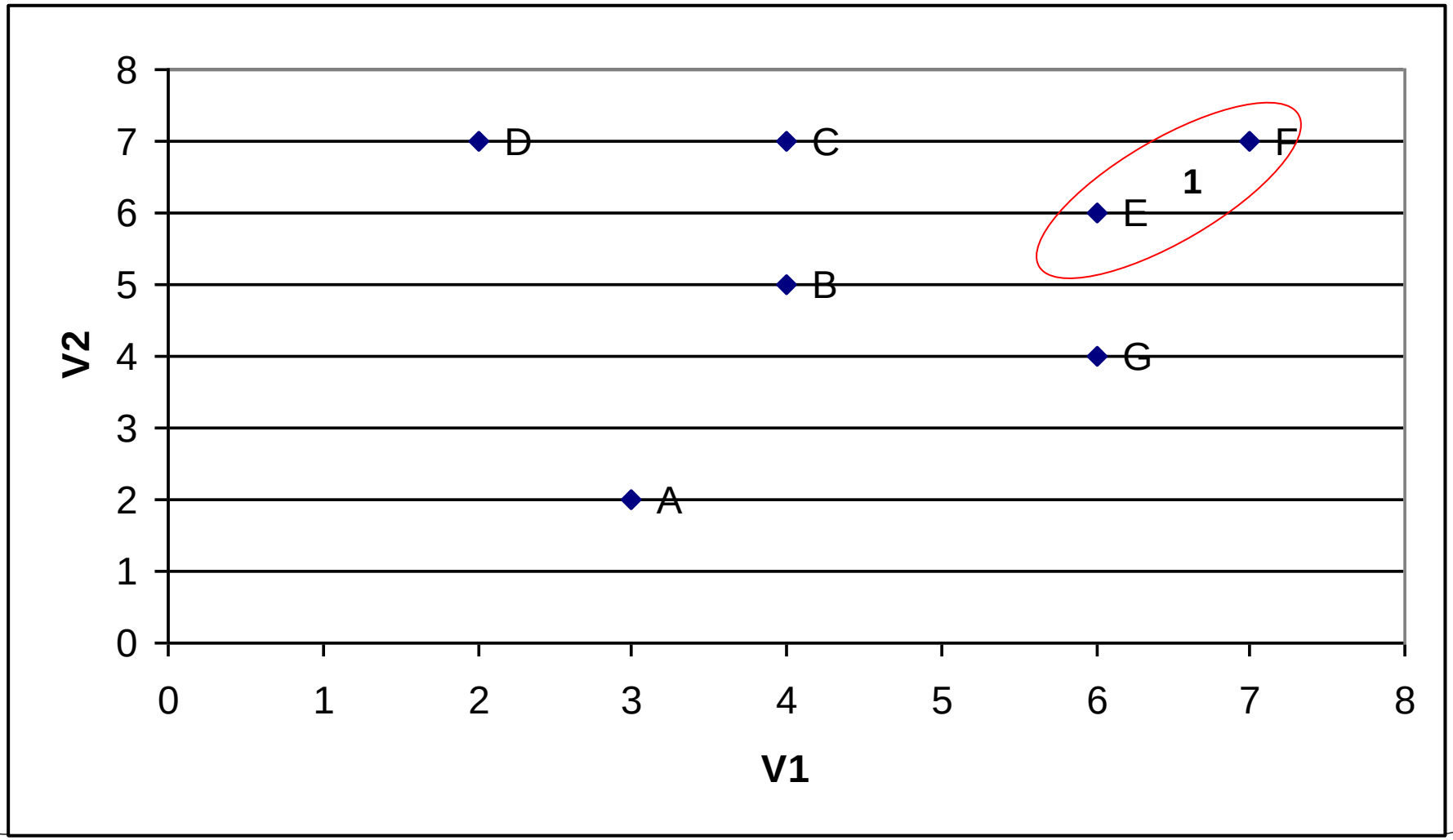


Euclideani distantsid

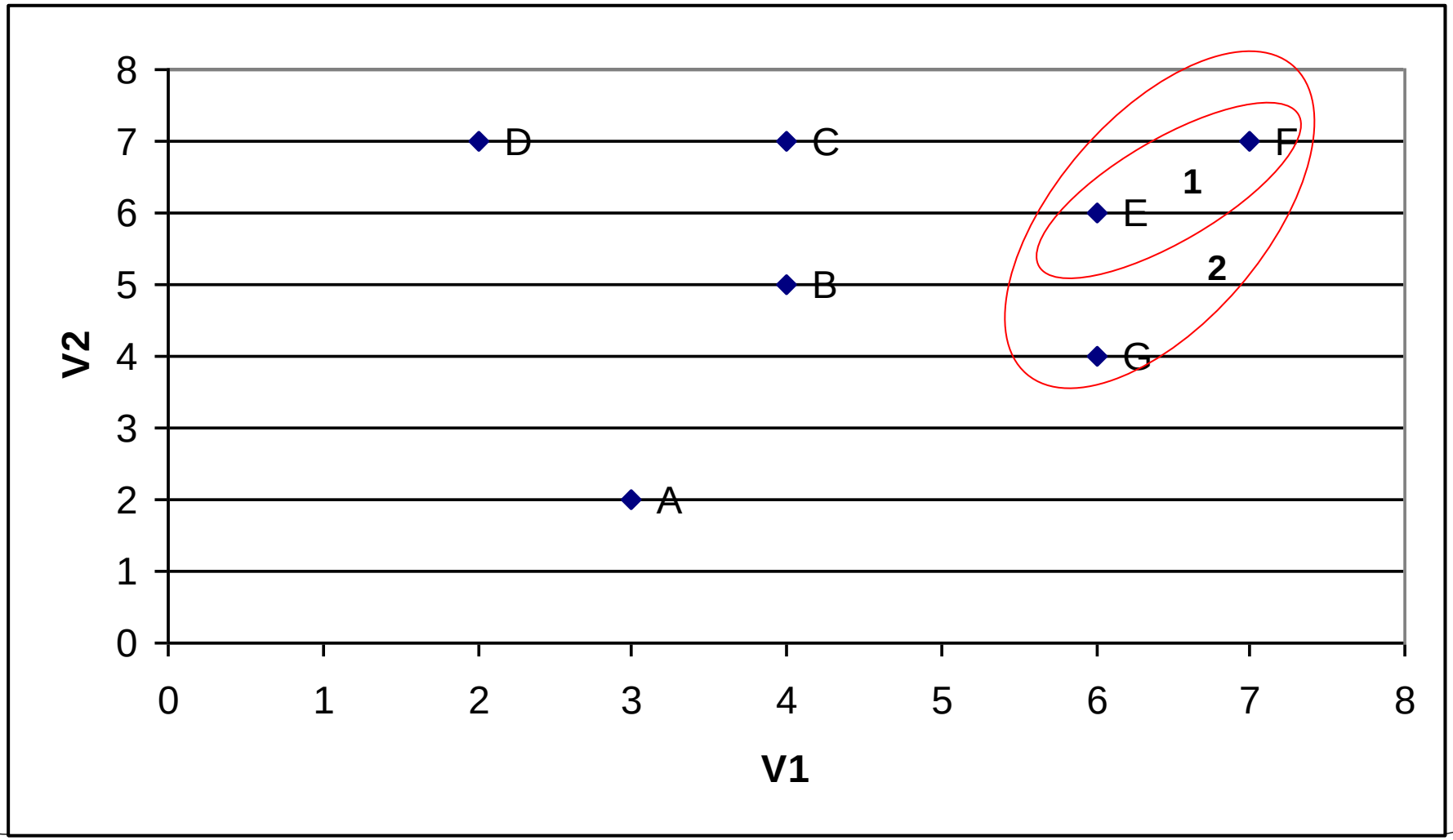
	Vaatlus						
Vaatlus	A	B	C	D	E	F	G
A	-						
B	3.162	-					
C	5.099	2.000	-				
D	5.099	2.828	2.000	-			
E	5.000	2.236	2.236	4.123	-		
F	6.403	3.606	3.000	5.000	1.414	-	
G	3.606	2.236	3.606	5.000	2.000	3.162	-

NB! Väiksem distants viitab objektide suuremale sarnasusele!

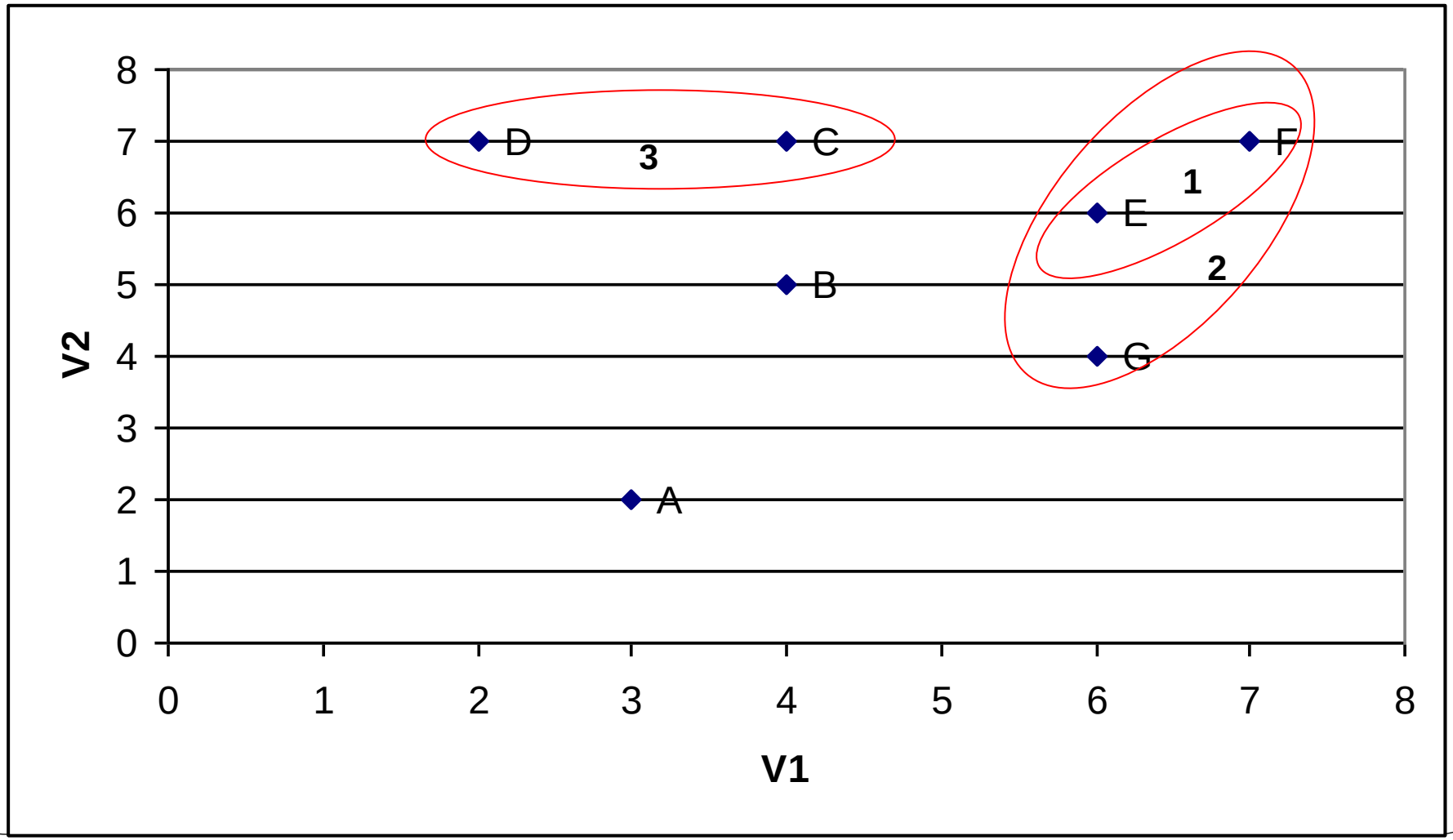
Klastitesse jagamise graafiline esitus (1)



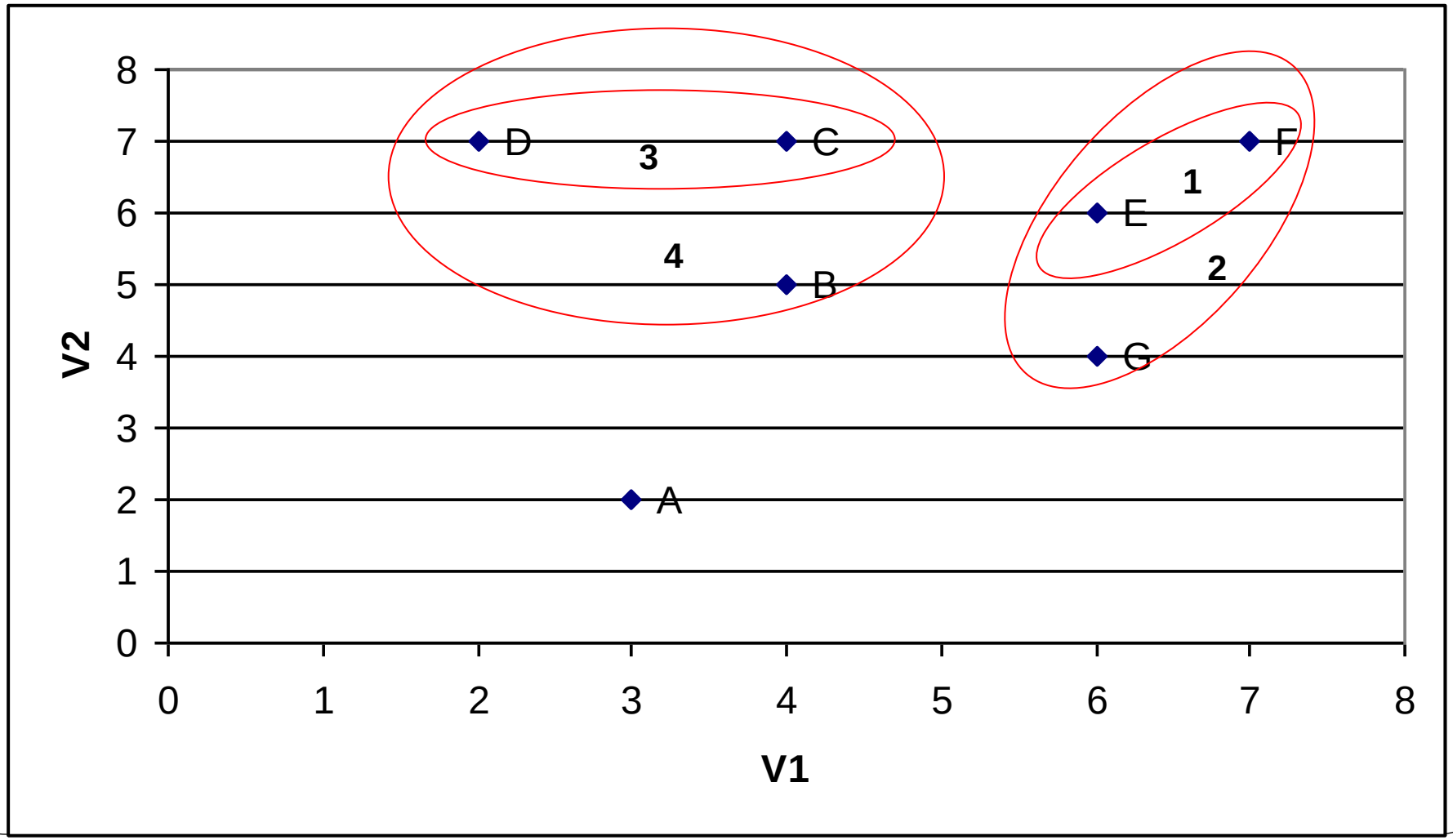
Klastitesse jagamise graafiline esitus (2)



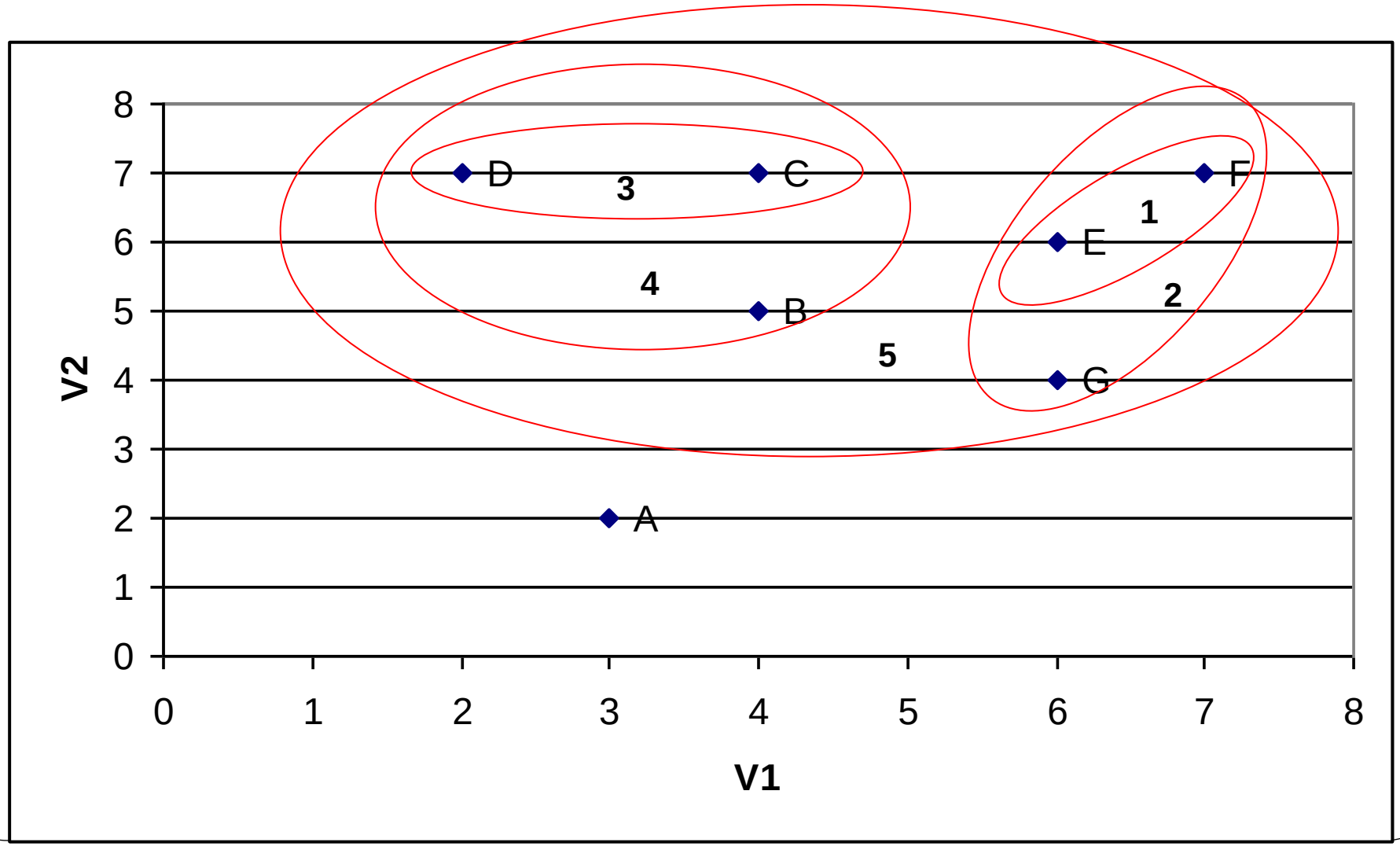
Klastitesse jagamise graafiline esitus (3)



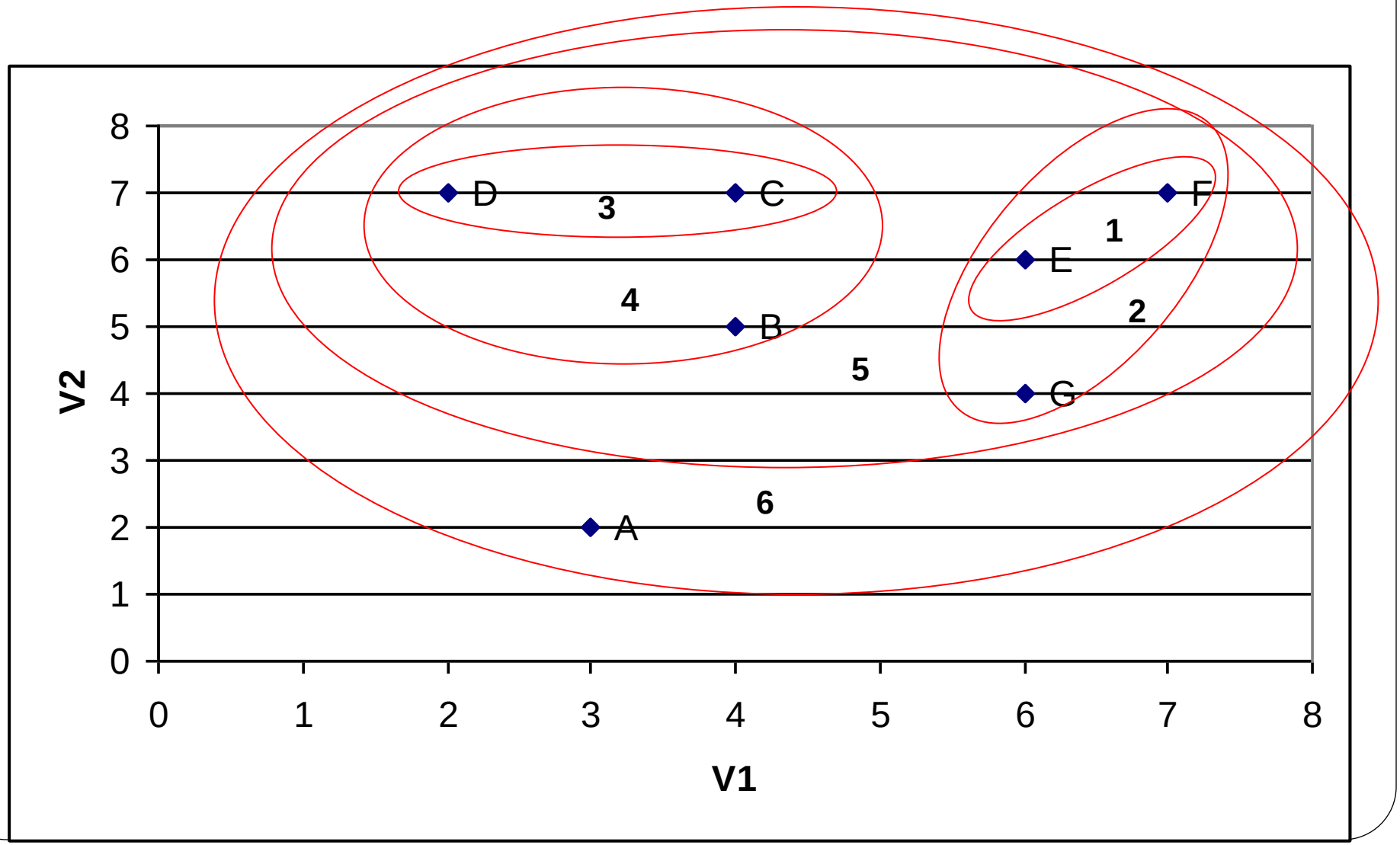
Klastitesse jagamise graafiline esitus (4)



Klastitesse jagamise graafiline esitus (5)



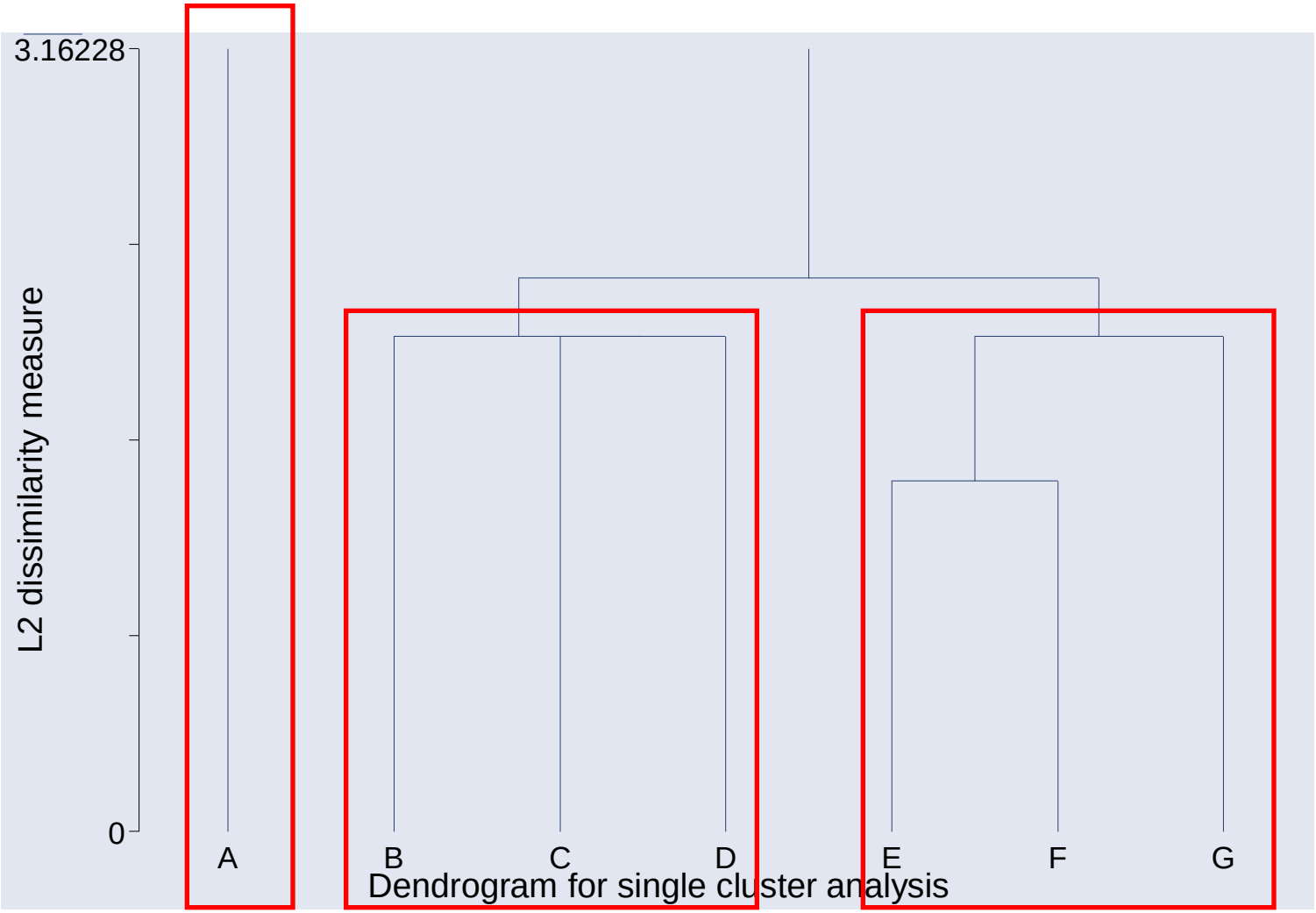
Klastitesse jagamise graafiline esitus (6)



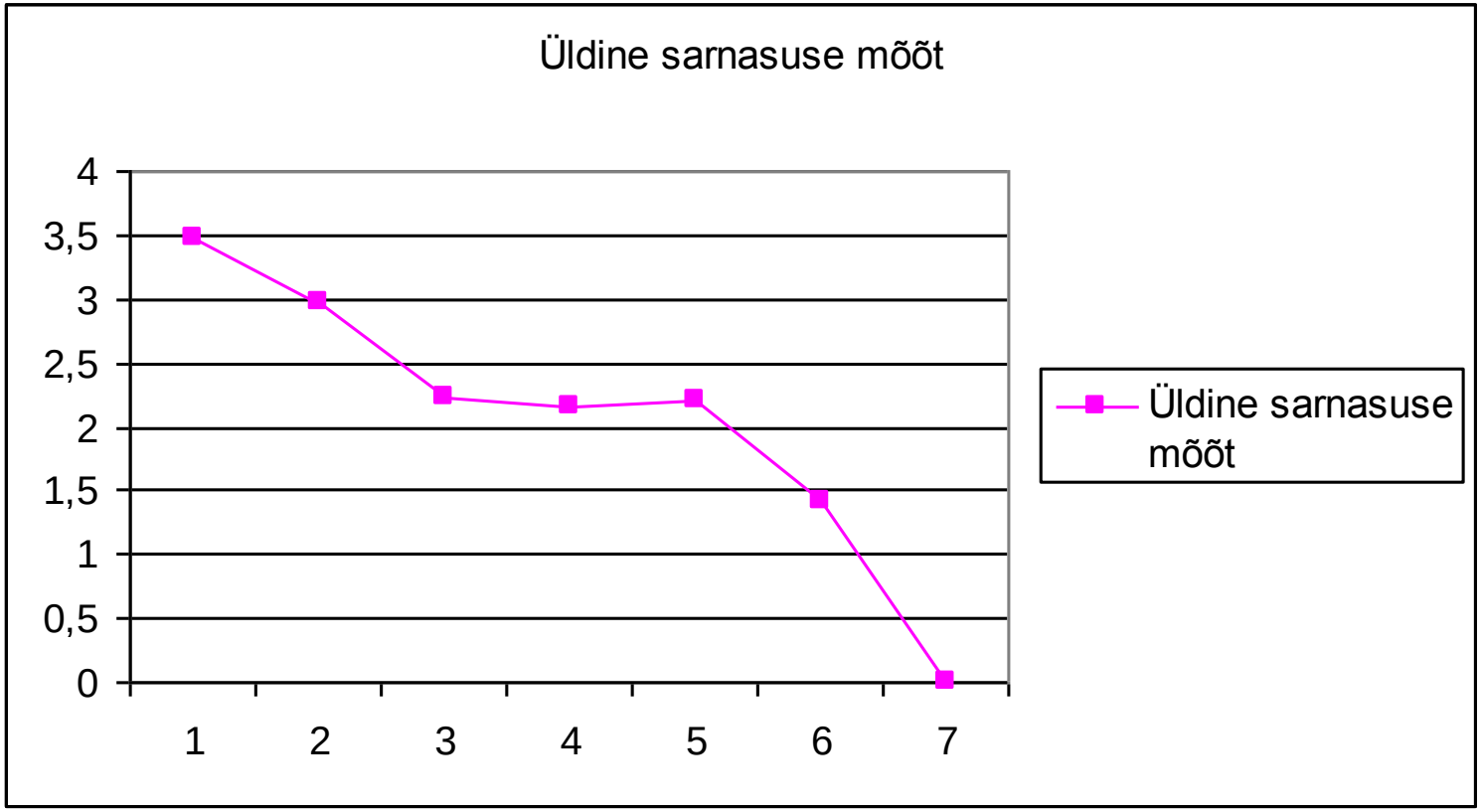
Objektide jaotamine klastritesse

	liitmisprotsess				
Aste	Min. distants	Vaatluste paar	Klastrite suhe	Klastrite arv	
			(A) (B) (C) (D) (E) (F) (G)	7	
1	1.414	E-F	(A) (B) (C) (D) (E-F) (G)	6	
2	2.000	E-G	(A) (B) (C) (D) (E-F-G)	5	
3	2.000	C-D	(A) (B) (C-D) (E-F-G)	4	
4	2.000	B-C	(A) (B-C-D) (E-F-G)	3	
5	2.236	B-E	(A) (B-C-D-E-F-G)	2	
6	3.162	A-B	(A-B-C-D-E-F-G)	1	

Dendrogramm



Sobiva klastrite arvu valik. Scree plot



Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Kahesammuline klasteranalüüs

- Kasutatakse, et tuvastada andmestikus gruppe (klastreid)

Head küljed:

- Klastrite moodustamisel on võimalik aluseks võtta nii kategoorilisi kui pidevaid muutujaid
- Klastrite arv valitakse automaatselt
- Võimalik on leida suurtes andmemassiivides teatud süsteemsust, mustreid, sarnaseid objekte

NB! Silmas tuleb pidada, et klasteranalüüsi korral tehakse teatud eeldusi:

- eeldatakse, et klastrimudelis on muutujad sõltumatud;
- eeldatakse, et pidevad muutujad on normaaljaotusega ja kategoorilised muutujad multinoomse jaotusega

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Kahesammuline klasteranalüüs

BIC – mida madalam, seda parem…
… aga oluline on ka muudu suurus – hea lahendus on see, kus BICi muutus on arvestatav
–> 3 klastrit

Auto-Clustering

Number of Clusters	Schwarz's Bayesian Criterion (BIC)	BIC Change[a]	Ratio of BIC Changes[b]	Ratio of Distance Measures[c]
1	1214,377			
2	974,051	-240,326	1,000	1,829
3	885,924	-88,128	,367	2,190
4	897,559	11,635	-,048	1,368
5	931,760	34,201	-,142	1,036
6	968,073	36,313	-,151	1,576
7	1026,000	57,927	-,241	1,083
8	1086,815	60,815	-,253	1,687
9	1161,740	74,926	-,312	1,020
10	1237,063	75,323	-,313	1,239
11	1316,271	79,207	-,330	1,046
12	1396,192	79,921	-,333	1,075
13	1477,199	81,008	-,337	1,076
14	1559,230	82,030	-,341	1,301
15	1644,366	85,136	-,354	1,044

a. The changes are from the previous number of clusters in the table.

b. The ratios of changes are relative to the change for the two cluster solution.

c. The ratios of distance measures are based on the current number of clusters against the previous number of clusters.

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Kahesammuline klasteranalüüs

Cluster Distribution

		N	% of Combined	% of Total
Cluster	1	62	40,8%	39,5%
	2	39	25,7%	24,8%
	3	51	33,6%	32,5%
	Combined	152	100,0%	96,8%
Excluded Cases		5		3,2%
Total		157		100,0%

Vehicle type

		Automobile		Truck	
		Frequency	Percent	Frequency	Percent
Cluster	1	61	54,5%	1	2,5%
	2	0	,0%	39	97,5%
	3	51	45,5%	0	,0%
	Combined	112	100,0%	40	100,0%

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

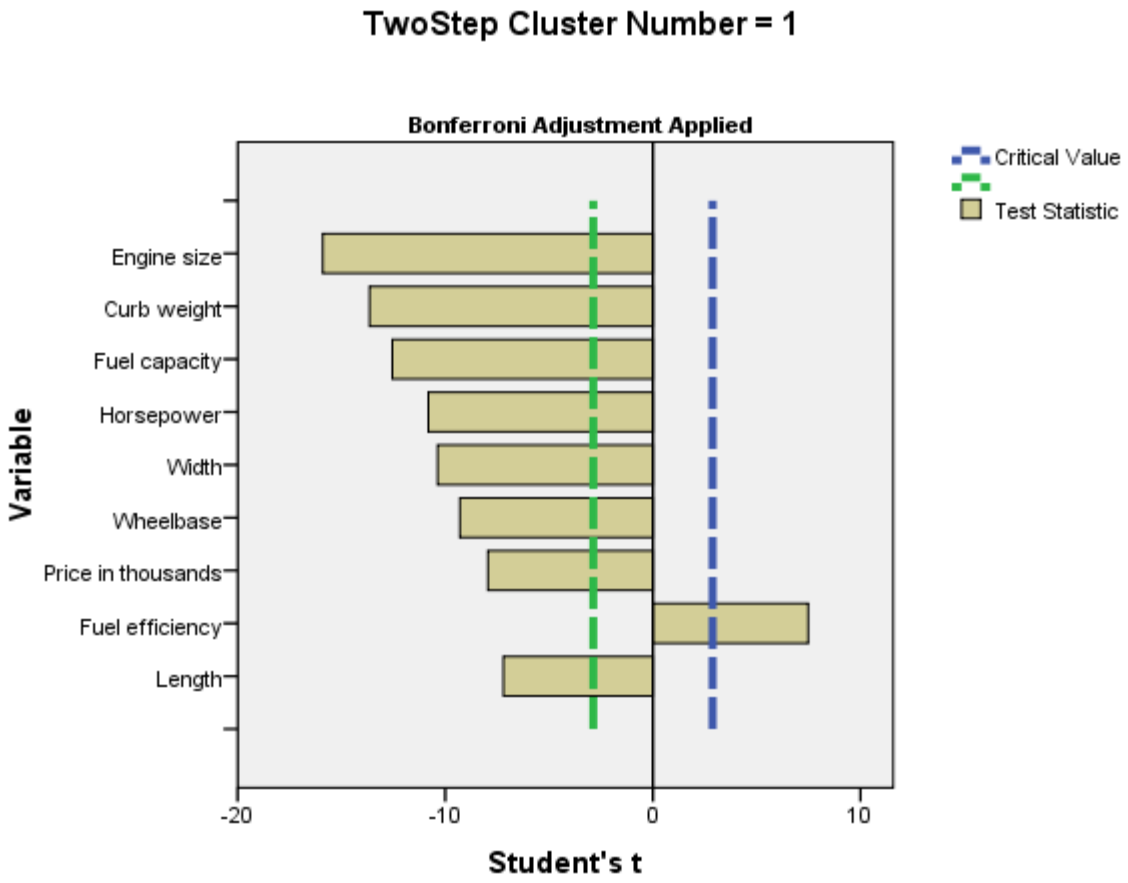
Kahesammuline klasteranalüüs

Centroids

		Cluster			
		1	2	3	Combined
Price in thousands	Mean	**19,61671**	26,56182	**37,29980**	27,33182
	Std. Deviation	7,644070	10,185175	17,381187	14,418669
Engine size	Mean	2,194	3,559	**3,700**	3,049
	Std. Deviation	,4238	,9358	,9493	1,0498
Horsepower	Mean	143,24	187,92	**232,96**	184,81
	Std. Deviation	30,259	39,049	54,408	56,823
Wheelbase	Mean	102,595	112,972	109,022	107,414
	Std. Deviation	4,0799	9,6537	5,7644	7,7178
Width	Mean	68,539	72,744	72,924	71,089
	Std. Deviation	1,9366	4,1781	2,1855	3,4647
Length	Mean	178,235	191,110	194,688	187,059
	Std. Deviation	9,6534	14,4415	10,3512	13,4712
Curb weight	Mean	2,83742	3,96759	3,57890	3,37618
	Std. Deviation	,310867	,671766	,297204	,636593
Fuel capacity	Mean	14,979	**22,064**	18,443	17,959
	Std. Deviation	1,8699	4,2894	2,0445	3,9376
Fuel efficiency	Mean	**27,24**	19,51	23,02	23,84
	Std. Deviation	3,578	2,910	2,060	4,305

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Kahesammuline klasteranalüüs



Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Hierarhiline klasteranalüüs

- Kasutatakse, et tuvastada andmestikus gruppe (klastreid)
- Kasutatakse, kui objekte on vähe
- Alustatakse sellest, et iga objekt moodustab omaette klastri. Esmalt ühendatakse kaks kõige sarnasemat objekti/muutujat, seejärel taas kaks kõige sarnasemat jne. Protsess lõppeb sellega, et kõik muutujad on koondatud ühte klastrisse.
- Protsessi graafilist esitlust nimetatakse **dendrogramm**iks

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

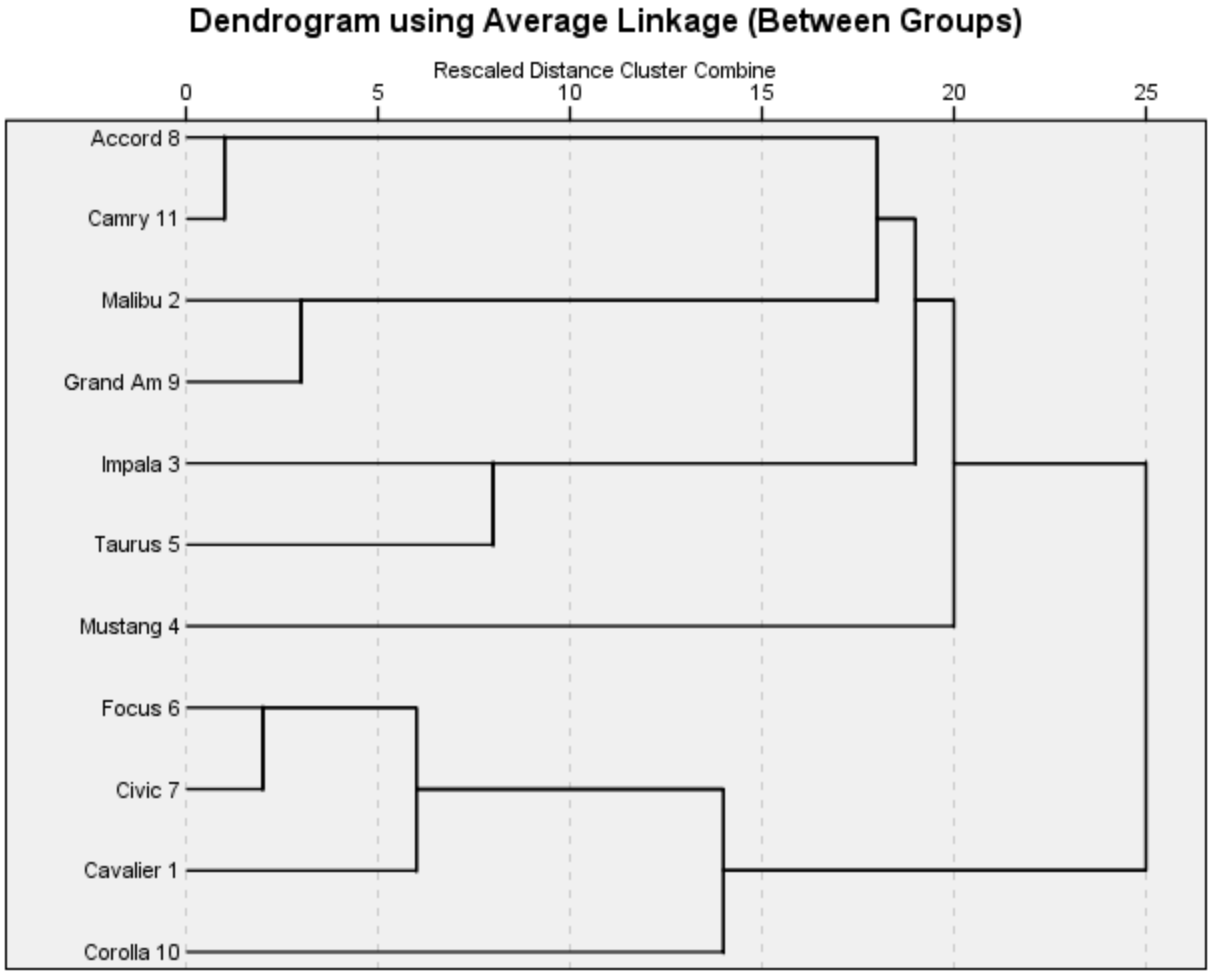
Hierarhiline klasteranalüüs

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	8	11	1,260	0	0	7
2	6	7	1,579	0	0	4
3	2	9	1,625	0	0	7
4	1	6	2,318	0	2	6
5	3	5	2,619	0	0	8
6	1	10	3,670	4	0	10
7	2	8	4,420	3	1	8
8	2	3	4,505	7	5	9
9	2	4	4,774	8	0	10
10	1	2	5,718	6	9	0

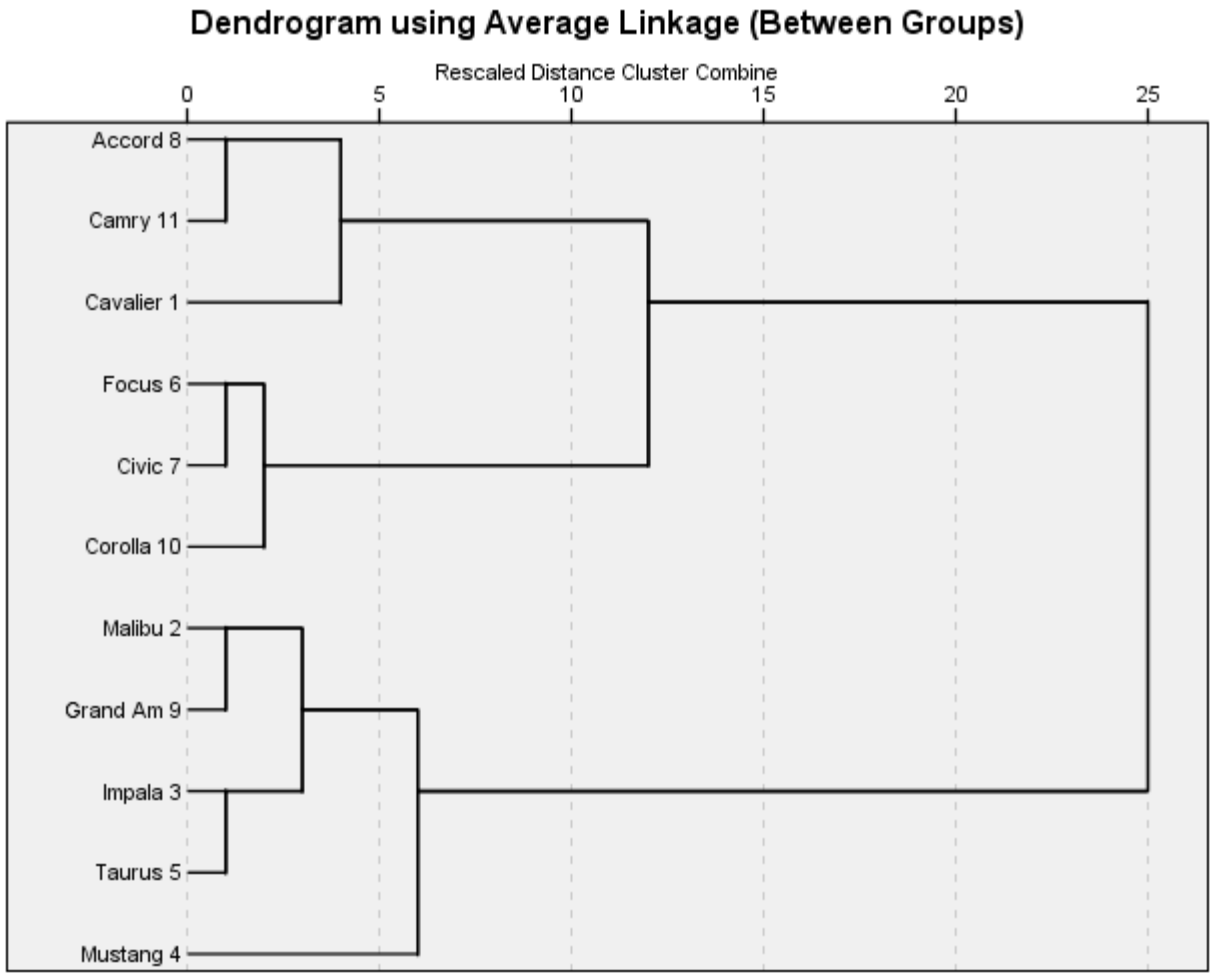
Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Hierarhiline klasteranalüüs



Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Hierarhiline klasteranalüüs



Sotsiaalteaduslike rakendusuuringute keskus [*RAKE*]

Mittehierarhiline ehk K-means klasteranalüüs

- Kasutatakse, et tuvastada andmestikus gruppe (klastreid)
- Kasutatakse, kui meil on klastrite arv teada
- Kasutatakse, kui objekte palju

Protsess:

- Määratakse esialgsed klastri keskmed (cluster centres)
- Objektid jagatakse klastritesse, võttes aluseks nende kauguse klastri keskmest
- Leitakse uued klastri keskmed jne

Iteratiivne protsess

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Mittehierarhiline klasteranalüüs

Iteration History[a]

Iteration	Change in Cluster Centers		
	1	2	3
1	4,565	3,899	3,975
2	,699	,197	1,823
3	,389	,182	,601
4	,193	,091	,597
5	,078	,090	,351
6	,000	,141	,443
7	,000	,035	,117
8	,000	,000	,000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is ,000. The current iteration is 8. The minimum distance between initial centers is 10,326.

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Mittehierarhiline klasteranalüüs

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Zscore: Price in thousands	30,387	2	,615	149	49,408	,000
Zscore: Engine size	57,271	2	,255	149	224,832	,000
Zscore: Horsepower	43,110	2	,439	149	98,166	,000
Zscore: Wheelbase	27,649	2	,663	149	41,722	,000
Zscore: Width	42,256	2	,454	149	93,126	,000
Zscore: Length	30,492	2	,610	149	49,979	,000
Zscore: Curb weight	50,179	2	,360	149	139,556	,000
Zscore: Fuel capacity	45,711	2	,426	149	107,323	,000
Zscore: Fuel efficiency	37,356	2	,522	149	71,497	,000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Mittehierarhiline klasteranalüüs

Final Cluster Centers

	Cluster		
	1	2	3
Zscore: Price in thousands	-,58809	,12022	1,34528
Zscore: Engine size	-,86360	,23733	1,74226
Zscore: Horsepower	-,79218	,24963	1,42275
Zscore: Wheelbase	-,66949	,28632	1,01185
Zscore: Width	-,80908	,29864	1,33192
Zscore: Length	-,73877	,34766	,93731
Zscore: Curb weight	-,86234	,33622	1,47717
Zscore: Fuel capacity	-,75885	,22229	1,57030
Zscore: Fuel efficiency	,74852	-,30752	-1,25325

Number of Cases in each Cluster

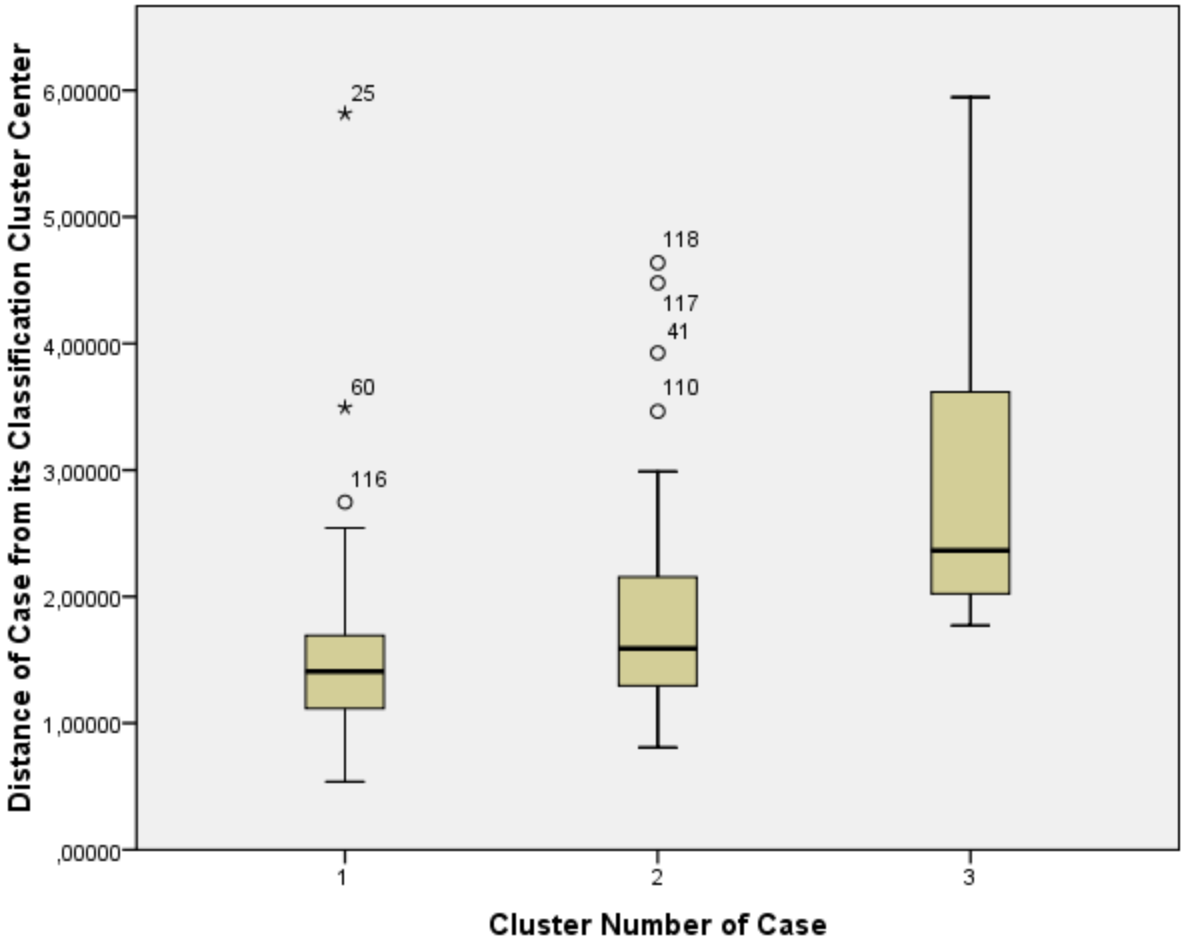
Cluster	1	63,000
	2	68,000
	3	21,000
Valid		152,000
Missing		5,000

Distances between Final Cluster Centers

Cluster	1	2	3
1		3,104	6,369
2	3,104		3,331
3	6,369	3,331	

Sotsiaalteaduslike
rakendusuuringute keskus
[*RAKE*]

Mittehierarhiline klasteranalüüs



Tänan tähelepanu eest!