TARTU ÜLIKOOL
EUROOPA KOLLEDŽ

e-õppe programm

Euroopa Liit
Euroopa Sotsiaalfond

Eesti tuleviku heaks

STATISTILISE ANALÜÜSI TEOSTAMINE EXCELI JA SPSSI ABIL (P2EC.00.146)

Konspekt

Lektor: Kerly Krillo

Tartu 2010

I LOENG

Diagrammide tegemine Excelis

Microsoft Office Excelis 2007 on võimalik kasutada mitmeid diagrammitüüpe, mis võimaldavad andmeid selgemalt esitada. Diagrammide loomise kohta leiate lisateavet teemast Diagrammi loomine (lisatud Moodle'is lingina). Käesolevas konspekti koostamisel on valdavalt kasutatud MS veebilehel olevat infot.

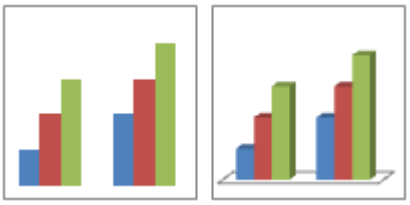
Sisukord

1. Tulpdiagrammid

Töölehel veergudesse või ridadesse korraldatud andmed saab kanda tulpdiagrammile. Tulpdiagrammid sobivad eelkõige mingi perioodi jooksul andmetes toimunud muutuste näitamiseks või üksuste võrdluse illustreerimiseks. Tulpdiagrammidel on kategooriad tavaliselt paigutatud horisontaalteljele ning (arvulised) väärtused vertikaalteljele. Tulpdiagrammidel on järgnevad alamtüübid.

1.1. Kobartulpdiagramm ja ruumiline kobartulpdiagramm

Kobartulpdiagrammid võrdlevad väärtusi kategooriate lõikes. Väärtused kuvatakse tasapinnaliste vertikaalsete ristkülikutena. Ruumilises kobartulpdiagrammis esitatakse andmed küll ruumilises vaates, kuid kolmandat väärtustelge (sügavustelge) ei kasutata.

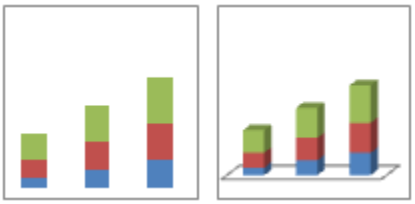


Kobartulpdiagrammi kasutatakse järgnevate kategooriate puhul:

- väärtustevahemikud (nt üksuste loendid);
- kindlad skaalakorraldused (nt Likerti skaala kirjed: täiesti nõus, nõus, neutraalne, ei ole nõus, ei ole üldse nõus);
- nimed, mis ei ole kindlas järjestuses (nt üksuste nimed, geograafilised nimed või inimeste nimed).

1.2. Virntulpdiagramm ja ruumiline virntulpdiagramm

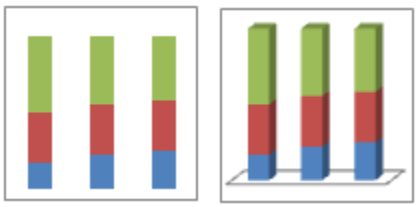
Virntulpdiagrammidel kuvatakse üksikute elementide seos tervikuga, võrreldes eri kategooriate kõigi väärtuste osakaalu kogusummas. Väärtused kuvatakse kahemõõtmeliste vertikaalsete virnastatud ristkülikutena. Ruumilises virntulpdiagrammis esitatakse andmed küll ruumilises vaates, kuid kolmandat väärtustelge (sügavustelge) ei kasutata.



Virntulpdiagrammi kasutatakse mitme andmesarja puhul ja siis, kui soovitakse rõhutada kogusummat.

1.3. 100% virntulpdiagramm ja ruumiline 100% virntulpdiagramm

100% virntulpdiagramm ja ruumiline 100% virntulpdiagramm võrdlevad eri kategooriate kõigi väärtuste protsentuaalset osakaalu kogusummas. 100% virntulpdiagrammi puhul kuvatakse väärtused vertikaalsete tasapinnaliste 100% virnastatud ristkülikutena.



100% virntulpdiagrammi on otstarbekas kasutada kolme või enama andmesarja puhul ning siis, kui tahetakse rõhutada väärtuste osakaalu kogusummas, eriti kui kogusumma on igas kategoorias sama.

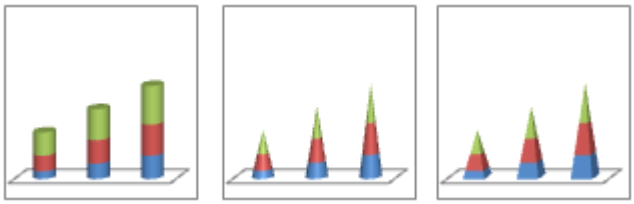
1.4. Ruumiline tulpdiagramm

Ruumilisel tulpdiagrammil on kolm muudetavat telge (horisontaal-, vertikaal- ja sügavustelg) ning see võrdleb andmepunkte horisontaal- ja sügavusteljel.

Ruumilist tulpdiagrammi kasutatakse andmete võrdlemiseks korraga nii kategooriate kui ka sarjade lõikes, sest selle diagrammitüübi puhul on kuvatud kategooriad nii horisontaal- kui ka sügavusteljel ning väärtused vertikaalteljel.

1.5. Silinder, koonus ja püramiiddiagrammid

Silinder-, koonus- ja püramiiddiagrammid on saadaval samade kobar-, virn-, 100% virn- ja ruumiliste diagrammitüüpidena kui ristkülikutega tulpdiagrammid ning neil kuvatakse ja võrreldakse andmeid täpselt sama moodi. Ainuke erinevus seisneb diagrammide kujus (ristküliku asemel kas silindri-, koonuse- või püramiidikujuline).



2. Joondiagrammid

Töölehel veergudesse või ridadesse korraldatud andmed saab kanda joondiagrammile. Joondiagrammidel kuvatakse ajaliselt järjestikused andmed ühisel skaalal, seega sobivad need hästi andmete trendi näitamiseks võrdsete ajavahemike tagant. Joondiagrammil on kategooriaandmed jaotatud ühtlaselt horisontaalteljele ning väärtuste andmed ühtlaselt vertikaalteljele. Joondiagrammi on soovitav kasutada siis, kui kategooriasiltideks on tekst, mis tähistab ühtlase vahemikuga väärtusi (nt kuud, kvartalid, finantsaastad). Seda eriti juhul, kui tegemist on mitme andmesarjaga (ühe andmesarja jaoks võiksite kasutada kategooriadiagrammi). Joondiagrammi on soovitav kasutada ka siis, kui on mitu ühtlase vahega arvulist kategooriasilti (eriti aastad). Rohkem

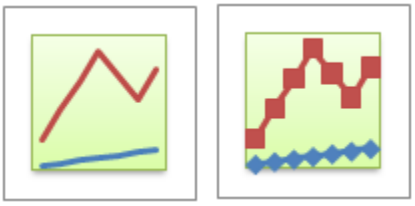
kui kümne arvsildi puhul tuleks kasutada punktdiagrammi. Joondiagrammidel on järgnevad alamtüübid.

2.1. Joondiagramm ja tähistega joondiagramm

Joondiagramme saab kuvada nii eraldi andmeväärtusi tähistavate tähistega kui ka ilma. Joondiagrammid sobivad trendide kuvamiseks ajaliselt või järjestatud kategooriate kaupa, eriti kui andmepunkte on palju ja nende esitamise järjestus on oluline. Mitme kategooria või ligikaudsete väärtuste puhul on soovitatav kasutada tähisteta joondiagrammi.

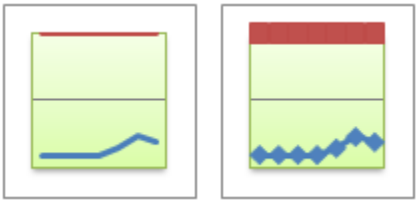
2.2. Virnjoondiagramm ja tähistega virnjoondiagramm

Virnjoondiagramme saab kuvada nii eraldi andmeväärtusi tähistavate tähistega kui ka ilma. Virnjoondiagramme saab kasutada iga väärtuse osakaalu trendi kuvamiseks ajaliselt või järjestatud kategooriate kaupa, kuid kuna pole lihtne näha, et jooned on virnastatud, eelistatakse neile siiski enamasti muud tüüpi joondiagrammi või virnkihtdiagrammi.



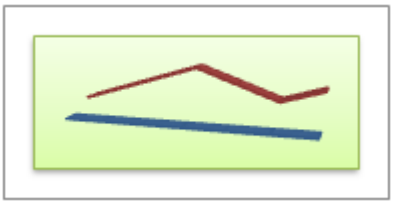
2.3. 100% virnjoondiagramm ja 100% tähistega virnjoondiagramm

Virnjoondiagramme saab kuvada nii eraldi andmeväärtusi tähistavate tähistega kui ka ilma. 100% virnjoondiagrammid sobivad iga väärtuse protsentuaalse osakaalu trendi kuvamiseks ajaliselt või kategooriate kaupa. Mitme kategooria või ligikaudsete väärtuste puhul on soovitatav kasutada tähisteta 100% virnjoondiagrammi. Seda tüüpi andmete esitamiseks võib paremini sobida 100% virnkihtdiagramm.



2.4. Ruumiline joondiagramm

Ruumilisel joondiagrammil kuvatakse kõik andmeread või -veerud ruumilise ribana. Ruumilisel joondiagrammil on muudetav horisontaal-, vertikaal- ja sügavustelg.



3. Sektordiagrammid

Töölehel ainult ühte veergu või ritta korraldatud andmed saab kanda sektordiagrammile. Sektordiagrammidel kuvatakse ühe andmesarja elementide maht kõigi elementide kogusumma suhtes. Sektordiagrammil kuvatakse andmepunktid protsendina tervikust.

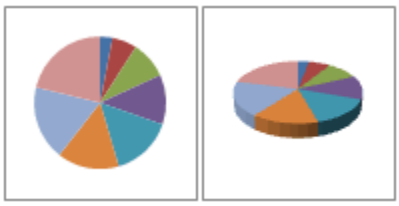
Sektordiagrammi kasutatakse enamasti järgmistel juhtudel:

- 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.

Sektordiagrammidel on järgnevad alamtüübid.

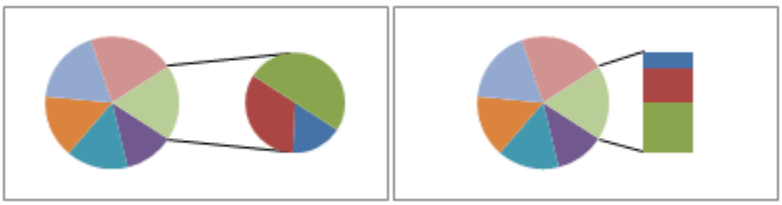
3.1. Sektordiagramm ja ruumiline sektordiagramm

Sektordiagrammides kuvatakse kõigi väärtuste osakaal kogusummas tasapinnalise või ruumilisena. Sektoreid saab diagrammist nende rõhutamiseks käsitsi eraldada.



3.2. Sektordiagramm sektordiagrammist ja lintdiagramm sektordiagrammist

Sektordiagrammil sektordiagrammist või lintdiagrammil sektordiagrammist kuvatakse kasutaja määratud väärtustega sektordiagrammid, mis on pärit põhi-sektordiagrammist ning lisatud kas teise sektordiagrammi või virnlintdiagrammi. Neid diagrammitüüpe saab kasutada ka juhul, kui soovitakse põhi-sektordiagrammi väikesi sektoreid hõlpsasti eristada.



3.3. Irdsektordiagramm ja ruumiline irdsektordiagramm

Irdsektordiagrammidel kuvatakse iga väärtuse osakaal kogusummas, rõhutades üksikuid väärtusi. Irdsektordiagramme saab kuvada ka ruumilisena. Irdsektordiagrammides saab muuta nii kõigi sektorite kui ka üksikute sektorite eraldamisviisi, kuid sektoreid ei saa käsitsi eraldada. Kui soovitakse sektoreid käsitsi eraldada, tuleks kasutada sektordiagrammi või ruumilist sektordiagrammi.

4. Lintdiagrammid

Töölehel veergudesse või ridadesse korraldatud andmed saab kanda lintdiagrammile. Lintdiagrammid sobivad üksikute elementide võrdluste illustreerimiseks.

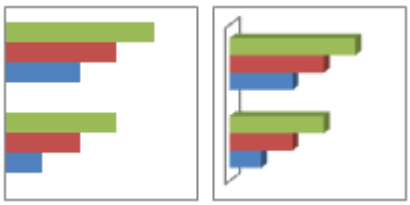
Lintdiagrammi kasutatakse järgmistel juhtudel:

- teljesildid on pikad;
- kuvatavad väärtused on millegi kestused.

Lintdiagrammidel on järgnevad alamtüübid.

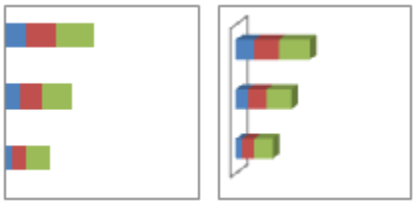
4.1. Kobarlintdiagramm ja ruumiline kobarlintdiagramm

Kobarlintdiagrammid võrdlevad väärtusi kategooriate lõikes. Kobarlintdiagrammil asuvad kategooriad tavaliselt vertikaalteljel ning väärtused horisontaalteljel. Ruumilises kobarlintdiagrammis kuvatakse ruumilisena ainult horisontaalsed ristkülikud. Andmeid kolmel teljel ei kuvata.



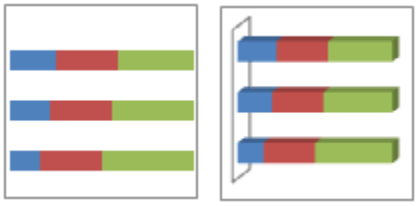
4.2. Virnlintdiagramm ja ruumiline virnlintdiagramm

Virnlintdiagrammidel kuvatakse üksikute elementide seos tervikuga. Ruumilises virnlintdiagrammis kuvatakse ruumilisena ainult horisontaalsed ristkülikud. Andmeid kolmel teljel ei kuvata.



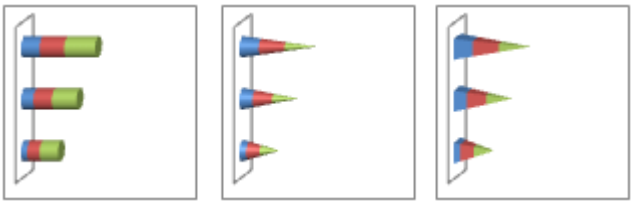
4.3. 100% virnlintdiagramm ja ruumiline 100% virnlintdiagramm

Seda tüüpi diagramm võrdleb eri kategooriate kõigi väärtuste protsentuaalset osakaalu kogusummas. Ruumilises 100% virnlintdiagrammis kuvatakse ruumilisena ainult horisontaalsed ristkülikud. Andmeid kolmel teljel ei kuvata.



4.4. Horisontaalsed silinder-, koonus- ja püramiiddiagrammid

Need diagrammid on saadaval samade kobar-, virn-, 100% virndiagrammitüüpidena kui ristkülikutega lintdiagrammid ning neil kuvatakse ja võrreldakse andmeid sama moodi. Ainuke erinevus seisneb kuvatavates kujundites (horisontaalse ristküliku asemel kuvatakse kas silinder, koonus või püramiid).

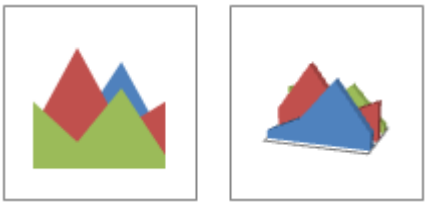


5. Kihtdiagrammid

Töölehel veergudesse või ridadesse korraldatud andmed saab kanda kihtdiagrammile. Kihtdiagrammid rõhutavad aja jooksul toimunud muutuste suurusjärku ning neid saab kasutada tähelepanu juhtimiseks kogusummade trendile. Näiteks saab andmeid, mis esindavad kasumit ajas, kanda kasumi kogusumma rõhutamiseks kihtdiagrammile. Lisaks diagrammile kantud väärtuste kogusumma kuvamisele näitab kihtdiagramm osade seost tervikuga. Kihtdiagrammidel on järgnevad alamtüübid.

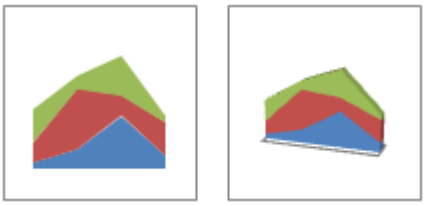
5.1. Tasapinnaline ja ruumiline kihtdiagramm

Olenemata sellest, kas andmed kuvatakse tasapinnalise või ruumilisena, kihtdiagrammidel kuvatakse väärtuste trendi ajas või muude kategooriaandmetena. Ruumilistes kihtdiagrammides kasutatakse kolme muudetavat telge (horisontaal-, vertikaal- ja sügavustelg). Reeglina tuleks mittevirnastatud kihtdiagrammi asemel kasutada joondiagrammi, kuna ühe sarja andmed võivad jääda mõne muu sarja andmete tõttu varjatuks.



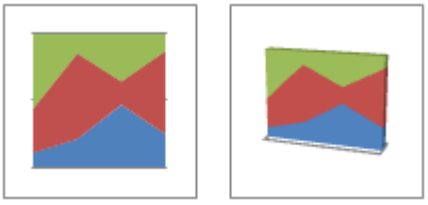
5.2. Virnkihtdiagramm ja ruumiline virnkihtdiagramm

Virnkihtdiagrammidel kuvatakse iga väärtuse osakaalu trendi ajas või muude kategooriaandmetena. Ruumilisel virnkihtdiagrammil kuvatakse samad andmed ruumilises vaates. Ruumiline vaade pole päris ruumiline diagramm, kuna kolmandat väärtustelge (sügavustelge) ei kasutata.



5.3. 100% virnkihtdiagramm ja ruumiline 100% virnkihtdiagramm

100% virnkihtdiagrammil kuvatakse iga väärtuse protsentuaalse osakaalu trend ajas või muude kategooriaandmetena. Ruumilisel 100% virnkihtdiagrammil kuvatakse samad andmed ruumilises vaates. Ruumiline vaade pole päris ruumiline diagramm , kuna kolmandat väärtustelge (sügavustelge) ei kasutata.



6. XY-diagrammid (punktdiagrammid)

Töölehel veergudesse ja ridadesse korraldatud andmed saab kanda XY-diagrammile (punktdiagrammile). Punktdiagrammides kuvatakse mitme andmesarja arvväärtuste seosed või kantakse diagrammile kaks arvude rühma ühe x- ja y-koordinaatide sarjana.

Punktdiagrammil on kaks väärtustelge. 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 (nt teaduslike, statistiliste ja tehniliste andmete puhul).

Punktdiagrammi kasutatakse järgmistel juhtudel:

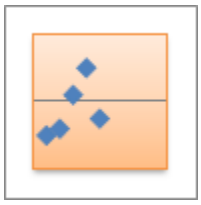
- horisontaalteljele paigutatavad väärtused pole ühtlaste vahedega
- horisontaalteljel on palju andmepunkte
- töölehe andmete (sh andmepaaride või rühmitatud andmekomplektide) tõhusamaks kuvamiseks soovitatakse kohandada punktdiagrammi sõltumatuid skaalasid, et anda rühmitatud andmete kohta rohkem teavet
- andmepunktide erinevuse kuvamise asemel soovitakse tuua välja sarnasusi suurte andmekomplektide vahel
- Soovitakse võrrelda mitmeid andmepunkte ilma ajalise tegurita – mida rohkem andmepunkte punktdiagrammile kantakse, seda paremini saate neid võrrelda

Töölehe andmete korraldamiseks punktdiagrammile kandmise jaoks peaksite x-telje väärtused paigutama ühte ritta või veergu ning seejärel sisestama vastavad y-telje väärtused külgnevatesse ridadesse või veergudesse.

Punktdiagrammidel on järgnevad alamtüübid.

6.1. Tähistega punktdiagramm

Seda tüüpi diagramm võrdleb väärtuste paare. Kasutage punktdiagrammi andmetähistega, kuid ilma joonteta. Suure hulga andmepunktide ja ühendusjoonte kasutamine muudab andmed raskesti loetavaks. Seda tüüpi diagrammi saate kasutada ka juhul, kui te ei pea andmepunktidevahelist ühendust näitama.



6.2. Sujuvjoontega punktdiagramm ning tähiste ja sujuvjoontega punktdiagramm

Seda tüüpi diagrammil kuvatakse andmepunkte ühendav sujuvjoon. Sujuvjooni saab kuvada koos tähistega või ilma. Kasutage tähisteta punktdiagrammi siis, kui andmepunkte on palju.

6.3. Sirgjoontega punktdiagramm ning sirgjoonte ja tähistega punktdiagramm

Seda tüüpi diagrammis kuvatakse andmepunkte ühendavad sirgjooned. Neid jooni saab kuvada tähistega või ilma.

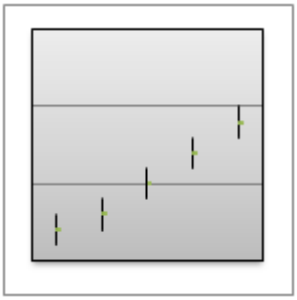
7. Börsidiagrammid

Töölehel veergudesse või ridadesse korraldatud andmed saab kanda börsidiagrammile. Nagu nimi viitab, kasutatakse börsidiagrammi enamasti aktsiahindade kõikumise illustreerimiseks. Seda saab aga kasutada ka teaduslike andmete puhul. Näiteks võite börsidiagrammi kasutada päeva- või aastatemperatuuride kõikumiste näitamiseks. Börsidiagrammide loomiseks peavad andmed olema õiges järjestuses.

See, kuidas börsidiagrammi andmed on töölehel korraldatud, on väga oluline. Näiteks lihtsa kõrge-madala-sulgemishinna börsidiagrammi loomiseks peaksid andmed olema korraldatud järjestikustesse veergudesse päistega Kõrge, Madal ja Sulgemishind. Börsidiagrammidel on järgnevad alamtüübid.

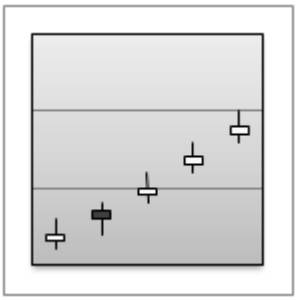
7.1. Kõrge-madal-sulgemishind

Selliseid börsidiagramme kasutatakse enamasti aktsiahindade illustreerimiseks. Selle jaoks on vaja kolme väärtussarja järgmises järjestuses: kõrge, madal ning sulgemishind.



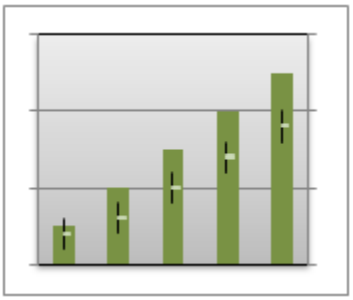
7.2. Avamis-kõrge-madal-sulgemishind

Seda tüüpi börsidiagrammi jaoks on vaja nelja õiges järjestuses väärtussarja (avamis-, kõrge, madal ja sulgemishind).



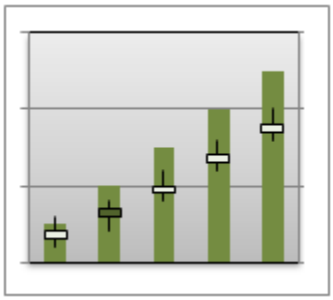
7.3. Maht-kõrge-madal-sulgemishind

Seda tüüpi börsidiagrammi jaoks on vaja nelja õiges järjestuses väärtussarja (maht, kõrge, madal ja sulgemishind). Diagrammil kuvatakse mõõdetud maht, kasutades kahte väärtustelge: ühte telge mõõdetud mahu veergude ja teist aktsiahindade jaoks.



7.4. Maht-avamis-kõrge-madal-sulgemishind

Seda tüüpi börsidiagrammi jaoks on vaja viite õiges järjestuses väärtussarja (maht, avamis-, kõrge, madal ja sulgemishind).

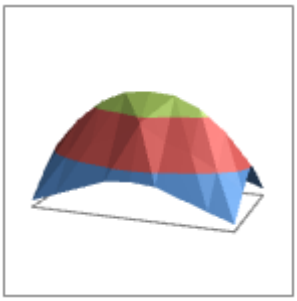


8. Pinddiagrammid

Töölehel veergudesse või ridadesse korraldatud andmed saab kanda pinddiagrammile. Pinddiagramm on vajalik siis, kui soovite leida kahe andmekogumi vahelisi optimaalseid kombinatsioone. Nii nagu topograafilisel kaardil, näitavad värvid ja mustrid alasid, mis on samas väärtustevahemikus. Pinddiagrammi saate kasutada siis, kui nii kategooriad kui ka andmesarjad on arvväärtused. Pinddiagrammidel on järgnevad alamtüübid.

8.1. Ruumiline pinddiagramm

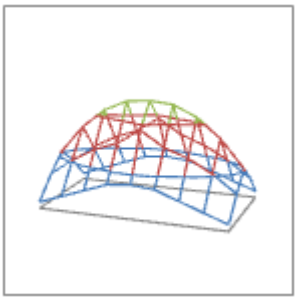
Ruumiline pinddiagramm kuvab väärtuste trendid kahe mõõtme lõikes ühe pideva kõverana. Pinddiagrammi värviribad ei tähista andmesarju, vaid erinevusi väärtuste vahel. Sellel diagrammil kuvatakse andmete ruumiline vaade, mis näeb välja nagu üle ruumilise tulpdiagrammi venitatud kummilina. Seda kasutatakse tavaliselt suurte andmehulkade vaheliste suhete näitamiseks, mida muidu oleks raske vaadata.



8.2. Ruumiline sõrestikpinddiagramm

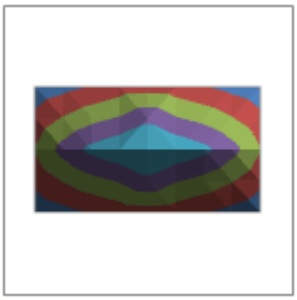
Ruumilist pinddiagrammi, mille pinnal värve pole, nimetatakse ruumiliseks sõrestikpinddiagrammiks. Sellel diagrammil kuvatakse ainult jooned. Ruumilist pinddiagrammi, mille ühelgi pinnal pole värve, nimetatakse ruumiliseks sõrestikpinddiagrammiks. Sellel diagrammil kuvatakse ainult jooned.

Ruumiline sõrestikdiagramm on raskesti loetav, kuid seda tüüpi diagrammi on mõistlik kasutada diagrammile andmete kiiremaks kandmiseks või suurte andmekogumite korral.



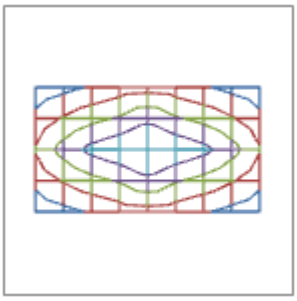
8.3. Kontuurdiagramm

Kontuurdiagrammid sarnanevad pealtvaates tasapinnaliste topograafiliste kaartidega, mille värviribad tähistavad mingeid kindlaid väärtusvahemikke. Kontuurdiagrammi jooned ühendavad võrdse väärtusega interpoleeritud punkte.



8.4. Sõrestikkontuurdiagramm

Sõrestikkontuurdiagrammid on ka pealtvaates pinddiagrammid. Sõrestikkontuurdiagrammi pind on värviribadeta ning seal kuvatakse ainult jooned. Sõrestikkontuurdiagramm on raskesti loetav. Selle asemel võiksite kasutada ruumilist pinddiagrammi.



9. Rõngasdiagrammid

Töölehel ainult veergudesse või ridadesse korraldatud andmed saab kanda rõngasdiagrammile. Nagu sektordiagrammgi, näitab rõngasdiagramm osade seost tervikuga, kuid võib sisaldada rohkem kui ühte andmesarja. Rõngasdiagrammidel on järgnevad alamtüübid.

9.1. Rõngasdiagramm

Rõngasdiagrammidel kuvatakse andmed rõngastena, kusjuures iga rõngas tähistab ühte andmesarja. Kui protsendid kuvatakse andmesiltidel, moodustavad diagrammi rõngad kokku 100%.

9.2. Irdrõngasdiagramm

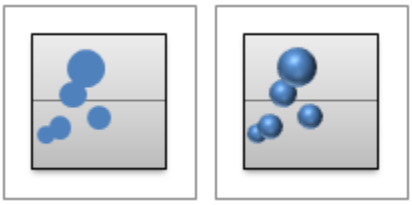
Sarnaselt irdsektordiagrammiga kuvatakse irdrõngasdiagrammil iga väärtuse osakaal kogusummas, rõhutades üksikuid väärtusi, kuid irdrõngasdiagrammid võivad sisaldada mitut andmesarja.

10. Mulldiagrammid

Mulldiagrammile saab kanda andmed, mis on töölehel veergudesse korraldatud nii, et x-telje väärtused on loendatud esimeses veerus ja neile vastavad y-telje väärtused ja mulli suuruse väärtused külgnevates veergudes. Näiteks saate andmed korraldada nii nagu järgmises näites. Mulldiagrammidel on järgnevad alamtüübid.

10.1. Mulldiagramm või ruumilise efektiga mulldiagramm

Mõlemat tüüpi diagrammid võrdlevad kahe väärtustekogumi asemel kolme. Kolmas väärtus määrab mulltähise suuruse. Saate valida, kas mullid kuvatakse tasapinnaliste või ruumilistena.



11. Radiaaldiagrammid

Töölehel veergudesse või ridadesse korraldatud andmed saab kanda radiaaldiagrammile. Radiaaldiagrammid võrdlevad mitmete andmesarjade kokkuvõtteväärtusi. Radiaaldiagrammidel on järgnevad alamtüübid.

11.1. Radiaaldiagramm ja tähistega radiaaldiagramm

Radiaaldiagrammidel kuvatakse andmete muutused keskpunkti suhtes, kas siis koos andmepunktide tähistega või ilma.


11.2. Täidetud radiaaldiagramm

Täidetud radiaaldiagrammis täidetakse andmesarja kaetud ala värviga.

12. Kokkuvõtteks

Diagramm	Andmete korraldamine
Tulp-, lint-, joon-, kiht-, pind- või radiaaldiagramm	Veergudes või ridades. Nt. Lorem Ipsum 1 2 3 4 Või Lorem 1 3 Ipsum 2 4
Sektor- või rõngasdiagramm	Ühe andmesarja (andmesari: seostuvad andmepunktid, mis on kantud diagrammile. Igal diagrammi andmesarjal on ainuvärv või -muster, mida kirjeldatakse diagrammi legendis. Diagrammile saab kanda ühe või mitu andmesarja. Sektordiagrammis saab olla ainult üks andmesari.) korral ühes andmeveerus või -reas ja ühes andmesiltide veerus või reas. Nt. A 1 B 2 C 3 Või A B C 1 2 3 Mitme andmesarja korral mitmes andmeveerus või -reas ja ühes andmesiltide veerus või reas. Nt.

<table>
<tr><td></td><td>
<table>
<tr><td>A</td><td>1</td><td>2</td></tr>
<tr><td>B</td><td>3</td><td>4</td></tr>
<tr><td>C</td><td>5</td><td>6</td></tr>
</table>
Või
<table>
<tr><td>A</td><td>B</td><td>C</td></tr>
<tr><td>1</td><td>2</td><td>3</td></tr>
<tr><td>4</td><td>5</td><td>6</td></tr>
</table>
</td></tr>
<tr><td>XY- (punktdiagramm) või mulldiagramm</td><td>X-telje väärtused esimeses veerus ning vastavad y-telje väärtused ja mulli suuruse väärtused külgnevates veergudes. Nt.
<table>
<tr><td>x</td><td>y</td><td>Mulli suurus</td></tr>
<tr><td>1</td><td>2</td><td>3</td></tr>
<tr><td>4</td><td>5</td><td>6</td></tr>
</table>
</td></tr>
<tr><td>Börsidiagramm</td><td>Veergudes või ridades järgmises järjekorras (kasutades nimesid või kuupäevi siltidena):
suured väärtused, väiksed väärtused ja lõppväärtused.
Nt.
<table>
<tr><td>Kuupäev</td><td>Suur</td><td>Väike</td><td>Lõpp</td></tr>
<tr><td>1/1/2002</td><td>46,125</td><td>42</td><td>44,063</td></tr>
</table>
Või
<table>
<tr><td>Kuupäev</td><td>1/1/2002</td></tr>
<tr><td>Suur</td><td>46,125</td></tr>
<tr><td>Väike</td><td>42</td></tr>
<tr><td>Lõpp</td><td>44,063</td></tr>
</table>
</td></tr>
</table>

Statistikafunktsioonid

Funktsioon	Kirjeldus
AVEDEV	Annab vastuseks andmepunktide keskmise absoluuthälbe keskväärtuse põhjal.
AVERAGE	Annab vastuseks oma argumentide keskväärtuse.
AVERAGEA	Annab vastuseks oma argumentide keskväärtuse, k.a arv-, teksti- ja loogikaväärtused.
AVERAGEIF	Annab vastuseks kõigi mitmele kriteeriumile vastavas vahemikus olevate lahtrite keskmise (aritmeetilise keskmise).
AVERAGEIFS	Annab vastuseks kõigi mitmele kriteeriumile vastavate lahtrite keskmise (aritmeetilise keskmise).
BETADIST	Annab vastuseks beetajaotuse tihedusfunktsiooni väärtuse.

BETAINV	Annab vastuseks beetajaotuse tihedusfunktsiooni pöördfunktsiooni väärtuse.
BINOMDIST	Annab vastuseks üksikliikme binoomjaotuse tõenäosuse.
CHIDIST	Annab vastuseks X2-jaotuse tõenäosuse ühepoolse piiranguga tõenäosuse.
CHIINV	Annab vastuseks X2-jaotuse ühepoolse piiranguga tõenäosuse pöördfunktsiooni väärtuse.
CHITEST	Annab vastuseks sõltumatusetesti tulemuse.
CONFIDENCE	Annab vastuseks valimikeskmise usaldusvahemiku.
CORREL	Annab vastuseks kahe andmekogumi korrelatsioonikordaja.
COUNT	Loendab argumentide loendis olevaid arve.
COUNTA	Loendab argumentide loendis olevaid väärtusi.
COUNTBLANK	Loendab vahemiku tühjad lahtrid.
COUNTIF	Loendab antud kriteeriumidele vastava vahemiku lahtrite arvu.
COUNTIFS	Loendab mitmetele kriteeriumidele vastava vahemiku lahtrite arvu.
COVAR	Annab vastuseks kovariatsiooni, andmepunktipaaride hälvete korrutiste keskmise.
CRITBINOM	Annab vastuseks väikseima väärtuse, mille puhul on kumulatiivne binoomjaotus väiksem või võrdne kriteeriumi väärtusega.
DEVSQ	Annab vastuseks hälvete ruutude summa.
EXPONDIST	Annab vastuseks eksponentjaotuse.
FDIST	Annab vastuseks F-tõenäosuse jaotuse.
FINV	Annab vastuseks F-tõenäosuse pöördjaotuse.
FISHER	Annab vastuseks Fisheri teisenduse.
FISHERINV	Annab vastuseks Fisheri pöördteisenduse.
FORECAST	Annab vastuseks väärtuse, eeldades lineaarset trendi.
FREQUENCY	Annab vastuseks andmete esinemissageduse jaotuse vertikaalse massiivina.
FTEST	Annab vastuseks F-testi tulemi.
GAMMADIST	Annab vastuseks gammajaotuse väärtuse.
GAMMAINV	Annab vastuseks gammajaotuse jaotusfunktsiooni pöördfunktsiooni väärtuse.
GAMMALN	Annab vastuseks gammafunktsiooni naturaallogaritmi (Γ(x)).
GEOMEAN	Annab vastuseks geomeetrilise keskmise.
GROWTH	Annab vastuseks väärtused, eeldades eksponentsiaalset trendi.

HARMEAN	Annab vastuseks harmoonilise keskmise.
HYPGEOMDIST	Annab vastuseks hüpergeomeetrilise jaotuse.
INTERCEPT	Annab vastuseks lineaarse regressioonisirge algordinaadi.
KURT	Annab vastuseks andmekogumi ekstsessi.
LARGE	Annab vastuseks suuruselt k-nda väärtuse andmehulgas.
LINEST	Annab vastuseks lineaarse trendi parameetrid.
LOGEST	Annab vastuseks eksponentsiaalse trendi parameetrid.
LOGINV	Annab vastuseks logaritmilise normaaljaotuse jaotusfunktsiooni pöördfunktsiooni.
LOGNORMDIST	Annab vastuseks logaritmilise normaaljaotuse jaotusfunktsiooni.
MAX	Annab vastuseks argumentide loendi suurima väärtuse.
MAXA	Annab vastuseks argumentide loendi suurima väärtuse, k.a arv-, teksti- ja loogikaväärtused.
MEDIAN	Annab vastuseks antud arvude mediaani.
MIN	Annab vastuseks argumentide loendi väikseima väärtuse.
MINA	Annab vastuseks argumentide loendi väikseima väärtuse, k.a arv-, teksti- ja loogikaväärtused.
MODE	Annab vastuseks andmekogumi kõige enam esineva väärtuse.
NEGBINOMDIST	Annab vastuseks negatiivse binoomjaotuse.
NORMDIST	Annab vastuseks normaaljaotuse jaotusfunktsiooni.
NORMINV	Annab vastuseks normaaljaotuse jaotusfunktsiooni pöördfunktsiooni väärtuse.
NORMSDIST	Annab vastuseks normaliseeritud normaaljaotuse jaotusfunktsiooni väärtuse.
NORMSINV	Annab vastuseks normaliseeritud normaaljaotuse jaotusfunktsiooni pöördfunktsiooni väärtuse.
PEARSON	Annab vastuseks Pearsoni korrelatsioonikordaja.
PERCENTILE	Annab vastuseks vahemiku väärtuste k-nda protsentiili.
PERCENTRANK	Annab vastuseks väärtuse protsentuaalse asukoha andmekogumis.
PERMUT	Annab vastuseks antud objektide arvu permutatsioonide arvu.
POISSON	Annab vastuseks Poissoni jaotuse väärtuse.
PROB	Annab vastuseks tõenäosuse, mil vahemiku väärtused on kahe piirväärtuse vahel.
QUARTILE	Annab vastuseks andmekogumi kvartiili.
RANK	Annab vastuseks arvu asukoha arvuloendis.

RSQ	Annab vastuseks Pearsoni korrelatsioonikordaja ruudu.
SKEW	Annab vastuseks jaotuse asümmeetriakordaja.
SLOPE	Annab vastuseks lineaarse regressioonisirge tõusu.
SMALL	Annab vastuseks väiksuselt k-nda väärtuse andmehulgas.
STANDARDIZE	Annab vastuseks normaliseeritud väärtuse.
STDEV	Arvutab valimi põhjal standardhälbe.
STDEVA	Arvutab valimi põhjal standardhälbe, k.a arv-, teksti- ja loogikaväärtused.
STDEVP	Arvutab standardhälbe kogu populatsiooni alusel.
STDEVPA	Arvutab standardhälbe kogu populatsiooni alusel, k.a arv-, teksti- ja loogikaväärtused.
STEYX	Annab vastuseks prognoositud y-väärtuse standardvea igale x-le regressioonis.
TDIST	Annab vastuseks Studenti t-jaotuse.
TINV	Annab vastuseks Studenti t-jaotuse pöördfunktsiooni väärtuse.
TREND	Annab vastuseks väärtused, eeldades lineaartrendi.
TRIMMEAN	Annab vastuseks andmekogumi ahendi keskväärtuse.
TTEST	Annab vastuseks Studenti t-testiga seotud tõenäosuse.
VAR	Annab vastuseks hinnangulise dispersiooni valimi alusel.
VARA	Annab vastuseks hinnangulise dispersiooni valimi alusel, k.a arv-, teksti- ja loogikaväärtused.
VARP	Arvutab dispersiooni terve populatsiooni alusel.
VARPA	Arvutab dispersiooni kogu populatsiooni alusel, k.a arv-, teksti- ja loogikaväärtused.
WEIBULL	Annab vastuseks Weibulli jaotuse.
ZTEST	Annab vastuseks ühepoolse piiranguga z-testi tõenäosusväärtuse.

Praktikumiülesanne

26.02.2010

I OSA: KIRJELDAV STATISTIKA JA JOONISED EXCELIS

1. Mis tüüpi andmestikuga on tegu (vt slaid „Andmete korraldamise viise“)?
2. Kodeerige andmed järgmiselt:

- sugu (1 – m; 2 – n)
- haridus (1 – põhi; 2 – üldkesk; 3 – keskeri; 4 – kõrg)
- sissetulek (1 – alla 5000; 2 – 5000-6999; 3 – 7000-9999; 4 – 1000-15000; 5 – üle 15000)

- külastamise sagedus (1 – harvem kui 1 kord nädalas; 2 – 1 kord nädalas; 3 – 2-4 korda nädalas; 4 – enam kui 4 korda nädalas)
- transpordivahend (1 – jalgsi; 2 – buss; 3 – auto).

3. Määratlege iga muutuja tüüp (nominaal, järjestus- või arvuline tunnus).
4. Otsustage, millised järgmistest kirjeldavatest statistikutest annavad iga muutuja korral sisukat informatsiooni ja leidke vastavad suurused, kasutades Exceli statistilisi funktsioone:

mediaan
mood
keskmine
dispersioon
standardhälve
variatsioonikoefitsient
variantsiooniamplituud
… maksimaalne
… minimaalne

5. Tehke sagedustabelid järgmiste tunnuste lõikes:

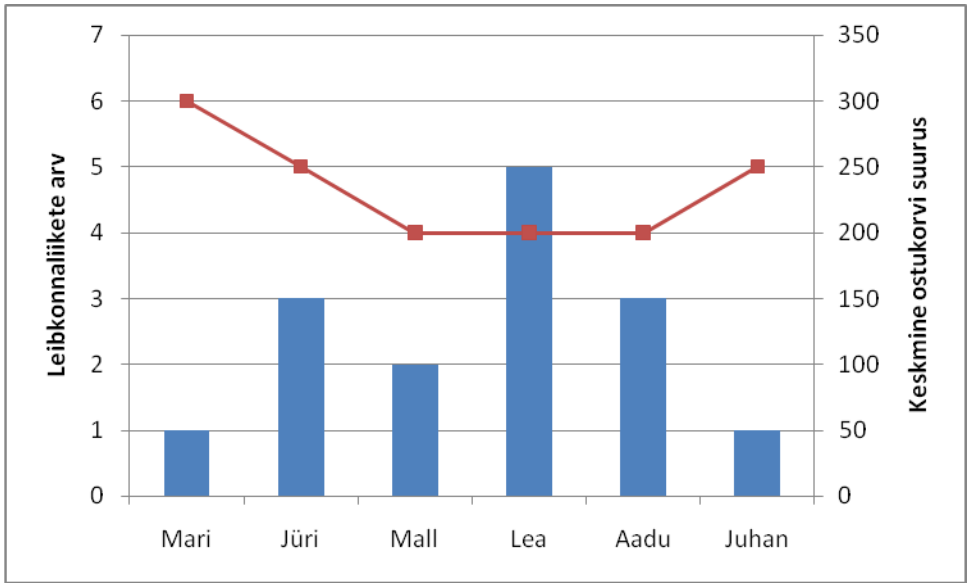
- sugu
- leibkonna suurus
- haridus
- sissetulek
- külastamise sagedus
- transpordivahend.

6. Tehke iga p-s 5 nimetatud muutuja lõikes joonis, valides enda arvates kõige paremini tulemusi illustreeriv joonise tüüp (tulp-, joon-, sektor- või lintdiagramm).

7. Tuginedes 7 esimese valimisse kaasatud inimese andmetele, tehke liitdiagramm, kus tulpdiagrammina oleks kujutatud leibkonna suurus ja joonena keskmine ostukorvi suurus:

- lisage paremale samuti skaala (0-350)
- vahendage joonena esitatud keskmise ostukorvi suuruse näitaja „punkti“ suurust
- muutke vasakut skaalat nii, et see varieeruks vahemikus 1-12
- lisage skaaladele nimetused (vasak skaala: „leibkonnaliikmete arv“; parem skaala: „keskmine ostukorvi suurus“)

  Lõpptulemus võiks olla selline:



II OSA: EUROSTAT

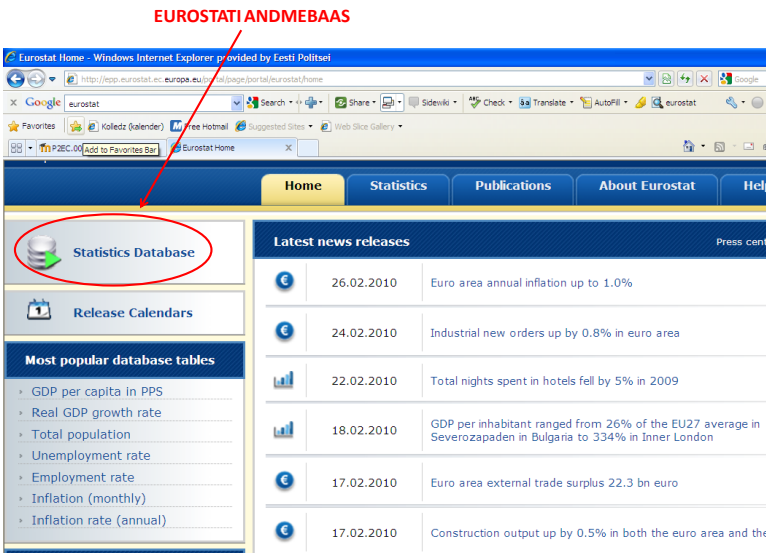
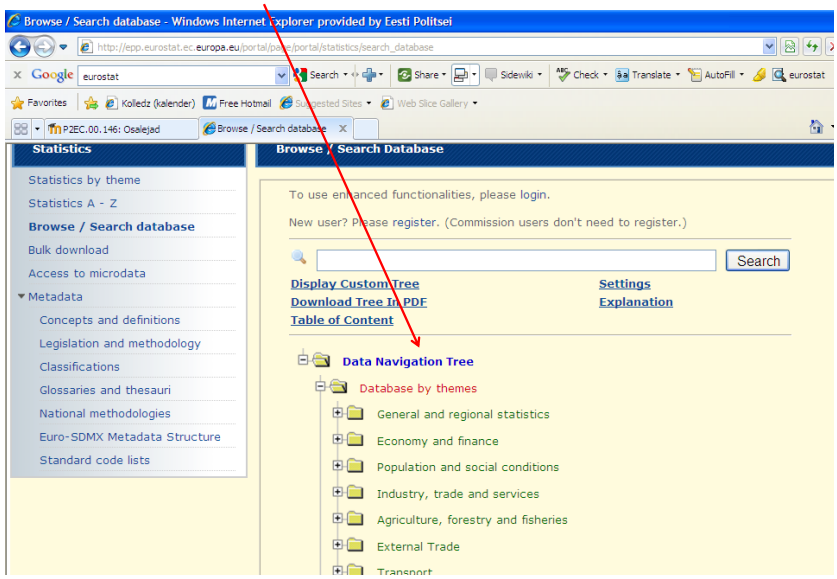
1. Link Eurostati veebilehele: http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/
2. Vaatame järgmisi sektsioone:
   - Most populaar database tables (NB! Infot saab vaid ühe klikiga kuvada ka graafikutel (sektsioon „Graph"; saab valida, milliseid andmeid soovitakse graafikul kuvada – vt sektsioon „Data"; andmeid saab sorteerida, kasutada riikide lühendeid – vt sektsioon „Sort and label" jne)
   - Country profiles (saab kiirülevaate riigi olulisematest näitajatest; saab kaht riiki võrrelda jne)
   - Sektsioon „Statistics" (eeldefineeritud tabelid)
3. Leidke ja kopeerige Eurostati andmebaasist Excelisse andmed kolme näitaja kohta.
4. Tehke andmeid illustreerivad joonised (sh üks liitdiagramm, üks aegridu ja üks riikidevahelisi erinevusi kajastav diagramm), leidke asjakohased kirjeldavate statistikute väärtused. Tõlgendage tulemusi.

Kodutöö 1 (max 10 punkti)

TUTVUMINE EUROOPA STATISTIKAAMETI EUROSTAT ANDMEBAASIGA JA KIRJEDAVA STATISTILISE ANALÜÜSI TEGEMINE

Eesmärk: tudengil on ülevaade Eurostati andmebaasi struktuurist, ta oskab iseseisvalt leida endale huvipakkuva valdkonna kohta vajalikud andmed,neid analüüsida ja tulemusi tõlgendada .

Minge Eurostati veebilehele http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home

Tutvu Eurostati andmebaasiga…	
… kasutades andmete navigeerimispuud (rohkem infot navigatsioonipuu kohta leiad failist „**Eurostati navigeerimispuu**“	

<u>Kodutöö ülesanded</u>:

1. Leia Eurostatist vähemalt viis Sinu lõputöö teema aspektist huvipakkuvat näitajat.
2. Kopeeri andmed punktis 1 nimetatud näitajate kohta EL-27 riikides (ilmtingimata ei ole vajalik, et andmed oleksid olemas kõikide ELi liikmesriikide kohta, kuid mida suurema arvu riikide kohta on andmed olemas, seda parem) vähemalt ühel aastal (võid kasutada ka enam kui ühe aasta näitajaid) Excelisse.
3. Mis tüüpi andmebaasiga (vt 1. loengu slaid „andmete korraldamise viise“) on tegemist?

Kas tegemist on üldkogumi või valimi andmetega (*näpunäide: sellest lähtuvalt tuleb valida punktis 5 õiged Exceli käsud, nt valida, kas standardhälbe käsk on STDEV või STDEVP*)?

Mis tüüpi andmetega (nominaal-, järjestus-, arv-) on tegemist (*näpunäide: ära unusta, et ka riikide nimetus ja aasta on muutujad!*)?

4. Püstita andmetest lähtuvalt uurimisküsimus (uurimisküsimus tuleks sõnastada nii, et töö lugeja saaks aru, mida ja mis eesmärgil uurija analüüsib).
5. Leia valitud andmete korral järgmised kirjeldavad statistikud (rohkem infot statistikute sisu kohta leiad vajadusel siit: www.mtk.ut.ee/doc/SSloeng2.doc), lähtudes sellest, millised annavad selle muutuja kohta sisukat informatsiooni:

- keskmised: aritmeetiline keskmine, mediaan, mood;
- variatsiooninäitarvud: variatsiooniamplituud, standardhälve;
- jaotuse karakteristikud: asümmeetriakordaja, ekstsess.

6. Tee tulemuste illustreerimiseks joonised (iga muutuja korral vähemasti üks), kasuta erinevaid joonise tüüpe (vt 1. loengu slaide ja Excelis jooniste tegemise konspekti).
7. Kirjuta tulemustest kokkuvõte (mitte vähem kui 3 lk), tuues ära kirjeldavad statistikud, tõlgendades neid ja illustreerides tulemusi joonistel.

NB! OLULINE!!!

Kodutöö tähtaeg on **16. märts**. Iga esitamisega viivitatud päeva eest kaotad ühe punkti, kuid **pärast 21. märtsi esitatud kodutöid ei aktsepteerita**.

Kodutöö võib teha **üksi** või **kahekesi**.

Kodutööd ootan **sisulist analüüsi**, st ei piisa sellest, kui kirjeldavad statistikud välja arvutada, neid peaks ka sisukalt tõlgendama, sama kehtib jooniste kohta.

Kodutööga koos palun esitada ka **Exceli faili**, kus on olemas kodutöös kasutatud **kirjeldavate statistikute arvutused** ja **joonised**.

Kui hätta jääd, ära kõhkle nõu küsimast (kerly.krillo@ut.ee).

Head statistika maailma avastamist!

II LOENG

Liigendtabelite tegemine Excelis

Sissejuhatuseks

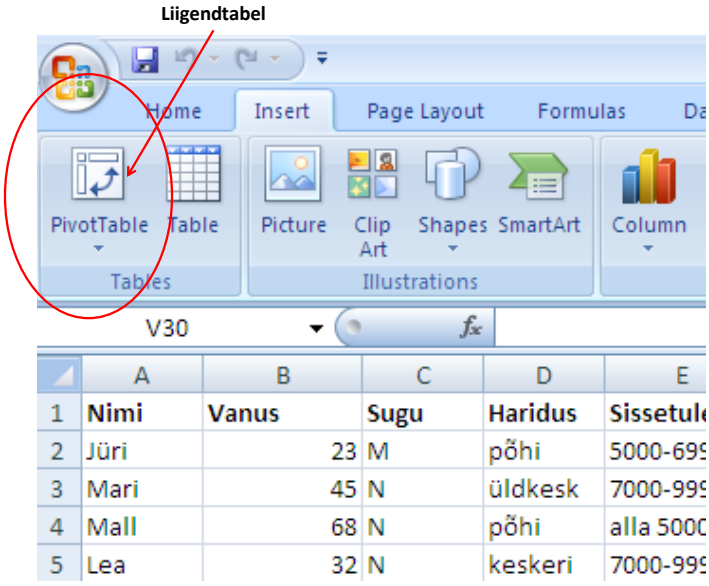
Liigendtabelid on Exceli funktsioon, mille abil saab andmeid hõlpsalt korraldada, süstematiseerida ja analüüsida. Eriti palju on liigendtabeli loomise võimalusest abi suurte andmemassiivide analüüsimiseks, kuigi loomulikult on see hea andmeanalüüsi vahend ka väiksemamahuliste andmebaaside korral.

Liigendtabelite abil saab vaid paari klikiga luua erinevaid tabeleid, teha jooniseid ja arvutada täiendavaid muutujaid. Ühe etteantud vormi asemel saab liigendtabelitega hõlpsalt luua vaid loetud sekundite jooksul üha uusi tabeleid, st on võimalik paindlikult teha just selliseid tabeleid/jooniseid, nagu uurija vajab.

Alustamine

Viisardi abil on võimalik defineerida, milliseid andmeid liigendtabelis kasutada soovitakse.

MS Exceli versioonis 2007 tuleb valida menüüst „Insert“ → „PivotTable“

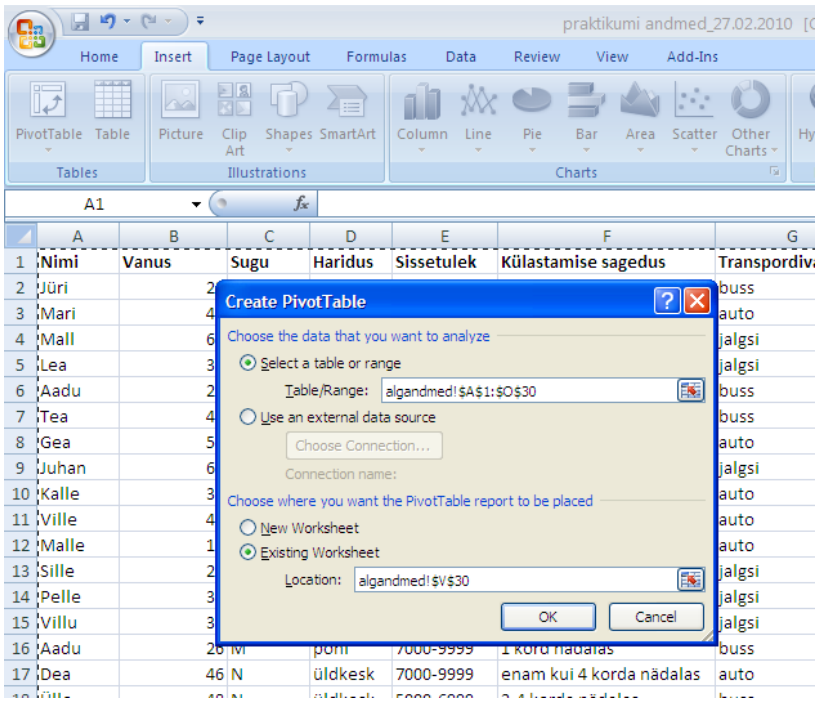


	A	B	C	D	E
1	Nimi	Vanus	Sugu	Haridus	Sissetule
2	Jüri	23	M	põhi	5000-699
3	Mari	45	N	üldkesk	7000-999
4	Mall	68	N	põhi	alla 5000
5	Lea	32	N	keskeri	7000-999

Excel 2007 avaneb seejärel dialoogiaken, kus tuleb teha vajalikud valikud – määratleda loodava liigendtabeli

1) andmete piirkond kas Exceli töölehel („Table/Range“) või välisest allikast („Use an external data source“) ja
2) paigutusala – valida saab, kas paigutada tulemus uuele töölehele („New Worksheet“) või samale töölehele („Existing Worksheet“), viimasel juhul tuleb märkida ka paigutuspiirkond (Location).

Excel 2007 on liigendtabeli viisard järgmine:



	A	B	C	D	E	F	G
1	Nimi	Vanus	Sugu	Haridus	Sissetulek	Külastamise sagedus	Transpordiv
2	Jüri	2					buss
3	Mari	4					auto
4	Mall	6					jalgsi
5	Lea	3					jalgsi
6	Aadu	2					buss
7	Tea	4					buss
8	Gea	5					auto
9	Juhan	6					jalgsi
10	Kalle	3					auto
11	Ville	4					auto
12	Malle	1					auto
13	Sille	2					jalgsi
14	Pelle	3					jalgsi
15	Villu	3					jalgsi
16	Aadu	26	M	põhi	7000-9999	1 kord nädalas	buss
17	Dea	46	N	üldkesk	7000-9999	enam kui 4 korda nädalas	auto

NB! Pea meeles. Liigendtabeli aruande koostamisel kasutatavad andmeid ei tohi sisaldada tühje veerge, samuti pole hea, kui piirkonnas on tühje ridu. Seega tühjad read/veerud, mis on mõeldud näiteks üht andmerühma teisest, tuleks enne liigendtabeli koostamist kustutada.

Excel 2003 viisard on samm-sammuline, kuid suuresti analoogiline. Liigendtabeli loomiseks MS Excel versioonis 2003 tuleb valida menüüst „Data"→"PivotTable and PivotChart Report". Vanevas viisardis tuleb teha järgmised valikud:

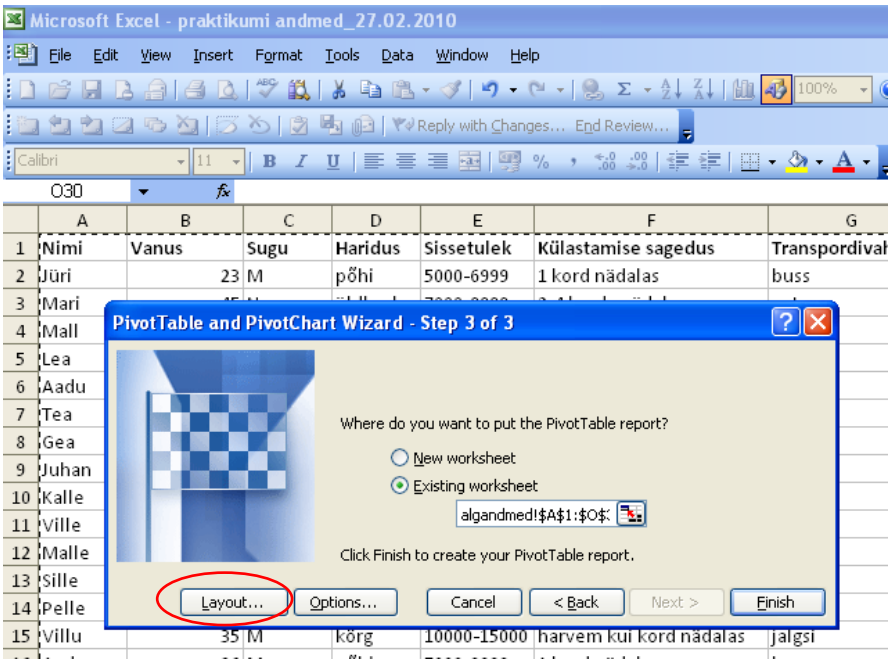
1. sammuna määrata

- andmete asukoht, mida soovitakse analüüsis kasutada („Where is the data that you want to analyze")
- defineerida, mis tüüpi raportit soovitakse luua (kas liigendtabel või liigendjoonis, *seda võimalust Excel 2007viisardis ei paku*);

2. sammuna määratleda andmete asukoht töölehel;

3. sammuna määrata, kuhu tabel/joonis paigutatakse („Where do you want to put the PivotTable report", *juhul, kui luuakse tabel*).

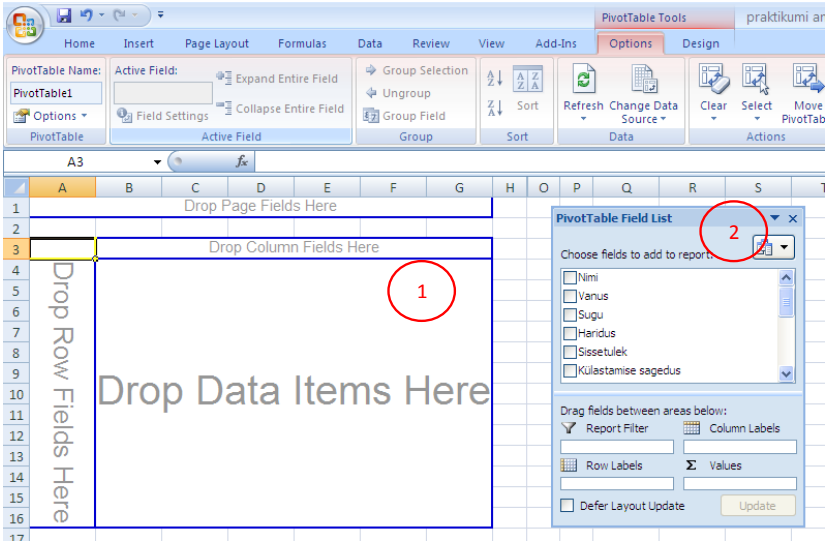
NB! Excel 2003 viisardi viimases aknas avaneb ikooni „Layout" (alloleval joonisel tähistatud punase ringiga**) vajutamisel dialoogiaken, kus saab luua liigendtabeli, lohistades huvipakkuvad muutujad vajalikele väljadele. Samas ei pea seda tegema kohe liigendtabeli loomisel, liigendtabeli võib muutujatega täita ka siis, kui esialgne liigendtabel on loodud.**



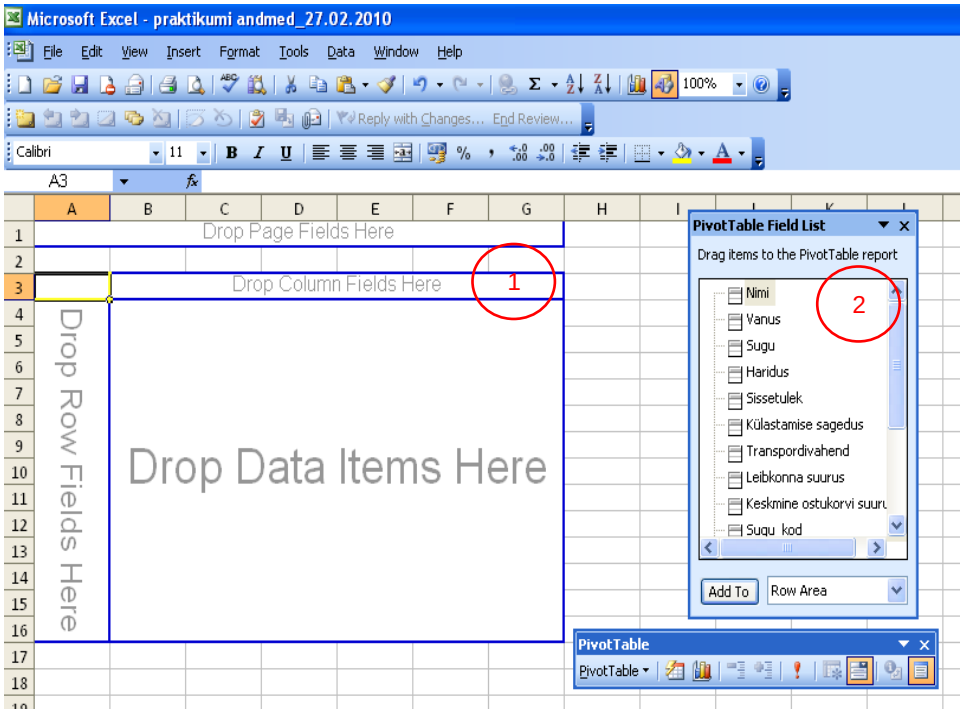
Liigendtabeli osad

Pärast liigendtabeli viisardis vajalike otsuste tegemist avaneb liigendtabeli vaade, kus vasakul on **liigendtabeli aruande paigutusala** (joonisel tähistatud 1-ga) ning paremal **liigendtabeli väljade loend** (joonisel tähistatud 2-ga). Viimases kuvatakse kõik väljad, mida saab liigendtabeli aruande koostamisel kasutada. Kuna väljade nimetused tulevad lähteandmete veerunimetustest, siis **on kindlasti vajalik, et igal veerul, mida liigendtabelis kasutatakse, oleks töölehel pealkiri**.

Excel 2007 on kuvatav vaade selline:



Excel 2003 on vaade järgmine:



Liigendtabeli aruande paigutusalal on neli sektsiooni:

1. read (ala nimetusega „Drop Row Fields Here“)
2. veerud (ala nimetusega „Drop Column Fields Here“)
3. andmed (ala nimetusega „Drop Data Fields Here“)
4. leheküljeala, mille abil on võimalik andmeid sorteerida (ala nimetusega „Drop Page Fields Here“).

Liigendtabeli aruande koostamine

Liigendtabeli aruande koostamiseks tuleb võtta vajalik muutuja liigendtabeli väljade loendist ja viia see vajalikule väljale paigutusalal. Seda saab teha, võttes hiire vasaku klahviga vajalikust muutujast ja lohistades see vajalikule väljale. Juhul, kui soovitakse aruannet muuta, tuleb võtta vajalikust muutujast paigutusalal ja lohistada see paigutusalalt välja. Kui klõpsata liigendtabeli paigutusalast väljaspool, kaob väljaloend. See tekib uuesti, kui klõpsata paigutusalal või aruandel.
NB! Ritta/veergu võib paigutada ka mitu muutujat. Leheküljealale on otstarbekas paigutada muutuja, mille abil soovitakse tulemusi filtreerida. Rea- ja veerualalt paigutatakse üldjuhul kategoorilised muutujad, andmete alale pidevad muutujad.

NÄIDE: moodustame tabeli, mis kajastaks soo ja hariduse lõikes keskmist sissetulekut.

Selleks tuleb:

1. lohistada soo ja hariduse muutujad ritta/veergu (pole sisulist vahet, kumb kuhu lohistada). Excel 2007 on väljaloendi osas „Drag fields between areas below“ näha, kuhu milline

muutuja on paigutatud.

NB! Excel 2007 on võimalik muutujad vedada ka mitte paigutusalale, vaid vajalikele väljadele (veerg, rida, andmed, filter) väljaloendi osas „Drag fields between areas below“.

2. Lohistada muutuja nimetus „Sissetulek“ andmete väljale.

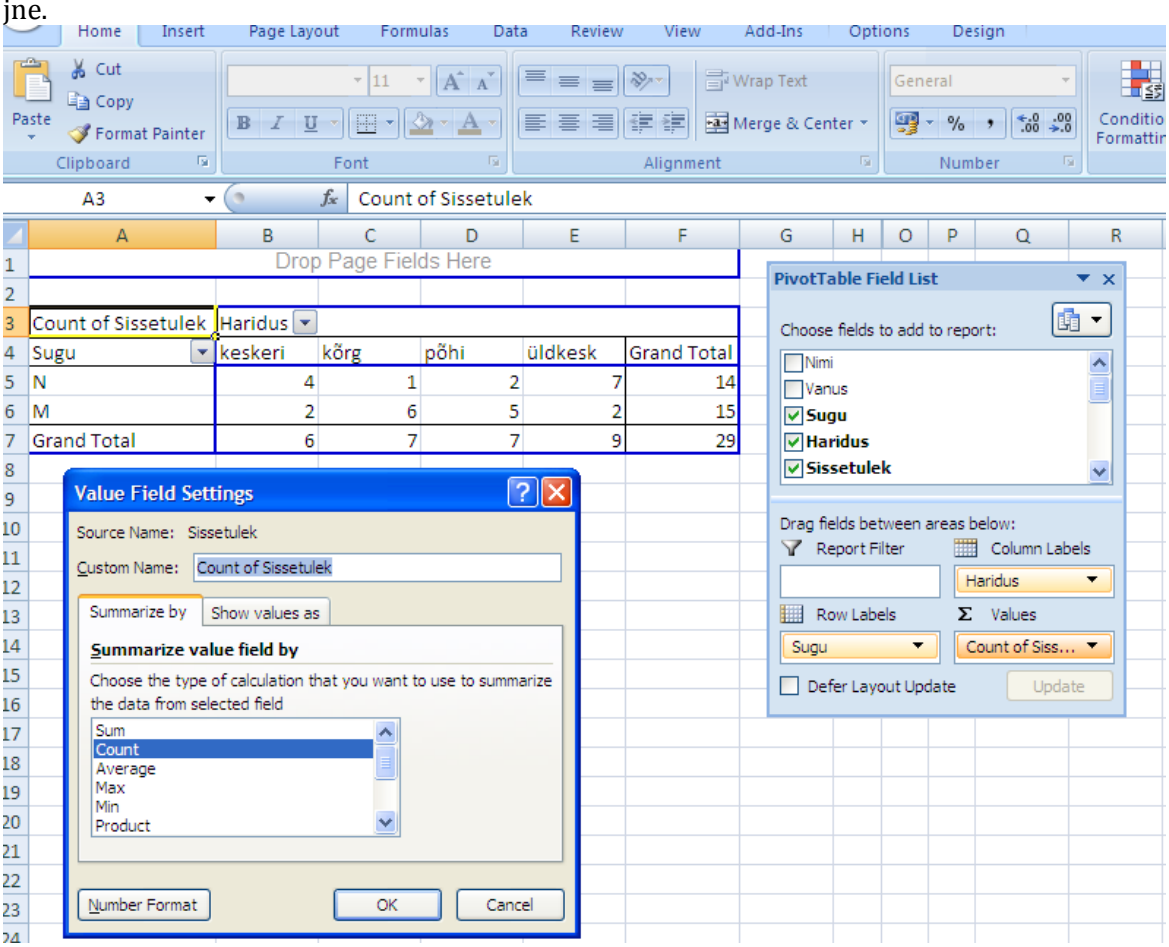
NB! Excel 2007 on võimalik see lohistada ka väljaloendi osa „Drag fields between areas below“ lahtrisse „Σ Values“.

Vaikimisi kuvatakse iga lahtri sagedus (nt alltoodud jooniselt on näha, et naisi, kellel on keskeriharidus, on 4, naisi, kellel on kõrgharidus, on 1, mehi, kellel on üldkeskharidus, on 2 jne). Selleks, et leida meile huvipakkuv keskmine, tuleb Excel 2007 vajutada väljaloendi sektsioonis „Drag fields between areas below“ noolele ∇ „Count on Sissetulek“ ning valida „Value Field Settings“.

Seal tuleks valida vajalik statistiline funktsioon (meie näites aritmeetiline keskmine ehk average):

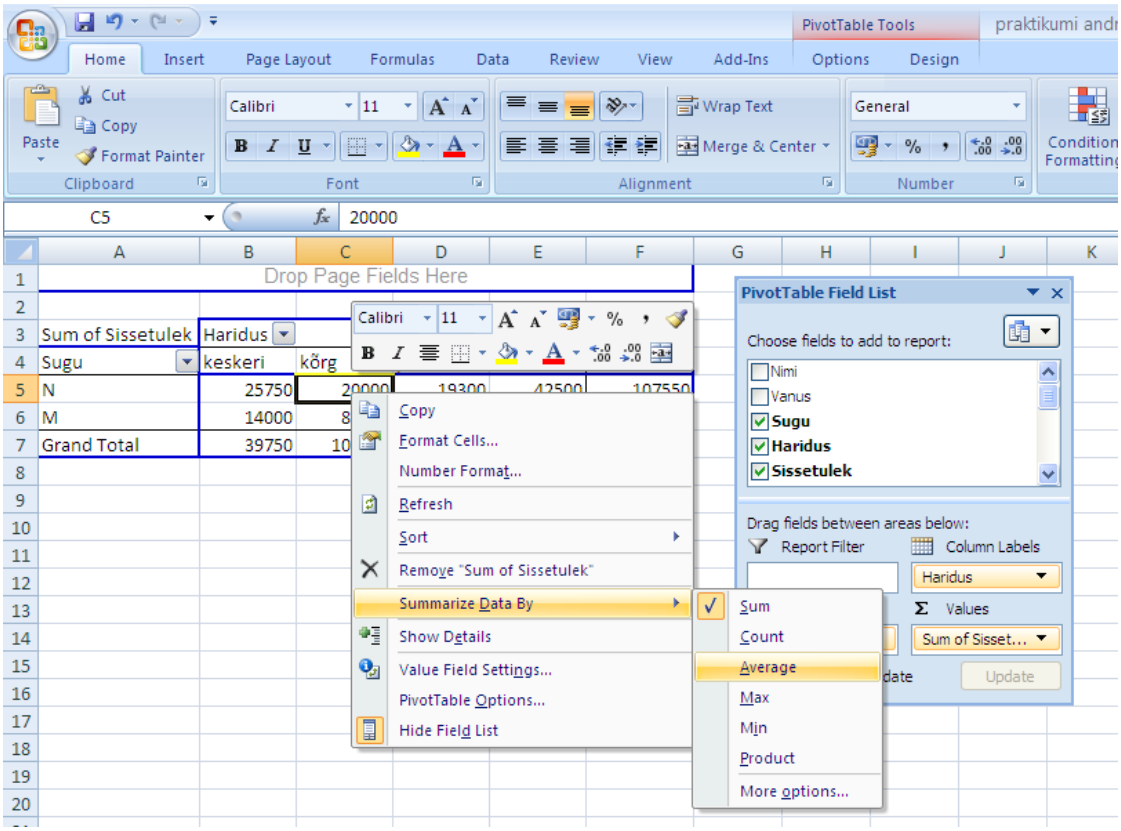
- summeerimine (sum), vaikimisi
- loendamine (count)
- aritmeetiline keskmine (average)
- maksimaalne väärtus (max)
- minimaalne väärtus (min)

jne.



NB! Dialoogiakna „Value Field Settings“ saab kuvada ka klikkides liigendtabeli paigutusalas hiire paremal klahvil.

Vajaliku statistilise funktsiooni saab Excel 2007 ka määrata lihtsamalt, nimelt tehes hiire paremkliki ja valides avanevast vaatest sobiva näitaja valikust „Summarize Data By“ (vt joonis).



Nagu näha, on hiire paremklikiga avanevas valikus võimalik

- kopeerida
- lahtreid formaatida (seda saab teha ka kuvatud menüüriba abil)
- numbreid formaatida (nt vähendada komakohti, muuta fonti, joondamist, värvi jne)
- tulemusi värskendada
- andmeid sorteerida
- muutujaid eemaldada
- detaile kuvada
- avada dialoogiaken „Value Field Settings“, millest oli juttu ülalpool
- kuvada liigendtabeli sätted (sh sakil „Layout & Format“ määratleda, mida kuvada tühjade lahtrite korral, kas uuendada automaatselt väljade laiusi juhul, kui tabelit uuendatakse; sakil „Totals & Filters“ määratleda, kas kuvatakse ka rea/veeru kokku-väärtused, vahesummad (subtotals) jne)
- peita väljade loendi.

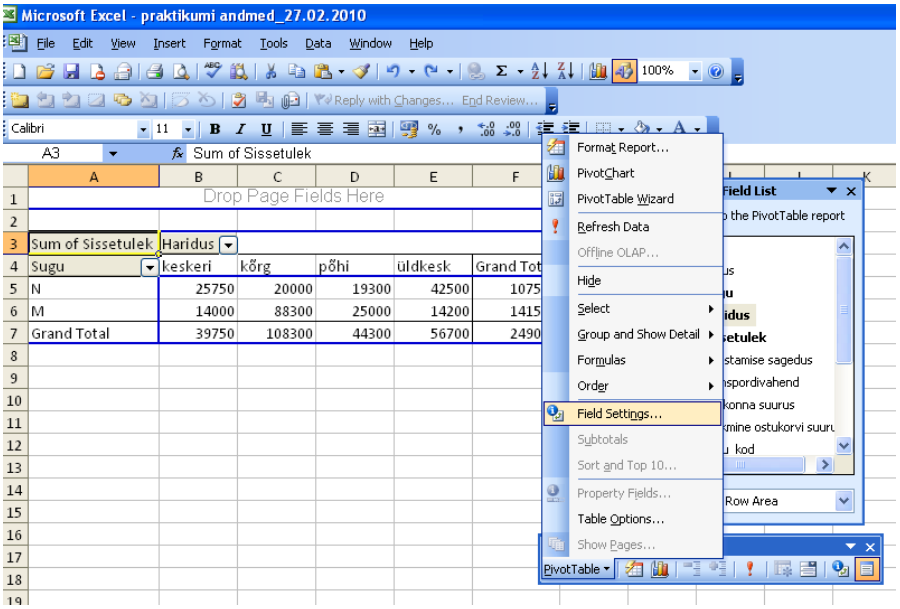
Excel 2003 on sama info võimalik kuvada mitmel viisil:

1. vajutades PivotTable menüüribal ikooni „Field settings“, mille järel avaneb dialoogiaken, kus saab määratleda muutuja (väli „Name“, meie näites „Sissetulek“), statistilise funktsiooni (meie näites „Average“) ja vahesummad („Subtotals“, võimalik on teha valik kolme alternatiivi vahel – „Automatic“, „Custom“, „None“)

NB! Vahekokkuvõtted on olulised seepärast, et kahe ja enama muutuja kasutamisel reas või veerus kuvatakse alati liigendtabelis vaikimisi vahekokkuvõtted. Teatud juhtudel need meid

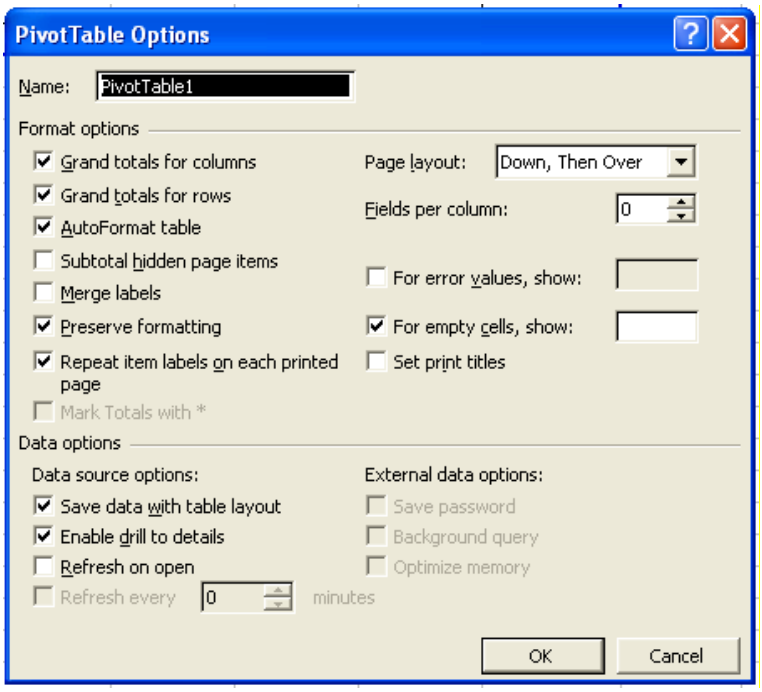
ei huvita, siis saab need hõlpsalt eemaldada. Selleks tuleb näiteks minna ühe „Total“ väärtuse peale, teha parem klikk ja valida „Field settings“ ning subtotals alt „None“.

2. vajutades PivotTable menüüribal ikooni „“PivotTable“ ja valides avanevast menüüst „Field Settings“



3. klikkida hiire parema klahviga liigendtabeli kõige vasakpoolsel ülemisel lahtril (meie näites „Sum on Sissetulek“) ning teha avanevas aknas valik „Field settings“)

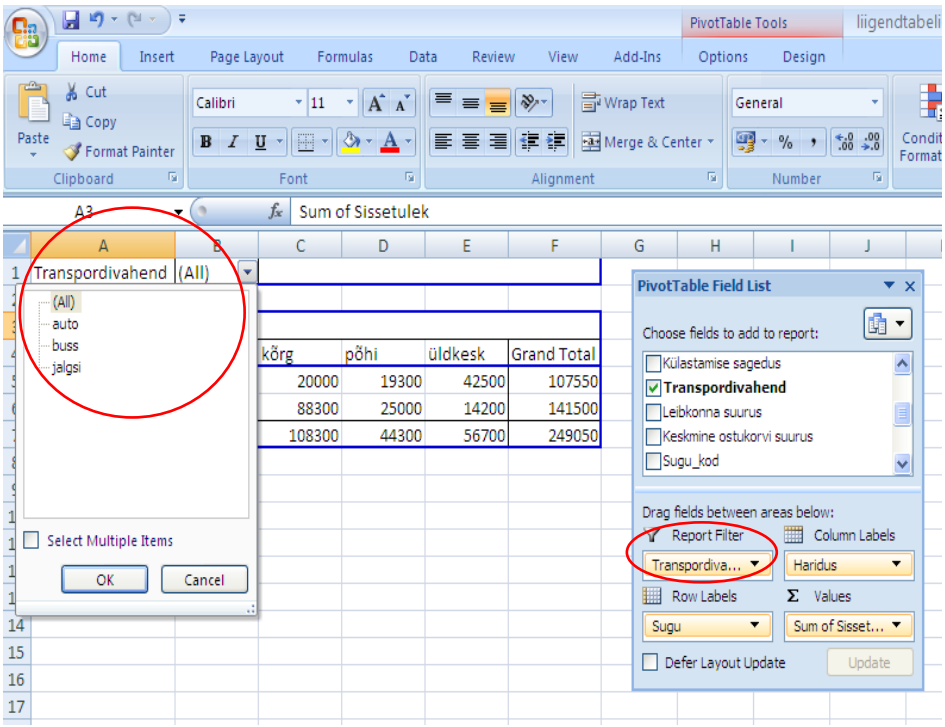
„Field setting“ all oleva valiku „Table options“ klikkimisel avaneb Excel 2003 dialoogiaken „PivotTable Options“.



Selles saab määratleda, kas liigendtabelis soovitakse kuvada „kokku“ rea/veeru „kokku“-väärtusi (Grand rows for rows/ for colums).
Funktsiooniga „AutoFormat table“ saab fikseerida, kas liigendtabelis veergusid kitsamaks/laiemaks nihutades iga järgmise liigutusega liigendtabel formaadi tabelit automaatselt ehk siis teeb veerud jälle laiemaks vastavalt pealkirja pikkusele (kuidas seda Excel 2007 teha, on käsitletud ülalpool).

NÄITE JÄTK. Oletame, et tahame sama analüüsi teha läbi vaid nende inimeste korral, kes kasutavad poes käimiseks autot.

Selleks on mõistlik lisada transpordivahendi muutuja filtrina, kas lohistades muutuja liigendtabeli alale „Drop Page Fields Here“, Excel 2007 on võimalik muutuja lohistada ka akna „PivotTable Field List“ alale (Report Filter“), vt ka allolevat joonist.

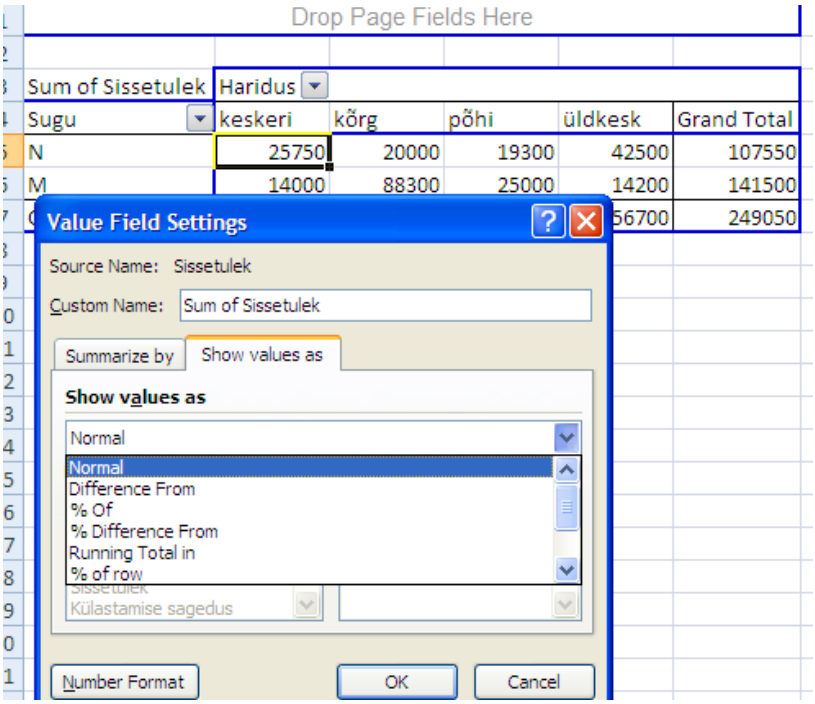


NÄITE JÄTK. Leiame, kuidas jagunevad valimisse kuuluvad mehed ja naised haridustasemete lõikes, st meid huvitab, kui suur osa naistest on põhiharidusega, kui suur osa keskharidusega jne.

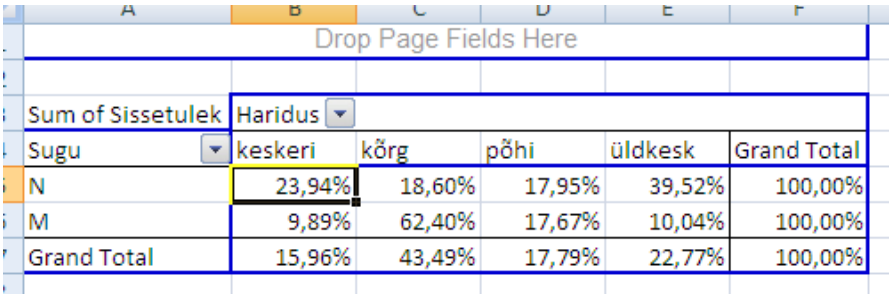
Selleks tuleb esmalt lohistada liigendtabeli aladele vajalikud väljad näiteks järgmiselt:

- ritta muutuja „sugu“
- veergu muutuja „haridus“
- andmete ossa mingi pidev muutuja (nt sissetulek)

Selleks, et leida suhtelisi sagedusi, tuleb kursor asetada mingisse lahtrisse andmete sektsioonis, teha hiire paremklikk, valida „Value Field Settings“ (sellest oli meil ülal juttu), seal valida sektsioon „Show values as“, misjärel avaneb järgmine vaade:



Siit tuleks meie uurimisprobleemist lähtuvalt valida „% of Row“, mille järel kuvatakse liigendtabelis tulemused nii, et iga rea väärtuste summa on kokku 100%.



Teised valikud:

- erinevus („Difference From“), siin tuleb ka täiendavalt määratleda, milline väli on erinevuste arvutamise aluseks („Base Field“) ja milline kategooria on erinevuste arvutamise aluseks („Base item“)
- osakaal („% of“), valikud on analoogilised eelmisega
- protsentuaalne erinevus („% Difference From“), valikud on analoogilised eelmisega
- osakaal veerust/kokku-väärtusest („% of Column“, „% of Total“)
  jne.

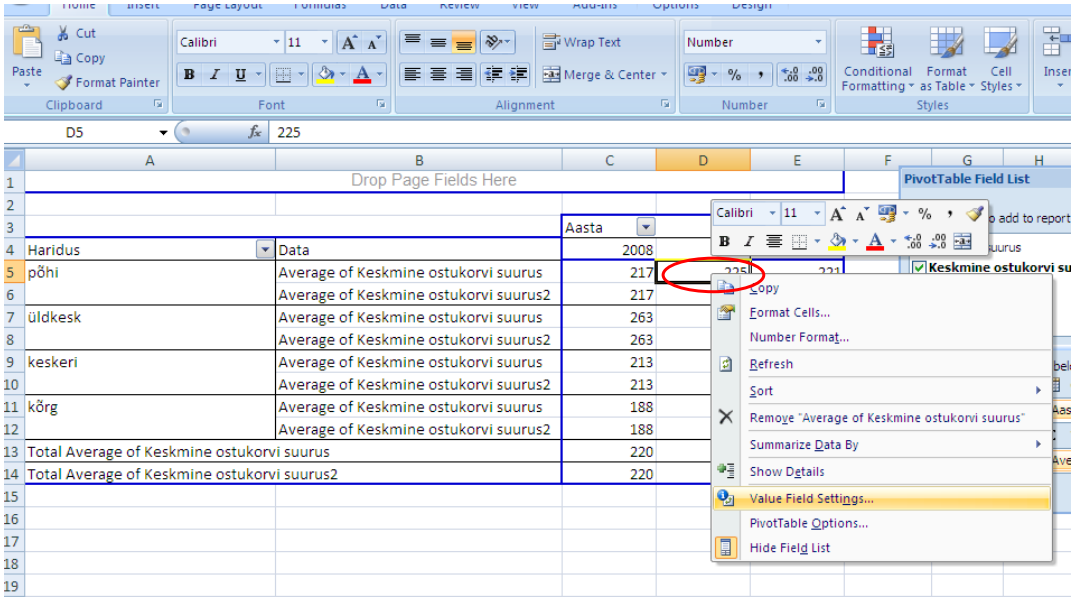
Excel 2003 tuleb minna andmete ossa, valida „Field Settings“ (kuidas seda teha, sellest oli juttu eespool, edasi tegutseda analoogiliselt äsja kirjeldatuga.

NÄITE JÄTK. Leiame, kas aastaga on ostukorvi suurus haridustasemete lõikes muutunud. Selleks lisame alguses tabeli veergu muutuja „Aasta“. Et arvutada protsentuaalne muutus nii, et tabelisse jääks nii ostukorvi suurus kui ka selle muutus, tuleb muutuja „Ostukorvi suurus“ tõsta tabelisse teist korda.

NB! Kui tõsta andmete lahtrisse tõsta rohkem kui üks muutujat, siis kuvatakse andmed vaikimisi reas:

			Aasta		
Haridus_kod	Haridus	Data	2008	2009	Grand Total
1	põhi	Average of Keskmine ostukorvi suurus	217	225	221
		Average of Keskmine ostukorvi suurus2	217	225	221
2	üldkesk	Average of Keskmine ostukorvi suurus	263	280	272
		Average of Keskmine ostukorvi suurus2	263	280	272
3	keskeri	Average of Keskmine ostukorvi suurus	213	325	250
		Average of Keskmine ostukorvi suurus2	213	325	250
4	kõrg	Average of Keskmine ostukorvi suurus	188	283	229
		Average of Keskmine ostukorvi suurus2	188	283	229
Total Average of Keskmine ostukorvi suurus			220	271	245
Total Average of Keskmine ostukorvi suurus2			220	271	245

Kuna see ei ole alati kõige mugavam vaade, siis selleks, et tõsta andmed veergu, tuleb teha järgmist. klikkida lahtril „Data“ lahtri peale, teha hiire paremklikk, valida „Order“, „Move to Column“ (analoogiliselt saab muutujaid tõsta algusesse, vasakule, paremale, lõppu). Muutuse arvutamiseks tuleb minna teist korda tõstetud muutuja peale (alloleval joonisel tähistatud punase ringiga).



Sakil „Show Values As“ tuleb valida „% Difference From“. „Base Field“ on muutuja, mille järgi muutuse arvutatakse, meie näites on selleks aasta. „Base item“ on eelmise muutuja väärtus, vis võetakse arvutamisel aluseks. Kuna meil aastal ainult 2 väärtust (2008 ja 2009), siis võime aluseks võtta kas

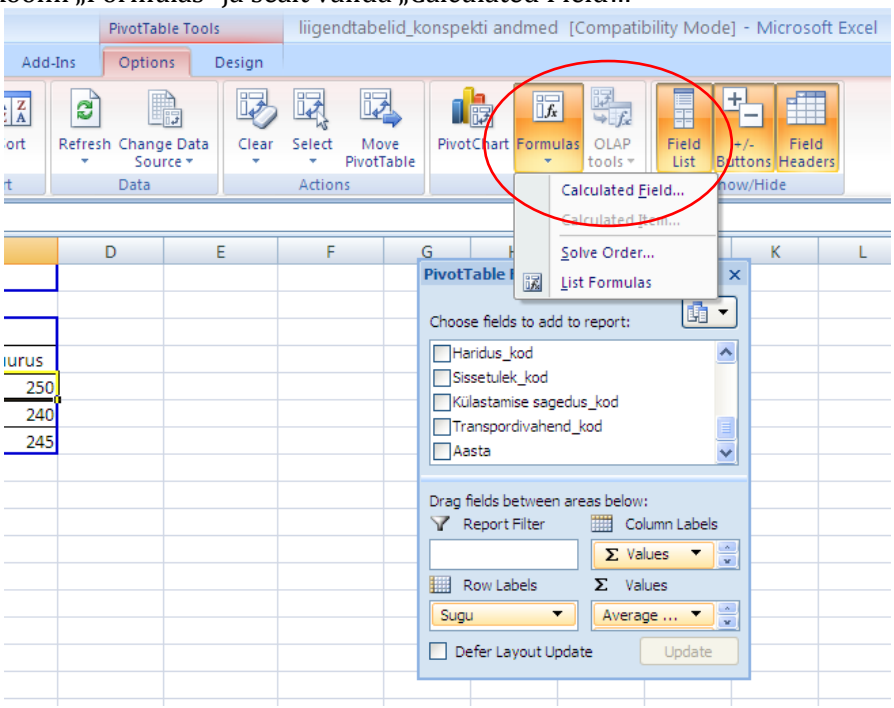
- 2008 või
- previous

NB! Mitme väärtuse korral oleks vahe (kui valida previous, siis on tegu ahelindeksiga, kui konkreetne aasta, siis on tegu baasindeksiga)

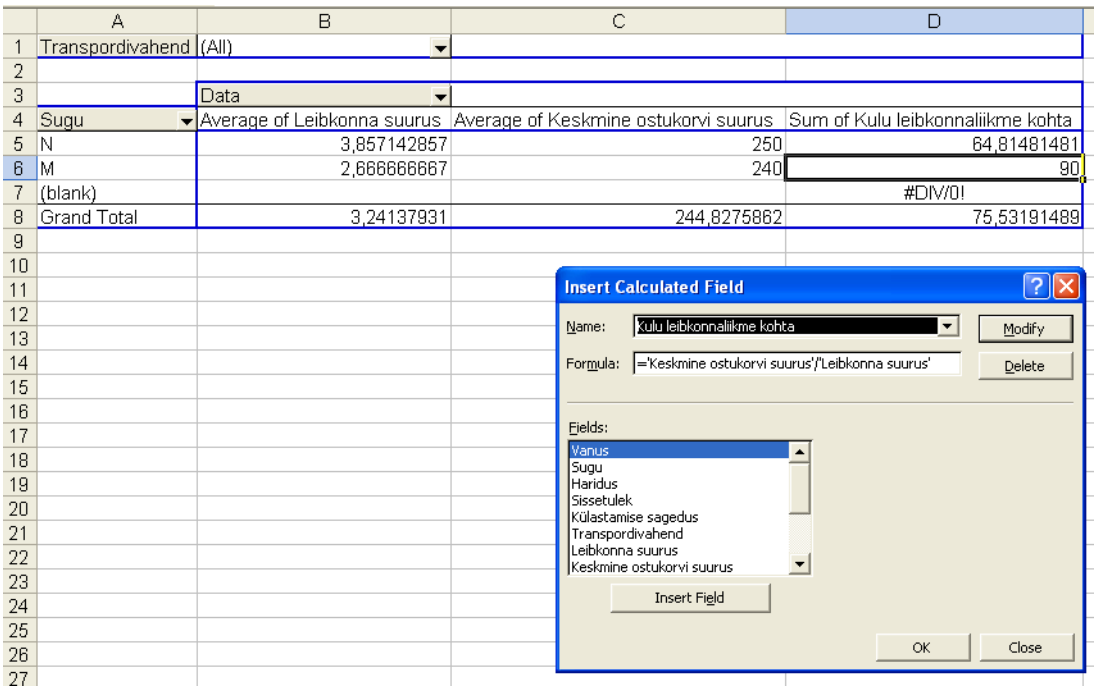
NÄITE JÄTK. Arvutame uue tunnuse „keskmine ostukorv leibkonnaliikme kohta“ meestele ja naistele.

Selleks tuleb liigendtabeli ritta lohistada muutuja „Sugu“ ning andmete ossa muutujad „Keskmine ostukorvi suurus“ ja „Leibkonnaliikmete arv“.

Excel 2007 tuleb klikkida menüüriba „PivotTable Tools“ sektsiooni „Options“ osas „Tools“ asuval ikoonil „Formulas“ ja sealt valida „Calculated Field…“



Excel 2003 tuleb liigendtabeli tööribalt teha hiireklikk valikul „PivotTable“ ning valida „Formulas“ ja „Calculated Field…“.



	A	B	C	D
1	Transpordivahend	(All)		
2				
3		Data		
4	Sugu	Average of Leibkonna suurus	Average of Keskmine ostukorvi suurus	Sum of Kulu leibkonnaliikme kohta
5	N	3,857142857	250	64,81481481
6	M	2,666666667	240	90
7	(blank)			#DIV/0!
8	Grand Total	3,24137931	244,8275862	75,53191489

Avanevas aknas tuleb määrata uue muutuja nimi (nt „Kulu leibkonnaliikme kohta“) ja valem.

NB! Käesolev konspekt ei ole kindlasti ammendav, kuna võimalusi vajalike tulemuste saamiseks on sageli enam kui üks.

Seega head liigendtabelite maailma avastamist!

Praktikumiülesanne

27.02.2010

1. Mitu meest ja mitu naist on andmebaasis?
2. Milline on keskmise ostukorvi suurus soo ja haridustaseme lõikes?
3. Kuidas jagunevad valimisse kuuluvad mehed ja naised haridustasemete lõikes?
4. Milline on keskmine ostukorvi suurus eri transpordivahendit kasutavatel inimestel?
5. Kuidas varieerub keskmine ostukorvi suurus külastamise sageduste lõikes?
6. Kas aastaga on ostukorvi suurus haridustasemete lõikes muutunud?
7. Leiame uue näitaja: keskmine ostukorv leibkonnaliikme kohta. Mida saab järeldada?

Kodutöö 2 (max 10 punkti)

LIIGENDTABELITE TEGEMINE

Eesmärk: tudengil oskab kasutada liigendtabeleid ning kirjeldava statistika tulemusi sisukalt tõlgendada.

Kodutöö andmebaas on lühendatud versioon Euroopa Sotsiaaluuringu (European Social Survey) andmebaasist.

Kodutöös palun vastata järgmistele küsimustele.

8. Mis tüüpi andmestikuga on tegu (vt loenguslaid „Andmete korraldamise viise")?
9. Tutvuge kodutöö andmebaasis sisalduvate muutujatega.
10. Püstitage viis uurimishüpoteesi (a la „soovime uurida andmebaasi kaasatud indiviidide jagunemist soo lõikes"; „soovime uurida andmebaasi kaasatud indiviidide keskmist vanust haridustasemes"). Iga hüpoteesi korral
    - määratlege kasutatavate muutujate tüübid (nominaal, järjestus- või arvuline tunnus);
    - kasutades Exceli liigendtabelite koostamise võimalust, koostage asjakohased liigendtabelid ja nende illustreerimiseks joonised (seejuures mõelge, milline joonise tüüp – sektor-, tulp-, joon- vms diagramm – on sobivaim), esitage tabelid/joonised kodutöös;
    - tõlgendage tulemusi.

Uurimishüpoteesid püsitage nii, et kasutate andmesektsiooni keskmist, loendeid, osakaalusid (% of row, % of column , % of total).

NB! OLULINE!!!

Kodutöö tähtaeg on **28. märts**. Iga esitamisega viivitatud päeva eest kaotad ühe punkti, kuid **pärast 2. aprilli esitatud kodutöid ei aktsepteerita**.

Kodutöö on rangelt soovitatav teha **kahekesi**.

Kodutöös ootan **sisulist analüüsi**, st ei piisa sellest, kui kirjeldavad statistikud välja arvutada, neid peaks ka sisukalt tõlgendama, sama kehtib jooniste kohta.

Kodutööga koos palun esitada ka **Exceli faili**, kus on olemas kodutöös kasutatud **kirjeldavate statistikute arvutused** ja **joonised**.

Kui hätta jääd, ära kõhkle nõu küsimast (kerly.krillo@ut.ee).

Head liigendtabelite maailma avastamist!

III LOENG, IV LOENG

Kodutöö 3 (max 10 punkti)

1. osa. Töötamine suurte andmehulkadega

Eesmärk: tudeng oskab kasutada Exceli funktsioone, filtreid, andmeid sorteerida ning luua liigendtabeleid

Kasuta praktikumi andmebaasi ning vasta järgmistele küsimustele.

1. Millises summas on müüdud harilikku piima? Kasuta funktsiooni SUMIF ja sea tingimus segmendi nime järgi. Kodutöösse esita vastus ja lahenduskäik funktsioonina.[1]
2. Millises summas on kaupluses D müüdud õunu? Kasuta Filtrit. Kodutöösse esita vastus.
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.
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).
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
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.

[1] Funktsiooni sisu esitamiseks kirjuta Excelis funktsiooni ette ’, näiteks ’=sum(A1:A4). Nii saad funktsiooni Wordi kopeerida.

2. osa. Andmete analüüs SPSSis

Eesmärk: tudeng oskab teha SPSSis lihtsamat statistilist analüüsi (kasutada menüü „Analyze“ osasid „Descriptive Statistics“, „Means“ and „Correlation“)

Andmebaas: „**Kodutoo 3_andmebaas**“

1. Andke ülevaade valimist, sh kirjeldage kirjeldavate statistikute abilvalimi soolist (muutuja „sex“), vanuselist („age“), hariduslikku („educ“) jaotust.

Selleks:

- leidke asjakohased kirjeldavad statistikud (selleks võite kasutada „Analyze“ menüü osas „Descriptive Statistics“ alajaotusi „Frequencies“), võttes arvesse tunnuse tüüpi (st mõelge, millised statistikud annavad sisukat infot pideva arvulise tunnuse korral nagu vanus ja millised nominaaltunnuse korral nagu sugu), tõlgendage tulemusi;
- valimi jaotumise illustreerimiseks soolises lõikes tehke sektordiagramm (nt „Analyze“ → „Descriptive Statistics“ → „Frequencies“ → „Charts“ → „Pie“ või „Graphs“ → „Chart Builder“ → „Pie/Polar“). Kandke joonisele ka jaotuse protsentuaalsed väärtused (selleks tehke joonisel topeltklikk valige „Show data labels“)
  ning vanuselises ja hariduslikus lõikes histogramm, kuhu kandke ka normaaljaotuse kõver (saate samast kohast). Mida järeldate?

2. Analüüsige ja võrrelge valge- ja mustanahaliste („race“) keskmiste kooliskäidud aastate arvu („educ“), õnnelikkuse taset (happy“) ja ametialalist („occat80“) jaotumust. Selleks:
- võrrelge esmalt kirjeldavate statistikute väärtusi kummaski grupis. Kõige hõlpsam on tulemusi genereerida, kui jaotate valimi rassi muutuja („race“) alusel kaheks. Seda saate teha järgmiselt:
  1) „Data“ →“Split File“ →“Compare Groups“ → (race)
  Seejärel võite nt kasutada „Analyze“ → „Descriptive Statistics“ → “Frequencies“/“Descriptives“
  NB! KUI OLETE VAJALIKUD ANDMETABELID JA JOONISED GENEREERINUD, ÄRGE UNUSTAGE „SPLIT FILE“-I MAHA VÕTTA!!!
  2) „Analyze“ → „Descriptive Statistics“ → „Explore“ (seal valige „Factor List“ muutujaks „race“ ja „Dependent List“ muutujateks „sex“, „educ“ ja „happy“).
- tehke muutuja „educ“ erinevuste illustreerimiseks rassi lõikes karpdiagramm (*boxplot*, siinkohal soovitan kasutada „Analyze“ → „Descriptive Statistics“ → „Explore“ võimalusi)

Millised on olulisimad järeldused?

3. Kontrollige, kas valgenahaliste keskmine kooliskäidud aastate arv on 12 (ehk teisisõnu, kas „keskmisel“ valgenahalisel on keskharidus). Selleks tuleb esmalt analüüsi kaasata vaid need vaatlused, kus muutuja „race“ väärtuseks on 1 – „white“. Seda saate teha järgmiselt:
   „Data“ → „Select Cases“ → “If condition is satisfied“ → race=1
   Seejärel „Analyze“ → „Compare Means“ → „One-Sample T-Test“
   Mida järeldate?
   Tehke sama analüüs läbi ka mustanahaliste korral.
   NB! KUI OLETE VAJALIKUD ANDMETABELID GENEREERINUD, ÄRGE UNUSTAGE „SELECT CASES“-I MAHA VÕTTA!!!

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);
- tehke sõltumatute valimite t-test („Analyze“ → „Compare Means“ → “Independent Samples t-test“ (võrreldavad grupid on „1“ – valgenahalised ja „2“ – mustanahalised). Mida järeldate

Levene'i testi põhjal (sh püstitage ka selle testiga kontrollitavad hüpoteesid)? Mida järeldate t-testi tulemuste põhjal?

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
- tehke dispersioonanalüüs, „Analyze" → „Compare Means" → "One-Way ANOVA" →

Dependent List: „age" ja „educ" NB! Dispersioonanalüüsis on sõltuv tunnus alati pidev!
Factor: „occcat80".
Mida järeldate?

- 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"
- Kasutades "One-Way ANOVA" võimalust „Contrasts", analüüsige, kas:
  1) erinevused juhtide (muutuja „occcat80" kategooria „1") ja lihttööliste (kategooria „6") keskmistes kooliskäidud aastate arvus on statistiliselt olulised
  2) erinevused teenindustöötajate (muutuja „occcat80" kategooria „3") ja põllumajandustöötajate (kategooria „4") keskmistes kooliskäidud aastate arvus on statistiliselt olulised.

6. Uurige korrelatsioonanalüüsi („Analyze" → „Correlate" → "Bivariare") abil, kas neil indiviididel, kelle vanemate kooliskäidud aastate arv on suurem (muutujad „paeduc" ja „maeduc", on kooliskäidud aastate arv suurem („educ"). Tehke ka asjakohased joonised. Millised on järeldused?

NB! OLULINE!!!

Kodutöö tähtaeg on **5. mai**. Iga esitamisega viivitatud päeva eest kaotad ühe punkti, kuid **pärast 9. aprilli esitatud kodutöid ei aktsepteerita**.

Kodutöö on soovitatav teha **kahekesi**.

Kodutööga koos palun esitada ka **Exceli faili**, kus on olemas vastuste lahenduskäik (1. osa). 2. osa puhul tuleks kõik asjakohased tabelid/joonised esitada kodutöö lahenduse tekstis.

Kui hätta jääd, ära kõhkle nõu küsimast (kerly.krillo@ut.ee).

Head statistika maailma avastamist!

V LOENG

Kodutöö 4 (max 10 punkti)

Eesmärk: tudeng oskab sisukalt tõlgendada t-testide ja dispersioon-, regressioon- ja klasteranalüüsi tulemusi.

3. T-testid ja dispersioonanalüüs analüüs SPSSis

Andmebaas: „**Kodutoo 3_andmebaas**“

7. Analüüsige t-testi abil, kas naiste keskmine vanus on 45. Selleks
   - püstitage null- ja alternatiivne hüpotees,
   - tehke ühe valimi t-test.

Mida järeldate?
Korrake sama analüüsi meeste korral.
NB! Nõuanne. Selleks, et aega vähem kuluks, võite andmebaasi muutuja „sex“ alusel jaotada ja seejärel teha t-testi üks kord. Selleks tuleb valida „Data“ → „Split File“ → „Compare Groups“ (muutujaks „sex“)
NB! Kui olete vajalikud andmetabelid genereerinud, ärge unustage „split file“-i maha võtta!!!

8. Analüüsige, kas erinevused meeste ja naiste keskmises vanuses on statistiliselt olulised. Selleks
- püstitage null- ja alternatiivne hüpotees (ehk H_0 ja H_1),
- tehke sõltumatute valimite t-test („Analyze“ → „Compare Means“ → “Independent Samples T-Test“.

Mida järeldate t-testi tulemuste põhjal?

9. Analüüsige, kes erineva eluga rahuolu tasemega inimestel („happy“) on keskmine vanus „age“ ja kooliskäidud aastate arv („educ“) erinev. Selleks teostage dispersioonanalüüs, sh
- püstitage uuritav hüpoteeside paar
- tehke dispersioonanalüüs, „Analyze“ → „Compare Means“ → “One-Way ANOVA“

Mida järeldate?
- Analüüsige, millistes eluga rahulolu 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“
- Kasutades “One-Way ANOVA“ võimalust „Contrasts“, analüüsige, kas:
  3) erinevused väga õnnelike (muutuja „happy“ kategooria „1“) ja mitte väga õnnelike (kategooria „3“) keskmistes kooliskäidud aastate arvus on statistiliselt olulised,
  4) erinevused väga õnnelike (kategooria „1“) ja keskmiselt õnnelike (kategooria „2“) keskmistes kooliskäidud aastate arvus on statistiliselt olulised.

5) erinevused keskmiselt õnnelike (kategooria „2“) ja mitte väga õnnelike (kategooria „3“) keskmistes kooliskäidud aastate arvus on statistiliselt olulised.

2. Regressioon- ja klasteranalüüs

1. Leidke kas
   a) Eurostatist Teid huvitavad muutujad (soovitatavalt võiks olla 4-6 sõltumatut muutujat, andmebaasi tuleks kaasata EL27 riikide andmed ühe aasta kohta (nt 2008))
   või
   b) kasutage mõnd meelepärast andmebaasi (nt andmeid, mida analüüsite oma magistritöö/kursusetöö vms raames), kus sisalduvad andmed,

mille vahelist seost soovite regressioonanalüüsi abil analüüsida.

Esialgu soovitan andmed salvestada Exceli faili ning seejärel tõsta SPSSi. SPSSi tõstmisel tuleb „Variable view“ aknas defineerida muutujate lühinimed „Name“ ja pikemad nimetused „Label“. Viimased kuvatakse väljundtabelites.

2. Teostage regressioonanalüüs. Selleks

- määratlege sõltuv muutuja ja sõltumatud muutujad
- teostage lineaarne regressioonanalüüs „Analyze“ → „Regression“ → “Linear“. Mida järeldate?

Soovitus tõlgendamiseks: kui kasutate analüüsis EL27 andmeid, siis tuleks mudeli parameetreid tõlgendada järgmiselt: riigis, kus muutuja x (see on muutuja, mille parameetrit tõlgendate) väärtus on ühe ühiku võrra kõrgem, on muutuja y (see on Teie mudeli sõltuv muutuja) väärtus parameetri väärtus võrra kõrgem/madalam. Kui jääte hätta, küsige nõu kas foorumis või otse minult.

- kontrollige, kas mudelis esineb multikollineaarsus (selleks tuleb teha aknas „Statistics“ linnukesed valiku „Part and partial correlations“ ja „Collinearity diagnostics“ ette).

Multikollineaarsuse ja selle testimise kohta saate vajadusel lisainfot siit: www.mtk.ut.ee/doc/OkonIVOsa.pdf.

- kontrollige, kas mudelis on heteroskedastiivsus. Selleks tuleb analüüsida seost mudeli hinnatud jääkliikme ja sõltuva muutuja vahel. Analüüsi teostamiseks:
  a) salvestage jääkliikmete väärtused (tehke lineaarse regressiooni aknas „Save“ linnuke grupis „Residuals“ valiku „Standardized“ ette)
  b) tehke joonis, kus x-teljel on punktis a) salvestatud jääkliikmete väärtused (see muutuja kuvatakse andmebaasis viimasena) ja y-teljel Teie poolt valitud sõltuv muutuja. Kui on näha, et varieeruvus kasvab/kahaneb, siis on tegemist heteroskedastiivsusega.

  Heteroskedastiivsuse ja selle testimise kohta saate vajadusel lisainfot siit: www.mtk.ut.ee/doc/OkonIVOsa.pdf.
- kontrollige, kas jääkliikmed on normaaljaotusega. Selleks tehke lineaarse regressiooni dialoogiakna „Plots“ sektsioonis „Standardized Residual Plots“ linnuke valiku „Histogram“ ees.

Mida järeldate?

LISAÜLESANNE, MIS ANNAB MAKSIMAALSELT 5 LISAPUNKTI JA MILLE LAHENDAMINE ON TEGELIKKUSES VÄGAGI SOOVITATAV, ET TEEKSITE TUTVUST KLASTERANALÜÜSI MAAILMAGA ☺

3. Kasutades oma andmebaasi, analüüsige hierarhilise klasteranalüüsi abil, millised andmebaasi objektid on sarnased, millised erinevad. Selleks

- valige analüüsi kaasatavad muutujad
- teostage klasteranalüüs „Analyze" → „Classify" → "Hierarchical Cluster"

NB! Juhul, kui kasutate andmeid, mis varieeruvad erinevalt, ärge unustage andmeid standardiseerida (selleks tuleb dialoogiakna „Method" sektsioonis „Transform Values" valida Standardize: Z scores)

- Valige jooniste sektsioonist dendrogramm.

Mida järeldate? Milline võiks olla sobiv klastrite arv?

- Analüüsige, mille poolest klastrid üksteisest erinevad. Selleks
  a) Genereerige uus muutuja, mis näitab, millisesse gruppi iga vaatlus kuuub. Seda saab teha järgmiselt: dialoogiaknas „Save" tuleb teha linnuke valiku „Single solution" eest ja kastikesse märkida see klastite arv, mis tundub sobiv.
  b) Jaotage andmebaas loodud muutuja alusel gruppideks („Data" → „Split File")
  c) Analüüsige analüüsi kaasatud muutujate keskmisi erinevates klastrites „Analyze" → „Descriptive Statistics" → "Descriptives"

  Mida järeldate?

NB! Juhul, kui kasutate analüüsis enda andmebaasi, siis soovitan klasteranalüüsi teostamiseks kasutada 15-20 vaatlust. Vastasel juhul läheb nö pilt väga kirjuks ning keerukas on dendrogrammilt midagi mõttekat välja lugeda.

NB! OLULINE!!!

Kodutöö tähtaeg on **4. juuni**. Iga esitamisega viivitatud päeva eest kaotad ühe punkti, kuid **pärast 9. juunit esitatud kodutöid ei aktsepteerita**.

Kodutöö on soovitatav teha **kahekesi**.

Kui hätta jääd, ära kõhkle nõu küsimast (kerly.krillo@ut.ee).

Head statistika maailma avastamist!

Iseseisev lugemine

F-TEST

F-testi abil kontrollitkse kahe valimi dispersioonide võrdsust.

Näiteks võib F-Testi tööriista kasutada kahe klassi õpilaste pikkuste (hinnete vms) võrdlemiseks. Tulemi saamiseks võrdleb tööriist nullhüpoteesi, et mõlemad valimid pärinevad võrdse dispersiooniga jaotustest, alternatiivse hüpoteesiga, et vastavate jaotuste dispersioonid pole võrdsed.

Tööriist arvutab F-statistiku (ehk F-suhte) f-väärtuse. Kui f-väärtus on 1 lähedal, siis osutab see, et aluseks olevad populatsioonidispersioonid on võrdsed. Kui väljundtabelis on f < 1 "P(F <= f) ühepoolne", annab see tõenäosuse, et võrdsete populatsioonidispersioonide puhul on F-statistiku väärtus vaatlemisel väiksem kui f, ning "F-statistiku kriitiline ühepoolne" annab valitud olulisuse nivoo alfa puhul tulemiks, et kriitiline väärtus on väiksem kui 1. Kui f > 1, "P(F <= f) ühepoolne", annab see tõenäosuse, et võrdsete populatsioonidispersioonide puhul on F-statistiku väärtus vaatlemisel suurem kui f, ning "F-statistiku kriitiline ühepoolne" annab alfa kriitiliseks väärtuseks rohkem kui 1.

Rohkem infot leiate siit:

http://www.eau.ee/~ktanel/kool_ja_too/stat_excelis/hypot_Ftest.html

http://www.sauga.pri.ee/audentes/download/ps_konspekt_lk31_42.pdf

KORRELATSIOONANALÜÜS

Sissejuhatus. Võimalikke seosekujusid

Mitmemõõtmelises statistikas ei huvita uurijat sageli mitte niivõrd iga muutuja analüüs eraldi, vaid sageli soovitakse analüüsida muutujate (võimalikku) seost. Sageli pakub huvi, kas ühe näitaja kõrge

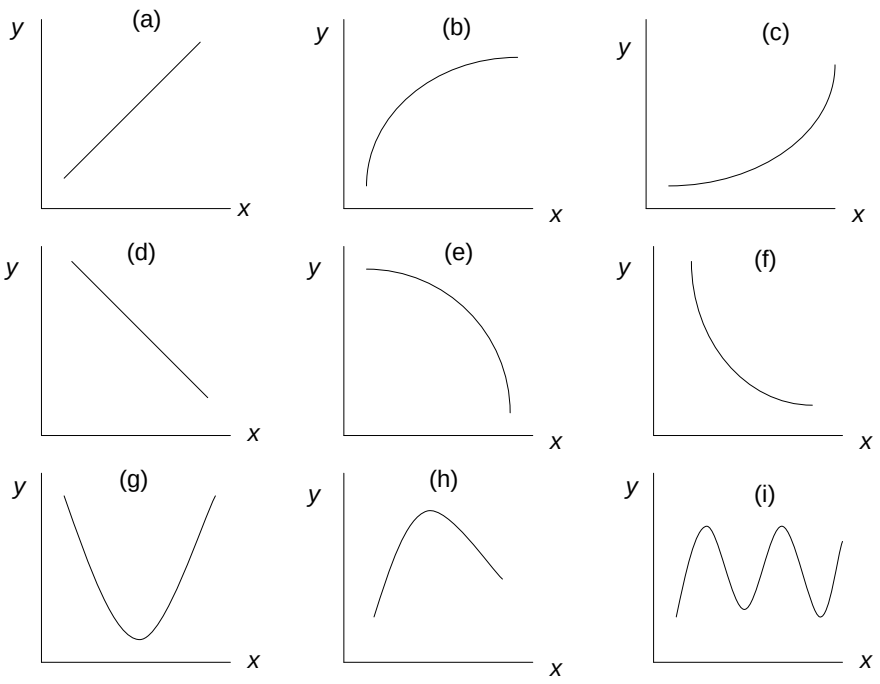
tasemega kaasneb teise näitaja kõrge/madal tase või mitte[1]. Tavaliselt võetakse taoliste küsimuste analüüsimisel aluseks alljärgnev tabel:

Tabel 1. Kahe muutuja vaheliste seoste analüüsimisel kasutatavad algandmed.

objektid	muutujad	
	x	y
objekt 1	x_1	y_1
objekt 2	x_2	y_2
objekt 3	x_2	y_3
…		
objekt n	x_n	y_n

Kui ühe muutuja kõrgete väärtustega kaasnevad teise muutuja kõrged ning madalatega madalad väärtused, siis öeldakse, et muutujate vaheline seos on positiivne ehk tegemist on positiivse korrelatsiooniga (vt joonis 1 osasid *a, b* ja *c*). Negatiivse seosega on tegemist juhul, kui ühe muutuja kõrgete väärtuste korral on teise muutuja väärtused madalad ning vastupidi (vt joonisel 1 osasid *d, e* ja *f*).

[1] Näiteks võib meid huvitada küsimus, kuidas eristuvad piirkonnad tööpuuduse ja kuritegevuse taseme poolest: kas piirkondades, kus on tööpuuduse tase kõrgem, on ka kuritegevuse tase kõrgem.



Joonis 1. Kahe muutuja vahelisi võimalikke seosekujusid.

Kolmas võimalus on, et muutujate teatud väärtuste korral on tegemist positiivse, teatud piirkonnas aga negatiivse seosega (vt joonisel 1 osasid *g, h* ja *i*). Sellist tüüpi seoseid nimetatakse sageli mittemonotoonseteks (*non-monotonic*). Joonisel *1g* kujutatud seos eksisteerib sageli toodangumahu ja keskmiste kulude vahel, joonisel *1h* kujutatud seos aga kasumi taseme ning reklaamikulude vahel. Joonisel *1i* kujutatud seost nimetatakse sageli tsükliliseks seoseks ning sellist tüüpi seos esineb kõige sagedamini juhtudel, kui muutuja *x* väljendab aega (*y* võib olla tööpuuduse määr vms). Juhul kui kahe muutuja väärtuste vahel ei ole süstemaatilist seost, siis öeldakse, et muutujad ei ole seotud, ei ole korreleeritud; on ortogonaalsed või sõltumatud.

Antud osa lõpetuseks tuleb mainida, et joonisel 1 esitatud seosed on suures osas teoreetilised ning reaalses elus nii siledaid ja pidevaid seoseid reeglina ei eksisteeri. Sellest tulenevalt on uurija ülesandeks välja selgitada, kas kahe muutuja vaheline seos on ligikaudu kirjeldatav ühega joonisel esitatutest ning kui on, siis millisega.

Kahe muutuja vahelise korrelatsiooni uurimine

Kahe muutuja vahelise korrelatsiooni uurimiseks on otstarbekas enne korrelatsioonikoefitsiendi väljaarvutamist analüüsida pisut ka andmeid. Seda on võimalik teha näiteks kahedimensioonilise tabeli (*2x2 contingency table*) ning punktdiagrammi (*scatter diagram*) abil.

Kuna kahe muutuja vaheline korrelatsioon peegeldab seda, mil määral ühe muutuja kõrged väärtused on seotud teise muutuja kõrgete/madalate väärtustega, siis võib analüüsi alustada sellest,

et hinnatakse iga objekti ning tuvastatakse, kas vaatlusaluste muutujate väärtused on tema puhul kõrged või madalad (need võib tähistada näiteks vastavalt plussi ja miinusega). Seejärel saadakse järgmine **kahedimensiooniline tabel**.

Tabel 2. Korrelatsiooni uurimine kahedimensioonilise tabeli abil.

muutuja 1	muutuja 2	
	kõrge	madal
kõrge	+ +	+ -
madal	- +	- -

Seejärel analüüsitakse, kui palju objekte kuulub igasse nelja gruppi. Juhul kui seos kahe muutuja tasemete vahel (st korrelatsioon) puudub, peaks igasse gruppi kuuluma enam-vähem ühepalju objekte. Kui aga on täheldatav mingi seaduspära, siis on näitajad ilmselt korreleeritud.

Selline analüüs on väga üldine, sest vaadeldakse üksnes, kas muutuja väärtus on mediaanist kõrgem või madalam, erinevuse arvuline suurus aga jäetakse vaatluse alt välja. Seepärast on järgmise sammuna otstarbekas vaadelda andmeid **punktdiagrammi** abil.

Kuigi punktdiagramm annab andmetest sageli ülevaatlikuma pildi kui tabel ning võimaldab visuaalse vaatluse põhjal uurijal teha oletusi seose olemasolu ja suuna kohta, annaks märksa enam informatsiooni kahe näitaja vahelist seost mõõtev numbriline indeks. Kõige sagedamini kasutatakse selleks **korrelatsioonikoefitsienti**. Kui tegemist on valimiga, tähistatakse korrelatsioonikoefitsient tavaliselt tähega *r* ning see on hinnang ülekogumi korrelatsioonikoefitsiendile ρ. Korrelatsioonikoefitsient omandab väärtusi vahemikus [-1;+1]. Korrelatsioonikoefitsiendi arvutusvalem on järgmine:

$$(1) \qquad r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y},$$

kus s_x ja s_y tähistavad vastavalt muutujate *x* ja *y* standardhälbeid ning n on vaatluste arv.

Kui *r = 1*, siis on tegemist perfektse positiivse korrelatsiooniga ning *r = -1* tähistab perfektset negatiivset korrelatsiooni. Kui *r = 0*, siis kahe näitaja vahel korrelatsioon puudub.

Vahemärkus: korrelatsioonikoefitsienti on võimalik leida ka standardiseeritud väärtuste abil. Sellisel juhul defineeritakse kaks uut muutujat:

$$z_x = \frac{(x_i - \bar{x})}{s_x} \text{ ning } z_y = \frac{(y_i - \bar{y})}{s_y}.$$

Sel juhul

(2) $$r = \frac{\sum z_x z_y}{n-1}.$$

Olulised eeldused korrelatsioonikoefitsiendi leidmisel

1. korrelatsioonikoefitsient r mõõdab üksnes kahe muutuja vahelist ***lineaarset*** seost. Juhul kui kahe muutuja vaheline seos on mittelineaarne, on võimalik, et hoolimata seose eksisteerimisest saame korrelatsioonikoefitsiendi väärtuseks nullilähendase arvu.
2. Analüüsitavad kaks muutujat on ***ühise normaaljaotusega*** (*joint normal distribution*).

Kui korrelatsioonikoefitsient on leitud, tuleb kontrollida selle statistilist olulisust. Selleks kasutatakse vastavat kriitiliste väärtuste tabelit. Vabadusastmete arvuks on korrelatsioonikoefitsiendi leidmisel *n-2* (st *df = n – 2*).

Vahemärkus. Korrelatsioonikordaja leidmisel tuleb, nagu eespool mainitud, arvestada mitmete piiravate eeldustega. Seetõttu kasutatakse lisaks eespool käsitletud lineaarsele korrelatsioonikoefitsiendile mitmeid teisi näitajaid. Näited[2]: *rank correlation coefficient, biserial correlation coefficient, point biserial correlation coefficient, tetrachoric correlation coefficient, four-fold point correlation coefficient, correlation ratio* jne.

Korrelatsioonitulemuste tõlgendamine

NB! Korrelatsioon ei tähenda kausaalsust!

Korrelatsioonanalüüsi tulemuste tõlgendamisel tuleb arvestada sellega, et korrelatsiooni olemasolu kahe muutuja vahel ei anna uurijale informatsiooni muutujatevahelise kausaalsuse ehk põhjuslikkuse kohta ning viimase väljaselgitamiseks tuleb kasutada teisi meetodeid (nt Grangeri kausaalsuse test vms).

Segavad muutujad

Korrelatsioonanalüüsis tuleb alati arvestada võimalusega, et eksisteerivad n.-ö. segavad muutujad, mis moonutavad analüüsitulemusi. Segavateks nimetatakse muutujaid, mida ei ole otseselt analüüsi kaasatud, kuid mis mõjutavad analüüsitavaid/analüüsitavat muutujaid/muutujat.

Korrelatsioonanalüüsi rakendusvõimalused

- Oluliste muutujate ning nendevahelise seose olemasuolu väljaselgitamine – korrelatsioonanalüüs annab uurijale hea ülevaate olemasolevatest andmetest. Kui kahe muutuja vahel korrelatsioon puudub, on see analüütikule oluliseks lisainformatsiooniks, sama kehtib ka seose olemasolu korral.

[2] Lugejal on soovitatav iseseisvalt erinevate korrelatsioonikoefitsientide olemusega tutvuda.

- Sisendite varieerimine – kui korrelatsioonanalüüs näitab, et uuritavate muutujate vahel on statistiliselt oluline korrelatsioon, siis saab järgmise sammuna andmeid varieerides analüüsida nendevahelist põhjuslikkust jms.
- Hüpoteeside testimine – teadmine, et kaks muutujat on/ei ole korreleeritud, loob aluse edasiste hüpoteeside testimiseks.
- Prognoosimine – kuigi prognoosimine seondub peamiselt regressioonanalüüsiga, mis ei ole antud loengu teema, sõltub prognoosimise täpsus (*accuracy*) vaatlusaluste muutujate vahelisest korrelatsioonist.
- Usaldusväärsuse hindamine – usaldatavuse all mõeldakse käesoleval juhul taasteostatavust (*reproducibility*). Sisuliselt otsitakse vastust küsimusele, kas saadud tulemused on juhuslikud või saadaks samad tulemused ka juhul, kui mõõtmised uuesti teostataks. Juhul kui kahel mõõtmisel saadakse ligikaudu samad tulemused, öeldakse, et mõõtmistulemused on usaldusväärsed (*test-retest stability*). Sõltuvalt kahe mõõtmise vahelisest ajaperioodist võivad tulemused üksteisest küll mõnevõrra erineda, kuid üldjuhul peaks kahe muutuja vaheline korrelatsioon (*test-retest correlation*) olema kõrgem kui 0.90.
- Valiidsuse hindamine – valiidsuse all mõeldakse käesoleval juhul seda, mil määral mõõtmistulemused peegeldavad neile seatud eesmärke. Näiteks kui tegemist on testiga, mille abil hinnatakse tulevast tööturul osalemise edukust ning selgub, et testitulemused ei ole edukusega korreleeritud, siis ei ole antud test valiidne: see ei täida koostaja poolt seatud eesmärke.

NB! Mõtiskleda iseseisvalt usaldusväärsuse ja valiidsuse seoste üle (st kas muutuja saab olla usaldusväärne ilma, et oleks valiidne või vastupidi)!

- Surrogaatmuutujate identifitseerimine – korrelatsioonanalüüsi abil on võimalik välja selekteerida muutujaid, mis on teistele muutujatele asendajateks (nt kuluefektiivsuse vms alusel).

Mitmemõõtmeline korrelatsioonanalüüs

Mitmemõõtmelise korrelatsioonanalüüsi korral on muutujaid enam kui kaks ning kõikidel objektidel mõõdetakse kõikide muutujate väärtused. Seejärel arvutatakse sarnaselt eelnevale analüüsile välja **kõikide** muutujate vahelised korrelatsioonikoefitsiendid. Nende koondamisel tabelisse saadakse **korrelatsioonimaatriks**.

Korrelatsioonimaatriks on diagonaali suhtes sümmeetriline ruutmaatriks, kus igas reas on üks vaatlusalune muutuja. Sama kehtib ka veeru kohta. Maatriksi igas lahtris esitatakse aga kahe muutuja vahelise seose tugevust väljendav korrelatsioonikoefitsient. Diagonaalil asuvad korrelatsioonikoefitsiendid on võrdsed ühega.

Osakorrelatsioon

Osakorrelatsioon mõõdab kahe muutuja vahelise korrelatsiooni tugevust eeldusel, et kolmanda muutuja mõju on neutraliseeritud. Vastav arvutusvalem on järgmine:

$$r_{12.3} = \frac{r_{12} - r_{13} \cdot r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}$$

kus $r_{12.3}$ – muutujate 1 ja 2 vaheline korrelatsioon eeldusel, et muutuja 3 on konstantne,

r_{ij} – muutujate *i* ja *j* vaheline korrelatsioonikoefitsient.

Järjestikune korrelatsioon (*serial correlation*)

Järjestikuse korrelatsiooniga ehk autokorrelatsiooniga on tegemist juhul, kui mõõdetakse korrelatsiooni tugevust näitaja ning selle viitaja (*lag*) vahel. Korrelatsioonikoefitsientide vastavat graafilist esitust nimetatakse korrelogrammiks.

NÄIDE

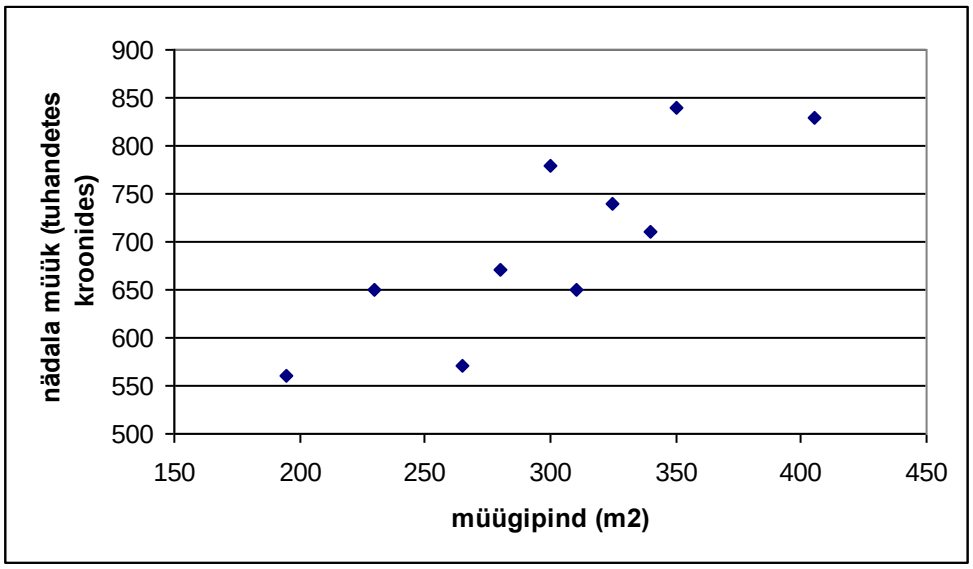
Mööbli maaletooja soovib teada, kas tema toodete läbimüük on seotud kaupluste pinnaga. Selle väljaselgitamiseks valib ta (juhuvaliku põhimõttel) välja 10 kauplust, milledes mõõdab kaupluse pinna ning uurib välja nädala müüginumbrid. Tulemused on koondatud tabelisse 3, mis on analoogne tabeliga 1.

Tabel 3. Andmed kümne kaupluse müügipinna ja läbimüügi kohta.

kaupluse number	müügipind (m^2)	nädala müük (tuhandetes kroonides)
1	340	710
2	230	650
3	405	830
4	325	740
5	280	670
6	195	560
7	265	570
8	300	780
9	350	840

10	310	650

Sageli (eriti mahukate tabelite puhul) on üksnes algandmete visuaalse vaatluse põhjal keeruline tuvastada, kas andmete vahel on seos või mitte. Sel juhul on otstarbekas informatsioon muutujate kohta esitada punktdiagrammil, mis annab andmetest ülevaatliku pildi. Tabelis 3 esitatud andmete põhjal koostatud punktdiagramm on esitatud joonisel 2. Iga punkt joonisel esindab üht kümnest kauplusest.



Joonis 2. Punktdiagramm mööblikaupluste läbimüügi ja müügipinna kohta.

Punktdiagrammi põhjal tundub, et kahe näitaja vahel eksisteerib positiivne seos. Selles veendumiseks arvutame välja korrelatsioonikoefitsiendi, tuginedes valemitele 1 ja 2.

Tabel 4. Korrelatsioonikoefitsiendi leidmine.

kaupluse nr	müügipind (m^2)	nädala müük (tuh kr)	z(x)	z(y)	z(x)*z(y)
1	340	710	0,657	0,102	0,067
2	230	650	-1,149	-0,509	0,584
3	405	830	1,724	1,322	2,279
4	325	740	0,410	0,407	0,167
5	280	670	-0,328	-0,305	0,100
6	195	560	-1,724	-1,424	2,454
7	265	570	-0,575	-1,322	0,760

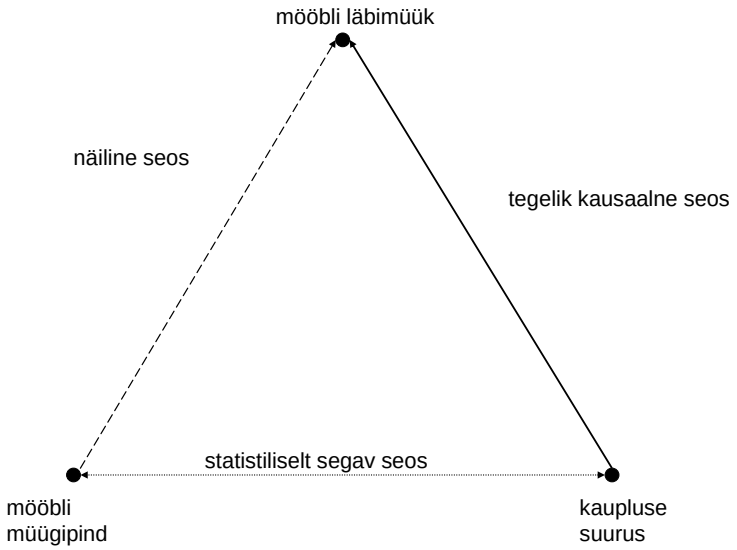
8	300	780	0,000	0,814	0,000
9	350	840	0,821	1,424	1,169
10	310	650	0,164	-0,509	-0,083
summa	3000	7000	0,000	0,000	7,496
keskmine	300	700	0	0	
standardhälve	60,92	98,32	1,00	1,00	

Korrelatsioonikoefitsiendi *r* väärtuseks saame 0.833. Kriitiliste väärtuste tabelist näeme, et vabadusastmete arvu 8 ning olulisuse nivool 0.1 korral on kriitiliseks väärtuseks 0.715 (st kui korrelatsioonikoefitsiendi väärtus on suurem kui 0.715, siis tuleb nullhüpotees kahe muutuja vahelise korrelatsiooni puudumisest ümber lükata). See aga tähendab, et eksisteerib statistiliselt oluline positiivne korrelatsioon kaupluse müügipinna ja mööbli läbimüügi vahel.

Tulemuste interpreteerimine

Nagu eespool mainitud, et saa korrelatsioon põhjal teha järeldusi kahe muutuja vaheliste seoste põhjuslikkuse kohta. See tähendab, et ei saa väita, et poodides, kus on müügipinda rohkem, on läbimüük suurem või et suurema läbimüügiga kauplustes on müügipinda rohkem.

Antud näites võib segavaks muutujaks olla kaupluse kogupind. Näiteks võib tegelikkuses olla olukord selline, et suuremates poodides on rohkem pinda mööbli all. Sellisel juhul on kaupluse suurus segatud mööblialuse pinnaga (vt joonis 3).



Joonis 3. Segava muutujaga situatsiooni graafiline esitus.

Jooniselt on näha, et suurema mööblimüügi põhjuseks on suurem kaupluse suurus ning mööblialuse pinna ning läbimüügi seos on näiline, st põhjustatud eelmisest seosest.

DISPERSIOONANALÜÜS *(*Analysis of Variance*)*

Sissejuhatuseks

Dispersioonanalüüs võimaldab identifitseerida seoseid pideva sõltuva muutuja (*criterion variable*) ja kategoorilis(t)e sõltumatu(t)e muutuja(t)e (*predictor variable*) vahel. Kuigi teatud juhtudel oleks võimalik dispersioonanalüüsi asemel kasutada regressioonanalüüsi, kaasates analüüsi fiktiivseid muutujaid, on esimene olemuselt laiem, võimaldades analüüsis kasutada nii kvantitatiivseid kui kvalitatiivseid prognoosimuutujaid.

Dispersioonanalüüs (*analysis of variance*, sageli kasutatakse lühendit ANOVA) hõlmab väga laia hulka tehnikaid, mis võimaldavad identifitseerida ning mõõta andmete varieeruvust. Kuna tegemist on meetodiga, mille tundmaõppimine eeldab uurijalt märksa põhjalikumat süvenemist, siis on antud loengu eesmärgiks anda kuulajatele eelkõige põhiteadmised ANOVA olemusest ning teemast sügavamalt huvitatud tudengitel on soovitatav iseseisvalt teemaga põhjalikumalt tutvuda.

Kuna dispersioonanalüüs baseerub F-jaotusel, siis tuletatakse esmalt lugejale põgusalt meelde nimetatud jaotuse põhiolemus. Seejärel tutvustatakse dispersioonanalüüsi ning saadud teadmiste kinnistamiseks tehakse vastavad arvutused läbi konkreetse näite varal.

F-jaotuse põhiolemus

Oletame, et tegemist on normaaljaotusega üldkogumiga, mille teatud karakteristiku keskväärtus on μ ning dispersioon on σ^2. Üldkogumist genereeritakse kaks valimit valimimahuga vastavalt n_1 ja n_2 ning arvutatakse mõlema valimi dispersioonid, mis tähistatakse kui s_1^2 ning s_2^2. Kui defineerida uus statistik F, mis avaldub kujul

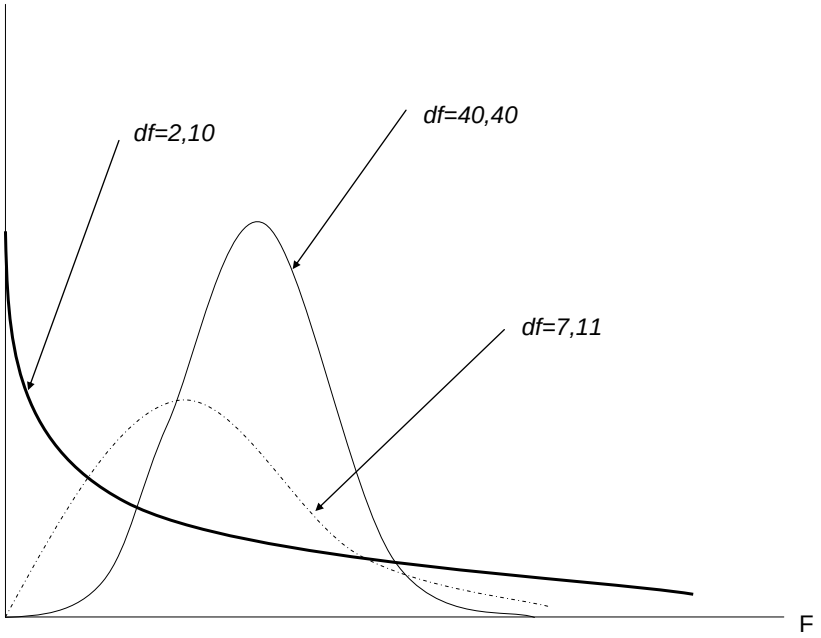
$$F = \frac{s_1^2}{s_2^2},$$

siis milline peaks eelduste kohaselt olema F-i väärtus? Kuna nii s_1^2 kui s_2^2 on eelduste kohaselt nihketa hinnangud üldkogumi dispersioonile σ^2, siis peaks F-i keskväärtus olema võrdne ühega, st

$$E(F) = E(\frac{s_1^2}{s_2^2}) = \frac{E(s_1^2)}{E(s_2^2)} = \frac{\sigma^2}{\sigma^2} = 1.$$

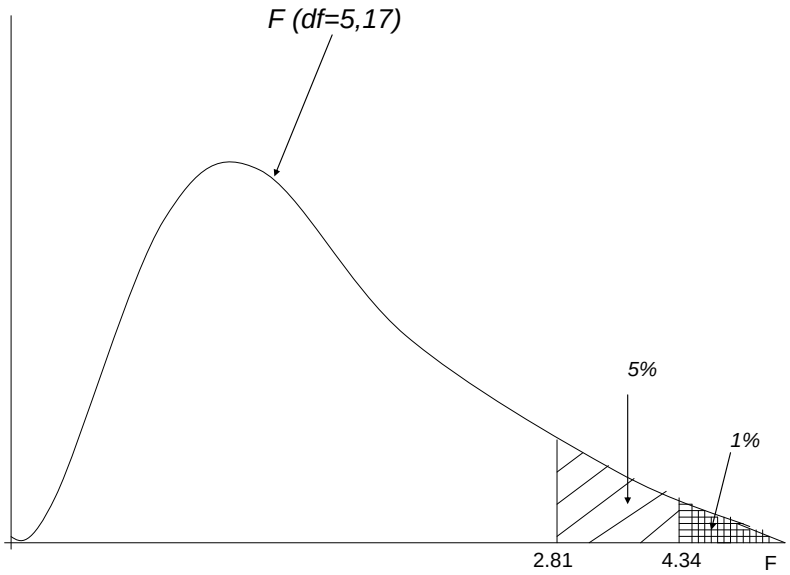
Seega, kui me genereeriksime normaaljaotusega üldkogumist lõpmata palju valimeid ning arvutaksime iga kord välja F-statistiku väärtuse, siis oleks selle väärtus keskmiselt (NB! mitte alati!) võrdne ühega.

Kuna F-statistik on funktsioon valimi dispersioonidest, on ka F-statistiku jaotus nimetatud näitajatega seotud. Täpsemalt, F-statistiku jaotus sõltub valimi mahtudest – mida suurem on valimi maht, seda väiksem on dispersioon. Selle tulemusena on F-statistiku jaotus erinev sõltuvalt analüüsitavate valimite suurustest. Täpsemalt formuleerides – F-statistiku jaotus varieerub olenevalt valimite põhjal arvutatavatest vabadusastmete arvudest. Vabadusastmete arv on ühe võrra väiksem valimi mahust. Kui vabadusastmete arv on suur, läheneb F-statistiku jaotus normaaljaotusele (vt joonis 1).



Joonis 1. Mõned näited F-jaotuse realisatsioonide kohta eri vabadusastmete paaride korral.

Teemast huvitatud lugejal on tungivalt soovitatav F-jaotuse olemusega iseseisvalt põhjalikumalt tutvuda, kuna loengumahu piiratuse tõttu ei ole võimalik sellel pikemalt peatuda. Küll aga võib tõdeda, et tegemist on n.-ö. hästi käituva tõenäosusjaotusega, st sarnaselt normaaljaotusele on võimalik välja arvutada, milline on tõenäosus, et F on teatud väärtusest kõrgem. Antud teema kontekstis pakub huvi see, et F-statistiku kriitilised väärtused on vabadusastmete korral tabuleeritud. Näiteks vabadusastmete 5 ja 17 korral (st *df=5,17*) on F-statistiku kriitiline väärtus kõrgem kui 2.81 tõenäosusega 0.05; 1% olulisuse nivool on vastav näitaja 4.34 (vt joonist 2).



Joonis 2. F-statistiku kriitilised väärtused 5% ja 1% olulisuse nivool vabadusastmete arvu 5 ja 17 korral.

Dispersioonanalüüsi olemus

Nagu eespool mainitud, kasutatakse dispersioonanalüüsi selgitamaks, kas üldkogumist moodustatud kahe või enama valimi teatud muutujate keskväärtused erinevad statistiliselt olulisel määral. Kuigi nimetus dispersioonanalüüs viitab sellele, et analüüsi keskmes on varieeruvused, on lihtsam käsitleda ANOVAt (ning ka MANOVAt) kui tehnikat, mille abil võrreldakse sõltumatu(te) muutuja(te) väärtuste erinevusi keskmisest

Kui tegemist on *k* valimiga, siis testitakse hüpoteesipaari:

H_0: $\mu_1 = \mu_2 = ... = \mu_k = \mu$

H_1: H_0 ei kehti, st vähemalt ühe valimi põhjal arvutatud muutuja keskväärtus erineb statistiliselt oluliselt teistest.

Testimaks, kas jääda nullhüpoteesi juurde või tuleks vastu võtta alternatiivne hüpotees, mille kohaselt kõikide gruppide antud näitaja keskväärtused ei ole võrdsed, arvutatakse välja gruppidesisesed ning –vahelised hälbed, mille arvutusalgoritm on esitatud tabelis 1. Neile tuginedes leitakse F-statistiku väärtus, mida võrreldakse vastava kriitilise väärtusega.

Tabel 1. Dispersioonanalüüsis kasutatavate näitajate arvutamisvalemid ühe sõltumatu muutuja korral.

dispersiooni liik	hälvete ruutude summa (*SS*)	df[3]	dispersiooni hinnang
kokku	$SS_T = \sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}^2 - \frac{(\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij})^2}{\sum_{j=1}^{k} n_j}$	$\sum_{j=1}^{k} n_j - 1$	-
gruppidevaheline	$SS_{bg} = \sum_{j=1}^{k} \frac{(\sum_{i=1}^{n_j} x_{ij})^2}{n_j} - \frac{(\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij})^2}{\sum_{j=1}^{k} n_j}$	*k-1*	$s_{bg}^2 = \frac{SS_{bg}}{k-1}$
gruppidesisene	$SS_w = SS_T - SS_{bg}$	$\sum_{j=1}^{k} n_j - k$	$s_w^2 = \frac{SS_w}{\sum_{j=1}^{k} n_j - k}$

Tabelis 1 esitatud arvutusalgoritmid on toodud ühe sõltumatu muutuja korral, sellist analüüsi nimetatakse „ühesuunaliseks" (*one-way*) või „lihtsaks" (*simple*). Analoogselt saab näitajad leida ka kahe ja enama sõltumatu muutuja korral ning sageli on just see uurija eesmärgiks. Selline dispersioonanalüüsi disain on tuntud kui faktoriline (*factorial design*) ning sellel peatutakse pikemalt käesoleva konspekti järgmises osas.

Faktoriline disain

Kui vaatluse all on vaid ühe sõltumatu muutuja mõju sõltuvale muutujale, püsib sarnaselt regressioonanalüüsiga oht, et eksisteerivad täiendavad, analüüsi mittekaasatud sõltumatud muutujad. Viimaste mõju uurimine oleks vajalik, kuna sellisel juhul võivad eksisteerida vastastikuse mõju efektid (*interacton effects*), st analüüsi mittekaasatud muutujad võivad mõjutada sõltumatu ja sõltuva muutuja väärtuseid.

Faktorilist disaini saab leida ükskõik kui suure arvu sõltumatute muutujate korral. Kui analüüsi kaasatakse kaks sõltumatut muutujat, on tegemist kahesuunalise (*two-way*) ANOVAga, kolme sõltumatu muutuja korral kolmesuunalisega jne. Faktorilise disaini korral leitakse lisaks sõltumatute muutuja ruutude summale ka nendevaheline vastastikuse seose efekt (interaktiivne efekt) ning saadud näitajate alusel arvutatakse välja F-statistikute väärtused ning olulisuse

[3] *df* – vabadusastmete arv

tõenäosused. Juhul kui interaktiivne efekt on statistiliselt oluline, tuleks saadud tulemustesse suhtuda ettevaatlikkusega ning võtta seda arvesse ka tulemuste tõlgendamisel. Kui aga vastastikuse mõju efekt on statistiliselt ebaoluline, siis viitab see asjaolule, et ühe sõltumatu muutuja mõju sõltuvale muutujale on teistest sõltumatutest muutujatest sõltumatu. Kuna enam kui ühe sõltumatu muutuja korral on näitajate arvutuskäik märksa keerukam ja võimalik on kasutada mitmeid alternatiivseid arvutusalgoritme ning seetõttu antud ainekursuses ei peatuta sellel pikemalt. Asjast huvitatud lugejal on soovitatav dispersioonanalüüsi kasutamisel iseseisvalt tutvuda kasutatavas tarkvarapaketis olemasolevate võimaluse ning nende leidmisalgoritmidega.

Näide

Alljärgnevalt analüüsitakse dispersioonanalüüsi olemust näite varal. Oletame, et majapidamiskaupade turustaja soovib teada, millised on tarbija eelistused seebikarpide värvi osas. Tarbija soovide tundmaõppimiseks kaasatakse analüüsi 12 kauplust. Neist neljas müüakse siniseid (*B*), neljas punaseid (*R*) ning neljas rohelisi (*G*) seebikarpe, kusjuures poed on valitud *juhuvaliku põhimõttel*. Karbid erinevad üksnes värvi poolest ning on muudelt omadustelt identsed. Esimese müüginädala tulemused, moodustatud valimi keskmised müügitulemused ning vastavad dispersioonid on toodud tabelis 2.

Tabel 2. Seebikarpide müük vastavalt värvile.

	Sinine	Punane	Roheline	
	6	18	7	
	14	11	11	
	19	20	18	Üldine
	17	23	10	keskmine
Valimi keskmine	14	18	11.5	14.5
Valimi dispersioon	32.7	26.0	21.7	

Testitakse, kas kehtib nullhüpotees

H_0: $\mu_B = \mu_R = \mu_G$, st kas eri värvi seebikarpide keskmine müük erineb statistiliselt olulisel määral või on erinevused üksnes juhuslikud. Sisuliselt on antud olukord analoogne juhuga, kui me kontrolliksime, kas kehtib hüpotees, mille kohaselt puudub seos seebikarbi värvi ja müügiarvu vahel. Hüpoteesi kehtivust kontrollitakse F-jaotusele tuginedes.

Gruppidesisene varieeruvus/dispersioon

Et F-statistikut arvutada, on vaja leida kaks dispersiooni hinnangut. Tabelis 2 on viimases reas esitatud andmed siniste, punaste ning roheliste seebikarpide müügi dispersiooni kohta, kuid kuna ükski neist ei ole teistest mingil objektiivsel põhjusel teistest eelistatum, tuleks nimetatud kolm näitajat ühendada ning luua näitaja, mis baseeruks gruppidesisesel varieeruvusel.

Kuna kõikides gruppides on valmi maht võrdne, on antud juhul võimalik suhteliselt hõlpsalt leida kolme sõltumatu dispersiooni hinnangu põhjal *ühendatud gruppidesisese dispersiooni hinnang* s_w^2 :

$$s_w^2 = \frac{32.7 + 26 + 21.7}{3} = 26.8$$

Juhul kui valimite mahud on erinevad (st antud grupi näitel igas grupis mitte neli poodi), siis tuleb s_w^2 leidmisel kasutada kaalutud keskmist.

NB! Ühendatud gruppidesisesel dispersioonil baseeruv valimi dispersioon ei ole võrdne kõigi 12 müügiarvu põhjal arvutatava dispersiooniga. Viimasel juhul leitakse dispersioon valimi keskmisele (mis antud näites on 14.5) tuginedes.

Gruppidevaheline varieeruvus/dispersioon

Antud näitaja hindab gruppide keskväärtuste (mis antud juhul on 14, 18 ning 11.5) varieeruvust üldisest keskmisest (14.5). Mida suurem on varieeruvus, seda tõenäolisem on, et vaatlusalused muutujad (antud näites seebikarpide müük ja värv) on seotud. Saadud näitajat kasutatakse gruppidevahelise dispersiooni hinnangu leidmisel. Kuna saadud näitaja pole mitte üldkogumi dispersiooni σ^2 hinnang, vaid keskväärtuste dispersiooni hinnang, siis tähistame selle $s_{\bar{X}}^2$ -ga (keskväärtuste standardvea ruut) ning antud näites leitakse see järgmiselt:

$$s_{\bar{X}}^2 = \frac{(14-14.5)^2 + (18-14.5)^2 + (11.5-14.5)^2}{2} = 10.75.$$

Varasematest statistika kursustest on teada, et $s_{\bar{x}}^2 = \frac{s^2}{n}$ ning siit saame, et $s^2 = n \cdot s_{\bar{x}}^2$.

Analoogselt leitakse ka gruppidevaheline dispersiooni hinnang, mis antud juhul on järgmine:

$$s_{bg}^2 = n \cdot s_{\bar{x}}^2 = 4 \cdot 10.75 = 43$$

NB! Gruppidevaheline dispersiooni hinnang ei peegelda erinevalt gruppidesisesest dispersiooni hinnangust mitte üksnes juhuslikku varieeruvust poodide vahel, vaid lisaks ka võimalikke varieeruvusi andmetes, mille põhjuseks on n.-ö. töötlemisefektid (*treatment effects*), antud juhul seebikarpide värvuse erinevused.

F-statistik

F-statistik avaldub gruppidesisese ja –vahelise dispersiooni hinnangu jagatisena:

$F = \frac{s_{bg}^2}{s_w^2}$. Käsitletavas näites on vabadusastmete arvuks 2 ja 9 (*df=2,9*) ning *F=1.60*.

Vabadusastmete arv

F-statistiku leidmisel kasutatavatel gruppidesisese ja –vahelise dispersiooni hinnangutel on kummalgi oma vabadusastmete arv, mille põhjal võrreldakse saadud statistiku väärtust vastava kriitilise väärtusega. Gruppidevahelise dispersiooni hinnangu s_{bg}^2 korral on vabadusastmete arv võrdne suurusega *k-1*, so ühe võrra väiksem gruppide arvust. Seebikarpide näites on s_{bg}^2 vabadusastmete arv võrdne kahega. Gruppidesisese dispersiooni hinnangu vabadusastmete arv on võrdne suurusega $\sum_j n_j - k$, kus n_j tähistab j-nda grupi valimi mahtu. Teisisõnu on gruppidesisese dispersiooni hinnangu vabadusastmete arv võrdne gruppide valimimahtude summa (mis loomulikult on võrdne kogu valimi mahuga) ja gruppide arvu vahega. Antud juhul on gruppidesisese dispersiooni hinnangu vabadusastmete arv võrdne üheksaga. Seega F-statistiku vabadusastmete arvud on antud näites 2 ja 9, mida lühidalt tähistatakse järgmiselt: *df=2,9*.

F-testi olulisus

Nagu eespool mainitud, peegeldab gruppidevaheline hinnang S_{bg}^2 nii varieeruvust, mis tuleneb juhuslikest poodidevahelistest erinevustest kui töötlemisefekte (st antud näites seebikarpide arvust tulenevat varieeruvust), samas kui gruppidesisene dispersiooni hinnang s_w^2 võtab arvesse üksnes esimest. Seega juhul, kui töötlemisefektid puuduksid, oleks F-statistik võrdne ühega ning kõrgem väärtus peegeldab võimalikke töötlemisefekte. Siinkohal tuleb aga meeles pidada, et F-statistiku väärtus võib olla ühest suurem ka lihtsalt juhuse tõttu, ilma, et töötlusefektidel oleks reaalne mõju. Antud näite korral tähendaks see, et me oleksime saanud samad müügimahud poodides ka juhul, kui kõik seebikarbid oleksid olnud näiteks sinised.

F-statistiku olulisuse leidmiseks kasutatakse vastavat kriitiliste väärtuste tabelit. Olulisusnivoo 0.05 ning *df=2,9* korral on F-statistiku kriitiline väärtus F_c*=4.26*. Kuna kriitiline väärtus ületab näites saadud F-statistiku väärtust, jõuame järeldusele, et seebikarbi värv ei mõjuta statistiliselt oluliselt selle müüki, st siniste, punaste ja roheliste seebikarpide müügiarvude keskväärtused ei erine üksteisest statistiliselt olulisel määral.

Kokkuvõtlikult: kuigi seebikarpide müüginumbrid viitasid, et müük varieerub mõnevõrra värvi lõikes, selgus dispersioonanalüüsi käigus, et erinevused ei olnud statistiliselt olulised.

Siinkohal tuleb aga juhtida lugeja tähelepanu asjaolule, et nullhüpoteesi juurde jäämine ei tähenda ilmtingimata, et saadud tulemus on õige. Sarnaselt varasemates ainekursustes õpitule võib ka

dispersioonanalüüsis teha 1. ja 2. järku vigasid. Lõpetuseks tuleb märkida, et nullhüpoteesi ümberlükkamise põhjuseks võib olla mitte erinevus keskväärtustes, vaid erinevus dispersioonides. Selle vältimiseks eeldatakse dispersioonanalüüsis F-statistiku leidmisel, et täidetud on homoskedastiivsuse nõue. See tähendab, et juhul, kui erinevused gruppide dispersioonides on statistiliselt olulised, on see tingitud erinevustest keskväärtustes, mitte dispersioonides (seejuures eeldatakse vaikimisi, et analüüsitav populatsioon on normaaljaotusega).

Statistikapakettides esitatakse dispersioonanalüüsi tulemus tavaliselt tabelina, kus veergudes on hälvete ruutude summad, vabadusastmete arv, dispersioonihinnang ning eelnevate näitajate põhjal leitud F-statistik. Viimased kolm näitajat on näites leitud, hälvete ruutude summad aga leitakse järgnevalt:

Siniste grupis: $(6-14)^2 + (14-14)^2 + (19-14)^2 + (17-14)^2 = 98$

Punaste grupis on näitaja 78 ning rohelistel 65. Ühendatud gruppidesisene hälvete ruutude summa avaldub kujul $\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_{.j})^2$,

kus x_{ij} tähistab i-ndat vaatlust j-ndas grupis ning $\bar{x}_{.j}$ analüüsitava näitaja keskväärtust j-ndas grupis.

Seebikarpide näites on ühendatud gruppidesisene hälvete ruutude summa 241.

Gruppidevaheline hälvete ruutude summa leitakse kaalutud keskmisena gruppide mahtude suhtes järgmiselt:

$$\sum_{j=1}^{k} n_j(\bar{x}_{.j} - \bar{x}_{..})^2 = 4\cdot(14-14.5)^2 + 4\cdot(18-14.5)^2 + 4\cdot(11.5-14.5)^2 = 86.0,$$

kus $\bar{x}_{..}$ tähistab üldist, kõikide andmete põhjal arvutatud keskmist.

Hälvete ruutude summa arvutatakse järgmiselt:

$$\sum_{j=1}^{k}\sum_{i=1}^{n}(x_{ij} - \bar{x}_{..})^2 = (6-14.5)^2 + (14-14.5)^2 + \ldots + (10-14.5)^2 = 327.0$$

Tehtud arvutused saame koondada tabelisse (nii esitatakse reeglina dispersioonanalüüsi tulemused ka statistikapakettides):

Tabel 3. Dispersioonanalüüsi tulemused

dispersiooni liik	hälvete ruutude summa	vabadusastmete arv	dispersiooni hinnang	F-statistik

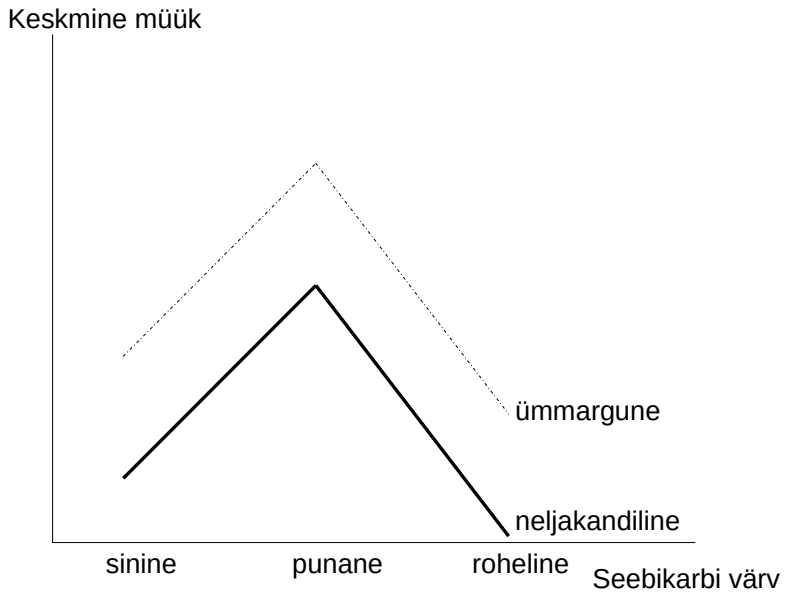
gruppidevaheline	86.0	2	43.0	1.60
gruppidesisene	241.0	9	26.8	
kokku	327.0	11		

Eelnevas näites kaasati iga kauplus analüüsi vaid üks kord. See ei ole üldine reegel, kasutada võib ka lähenemist, kus igal objektil mõõdetakse kõikide sõltumatu muutuja väärtused. Näiteks võiks sellist lähenemist kasutada juhul, kui soovitakse mõõta erinevate palavikualandajate mõju. Kuna iga indiviidi reaktsioon ravile sõltub tema füsioloogilistest eripäradest, on usaldusväärse tulemuse saamiseks otstarbekas mõõta, kuidas patsient reageerib eri ravimitele. Sellise disainiga eemaldatakse objektidevahelised juhuslikud varieerumised –on ootuspärane, et ühe indiviidi poolt kolmele reklaamile antud hinnang varieerub mõnevõrra vähem kui kolme eri indiviidi hinnang kolmele reklaamile.

Vastastikuse mõju efektid (*interaction effects*)

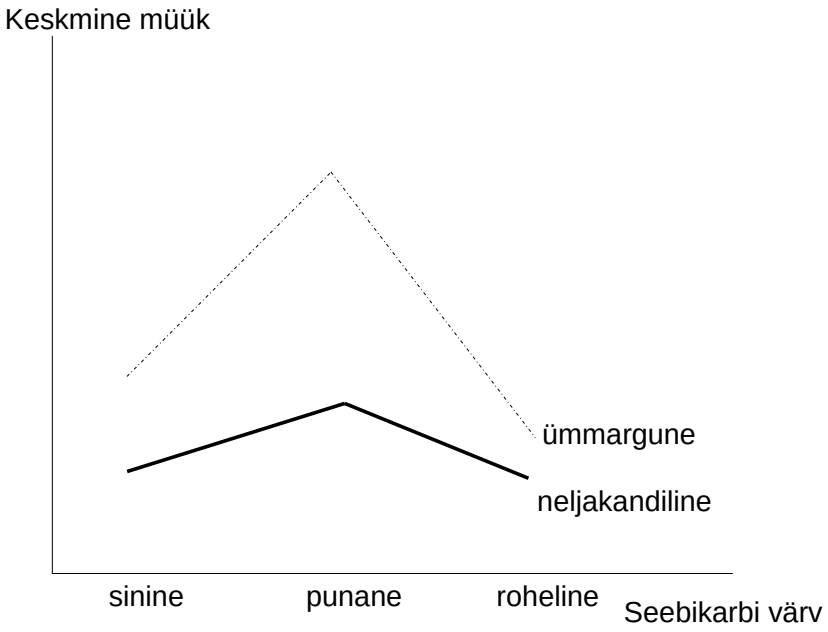
Juhul kui analüüsitakse kahe või enama sõltumatu muutuja mõju sõltuvale muutujale, tuleb alati arvesse võtta võimalust, et tegemist on vastastikuse mõju efektidega, st tuleb välja selgitada, kas ühe sõltumatu muutuja mõju sõltuvale muutujale oleneb teis(t)e sõltumatu(te) muutujate väärtustest. Kõige hõlpsam on seost kindlaks teha graafilise analüüsi abil.

Eelnevat näidet jätkates oletame, et kaasame analüüsiprotsessi täiendava muutuja – seebikarbi kuju, mis võib omandada kaht väärtust (ümmargune või nelinurkne). Joonis 3 kujutab juhtu, kus vastastikuse mõju efekte seebikarbi värvi ja kuju vahel ei ole. Kõige enam ostetakse punaseid seebikarpe ning tarbijad eelistavad ümmargusi seebikarpe.



Joonis 3. Kahefaktorilise analüüsi tulemus juhul, kui vastastikuse toime efekt puudub (jooned ühendavad andmepunkte, andmaks ülevaadet keskväärtuste erinevustest).

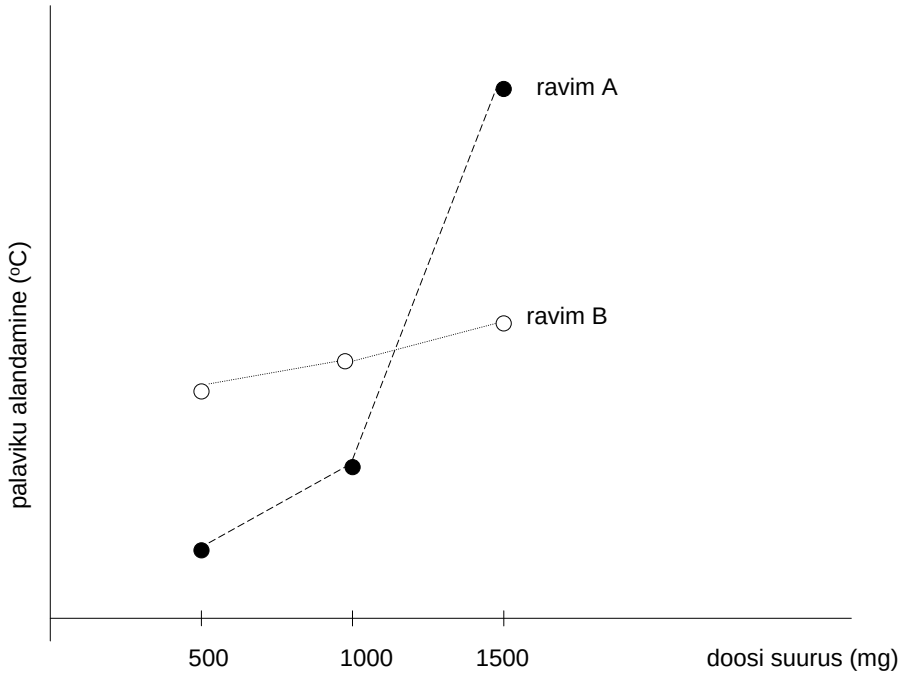
Joonis 4 kujutab olukorda, kus kahe näitaja vahel esinevad vastastikuse mõju efektid.



Joonis 4. Kahefaktorilise analüüsi tulemus juhul, kui esineb vastastikuse toime efekt (jooned ühendavad andmepunkte, andmaks ülevaadet keskväärtuste erinevustest).

Joonisel 4 esitatud juhul mõjutab värv (esimene sõltumatu muutuja) seebikarpide müüki sõltuvalt seebikarbi kujust (teine sõltumatu muutuja). Müük erineb värvi lõikes üksnes ümmarguste seebikarpide korral, neljakandiliste seebikarpide müük on kolme värvivaliku lõikes suhteliselt sarnane. Jooniselt on näha, et punaste seebikarpide korral erineb müük sõltuvalt karbi kujust kõige märgatavamalt, kõige tagasihoidlikum on erinevus roheliste seebikarpide korral.

Vastastikuse mõju efekti olemasolu väljaselgitamine on oluline kahel põhjusel. Esiteks baseerub enamik mitmefaktorilisi mudeleid eeldusel, et vastastikuse mõju efekte ei ole. Näiteks, kui analüüsi alusena kasutataks joonisel 5 esitatud andmeid, siis suure tõenäosusega näitaks F-test, et ravimite A ja B tõhusused palaviku alandamisel ei ole statistiliselt erinevad. Põhjuseks on asjaolu, et F-test põhineb näitajate keskmistel tasemetel.



Joonis 5. Vastastikuse toime efektid ravimi tüübi ning doosi suuruse vahel.

Siiski on jooniselt selgelt näha, et ravimite A ja B tõhusus on erinev sõltuvalt doosi suurusest, st esineb vastastikuse toime efekt ravimi tüübi ning toosi suuruse vahel. Ning see on näiteks teise põhjuse kohta, miks vastastikuse toime efektide avastamine on oluline. Nimelt osutub, et üksnes koguefektide vaatlemine (mitte võttes arvesse teiste muutujate väärtusi) võib viia uurija valedele järeldustele.

KOVARIATSIOONANALÜÜS *(*Analysis on Covariance*)*

Sissejuhatuseks. ANCOVA olemus

Kovariatsioonanalüüs (ANCOVA) on dispersioonanalüüsi laiendus. ANCOVA korral hinnatakse sõltumatute muutujate (IV) mõju sõltuvale muutujate pärast seda, kui viimast on kohandatud erinevuste suhtes, mis tulenevad sõltuva muutuja (DV) ja täiendavate muutujate, mida nimetatakse kovariantideks (*covariates*, CV) korreleerumisest.

ANCOVA korral otsitakse vastust küsimusele, kas sõltumatu muutuja gruppides on erinevused kohandatud sõltuva muutuja keskväärtuste vahel juhuslikud või leidub teatud seaduspära. Näiteks saab ANCOVA abil analüüsida, kas pärast ravimi manustamist on erinevused teatud kõrgveretõverohtu saanud grupi ning kontrollgrupi keskmistes vererõhu tasemetes (DV) võrdsed, kui me oleme sõltuvat muutujat kohandanud erinevuste suhtes testieelses vererõhu tasemes (CV).

ANCOVA kasutamisel on kolm peamist eesmärki. Esiteks saab ANCOVA abil vähendada vealiiget: vealiiget kohandatakse (ning loodetavasti vähendatakse), võttes arvesse DV ja CV seost. ANCOVA suurendab F-testi võimsust, eemaldades analüüsiprotsessist CV-dega seotud prognoositava varieeruvuse. See tähendab, et CV-sid kasutatakse „müra" hindamiseks, kus „müra" tähendab DV soovimatut varieerumist (st individuaalseid erinevusi), mida hinnatakse CV-de abil.

Teiseks võimaldab kõnealune meetod kohandada DV keskmisi väärtusi võttes arvesse CV-de väärtusi. Selline lähenemine on sageli kasulik olukorras, kus uurija analüüsib mitteeksperimentaalseid andmeid, nt patsiente ei saa juhuslikult jagada ravi saavateks ja mittesaavateks. Sellisel juhul võimaldab ANCOVA kasutamine eemaldada CV-dest tulenevad erinevused subjektide vahel nii, et alles jäävad üksnes IV-dega seotud erinevused. Siinkohal tuleb meeles pidada, et tegemist on kirjeldava analüüsiga, mille abil ei ole võimalik tuvastada kausaalsust.

ANCOVA olemuse mõistmiseks toome järgmise näite. Oletame, et soovitakse analüüsida, kas poliitilised eelistused varieeruvad regiooniti. DV on muutuja, mis väljendab liberalismi-konservatismi. Regioonid olgu analüüsitava riigi piirkonnad, nt põhja-, ida-, lõuna- ja lääneosa. Kaks muutujat, mis võivad eelduste kohaselt regiooniti ning poliitilise eelistuse lõikes varieeruda, on indiviidi sugu ning sotsiaalmajanduslik staatus. Need muutujad võib analüüsi kaasata kovariaatidena. ANCOVA abil püütakse antud näites leida vastus küsimusele, kas kehtib nullhüpotees, mille kohaselt poliitiline orientatsioon ei erine geograafiliste piirkondade lõikes pärast seda, kui arvesse on võetud vastaja sotsiaalmajanduslik staatus ning vanus.

Viimaks, ANCOVA abil saadud teadmisi saab edasi kasutada MANOVA (mitmemõõtmelise dispersioonanalüüsi) teostamisel.

Kõigil kolmel juhul on analüüsiprotsess sama ning sarnaselt ANOVAga jagatakse varieeruvus gruppidesiseseks ning –vaheliseks. Enne seda aga hinnatakse regressioon DV ja ühe või mitme CV vahel. Seejärel kohandatakse DV väärtusi ning keskväärtusi, eemaldamaks CV-de lineaarseid efekte. Viimaks rakendatakse kohandatud muutujatele ANOVA analüüsimeetodeid.

ANCOVA analüüsiprotsess

Sarnaselt ANOVAga jagatakse ka ANCOVA korral hälvete ruutude summa kaheks eraldi komponendiks: gruppidesiseseks ning –vaheliseks. See tähendab, et muutuja *Y* tasemete (sõltuv muutuja, DV) ning üldkeskmise (*GM*, st kõikide objektide lõikes leitud keskmise) hälvete ruutude summa koosneb kahest eraldi komponendist:

1. *j*-nda grupi keskmise ning üldkeskmise hälbe ruutude summa
2. indiviidide sõltuva muutuja taseme ning vastava grupi keskmise hälbe ruutude summa.

Vastav arvutusvalem on järgmine:

$$(1) \qquad \sum_i \sum_j (Y_{ij} - GM)^2 = n \sum_j (\bar{Y}_j - GM)^2 + \sum_i \sum_j (Y_{ij} - \bar{Y}_j)^2$$

ehk

(2) $SS_{total} = SS_{bg} + SS_{wg}$.

ANCOVA korral tuleb teha kaks täiendavat liigendust. Esiteks jagatakse CV-de korral hälvete ruutude summa samuti kaheks komponendiks – gruppidesiseseks ning -vaheliseks:

(3) $SS_{total(x)} = SS_{bg(x)} + SS_{wg(x)}$.

Teiseks jagatakse ka kovariatsioon (st lineaarne seos DV ja CV vahel) kaheks: gruppidesiseseks ning –vaheliseks tulemuste summaks (*SP, sum of products*):

(4) $SP_{total} = SP_{bg} + SP_{wg}$.

NB! Ruutude summa korral arvutatakse hälve keskmisest (st kas $X_{ij} - X_j$ või $Y_{ij} - Y_j$), võetakse saadud näitajad ruutu ning summeeritakse need üle objektide. Tulemuste summa korral arvutatakse iga objekti korral välja mõlema hälbed (st nii $X_{ij} - X_j$ kui $Y_{ij} - Y_j$), korrutatakse need (mitte ei võeta ruutu) ning seejärel liidetakse üle objektide.

Kui igas grupis on võrdne arv objekte, saab eespool toodud näitajate arvutamiseks kasutada tabelis 1 toodud arvutusvalemeid.

Tabel 1. Olulisemad arvutamisvalemid.

	gruppidevaheline	gruppidesisene
Y (DV) ruutude summa	$SS_{bg} = \frac{\sum^{k}(\sum^{n} Y)^2}{n} - \frac{(\sum^{k}\sum^{n} Y)^2}{kn}$	$SS_{wg} = \sum^{k}\sum^{n} Y^2 - \frac{\sum^{k}(\sum^{n} Y)^2}{n}$
X (CV) ruutude summa	$SS_{bg(x)} = \frac{\sum^{k}(\sum^{n} X)^2}{n} - \frac{(\sum^{k}\sum^{n} X)^2}{kn}$	$SS_{wg(x)} = \sum^{k}\sum^{n} X^2 - \frac{\sum^{k}(\sum^{n} X)^2}{n}$
tulemuste summa (SP)	$SP_{bg} = \frac{\sum^{k}(\sum^{n} Y)(\sum^{n} X)}{n} - \frac{(\sum^{k}\sum^{n} Y)(\sum^{k}\sum^{n} X)}{kn}$	$SP_{wg} = \sum^{k}\sum^{n}(XY) - \frac{\sum^{k}(\sum^{n} Y)(\sum^{n} X)}{n}$

Märkus: k = gruppide arv

n = objektide arv grupis

Võrrandeid (3) ja (4) kasutatakse gruppidevahelise ruutude summa kohandamiseks:

$$SS'_{bg} = SS_{bg} - \left[\frac{(SP_{bg} + SP_{wg})^2}{SS_{bg(x)} + SS_{wg(x)}} - \frac{(SP_{wg})^2}{SS_{wg(x)}} \right]$$

ning gruppidesisese ruutude summa kohandamiseks:

$$SS'_{wg} = SS_{wg} - \frac{(SP_{wg})^2}{SS_{wg(x)}}$$

Kui vastavad kohandused on tehtud, leitakse ülejäänud näitajad nagu tavaliselt, kuid meeles tuleb pidada, et ANCOVA korral on gruppidevahelise hinnangu korral vabadusastmete arv $k - 1$ ning gruppidesisese hinnangu korral $N - k - c$ (N – koguvalimi suurus, k – IV tasemete arv, c – CV-de arv).

Nullhüpoteesi kehtivust (mille kohaselt gruppide vahel puuduvad erinevused) testitakse F-statistiku abil, jagades kohandatud gruppidevahelise MS-i (keskruut, *mean square*) gruppidesisese MS-ga. Seejärel kontrollitakse vastava kriitiliste väärtuste tabeli abil, kas saadud tulemus on ka statistiliselt oluline.

Kui sõltumatu muutuja mõjutab statistiliselt oluliselt kohandatud sõltuvat muutujat, siis kasutatakse seose tugevuse hindamiseks η^2:

$$\eta^2 = \frac{SS'_{bg}}{SS'_{bg} + SS'_{wg}}$$

Piirangud ANCOVA kasutamisel

Teoreetilised piirangud

Sarnaselt ANOVAga ei ole ka käesoleva meetodi korral statistiliste testide abil võimalik kindlaks määrata, kas muutused DV-s on põhjustatud IV poolt. Järeldused kausaalsuse suuna kohta on pigem loogikal põhinevad, nagu ka CV-de valik. Üldreeglina soovitatakse, et CV-de arv peaks olema nii väike kui võimalik, kõik CV-d peaksid olema korreleeritud DV-ga ning ei tohiks olla üksteisega korreleeritud. Lisaks tuleb arvestada, et iga CV kaasamisel analüüsiprotsessi kaotatakse üks vabadusaste.

Igal juhul tuleb silmas pidada, et kohandatud keskmiste interpreteerimine on keerukas. Kohandatud keskmised näitavad, millised oleksid olnud keskmiste väärtused, kui kõikidel indiviididel oleksid CV-de väärtused samad.

Erindite puudumine

Analüüsitulemusi võib mõjutada erindite olemasolu. DV või CV-de hulgas esinevad erindid võivad põhjustada olukorra, kus nullhüpoteesi ei suudeta ümber lükata või saadakse analüüsi tulemusteks põhjendamatud kohandatud DV väärtused.

Multikollineaarsuse ning singulaarsuse puudumine

Kui analüüsiprotsessi kaasatakse mitu CV-d, ei tohiks need olla kõrgelt korreleeritud. **Kõrgelt korreleeritud CV-d tuleks analüüsiprotsessist eemaldada!**

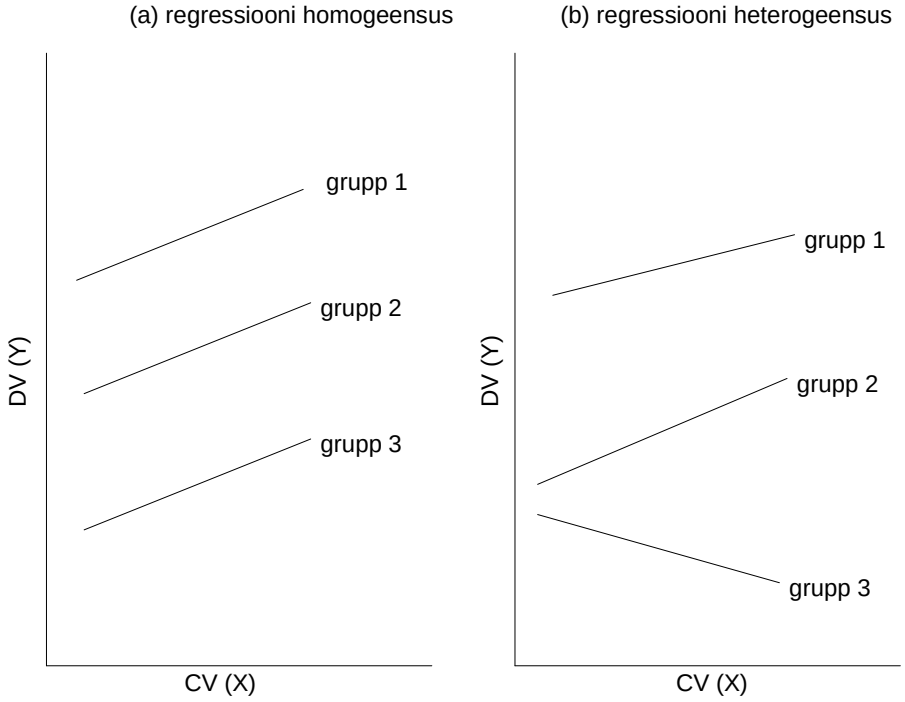
Lineaarsus

ANCOVA mudel baseerub eeldusel, et CV-de ja DV vaheline seos on lineaarne. Kui see nõue ei ole täidetud, vahendab see sarnaselt mitmese regressiooniga statistiliste testide võimsust

Regressiooni homogeensus

Eeldatakse, et DV ja CV(-de) vahelise regressiooni tõus on kõikidel gruppide võrdne (vt joonis 1). Kui see nõue ei ole täidetud, ei anna ANCOVA rakendamine õigeid tulemusi ning seda ei tohiks kasutada.

Regressiooni homogeensuse kontrollimiseks on olemas formaalsed testid, mis on aga käsitsi arvutamiseks liialt aeganõudvad. Seepärast kasutatakse antud probleemi testimiseks enamasti arvutiprogrammide abi.



Joonis 1. DV-CV regressioonijooned homogeensel (paneel (a)) ning heterogeensel (paneel (b)) juhul.

Muutujate usaldusväärsus

ANCOVA korral eeldatakse, et analüüsi kaasatud CV-de väärtused on mõõdetud veata. Mitmete muutujate (nt sugu, vanus jne) korral on antud eeldus tavaliselt täidetud, kuid teatud muutujate korral võib probleeme esineda.

Näide

Olgu vaatluse all üheksa õpiraskustega last. Sõltumatuks muutujaks (IV) on ravimeetodi tüüp, millel on kolm kategooriat (meetod I, meetod 2 ja kontroll), igasse neist kaastakse kolm last. Kõigil üheksal lapsel mõõdetakse kahe näitaja väärtused: lugemiskiirus enne ravi (CV) ning sama näitaja pärast ravi (DV). Uurimisküsimus on järgmine: kas rakendatavatel ravimeetoditel on erinev mõju lapse arenemiskiirusele pärast seda, kui arvesse on võetud erinevused laste testieelses lugemisvõimes?

Analüüsi aluseks olevad andmed on toodud tabelis 2.

Tabel 2. Algandmed

	grupid					
	ravimeetod 1		ravimeetod 2		kontroll	
	ravieelne	ravijärgne	ravieelne	ravijärgne	ravieelne	ravijärgne
	85	100	86	92	90	95
	80	98	82	99	87	80
	92	105	95	108	78	82
summa	257	303	263	299	255	257

Vajalikud näitajad arvutatakse antud näites järgmiselt:

$$SS_{bg} = \frac{(303)^2 + (299)^2 + (257)^2}{3} - \frac{(859)^2}{3 \cdot 3} = 432.899$$

$$SS_{wg} = (100)^2 + (98)^2 + (105)^2 + (92)^2 + (99)^2 + (108)^2 + (95)^2 + (80)^2 + (82)^2 -$$
$$- \frac{(303)^2 + (299) + (257)^2}{3} = 287.333$$

$$SS_{bg(x)} = \frac{(257)^2 + (263)^2 + (255)^2}{3} - \frac{(775)^2}{3 \cdot 3} = 11.556$$

$$SS_{wg(x)} = (85)^2 + (80)^2 + (92)^2 + (86)^2 + (82)^2 + (95)^2 + (90)^2 + (87)^2 + (78)^2 -$$
$$- \frac{(257)^2 + (263)^2 + (255)^2}{3} = 239.333$$

$$SP_{bg} = \frac{(257)(303) + (263)(299) + (255)(257)}{3} - \frac{(775)(859)}{3 \cdot 3} = 44.889$$

$$SP_{wg} = (85)(100) + (80)(98) + (92)(105) + (86)(92) + (82)(99) + (95)(108) + (90)(95) +$$
$$(87)(80) + (78)(82) - \frac{(257)(303) + (263)(299) + (255)(257)}{3} = 181.667$$

Saadud tulemused saab esitada maatriksitena:

$$S_{bg} = \begin{bmatrix} 11.556 & 44.889 \\ 44.889 & 432.889 \end{bmatrix}$$

ning

$$S_{wg} = \begin{bmatrix} 239.333 & 181.667 \\ 181.667 & 287.333 \end{bmatrix}$$

Kohandatud näitajad leitakse järgmiselt:

$$SS'_{bg} = 432.889 - \left[\frac{(44.889 + 181.667)^2}{11.556 + 239.333} - \frac{(181.667)^2}{239.333} \right] = 366.202$$

$$SS'_{wg} = 287.333 - \frac{(181.667)^2}{239.333} = 149.438$$

Saadud väärtused kantakse järgmisse tabelisse (vt tabel 3).

Tabel 3. ANCOVA tabel

variatsiooni liik	kohandatud hälvete ruutude summa	vabadusastmete arv	MS	F-statistik
gruppidevaheline	366.202	2	183.101	6.13*
gruppidesisene	149.439	5	29.888	

*p<0.05.

Käesolevas näites on F-statistiku kriitiliseks väärtuseks 5.79 ($\alpha = 0.05, \quad df = 2,5$) ning seega lükatakse nullhüpotees ümber, st ravi liik mõjutab ravijärgset lugemiskiirust.

Seose tugevus on arvutatav järgmiselt:

$$\eta^2 = \frac{366.202}{366.202 + 149.438} = 0.71$$

ning see näitab, et 71% kohandatud sõltuva muutuja väärtuste hajuvusest on seletatav ravi liigiga.

Antud andmetele vastav ANOVA tabel on järgmine.

Tabel 4. ANOVA tabel

variatsiooni liik	Hälvete ruutude summa	vabadusastmete arv	MS	F-statistik
gruppidevaheline	432.889	2	216.444	4.52
gruppidesisene	287.333	6	47.889	

Kaht tabelit võrreldes on näha, et ANOVA korral on hälvete ruutude summa suurem kui ANCOVA puhul. Ka vabadusastmete arv on gruppidevahelisel hinnangul ühe võrra suurem, kuna puudub CV. Kõige olulisem on aga see, et ANOVAt kasutades jäädaks antud korral nullhüpoteesi juurde, ANCOVAt aluseks võttes aga lükatakse nullhüpotees ümber. Seega on CV kasutamine antud näites vealiikme „müra" vähendanud.

Kokkuvõtteks

Kovariatsioonanalüüs on dispersioonanalüüsi edasiarendus, mis võimaldab arvesse võtta ka täiendavate pidevate sõltumatute muutujate, mida nimetatakse kovariantideks, mõju sõltuvale

muutujale. Teatud juhtudel erineb kovariatsioonanalüüsi järeldus dispersioonianalüüsi tulemustest. Sel juhul on kovariantide lisamine õigustatud.

KLASTERANALÜÜS (Cluster Analysis)

Sissejuhatuseks

Klasteranalüüs[4] on mitmemõõtmelise statistika meetod, mille peamiseks eesmärgiks on jagada objektid (nt indiviidid, ettevõtted, tooted, käitumine) teatud nendele omaste tunnuste alusel gruppidesse[5], mida nimetatakse **klastriteks**. Klasteranalüüs sarnaneb mõneti samas ainekursuses käsitletava faktoranalüüsiga, kuid peamiseks erinevuseks nende kahe analüüsitehnika vahel on asjaolu, et klasteranalüüs grupeerib objekte, samal ajal kui faktoranalüüs on suunatud muutujate grupeerimisele.

Klasteranalüüsi käigus grupeeritakse objektid klastritesse nii, et ühte klastrisse kuuluvad objektid on teatud (eelnevalt kindlaksmääratud) valikukriteeriumi põhjal sarnasemad samasse klastrisse kuuluvate kui teistes klastrites asuvate objektidega. See tähendab, et eesmärgiks on ühtaegu maksimeerida nii klastritesisest objektide homogeensust kui klastritevahelist heterogeensust. Seeläbi on võimalik uurijal saada minimaalse informatsioonikaoga sisutihedamat ning kergemini haaratavat informatsiooni vaatluste kohta.

Kuna klasteranalüüsi tulemused ning nende tõlgendamised on suuresti uurijaspetsiifilised, siis on vahel väidetud, et **klasteranalüüs on pigem kunst kui teadus**. Siiski on tegemist meetodiga, mis kogub ühe enam populaarsust oma lihtsuse ning hõlpsa ning ülevaatliku graafilise esituse tõttu.

Klasteranalüüsi olemus ning muutujate valik

Klasteranalüüsi eesmärk on tihedalt seotud selles kasutatavate muutujate valimisega. Analüüsi käigus formuleeritud klastrid peegeldavad ju vastavalt valitud muutujatele andmetele omast struktuuri. Paraku ei ole klasteranalüüsi protsessi sisemiselt sisseehitatud mehhanisme oluliste ning ebaoluliste muutujate eristamiseks. Muutujate valikul tuleb uurijal seetõttu arvestada teoreetiliste, kontseptuaalsete ning praktiliste kaalutlustega ning muutujate valikul saab toetuda näiteks varasematele uurimustele või konkreetsele teooriale. Tuleb silmas pidada, et valitavad muutujad peaksid 1) iseloomustama klastritesse jagatavaid objekte ning 2) olema tihedalt seotud klasteranalüüsis püstitatud eesmärkidega. Ebaoluliste muutujate kaasamine suurendab erindite

[4] Klasteranalüüsi sünonüümidena kasutatakse mõisteid Q-analüüs (*Q- analysis*), numbriline süstemaatika (*numerical taxonomy*), tüpoloogia konstruktsioon (*typlology construction*), klassifikatsioonianalüüs (*classification analysis*).

[5] Kuigi antud peatükis tutvustatakse eelkõige klasteranalüüsi kasutusvõimalusi majanduses, on sama analüüsivõtet võimalik ning sageli vajalik kasutada ka teistel elualadel, nimetagem siinkohal näiteks reaal- või sotsiaalaineid.

esinemise võimalust ning seeläbi ka tulemuste mitteadekvaatsuse ohtu. Seega ei tohiks muutujaid kunagi analüüsi kaasata valimatult, kvaliteetse tulemuse saamiseks on oluline valida muutujaid hoolikalt ning valida edasise uurimise tarvis üksnes need, mis täidavad eelnevat kaht nõuet.

Klasteranalüüsis on oluliseks mõisteks **klastermuutuja** (*cluster variate*). Klastermuutuja puhul on tegemist kombineeritud muutujaga (nagu faktoranalüüsi, MANOVA ning diskriminantanalüüsi korral), st tegemist on teatud muutujate kombinatsiooniga. Kombineeritud muutuja aitab objekte võrrelda ning selle alusel toimub objektide jaotamine klastritesse. Siinkohalt tuleb rõhutada, et klasteranalüüs on ainus mitmemuutujatehnika, mis ei hinda kombineeritud muutuja saamiseks muutujate süsteemi.

Klasteranalüüsi eeldused

Klasteranalüüs pole erinevalt enamikest antud ainekursuse raames tutvustatavatest analüüsitehnikatest, kus valimi põhjal saadud parameetrite hinnangute põhjal tehakse järeldusi üldkogumi kohta, suunatud statistilise järelduste tegemisele. Seetõttu kasutatakse klasteranalüüsi käigus kindlat matemaatilist aparatuuri, kuid statistilisi eeldusi, mille kontrollimine on enne analüüsi algust oluline, on vähe. Viimastest on väga olulised valimi esinduslikkus, multikollineaarsus ning erindite puudumine.

Valimi esinduslikkus

Valdaval enamikel juhtudel on uurija käsutuses valimi, mitte üldkogumi andmed. Kuna üldiselt eeldatakse, et valimi andmetele tuginedes moodustatud klastrid esindavad piisava adekvaatsusega üldkogumit, siis peab uurija olema veendunud, et valim on üldkogumi suhtes samuti esindav.

Multikollineaarsuse puudumine

Klasteranalüüsi puhul toob multikollineaarsete muutujate kasutamine kaasa olukorra, kus multikollineaarseid muutujaid kaalutakse analüüsi käigus tugevamalt. Seega peab uurija analüüsima, kas kaasatud muutujate puhul on olemas multikollineaarsuse oht. Sellise ohu tuvastamisel on otstarbekas kõrgelt korreleeruvatest muutujatest üks analüüsist välja jätta või kasutada sarnasuse mõõduna Mahalanobise distantsi, mis võtab muutujate korreleeritust arvesse (vt peatükki sarnasuse mõõtmine).

Erindite puudumine

Andmetes struktuuri otsimisel on klasteranalüüs väga tundlik ebaoluliste muutujate ning erindite kaasamise suhtes, kuna erindid moonutavad andmete tõelist struktuuri. Seejuures ei tule mitte alati erindid analüüsiprotsessist automaatselt eemaldada, kuna erindid võivad peegeldada:

1. üldkogumi andmete tegelikku struktuuri
2. „hälbeid", st nad ei peegelda üldkogumi struktuuri.

Seega, kui erindid tuvastatakse, peab uurija analüüsima, kas need iseloomustavad uuritavat kogumit ning tuleks analüüsiprotsessi jätta, kuna nende eemaldamine tooks kaasa nihkega tulemused või tuleks need analüüsist eemaldada, kuna nad ei ole uuritavale kogumile tüüpilised.

Andmete standardiseerimine klasteranalüüsis

Muutujate standardiseerimine

Klasteranalüüsi teostamisel tuleb arvestada, et enamik distantse on tundlikud muutujate erinevate skaalade valiku suhtes ning sellest tulenevalt võib skaala valik järjestust muuta. Nii võib järjestus sõltuda sellest, kas kasutada analüüsis sekundeid, minuteid või tunde; meetreid või kilomeetreid jne. Üldjuhul on suurema dispersiooniga muutujatel on lõplikule sarnasuse mõõdule väärtusele suurem mõju. Mõõtmisviisist sõltuvate tulemuste vältimiseks võib osutuda otstarbekaks andmeid enne klastrite moodustamist standardiseerida (vt lisa 1). Alternatiiviks on leida näitajad mitme erineva skaala korral ning võrrelda saadud tulemusi ka teoreetilisest aspektist lähtuvalt.

Kõige tavalisem võimalus muutujate standardiseerimiseks on iga muutuja arvestada ümber standardsesse skoori (tuntud ka kui Z-skoor) lahutades muutuja väärtusest keskväärtuse ning jagades selle muutuja standardhälbega. Sellisel juhul konverteeritakse algsed andmed standardiseeritud väärtusteks keskväärtusega 0 ning standardhälbega 1. Standardiseerimine tagab, et skaala muutudes standardiseeritud skoorid ei muutu. Seega, olenemata sellest, kas mõõta näiteks teleri vaatamise aega minutites või tundides, nende standardiseeritud skoorid on samad ning seega on identsed ka klasteranalüüsi tulemused.

Samas ei ole standardiseerimine alati vajalik ning sellest tuleks hoiduda, kui muutujate skaalas on „loomulik" seos. Otsus, kas kasutada standardiseerimist või mitte, peaks tuginema nii empiirilistel kui kontseptuaalsetel kaalutlustel.

Objektide standardiseerimine

Lisaks muutujatele on teatud juhtudel otstarbekas standardiseerida ka objekte (indiviide, ettevõtteid jne). Nimelt osutub, et indiviidide vastused näiteks rahuloluküsimustele kipuvad tihti olema sarnased: teatud osa vastajatest on kõigega rahul (n.-ö. jah-tüüpi inimesed), teatud osa on kõikide küsitud aspektide osas negatiivselt meelestatud (n.-ö. ei-tüüpi inimesed). See tähendab, et kasutades küsitlustes jah/ei tüüpi küsimusi, võib saadavate vastuste põhjal sageli näha **vastukaja-stiili efekte** (*response-style effects*), st teatud indiviidid kalduvad vastama küsimustele positiivselt, teised negatiivselt.

Otsustamaks, kas kasutada objektide standardiseerimist või mitte, tuleb uurijal määratleda, milliseid tulemusi saada soovitakse. Kui gruppe soovitakse formuleerida küsitletute vastukaja-stiili

efektide põhjal, pole standardiseerimine vajalik. Siiski soovitakse enamikul juhtudel teada saada ühe muutuja olulisust teise suhtes ning sel juhul on otstarbekas vaatlusi standardiseerida. Sellisel juhul ei kasutata standardiseerimisel mitte objektide, vaid objekti muutujate keskmist skoori. Seda tuntakse kui **juhtumisisest** (*within-case*) või **reakeskset** (*row-centering*) **standardiseerimist**.

Klastrite formuleerimise protsess

Klasteranalüüsi teostamisel tuleb uurijal leida vastused järgmisele küsimustele:

1. kuidas mõõta objektidevahelist sarnasust?

2. millistel alustel formuleerida klastrid?

3. mitu klastrit luua?

Objektide vahelise sarnasuse mõõtmine

Kuna sarnasuse kontseptsioon on klasteranalüüsi aluseks, on oluline mõista, kuidas mõõta objektidevahelist sarnasust. Selleks on sisuliselt võimalik kasutada ühe kahest võimalusest: **sarnasuse** või **erinevuse** mõõtu. Üheks tuntuimaks sarnasuse mõõduks on **korrelatsioonikoefitsiendid**[6]. Kuna korrelatsioonanalüüsi on antud loengukursuse raames käsitletud, siis antud teema korral peatutakse pikemalt erinevuse mõõtudel, mida sageli nimetatakse **distants**iks (*distance*). Distantsi abil mõõdetakse tegelikkuses objektide erinevust, näitaja suuremad väärtused viitavad ulatuslikematele erinevustele (ehk teisisõnu väiksemale sarnasusele) objektide vahel. Klasteranalüüsi raames tuntakse erinevaid distantsi mõõtusid, neist enimkasutatav on **Euclideani distants**, mille arvutamisalgoritm kahe objekti (*p* ja *q*) ning *n* muutuja korral on järgmine:

$$Dist_{EUCL} = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + ... + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}\ .$$

Antud valemi põhjal saadakse lihtne Euclideani distants, mille põhjal võib hõlpsalt leida näitaja ruudu või absoluutväärtuse. Euclideani distantsi ruutu kasutatakse Wardi ning tsentroidi meetodi korral (vt edaspidi).

***City-block* ehk *Manhattani* ehk absoluutse distantsi** korral kasutatakse erinevuse mõõdu arvutamisel näitajate absoluutsete erinevuste summat (selgituseks: Euclideani distantsi korral kasutatakse erinevuste ruutude summat).

---

[6] Sellisel juhul kasutatakse objektidevahelise sarnasuse tuvastamisel mitut muutujat. Objektide eristamiseks kasutatavad muutujad koondatakse sel juhul maatriksisse, kus veerud esindavad objekte ning read muutujaid. Seega mõõdab kahe veeru vaheline korrelatsioonikoefitsient kahe objekti vahelist sarnasust (korrelatsiooni). Nagu ikka, kõrge korrelatsioonikoefitsient viitab objektide sarnasusele ning madal korrelatsioon selle puudumisele. Üldiselt kasutatakse korrelatsioonikoefitsiente klasteranalüüsi teostamisel harva, kuna objektide sarnasuse tuvastamiseks eksisteerivad paremad meetodid.

$$Dist_{CB} = |p_1 - q_1| + |p_2 - q_2| + ... + |p_n - q_n| = \sum_{i=1}^{n} |p_i - q_i|$$

City block distantsi kasutamine seondub mitmete probleemidega, kirjanduses on välja toodud, et nimetatud sarnasuse mõõt eeldab muuhulgas, et muutujad ei ole korreleeritud, mis võib teatud juhtudel osutuda suhteliselt piiravaks eelduseks. Kui näitajad on korreleeritud, on city-block distantsi kasutamisel saadud klastrid mittevaliidsed.

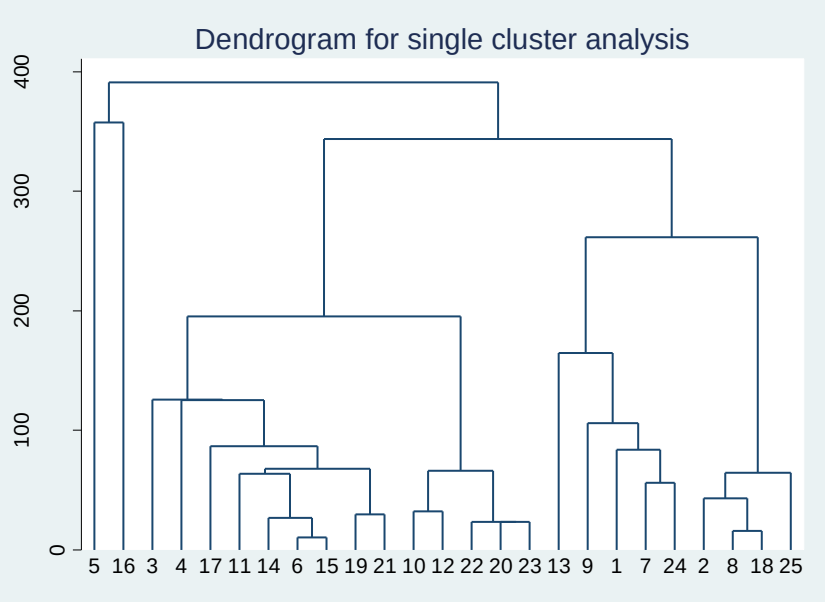
Mahalanobise distants (D^2) toetub otseselt standardiseeritud andmetele, kasutades kaaludena andmete standardhälbeid; samuti summeerib näitaja gruppidesisese dispersiooni-kovariatsiooni, mis kohandab muutujatevahelist korrelatsiooni, olles seega analoogne regressioonanalüüsis kasutatava R^2 **–**ga. Kõrgelt korreleeritud muutujad võivad kaasa tuua olukorra, kus klasteranalüüsis ülehinnatakse teatud muutujate mõju ning Mahalanobise distants võimaldab seda probleemi vältida.

Kui tegu on kategooriliste muutujatega, siis kasutatakse objektide klasterdamisel ühendumise **sarnasuse mõõtu**. Näiteks jah/ei vastuste puhul näitab ühendumise mõõt, kui suur osa küsitletavate paaridest nõustus (mõlemad vastasid jaatavalt või eitavalt. NB! TEOSTATAKSE PAARIDE KAUPA!). Loomulikult eksisteerib ka märksa keerukamaid viise sarnasuse mõõtmiseks.

Objektide jagamine klastritesse

Objektide jagamisel klastritesse võib kasutada kas **hierarhilist** või **mittehierarhilist meetodit**. Hierarhilise meetodi korral ei ole klastrite arv a priori uurija poolt kindlaks määratud ning uued klastrid moodustatakse alati olemasolevate põhjal. Klastrite formuleerimise meetodist olenevalt eristatakse hierarhilise meetodi korral liitmis- ja jaotamismeetodit. **Liitmismeetodi** (*agglomerative hierarchical clustering*) korral moodustab protsessi alguses iga objekt omaette klastri ning integreeritakse samm-sammult kaks kõige lähedasemat klastrit üheks, kuni kõik vaatlused kuuluvad samasse klastrisse. Seega moodustatakse uued klastrid olemasolevate põhjal. Kui klastritesse jagamist alustatakse liitmisprotsessile vastupidiselt, siis on tegemist **jaotusmeetodiga (***divisive hierarchical clustering***)**. Sel juhul alustatakse ühest suurest klastrist, kuhu kuuluvad kõik vaatlusalused objektid. Järgnevatel etappidel tuvastatakse kõige erinevamad objektid ning selle põhjal moodustatakse väiksemad klastrid. Protsess jätkub seni, kuni iga vaatlus moodustab omaette klastri. Üldiselt kasutatakse rohkem siiski liitmismeetodit.

Hierarhilist klastritesse jagamise protsessi saab kujutada ka graafiliselt ning selleks on mitmeid võimalusi. Näiteks saab seda näidata dendrogrammi abil. Dendrogramm sarnaneb puuga, kus horisontaalteljel tuuakse objektid ning vertikaalteljel distantsi väärtus.

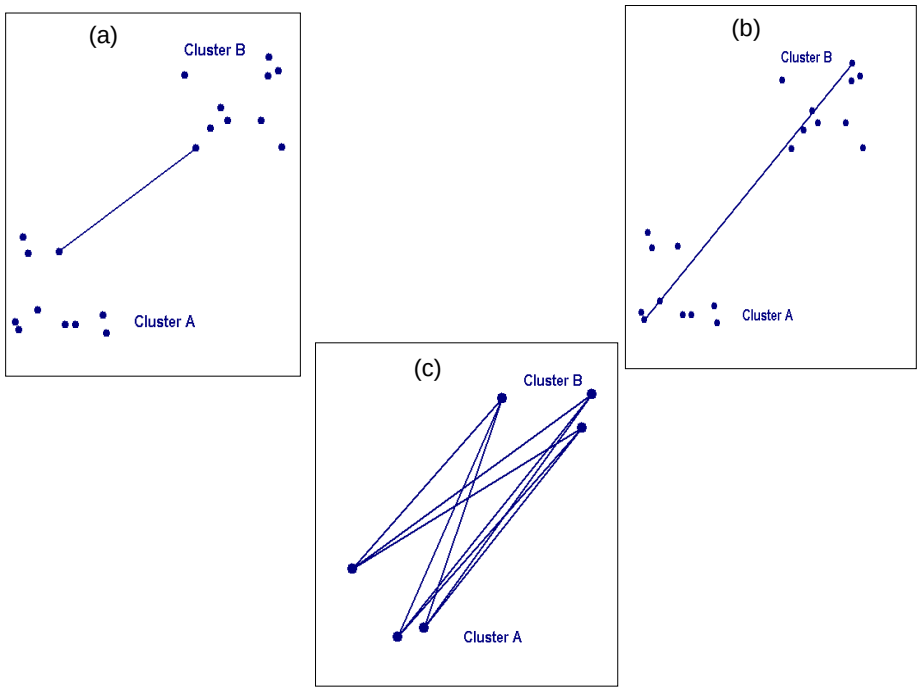


Joonis 1. Dendrogramm.

Hierarhilistest meetoditest kasutatakse praktikas mitmeid meetodeid, mis eristuvad klastrite distantsi definitsiooni põhjal. Enimkasutatavad on

- minimaalse seose (*single linkage*) meetod – klastritevaheliseks distantsiks on vastavate klastrite kõige lähemalasuvate objektide vaheline distants. Teatud juhtudel tekitab üksikseostatuse meetod ussiga sarnanevaid klastreid, kus ühte klastrisse kuuluvad objektid on üsna erinevad.

- täieliku seose (*complete linkage*) meetod –klastritevaheliseks distantsiks on kõige kaugemate liikmete vaheline distants. Täieliku seostatuse meetod võimaldab vältida pikkade klastrite teket.
- keskmise seose (*average linkage*) meetod – klastritevaheliseks distantsiks on klastri keskmine distants
- tsentroidi seose (*centroid linkage*) meetod – klastritevaheliseks distantsiks on klastrite tsentroidide (keskmiste vektorite) vaheline distants
- Wardi meetod – erineb kõikidest eespool toodud meetoidest, kuna klastritevahelise distantsi mõõtmisel kasutatakse dispersioonanalüüsi. Wardi meetodi korral minimeeritakse igal etapil kahe (hüpoteetilise) klastri hälvete ruutude summat. Eeldab, et klastrite arvu kasvades muutub informatsioon ebatäpsemaks, kuna ühendatakse üha erinevamaid objekte.



Joonis 2. Klastrite distantside mõõtmine (a) minimaalse seose, (b) täieliku seose ning (c) keskmise seose meetodil.

Mittehierarhilised klastritesse jagamise meetodid (tuntuimad *K-means clustering* ja *K-medians clustring*) ei kasuta puule sarnanevat konstruktsiooniprotsessi, objektide jaotamine klastritesse toimub pärast seda, kui klastrite arv on kindlaks määratud. See tähendab, et näiteks seitsme vaatluse põhjal formuleeritud kuueklastriline lahend ei ole saadud mitte kahe kõige sarnasema objekti klastrisse koondamise tulemusena nagu hierarhilise meetodi korral, vaid kuueklastriline lahend leitakse nii, et see oleks teatud kriteeriumide alusel parim. Selleks leitakse esmalt **klastri seeme** (*cluster seed*) kui esialgne klastri keskpunkt ning seejärel kaasatakse klastrisse kõik objektid, mis asuvad klastri keskpunktist teatud kaugusel. Seejärel määratakse kindlaks järgmise klastri keskpunkt ja klastrisse koondatakse taas objektid, mis asuvad keskpunktist teatud kaugusel jne. Sisuliselt on tegu iteratiivse protsessiga, kus lõpplahend tagab etteantud arvu klastrite korral parima võimaliku objektide jaotuse (st minimaalse klastritesisese ning maksimaalse klastritevahelise varieeruvuse). Tuntuimad mittehierarhilised meetodid on järjestikuse moodustamise (*sequential treshold*), paralleelse moodustamise (*parallel treshold*) ning optimiseerimise meetod.

Järjestikuse moodustamise meetodi korral valitakse protsessi alguses välja üks klastri seeme ja koondatakse loodud klastrisse kõik objektid, mis asuvad seemnest eelnevalt kindlaksmääratud kaugusel. Kui klaster on moodustatud, valitakse järgmine seeme ja korratakse protsessi. Kui objekt on kaasatud ühte klastrisse, siis seda järgmiste klastrite moodustamisel enam ümber ei paigutata. Paralleelse moodustamise meetodi korral valitakse simultaanselt mitu klastri seemet ning objekt kaasatakse klastrisse, mille kaugus klastri seemnest on vähim. Protsessi jooksul on võimalik

distantse kohandada. Teatud juhtudel võib aga mõni objekt jääda klasterdamata, kui see on kõikidest klastri seemnetest liiga kaugel.

Optimiseerimise meetod sarnaneb eespool käsitletud mittehierarhilise klasteranalüüsi meetoditega, kuid lubab objektide ümberpaigutumist.

Ühest vastust küsimusele, kas kasutada hierarhilist või mittehierarhilist klastrite moodustamise meetodit, anda ei saa. Kumb osutub otstarbekamaks, sõltub konkreetsest uurimisprobleemist. Põhjuseks, miks ajalooliselt on rohkem kasutatud hierarhilisi meetodeid, on asjaolu, et hierarhiliste meetodite korral on klastrite moodustamine märksa hõlpsam, samas ei ole tänapäeva arvutitele probleemiks ka mittehierarhilise klasteranalüüsi teostamine. Hierarhilise meetodi puudustena on välja toodud ka seda, et tulemused on tundlikud erindite kaasamise suhtes ning erindite olemasolul võivad tulemused olla tegelikkust mittepeegeldavad. Samuti ei sobi hierarhilised meetodid suurte valimite korral klasteranalüüsi teostamiseks. Mittehierarhiliste meetodite kasutamisel on põhiküsimuseks klastri seemnete õige määramine, samas on meetod märksa vähemtundlikum erindite, mitteoluliste muutujate ning kasutatava sarnasuse mõõdu suhtes. Mittehierarhilise meetodi puudusena võib välja tuua ka asjaolu, et juhul kui klastri seeme määratakse juhuslikult, on iga kord tulemus mõnevõrra erinev. Seepärast kasutatakse sageli analüüsis mõlema meetodi kombinatsiooni.

Klastrite arvu valik

Hierarhilise klastritesse jaotamise puhul on võimalik valida suure arvu klastrite vahel. Seega tõstatub küsimus, mitu klastrit oleks antud juhul mõttekas moodustada. On teada, et mida vähem klastreid moodustada, seda hõlpsam on tulemusi tõlgendada, kuid seda erinevamad on samasse klastrisse kuuluvad objektid. Uurijal tuleb seega teha teataval määral kompromiss klastrite arvu ning nende homogeensuse vahel: mida vähem on klasterid, seda lihtsam on teha järeldusi, samas seda suurem on ka klastritesisene heterogeensus.

Klastrite arvu valimise hõlbustamiseks on välja töötatud mitmeid formaalseid teste, kuid lõpliku otsuse peaks uurija siiski tegema ka praktilistele kaalutlustele tuginedes.

Klastrite tõlgendamine

Kui klastrid on moodustatud, on järgmiseks oluliseks etapiks saadud tulemuste sisuline tõlgendamine. Interpretatsioonietapis analüüsitakse iga klastrit, et anda klastritele nende olemust võimalikult täpselt iseloomustav nimetus. Tõlgendamine tähendab antud kontekstis eelkõige klastrite olemuse täpset ning adekvaatset kirjeldamist. Näiteks, kui turuanalüütik uurib inimeste suhtumist dieet- ja harilikesse karastusjookidesse, siis tuleks iga grupi puhul leida **klastri tsentroid** (antud näitaja keskmine väärtus vaatlusaluses grupis). Siinkohal tuleb meeles pidada, et kui eelnevas analüüsis on kasutatud standardiseeritud andmeid, tuleks tõlgendusetapis kasutada originaalandmeid. Selle abil saab hinnata, milline on igale klastrile omane suhtumine dieetjookidesse, milliste karakteristikutega inimesed eelistavad suhkruga karastusjooke jne. Sellest

lähtuvalt peaks analüütik valima igale klastrile selle olemust võimalikult hästi kirjeldava üldnimetuse.

Klastrite tõlgendamine ei seisne üksnes neile nimetuste andmises, vaid hõlmab ka uurijapoolset analüüsi, kas saadud tulemused on kooskõlas teooria ja varasemate uurimuste tulemusega. Erinevuste, aga ka sarnasuste korral tuleb uurida võimalikke põhjusi.

Klastrite hindamine ja profileerimine

Hindamisprotsessi käigus peab uurija hindama, kas valimi põhjal saadud tulemused on omased kogu üldkogumile, kas tulemusi võib üldistada ning kas need on ajas stabiilsed. Kõige lihtsam võimalus selle kontrollimiseks on analüüsida erinevaid valimeid. Kuna see on sageli aja- ja rahamahukas ning samuti võib nappida sobivaid objekte, keda (mida) analüüsi kaasata, siis praktikas jagatakse olemasolev valim sageli lihtsalt kaheks ning võrreldakse saadud tulemusi omavahel.

Profileerimisetapp algab pärast klastrite moodustamist ning on suunatud peamiselt klastrite erinevuste põhjuste väljaselgitamisele. Profileerimisetapil kasutab uurija varasemas analüüsis väljajäetud informatsiooni (näiteks demograafilised karakteristikud, psühholoogilised faktorid, tarbimismudelid jne). Kasutades diskriminantanalüüsi, selgitatakse, kuidas nimetatud tegurid mõjutavad erinevatesse klastritesse kuuluvaid inimesi. Jätkates varasemat näidet, võib profileerimisetapil selguda, et tervislikke tooteid eelistavad paremini haritud ning kõrgema sissetulekuga inimesed; et noorte jaoks on oluline toodete sisaldus, samas vanemad inimesed pööravad sellele vähem tähelepanu jne.

Kokkuvõtteks

Klasteranalüüs on mitmemõõtmelise statistika meetod, mis võimaldab jagada objekte teatud nendele omaste karakteristikute põhjal gruppidesse, mida nimetatakse klastriteks. Seeläbi on võimalik leida olemasolevates andmetes struktuur, mida seal korrastamata andmete alusel ei leita.

Klastrite arvu valik nõuab autori otsust ning on selgelt subjektiivne. Sellest tulenevalt on mitmed teoreetikud leidnud, et klasteranalüüsi, nagu ka faktoranalüüsi näol, on märksa enam tegemist kunsti kui teadusega ning analüüsiprotsessi igal etapil tuleb uurijal teha iseseisvalt otsuseid, mis muudab kogu tulemuse selgelt sõltuvaks konkreetsest uurijast.

Näide

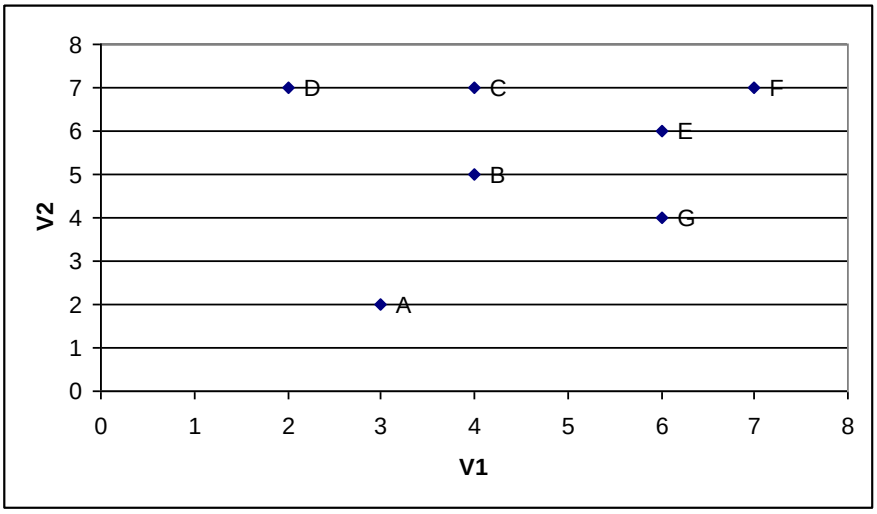
Klasteranalüüsi olemust selgitame järgneva kahemuutujalise näitega. Oletagem, et turuanalüütik soovib määratleda teatud kaupluse külastajate lojaalsust tootele ning lojaalsust kauplusele. Selleks kasutatakse pilootvaatluse korras väikest valimit (seitse inimest) kaupluse klientidest. Kasutatakse kaht lojaalsust näitavat muutujat - lojaalsust brändile (V_1) ning lojaalsust kauplusele (V_2) ning kummagi näitaja puhul kasutatakse skaalat nullist kümneni.

Kõigi seitsme vastaja hinnangud kummalegi näitajale on toodud tabelis 1.

Tabel 1. Näite algandmed

	vastaja						
klastermuutuja	A	B	C	D	E	F	G
V1	3	4	4	2	6	7	6
V2	2	5	7	7	6	7	4

Vastav punktdiagramm on toodud joonisel 3.



Joonis 3. Küsitletavate hinnangud muutujatele V_1 ja V_2.

Tabel 3 sisaldab seitsme küsitletu vastuste läheduse hinnanguid. Kasutades distantsi vastuste läheduse mõõduna, tuleb meeles pidada, et väiksem distants viitab kõrgemale sarnasusele, seega näiteks vastajad E ja F on valitud karakteristikute (lojaalsus brändile ja kauplusele) põhjal kõige sarnasemad ning A ja F kõige erinevamad.

Tabel 2. Vaatluste vahelised 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	-

Tabelis 3 on näidatud, kuidas seitse klastrit kombineeritakse üheks.

Tabel 3. Objektide jaotamine klastritesse.

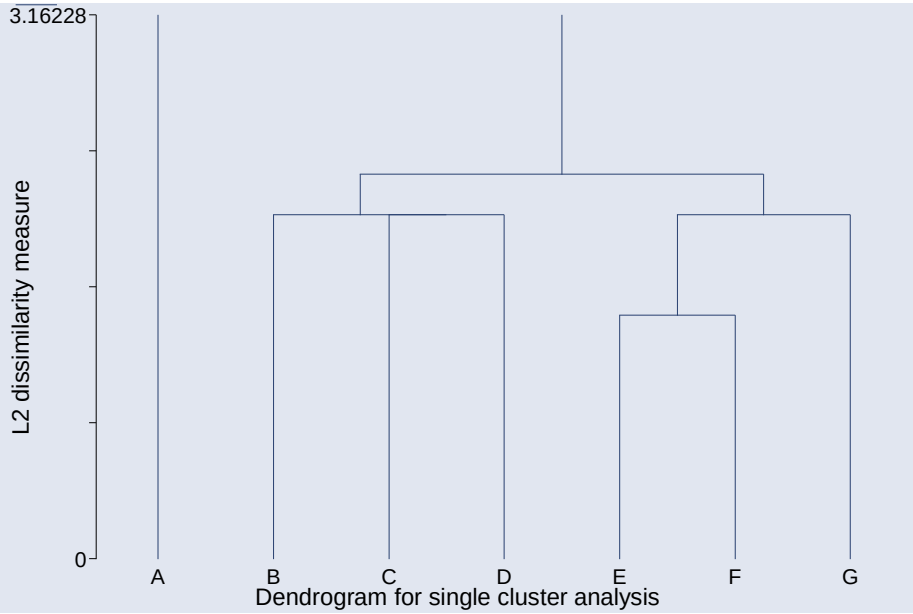
	liitmisprotsess		Klastriotsus		
Aste	Min. distants [7]	Vaatluste paar	Klasterite suhe	Klastrite arv	Üldine sarnasuse mõõt[8]
			(A) (B) (C) (D) (E) (F) (G)	7	0
1	1.414	E-F	(A) (B) (C) (D) (E-F) (G)	6	1.414
2	2.000	E-G	(A) (B) (C) (D) (E-F-G)	5	2.192
3	2.000	C-D	(A) (B) (C-D) (E-F-G)	4	2.144
4	2.000	B-C	(A) (B-C-D) (E-F-G)	3	2.234
5	2.236	B-E	(A) (B-C-D-E-F-G)	2	2.896
6	3.162	A-B	(A-B-C-D-E-F-G)	1	3.420

Esimesel sammul identifitseeritakse kaks kõige lähedasemat vaatlust (E ja F) ning kombineeritakse need üheks klastriks, vähendades klastrite arvu seitsmelt kuuele. Teisel sammul leitakse järgmised kõige lähedasemad vaatlused. Antud näite puhul on kolm paari sama distantsiga 2.000 (E-G, C-D, B-C) ning ei ole vahet, millisest paarist alustada. Kui alustada näiteks paarist E-G, siis G on üksik

[7] Minimaalne distants klastrisse mittekuuluvate vaatluste vahel, kasutatakse tabelis 3 toodud Euclideani distantse.
[8] Keskmine klastritesisene distants.

klaster, kuid E on juba eelmisel sammul ühendatud F-ga. Seega moodustub teisel etapil uus kolmest vaatlusest koosnev klaster jne. Sel viisil ühendamist jätkates vähendatakse klastrite arvu seitsmelt ühele. Kuuendal sammul ühendatakse vaatlus A ülejäänud ühte klastrisse koondatud vaatlustega. Kui vaadata tabelit 2, siis on näha, et teatud tegurite vaheline distants on väiksem kui 3.162, kuid kõik need vaatlused on juba koondatud ühte klastrisse ning neid ühendamisel ei kasutata. See lähenemine võimaldab kergesti eristada ka erindeid (käesoleval juhul on selleks vaatlus A).

Joonisel 4 on esitatud tulemusi illustreeriv dendrogramm.



Joonis 4. Dendrogramm.

Klastrite arvu valikul tuleb meeles pidada, et uurija eesmärgiks on leida kõige lihtsam objektide struktuur, mis seejuures esindaks võimakult hästi homogeenseid gruppe. Ka näite puhul oli näha, et kui vaadelda üldist sarnasuse näitajat (tabel 3 viimane veerg), siis klastrite arvu vähenedes näitab üldise sarnasuse mõõdu järsk kasv, et kaks klastrit, mis ühendati, et ole väga sarnased. Toodud näite puhul kasvab üldine sarnasuse mõõt märgatavalt esimesel etapil, kui ühendatakse esmakordselt kaks vaatlust ning samuti kasvab näitaja märgatavalt teisel sammul. Järgmisel kahel etapil (3. ja 4. etapil) ei muutu näitaja oluliselt, kuid 5. etapil, kui ühendatakse kaks kolmeliikmelist klastrit, on taas täheldatav näitaja suurem kasv. Sama toimub ka 6. etapil, mis viitab sellele, et kuigi vaatlus A ühendati ülejäänutega alles viimasel etapil, muudab tema lisamine klastrisse oluliselt klastri homogeensust. Seega, arvestades vaatluse A unikaalsust, oleks see ilmselt mõttekas koondada **entroopia gruppi**, kuhu kuuluvad erindid ning vaatlused, mis on olemasolevatest klastritest sõltumatud. Eelnevast tulenevalt oleks antud näite puhul kõige otstarbekam moodustada kaks kolmeliikmelist ning üks üheliikmeline klaster (etapp 4).

Näide (andmete standardiseerimine)

Kõigi distantsimõõtude korral tekitab raskusi see, et standardiseerimata andmete korral võib muutujate skaalade vahetamine muuta ka järjestust. Selgitame seda näite varal ning samas näitame, kuidas andmete standardiseerimisega on võimalik probleemi vältida.

Näiteks, oletagem, et teame kolme indiviidi A, B ja C kohta on meil olemas andmed nende brändi X ostmise tõenäosuse ning brändi X telereklaamide vaatamisele kulutatud aja kohta. Need andmed on esitatud tabelis 1.

Tabel 1. Algandmed

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

Toodud andmete põhjal saame arvutada distantsimõõdud, mis on esitatud tabelis 2.

Tabel 2. Ostmise tõenäosusel ning reklaamide vaatamisele (minutites mõõdetud) pühendatud ajal baseeruvad distantsid.

Objektide paar	Lihtne Euclediani distants		Ruutu tõstetud ehk absoluutne Euclediani 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

NB! Tulemuste tõlgendamisel tuleb silmas pidada, et distantside väiksemad väärtused näitavad suuremat lähedust ning sarnasust.

Tabel 3. Ostmise tõenäosusel ning reklaamide vaatamisele (sekundites mõõdetud) pühendatud ajal baseeruvad distantsid.

Objektide	Lihtne Euclediani	Ruutu tõstetud ehk	City-block distants

paar	distants		absoluutne Eucledіani 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

Tabel 4. Standardiseeritud andmetega distantsid.

Objektide paar	Standardiseeritud väärtused		Lihtne Eucledіani distants		Ruutu tõstetud ehk absoluutne Eucledіani 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

ARVESTUSTÖÖ I OSA

Question 1

Punktid: 2

Asümmeetriakordaja annab informatsiooni muutuja

a. jaotuse

b. keskväärtuse

c. tunnuse tüübi

Question 2

Punktid: 2

Dispersioonanalüüsi kasutatakse

a. kahe grupi dispersioonide võrdsuse testimiseks

b. kahe või enama grupi keskmiste võrdsuse testimiseks

c. kahe või enama grupi dispersioonide võrdsuse testimiseks

d. kahe grupi keskmiste võrdsuse testimiseks

Question 3

Punktid: 2

Tunnus, mille väärtused on järgmised: "alati", "enamasti", "aeg-ajalt", "harva", "üldse mitte", on

Vali üks või enam vastust.

- a. nominaaltunnus
- b. ordinaaltunnus
- c. pidev tunnus

Question 4

Punktid: 2

Kui tunnuse jaotus on paremale välja venitatud, st "saba" on paremal pool, on asümmeetriakordaja

Vali üks vastus.

- a. negatiivne
- b. positiivne

Question 5

Punktid: 2

F-testi abil kontrollitakse kahe grupi/valimi

Vali üks või enam vastust.

- a. dispersioonide võrdsust
- b. keskväärtuste võrdsust
- c. jaotuste sarnasust

Question 6

Punktid: 2

Tunnuse kõige sagedamini esinevat väärtust nimetatakse

Vali üks või enam vastust.

- a. mediaaniks
- b. dispersiooniks
- c. moodiks
- d. ekstsessiks
- e. keskmiseks

Question 7

Punktid: 2

Kui tegu on normaaljaotusega, siis on

Vali üks vastus.

- a. asümmetriakordaja nullilähedane ja ekstsess ligikaudu tunduvalt kõrgem
- b. nii asümmetriakordaja kui ekstsess nullilähedased
- c. asümmetriakordaja oluliselt nullist erinev ja ekstsess ligikaudu null

Question 8

Punktid: 2

Aritmeetiline keskmine ja mediaan annavad nominaaltunnuste puhul vähem sisukat infot kui arvuliste tunnuste puhul

Vastus:

- Õige

Vale

Question 9

Punktid: 2

Korrastatud rea keskmist liiget nimetatakse

Vali üks või enam vastust.

- a. moodiks
- b. dispersiooniks
- c. ekstsessiks
- d. korrelatsiooniks
- e. mediaaniks

Question 10

Punktid: 2

Kui ühe muutuja kõrgete väärtustega kaasnevad teise muutuja madalad väärtused, on korrelatsioonikoefitsient

Vali üks vastus.

- a. negatiivne
- b. positiivne
- c. nullilähedane

Question 11

Punktid: 2

T-testi abil kontrollitakse kahe grupi/valimi

Vali üks või enam vastust.

- a. dispersioonide võrdsust
- b. keskväärtuste võrdsust
- c. jaotuste sarnasust

Question 12

Punktid: 2

Korrelatsioonanalüüs võimaldab analüüsida muutujatevaheliste seoste kausaalsust

Vastus:

- Õige
- Vale

Question 13

Punktid: 2

Juhul, kui keskmine on tunduvalt kõrgem kui mediaan, on tegemist

Vali üks või enam vastust.

- a. vasakkaldelise reaga
- b. paremkaldelise reaga

Question 14

Punktid: 2

Liigendtabelid on Exceli analüüsivahend

Vali üks või enam vastust.

- a. tabelite hõlpsaks genereerimiseks
- b. tulemuste hõlpsaks visualiseerimiseks
- c. mõlemad eelpool nimetatud on õiged

Question 15

Punktid: 2

Kui samu inimesti küsitletakse nii 2008. kui 2009. aastal on tegu
Vali üks või enam vastust.

- a. ristlõikeandmestikuga
- b. longituudse uuringuga

ARVESTUSTÖÖ II OSA

1. Oletame, et soovitakse testida, kas meeste keskmine vanus andmebaasis on 45. Selleks tuleks kasutada
   a) paarisvalimi t-testi
   b) sõltumatute valimite t-testi
   c) ühe valimi t-testi
   d) dispersioonanalüüsi

2. Oletame, et tahame analüüsida, kas ravim mõjutab inimese vererõhu taset. Selleks mõõdame igal valimisse kaasatul vererõhu taseme enne ja pärast ravimi manustamist. Testimaks, kas ravimil on mõju vererõhu tasemele, tuleks kasutada
   a) paarisvalimi t-testi
   b) sõltumatute valimite t-testi
   c) dispersioonanalüüsi

3. Oletame, et tahame analüüsida, kas töötute, hõivatute ja mitteaktiivsete keskmine haridustase on sarnane. Selleks tuleks kasutada
   a) paarisvalimi t-testi
   b) sõltumatute valimite t-testi
   c) dispersioonanalüüsi

4. Oletame, et tahame analüüsida, kas meeste ja naiste keskmine abiellumise iga on sarnane. Selleks tuleks kasutada
   a) paarisvalimi t-testi
   b) sõltumatute valimite t-testi
   c) dispersioonanalüüsi

5. Hierarhilise klasteranalüüsi graafikut nimetatakse…
6. Kui jaotus on paremkaldeline, siis tähendab see, et
   a) Rohkem on muutuja madalaid väärtusi
   b) Rohkem on muutuja kõrgeid väärtusi
7. Levene'i testiga kontrollitakse
   a) kas vaatlusalustes kategooriates on muutuja keskmine sarnane
   b) kas vaatlusalustes kategooriates on muutuja varieeruvus sarnane

c) kas vaatlusalustes kategooriates on muutuja jaotus ligikaudu normaaljaotusega

8. karpdiagrammi keskmine joon on
   a) aritmeetiline keskmine
   b) mediaan
   c) mood
   d) asümmeetriakordaja
9. karpdiagrammi kasti ülemine ja alumine serv on
   a) 25 ja 75 kvartiil
   b) Minimaalne ja maksimaalne väärtus
10. Selgita hierarhilise ja mittehierarhilise klasteranalüüsi olemust ja erinevusi.
11. Oletame, et oleme hinnanud mudeli, kus sõltuvaks muutujaks on tarbimine ja sõltumatuks muutujaks palk. Hinnatud regressioonivõrrandi vabaliikme hinnang on 1000. Mida see sisuliselt tähendab?
12. Oletame, et oleme hinnanud mudeli, kus sõltuvaks muutujaks on tarbimine ja sõltumatuks muutujaks palk. Hinnatud regressioonivõrrandi sõltumatu muutuja parameetri hinnang on 0,8. Mida see sisuliselt tähendab?
13. Selgita kahe lausega (oma sõnadega, mitte copy-paste konspektist) heteroskedastiivsuse olemust. Millised märgid viitavad, et mudelis võib olla heteroskedastiivsus? Mis juhtub, kui mudelis on heteroskedastiivsus?
14. Selgita kahe lausega (oma sõnadega, mitte copy-paste konspektist) multikollineaarsuse olemust. Millised märgid viitavad, et mudelis võib olla multikollineaarsus? Mis juhtub, kui mudelis on multikollineaarsus?
15. Selgita kahe lausega (oma sõnadega, mitte copy-paste konspektist) autokorrelatsiooni olemust. Millised märgid viitavad, et mudelis võib olla autokorrelatsioon? Mis juhtub, kui mudelis on autokorrelatsioon?