

Biomeetria
2. praktikum

Uuritava tunnuse jaotuse kirjeldamine

Tänases praktikumis üritame uurimise alla võtta ühe uuritava tunnuse korraga ja proovime kirjeldada, millised on uuritava tunnuse väärtused.

Harjutamiseks kasutame juba eelmisest praktikumist tuttavat tudengite küsitlusuuringu andmeid. Loeme vastava andmestiku ka R'i:

```
load(url("http://www.ms.ut.ee/mart/biomeetria2012/andmefail.RData"))
```

Oleme sisse lugenud andmestiku kokku. Selleks, et me ei peaks tunnuste väärtuste nägemiseks (näiteks tudengite soo vaatamiseks) alati andmestiku nime ette kirjutama (kokku\$sugu), peame andmestikku kasutusele võtma, attachima:

```
attach(kokku)
```

Nüüd on algandmed kergelt kättesaadavad ja võime tööga alustada.

Nominaalne / järjestus- / (väheste väärtustega) diskreetne tunnus

Nominaalne tunnus on tunnus, mille väärtuseid pole võimalik mõistlikul viisil järjestada – üks võimalikest järjestustest pole iseenesest parem või halvem kui mingi teine järjestus. Antud andmestikus on nominaalseks tunnuseks (ehkki nähtavasti mitte parim võimalik nominaalse tunnuse näide) tunnus perekonnaseis, mis on kodeeritud järgmiselt:

- 1 – vallaline
- 2 – abielus
- 3 – vabaabielu
- 4 – lahutatud või elate lahus
- 5 – lesk

Sagedustabel ja jaotustabel

Esimese asjana vaatame, kui palju vallalisi, abielus jne tudengeid oli valimis

```
table(perekonnaseis)
```

Sagedustabel

1	2	3	4
579	17	61	4

Leitud sagedustabelist on näha, et päris palju on uuritud tudengite seas olnud vallalisi (579), natuke on olnud abielus tudengeid (17) jne. Sagedustabel kirjeldab eelkõige valimit. Me ei eeldagi, et üldkogumis (mis on enamasti palju-palju suurem) oleks ka vallalisi 579 tükki (või isegi ligilähedaselt 579). Ka kellegi teise teadlase erineva suurusega valim Tartu Ülikooli tudengitest ei anna tulemuseks ligilähedaselt samasuguseid numbreid.

Populatsiooni kirjeldamiseks (ja erineva suurusega valimite võrdlemiseks) on sobivam jaotustabel:

```
prop.table(table(perekonnaseis))
```

Jaotustabel

1	2	3	4
0.876	0.026	0.092	0.006

Jaotustabelist näeme, et vallaliste tudengite osakaal valimis on 0,876 (ehk 87,6%). Jaotustabel kirjeldab juba uuritava tunnuse jaotust populatsioonis, iseloomustab populatsiooni. Ka populatsioonis tervikuna võiks vallaliste osakaal olla ligikaudu 0,876.

Julgematele.

Proovi mõista, mida teevad järgmised programmikused:

Programm 1

```
a=table(perekonnaseis)
transform(a, protsent = round(prop.table(Freq)*100,1))
```

Programm 2

```
a=table(perekonnaseis)
data.frame(
  a, osakaal=paste(
    "(", round(prop.table(a)*100,1), "%)", sep=""
  )
)
```

Programm 3

```
a=table(perekonnaseis)
print(
  rbind(a, paste( "(", round(prop.table(a)*100,1), "%)", sep="")),
  quote=FALSE)
```

Saadud sagedustabelites/jaotustabelites esineb praegu mitu väiksemat iluviga. Esiteks pole saadud tabelitest võimalik näha, mitu leske valimis oli (mis sa arvad, miks?). Teiseks esinevad sagedustabelis tunnuse perekonnaseis kodeerimiseks kasutatud koodid (1, 2, 3, ...), mitte aga lugejale arusaadavad tähistused (vallaline/abielus/vabaabielus/...). Proovime väärtuste tähistused paika panna, tekitades uue nn faktortunnuse (nimega perekonnaseisF):

```
perekonnaseisF=factor(perekonnaseis, levels=1:5,
  c("vallaline", "abielus", "vabaabielus", "lahutatud", "lesk"))
```

Tunnuse lubatud
väärtused (kodeeringud)

kasutades nüüd tabelite tegemiseks tunnuse perekonnaseis asemel tunnust perekonnaseisF, saame veidi paremini loetavaid tabelleid:

```
table(perekonnaseisF)
```

Kodeeringute tähendused,
samas järjekorras kui esitatud
kodeeringute vektoris

Tulpdiagramm (sagedus- või jaotustabeli visualiseerimine)

Sagedus- või jaotustabelit on vahel mõistlik esitada graafikuna:

```
barplot(table(perekonnaseisF))
```

Enamasti tuleb soovitud välimusega graafiku saamiseks siiski veidi enam tööd teha – tuleb anda telgedele mõistlikud tähistused, valida meelepärased värvid jne:

```
barplot(table(perekonnaseisF),  
        col=c("skyblue", "midnightblue", "blue", "dodgerblue", "blue2"),  
        xlab="Perekonnaseis", ylab="Sagedus", main="Tulpdiagramm")
```

Värve saab määrata mitmel erineval viisil:

- Värvide nime järgi: käsk `colors()` trükkib välja kõigi R'le teadaolevate värvide nimed, võid vaadata ka veebilehelt

<http://www.ms.ut.ee/mart/R/varvid2.html>

- Saab kasutada värvipaletti, kuhu on juba koondatud teineteisega mingil kujul kokkusobivad värvid, proovi näiteks

```
barplot(1:10, col=terrain.colors(10))  
barplot(1:10, col=topo.colors(10))  
barplot(1:10, col=rainbow(10))  
barplot(1:10, col=heat.colors(10))
```

Graafikute jaoks sobivaid ja kauneid palette leidub ka R'i lisamoodulites, näiteks lisamoodulis `RcolorBrewer`.

- Saab kasutada värvi numbreid:

```
barplot(table(perekonnaseisF), col=1:5)
```

- Võib värve määrata kas halltooni või segades värve ise põhitoonidest kokku:

```
barplot(1:10, col=gray((1:10)/10))  
barplot(1:10, col=rgb((1:10)/10,1,0))
```

Proovi järgi, mida teevad järgmised käsud:

```
barplot(table(perekonnaseisF), horiz=TRUE)  
barplot(table(perekonnaseisF), legend.text=names(table(perekonnaseisF)),  
        col=topo.colors(7), names.arg="")
```

Kakuke

Teine lihtrahva poolt armastatud viis jaotus- või sagedustabelit graafiliselt esitada on kakuke ehk sektordiagramm. Alljärgnevad paar näidet kakukese küpsetamise võimaluste kohta R'is. Näidetes kasutatud värvivalik ei ole eeskujuna mõeldud!

```
pie(table(perekonnaseisF))
```

```
pie(table(perekonnaseisF, exclude="lesk"),  
    col=c("gold", "gold4", "red2", "pink"))
```

```
sagedustabel=table(perekonnaseisF, exclude="lesk")  
pie(sagedustabel, col=c("gold", "gold4", "red2", "pink"),  
    labels=paste(names(sagedustabel), "\n(",  
                  round(prop.table(sagedustabel)*100), "%)", sep=""))
```

Väärtuste järjekorra muutmine tabelites/graaafikutel

Järjestustunnuse väärtused peaksid tabelites/joonistel olema õiges järjekorras. Kui tunnused on sisestatud tekstiliselt, võib väärtuste järjekord osutuda valeks:

```
x=c("Vähe", "Vähe", "Palju", "Keskmiselt", "Kõige rohkem", "Palju")
table(x)
```

Keskmiselt	Kõige rohkem	Palju	Vähe
1	1	2	2

Lahendus – faktortunnuse defineerimisel määrame väärtuste soovitud järjekorra. NB! Kontrollige äärmiselt hooliga, et ei tee väärtuste sisestamisel kirjavigu!!!

```
fx=factor(x, levels=c("Vähe", "Keskmiselt", "Palju", "Kõige rohkem"))
table(fx)
```

Vähe	Keskmiselt	Palju	Kõige rohkem
2	1	2	1

Puuduvad väärtused

Vahel soovime / on mõistlik lisada sagedustabelisse ka puuduvad väärtused – kui paljudel juhtudel on meil uuritava tunnuse tegelik väärtus teadmata. Seda saab teha lisades table-käsule lisaparametri `exclude=NULL`:

```
table(kiirabi, exclude=NULL)
kiirabi
  0    1 <NA>
408  73 180
```

Ülesanne 1

Küsimusele „Kuidas hindate oma tervist antud hetkel?“ antud vastuste põhjal on moodustatud tunnus tervis. Väärtused tunnuses tervis on kodeeritud järgmiselt:

- 1 – väga hea
- 2 – hea
- 3 – keskmine
- 4 – halb
- 5 – väga halb

- a) Koosta sagedustabel, kus oleks näha kõigi ülaltoodud väärtuste esinemissagedus ja puuduvate vastuste esinemissagedus. Loo uus faktortunnus, mida kasutades näeksime mitte kodeeringut (väärtuseid 1,2,..5) vaid nende kodeeringute tähendusi („väga hea“, „hea“, ...). Loodud faktortunnust kasutades iseloomusta tunnuse jaotust tulpdiagrammi abil (joonista ilusates värvides tulpdiagramm, vaata et kõik teljed kannaksid ka mõistlikuid tähiseid).
- b) Koosta loodud faktortunnust kasutades jaotustabel, kus oleksid antud tunnuse väärtuste osakaalud esitatud protsentides täpsusega 2 kohta peale koma.
- c) Koosta eraldi jaotustabelid meestudengite ja naistudengite tervisliku seisundi iseloomustamiseks. Kas tunnuse tervis jaotus tundub olevat sarnane nii meestel kui ka naistel?

Hinnatud jaotuse täpsuse iseloomustamine ehk usaldusintervalli lisamine

Üks asi on oma vaatluste – valimi – jaotuse kirjeldamine, teine asi on populatsiooni jaotuse kirjeldamine. Juhuslikus valimis on küll uuritava tunnuse jaotus ligilähedane uuritava tunnuse jaotusele populatsioonis, kuid eriti väiksemate valimite korral võivad kõrvalekalded olla siiski märkimisväärsed. Kui eesmärgiks on kirjeldada, milline on uuritava tunnuse jaotus populatsioonis, siis on heaks mõtteks kasutada usaldusintervalle iseloomustamiseks seda, milline ikkagi võiks olla ühe või teise väärtuse tegelik osakaal uuritavas populatsioonis. Selleks tuleks lisada jaotustabelile (või jaotustabeli põhjal tehtud joonisele) usaldusintervallid, mis kirjeldavad, kui palju võiks populatsiooni jaotus erineda valimi jaotusest. Usaldusintervalli tõenäosusele (ja ka protsendile, kui korrutame tõenäosusele saadud usaldusintervalli 100-ga) saab R-is arvuda funktsiooni `binom.test` abil:

```
binom.test(12, 98)
```

soodsa sündmuse
toimumiskordade arv
valimis (antud väärtuse
esinemissagedus)

kõigi katsete arv
(valimi suurus)

Usaldusintervall tõenäosusele
(usaldusintervalli saamiseks
protsendile korruta 100-ga)

Käsu `binom.test` väljund:

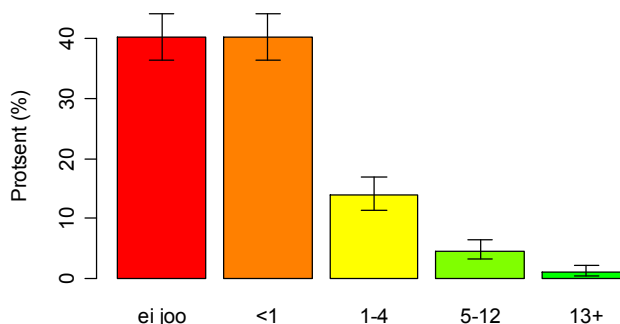
```
Exact binomial test
data: 12 and 98
number of successes = 12, number of trials = 98, p-value = 5.933e-15
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.06490049 0.20412687
sample estimates:
probability of success
      0.1224490
```

Sel viisil saab kõigi võimalike väärtuste esinemistõenäosuste jaoks leida usaldusintervallid ja lisada need soovi korral sagedustabelisse.

Sellist valmismeisterdatud käsku, mis võimaldaks joonistada uuritava tunnuse jaotuse koos usalduspiiridega R-is pole (vähemalt pole ma leidnud sellist varianti, mis mulle meelepärane oleks).

Soovitud graafiku tegemiseks tuleks esmalt leida usalduspiirid (näiteks `binom.test`-käsu abil) ja need siis `arrows`-käsku kasutades joonisele lisada. Samas on R aga täiendatav ja me võime kerge vaevaga sellise käsu ise juurde meisterdada. Järgmine programmilõik defineerib uue käsu (tulpdia) mis joonistab tulpdiagrammi protsendile koos 95%-usalduspiiridega (kõigest programmis toimuvast ei pea aru saama...):

Õllejoomise jaotus koos 95%-usaldusintervalliga



```
# Usaldusintervallidega tulpdiagrammi joonistav funktsioon

tulpdia=function(tunnus, ...){
  alpha=0.025
  x=table(tunnus); n=sum(x)
  al=qbeta(alpha, x, n - x+1)
  yl=qbeta(1-alpha, x+1, n - x)
  a=barplot(prop.table(x)*100, ylab="Protsent (%)",
            ylim=c(0, max(yl*100)), ...)
  arrows(a, al*100, a, yl*100, code=3, length=0.1, angle=90)
}
```

Nüüd kui soovitud käsk on R-ile lisatud (üalaloodud programm on korra käivitatud), võime vaevata teha soovitud joonised:

```
tulpdia(perekonnaseisF)

tulpdia(olu, col=rainbow(12),
        main="õlletarbijate jaotus koos 95%-usaldusintervalliga")

tulpdia(sugu)
```

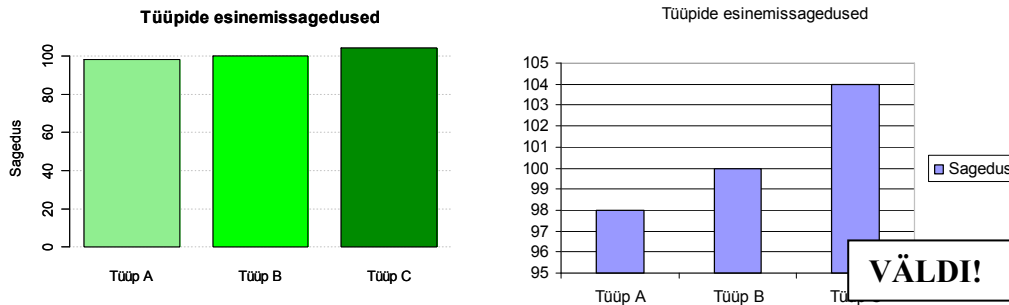
Julgetele

Vaata, mida teeb järgmine programm

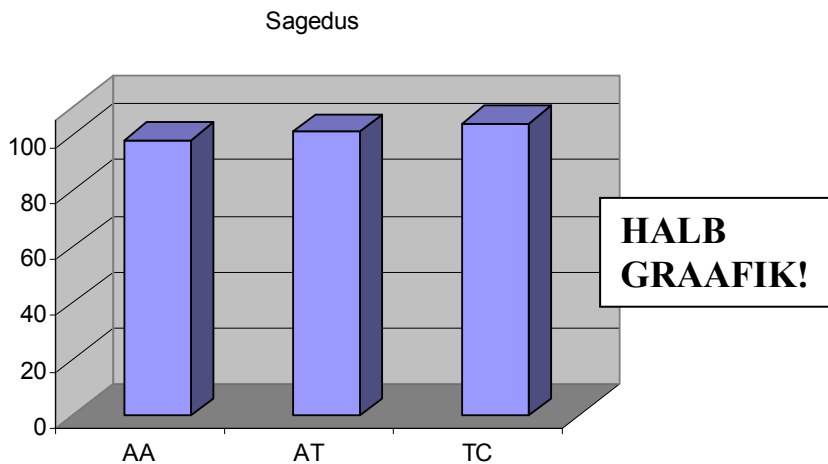
```
ajut=prop.table(table(perekonnaseisF))*100
x=barplot(ajut, ylim=c(0, 100), ylab="Protsent (%)")
text(x,ajut+2, paste(round(ajut,1), "%", sep=""), adj=c(0.5, 0) )
```

Soovitused

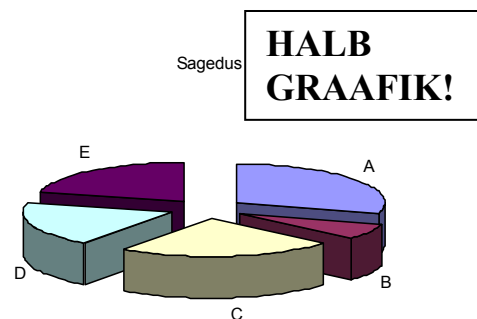
A. Võimaluse korral väldi tulpdiagrammi tegemisel telgi, mis ei alga 0-st (R-i barplot alustab alati nullpunktist, kuid mõned teised programmid võivad teisiti käituda). Põhjus – graafikut vaatav inimene võib kergesti jõuda ekslikele järeldustele (oi, kui suur erinevus!) kui telgedega mängida. Seda tasub meeles pidada ka graafikute vaatamisel ja interpreteerimisel – ära lase end eksitada! All on toodud sama sagedustabeli (samad andmed) kaks visualiseeringut.



B. **Ära kasuta 3D-graafikuid!** Ruumilistelt graafikutelt on väga raske informatsiooni maha lugeda, eksimused on kerged tulema! Kasuta vaid siis, kui eesmärgiks on statistika abil valetada (auditooriumit eksitada)! Vaata näiteks alltoodud Excelis tehtud joonist. Kas loed ikka jooniselt välja, et variandi „AT“ (keskmine tulp) esinemissagedus on üle 100 (tegelik väärtus – 101)? Või oleksid joonise põhjal arvanud midagi muud?



Kas A-d esineb rohkem kui C-d? Siit graafikult on väga raske näha, et A-d esineb 25% rohkem kui C-d!



Pidev või paljude väärtustega diskreetne jaotus

Sagedustabel

Pideva tunnuse väärtused tuleb sagedustabeli koostamisel jagada sobivateks vahemikeks. Vahemikesse jagamist saab kõige mugavamalt teha *cut*-käsu abil:

```
> pikkusKL=cut(pikkus, seq(140, 210, 10))
```

Midagi, mis on väiksem kui tunnuse väikseim võimalik väärtus

Maksimaalne väärtus või veidi suurem

samm – vahemiku laius

Peale vahemikesse jagamist võib juba teha sagedustabeli:

```
> table(pikkusKL)
pikkusKL
(150,160] (160,170] (170,180] (180,190] (190,200] (200,210]
      53       294       220       76        15         1
```

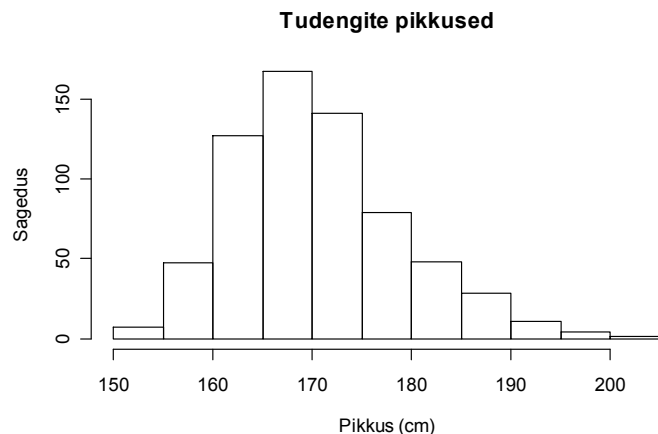
Histogramm

Histogrammi saab joonistada *hist*-käsu abil.

```
hist(pikkus,
     main="Tudengite pikkused",
     xlab="Pikkus (cm)",
     ylab="Sagedus")
```

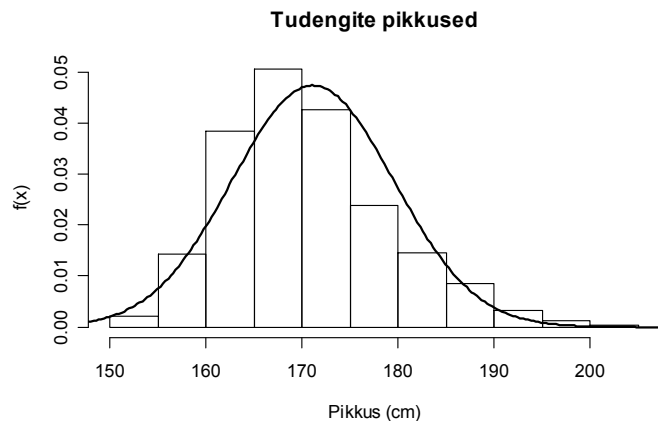
Vahel soovime joonisele lisada ka andmetega kõige paremini sobiva normaaljaotuse tihedusfunktsiooni graafiku (võrdlemiseks – kas on normaaljaotuse moodi või ei).

Selleks lisame joonistatud histogrammile *lines*-käsu abil normaaljaotuse tihedusfunktsiooni.



```
hist(pikkus,
     main="Tudengite pikkused",
     xlab="Pikkus (cm)",
     ylab="f(x)", freq=F)

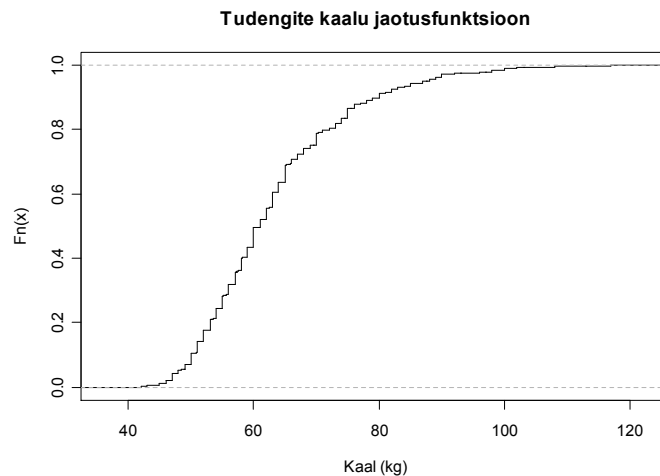
x=seq(140, 220, length=300)
y=dnorm(x,
     mean=mean(pikkus, na.rm=T),
     sd=sd(pikkus, na.rm=T))
lines(x,y, lwd=2)
```



Jaotusfunktsiooni graafik

Harvem, aga vahel siiski, kasutatakse jaotuse iseloomustamiseks ka jaotusfunktsiooni graafikut.

Jaotusfunktsiooni kasutatakse siis, kui huvitatakse küsimustest nagu „Kui kõrgele pean maja ehitama, et suurvesi ujutaks maja üle harvem kui kord saja aasta jooksul?“; „Kui kõrge peab olema auditooriumi uks, et 99% tudengiteks pääseks sisse ilma kummardamata?“ jne. Teisisõnu öeldes – olukordades, kus lugejat huvitab eelkõige kvantiilide leidmine/uurimine.

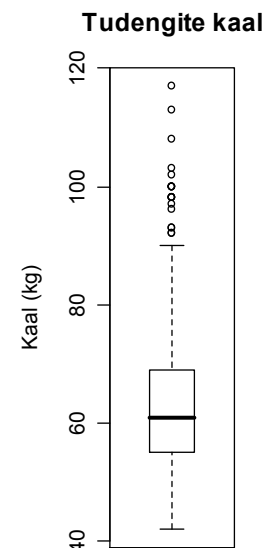


```
plot(ecdf(kaal),
     do.points=F, verticals=T,
     xlab="Kaal (kg)",
     main="Tudengite kaalu jaotusfunktsioon")
```

Karp-vurrud diagramm (boxplot)

Karp-vurrud diagrammi kasutatakse jaotuse iseloomustamiseks eelkõige seoste uurimisel (naiste/meeste erinevus; kontrollgrupp vs „töödeldud“ grupp). Samas võib antud graafiku tüüp osutada kasulikuks jaotuse iseloomustamiseks ka siis, kui vaatluseid on vähe. Mõistliku histogrammi saamiseks peab sageli olema palju vaatluseid, väikeste valimite jaoks ei tasugi histogrammi joonistama hakata. Samas võib karpdiagramm anda hea ettekujutuse andmetest ka kümnekonna vaatluse olemasolul.

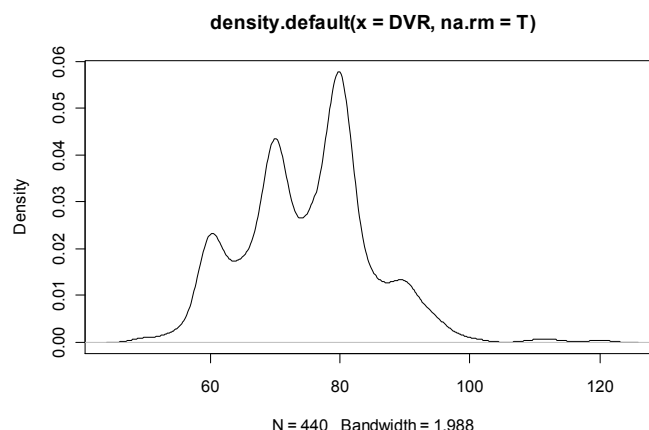
```
boxplot(kaal, main="Tudengite kaal", ylab="Kaal (kg)")
```



Tihedusfunktsiooni hindamine

Uuritava pideva tunnuse tihedusfunktsiooni on võimalik hinnata ka nn tuumameetodil (kernel density estimation). Mainitud meetod annab võrreldes histogrammiga täpsema hinnangu uuritava pideva tunnuse tihedusfunktsioonile. Samas võib mainitud meetodi nõrku kohti ja käitumist mitteteadev inimene end kergesti eksitada lasta ja graafikut valesti interpreteerimisel – sestap kasuta antud meetodit ettevaatlikult, ole skeptiline ja võrdle tulemusi sulle tuttavale meetodile saadud graafikuga – näiteks histogrammiga!

```
plot(density(DVR, na.rm=T))
```



Jaotust iseloomustavad statistikud

Üks võimalus pideva tunnuse jaotuse iseloomustamiseks on anda paari-kolme statistiku väärtused. Peamiselt raporteeritakse keskmist ja standardhälvet; tugevalt asümmeetriliste jaotuste (üksikud väga suured või väga väikesed väärtused) tasub kaaluda lisaks keskmisele ka mediaani raporteerimisest (või keskmise asemel). Keskmist ja standardhälvet teades saab uuritava tunnuse väärtustest juba mingi ettekujutuse, see on sageli sobiv lahendus olukorras, kus tunnus pole piisavalt oluline ruumi nõudva histogrammi lisamiseks (st. sobiv viis „taustatunnuste“ jaotuse kirjeldamiseks). Peamised ühe tunnuse jaotust iseloomustavad statistikud on R-is leitavad järgmiste käskude abil:

```
mean      – keskmine
median    – mediaan
sd        – standardhälve
var       – dispersioon
min       – miinimum
max       – maksimum
quantile  – kvantiilid
summary   – käsk võimaldab leida korraga peamised jaotust iseloomustavad
           statistikud
```

Kõigi ülaltoodud funktsioonide puhul (välja arvatud summary-käsk) antakse puuduvate vaatluste olemasolu korral vastuseks „NA“ – õige vastus on teadmata. Kui soovime, et puuduvaid väärtuseid arvutuste tegemisel ignoreeritakse, tuleb lisada käsule lisaparameter `na.rm=T`.

```
> mean(pikkus, na.rm=T)
[1] 171.1167
> quantile(pikkus, 0.25, na.rm=T)
25%
165
> summary(pikkus)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
150.0   165.0   170.0   171.1   176.0   201.0     1.0
```

Näide kirjandusest: jaotust iseloomustavate statistikute esitus artiklis/raportis.

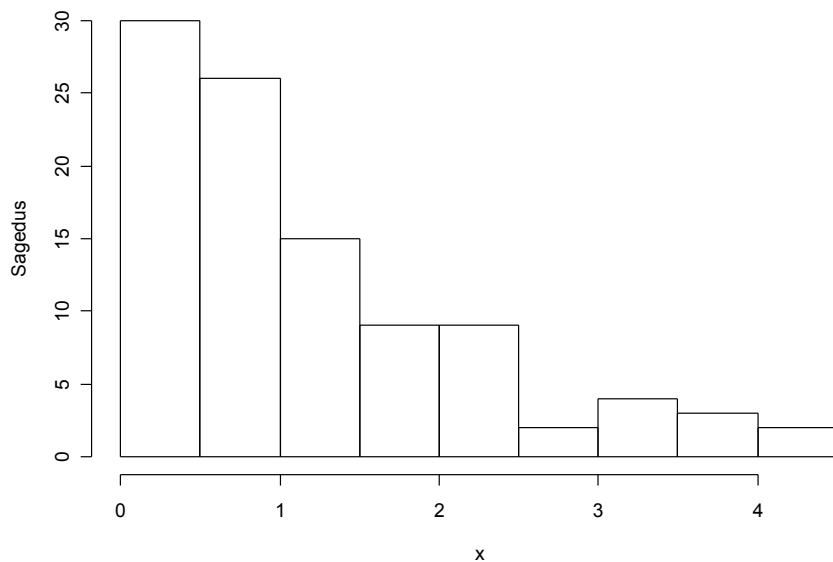
Table 6 Baseline characteristics; overall and divided by *ACE2* -381 T>C genotype

Characteristic	Overall Population (N= 147)	
	Mean	Std. Dev.
AGE	70.3	6.84
%Male	51%	
%European American	83%	
New York Heart Assoc. Class	2.03	1.36
Systolic Blood Pressure	145.96	22.26
Body Mass Index	29.79	6.74
Hemoglobin	13.48	1.54
Blood Urea Nitrogen	19.81	6.96
Estimated Creatinine Clearance ^a	84.04	30.90
Echocardiographic E:A ratio	0.95	0.37

Histogrammi interpreetimisest

Kas oskad pideva tunnuse jaotust iseloomustava histogrammi abil öelda, milline on uuritava tunnuse keskmine, standardhälve, mediaan? Vaata alljärgnevaid jooniseid ja tee oma pakkumine!

1. pilt (histogramm)

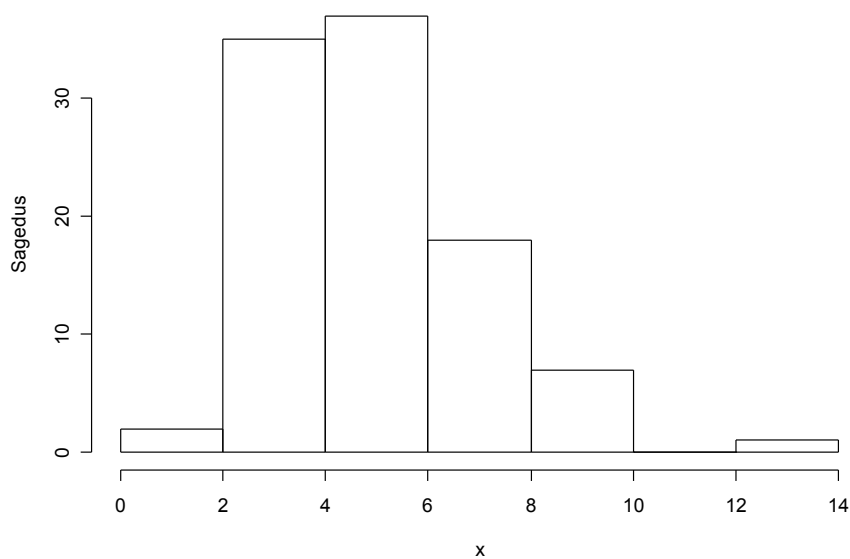


$\bar{x}$ =

med =

s =

2. pilt (histogramm)



$\bar{x}$ =

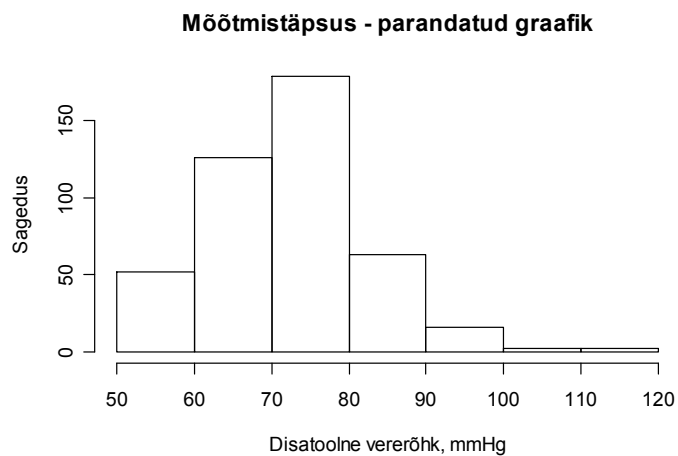
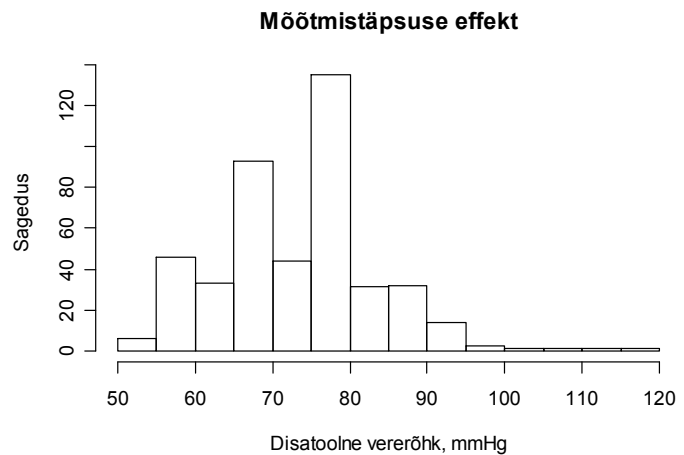
med =

s =

Võimalikud keerukamad situatsioonid

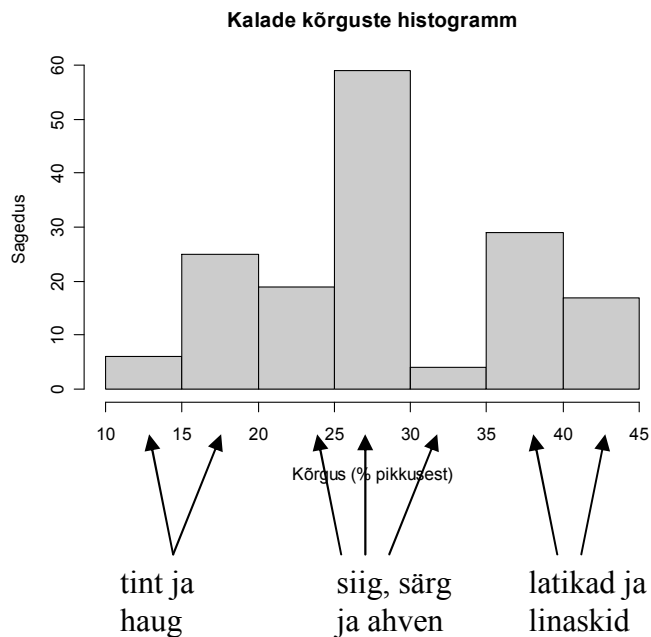
Probleemid ebapiisava mõõtmistäpsusega

Sageli juhtub, et pideva tunnuse väärtuseid ei saa praktikas kuigi täpselt mõõta. Võib-olla on kaalud mõõdetud kõigest grammi täpsusega, pikkused vaid mm täpsusega jne. Vahel võib mõõtmistulemuste „ümardamisest“ tingitud mõju ilmned ka histogrammil. Alltoodud joonis kajastab ühte sellist olukorda – vererõhk on enamasti mõõdetud kui „60“, „70“ või „80“. Mõõtmistäpsusest tingituna ilmnevad histogrammil „hambad“ – väärtuseid „60“, „70“ ja „80“ on palju, aga väärtuseid „71“, „83“ jne pole. Juhul, kui eesmärgiks on mõõtmistäpsuse uurimine, võib muidugi taolist hammastega graafikut esitada (võiks isegi vahemike arvu suurendades hambad selgemalt esile tuua), kui aga eesmärgiks on uuritava tunnuse jaotuse kirjeldamine, tuleb vahemikud valida selliselt, et histogrammi tulba laius kataks täisarv kordi „mõõdetavaid ühikuid“ – grammides kaalu mõõtes peaks vahemikud olema täisarv „gramme“ pikad. Probleemi illustreerimiseks toodud graafiku puhul peaks vahemiku laiuseks võtma 10.



Alamgrupid

Vahel võib histogrammi silmitsedes märgata mitut „kühmu“ ehk uhkemas keeles öeldes – jaotus on multimodaalne (bimodaalne jaotus – uuritava tunnuse jaotus on kahe „kühmuga“). Enamasti tasub aru saada ja mõista, mis siis ikka tekitab multimodaalset jaotust (noorloomad/täiskasvanud; isased/emased; erinevad alamliigid vms). Soovitav oleks multimodaalsuse põhjust kommenteerida juba uuritava tunnuse jaotust kirjeldades.



Ülesanne 2

- Joonista histogrammid nii meetudengite pikkusele kui ka naistudengite pikkusele.
- Muuda joonistatud histogrammid sellisteks, et jooniste teljed oleksid võrreldavad (naistudengite graafikul ulatugu pikkused sama kaugele kui meestudengite joonisel). Selleks proovi hist- käsule lisada lisaparametrit `xlim=c(telje algus, telje lõpp)`, vaata ja proovi ka alltoodud näidet:

```
hist(pikkus, xlim=c(0, 250))
```