# Data-analyysi R-ohjelmistolla, kevät 2015

# Viikko 4

Tällä viikolla tutkitaan aineistoa luokittelun, summamuuttujien ja luottamusvälien avulla. Koska esimerkit vaativat jo jonkinlaisen yksinkertaisen aineiston käyttöä, otetaan tarkasteluun taulukko `A`. Voit luoda kyseisen taulukon omia kokeilujasi varten seuraavalla koodilla:

```
A <- data.frame(matrix(1:100, nrow=20))
colnames(A) <- c("A1", "A2", "A3",
     "A4", "A5")
```

Taulukossa on siis viisi saraketta, nimeltään `A1 - A5`, ja 20 riviä.

Neljännen viikon tehtävät vaativat paikoin pohdintaa ja edellisten viikkojen osaamisen yhdistämistä.

# 1 Osa-aineistojen valinta, osa 3

## 1.1 Sarakkeiden valinta nimellä

Viikolla 3 tutustuttiin `subset()`-funktion käyttöön osa-aineistojen valinnassa. Seuraavaksi katsotaan miten `subset()` toimii osa-aineistojen valinnassa laajemmin.

**Esimerkki 1.** Valitaan taulukosta `A` sarakkeet `A2` ja `A3` ja tallennetaan valittu osa-aineisto muuttujaan `A_23`

```
A_23 <- subset(A, select=c("A2", "A3"))
```

**Esimerkki 2.** Valitaan seuraavaksi taulukosta `A` sarakkeista `A2` ja `A3` ne rivit, joissa `A4` $> 70$. Tallennetaan valittu osa-aineisto muuttujaan A__23__yli70

```
A_23_yli70 <- subset(A, A4 > 70, select=c("A2", "A3"))
```

Monien sarakkeiden valinta kirjoittaen jokaisen valittavan sarakkeen nimi käsin on saattaa olla huomattavan työlästä. Usein aineistossa sarakkeet nimetään ja numeroidaan jollakin loogisella tavalla, joka noudattaa helposti toistettavaa kaavaa. Käydään seuraavaksi läpi yksi nopeahko tapa valita tällaisessa tapauksessa useita nimiä samalla kerralla.

**Esimerkki 3.** Katsotaan ensin miten toimii `paste()`-funktio:

```
> paste("Sarake",1:5)
[1] "Sarake 1" "Sarake 2" "Sarake 3" "Sarake 4" "Sarake 5"
> paste("Sarake",1:5, sep="")
[1] "Sarake1" "Sarake2" "Sarake3" "Sarake4" "Sarake5"
> paste("Sarake",1:5, sep="-")
[1] "Sarake-1" "Sarake-2" "Sarake-3" "Sarake-4" "Sarake-5"
```

Tässä siis argumentti `sep` kertoo mitä tekstin ja numeron väliin tulee. Voidaan myös muodostaa sarakenimiä, joissa on kirjaimia nimen ja numeron perässä.

**Esimerkki 4.** Katsotaan miten `paste()`-funktiolla voidaan muodostaa hivenen monimutkaisempia sarakenimiä:

```
> paste("Sarake", 1, letters[1:5], sep="")
[1] "Sarake1a" "Sarake1b" "Sarake1c" "Sarake1d" "Sarake1e"
> paste("Sarake", 1, LETTERS[1:5], sep="")
[1] "Sarake1A" "Sarake1B" "Sarake1C" "Sarake1D" "Sarake1E"
```

**Esimerkki 5.** Käytetään nyt `paste()`-funktiota muodostamaan sarja kiinnostavien sarakkeiden nimiä esimerkkitaulukosta `A`. Nyt esimerkkitaulukossa `A` on vain viisi saraketta, mutta valitaan ne kaikki:

```
# Valittavat sarakkeet:
> paste("A",1:5, sep="")
[1] "A1" "A2" "A3" "A4" "A5"
# Sitten valitaan:
subset(A, select=paste("A",1:5, sep=""))
```

**Esimerkki 6.** Jatketaan vielä edellistä esimerkkiä ja valitaan sarakkeista `A1, A3, A5` ne rivit, joilla `A2` on parillinen:

```
> subset(A, A2%%2 == 0, select=paste("A",seq(1,5,by=2), sep=""))
   A1 A3  A5
2   2 42  82
4   4 44  84
6   6 46  86
8   8 48  88
10 10 50  90
12 12 52  92
14 14 54  94
16 16 56  96
18 18 58  98
20 20 60 100
```

# 2 Aineiston luokittelu

Aineiston luokittelu tulee tarpeen erityisesti jatkuvien muuttujien (ikä, pituus, paino...) tapauksessa. Luokittelu voidaan tehdä käyttäen funktiota `cut()`, joka palauttaa annetun aineiston factorina. Tarkastellaan seuraavaksi `cut()`-funktion toimintaa esimerkein:

**Esimerkki 7.** Luokitellaan nyt taulukon `A` sarake `A1` kahteen luokkaan:

```
> cut(A$A1, breaks=c(0,11,20))
 [1] (0,11]  (0,11]  (0,11]  (0,11]  (0,11]  (0,11]  (0,11]  (0,11]  (0,11]  (0,11]
[11] (0,11]  (11,20] (11,20] (11,20] (11,20] (11,20] (11,20] (11,20] (11,20] (11,20]
Levels: (0,11] (11,20]
```

Huomaa: Mikäli jonkin solun arvo ei ole luokitteluun annetulla välillä, solun arvoksi tulee `NA`.

**Esimerkki 8.** Jatketaan edellisen esimerkin luokittelua ja jaetaan taulukon `A` sarake `A1` tällä kertaa kolmeen ja viiteen luokkaan. Katsotaan tuloksena saatavaa vektoria selvyyden vuoksi `table()`-komennon avulla:

```
> table(cut(A$A1, breaks=c(0,11,15,20)))

 (0,11] (11,15] (15,20]
     11       4       5
> table(cut(A$A1, breaks=c(0,6,10,15,18,20)))

  (0,6]  (6,10] (10,15] (15,18] (18,20]
      6       4       5       3       2
```

**Esimerkki 9.** Jatketaan edellistä esimerkkiä ja lasketaan kolmeen luokkaan jaetusta aineistosta luokkakohtaiset otoskeskihajonnat sarakkeelle `A3` käyttäen `tapply()`-funktiota:

```
> luokittelu <- cut(A$A1, breaks=c(0,11,15,20))
> tapply(A$A3, luokittelu, sd)
  (0,11]  (11,15]  (15,20]
3.316625 1.290994 1.581139
```

# 3 T-luottamusväli

Oletetaan, että käytettävän aineiston havainnot ovat otos normaalijakaumasta tuntemattomin parametrein $\mu$ ja $\sigma^2$. Tutustutaan estimaatin $\hat{\mu}$ luottamusvälin laskemiseen. Lisätietoa estimaateista ja luottamusväleistä löytyy samaan aikaan käynnissä olevan Johdatus tilastolliseen päättelyyn -kurssin luentomonisteesta.

**Esimerkki 10.** Olkoon nyt havaintovektori:

```
havainnot <- c(4,5,6,5,4,3,4,5,7,6,3,4,5,3)
```

Kaksisuuntainen t-luottamusväli, monen muun asian lisäksi, saadaan lasketttua komennolla `t.test()` seuraavasti:

```
> t.test(havainnot, conf.level = 0.99)

One Sample t-test

data:  havainnot
t = 13.9916, df = 13, p-value = 3.247e-09
alternative hypothesis: true mean is not equal to 0
99 percent confidence interval:
 3.587237 5.555620
sample estimates:
mean of x
 4.571429
```

Tässä käytettiin luottamustasoa 0.99 (`conf.level=0.99`). Luottamusvälin ylä- ja alarajat voidaan lukea kohdasta confidence interval. Tässä tapauksessa luottamusväli on siis pyöristettynä kahteen desimaaliin $[3.59, 5.56]$.

Muihin tämän funktion antamiin tuloksiin palataan myöhemmin.

**Esimerkki 11.** Edellisen esimerkin luottamusväli saadaan myös seuraavasti:

```
> a <- t.test(havainnot, conf.level = 0.99)
> a$conf.int
[1]  3.587237 5.555620
attr(,"conf.level")
[1] 0.99
```

Tämä tapa on huomattavasti kätevämpi silloin, kun ollaan kiinnostuttu yksinomaan luottamusvälistä tai halutaan päästä käsittelemään luottamusvälin ylä- ja alarajoja.

# 4 Kuvien piirtämisestä

Kuvien piirtäminen on aineiston analysoinnissa tärkeä vaihe. Kaikkia kuvia ei voi kuitenkaan piirtää `plot()`-funktiolla suoraan, vaan kuva joudutaan piirtämään osissa. Tällaisia tilanteita tulee esimerkiksi silloin, kun halutaan piirtää useita asioita samaan kuvaan.

Tutkitaan seuraavaksi miten `plot()`-funktiota voidaan käyttää piirtämään tyhjä kuva ja miten siihen voidaan lisätä asioita jälkikäteen käyttäen funktiota `lines()`.

**Esimerkki 12.** Luodaan tyhjä kuvaaja. On kuitenkin syytä kertoa `plot()`-funktiolle x- ja y-akselien minimit ja maksimit. Tämä voidaan tehdä käyttäen argumentteja `xlim` ja `ylim`:

```
plot(NULL, xlim=c(0,100), ylim=c(-1,1))
```

Tällä saadaan kuvaaja, joka on muuten tyhjä, mutta siihen on piirretty x-akseli välille [0,100] ja y-akseli välille [-1,1].

**Esimerkki 13.** Luodaan nyt tyhjä kuvaajaikkuna ja piirretään siihen pisteitä sini-funktiosta:

```
plot(NULL, xlim=c(0,2*pi), ylim=c(-1,1))
x_points <- seq(0,2*pi, length.out=50)
points(x=x_points, y=sin(x_points))
```

**Esimerkki 14.** Piirretään nyt esimerkin 12 koodilla luotavaan tyhjään kuvaajaan sini- ja kosini -käyrät:

```
x <- sapply(1:100, function(x) c(sin(x/10)+1, cos(x/10)-1))
plot(NULL, xlim=c(0,100), ylim=c(-1,1))
lines(x[1,], type="l")
lines(x[2,], type="l")
```

Tällä saadaan kuvaaja, joka on muuten tyhjä, mutta siihen on piirretty x-akseli välille [0,100] ja y-akseli välille [-1,1].

Katsotaan vielä, miten voidaan piirtää viivoja pisteiden välille. Seuraava esimerkki on mielenkiintoinen esimerkiksi luottamusvälin toiminnan havainnollistamisessa. Soveltaminen luottamusväliin jätetään kuitenkin tehtäväksi, joten sovelletaan tätä nyt edellisen esimerkin sini- ja kosini- funktioihin.

**Esimerkki 15.** Piirretään pystyviivoja sini-funktiosta kosini-funktioon:

```
> x <- sapply(1:100, function(x) c(sin(x/10)+1, cos(x/10)-1))
> plot(NULL, xlim=c(0,100), ylim=c(-1,1))
> segments(x0 = 1:100, y0 = x[1,], y1 = x[2,])
```

Tässä `segments()`-funktion argumentti `x0` on viivan paikka x-akselilla ja `y0`, `y1` ovat viivan päätepisteet y-akselilla.

# 5 Tehtäviä

1. Luodaan vastaajien tyytyväisyyttä Helsingin pyöräilyolosuhteisiin kuvaava summamuuttuja.

   a) Valitse osa-aineisto, joka sisältää muuttujat `Aq6-Aq15`, jotka kuvaavat vastaajien tyytyväisyyttä erilaisiin pyöräilyä koskeviin asioihin Helsingissä, ja tallenna se taulukkoon. Seuraavien kohtien tehtävät on helpompi tehdä tätä aputaulukkoa käyttäen, jolloin muuttujia ei tarvitse joka kerta valita erikseen. Vihje: Esimerkki 5.

   b) Tulosta muuttujien `Aq6-Aq15` frekvenssitaulut. Vihje: `apply()`.

   c) Muuta muuttujien `Aq6-Aq15` ”En osaa sanoa”-vastaukset puuttuviksi. Tarkista onnistuminen tulostamalla muuttujien frekvenssitaulut uudelleen.

   d) Laske vastausten `Aq6-Aq15` keskiarvot *riveittäin* (siis jokaiselle rivil-le lasketaan vastaajan vastausten keskiarvo kysymyksistä `Aq6-Aq15`) ja tallenna tulos alkuperäiseen taulukkoon `pb` uudeksi muuttujaksi `tyytyvaisyys`. Vihje: `apply()` ja `cbind()`.

2. Aineistossa on valmis viiteen luokkaan luokiteltu ikä-muuttuja `Aikalk1`. Luodaan vastaava muuttuja itse.

   a) Luo aineistoon uusi muuttuja `ika5`, jonka luokittelee vastaajat viiteen ikäryhmään *täsmälleen* samalla tavalla kuin `Aikalk1`. Vihje: `cut()` ja `cbind()`.

   b) Testaa onnistuminen ristiintaulukoimalla luomasti `ika5` ja `Aikalk1` (käytä `useNA`-argumenttia, niin näet myös mahdolliset puuttuvat arvot). Tuloksena tulisi olla taulu, jossa kaikki muut paitsi diagonaalialkiot ovat nollia. Jos epäonnistuit, palaa a-kohtaan ja yritä uudelleen.

   c) Nimeä muuttujan `ika5` tasot uudelleen jotta saat siistimmän tulosteen, esimerkiksi samoiksi kuin muuttujan `AikAlk1` tasot. Tulosta muuttujan `ika5` frekvenssitaulu

   d) Laske `tyytyvaisyys`-summamuuttujan keskiarvot ikäluokittain. Vihje: Esimerkki 9.

3. Luottamusvälin laskeminen.

   a) Laske kaksisuuntainen t-luottamusväli muuttujan `tyytyvaisyys` keskiarvolle 95%:n luottamustasolla sijoittamalla oikeat arvot luottamusvälin kaavaan

   $$\left[\bar{y} - t_{n-1}(0.025)\frac{s}{\sqrt{n}}, \quad \bar{y} + t_{n-1}(0.025)\frac{s}{\sqrt{n}}\right],$$

   missä $\bar{y}$ on muuttujan `tyytyväisyys` keskiarvo, $s$ sen keskihajonta ja $n$ otoskoko (huom. ei sama kuin koko aineiston otoskoko, koska muuttujalla `tyytyväisyys` puuttuvia arvoja). Merkintä $t_{n-1}(0.025)$ tarkoittaa vapausasteen $n-1$ t-jakauman 0.025-yläkvantiilia, eli pistettä, jonka oikealla puolella on 2.5% jakauman todennäköisyysmassasta, eli vasemmalla puolella on 97,5% jakauman todennäköisyysmassasta, eli kyseessä on 0.975-kvantiili. Vihje `qt()`.

   b) Tarkista tulos komennolla `t.test(pb$tyytyvaisyys)$conf.int`.

4. Tarkastellaan kysymyksen Aq3 (tyytyväisyys Helsinkiin pyöräilykaupunkina) vastauksia sukupuolittain. Laske 95% kaksisuuntaiset t-luottamusvälit miehille ja naisille erikseen. Vertaa luottamusvälien pituuksia toisiinsa.

5. Pohditaan seuraavaksi kurssin Johdatus Tilastolliseen päättelyyn harjoitusten 3 tehtävää 2. Tehtävässä tutkitaan auton polttoaineen keskikulutusta kaupunkiajossa. Ajoa suorittaa kuusi eri kuljettajaa, joiden keskikulutukseksi havaitaan: 6.1, 5.7, 5.9, 6.8, 6.7 ja 6.0.

   a) Tallenna annetut mittaustulokset vektoriin ja laske sen keskiarvo ja keskihajonta.

   b) Laske keskikulutukselle 95% luottamusväli käyttäen funktiota `t.test()`.

6. Tutkitaan nyt tapausta jossa $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$ ovat riippumattomia. Parametrin $\mu$ suurimman uskottavuuden estimaattori on tunnetusti satunnaismuuttujien $Y_i$, $i = 1, \ldots, n$ keskiarvo, siis $\hat{\mu} = \overline{Y} = n^{-1}\sum_{i=1}^{n} Y_i$.

   Tutkitaan nyt estimaattorin $\hat{\mu}$ kaksisuuntaisen 95% t-luottamusvälin merkitystä, kun molemmat parametrinarvot ovat tuntemattomia. Käytetään simuloinnissa jakaumaa N(0,1).

   a) Simuloi 200 otosta joissa jokaisessa otoskoko $n = 100$, laske otoksista kaksisuuntaiset 95% t-luottamusvälit odotusarvolle ja tallenna ne muuttujaan `luottamus`. (Vihje: `sapply()` ja esimerkki 10). *Voit myös halutessasi laskea ylä- ja alarajat omiin vektoreihinsa.*

   b) Luo tyhjä `plot()`-ikkuna, jossa y-akseli on välillä $[-1, 1]$ ja x-akseli välillä $[0, 200]$. Vihje: Esimerkki 12.

   c) Piirrä kuvaan pystyviivat luottamusvälin ylärajasta ja alarajaan käyttäen funktiota `segments()`. Vihje: Esimerkki 15.

   d) Piirrä kuvaan punainen viiva tasolle y=0. Vihje: `abline()`.

   e) Tarkastele silmämääräisesti: Peittävätkö luottamusvälit todellisen parametrinarvon noin noin 95%:ssa tapauksista?

7. Tarkastellaan tehtävässä 6 piirrettyjä luottamusvälejä, jotka tallennettiin muuttujaan `luottamus` (*tai ylä- ja alarajat omiin vektoreihinsa*). Laske kuinka suuri osuus luottamusväleistä peittää parametrin todellisen arvon $\mu = 0$.

8. Piirrä tehtävän 6 kuva uudestaan, nyt käyttäen ensin otoskokoa $n = 50$ ja sitten $n = 500$. Mitä huomaat?