Tuomas Kujala

Katsaus “R”:ään, tilastolliseen ohjelmointiympäristöön

1 Johdanto

Halusin tehdä harjoitustyöni kertomalla vaihtoehtoisesta tilastollisesta ohjelmasta, sillä käytän paljon Linux-ympäristöä, jolle kurssilla käytettävää SPSS-ohjelmistoa ei ole ollut vielä kovin kauaa saatavilla. SPSS-ohjelmiston kotisivuilta saa ladattua nykyisin myös Linux-alustalla toimivan version, muta tämä versio on rajoitettu demo-versio, jota voi käyttää ainoastaan kahden viikon ajan (kuten myös Windowsissa).

Esittelemäni R on – kuten otsikkokin kertoo – enemmänkin tilastollinen ohjelmointiympäristö kuin pelkkä tilastollinen laskentaohjelma. R on täysin vapaa monella eri järjestelmällä toimiva ohjelmisto, joka on julkaistu lisenssin “GNU General Public License” alla. Jos haussa on siis tehokas ja ilmainen vaihtoehto esim. SPSS:lle, kannattaa tutustua tarkemmin R:ään.

Pieni varoituksen sana; R:n käyttöliittymä on oletusarvoisesti täysin komentorivipohjainen, joten minkäänlaista graafista käyttöliittymää ei ole tarjolla, ellei sellaista itse asenna. Näin ollen R:n käyttö saattaa tuntua aluksi hieman haasteelliselta, varsinkin jos käyttäjällä on aikaisempaa kokemusta ainoastaan graafisista käyttöliittymistä. Jos komentorivin käyttö ei tunnu tutulta ja pelkän näppäimistön käyttö arvelluttaa, osoiteesta `http://www.sciviews.org/_rgui/` löytyy erilaisia käyttöliittymiä R:lle. Koska erityyppisten käyttöliittymien kirjo on melko suuri, jätetään niihin tutustuminen lukijan vastuulle. Tässä tekstissä keskitytään siis ainoastaan R:n komentorivipohjaiseen käyttöön.

2 Perustietoja R:stä

Mikä R sitten oikein on? Kuten edellä mainittiin, ei se ole pelkästään tilastollinen laskentaohjelma, vaan paljon muutakin. R:n ydin on tulkattava ohjelmointikieli, joka mahdollistaa haarautumisen, toistorakenteet ja omien funktioiden kirjoittamisen. Ydin itsessään ei mahdollista esim. tilastollista analyysia, vaan kaikenlainen ylimääräinen toiminnallisuus toteutetaan R:ään liitettävillä eräänlaisilla moduuleilla. Näitä moduuleita on

tarjolla monenlaisia tehtäviä varten, ja jos omiin tarpeisiin sopivaa moduulia ei löydy ja taitoa löytyy, moduulin voi toki ohjelmoida itsekin. R:n asennuspaketti sisältää jo valmiiksi monia tärkeitä ja useimmiten käytettyjä moduuleja, joten moduulien asentamiseen ja hallintaan ei perehdytä tässä tarkemmin.

Käytössäni olevaan Linuxiin (Ubuntu – versio 8.04, Hardy Heron) R:n saa asennettua helposti Ubuntun oman pakettienhallintaohjelman (Synaptic) kautta. Asennuspaketin nimi on `r-base`. Ubuntun pakettienhallinnasta löytyy myös todella paljon moduuleita R:ään. Moduulit tunnistaa pakettienhallinnassa nimen alkuosasta "`r-cran`". Jos R:ää ei saa asennettua jostain syystä Linux-jakelunsa pakettienhallinnan kautta, tai käytössä on Windows, voi R:n tuoreimman version asennuspaketin (tai lähdekoodit) ja puuttuvat moduulit noutaa myös internetistä osoitteesta `http://cran.r-project.org/`.

3 R:n peruskäyttö

Asennuksen jälkeen R käynnistetään Linuxissa kirjoittamalla komentoriville `R` (Huomioi, että kirjaimen tulee olla iso `R`, ei `r`) Windowsissa asennusohjelma luo työpöydälle käynnistyskuvakkeen. Käynnistymisen jälkeen näytölle tulee seuraavankaltainen teksti (mm. versionumero voi vaihdella, itselläni on käytössä versio 2.6.2):

```
R version 2.6.2 (2008-02-08)
Copyright (C) 2008 The R Foundation for Statistical Computing
ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
```

```
>
```

Windows-versiossa R avautuu ikkunaan, johon yllämainittu teksti aukeaa. Windows-version ikkunan yläreunassa on valmiina jo joitakin peruskäyttöön liittyviä komentoja, kuten vanhan työtilan lataaminen ja ohjelmasta poistuminen. Tästä eteenpäin en käsittele laajemmin ohjelman toimintaa Windowsilla, sillä se on hyvin samankaltaista Linux-version kanssa. Kaikki esimerkkikomennot on kirjoitettu tasavälisellä fontilla `Courier New`, jotta ne pystyisi helposti erottamaan leipätekstistä.

Ensimmäinen komento, joka on hyvä opetella, on R:stä poistuminen. Antamalla komennon `q()`, järjestelmä sulkee R:n ja palaa komentoriville (Windowsissa työpöydälle). Komennot tunnistaa R:ssä siitä, että ne loppuvat aina sulkuihin. Sulkujen sisältö voi olla tyhjä, kuten poistumiskomennossa, tai sitten ne voivat sisältää joitakin kyseiseen komentoon liittyviä parametreja. Komennon antamisen jälkeen R kysyy: "`Save workspace image? [y/n/c]:`" Vastaamalla kysymykseen myönteisesti (`y` = yes) R tallentaa käytössä olleen työtilan tulevaa käyttöä varten. Käytössä olleen työtilan voi luonnollisesti hylätä kielteisellä vastauksella (`n` = no) ja R:ään voi palata takaisin vastauksella `c` (cancel). Työtilan tallennus on hakemistokohtainen, joten jos haluat ladata viimeksi tallentamasi työtilan, muista käynnistää R samasta hakemistosta kuin edellisellä kerralla.

Yksi erittäin hyödyllinen komento R:ssä on `help()`. Tämän komennon antamalla R antaa hyödyllisiä ohjeita ohjelman yleisestä käytöstä. Tietystä komennosta saa ohjeita kirjoittamalla `help(komento)` tai `?komento`, joissa sana "komento" korvataan tietenkin tarvittavan komennon oikealla nimellä (tässä tapauksessa komentoa yleensä seuraavat sulut jätetään pois). Esimerkiksi `help(q)` antaa tietoja ohjelman sulkemiseen käytettävästä `q`-komennosta. Jos ei ole aivan varma komennon koko nimestä, voi osalla komennon nimestä hakea sanaa muistuttavia komentoja, esimerkiksi `help.search("apply")` palauttaa kaikki komennot, joiden nimessä sana "apply" esiintyy. Mitkä tahansa ohjeet voi sulkea painamalla `q`.

Kaikkia eri komentoja ja niiden parametreja ei liene tarpeen luetella tässä, sillä käytettävissä olevat komennot riippuvat käytössä olevista moduuleista. Käytössä olevat moduulit näkee komennolla `library()`, ja tietyn moduulin sisältämät komennot saa esille komennolla `library(help=moduuli)`, missä "moduuli" on tarkasteltavan

moduulin nimi. Esimerkkien yhteydessä mahdollisia uusia komentoja käytettäessä kerrotaan, mitä kyseinen komento tekee, mistä moduulista se löytyy ja miten sitä käytetään.

4 Aineiston lataaminen R:ään

Käsiteltävän aineiston voi syöttää käsin R:ään, mutta kyseinen tapa on melko vaivalloinen ja virhealtis, eikä käsin tallennukseen ole yleensä syytä jos aineisto on jo tallennettu jossain muussa muodossa. Kaksi tässä käsiteltävää tapaa syöttää aineistoja R:ään on lukea aineisto SPSS-tiedostosta (tiedostopääte "`.sav`") ja CSV-tiedostosta. CSV-tiedostot ovat perinteisesti tavallisia tekstitiedostoja, joiden riveillä tietoalkiot ovat eroteltu toisistaan jollakin merkillä, esimerkiksi pilkulla. Käsiteltäessä tilastoaineistoja CSV-tiedoston yksi rivi kuvaa yhtä havaintoyksikköä. Havaintoyksikön eri muuttujien tiedot on siis eroteltu toisistaan jollain merkillä.

SPSS-tiedostojen lukemiseen käytetty komento ei ole automaattisesti käytössä, vaan R:n käyttöön pitää ladata moduuli "foreign". Lataaminen tapahtuu kirjoittamalla `library(foreign)`. Tätä kyseistä moduulia ei siis tarvitse ladata internetistä, vaan se kuuluu ainakin Ubuntu-version oletusmoduuleihin. Moduuli vain ladataan R:n käyttöön levyltä.

Moduulin lataamisen jälkeen voidaan ladata itse aineisto SPSS-tiedostosta. Tämä onnistuu komennolla `muuttuja = read.spss("tiedoston_sijainti", to.data.frame = TRUE)`. Aivan ensimmäiseksi valitaan muuttujan nimi (ja kirjoitetaan sijoitusoperaattori "=" tai "`<-`"), johon ladatun aineiston tiedot tallennetaan. Tämän jälkeen komennolle annetaan parametreiksi tiedoston sijainti. Huomioi, että Windows-versiossa tiedostopolkuja kirjoitetaessa kenoviiva "`\`" tulee kirjoittaa tuplana "`\\`", sillä yksittäistä kenoviivaa käytetään ohjausmerkkinä. Parametri "`to.data.frame = TRUE`" on hyödyllinen, sillä se tallentaa aineiston helposti käsiteltävään muotoon. Komennon suorittamisen jälkeen aineisto on valmiina käyttöä varten.

Jos aineisto on tallennettu tekstitiedostoon CSV-muotoon, voidaan tiedot ladata (esimerkki)komennolla `muuttuja = read.table("tiedoston_sijainti", sep=";", dec=".", header=TRUE)`. Komennon `read.table()` sisältävä moduuli on ladattu käyttövalmiiksi jo R:n käynnistyksessä, joten sitä ei tarvitse ladata erikseen kuten SPSS-tiedostojen tuonnissa. Kuten edellä, myös tämän komennon

yhteydessä määritellään aluksi muuttujan nimi, sijoitusoperaattori, itse komennon nimi sekä ensimmäisenä parametrina tiedoston sijainti. Tämän lisäksi kannattaa määritellä parametrit "`sep`" ja "`dec`". Näistä "`sep`" määrittelee, mikä merkki erottaa havaintoyksikön muuttujat toisistaan tiedostossa. Parametri "`dec`" taas ilmaisee, millä merkillä mahdollisten desimaalilukujen kokonaisosa on eroteltu desimaaliosasta. Jos CSV-tiedoston ensimmäinen rivi sisältää muuttujien nimet, voi ne tuoda R:ään parametrilla "`header=TRUE`". Lisätietoja aineiston tuonnista saa komennolla "`?read.table`".

Jos kaikki meni hyvin aineistoa tuotaessa, ei R ilmoita asiasta erikseen. Vikatilanteissa näytölle ilmestyy enemmän tai vähemmän informatiivinen virheilmoitus.

5 Aineiston perustarkastelu

Aineiston tuonnin jälkeen on hyvä tarkistaa, että tiedot tuotiin R:ään oikein. Huomioi, että vaikka R ei olisikaan tuonnin yhteydessä ilmoittanut virheestä, eivät kaikki tiedot ole välttämättä tallennettu oikein. R voi esimerkiksi lukea virheellisesti kvantitatiivisen muuttujan kvalitatiiviseksi, joka saattaa johtaa myöhemmin vaikeasti havaittavaan ongelmakäyttäytymiseen.

Komennolla `str(aineisto)` saa tietoja tuodun aineiston rakenteesta (str = structure). Komennossa parametri "aineisto" korvataan tietenkin muuttujan nimellä, johon aineisto tuotiin. Aineiston rakenteesta näkee, millaisiksi R on tuodut muuttujat luokitellut. Luokitteluja ovat esimerkiksi: factor (luokitteluasteikollinen), ordered factor (järjestysasteikollinen), int/num (kvantitatiivinen) ja merkkijonomuuttuja (char). Jos tuodussa aineistossa on koodattu luokitteluasteikollinen muuttuja numeerisesti kokonaisluvuilla, voidaan tällainen muuttuja muuttaa numeerisesta takaisin luokitteluasteikolliseksi "factor"-komennolla. Seuraavalla esimerkkikomento muuttaa "ihmiset"-nimisen aineiston muuttujan "sukupuoli" luokitteluasteikolliseksi ja korvaa aineiston vanhan kvantitatiivisen muuttujan: `ihmiset = transform(ihmiset,sukupuoli=factor(sukupuoli,labels=c("Nainen","Mies")))`. Muuttujan voi muuttaa järjestysasteikolliseksi lisäämällä factor-komentoon parametriksi "`ordered=TRUE`". Komennossa käytetyn "transform":in avulla aineistoon voidaan lisätä muuttujia. Lisätietoja molemmista komennoista löytyy R:n ohjeista.

Luokitteluasteikollisen muuttujan voi luokitella uudelleen komennolla recode, jolloin muuttujan luokkia voi yhdistää suuremmiksi kokonaisuuksiksi (ohjeet R:ssä).

Komento recode ei löydy R:n oletusmoduuleista, vaan komennon käyttö vaatii moduulin "car" asennuksen (Ubuntun pakettienhallinnassa nimellä "`r-cran-car`"). Myös jatkuva muuttuja on mahdollista muodostaa luokitteluasteikolliseksi. Tämä tapahtuu komennolla "cut".

6 Tunnuslukuja aineistosta

Tavallisimpia tunnuslukuja aineistosta saa esille komennolla `summary(aineisto)`. Tämä tulostaa näytölle kaikista kvantitatiivisista muuttujista muuttujan minimin, alakvartiilin, mediaanin, keskiarvon, yläkvartiilin sekä maksimin. Kvalitatiivisista muuttujista tulostetaan muuttujan luokkien frekvenssit. Jos puuttuvia arvoja on, ne ilmoitetaan. Aineiston tietystä yksittäisestä muuttujasta samat tunnusluvut saa komennolla `summary(aineisto$muuttuja)`. Yleisestikin aineiston tiettyyn muuttujaan viitataan merkinnällä "`aineisto$muuttuja`".

Muuttujien frekvenssit saa esiin komennolla `table(aineisto$muuttuja)`. Tätä komentoa ei voi kuitenkaan käyttää kerralla koko aineistoon, eikä "table" näytä puuttuvia arvoja.

Kvantitatiivisten muuttujien tutkimiselle on olemassa myös monta yksittäistä komentoa, joihin ei perehdytä tässä tarkemmin. Näitä ovat mm. `min` (minimi), `max` (maksimi), `range` (arvoalue), `mean` (keskiarvo), `median` (mediaani), `quantile` (4-kvantiili), `sd` (keskihajonta) ja `var` (varianssi).

7 Graafisia esityksiä

Aineistoa voi tutkia myös mainiosti erilaisten graafisten esitysten avulla. R mahdollistaa monenlaisten, muunneltavien grafiikoiden piirtämisen. Grafiikat esitetään oletusarvoisesti suoraan näytöllä, mutta ne voi myös tulostaa tai tallentaa tiedostoon myöhempää käyttöä varten. Tulostamiseen suoraan R:stä ei perehdytä tässä, sillä R:stä tallennetun grafiikan voi myös tulostaa erikseen.

Muuttujien perustarkasteluun sopivat esimerkiksi komennot `barplot` (pylväsdiagrammi), `pie` (piirakkadiagrammi) ja `hist` (histogrammi). Komento pie vaatii

parametrikseen muuttujan, joka sisältää ainoastaan positiivisia lukuja. Luokitteluasteikollisista muuttujista voi kuitenkin tehdä piirakkadiagrammin, kunhan antaa piirrettävästä muuttujasta luokkien frekvenssit R:n käyttöön. Tämä tapahtuu yksinkertaisesti komennolla `pie(table(muuttuja))`. Komento `table(muuttuja)` siis laskee muuttujan frekvenssit, kuten edellisessäkin luvussa todettiin.

Toinen kätevä esitysmuoto on pylväsdiagrammi, eli `barplot`. Barplot toimii samalla periaatteella kuin `pie`, eli luokitteluasteikollisia muuttujia ei saa suoraan piirrettyä komennon avulla, vaan avuksi tarvitaan `table`:a.

Komento hist piirtää numeerisen muuttujan histogrammin. Komennolla voi antaa parametriksi sanan "`breaks=n`", jossa n kertoo, kuinka monta palkkia histogrammiin piirretään. Esimerkkikomento voisi olla vaikkapa `hist(pituus, breaks=8)`.

Sirontakuvioiden piirtäminen onnistuu komennoilla `pairs` ja `plot`. Käyttäjä voi tutkia aineiston kaikkien muuttujien riippuvuutta toisistaan komennolla `pairs(aineisto)`. Koko aineiston tutkiminen samalla kertaa sirontakuvion avulla ei ole suositeltavaa, jos muuttujia on kovin monta, sillä piirtäminen on hidasta ja yksittäisistä graafeista tulee niin pieniä, ettei niistä tahdo saada selvää. Kahden muuttujan välisen pisteparven voi piirtää yksinkertaisesti komennolla `plot(muuttuja1, muuttuja2)`.

Komento plot on piirtämisessä kätevä työkalu. Se on geneerinen komento (kuten monet muutkin R:n komennoista), joka hyväksyy monia erilaisia parametreja. Komennon ohjesivut kannattaa lukea tarkkaan, sillä ne sisältävät monia esimerkkejä ja parametreja, joita tässä ei käsitelty.

Oletuksena grafiikka piirtyy tietokoneen näytölle, eikä sitä voi näytön kautta tallentaa mitenkään (muuten kuin PrintScreenin avulla).[1)] Seuraavaksi käsitellään hieman grafiikoiden tallennusta tiedostoihin. Grafiikkaa voi tallentaa eri formaatteihin, ja seuraavan käsitellään grafiikan tallentamista yleisellä tasolla yhden esimerkin avustuksella.

R osaa tallentaa grafiikat moneen eri tiedostomuotoon, joista yleisimpiä ovat jpg, png, pdf, postscript, bmp ja tiff. Kaikkiin tiedostomuotoihin tallentamisessa on sama periaate; ensiksi annetaan komennon nimenä tallennusmuoto (jpeg (sic), bmp, pdf, bmp, tiff tai postscript), jonka jälkeen annetaan parametrina tiedostonimi. Tämä komento vasta alustaa tiedostoon tallentamisen, mitään ei ole vielä kuitenkaan tehty. Alustamisen jälkeen määritellään (l. annetaan komentona) itse grafiikka, joka tallennetaan. Tämä voi olla mikä tahansa graafinen esitys, jonka R oletuksena esittäisi näytöllä. Vielä grafiikan

1) Tallentaminen onnistunee vahvistamattoman tiedon mukaan Windows-versiossa suoraan itse grafiikkaikkunasta.

määrittelynkään jälkeen tiedostoa ei ole tallennettu levylle. Kahden edellisen vaiheen jälkeen tulee antaa vielä komento `dev.off()`, joka sulkee käytössä olevan grafiikkalaitteen, joka on tässä tapauksessa tallennettava tiedosto. Näiden kolmen vaiheen jälkeen grafiikka tallentuu annetulla tiedostonimellä hakemistoon, josta R käynnistettiin. Lähes kaikki grafiikkamuodot tukevat myös kuvan koon, resoluution ja monen muun parametrin asettamista. Näistä lisää kunkin komennon ohjeistuksessa. Alla esimerkkikomentosarja pituus-paino -sirontakuvion tallentamisesta levylle jpg-muotoon:

```
jpeg(“sironta.jpg”, width = 640, height = 640, quality = 80)
plot(pituus, paino)
dev.off()
```

8 Aineiston analysointi ja testaus

Aineiston perustarkastelun jälkeen on aika syventyä hieman datan tarkempaan analysointiin. Tässä kappaleessa esitellään joitakin tilastollisia menetelmiä aineiston tarkempaa tutkimista varten.

Ristiintaulukointi onnistuu R:ssä jo aikaisemminkin käytetyllä monitoimikomennolla `table`. Muodostettaessa ristiintaulukko tulee käsiteltävien muuttujien olla kvalitatiivisia. Jos aineiston kaikki muuttujat ovat tuonnin jäljiltä numeerisia, onnistuu niiden muuntaminen luokitteluasteikollisiksi kappaleen 5 ohjeiden mukaan. Muodostettava ristiintaulukko kannattaa tallentaa erilliseen muuttujaan, jolloin sen jatkokäsittely on helpompaa. Ristiintaulukon muodostaminen onnistuu helposti komennolla `risttaul = table(muuttuja1, muuttuja2, dnn=list(“muuttuja1:n kuvaus”, “muuttuja2:n kuvaus”))`. Kuten nähdään, parametrilla `dnn` voi kontrolloida muodostettavan ristiintaulukon dimensioiden nimiä. Huomioi, että puuttuvat arvot jätetään pois ristiintaulukosta. Komennolla `prop.table(risttaul, 1)` saa ristiintaulukon prosenttiosuudet selittävän muuttujan mukaan. Komennossa parametri `risttaul` on luonnollisesti edellä luotu ristiintaulukko. Muuttamalla komennon jälkimmäinen parametri “2”:ksi prosenttiosuudet lasketaan selitettävän muuttujan mukaan.

Ristiintaulukoiden yhteydessä tutkitaan useasti taulukon kahden muuttujan välistä riippuvuutta. Tämän riippuvuuden tarkastelu onnistuu X^2 -riippumattomuustestillä (lue: khi toiseen). X^2 -riippumattomuustesti onnistuu helposti komennolla

`chisq.test(risttaul)`. Huomioi, että R ei testaa automaattisesti testin oikeellisuutta, vaan käyttäjän tulee itse tarkistaa komennolla `chisq.test(risttaul) $expected`, että kunkin ristiintulon solun odotettu frekvenssi on yli 1, ja soluja, joiden odotettu frekvenssi on alle 5, on alle 20% kaikista soluista.

R:llä voidaan testata myös odotusarvojen yhtäsuuruutta riippumattomien otosten t-testillä. Ennen itse t-testiä on kuitenkin hyvä suorittaa esim. F-testi varianssien yhtäsuuruudelle komennolla `var.test(muuttuja1, muuttuja2)`. Varianssien yhtäsuuruuden testaamisen jälkeen voidaan edetä itse t-testiin. Itse testaus tapahtuu yksinkertaisesti komennolla `t.test(muuttuja1, muuttuja2, var.equal = TRUE)`. Äskeisessä on oletettu varianssit yhtäsuuriksi; jos varianssit eivät ole yhtä suuria, vaihdetaan parametrin `var.equal` arvoksi `FALSE`.

Luottamusvälin laskeminen onnistuu myös samalla komennolla: `t.test(paino)`. Testaukseen voi liittää hypoteesin komennolla `t.test(paino, mu = 80, conf.level = 0.01)`. Edellä siis oletetaan, että havaintoryhmän paino on keskimäärin 80 (kiloa). Lisäksi luottamusväliksi on asetettu 99% (100 * [1 – 0.01]).

Kahden muuttujan välisen korrelaatiokertoimen löytää yksinkertaisesti komennolla `cor(muuttuja1, muuttuja2)`.

Prosenttiosuuden luottamusvälin pitäisi saada laskettua yksinkertaisesti komennolla `prop.test(suotuisat, kaikki)`, missä `kaikki` on otoksen suuruus kokonaislukuna ja `suotuisat` on suotuisat tapaukset koko otoksesta. Luottamusvälin voi määritellä 95%:stä poikkeavaksi edellämainitun esimerkin tapaan parametrilla `conf.level`. Komento `prop.test` antaa jostain syystä hieman esim. luentomonisteen esimerkeistä poikkeavia lukuja. Tämä käytös jäi hieman mysteeriksi.

9 Oman aineiston esimerkinomainen tutkiminen

Tässä kappalessa esittelen omaa harjoitustyöhön liittyvää aineistoani ja joidenkin edellä esiteltyjen menetelmien soveltamista tähän kyseiseen aineistoon.

Tarkoitukseni oli alunperin suoritaa kurssin harjoitustyö perinteisellä tavalla, joten kerkesin kerätä oman aineistoni ennen tämän tekstin kirjoittamista. Aineistoni on peräisin Forbes-talouslehden maailman 2000 suurimman yrityksen listauksesta. Keräsin omaan dataani edellämainitusta listauksesta 500 suurinta yritystä. Vaikka tämä data ei aivan satunnaisotoksen vaatimuksia täytäkään, voidaan näin kuitenkin kuvitella, jotta testaukset

voidaan suorittaa. Forbesin listalla yrityksistä on tallennettu seuraavat tiedot: sija listalla, nimi, kotimaa, ammattiala, edellisen vuoden myynti miljardeina dollareina, edellisen vuoden tuotto miljardeina dollareina, sijoitukset miljardeina dollareina sekä markkina-arvo miljardeina dollareina.

Lähden ilman pidempiä puheita analysoimaan aineistoani. Kerron joka kohdassa, mitä olen tekemässä ja liitän tähän dokumenttiin teksti- ja kuvakaappaukset (grafiikoita käsiteltäessä) R:stä, jotta lukija näkee millaista R:n käyttö oikeastaan on.

Aloitetaan lataamalla aineisto tiedostosta “firmat.csv” R:ään ja tutkimalla muuttujaa, jossa tiedot sijaitsevat. Tämän jälkeen tulostetaan aineistosta joitakin tavallisimpia tunnuslukuja ja keskihajonnat:

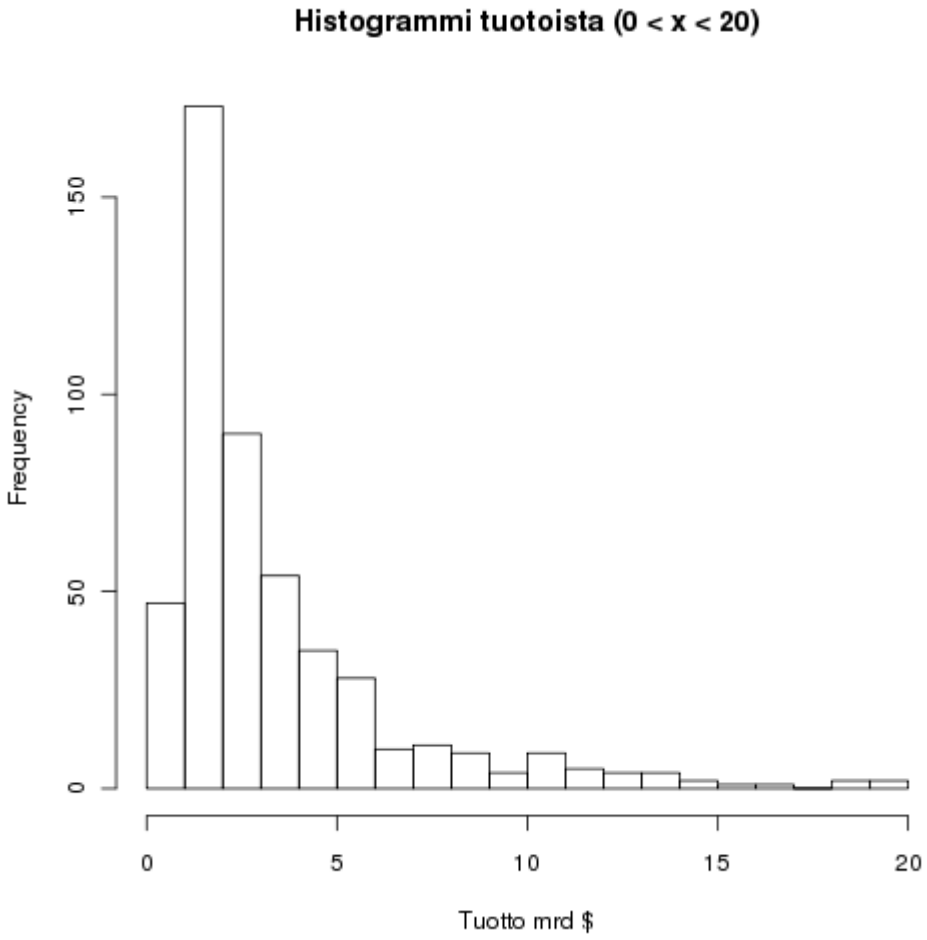
```
> firmat = read.table("firmat.csv", header=T, sep=",", dec=".")
> str(firmat)
'data.frame':     500 obs. of  7 variables:
 $ SIJA         : int  1 2 3 4 5 6 7 8 9 10 ...
 $ MAA          : Factor w/ 43 levels "Alankomaat","Alankomaat/Yhdistynyt
kuningaskunta",..: 40 43 43 43 43 1 40 18 1 43 ...
 $ AMMATTIALA   : Factor w/ 27 levels "Elintarvikkeet & savukkeet",..: 18
15 18 18 17 17 17 9 27 14 ...
 $ MYYNTI       : num  146 173 119 116 359 ...
 $ TUOTTO       : num  19.1 22.2 15.0 15.4 40.6 ...
 $ SIJOITUKSET  : num  2349  795 1716 1562  242 ...
 $ MARKKINA_ARVO: num  181 331 177 137 466 ...

> summary(firmat)
      SIJA                                MAA                     AMMATTIALA
 Min.   :  1.0   Yhdysvallat             :164   Pankkitoiminta      : 87
 1st Qu.:125.8   Japani                  : 47   Öljy- & kaasuala    : 47
 Median :250.5   Yhdistynyt kuningaskunta: 34   Julkiset palvelut   : 40
 Mean   :250.4   Ranska                  : 32   Vakuutusala         : 40
 3rd Qu.:375.2   Saksa                   : 27   Monialainen rahoitus: 28
 Max.   :500.0   Kanada                  : 20   Telekommunikaatioala: 24
                 (Other)                 :176   (Other)             :234
     MYYNTI           TUOTTO          SIJOITUKSET      MARKKINA_ARVO
 Min.   :  3.96   Min.   :-10.590   Min.   :  10.11   Min.   :  8.17
 1st Qu.: 13.79   1st Qu.:  1.347   1st Qu.:  27.25   1st Qu.: 19.55
 Median : 23.56   Median :  2.400   Median :  50.31   Median : 32.65
 Mean   : 39.11   Mean   :  3.564   Mean   : 186.90   Mean   : 51.60
 3rd Qu.: 45.45   3rd Qu.:  4.075   3rd Qu.: 139.45   3rd Qu.: 57.88
 Max.   :378.80   Max.   : 40.610   Max.   :3807.51   Max.   :546.14

> sd(firmat$MYYNTI)
[1] 44.29029
> sd(firmat$TUOTTO)
[1] 4.139965
> sd(firmat$SIJOITUKSET)
[1] 393.4242
> sd(firmat$MARKKINA_ARVO)
[1] 57.18308
```

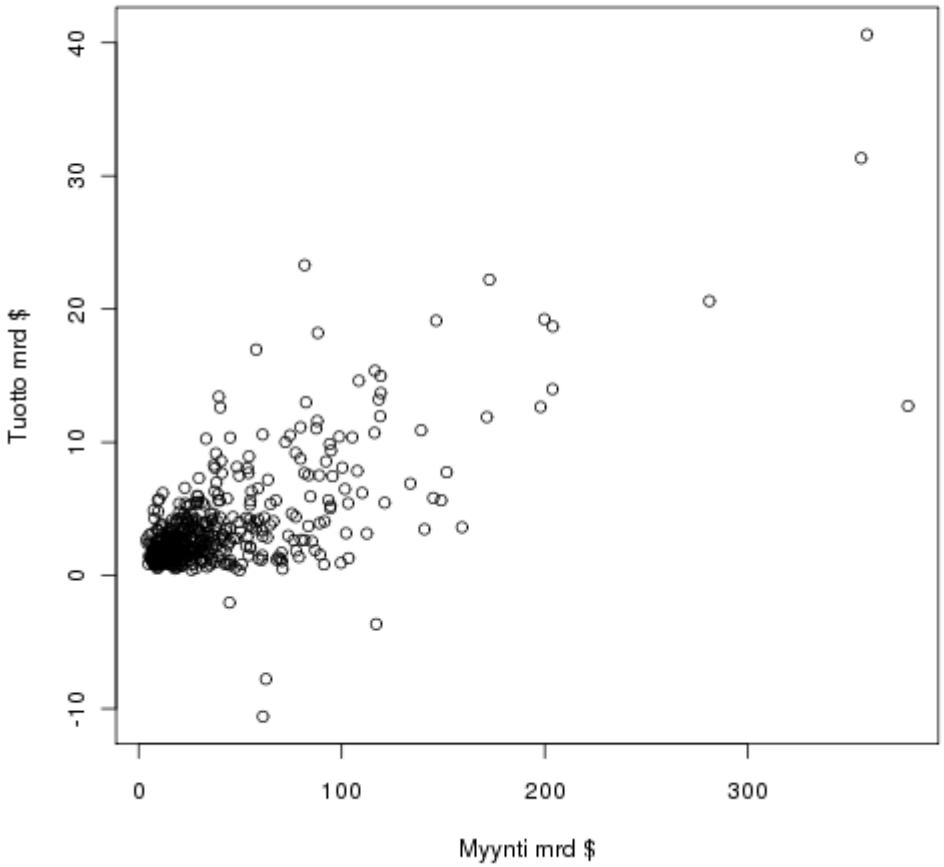
Tunnuslukujen jälkeen tutkimme joitakin graafisia esityksiä aineistosta. Ensimmäisenä yrityksen tuottojen histogrammi rajattuna välillä 0 – 20 mrd $ esitettynä 20 pylväällä:

```
> png("histogrammi_tuotto.png", width = 500, height = 500)
> hist(firmat$TUOTTO[firmat$TUOTTO > 0 & firmat$TUOTTO < 20], breaks = 20,
main = "Histogrammi tuotoista (0 < x < 20)", xlab = "Tuotto mrd $")
> dev.off()
```



Seuraavaksi sirontakuvio yrityksen myynnin vaikutuksesta tuottoon:

```
> plot(firmat$MYYNTI, firmat$TUOTTO)
```



Grafiikoiden esittelyn jälkeen tutkimme hieman aineiston analysointia. Tutkitaan, mikä on yllä esitetyn kuvan myynnin ja tuoton korrelaatiokerroin:

```
> cor(firmat$MYYNTI, firmat$TUOTTO)
[1] 0.7167398
```

Lasketaan firmojen sijoitusten odotusarvolle 99%:n luottamusväli:

```
> t.test(firmat$SIJOITUKSET, conf.level = 0.99)

      One Sample t-test

data:  SIJOITUKSET
t = 10.6228, df = 499, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
99 percent confidence interval:
 141.4075 232.3962
sample estimates:
mean of x
 186.9019 > cor(firmat$MYYNTI, firmat$TUOTTO)
```

Sijoitusten arvojen 99%:n luottamusväliksi saadaan 141 – 232 miljardia dollaria.

Tutkitaan, eroavatko USA:ssa ja Yhdistyneen kuningaskunnassa kotimaataan pitävien yritysten sijoitukset toisistaan. Ensiksi kummankin maan sijoitustiedot kannattaa tallentaa uusiin muuttujiin testauksen helpottamiseksi. Tämän jälkeen tutkitaan, eroavatko sijoitusten varianssit toisistaan. Sen jälkeen tehdään itse testaus:

```
> usa = firmat$TUOTTO[firmat$MAA == "Yhdysvallat"]
> uk = firmat$TUOTTO[firmat$MAA == "Yhdistynyt kuningaskunta"]

> var.test(usa, uk)

      F test to compare two variances

data:  usa and uk
F = 0.7175, num df = 163, denom df = 33, p-value = 0.1835
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.4000103 1.1695033
sample estimates:
ratio of variances
         0.7175199

> t.test(usa, uk, var.equal = T)

      Two Sample t-test
```

```
data:  usa and uk
t = -0.5175, df = 196, p-value = 0.6054
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.295338  1.341076
sample estimates:
mean of x mean of y
 3.679634  4.156765
```

Testauksessa saatu luottamusväli sisältää nollan, joten USA:n ja Yhdistyneen kuningaskunnan yritysten sijoituksissa ei ole eroja.

10 Yhteenveto

Vaikka tästä harjoitustyöstä tulikin odotettua pidempi, on se silti vain todella pieni pintaraapaisu R:n maailmaan. Toiminnallisuudesta tuli käsiteltyä ainoastaan oleellisimmat osat ja komennoista nähtiin ainoastaan jäävuoren huippu. Tällä oppaalla päässee kuitenkin alkuun R:n käytössä, ja lisää tietoa löytyy tämän työn liitteessä sijaitsevista linkeistä.

Soveltuuko R sitten käytettäväksi Tampereen yliopiston Tilastotieteen johdantokurssille? Vastaus on kyllä ja ei. R vaatii paljon enemmän opettelemista kuin esimerkiksi SPSS, ja se ei ole intuitiivinen käyttää komentoriviin tottumattomille. Toisaalta jos tietokone- ja ohjelmistomaailma ei ole vieras ja hakusassa on tehokas ja moneen pystyvä (ilmainen) tilastollinen ohjelma, on R varmasti oiva valinta myös Tilastotieteen kursseille käytettäväksi. Soveltuvuus ei siis ole aivan yksiselitteinen asia, vaan se riippuu paljon yksittäisestä opiskelijasta ja hänen taustoistaan ja kiinnostuksistaan. Katsoisin R:n olevan kuitenkin vahva vaihtoehto mihin tahansa tilastolliseen laskentaan.

Liitteet

R: opas ekologeille:
`http://cc.oulu.fi/~jarioksa/opetus/rekola/`

Helsingin yliopiston kurssimateriaalia – johdatus R-ohjelmiston käyttöön:
`http://www.rni.helsinki.fi/~pek/r-johdatus/kmat.html`

Kuopion yliopiston tietotekniikkakeskuksen oppaat – R-opas:
`http://www.uku.fi/tike/oppaat.shtml`