

R: keel ja keskkond statistilisteks arvutusteks

Sissejuhatus

Krista Fischer

Teemad

- Alustamine: tutvumine programmiga
- Nõuded andmefailile, andmete ja tunnuste tüübid
- Andmete lugemine R-i: `read.table`, `read.csv`
- Andmestikud ja tunnused R-is: `data.frame`, `matrix`, pidevad tunnused, faktortunnused, stringid (`head`, `str`, `summary`)
- Lihtne ülevaade andmetest: `summary`, `mean`, `sd`, `hist`, `plot`
- Tabelid, grupeeritud statistikud: `table`, `stat.table`, `barplot(table())`
- Graafikute kujundamine: `plot()`, `lines()`, `col`

Mis on R?

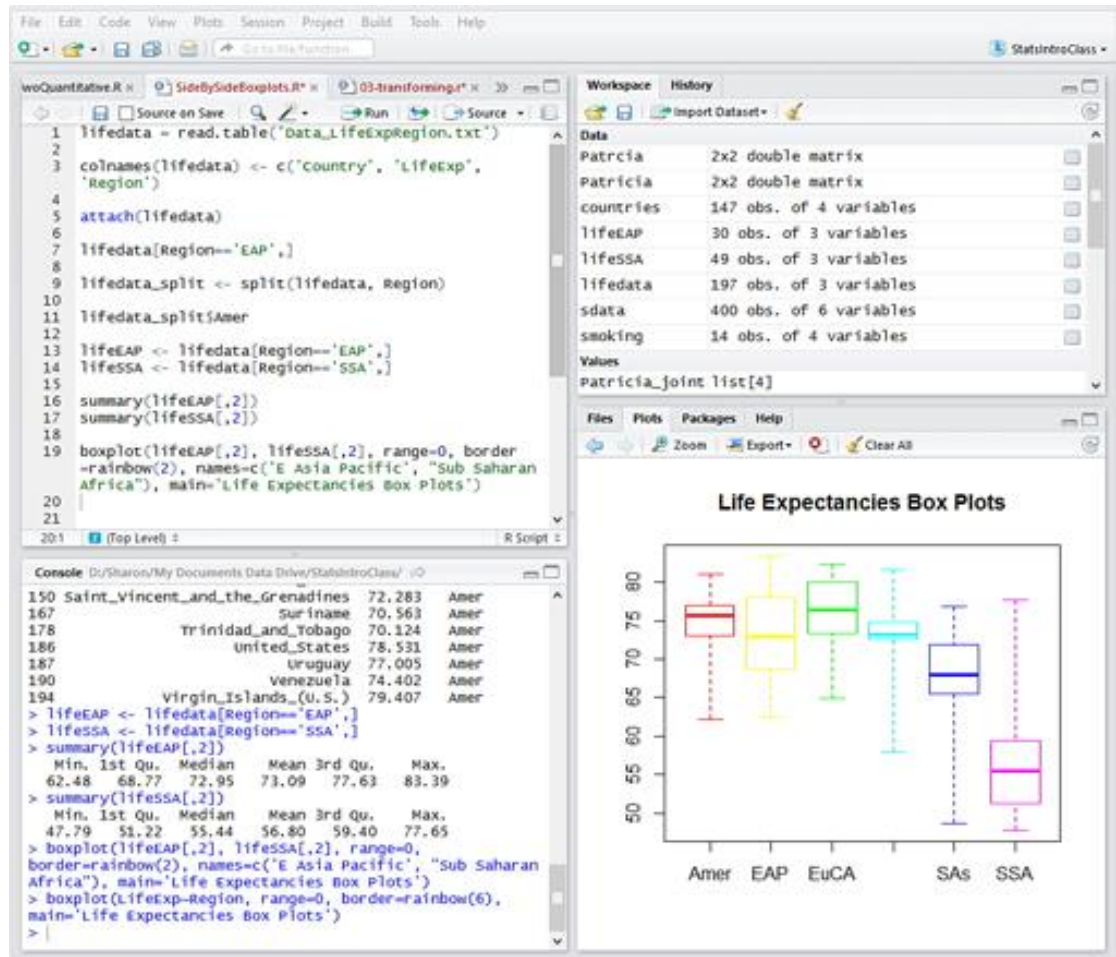
www.r-project.org

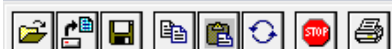
- R on keel ja keskkond statistilisteks arvutusteks
- R on vabavara (<http://cran.r-project.org>), mille põhiosa arendamisega tegeleb mitukümmend teadlast (statistikut) üle maailma – “The R core team”
- Lisaks saab igaüks kirjutada R-ile lisapakette – hetkel on neid kõigile kättesaadavaks tehtud >13000
- R töötab “command line” põhimõttel – R-i poole pöördumine toimub käsureale koodi trükkimise kaudu (võib kasutada ka skriptiakent, kuhu kirjutatud käsked saab ridade või lõikude haaval käivitada)
- R on väga paindlik ja kohandatav

RStudio

- Populaarseim R-i kasutajakeskkond
- Aitab organiseerida tööd konkreetsete projektidena
- Integreeritud R-markdown keel võimaldab koostada dünaamilisi dokumente

www.rstudio.com





```
R version 2.15.0 (2012-03-30)
Copyright (C) 2012 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: i386-pc-mingw32/i386 (32-bit)
```

```
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
Type 'contributors()'
'citation()' on how
```

```
Type 'demo()' for some
'help.start()' for an
Type 'q()' to quit R.
```

```
> 23+45
[1] 68
> (23+45)*6
[1] 408
> 285 + (23+45)*6
[1] 693
> |
```

R on suurepärase
kalkulaator: arvutustehteid
saab sisestada käsurealt,
“enter” suunab käsu
täitmisele

Ülemise nooleklahvi abil saab
eelnevalt sisestatud käske
käsureale kopeerida, kus neid siis
vajadusel modifitseerida ja uuesti
käivitada saab

Kasulikke algteadmisi

- UNIXis käivitub R käsurealt 'R' sisestamisel ja R-ist väljuda saab käsu `q()` abil.

`q("yes")` salvestab töökeskkonna, `q("no")` väljub salvestamata.

- R-i käsurealt võib veel sisestada:
 - Arvutustehteid
 - Pöördumisi R-i funktsioonide poole
 - Objektide nimetusi
 - Ülaloodud elementide erinevaid kombinatsioone
- `help(funktsioon)` või `?funktsioon` avavad funktsiooni abi lehekülje

Näide: sisestame 5 arvu ja arvutame nende keskmise

Omistamine: `<-` võib asendada ka võrdusmärgiga `=`

```
> x <- c(0, 2, 5, 6, 3)
```

```
> x
```

```
[1] 0 2 5 6 3
```

```
> mean(x)
```

Funktsiooni *argument* (või *argumendid*) alati sulgudes

```
[1] 3.2
```

```
> m <- mean(x)
```

```
> m
```

```
[1] 3.2
```

```
> x-m
```

Arv m lahutatakse vektori x igast elemendist

```
[1] -3.2 -1.2 1.8 2.8 -0.2
```

```
> ls()
```

```
[1] "m" "x"
```

Töökeskkonnas salvestatud objektide nimekiri

Tehted kahe vektoriga (tunnusega)

```
> x <- c(0, 2, 5, 6, 3)
```

```
> y <- c(1, 1, 1, 3, 3)
```

```
> x-y
```

```
[1] -1  1  4  3  0
```

Kui tehakse tehteid sama pikkade vektoritega, siis need tehted sooritatakse elementhaaval

Nõuded andmefailile (et seda saaks R-i abil analüüsida)

Kood	Vanus liitumisel	Nohu või köha?
mehed		
A01234	34	Nohu
A09461	21,5	Nohu, köha
B98243	54	Ei ole
Naised		
C00001	23.5	Köha
C43278	45	Köha, nohu
C09375	21a, 6 kuud	Ei tea

või

Kood	vanus	sugu	nohu	köha
A01234	34	1	1	0
A09461	21.5	1	1	1
B98243	54	1	0	0
C00001	23.5	2	0	1
C43278	45	2	1	1
C09375	21.5	2	NA	NA



Nõuded andmefailile

(et seda saaks R-i abil analüüsida)

Andmefail peab olema organiseeritud maatriksi (2-mõõtmelise tabeli) kujul, kus:

- Iga rida vastab ühele *indiviidile* (objektile).
Erandjuhtudel võib ühele indiviidile vastata mitu rida, enamasti vastab ühele indiviidile üks rida
- Iga veerg vastab ühele *tunnusele* (objekti omadusele)
- Esimeses reas (ainult ühes reas!) võivad olla tunnuste nimed
- Ühes lahtris võib olla ainult 1 number või 1 *string* (tekstiline fraas)
- Samas veerus olevate väärtuste korral on kasutatud sama formaati
- Puuduvad väärtused: eraldi tähistuse (NA) kasutamine on soovitatav

Andmete sisestamine R-i

1. Salvesta (nt Excelis) andmefail tekstikujul: (.csv või .txt)

(Menüüst: *Save as....Other formats*;

Sealt edasi: *Save as type....CSV (Comma delimited)*)

2. R-is kasuta `read.table()` käsku andmete sisselugemiseks

Andmete sisselugemine R-i

Loodava
andmestiku
nimi

failinimi

Mis eraldab väljasid?
(“;” - eesti settingutega arvuti,
“,” - USA settingutega arvuti,
“\t” – tab, “ ” - space)

```
andmed<- read.table("valim.csv",header=T,sep=";",dec=","")
```

Kas tunnuste nimed on
1. reas (T: jah, F: ei)

Mis eraldab komakohti?
(", " - eesti settingutega arvuti,
"." - USA settingutega arvuti)

Vt ka `help(read.table)`,
`read.csv`, `read.csv2`
(muud võimalused, sh puuduvate
väärtuste tähistamine)

Andmete sisselugemine R-i – kuidas anda ette andmefaili asukoht

```
andmed<- read.csv("valim.csv")  
# töötab näitefailiga, kui fail on Ri töökataloogis  
andmed<- read.csv("C:/Users/Krista/Documents/Stat1/valim.csv")  
# kui ette anda õige kataloog  
andmed<- read.csv(file.choose())  
# otsi fail ise üles!  
  
andmed<-  
read.csv("https://courses.ms.ut.ee/2018/sissejuhmat/fall/uploads/Main/valim.csv")  
# töötab ka URL-iga (kui see on õige)
```

NB! „Teistpidine“ kaldkriips

Vt ka `help(read.table)`, `read.csv`, `read.csv2`
(muud võimalused, sh puuduvate väärtuste tähistamine)

Kas andmed said õigesti sisse loetud?

Kontrolliks võib kasutada:

`dim (andmed)` mitu rida, mitu veergu

`str (andmed)` objekti struktuur

`head (andmed)` esimesed 5 rida

`andmed [1:3, 1:3]` esimesed 3 rida, 3 veergu

Töö R-is: üldisi nõuandeid

- Vali sobiv töökataloog ([Rstudio](#) abil lihtsam)

File – Change dir või `setwd("C:/kataloog/alamkataloog")`
(töökataloogi nägemiseks kasuta `getwd()`)

- Kasuta skripti-editori (File – New script)
- Lisa skripti kommentaare sümboli `#` järele

- Salvesta andmed faili:

```
save(andmed, file="andmed.RData")
```

- Järgmisel korral kasuta kas

```
load("andmed.RData")
```

File – load Workspace ...andmed.RData

Andmestikud ja tunnused R-is

Andmete formaat R-is on `data.frame`

```
is.data.frame(andmed) # kas on data.frame?  
andmed <- as.data.frame(andmed)  
# muudame data.frame –ks (kui ei ole)
```

andmed\$**tunnus1**

tunnus "tunnus1" andmestikud andmed

Ülevaade andmetest

```
summary (andmed)
```

```
# ülevaade kõigist tunnustest andmestikus
```

```
summary (andmed$pikkus)
```

```
# ülevaade tunnusest 'pikkus'
```

```
mean (andmed$pikkus) # keskmine 'pikkus'
```

```
mean (andmed$pikkus, na.rm=T)
```

```
# keskmine pikkus pärast puuduvate väärtuste  
eemaldamist
```

```
sd (andmed$pikkus) # standardhälve
```

Ülevaade andmetest 2

```
mean (andmed$ pikkus) # keskmine 'pikkus'  
with (andmed, mean (pikkus))  
# samaväärne eelmisega
```

```
hist (andmed$ pikkus)  
# joonistab histogrammi
```

```
table (andmed$ maiust) # sagedustabel  
barplot (table (andmed$ maiust))  
# tulpdiaagramm sagedustabeli põhjal  
with (andmed, table (maiust, sugu))
```

Ülevaade andmetest 3: veel graafikuid

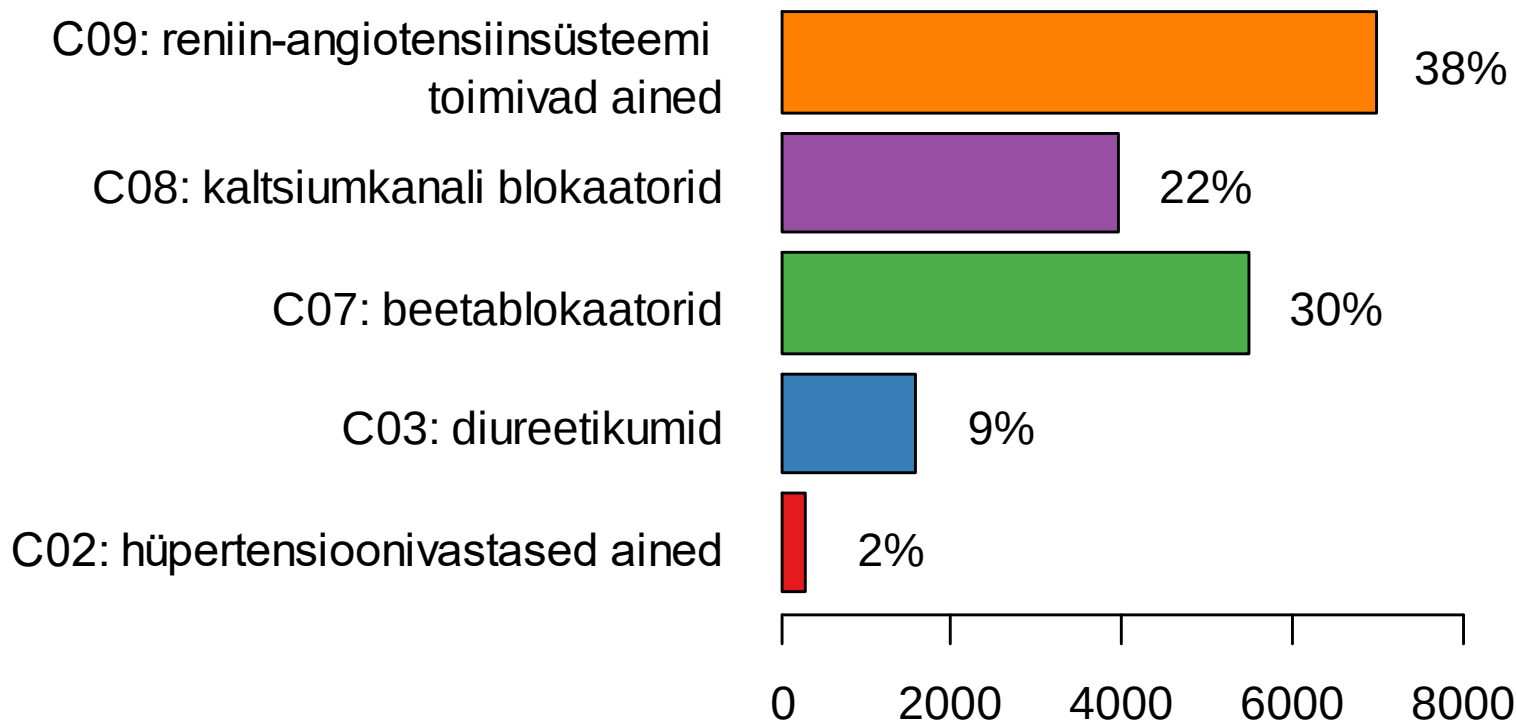
```
andmed$maiust <- factor(andmed$maiust,  
labels=c("ei söö", "1-2 korda", "3-5 korda", "6-7 korda"))  
  
barplot(table(andmed$maiust))  
  
barplot(table(andmed$maiust), horiz=T)  
  # tulbad teistpidi  
barplot(table(andmed$maiust), col=1:4) # värvid  
barplot(table(andmed$maiust), horiz=T, col=rainbow(4))  
  
with(andmed, plot(pikkus,kaal))  
with(andmed, plot(pikkus,kaal,col=sugu+1))
```

Uute tunnuste moodustamine

```
andmed$cpikkus <- with(andmed, pikkus -  
  mean(pikkus))  
  
  # tsentreeritud pikkus (erinevused keskmisest)  
andmed$zpikkus <- scale(andmed$pikkus)  
  # skaleeritud pikkus (erinevused keskmisest jagatud  
  standardhälbega) ehk pikkuse Z-skoorid  
  
andmed$bmi <- with(andmed, 10000*kaal/(pikkus^2))  
  
hist(andmed$bmi)  
with(andmed, boxplot(bmi~sugu))  
with(andmed, boxplot(bmi~maiust))
```

Graafikute näiteid

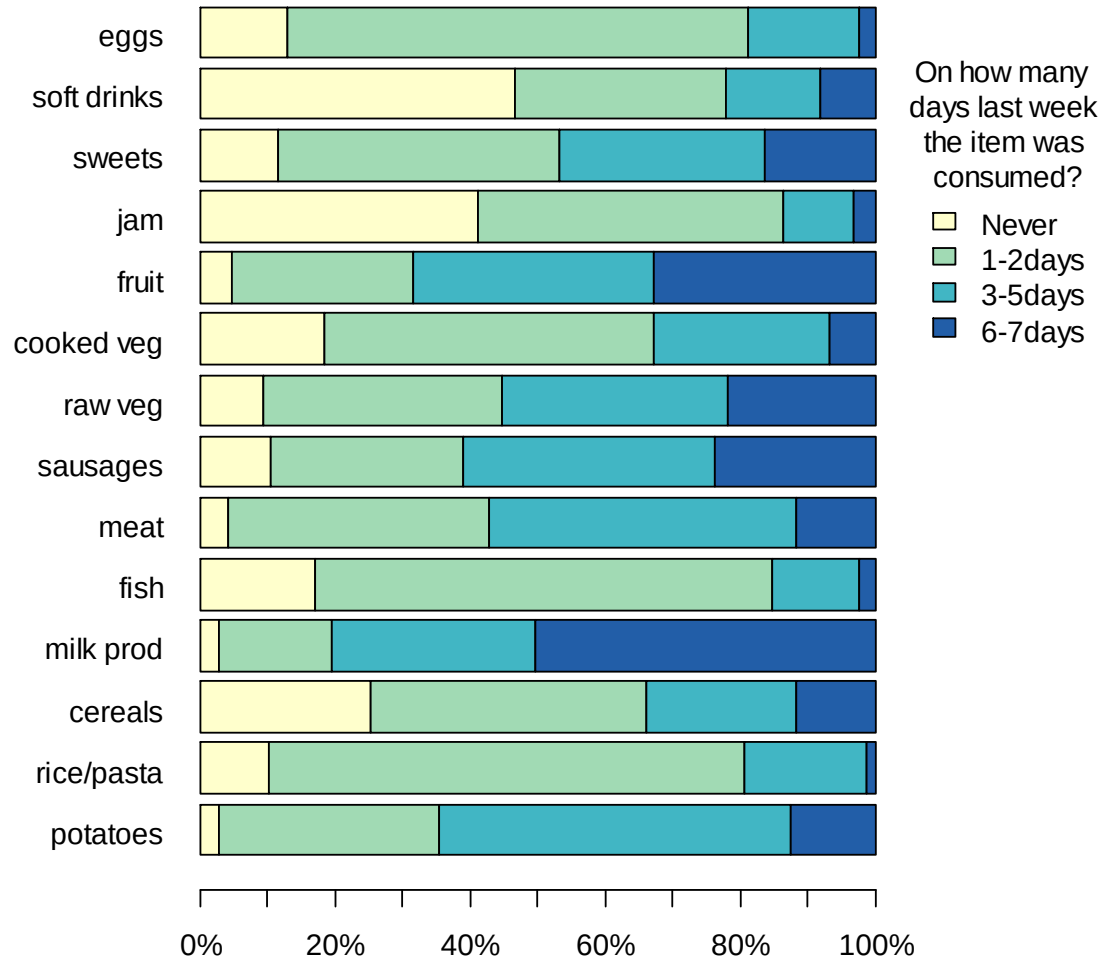
Hüpertensiooniravimite kasutamine TÜ Eesti Geenivaramu kohordis



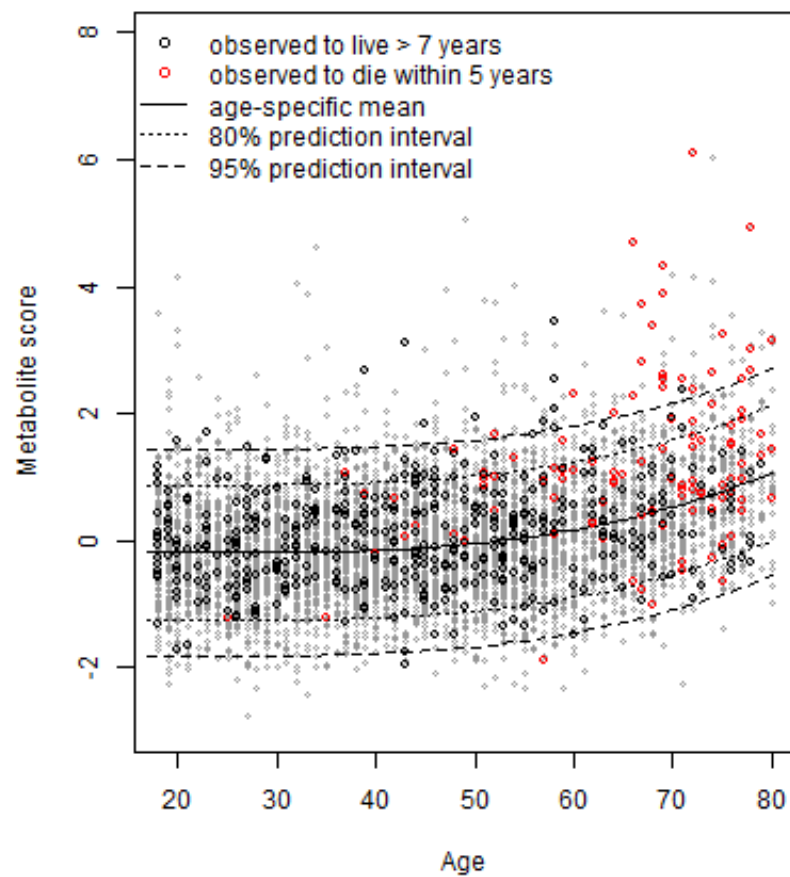
Kokku kasutab vähemalt ühte hüpertensiooniravimit 11820 geenidoonorit (23% kohordist)

Nutrition data

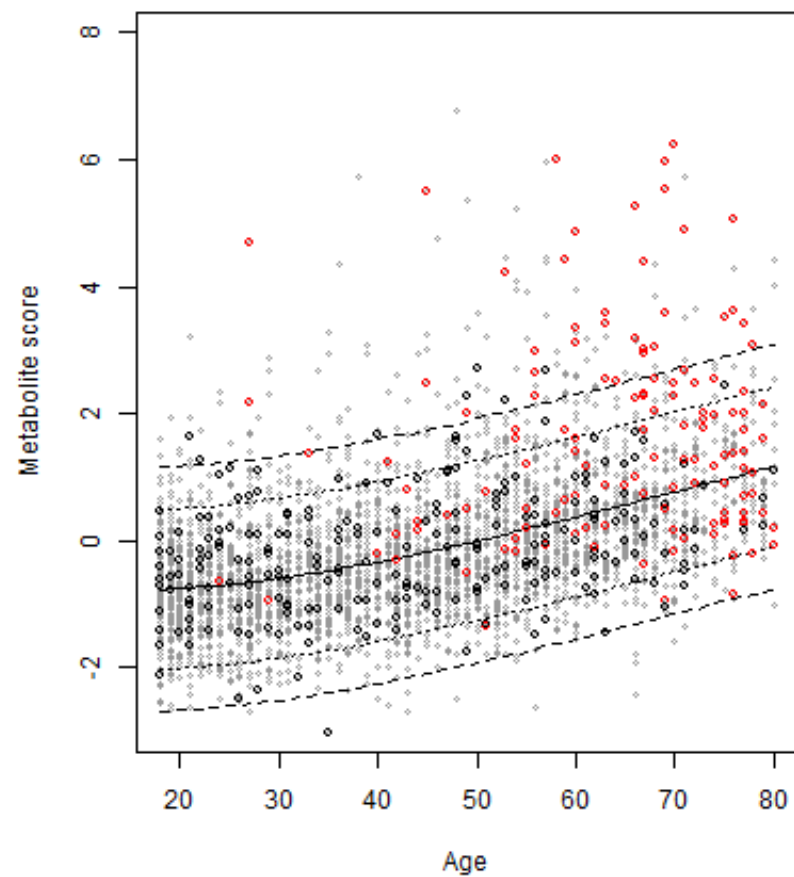
EGCUT food frequency data
N=50359



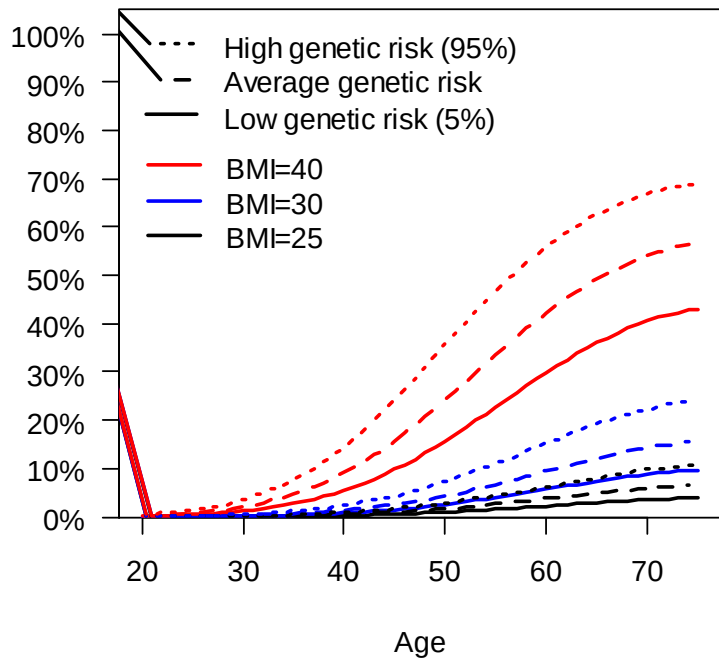
A Females



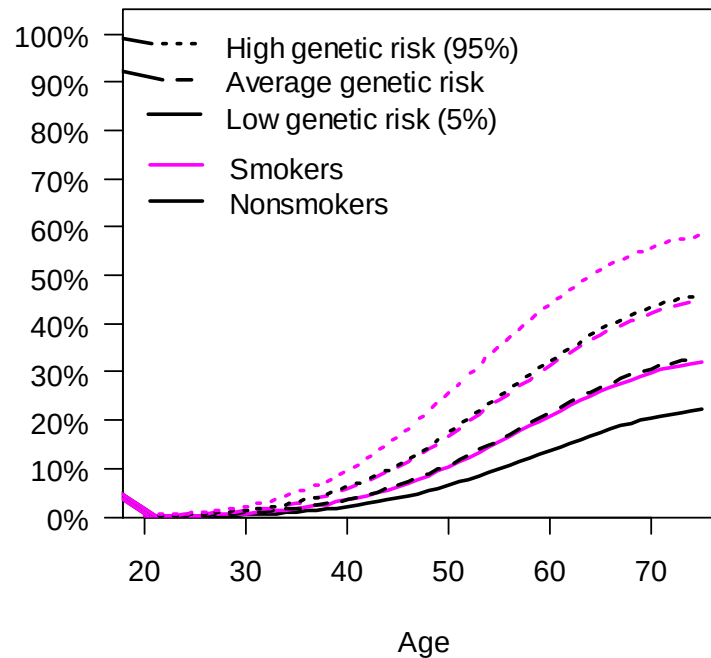
B Males



**T2D prevalence in nonsmoking males:
effect of BMI and genetic risk score**



**T2D prevalence in obese males (BMI=35):
effect of smoking and genetic risk**



- Vaata ka:

demo(graphics) - R-i keskkonnas sees

<http://www.ms.ut.ee/mart/R/Rgraafika.html>

(Autor: Märt Möls)

Kust saada abi/õpetust?

- Sissejuhatus R-i: Introduction to R
(pdf-fail kaasas standardinstallatsiooniga)
- www.r-project.org
- www.r-bloggers.com
- Datacamp – www.datacamp.com
- www.rseek.org
- Kursus „Rakendustarkvara R“

DataCamp – interaktiivne õppekeskkond, kus on ka tasuta sissejuhatavaid Ri kursuseid

 DataCamp

EXERCISE

Naming a vector (2)

If you want to become a good statistician, you have to become lazy. (If you are already lazy, chances are high you are one of those exceptional, natural-born statistical talents.)

In the previous exercises you probably experienced that it is boring and frustrating to type and retype information such as the days of the week. However, when you look at it from a higher perspective, there is a more efficient way to do this, namely, to assign the days of the week vector to a **variable**!

Just like you did with your poker and roulette returns, you can also create a variable that contains the days of the week. This way you can use and re-use it.

INSTRUCTIONS 100 XP

- A variable `days_vector` that contains the days of the week has already been created for you.
- Use `days_vector` to set the names of `poker_vector` and `roulette_vector`.

 Take Hint (-30 XP)

SCRIPT.R

```
1 # Poker winnings from Monday to Friday
2 poker_vector <- c(140, -50, 20, -120, 240)
3
4 # Roulette winnings from Monday to Friday
5 roulette_vector <- c(-24, -50, 100, -350, 10)
6
7 # The variable days_vector
8 days_vector <- c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday")
9
10 # Assign the names of the day to roulette_vector and poker_vector
11 names(poker_vector) <-
12 names(roulette_vector) <-
```

Run Code Submit Answer

R CONSOLE

> |