Seosed tunnuste vahel

Seos kahe nominaalse/järjestus- või väheste väärtustega diskreetse tunnuse vahel.

Tabelid

Uurides seost nominaalsete või järjestustunnuste vahel koostatakse sageli esmalt kahemõõtmeline sagedustabel. Sagedustabel ise pole küll eriti sobiv vahend seose olemuse mõistmiseks, aga kuna ta on algmaterialiks kõigi järgnevate jooniste ja arvutuste jaoks, võib osutuda soovitavaks sagedustabeli lisamine analüüsitulemuste sekka – see võimaldab ka teistel soovi korral tehtud analüüse korrata. Samuti võimaldab sagedustabel sageli leida üles andmete sisestamisel tehtud vigu.

Sagedustabelit saab R-is teha käsuga *table*. Vaatame sagedustabelit tunnustele sport (mitu korda nädalas teed sporti: 1-„ei tee sporti“; 2-„1-2 korda“; 3- „3-4 korda “; 4-„5 või enam“) ja sugu (1-„naine“; 2-„mees“):

```
> table(sport, sugu)
     sugu
sport   1   2
    1 103  21
    2 313  69
    3  80  38
    4  13  19
    5   2   2
```

Vastusevariant 5 puhul on tegemist sisestusveaga – arvatavasti on vastusevariandi „5 vï enam korda nädalas spodin“ sisestamisel eksitud ja sisestatud 4 asemel 5. Teeme vajalikud parandused:

```
sport2=sport
sport2[sport==5]=4
```

ja anname tunnuste väärtustele ka õiged nimed

```
suguF=factor(sugu, labels=c("naine", "mees"))
sportF=factor(sport2,
      labels=c("ei spordi", "1-2", "3-4", "5 ja enam"))
```

Teeme uue, korrektsema sagedustabeli:

```
> table(sportF,suguF)
            suguF
sportF       naine mees
  ei spordi   103   21
  1-2         313   69
  3-4          80   38
  5 ja enam    15   21
```

Muuseas, soovi korral saab sagedustabelile lisada veergude/ridade summad käsuga addmargins:

```
> addmargins(table(sportF,suguF))
           suguF
sportF      naine mees Sum
  ei spordi   103   21 124
  1-2         313   69 382
  3-4          80   38 118
  5 ja enam    15   21  36
  Sum         511  149 660
```

Parema tunnetuse seosest saame, kui vaatame nn. tinglikke jaotuseid – kuidas muutub meeste/naiste osakaal erinevalt sportivate inimeste seas või, alternatiivselt, võime uurida, kui palju naistest ei tee sporti/teeb kord nädalas sporti/... ja leiame võrdluseks kõrvale samasuguse jaotuse ka meeste kohta.

Tunnuse sugu protsentuaalne jaotus erineva spordihuviga tudengite seas:

```
> prop.table(table(sportF,suguF),1)*100
           suguF
sportF         naine     mees
  ei spordi 83.06452 16.93548
  1-2       81.93717 18.06283
  3-4       67.79661 32.20339
  5 ja enam 41.66667 58.33333
```

Näeme, et kui mittesportijatest on 83% naised siis 5 ja enam kordi nädalas sporti tegevate inimeste seas on naisi vaid 42%.

Soovi korral võime vaadelda ka tunnuse sport jaotust sugude kaupa:

```
> prop.table(table(sportF,suguF),2)*100
           suguF
sportF          naine      mees
  ei spordi 20.156556 14.093960
  1-2       61.252446 46.308725
  3-4       15.655577 25.503356
  5 ja enam  2.935421 14.093960
```

Näeme, et kõigest 3% naistudengitest teevad sporti 5 ja enam korda nädalas, samas meestudengitel on vastav protsent 14%.

Järeldus sagedustabelite uurimisest võiks olla järgmine: meestudengid teevad rohkem sporti kui naistudengid.

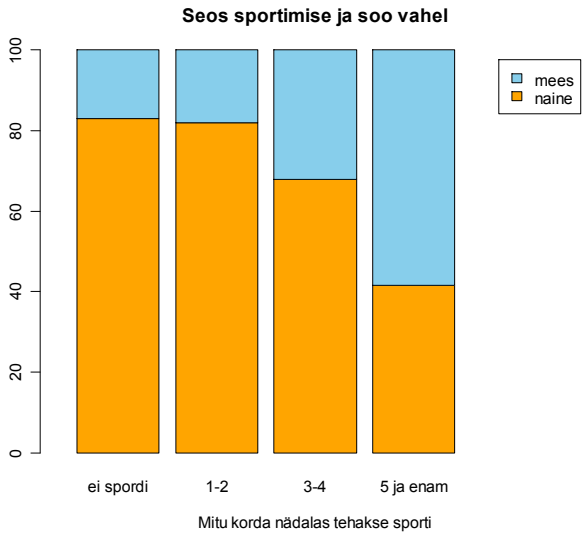
Sagedustabeli graafiline esitus

Sageli saab hästi valitud graafiku abil õige järelduseni jõuda kiiremini, sestap sobivad ka graafikud hästi oma ideede/avastuste edasiandmiseks laiale publikule. Vaatame lähemalt, kuidas kahemõõtmelisi sagedustabeleid saab visualiseerida. Parim võimalus selleks on kasutada tulpdiagrammi (barplot-käsku).

Nagu tinglike jaotuste vaatamisel, nii on ka tulpdiagrammi joonistamisel kaks võimalust – võime vaadata naiste/meeste jaotust eraldi iga sportimiskoguse korral või võime vaadelda tunnuse „sport“ jaotust meeste ja naiste korral.

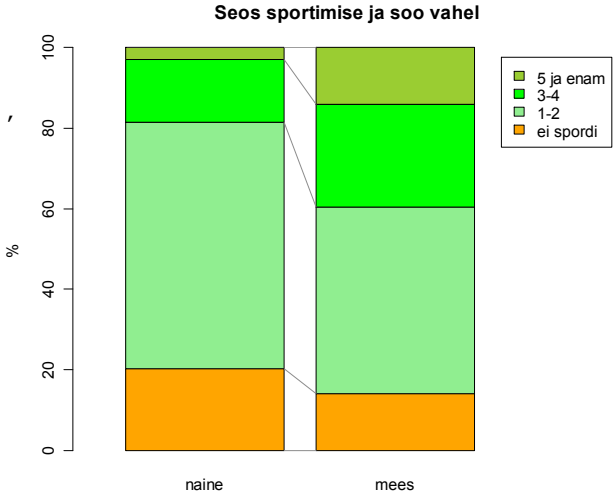
Varint 1 – tunnuse sugu (tinglik)jaotus iga tunnuse sport väärtuse korral

```
barplot(prop.table(table(suguF, sportF),2)*100,
  col=c("orange","skyblue"), ylab="%",
  main="Seos sportimise ja soo vahel",
  xlab="Mitu korda nädalas tehakse sporti",
  legend=T, xlim=c(0,6.2))
```



Varint 2 – tunnuse sport (tinglik)jaotus iga tunnuse sugu võimaliku väärtuse korral

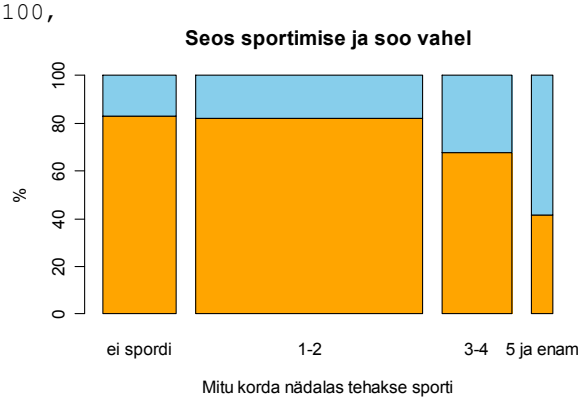
```
a=barplot(prop.table(table(sportF, suguF),2)*100,
 col=c("orange","lightgreen","green","yellowgreen"),
 ylab="%", main="Seos sportimise ja soo vahel",
 legend=T, xlim=c(0,3.2))

# Lisame hallid „abijooned“:

abi=prop.table(table(sportF, suguF),2)*100
abi1=cumsum(abi[,1])
abi2=cumsum(abi[,2])
arrows(rep(a[1]+0.5, length(abi1)+1),
    c(0,abi1), rep(a[2]-0.5, length(abi2)+1),
    c(0, abi2), code=0, col="gray55")
```



Vahel soovib inimene vaheldust – üks graafiku kuju, ükskõik kui hea, võib ajapikku tüütuks osutuda. Vaatame alljärgnevalt teisi võimalusi seose kirjeldamiseks (NB! tulpdiagramm jääb siiski esimeseks soovituseks!)

Variant 3 – tulpdiagramm, mille tulpade laiused sõltuvad vaatluste arvust

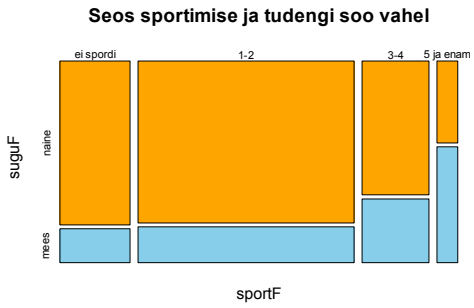
```
barplot(prop.table(table(suguF, sportF),2)*100,
  col=c("orange","skyblue"), ylab="%",
  main="Seos sportimise ja soo vahel",
  xlab="Mitu korda nädalas tehakse sporti",
  width=table(sportF))
```



Antud graafikul on tulp „1-2 korda nädalas“ kõige laiem, sest enamik tudengitest teeb 1-2 korda nädalas sporti. 5 ja rohkem kordi teevad sporti vaid mõned üksikud tudengid, sestap on ka vastav tulp üsna kleenuke.

Variant 4 – variatsioon eelmisest

Üsna sarnase pildi eelmisele saame, kui lihtsalt laseme R-l joonistada sagedustabeli põhjal graafiku. Tulba laius näitab, kui palju oli sellise intensiivsusega sportivaid tudengeid, piirjoon sugude vahel iseloomustab aga seda, kui palju siis antud intensiivsusega sportivatest tudengitest olid naised ja kui paljud olid mehed.

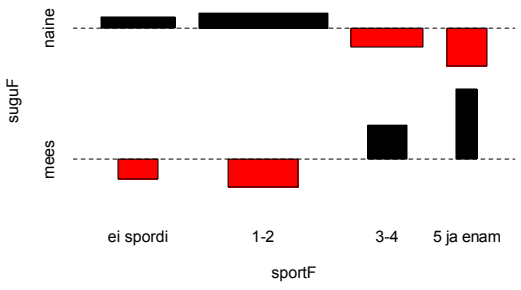


Ainsaks puuduseks antud graafiku puhul on see, et puudub y-telg ühikutega (saame küll üldpildi seosest endast, kuid midagi täpsemat on antud jooniselt raske välja lugeda).

```
plot(table(sportF, suguF), col=c("orange", "skyblue"),
     main="Seos sportimise ja tudengi soo vahel")
```

Variant 5 – võrdlus „keskmisega“

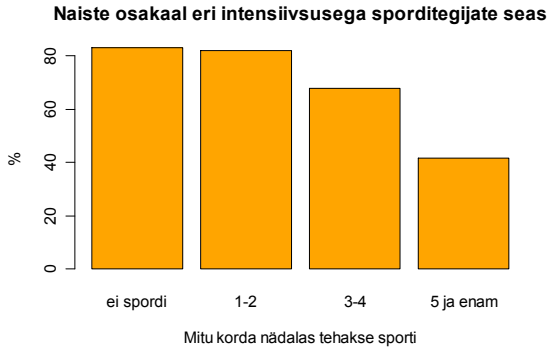
Vahel näeb inimsilm seost paremini, kui saame graafikul kujutada, kas ühes või teises grupis oli naisi liiga palju või liiga vähe võrreldes „keskmisega“ ehk võrreldes naiste proportsiooniga kõigi uuritavate seas. Antud joonisel on punktiirjoonest ülespoole tulbaga tegemist siis, kui antud soost tudengeid oli „liiga palju“ antud rühmas ja allapoole tulbaga on tegemist siis, kui antud soost tudengeid oli vaadeldud rühmas „liiga vähe“. Tulpade laius iseloomustab tudengite arvu – laiem tulp näitab, et antud „tüüpi“ tudengeid oli rohkem kui kitsa tulba korral.




```
assocplot(table(sportF, suguF))
```

Variant 6 – tulpdiagramm ainult naiste osakaalule!

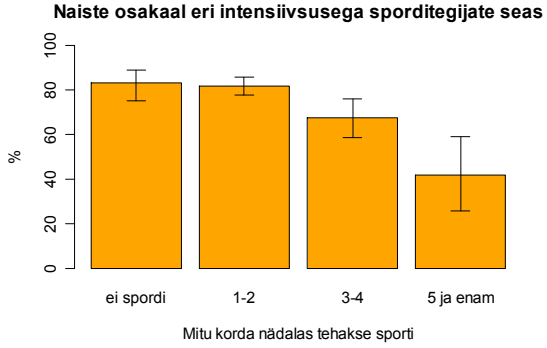
Kui ühel tunnusel (sugu) on vaid kaks võimalikku väärtust, siis võime vaadelda ka lihtsalt naiste (või meeste) osakaalu erineva intensiivsusega sporditegijate seas:




```
barplot(prop.table(table(suguF,
sportF),2)[1,]*100,
  col=c("orange"), ylab="%",
  main="Naiste osakaal eri intensiivsusega sporditegijate seas",
  xlab="Mitu korda nädalas tehakse sporti")
```

Variant 7 – tulpdiagramm koos usalduspiiridega

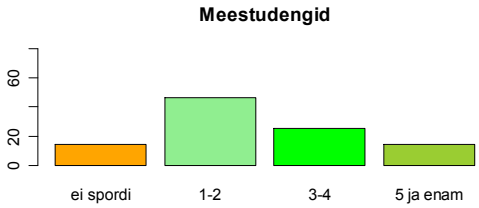
Eeltoodud joonisele võime lisada ka usalduspiirid – kui palju siis naiste osakaal ühes või teises klassis tegelikult olla võiks. NB! Usalduspiiride lisamisel peame *y*-teljele jätma täiendavat ruumi (*ylim*-lisaparameetrit kasutades)!

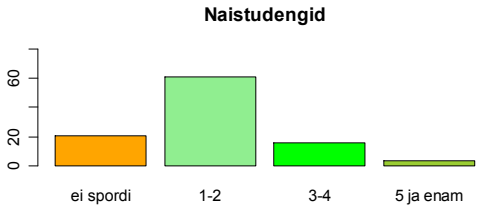



```
# Arvutame usalduspiirid naistudengi
# saamise tõenäosusele (võib teha ka
# käsitsi binom.test - funktsiooni abil!)
naised=table(suguF, sportF)[1,]
mehed=table(suguF, sportF)[2,]
n=naised+mehed
al=qbeta(0.025, naised, n - naised+1)
yl=qbeta(0.975, naised+1, n - naised)

# Joonistame tulpdiagrammi -
#   pane tähele omistamist a=barplot(...
#   (me salvestame tulpade asukohad
#   hilisemaks kasutamiseks)

a=barplot(prop.table(table(suguF, sportF),2)[1,]*100,
  col=c("orange"), ylab="%",
  main="Naiste osakaal eri intensiivsusega sporditegijate seas",
  xlab="Mitu korda nädalas tehakse sporti", ylim=c(0,100))

# Lisame joonisele väljaarvutatud usalduspiirid
arrows(a, al*100, a, yl*100, code=3, length=0.1, angle=90)
```

Variant 8 – Suvaline graafik, mida saab kasutada jaotuse iseloomustamiseks

Seose iseloomustamiseks saab kasutada tegelikult suvalist graafikut, mille abil võisime kirjeldada uuritava tunnuse jaotust. Lihtsalt tuleb teha mitu sellist graafikut – näiteks meeste ja naiste jaoks eraldi.






```
 Kaks graafikut üksteise alla
par(mfrow=c(2,1))

barplot(
   prop.table(table(
     sportF[suguF=="mees"]))*100,
     main="Meestudengid",
   col=c("orange","lightgreen",
         "green","yellowgreen"),
   ylim=c(0,80))

barplot(prop.table(table(
    sportF[suguF=="naine"]))*100,
    main="Naistudengid", ylim=c(0,80),
    col=c("orange","lightgreen","green","yellowgreen"))
```

Kahe graafiku joonistamisel kontrollige, et telgede pikkused mõlemal graafikul oleksid samad – antud juhul peaks y-telje pikkus mõlema graafiku jaoks olema sama (saavutatud ylim-lisaparameetri abil). Samuti peaksid mõlemal graafikul olema esindatud kõik kategooriad – kui 5 ja enam kordi nädalas sportivaid naistudengeid poleks olnud, siis oleks pidanud vaatama, et graafikule vastava tulba jaoks siiski koht jäetakse – vaata ka vastavat kirjeldust osast „jaotuse iseloomustamine“.

Seose olemasolu testimine

Kas seos, mida arvame nägevat tabelis või graafikul eksisteerib ikka tegelikult või võib olla tingitud valimi juhuslikusest? Sellele küsimusele vastamiseks peaksime testima (statistilise hüpoteeside kontrollimise abil) kas seos ikka eksisteerib. Üks universalsemaid võimalusi seose olemasolu kontrollimiseks on hii-ruut test.

Hii-ruut test (nominaalne tunnus vs nominaalne/järjestus/diskreetne tunnus)
Hii-ruut test on (pea)ainus võimalus seose olemasolu kontrollimiseks, kui üks uuritavatest tunnustest on nominaalne. Kui uurime seose olemasolu järjestustunnuste või diskreetse tunnuse ja järjestustunnuse vahel, siis tekib juurde teisigi võimalusi, kuid hii-ruut testi võime kasutada ka sellisel juhul.

```
> chisq.test(table(suguF, sportF))

        Pearson's Chi-squared test

data:  table(suguF, sportF)
X-squared = 39.2996, df = 3, p-value = 1.500e-08
```

Hii-ruut statistiku väärtus

vabadusastmete arv – kui ühel tunnusel on *k* erinevat väärtust ja teisel *l*, siis *df*=(*k*-1)(*l*-1). Seda tasuks kontrollida – sest kui arvuti teeb midagi valesti (saab meist valesti aru) nähtub see sageli ka vales vabadusastmete arvus.

olulisustõenäosus – kui on väiksem kui 0,05, siis oleme seose olemasolu tõestanud. Praegusel juhul 0,000000015<<0,05 ja järelikult nähtud seos tudengi soo ja sporditegemise vahel ei saa olla tingitud vaid valimi juhuslikkusest – see on tegelikult ka olemas.

Hii-ruut testi puhul peab meeles pidada, et testi tulemus (olulisustõenäosus) on õieti arvutatud vaid siis, kui sõltumatuse (seose puudumise) korral oleks kõigi võimalike kombinatsioonide oodatav esinemiste arv 5 või enam. Kui palju ühte või teist kombinatsiooni peaks esinema siis, kui tunnused on sõltumatud, võime vaadata järgmise käsuga:

```
> chisq.test(table(suguF, sportF))$expected
       sportF
suguF   ei spordi      1-2      3-4 5 ja enam
  naine  96.00606 295.7606 91.36061 27.872727
  mees   27.99394  86.2394 26.63939  8.127273
```

Näeme, et antud juhul kõik (sõltumatuse korral) oodatavad sagedused on suuremad kui 5 – seega probleemi pole.

Mida teha, kui mõni oodatavatest sagedustest oleks väiksem kui 5? Näiteks nagu siis, kui vaatame, kas on seost õlletabimise ja soo vahel:

```
> chisq.test(table(suguF, olu))$expected
       olu
suguF           1         2        3         4        5
  naine 206.03933 206.03933 71.26172 23.237519 5.422088
  mees   59.96067  59.96067 20.73828  6.762481 1.577912
Warning message:
Chi-squared approximation may be incorrect in:
chisq.test(table(suguF, olu))
```

Antud juhul on kaks võimalust. Esiteks võime sarnaseid kategooriaid kokku võtta. Antud juhul tähistab tunnuse olu väärtus 4 tudengit, kes joob 5-12 pudelit õlut nädalas ja väärtus 5 tähistab tudengit, kes joob 13 või enam pudelit nädalas. Antud juhul võime tunnuse ümber kodeerida ja tähistada väärtusega 4 tudengeid, kes joovad 5 või enam pudelit nädalas, st muuta väärtused 5 samuti väärtusteks 4:

olu2=olu; olu2[olu==5]=4
Peale ümberkodeerimist on kõigi väärtuste kombinatsioonide korral sõltumatuse korral oodatavad sagedused 5 või enam:

```
> chisq.test(table(suguF, olu2))$expected
       olu2
suguF           1         2        3         4
  naine 206.03933 206.03933 71.26172 28.659607
  mees   59.96067  59.96067 20.73828  8.340393
```

Ja hii-ruut testi tulemus (seos eksisteerib) seega usaldusväärne:

```
> chisq.test(table(suguF, olu2))

        Pearson's Chi-squared test

data:  table(suguF, olu2)
X-squared = 161.0025, df = 3, p-value < 2.2e-16
```

Alternatiivina ümberkodeerimisele võime lasta R-l hinnata tegelikku olulisustõenäosust algse tabeli puhul. Kuna täpne olulisustõenäosuse arvutus on väga aeglane, kasutab R õige olulisustõenäosuse hindamiseks ligikaudset meetodit, mille vastus võib igal katsel veidi erinev tulla:

```
> chisq.test(table(suguF, olu), B=100000, simulate.p.value=p)

Pearson's Chi-squared test with simulated p-value (based on 1e+05 replicates)

data:  table(suguF, olu)
X-squared = 161.021, df = NA, p-value = 1e-05
```

Tegemist on nn täpse hii-ruut testiga (Exact chi-square test) – otsuseks on, et seos eksisteerib (olulisustõenäosus on 0,00001). Ka täpse hii-ruut testi puhul tasub tähele panna, et kui oodatavad sagedused kipuvad väga väikesed tulema (paljudel juhtudel 1 või alla selle), peaks midagi kuskil siiski muutma või otsima sobivamat testi.

Fisheri täpne test (binaarne vs binaarne)
Kui uuritakse seose olemasolu kahe binaarse tunnuse vahel, näiteks tunnuste töötlus (töödeldi vs kontroll) ja tulemus (haigestus vs ei haigestunud) vahel, sobib seose olemasolu testimiseks kõige paremini Fisheri täpne test. Antud näites vaatame, kas eksisteerib seos tudengi soo (tunnus *sugu*) ja selle vahel, kas tudeng on vajanud viimasel ajal kiirabi abi (tunnus *kiirabi*).

```
> fisher.test(table(sugu, kiirabi))

        Fisher's Exact Test for Count Data

data:  table(sugu, kiirabi)
p-value = 0.6484
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.607045 2.120387
sample estimates:
odds ratio
   1.15598
```

Olulisustõenäosus on suurem kui 0,05, seega pole seose olemasolu tudengi soo ja kiirabi vajamise vahel võimalik antud andmete põhjal tõestada.

Kendali tau ja tema olulisustõenäosus (järjestustunnus vs järjestustunnus)
Kahe järjestustunnuse vahel eksisteerivat seost, juhul kui see seos on monotoonne (ühe tunnuse väärtuste kasvades kipuvad kasvama ka teise tunnuse väärtused) saab kontrollida iseloomustada kasutades nn Kendali tau –seosekordajat ja tema olulisust.

```
> cor.test(sport, olu, use="pairwise.complete.obs", method="kendall")

        Kendall's rank correlation tau

data:  sport and olu
z = 2.5445, p-value = 0.01094
alternative hypothesis: true tau is not equal to 0
sample estimates:
       tau
0.08779852
```

Olulisustõenäosus on väiksem kui 0,05 – seos sportimise ja õlletarbimise vahel eksisteerib.

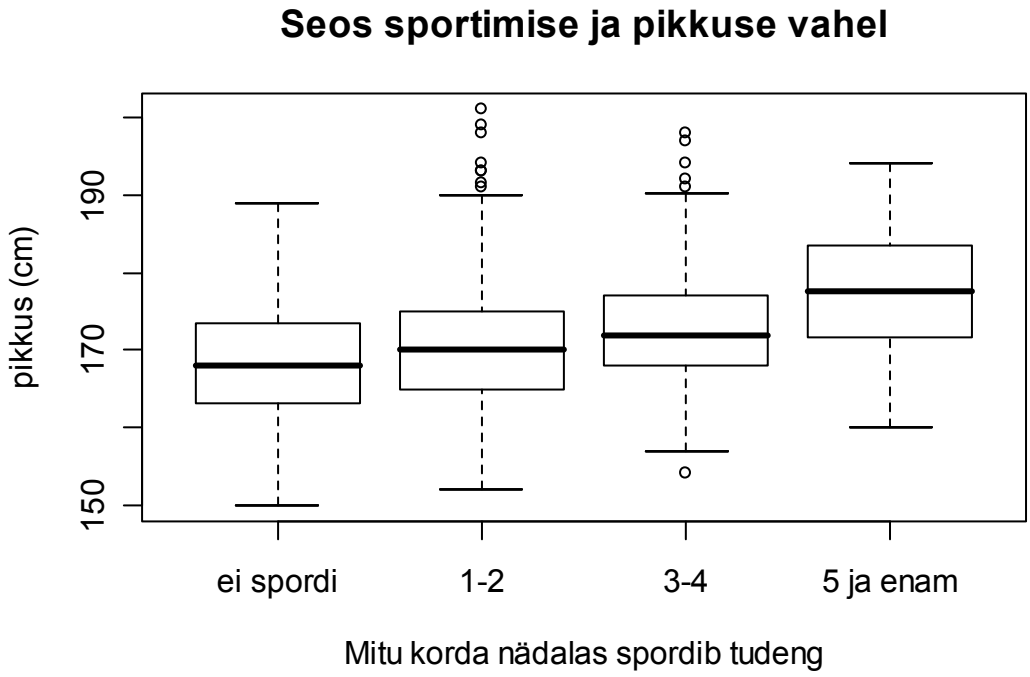
Kendali tau on positiivne – seega mida enam tudeng spordib, seda enam ta ka joob (või vastupidi – mida enam joob, seda enam spordib)

Märkus 1: Kendali tau-sid on mitmeid erinevaid, sestap võib numbrite võrdlemisel kergesti eksida.

Pidev tunnus vs nominaalne/järjestustunnus

Kui on tarvis iseloomustada seost nominaalse (või järjestustunnuse) ja pideva tunnuse vahel, on sageli kõige sobivamaks graafikuks karp-vurrud diagramm (*boxplot*).

```
boxplot(pikkus~sportF, main="Seos sportimise ja pikkuse vahel",
      xlab="Mitu korda nädalas spordib tudeng", ylab="pikkus (cm)")
```



Graafikult näeme, et rohkem sporti tegevad tudengid kipuvad olema pikemad (sest enam teevad sporti meestudengid ja mehed kipuvad olema pikemad).

Seose kirjeldamisel võime anda ka iga grupi jaoks põhistatistikute (näiteks keskmise ja standardhälbe või standardvea) väärtused.

Keskmised:

```
> by(pikkus, sportF, mean, na.rm=T)
INDICES: ei spordi
[1] 168.9363
------------------------------------------------------------------------------------
INDICES: 1-2
[1] 170.5530
------------------------------------------------------------------------------------
INDICES: 3-4
[1] 173.2644
------------------------------------------------------------------------------------
INDICES: 5 ja enam
[1] 177.4722
```

ja standardhälbed:

```
> by(pikkus, sportF, sd, na.rm=T)
INDICES: ei spordi
[1] 7.634052
------------------------------------------------------------------------------------------
INDICES: 1-2
[1] 8.096612
------------------------------------------------------------------------------------------
INDICES: 3-4
[1] 8.940492
------------------------------------------------------------------------------------------
INDICES: 5 ja enam
[1] 8.21666
```

Tulemus:

Sportimiskordade arv nädalas	Pikkus (cm) keskmine	standardhälve
0	168,9	7,6
1-2	170,6	8,1
3-4	173,3	8,9
5-	177,5	8,2

Vahel võime ka graafikul esitada gruppide jaoks mõne põhistatistiku väärtust – näiteks iseloomustada, kuidas muutub grupi keskmine pikkus sportimisintensiivsuse kasvades. Sellisele graafikule sobiks muidugi lisada ka usalduspiirid. Valmiskujul soovitud graafikut R-i põhipaketis pole, küll aga saame kerge vaevaga vastavat tüüpi joonist tegeva funktsiooni ise lisada. Käivitame järgmise programmi:

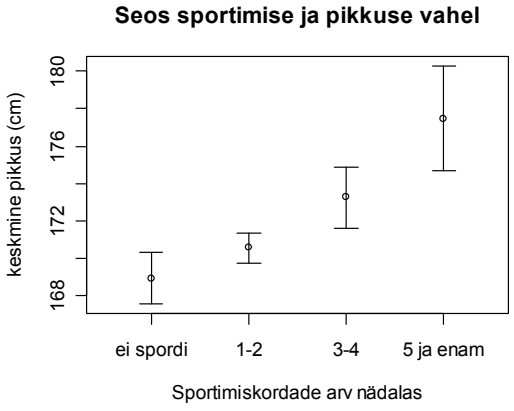
```
keskmUI=function(pidev, grupp,...){
  a=by(pidev, grupp, mean, na.rm=T)
  Ylem=function(pidev){return(t.test(pidev)$conf.int[1])}
  Alum=function(pidev){return(t.test(pidev)$conf.int[2])}
  aY=by(pidev, grupp, Ylem)
  aA=by(pidev, grupp, Alum)
  plot(1:length(a),a,  xaxt="n", ylim=range(c(a, aY, aA)),
       xlim=c(0.5, length(a)+0.5),...)
  axis(1, at=1:length(a), labels=names(table(grupp)))
  arrows(1:length(a), aY, 1:length(a), aA, code=3, length=0.1,
angle=90)
}
```

ja meile tekib lisaks käsk „keskmUI“, mida saame juba kasutada jooniste tegemisel:

```
keskmUI(pikkus, sportF,
       xlab="Sportimiskordade arv nädalas",
       ylab="keskmine pikkus (cm)",
       main="Seos sportimise ja pikkuse vahel")
```

või, kui huvituksime vererõhu ja õllejoomise vahelisest seosest:

```
keskmUI(SVR, olu)
```



Testimine

Pidev tunnus (normaaljaotusega jäägid) vs nominaalne või järjestustunnus

Levinuim viis on kasutada dispersioonanalüüsi (ANOVA-t) – kontrollida, kas leidub selliseid gruppe, mille keskväärtused pole võrdsed. Juhul, kui keskväärtused pole võrdsed, saame öelda, et seos nominaalse tunnuse (grupeeriva tunnuse) ja pideva tunnuse vahel eksisteerib. Samas võib seos nominaalse tunnuse ja pideva tunnuse vahel olla selline, mida dispersioonanalüüs avastada ei suuda (näiteks võib grupiti muutuda uuritava tunnuse hajuvus või näiteks võib uuritava tunnuse jaotus muutuda aina ebasümmeetrilisemaks). Dispersioonanalüüsi tegemine R-ga:

Dispersioonanalüüsi mudeli hindamine

```
m1=glm(pikkus~factor(olu))
```

Tulemuste vaatamine

```
summary(m1)
```

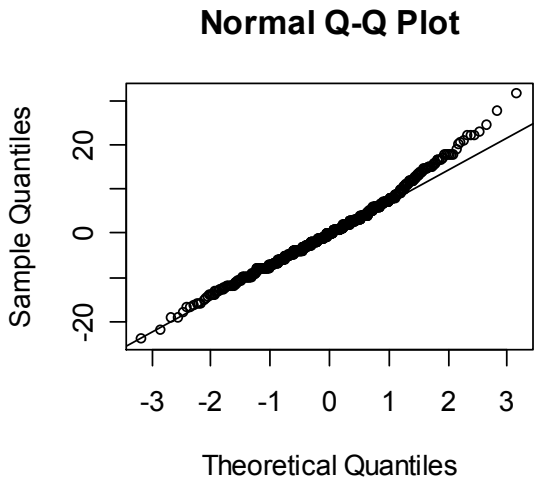
Ning seose olemasolu testimine:

```
> drop1(m1, test="F")
Single term deletions

Model:
pikkus ~ factor(olu)
            Df Deviance   AIC F value     Pr(F)
<none>           41098  4612
factor(olu)  4    46512  4685  21.568 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Näeme, et seos pikkuse ja õlletarbimise vahel on statistiliselt oluline (olulisustõenäosus on väiksem kui 0,05). Samas tuleb meeles pidada, et vastav test on korrektne eelkõige siis, kui nn mudeli jäägid on ligikaudu normaaljaotusega:

```
qqnorm(resid(m1))
qqline(resid(m1))
```

Jääkide normaalsuse eeldus pole täielikult rahuldatud – seega tuleks testitulemusse suhtuda mõningase ettevaatusega (kuigi sedavõrd väikese olulisustõenäosuse korral vaevalt et lõppotsus muutub ka korrektsema analüüsi korral).



Juhul, kui huvitutakse pigem hajuvuse võrdlemisest, sobib paremini Bartlett' test:

```
> bartlett.test(pikkus~factor(olu))

        Bartlett test of homogeneity of variances

data:  pikkus by factor(olu)
Bartlett's K-squared = 31.7351, df = 4, p-value = 2.167e-06
```

Antud juhul näeme, et Bartlett' testi arvates on uuritava tunnuse hajuvus grupiti erinev (õlut „parajalt" joovate tudengite pikkused on varieeruvamad, kui õlut mittetarbivate või väga palju õlut tarbivate tudengite pikkused – sest „parajalt" joojate seas on nii naisi kui mehi, seevastu mittetarbijad on põhiliselt naistudengid ja ohtralt joojad peamiselt mehed). Samas on (jääkide) normaaljaotuse eeldus praegu veidi kahtluse all (vaata graafikut eelmisel leheküljel) – Bartlett testi eeldused pole täidetud ja seega jääb veidi ebaselgeks, kas olulisustõenäosus on ikka õieti arvutatud.

Pidev tunnus (suvaline) vs nominaalne või järjestustunnus

Juhul, kui dispersioonanalüüsi (ANOVA-t) ei saa kasutada, sest mudeli jäägid pole normaaljaotusega, siis võib proovida kasutada nn. mitteparameetrilisi alternatiive. Kruskal-Wallise test kontrollib seose olemasolu pideva ja nominaalse/järjestustunnuse vahel, kusjuures test suudab eelkõige avastada selliseid juhte, kus mõnes grupis uuritava tunnuse väärtused kipuvad olema suuremad kui teises. Kontrollitava hüpoteesi täpne sõnastus on aga üsna keeruline (H1 tõestamine ei tähenda seda, et keskväärtused oleksid grupiti erinevad, ega tähenda ka seda, et mediaanid erineksid grupiti teineteisest, küll aga tähendab seose olemasolu). Kruskal-Wallise test R-is (kontrollime, kas eksisteerib seos pikkuse ja õlle tarbimise vahel):

```
> kruskal.test(pikkus~factor(olu))

        Kruskal-Wallis rank sum test

data:  pikkus by factor(olu)
Kruskal-Wallis chi-squared = 64.8768, df = 4, p-value = 2.732e-13
```

Otsus: Seos eksisteerib (kuna olulisustõenäosus on palju väiksem 0,05-st). Muuseas – kuigi Kruskal-Wallise test ei eelda normaaljaotust, on vajalik, et uuritav tunnus (pikkus) oleks pidev tunnus – korduvaid väärtuseid ei tohiks olla märkimisväärselt.

Kui soovime testida, kas uuritava tunnuse (pikkus) hajuvus grupiti (erineva õlletarbimise korral) on erinev, saab teha näiteks kasutades Fligner-Killeen'i testi.

```
> fligner.test(pikkus~factor(olu))

        Fligner-Killeen test of homogeneity of variances

data:  pikkus by factor(olu)
Fligner-Killeen:med chi-squared = 39.1426, df = 4, p-value = 6.51e-08
```

Näeme, et uuritava tunnuse varieeruvus tõepoolest grupiti erineb. P.S. – ettevaatust saadud tulemuse interpreteerimisel – Fligner-Killeeni testi mõistes hajuvuste erinevus ei pruugi tähendada dispersioonide erinevust!

Pidev tunnus vs binaarne tunnus

Juhul, kui meil on tegemist kahe grupiga, siis sobivad seose olemasolu testimiseks

a) Kolmogorov-Smirnovi test (kontrollib kõige seost kõige üldisemas mõttes – kas uuritavate tunnuste jaotused antud kahes grupis on erinevad või mitte);
b) t-test (kontrollib, kas uuritava tunnuse keskväärtused kahes grupis on samad või mitte – eeldab kas suurt valimit või normaaljaotusega uuritavat tunnust)
c) Wilcoxoni test (kontrollib, kas uuritava tunnuse väärtused ühes grupis „kipuvad olema suuremad" kui teises grupis – ei nõua normaaljaotusega uuritavat tunnust)

Graafikute koha pealt muutuseid pole – sobivad eelmises peatükis kirjeldatud graafikud.

Kolmogorov-Smirnovi test

Kolmogorov-Smirnovi test on kõige üldisem test (nn. omnibuss-test – kontrollib kõikvõimalikke erinevusi jaotuste vahel). Probleemid: uuritav tunnus peab olema pidev, st uuritaval tunnusel ei tohiks esineda korduvaid väärtuseid (või vähemalt ei tohiks korduvaid väärtuseid esineda märkimisväärselt). Näide Kolmogorov-Smirnovi testi kasutamisest (kas tudengite pikkuse ja soo vahel eksisteerib seos):

```
> ks.test(kaal[sugu==1], kaal[sugu==2])

        Two-sample Kolmogorov-Smirnov test

data:  kaal[sugu == 1] and kaal[sugu == 2]
D = 0.644, p-value < 2.2e-16
alternative hypothesis: two-sided

Warning message:
cannot compute correct p-values with ties in: ks.test(kaal[sugu == 1], kaal[sugu ==
2])
```

Olulisustõenäosus on selgelt väiksem kui 0,05 – seega peaks eksisteerima seos tudengite kaalu ja soo vahel. Samas antakse hoiatus – andmestikus eksisteerivad korduvad väärtused – sama kaalunumbrit esineb korduvalt – ehk teisisõnu – meil pole tudengite kaal mõõdetud piisava täpsusega. Sestap pole ka Kolmogorov-Smirnovi testi tulemus mitte päris usaldusväärne.

T-test
T-test kontrollib, kas uuritava tunnuse keskväärtus mõlemas grupis on samasuur või mitte. Vaikimisi kasutab R nn Welchi t-testi (mis lubab, et uuritava tunnuse hajuvus mõlemas grupis võib olla erinev).

```
> t.test(kaal~sugu)

        Welch Two Sample t-test

data:  kaal by sugu
t = -16.4634, df = 197.086, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -19.24219 -15.12544
sample estimates:
mean in group 1 mean in group 2
       59.19941        76.38322
```

Näeme, et olulisustõenäosus on väiksem kui 0,05 – järelikult on meestudengite kaalu keskväärtus ja naistudengite kaalu keskväärtus tõestatavalt erinevad (ja seega saame väita, et seos kaalu ja soo vahel eksisteerib).

Wilcoxoni (Mann-Whitney) test
Wilcoxoni test on alternatiiv t-testile. Wilcoxoni testi võib kasutada ka siis, kui uuritava tunnuse jaotus pole normaaljaotus (ja kui valim on väike). Paraku on tal ka puuduseid – juhul, kui Wilcoxoni test otsustab alternatiivse hüpoteesi kasuks – seos eksisteerib – siis võib osutuda üsna raskeks seletada, milles see seos siis ikkagi seisneb. Kuigi tunnetuslikult tähendab alternatiivse hüpoteesi kasuks otsustamine seda, et ühes grupis on uuritava tunnuse väärtused „suuremad“ kui teises, ei pruugi see veel tähendada, et uuritava tunnuse keskväärtused või mediaanid oleksid ühes grupis suuremad kui teises.

```
> wilcox.test(kaal~sugu)

        Wilcoxon rank sum test with continuity correction

data:  kaal by sugu
W = 8088, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Antud näites saame lugeda tõestatuks, et seos tudengi kaalu ja soo vahel eksisteerib – olulisustõenäosus (p-value) on märkimisväärselt alla 0,05. Seega võime oletada, et ühes grupis on uuritava tunnuse väärtused suuremad kui teises grupis – antud juhul on meestudengite kaalud suuremad kui naistudengite kaalud.

P.S. Wilcoxoni test eeldab samuti pidevat uuritavat tunnust ja väga paljude korduvate väärtuste korral (palju kokkulangevaid kaale näiteks) võib testi tulemus osutuda ebatäpselt arvutatuks. Wilcoxoni testi (ja ka paljude teiste mitteparameetriliste testide korral) „ohutuks“ kasutamiseks tuleks pidevat tunnust mõõta piisavalt täpselt, et saaks üheselt paika panna järjestuse – kelle kaal siis ikkagi on suurem ja kelle kaal on väiksem.