Biomeetria
2. praktikum

Harjumine R'iga jätkub. Uuritava tunnuse jaotuse kirjeldamine.

Esmatutvus andmetega algab enamasti uuritavate tunnuste jaotuste kirjeldamisega. Ühtlasi võimaldab uuritava tunnuse jaotuse vaatamine märgata näiteks sisestusvigu.

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/biomeetria2009/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.

Esmalt paar ülesannet eelmises praktikumis õpitu meenutamiseks.

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: .....

Nominaalse tunnuse jaotuse iseloomustamine

Kalade andmestikus on üks nominaalne tunnus – kala liik (*Species*). Leiame esmalt sagedustabeli:

```
table(Species)
```

Millist liiki kala esines kõige rohkem meie valimis? Sellele küsimusele vastamine osutub mõnevõrra ebamugavaks, sest tunnus *Species* on kodeeritud. Edaspidi soovime, et järgnevatel analüüsidel ja graafikutel näeksime uuritava tunnuse sõnalisi väärtuseid („ahven“, „lest“ jne), mitte tunnuse kodeeritud väärtuseid. Selleks loome uue faktortunnuse ja ütleme, milline väärtus mida tähendab.

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

Sagedustabelit kasutades võime joonistada ka mõne graafiku – näiteks tulpdiagrammi või kakukese:

```
pie(table(liik))
```

Ülesanne.
Enamasti on tulpdiagramm parem vahend nominaalse tunnuse jaotuse kirjeldamiseks (tulpdiagrammilt on kergem algset informatsiooni välja lugeda). Joonista tulpdiagramm tunnusele liik, kusjuures tee nii, et *y*-teljel poleks mitte sagedused vaid iga liigi protsentuaalne osakaal. Vajadusel vaata „Jaotuse kirjeldamine“-nimelist abimateriali. Soovitud graafiku saad järgmise käsu abil:

.....

Ülesanne.
Tee tunnuse *Sex* – kalade sugu – jaotust iseloomustav graafik või tabel. Kas mõlemast soost kalu on ligikaudu samapalju? Kui ei, siis miks (paku välja näiteks kolm võimalikku põhjendust)? Osaliselt võib segadus tunnusega-sugu olla tingitud ka puuduvatest väärtustest (miks?). Sestap soovime kindlasti näha sagedustabelis/graafikul ka puuduvaid väärtuseid:

```
table(Sex, exclude=NULL)
```

Tunnus *Sex* on kodeeritud, väärtustega 0/1. Loo uus (faktor)tunnus *sugu*, mille puhul sagedustabelites/graafikutel oleks nullide ja ühtede asemel „isased“ ja „emased“. Katseta loodud tunnust näiteks käsuga

```
pie(table(sugu))
```

NB! faktortunnuse loomiseks on kaks võimalust, sõltuvalt sellest, kas tahame puuduvaid väärtuseid näidata tabelites/graafikutel või mitte. Puuduvate väärtuste nähtavale toomiseks võime faktortunnuse sugu luua sellisel viisil:

```
sugu=factor(Sex, exclude=NULL, labels=c("emane","isane","PUUDU"))
```

Pideva tunnuse jaotus.

Heidame esmalt pilgu peale kalade pikkuste (tunnus Length3) jaotusele – kasutame selleks histogrammi:

```
hist(Length3)
```

või, korrektsemalt vormistatud tulemuse jaoks:

```
hist(Length3, ylab="Sagedus", xlab="Pikkus (cm)", main="Kalade pikkused",
       col="gold")
```

Saadud graafikul peaks olema näha mitu „tippu“ (kalade pikkuse jaotus on multimodaalne). Miks?

Vaata ka ahvenate pikkuste jaotust. Kas ahvenate pikkuste jaotus on unimodaalne (ühe tipuga)? Soovi korral põhjenda nähtud tulemust.

Vahel soovime pideva tunnuse jaotust küll graafiliselt iseloomustada, aga sedavõrd palju ruumi nagu nõuab histogramm me oma raportis/töös/artiklis antud tunnusele ka eraldada ei raatsi (näiteks siis, kui tahame võrrelda uuritava tunnuse jaotuste erinevust paljudes alamgruppides vms). Sellisel juhul võime kasutada jaotuse iseloomustamiseks karp-vurrud diagrammi (boxplot). vaatame näiteks kalade pikkuste jaotuseid liigiti:

```
boxplot(Length3~liik)
```

Loe graafiku pealt välja, milline on näiteks latikate pikkuste mediaan, anna ka vahemik, kuhu sattub 50% latikate pikkustest!

Histogramm on ilus küll, aga vahel soovime ka pideva tunnuse puhul esitada andmeid sagedustabeli abil. Selleks peame uuritava tunnuse enne „lõikama“ paraja pikkusega juppideks (kasutades cut-käsku):

```
pikkusklass=cut(Length3, breaks=10)
table(pikkusklass)
```

Tulemus on üsna kole, kas pole? Lahenduseks oleks, kui annaksime ise ette klassipiirid. R võimaldab klassipiirideks ette anda (suvalist) vektorit, näiteks võime kasutada breaks-käsu taga sellist vektorit:

```
seq(0,70,10)
```

Tulemus on palju parem, proovi ise järgi!

```
pikkuskl=cut(Length3, breaks=seq(0, 70, 10))
table(pikkuskl)
```

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?

Vahel võime eelistada uuritava tunnuse jaotuse kirjeldamiseks kasutada jaotusfunktsiooni graafikut. Joonistamegi kalade kaalude empiirilise jaotusfunktsiooni graafiku:

```
plot(ecdf(Weight))
```

või, ilusamini vormistatud pildi saamiseks:

```
plot(ecdf(Weight), do.points=F, verticals=T,
  xlab="Kaal (g)", main="Kala kaalu jaotusfunktsioon")
```

Vaata leitud graafikut ja ütle, millise tõenäosusega kaalub esimene kala, mille me Laenelmavesi järvest välja püüame, üle 1kg?

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?

Pidevate tunnuste puhul tahame sageli teada, kas uuritava tunnuse jaotus võiks olla normaaljaotus. Selle küsimuse juurde pöördume veel tagasi, aga alustuseks proovi joonistada tunnuse *Height* histogrammi ning lisada graafikule kõige paremini sobiva normaaljaotuse jaotusfunktsiooni. Vaata lisamaterialis „Jaotuste kirjeldamine“ toodud näidet ja kohanda programmi kasutamaks tunnuse *Height* jaoks. Tee sama läbi ka ahvenate kõrgustega. Kas sinu arvates võiks tunnuse *Height* jaotus (kõigi kalade jaoks/ahvenate jaoks) olla normaaljaotus?

Soovi korral proovi ka järgmist:
Uuritud kalasid oli kokku 159, uuritud kalade kõrguste keskmise ja dispersiooni saad kätte käsuga

```
mean(Height); sd(Height)
```

Me võime R-l võtta valimi suurusega 159 sellisest normaaljaotusega populatsioonist, mille keskväärtus ja dispersioon on samasugused kui meie valimi keskmine ja dispersioon:

```
valim=rnorm(159, mean(Height), sd(Height))
```

Joonistame selle garanteeritult normaaljaotusega populatsioonist võetud valimi jaoks histogrammi. Kas saadud histogramm on rohkem „kellukesekujuline“ kui meie valimi põhjal leitud histogamm? Soovi korral korda protseduuri, võta uus valim normaaljaotusega populatsioonist ja vaata jälle, millise histogrammi said.

Proovi samal viisil uurida ka ahvenate kõrguste jaotust. Pea vaid silmas, et ahvenaid pole 159, vaid vähem!