Biomeetria
2. praktikum

Ülesanne 0. Vaata lõpuni eelmise praktikumi material (sirvi läbi/katseta näiteid peatükist „Graafikast“).

Loeme sisse andmestiku fishcatch.dat:

andmed=read.table(“http://www.ms.ut.ee/mart/biomeetria2007/fishcatch.dat“,
header=TRUE)

Andmestiku lühikirjeldus:
Soomes Tampere lähedal asuvast Laenelmavesi järvest püüti 159 kala. Selgus, et püütud kalad on pärit 7 liigist. Mõõdetud tunnuste kirjeldused:
1. *Species* on kodeeritud tunnus kalaliikidest:
    1 latikas
    2 siig
    3 särg
    4 linask
    5 tint
    6 haug
    7 ahven
2. *Weight* on kala kaal grammides
3. *Length1* on kala pikkus ninast saba alguseni sentimeetrites.
4. *Length2* on kala pikkus ninast saba keskosani sentimeetrites
5. *Length3* on kala pikkus ninast saba tipuni sentimeetrites.
6. *Height* on maksimaalne kõrgus, mis antud protsendina *length3*-st.
7. *Width* on maksimaalne paksus, mis on samuti antud protsendina *length3*-st.
8. *Sex* on kala sugu, kus 1=isane ja 0=emane.

Ülesanne 1.
Vaata esimese 10 kala andmeid. Seda saab teha käsuga: .....

Ülesanne 2.
Andmestikust saab ühte tunnust, näiteks kaalu (*Weight*) vaatamiseks välja võtta käsuga *andmed$Weight*. Ühe ja sama andmestikuga töötades muutub andmestiku nime igale poole ettekirjutamine tülikaks. Milline käsk tuleb anda, et saaksime andmestiku *andmed* tunnuste poole pöörduda ilma, et peaksime igale poole andmestiku nime kirjutama? Anna see käsk arvutile!
    Kirjuta kasutatud käsk ka siia: .....

Ülesanne 3.
Kuidas saaksime leida isaste kalade kõrguseid sentimeetrites?

    Selleks sobib järgmine käsk:.....

Piilume kiiresti paari tunnuse jaotust.

Milline on tunnuse Species jaotus?

Alustame sagedustabelist:

table(Species)

üsna kole. Ilusad emakeelsed kalanimed oleksid ilusamad vaadata.

liik=factor(Species, labels=c("latikas","siig","särg","linask","tint","haug","ahven"))
table(liik)

See on sagedustabel – kalade arvud. Kuidas saada jaotuse hinnangut (osakaalud)?

prop.table(table(liik))

Ülesanne 4
Aga mida tuleb teha, et jaotust protsentides (%) saada? Tuleb lihtsalt anda järgmine käsk:

.....

Kasuta round-käsku selleks, et ümmardada saadud tulemused protsendi täpsuseni!

.....

NB! Kui sooviksid ümmardada protsendikümnendike täpsuseni, siis peaksid kasutama käsku round(<vektor või arv, mida soovid ümmardada>, 1).

Tabel on hea. Tark inimene põhjalikku infot otsides eelistab ikka tabelit. Aga mida pakkuda Sinu artiklit puhkehetkel (peaaegu uinunud mõistusega) lehitsevale inimesele? Vajame pilte!

barplot(prop.table(table(liik)))

või, veidi kirevam:

barplot(prop.table(table(liik)), col=rainbow(7))

või

barplot(prop.table(table(liik)), col=terrain.colors(7), main="kalad Laenelmavesi järves", xlab="Kalaliik", ylab="osakaal")

või, sama asi kaapekakukesena:

pie(table(liik), col=terrain.colors(9))

või, alternatiivina:

pie(table(liik), col=terrain.colors(9), labels=paste(names(table(liik)), " (", table(liik), ")", sep=""))

Piilume ka mõnda teist tunnust. Kui palju üks järvest väljaõngitsetud kala kah kaalub?

hist(Weight)
mean(Weight); median(Weight)

Miks tuleb keskmine kaal palju suurem kui kaalude mediaan?

Vaatame ahvenate kaalude jaotust (histogrammi):

hist(Weight[liik=="ahven"])

Miks on tegemist bimodaalse jaotusega? Mis põhjusel võiks selline „kahe tipuga“ jaotus tekkida?

Uurime ka tunnuse Height (kala kõrgus protsendina pikkusest) jaotust liigiti:

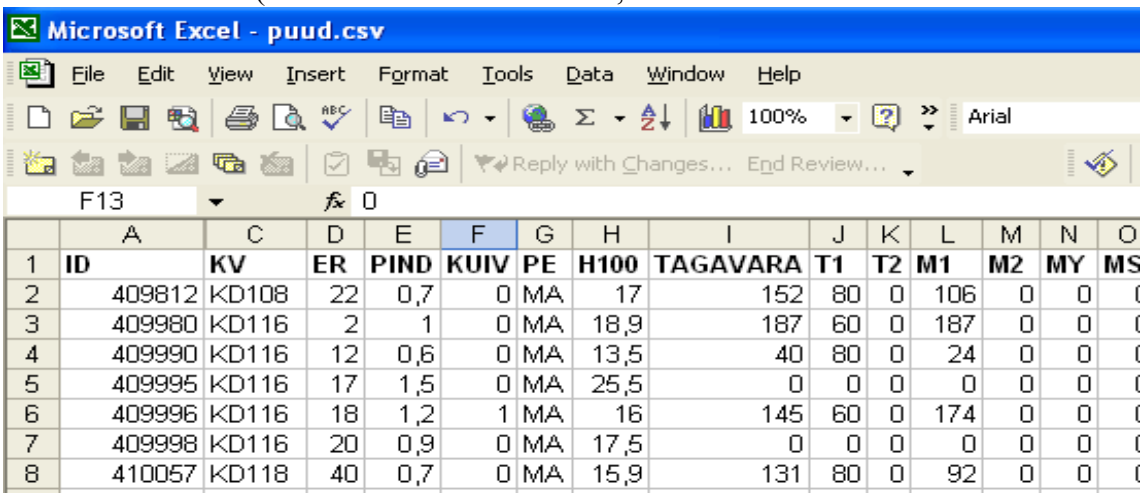
boxplot(Height~liik)

Kuidas antud joonist interpreteerida? Milliste kalade kõrgus tuli suur, millistel väike? Milliseid oletusi võiks teha selle joonise järgi?

Lisa 1. Andmete importimisest

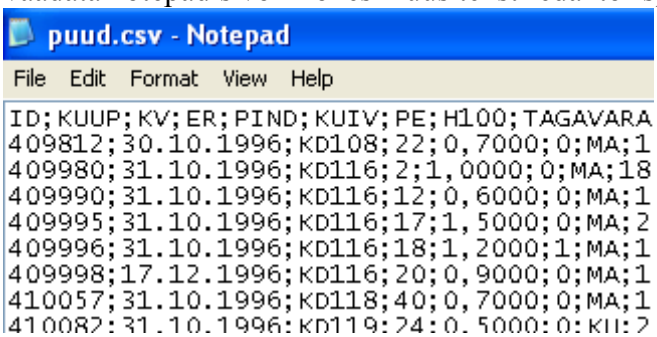
Suuremate andmestike sisestamiseks pole R just kõige sobivam vahend. Enamasti sisestataksegi andmeid kas spetsiaalselt andmete sisestamiseks mõeldud tarkvara abil või kasutades mõnda tabelarvutusprogrammi (näiteks Excelit). Andmestiku importimisel tabelarvutusprogrammist (Excelist) on soovitav andmefail esmalt salvestada CSV-formaadis (Comma Separated Values), näiteks faili "C:\puud.csv" (*File -> Save As ->* muuda *Save As type* aknas failitüüp „*CSV (Comma Delimited) (*.csv)*"-ks).

Andmestik Excelis (tunnuste nimed – lühikesed, soovitavalt ühesõnalised – esimeses reas)



	A	C	D	E	F	G	H	I	J	K	L	M	N	O
1	**ID**	**KV**	**ER**	**PIND**	**KUIV**	**PE**	**H100**	**TAGAVARA**	**T1**	**T2**	**M1**	**M2**	**MY**	**MS**
2	409812	KD108	22	0,7	0	MA	17	152	80	0	106	0	0	
3	409980	KD116	2	1	0	MA	18,9	187	60	0	187	0	0	
4	409990	KD116	12	0,6	0	MA	13,5	40	80	0	24	0	0	
5	409995	KD116	17	1,5	0	MA	25,5	0	0	0	0	0	0	
6	409996	KD116	18	1,2	1	MA	16	145	60	0	174	0	0	
7	409998	KD116	20	0,9	0	MA	17,5	0	0	0	0	0	0	
8	410057	KD118	40	0,7	0	MA	15,9	131	80	0	92	0	0	

Andmestik peale csv-faili salvestamist näeb välja järgmine (soovi korral saame csv-faili vaadata notepadis või mõnes muus tekstiredaktoris):



Salvestatud andmestiku lugemiseks R-i tuleb anda näiteks järgmine käsk:

```
puud=read.csv2("C:/puud.csv", header=T)
```

ja vaatamaks, kas andmete sisselugemine läks valutult:
head(puud)
names(puud)

Vahel tekivad probleemid – andmed ei loeta R-i õigel kujul, saame veateateid vm. ÜKs põhjus – andmefail sisaldas tekstikirjeid, mis sisaldasid keelatud märke (Näiteks sümbolit „;"). Vahel aga on põhjuseks see, et erinevad Excelid võivad (erinevates masinates) teha erinevaid csv-faile. Vahel pannakse andmeväljade vahele semikooloni (;) asemel näiteks koma (,) ja kümnendkohtade eraldaja arvus kasutatakse hoopis punkti (.) – seda näeme

vaadates tekkivat csv-faili notepadis/tekstiredaktoris. Sellisel juhul tuleb andmete sisselugemiseks R-i anda hoopis käsk

```
puud=read.csv2("C:/puud.csv", header=T, dec=".", sep=",")
```

Pange tähele:

- sisseloetud andmestik tuleb kuhugi salvestada, kui soovime teda hiljem ka kasutada! Siin salvestasime ta andmestikuks nimega "puud".
- Failinimes on kasutatud tagurpidi kaldkriipse („/")!
- Käsuga *sep=* määratakse sümbol, mis eristab eri tunnuste väärtuseid (Excel võib salvestamisel kasutada eraldajana nii sümbolit "," kui ka ";")
- *dec=* parameetri abil saab määratleda sümboli, mis tähistab arvus koma (Näiteks Excel kasutab kümnendkohtade tähistamiseks vahel sümbolit "." ja vahel sümbolit ",").
- Parameeter *header=T* ütleb arvutile, et tunnuste nimed on kirjas tekstifaili esimeses reas.

R-i andmestikku saab samuti salvestada tekstifaili kasutades käsku

```
write.table(andmestik, "C:/andmed/uustabel.csv",
     sep=",", dec=".", header=T, row.names=F)
```

Teine võimalus

Alternatiivne võimalus oleks installeerida lisamoodul *xlsReadWrite*, milles sisalduv funktsioon *read.xls* võimaldab ka otse Exceli faile lugeda. Lisamooduli lisamiseks (Vanemuise tänava arvutiklassis saab seda teha ainult administraator) valige R-is menüüst Packages->Install Packages -> <vali server> -> *xlsReadWrite*
Seejärel saate Exceli andmestiku sisse lugeda näiteks järgmise käsu abil:
andmed2=read.xls("E:/konsult/hunt/INSrynnakud.xls")

Märkused:

- Kasutades lisamoodulit foreign saab R-i (otse) lugeda ka statistikapakettide S-Plus, Stata, Minitab ja SPSS andmefaile.
- Täiendavat infot vaata R-i koduleheküljelt (Manuals->R Data Import/Export)