

# Data-analyysi R-ohjelmistolla

Tommi Härkänen

Terveyden ja hyvinvoinnin laitos (THL), Helsinki  
E-mail: [tommi.harkanen@helsinki.fi](mailto:tommi.harkanen@helsinki.fi)

Helsingin yliopisto, 11.-12.3.2014

## R-ohjelmisto: yleistä

R-ohjelmiston ominaisuudet sisältävät aineiston muokkaamisen, tilastollisen laskennan ja tulosten (graafisen) esittämisen:

- ▶ aineisto tallennetaan yleensä matriisi- tai listamuodossa
- ▶ aineistoa käsitellään erilaisten funktioiden avulla
- ▶ tulokset voidaan esittää monipuolisesti, esim. graafisesti ja tallentaa kuvatiedostoiksi
- ▶ tulokset voidaan viedä erilaisissa muodoissa käytettäväksi muissa ohjelmissa
- ▶ laajentaa R-ohjelmiston toiminnallisuutta ohjelmoimalla uusia funktioita

## Sisältö

Kurssiin liittyvät käytännön asiat

Lyhyt johdanto R-ohjelmistoon

## R-ohjelmisto: Tällä kurssilla aiheina

- ▶ johdatus R-ohjelmiston käyttöön ja S-kieleen
- ▶ aineiston syöttäminen, muokkaaminen ja tallentaminen
- ▶ kuvaileva tilastotiede, tilastolliset analyysit
- ▶ tilastografiikka
- ▶ tilastollisen päättelyn esimerkkejä R-ohjelmiston avulla
  - ▶ uskottavuusfunktio,
  - ▶ piste-estimointi,
  - ▶ luottamusvälit ja -joukot,
  - ▶ testiteoriaa

## Luennot, harjoitustehtävät ja harjoitustyö

- Tietokonealuokkaharjoitus (C128) tiistai 11.3. kello 14-16 ja 16-18;  
keskiviikko 12.3. kello 14-16 ja 16-18
- Luennot** (CK122) 25.3.-16.4. tiistaisin kello 16-18 ja keskiviikkoisin 14-16
- Moodle** Linkki Moodle-sivulle, josta löytyy kurssimateriaalia, löytyy kurssin wiki-sivulta<sup>1</sup>.
- Harjoitukset** (C128) Tehtävänannot löytyvät Moodlesta. Ensimmäisistä harjoitustehtävistä (palautettava 26.3. mennessä) saa pisteitä, jotka huomioidaan loppuarvosanassa.
- Harjoitustyö** Viimeisellä luentokerralla 16.4. esitellään tehtävänanto, ja materiaali löytyy moodlesta. Harjoitustyö palautetaan moodlen kautta. Harjoitustyö vaikuttaa eniten kurssiarvosanaan.
- Arvostelu** numeerinen, 0 – 5.

<sup>1</sup><http://wiki.helsinki.fi/pages/viewpage.action?pageId=113247055>

## Verkkosivustoja

- ▶ Pääsivusto: [r-project.org](http://r-project.org).
- ▶ Wikipedia: [http://en.wikipedia.org/wiki/R\\_\(programming\\_language\)](http://en.wikipedia.org/wiki/R_(programming_language)).
- ▶ R-aiheisia blogikirjoituksia löytyy <http://www.r-bloggers.com/> monista aiheista esim. opastukset R:n ominaisuuksiin, uudet ominaisuudet, uudet/päivitetyt lisäpaketit.
- ▶ Esimerkkejä R:n avulla tehdyistä kuvista: <http://gallery.r-enthusiasts.com/>.
- ▶ Blogi suomeksi: <http://www.r-ohjelmointi.org/>.
- ▶ Parannettu käyttöliittymä on **RStudio**: <http://www.rstudio.com/>. Tutustumisen arvoinen(!), mutta ei käsitellä tällä kurssilla.

## Historia

- ▶ S kieli
  - Old S** S on yksi tilastolaskentaan kehitetyistä kielistä (Bell Laboratories), ensimmäinen versio 1975-1976.
  - New S** Vuonna 1991, *Statistical Models in S* (White Book), jossa esiteltiin mm. datakehikko-olio, julkaistiin.
- ▶ Kaupallinen ohjelmisto S-plus ilmestyi 1988.
- ▶ Avoimen lähdekoodin R-ohjelmisto
  - Versio 0.16** 1.4.1997.
  - Versio 1.0.0** Tuotantokäyttöön soveltuva versio 29.2.2000.

## Dokumentaatio

- ▶ Manuaalisivut:
  - ▶ pdf- and html-muodossa: katso Help-valikosta, tai
  - ▶ komentoriviltä, esim. `?print`
- ▶ R-projektin sivusto <http://www.r-project.org/>:
  - ▶ Manuaalit
  - ▶ Postituslistat
  - ▶ Linkki CRAN-sivustolle
- ▶ CRAN: The Comprehensive R Archive Network
  - Download** R-ohjelmiston lataus Windows-, Mac OS X- ja Linux-käyttöjärjestelmille
  - Extension packages** Lisäpaketit, joiden avulla voi laajentaa tai parantaa R:n toiminnallisuutta (tällä hetkellä 5,274 pakettia)<sup>2</sup>.
  - CRAN Task Views** Eri aihealueiden mukaan esitelty tärkeimpiä lisäpaketteja. Tällä hetkellä aihealueita on 33.

<sup>2</sup>Asentaminen: Valikon *Packages / Install Packages* kautta tai funktiolla `install.package()`.