Jaotuse kirjeldamine

Antud materialis toodu näiteid on võimalik ka ise läbi teha, kui lugeda R'i sisse Tartu Ülikooli arstiteaduskonna tudengite küsitlemisel saadud andmestik *kokku*. Seda saab teha käskudega:

```
load(url("http://www.ms.ut.ee/mart/MC2007/kokku.Rdata"))
attach(kokku)
```

Nominaalne / järjestus- / (väheste väärtustega) diskreetne tunnus

Kui uuritaval tunnusel on suhteliselt vähe väärtuseid, siis on parimaks võimaluseks uuritava tunnuse jaotust kirjeldada sagedustabel (või sagedustabelit visualiseeriv graafik).

Eeltöö
Juhul, kui uuritav tunnus on kodeeritud, siis võiks enne tabelite koostamist/graafikute joonistamist panna paika „tähistused". NB! Vaadake, et tähistused antakse ette õiges järjekorras – tehke kasvõi sagedustabel tunnusele olu – `table(olu)` – ja vaadake, millises järjekorras R tunnuse algseid väärtuseid näeb!

```
oluF=factor(olu, labels=c("ei joo", "<1", "1-4","5-12","13+"))
```

Sagedustabel

Sagedustabel tunnusele oluF:

```
> table(oluF)
```

```
oluF
ei joo    <1    1-4   5-12    13+
   266   266    92     30      7
```

Erinevate väärtuste osakaalud:

```
> prop.table(table(oluF))
```

Protsent (vahel kutsutud ka kui jaotustabel):

```
> prop.table(table(oluF))*100
```

Protsent ümmardatult (täpsusega 1 koht peale koma):

```
> round(prop.table(table(oluF))*100,1)
```

```
oluF
ei joo    <1    1-4   5-12    13+
  40.2  40.2  13.9    4.5    1.1
```

Sagedustabeli esitamine raportis

Vahel ei esitata sagedustabelit tabelina, vaid antakse vajalikud arvud edasi tekstina („There were 9399 (61.9%) men and 5790 (38.1%) women in the database"). Sagedustabelit tabelina esitades võib (soovi korral) esitada nii sagedused (mitu sellist oli) kui ka protsendi:

Tabel 1. Tudengi poolt nädala aja jooksul joodud õllepudelite arvu jaotus.

	ei joo	<1	1-4	5-12	13+
sagedus	266 (40%)	266 (40%)	92 (14%)	30 (5%)	7 (1%)

Veel paar näidet ajakirjadest (kuhu, tõsi küll, on lisatud ka muud informatsiooni):

Table 8 Haplotype frequencies

Haplotype	African American (%)	European American (%)
T G	17 (35.4)	133 (55)
C A	22 (45.8)	103 (42.4)
T A	9 (18.8)	4 (1.6)
C G		2 (1)

Allele and Genotype Frequencies of the C(−260)→T Polymorphism in the CD14 Gene Promoter in Patients With Myocardial Infarction and in the Control Group

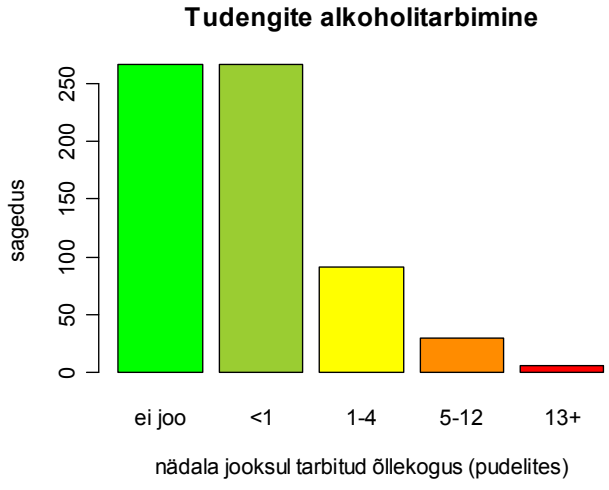
	Patients (n=178)		Controls (n=135)		
	n	%	n	%	P
Allele					
T	175	49.2	95	35.2	0.0005
C	181	50.8	175	64.8	
Genotype					
T/T	49	27.5	21	15.6	0.0049
C/T	77	43.3	53	39.3	
C/C	52	29.2	61	45.2	

Sagedustabeli visualiseerimine

Pea kõigil alltoodud näidete puhul võib table(oluF) asemel kasutada `prop.table(table(oluF))*100` – saamaks sageduste asemel protsente iseloomustava joonise.

Tulpdiagram (barplot)

Tulpdiagramm on üks soovitavamaid viise sagedustabelit visualiseerida. Soovi korral võib tulpdiagrammilt väikese vaevaga (ja üsna täpselt) välja lugeda algsed numbrilised väärtused – seega kõlbab ta kõige paremini asendama sagedustabelit ennast. Võib kaaluda ka viitejoonte lisamist graafikule – vaata lisad.




```
barplot(table(oluF))
```

või

```
barplot(table(oluF), xlab="nädala jooksul tarbitud õllekogus (pudelites)",
        ylab="sagedus", main="Tudengite alkoholitarbimine",
        col=c("green","yellowgreen","yellow","darkorange","red"))
```

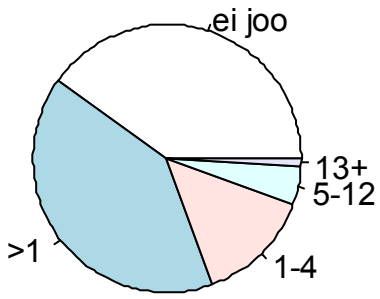
Ringdiagram ehk kakuke (piechart)

Kakukest sobib kasutada nominaalse tunnuse jaotuse iseloomustamiseks. Pildil toodud näide on paraku tehtud järjestustunnusega – ja ausalt öeldes pole just tegu parima valikuga võimalikest.

```
pie(table(oluF))
```

või

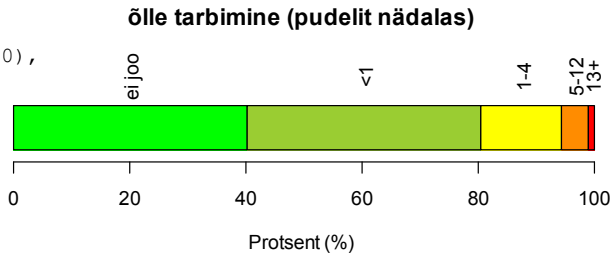
```
pie(table(oluF), main="Tudengite alkoholitarbimine")
```



Veel üks võimalus

Veel üks võimalus on kasutada nn stacked barchart'i. Kuna vastav graafik on eriti kasulik seal, kus tahame jaotuseid võrrelda, siis vaata lähemalt seoste iseloomustamist kirjeldavast osast.

```
par(xpd=NA)
barplot(as.matrix(prop.table(table(oluF))*100),
  ylim=c(0, 2.2), xlab="Protsent (%)",
  main="õlle tarbimine (pudelit nädalas)",
  col=c("green","yellowgreen","yellow",
        "darkorange","red"), horiz=T)
a=cumsum(prop.table(table(oluF))*100)
kohty=(a+c(0, a[-length(a)]))/2
text(kohty, 1.9, names(table(oluF)), srt=90)
```



Sagedamini esinevad probleemid

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 lisaparameetri `exclude=NULL`:

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

Andmestikus esindamata väärtused

Diskreetse tunnuse puhul peavad tabelites ja joonistel olema esindatud ka vahepealsed väärtused – mis siis, et meie valimis polnud ühtegi pesa 4 linnupojaga (aga olid 3 ja 5 linnupojaga pesad) – tabelisse ja graafikule tuleb ikkagi nelja linnupojaga variant kanda. Sama kehtib järjestustunnuse puhul – ka seda tuleb öelda, et vastsevariante „meeldis väga“ lihtsalt polnud vms. Kuidas seda teha?

Ütleme faktortunnuse defineerimisel, millised on tunnuse võimalikud väärtused. Seda saab teha `levels`-käsu abil:

```
oluF=factor(olu, levels=1:6,
      labels=c("ei joo", "<1", "1-4","5-12","13-100", "100+"))
table(oluF)
```

```
oluF
ei joo     <1    1-4   5-12    13+   100+
   266    266     92     30      7      0
```

Teine näide – sagedustabel tudengite vanustele

```
table(factor(vanus, levels=min(vanus):max(vanus)))
```

Väärtuste järjekorra muutmine tabelites/graafikutel

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 hoolega, 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
```

Populatsiooni jaotuse iseloomustamine ehk usaldusintervalli lisamine

Üks asi on oma vaatluste – valimi – jaotuse kirjeldmine, teine asi on populatsiooni jaotuse kirjeldamine. Populatsiooni jaotuse kirjeldamise peamiseks võimaluseks on lisada sagedustabelile (joonisele) usaldusintervallid, mis kirjeldavad, kui palju võiks populatsiooni jaotus erineda valimi jaotusest. Usaldusintervalli tõenäosusele (ja seega ka protsendile) saab R-is arvuda funktsiooni binom.test abil:

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

soodsa sündmuse toimumiskordade arv (antud väärtuse esinemissagedus)

kõigi katsete arv (valimi suurus)

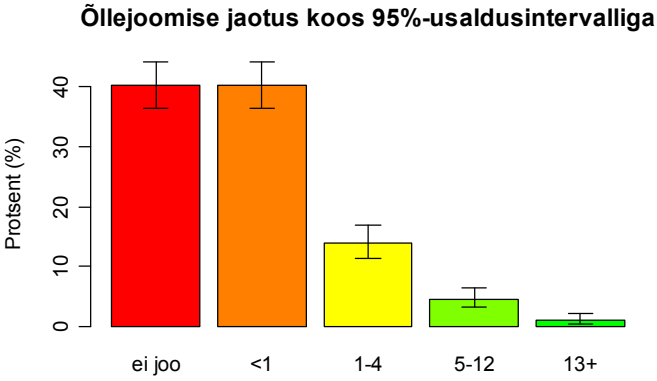
Usaldusintervall tõenäosusele (usaldusintervalli saamiseks protsendile korruta 100-ga)

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. 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:

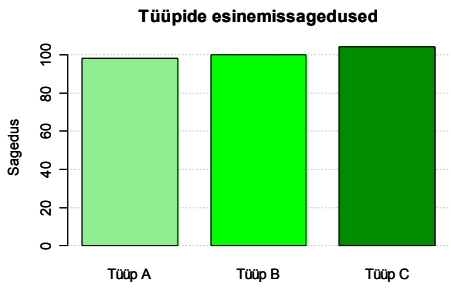
```
tulpdia=function(tunnus, ...){
  alpha=0.025
  x=table(tunnus); n=sum(x)
  al=qbeta(alpha, x, n - x+1)
  yl=qbeta(1-alpha, x+1, n - x)
  a=barplot(prop.table(x)*100, ylab="Protsent (%)",
      ylim=c(0, max(yl*100)), ...)
  arrows(a, al*100, a, yl*100, code=3, length=0.1, angle=90)
}
```

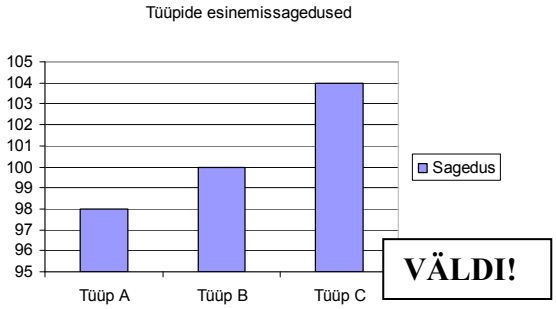
Nüüd kui soovitud käsk on R-ile lisatud, võime vaevata teha soovitud joonised:

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

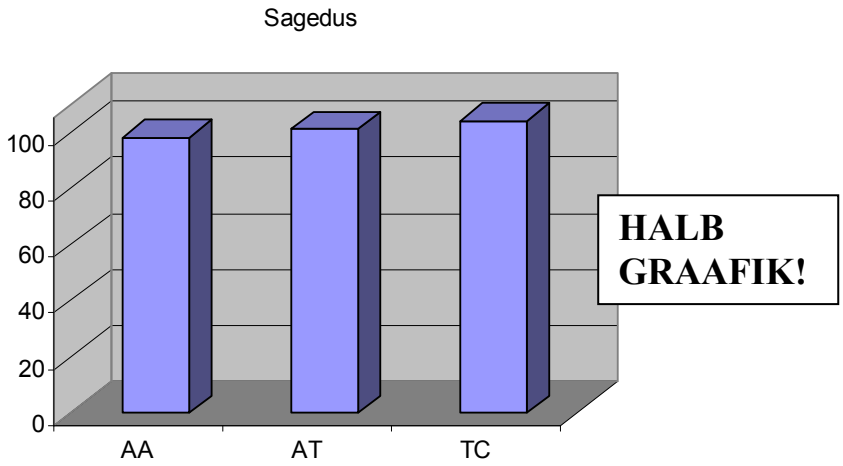
Soovitused

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.

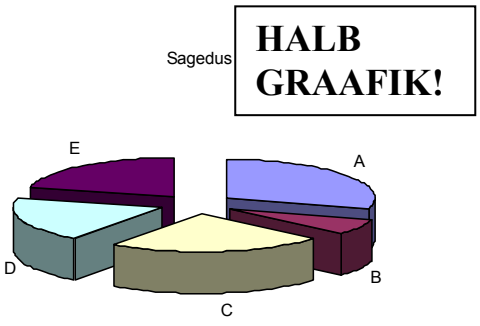




B. **Ära kasuta 3D-graafikuid**! Ruumilistelt graafikutelt on väga raske informtasiooni maha lugeda, eksimused on kerged tulema! Kasuta vaid siis, kui eesmärgiks on statistika abil valetada (auditooriumit eksitada)! Vaata näiteks alltoodud Excelis tehtud joonist. Kas loed ikka jooniselt välja, et variandi „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

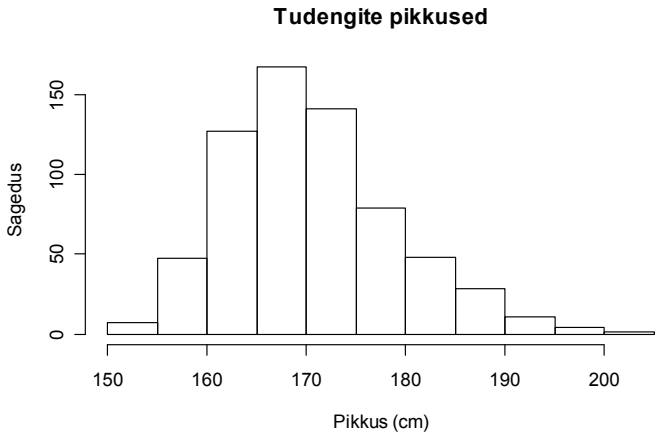
samm – vahemiku laius

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.

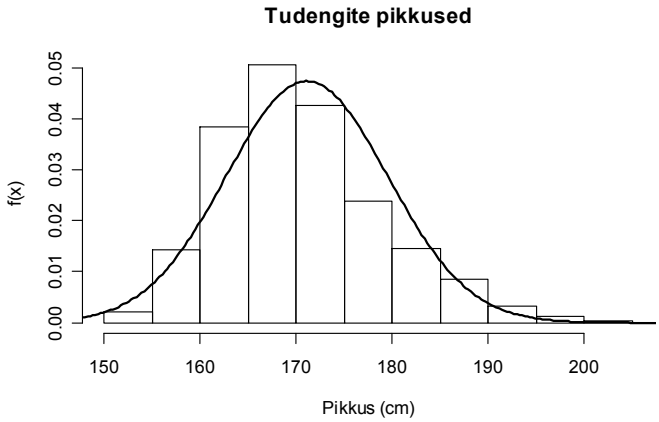
```
hist(pikkus,
  main="Tudengite pikkused",
  xlab="Pikkus (cm)",
  ylab="Sagedus")
```



Vahel soovime joonisele lisada ka andmetega kõige paremini sobiva normaaljaotuse tihedusfunktiooni graafiku (võrdlemiseks – kas on normaaljaotuse moodi või ei).
Selleks lisame joonistatud histogrammile lines-käsu abil normaaljaotuse tihedusfunktsiooni.

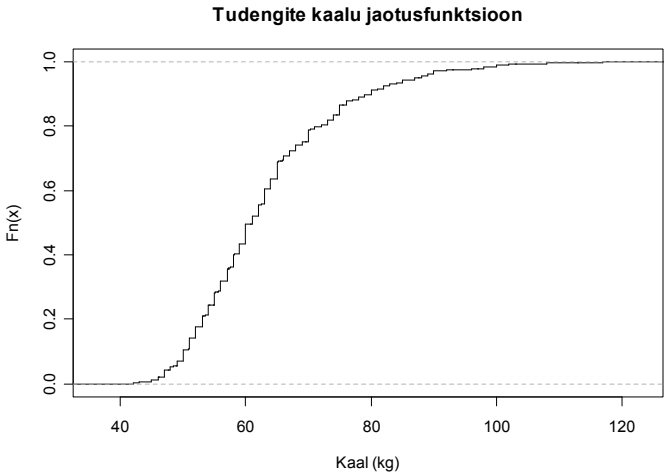
```
hist(pikkus,
   main="Tudengite pikkused",
   xlab="Pikkus (cm)",
   ylab="f(x)", freq=F)

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

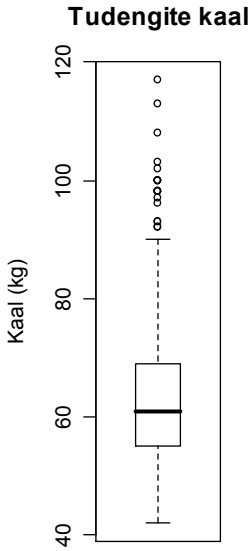
Harvem, aga vahel siiski, kasutatakse jaotuse iseloomustamiseks ka jaotusfunktsiooni graafikut. Jaotusfunktsiooni kasutatakse siis, kui huvitutakse küsimustest nagu „Kui kõrgele pean maja ehitama, et suurvesi ujutaks maja üle harvem kui kord saja aasta jooksul?“; „Kui kõrge peab olema auditooriumi uks, 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)

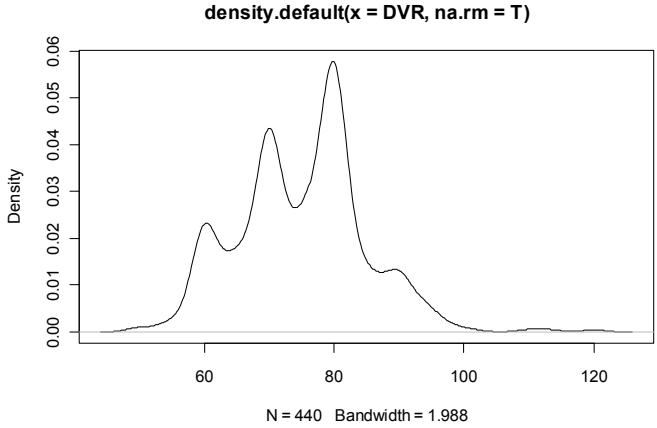
Karp-vurrud diagrammi kasutatakse jaotuse iseloomustamiseks eelkõige seoste uurimisel (naiste/meeste erinevus; kontrollgrupp vs „töödeldud“ grupp). Samas võib antud graafiku tüüp osutuda kasulikuks jaotuse iseloomustamiseks ka siis, kui vaatluseid on vähe. Mõistliku histogrammi saamiseks peab sageli 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 tuttaval meetodil 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 reporteerimisest (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` – kvantiil
`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 lisaparameeter `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, kuidas jaotust iseloomustavaid statistikuid saab artiklis/raportis esitada.

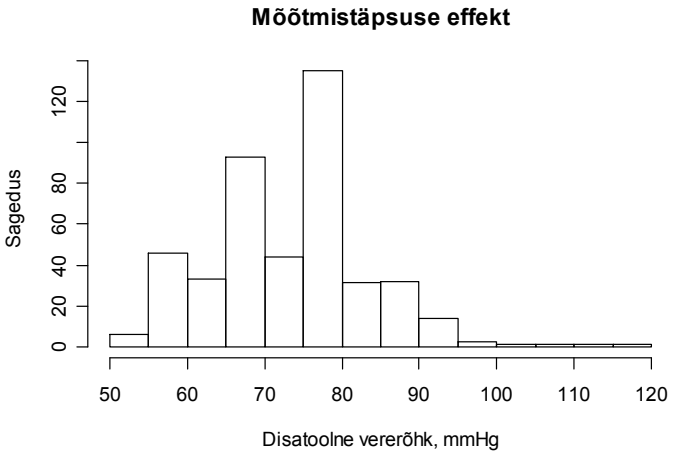
Table 6 Baseline characteristics

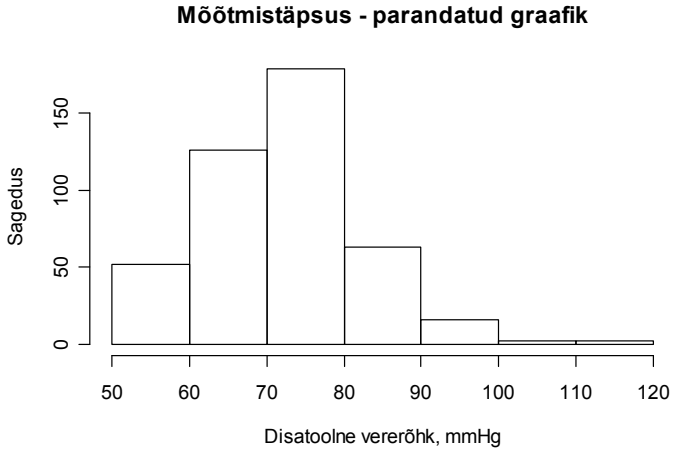
Characteristic	Overall Population (*N*= 147)	
	Mean	Std. Dev.
AGE	70.3	6.84
%Male	51%	
%European American	83%	
New York Heart Assoc. Class	2.03	1.36
Systolic Blood Pressure	145.96	22.26
Body Mass Index	29.79	6.74
Hemoglobin	13.48	1.54
Blood Urea Nitrogen	19.81	6.96
Estimated Creatinine Clearance[a]	84.04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

Histogrammi interpreteerimisest

Probleemid ebapiisava mõõtmistäpsusega

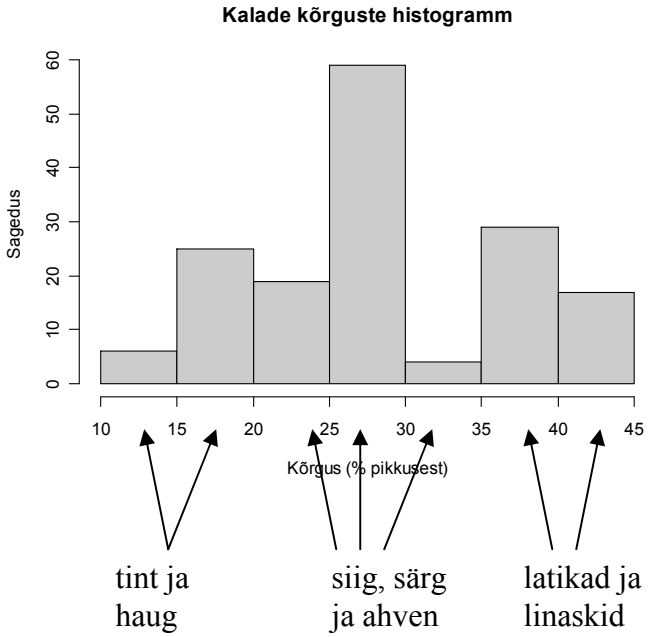
Sageli juhtub, et pideva tunnuse väärtuseid ei saa praktikas kuigi täpselt mõõta. Võib-olla on kaalud mõõdetud kõigest grammi täpsusega, pikkused vaid mm täpsusega jne. Vahel võib mõõtmistulemuste „ümmardamisest“ tingitud mõju ilmneda ka histogrammil. Alltoodud joonis kajastab ühte sellist olukorda – vererõhk on enamasti mõõdetud kui „60“, „70“ või „80“. Mõõtmistäpsusest tingituna ilmnevad histogrammil „hambad“ – väärtuseid „60“, „70“ ja „80“ on palju, aga väätuseid „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.



tint ja haug

siig, särg ja ahven

latikad ja linaskid