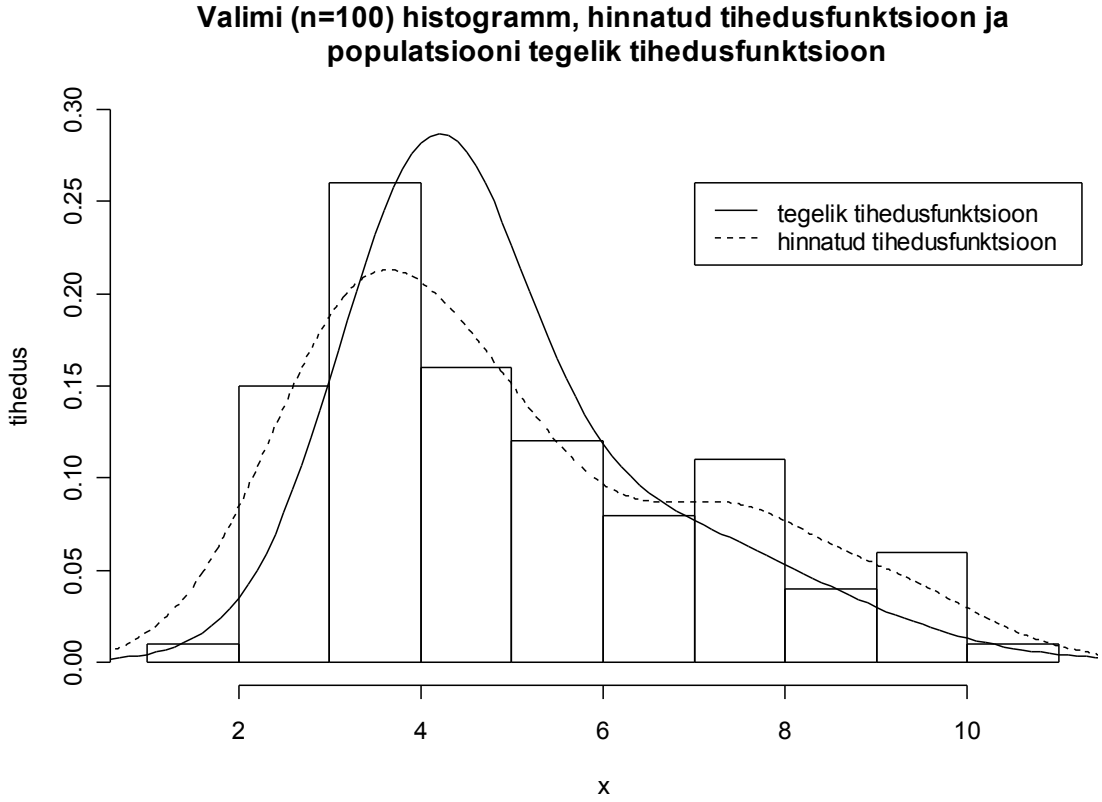
Biomeetria bioloogidele
3. praktikum

Näide: Tegelik jaotus, hinnatud jaotus ja jaotust kirjeldavad statistikud

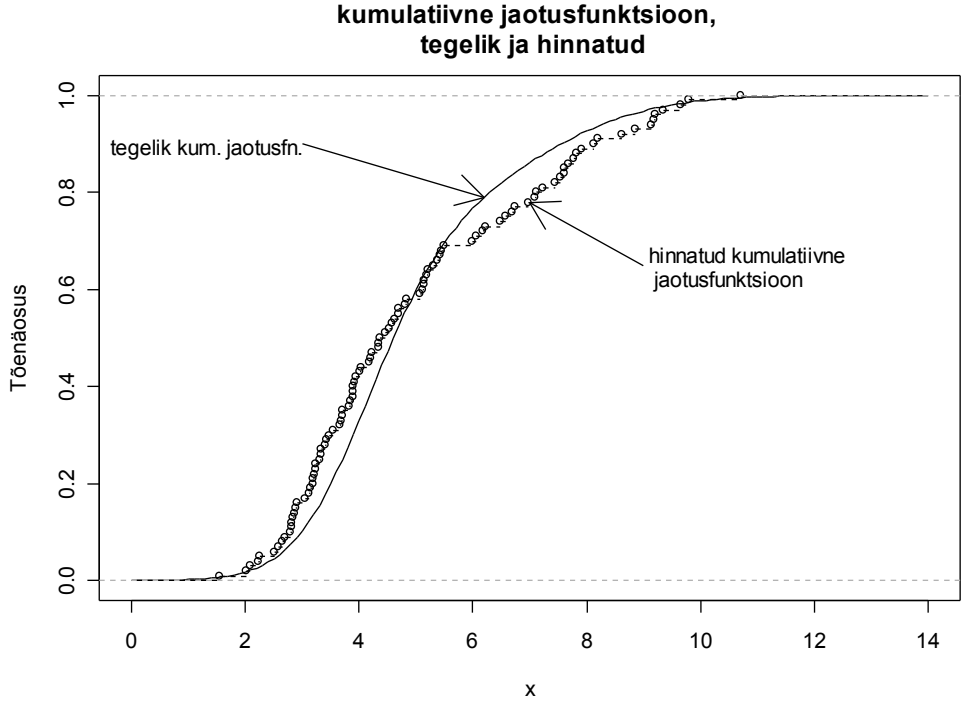


Tegelik (populatsiooni):

keskväärtus	EX = 4,94
mediaan	med = 4,60...
standardhälve	σ = 1,80...

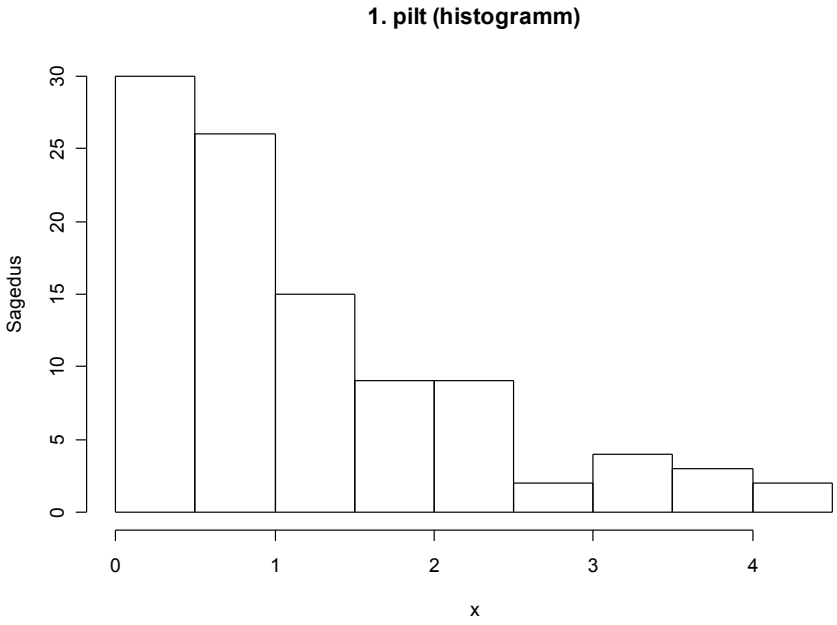
Valimi:

keskmine	$\bar{x}$ = 5,02...
mediaan	med = 4,40..
standardhälve	s = 2,14

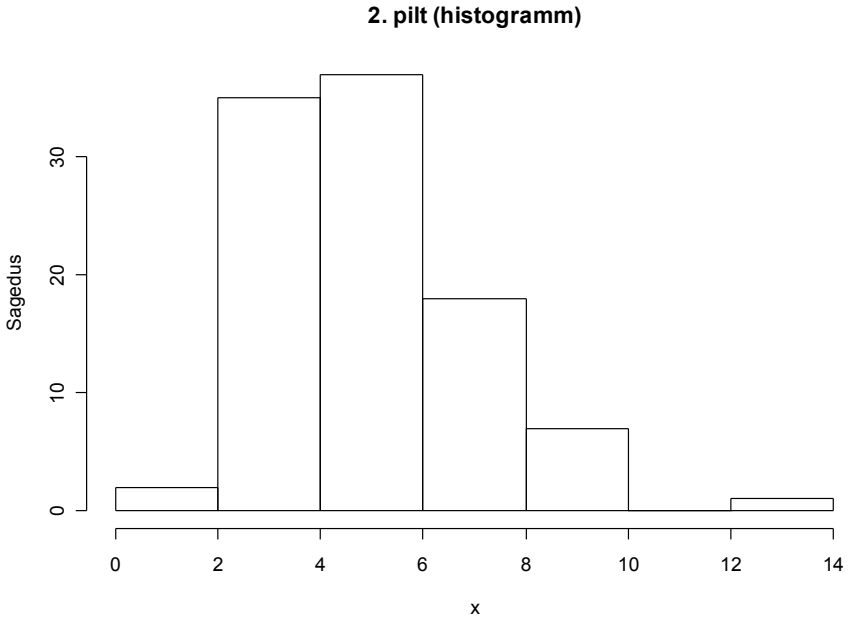


Ülesanne 1

Kasutades jooniseid ürita ära arvata statistikute väärtuseid

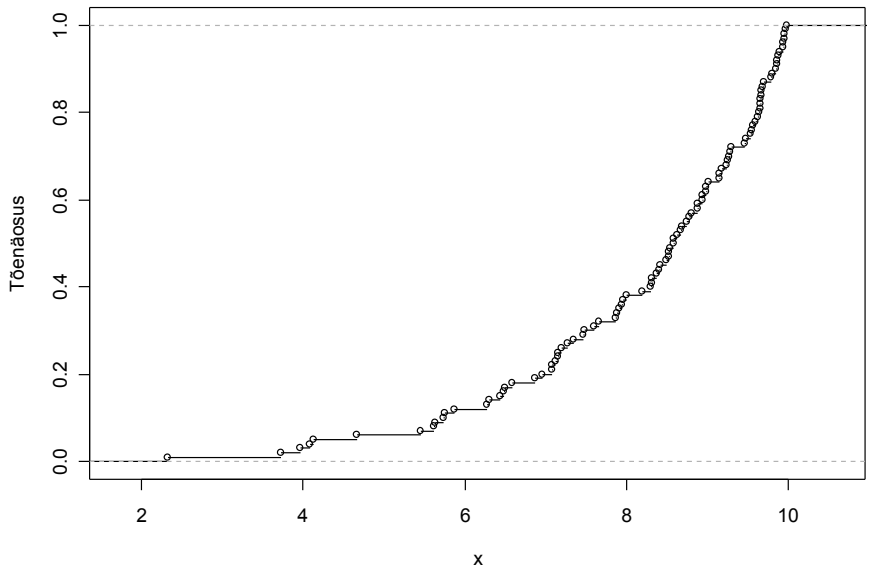


$\bar{x}$ = ...
med = ...
s = ...



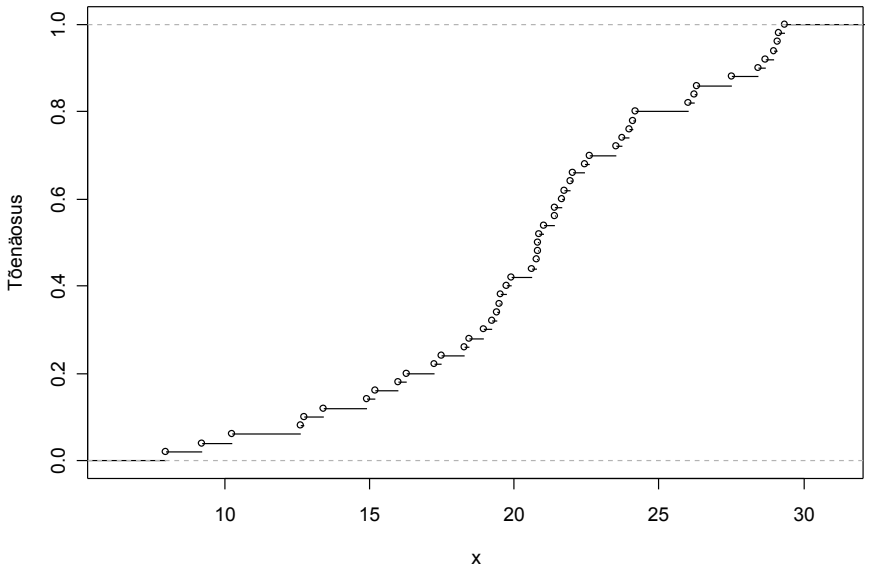
$\bar{x}$ = ...
med = ...
s = ...

3. pilt (kumulatiivne jaotusfunktsioon)



med = ...
$\bar{x}$ = ...
s = ...
0,05-kvantiil = ...
0,95-kvantiil = ...

4. pilt (kumulatiivne jaotusfunktsioon)

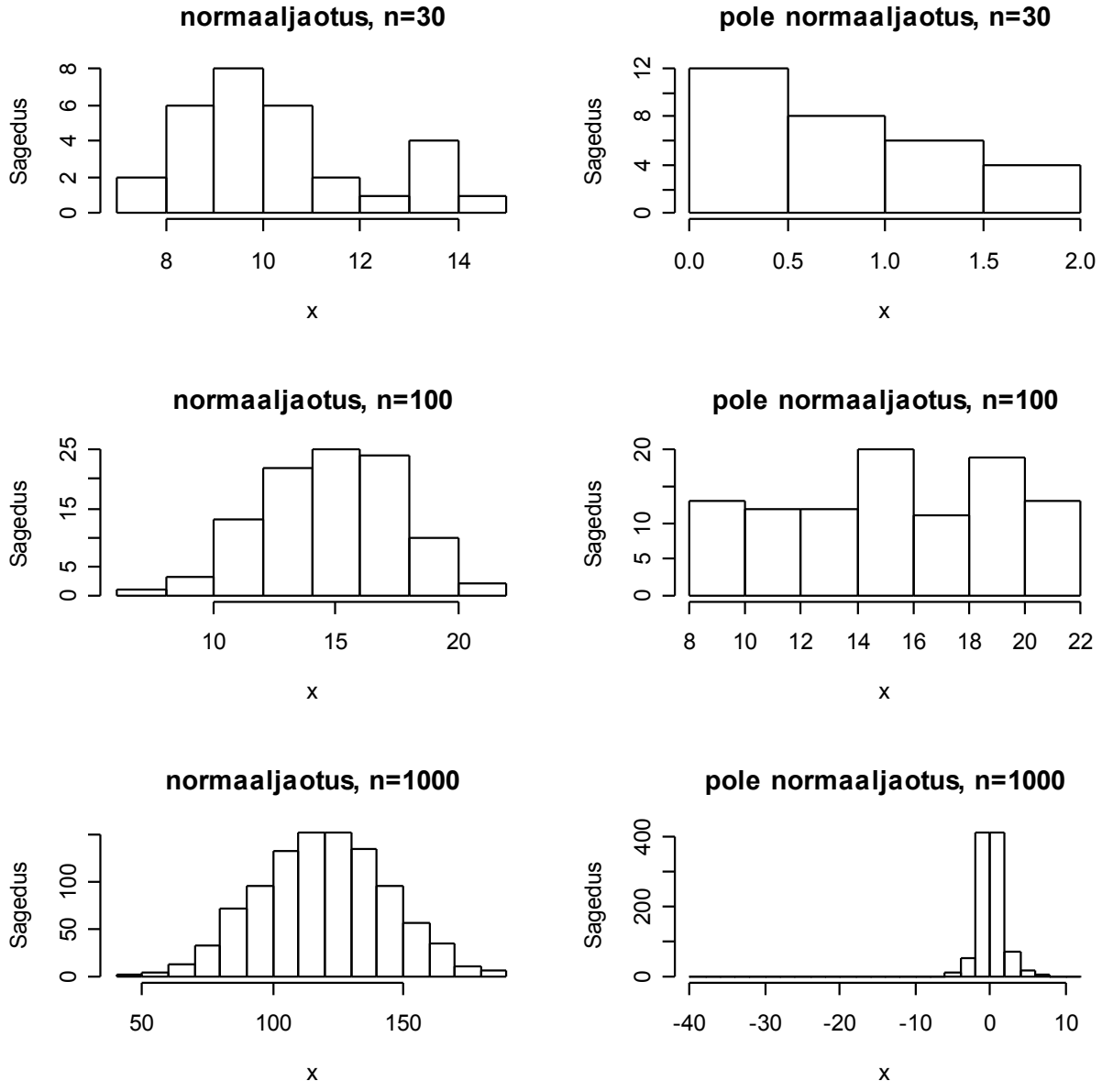


med = ...
$\bar{x}$ = ...
s = ...
0,05-kvantiil = ...
0,95-kvantiil = ...

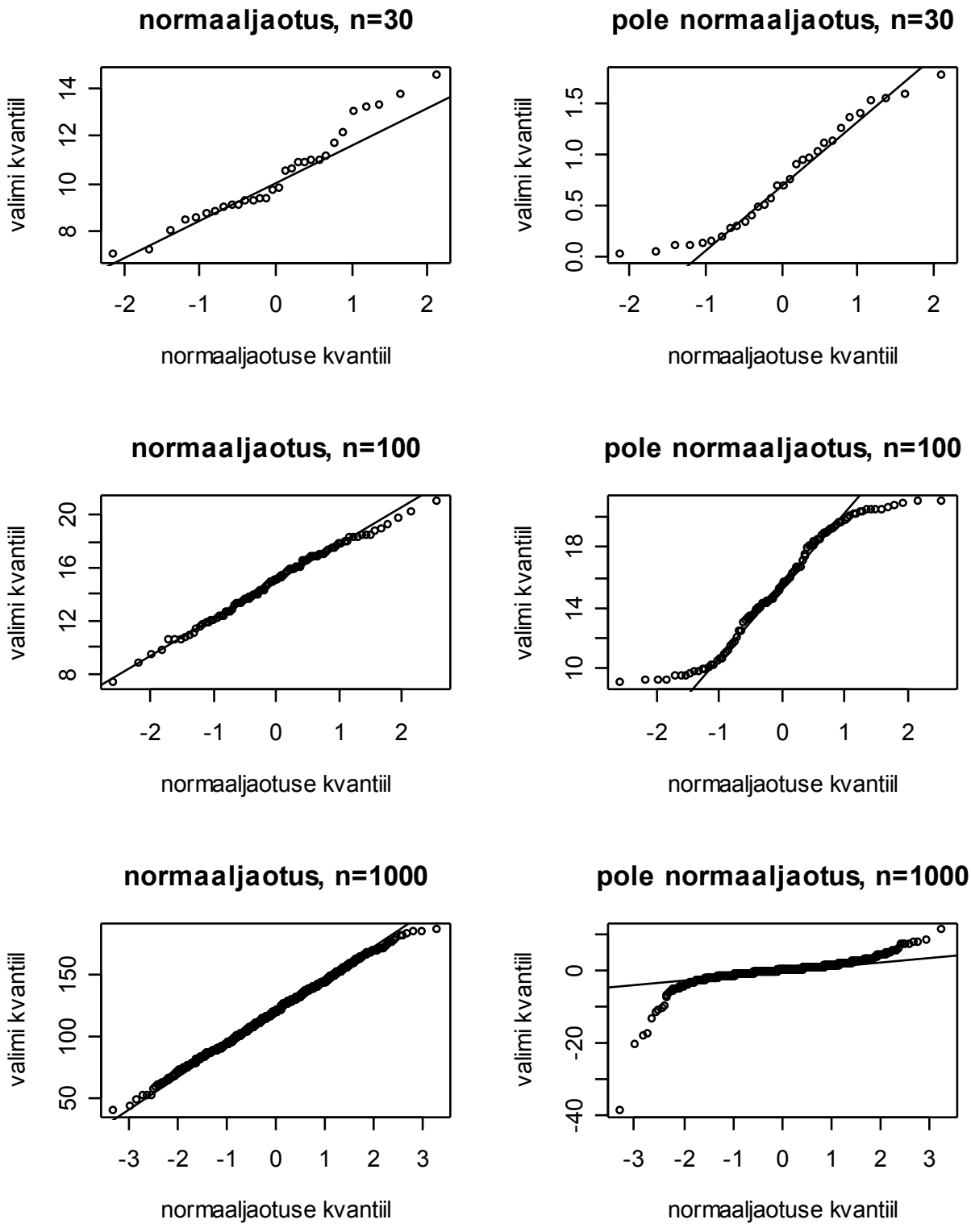
Kas uuritav tunnus on normaaljaotusega?

Vahel tuleb otsustada, kas uuritav tunnus võiks olla normaaljaotusega või mitte. Eriti väiksema valimi korral võib taolisele küsimusele vastamine keerukaks osutuda. Vaatame mõningaid võimalusi sellele küsimusele vastamiseks.

Üheks võimaluseks oleks teha otsus uuritava tunnuse histogrammi põhjal. Näiteid:



Histogramm ei pruugi olla kõige mugavam vahend normaaljaotuse tuvastamiseks. Kasutatakse ka tõenäosuspaberit (quantile-quantile plot, qq-plot) kus moodustatakse valimi kvantiilidest ja normaaljaotuse vastavatest kvantiilidest punktipaarid. Punktid kantakse graafikule ja juhul, kui uuritav tunnus on ka tegelikult normaaljaotusega, peaksid kõik punktid jääma enam-vähem ühele sirgele. Kui punktid kipuvad märgatavalt sirgest kõrvale kalduma, siis pole kardetavasti tegemist normaaljaotusega. Järgneval joonisel on antud tõenäosuspaberid kõigi 6 juhu tarvis, mille jaoks joonistasime ka histogrammid.



Kuidas joonistada tõenäosuspaberit tunnusele X?

Kasuta käske:
`qqnorm(X)` - joonistab punktid
`qqline(X)` - lisab pildile silma abistamiseks võrdlusjoone

Järgnevatele küsimustele vastates kontrolli enne, milliseid graafikuid võid saada, kui uuritav tunnus on tegelikult normaaljaotusega. Seda tee nii:

1. vaata, kui suur on valim (mitut vaatlust oled oma graafiku joonistamiseks kasutanud). Oletame järgmises näiteprogrammis, et vaatluseid oli 42.
2. Tekita 42 vaatlusega valim garanteeritult normaaljaotusega uuritavast tunnusest. Normaaljaotusega populatsioonist võtab soovitud suurusega valimi R-i käsk `rnorm`. Joonista oma genereeritud andmete pealt graafik.
3. Korda sammu 2 mitu korda ja leia, kuivõrd kaugele „normist“ võib sellise suurusega valimi korral joonistatav graafik jääda. Vaata, kas tegelike andmete pealt joonistatud graafik tundub veidram kui genereeritud andmete pealt joonistatud graafikud.

Seda tööd võiks teha järgmine programm. Programmi võiksid esmalt kirja panna tekstiredaktoris (notepad) ja sealt kogu programmi korraga kopeerida R-i.

Histogrammide joonistamine:

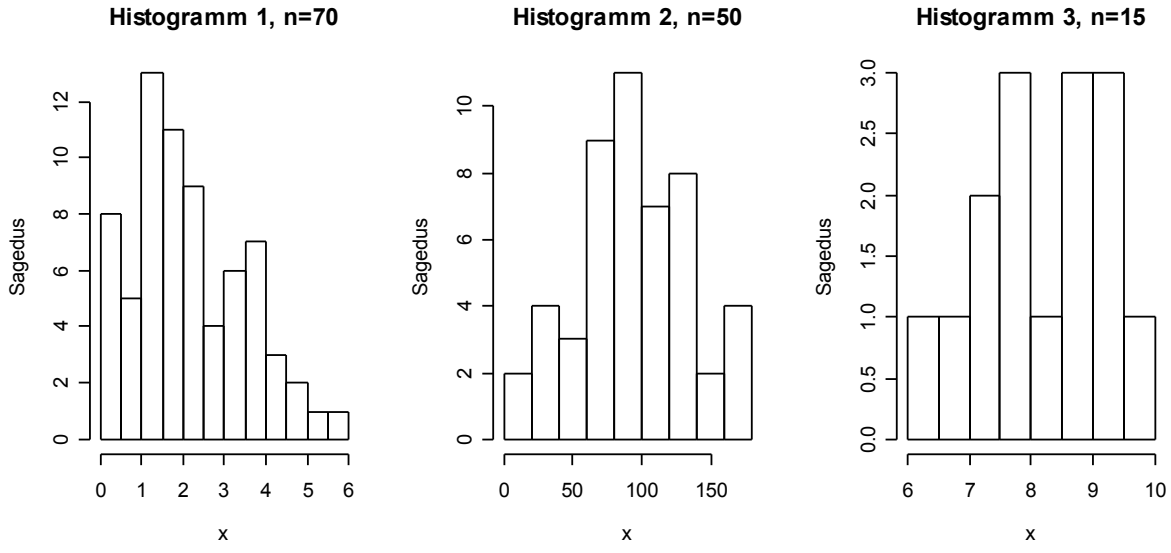
```
par(mfrow=c(2,2))
for (i in 1:4){
  x=rnorm(42)
  hist(x, main="normaaljaotus, n=42")
}
```

Tõenäosuspaberite joonistamine:

```
par(mfrow=c(2,2))
for (i in 1:4){
  x=rnorm(42)
  qqnorm(x, main="normaaljaotus, n=42")
  qqline(x)
}
```

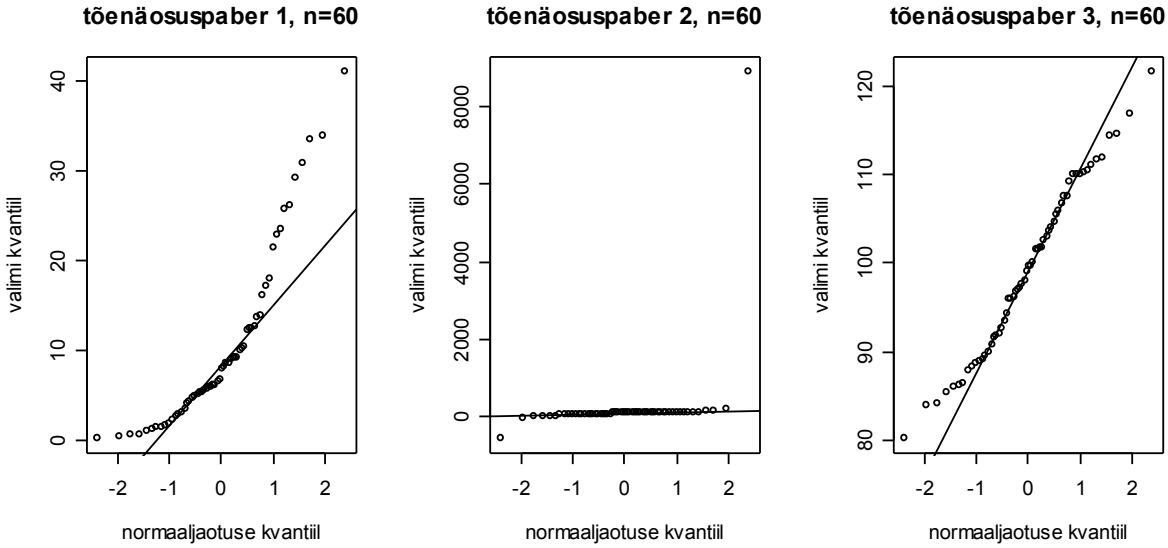
Ülesanne 2

Otsusta, kas järgmiste graafikute puhul võiks olla tegemist normaaljaotusega tunnustega:



Ülesanne 3

Vaata, millistel järgmistest graafikutest võiks olla uuritava tunnuse jaotuseks normaaljaotus?



Ülesanne 4

Loe sisse eelmises praktikumis kasutatud kalade andmestik. Seda sab teha käsuga:

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

Kontrolli, kas ahvenate (*Species*=7) kaal (*Weight*) on normaaljaotusega!

Usaldusintervalli leidmine (2 võimalust)

Loomulikult huvitab ühte asjalikku kalameest küsimus, milline võiks olla järvest püütavate kalade keskmine kaal (ja ega see aastati vähenenud pole, kas kalamehe keskmine on parem/halvem kui „järve" keskmine jne jne). Üritame leida usaldusintervalli järvest püütavate kalade keskmisele kaalule (täpsem väljendusviis: kaalu keskväärtusele). Seda saab teha kahel viisil – kasutades valmismeistertatud vahendeid või ise arvutades. Proovime mõlemat.

Alustuseks proovime usaldusintervalli ise leida. Hiljem kontrollime, kas meie poolt leitud usaldusintervall tuli sama, mis R-i enda valmisprotseduuri poolt pakutav.

(1-*a*)-usaldusintervalli saab leida kasutades valemit

$$\left[\bar{x} + \frac{s}{\sqrt{n}} t_{\alpha/2;df=n-1} \ldots \bar{x} + \frac{s}{\sqrt{n}} t_{1-\alpha/2;df=n-1}\right]$$

kus s on valimi standardhälve, n on valimi suurus ja $t_{a/2;n-1}$ ning $t_{1-a/2;n-1}$ on t-jaotuse $a/2$ ja (1-$a/2$)-kvantiilid. T-jaotuse mingit kvantiili (mingi vabadusastmete arvu korral), näiteks 0,03-kvantiili (kui vabadusastmete arv on 20) saab leida kasutades käsku `qt(0.03, df=20)`. Leiame 95%-usaldusintervalli kalade kaalude keskväärtusele:

standardviga

```
mean(Weight)+qt(0.025, df=158)*sd(Weight)/sqrt(159)
mean(Weight)+qt(0.975, df=158)*sd(Weight)/sqrt(159)
```

Ülesanne 5

Enamasti sellist pool-käsitsi arvutamist ei kasutata – kasutatakse mõnda valmisprotseduuri, R-is funktsiooni t.test. Aga vahel pole meil algandmeid käepärast ja peame usaldusintervalli leidma koondandmeid (artiklis tood põhistatistikuid) kasutades, ja siis on ainsaks võimaluseks eeltoodud arvutused ise läbi teha. Oletame, et teame (kirjandusest): uuritava tunnuse keskmine oli 15,6; standardhälve 4,2 (arvud saadud kasutades 120 vaatlust). Milline tuleb usaldusintervall keskväärtusele? Kas meie teooria väide (keskväärtus on 16,2) on mõeldav varasema uuringu valguses – st. kas meie teoreetiliste arutelude tulemusena leitud väärtus on selline, mis oleks olnud varasema uuringu tulemuste valguses mõeldav väärtus (langeb usaldusintervalli)?

R-is saab kõige väiksema vaevaga usaldusintervalli keskväärtusele leida käsu t.test abil (funktsioon t.test kontrollib hüpoteese keskväärtuse kohta kasutades t-testi – tean, tean, hüpoteeside kontrollimist pole me veel õppinud, aga usaldusintervalli võime ikka ju vaadata?). Näiteks latikate kaalu keskväärtusele saame usaldusintervalli (ja palju muud pahna) järgmisel viisil:

```
> t.test(Weight[Species==1])

        One Sample t-test

data:  Weight[Species == 1]
t = 17.9921, df = 34, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 552.0986 692.7014
sample estimates:
mean of x
    622.4
```

t-jaotuse kvantiilide leidmisel kasutatud vabadusastmete arv (*n*-1)

95%-usaldusintervall latikate kaalu keskväärtusele

Valimi keskmine

Kas antud juhul on võimalik, et latikate kaalu keskväärtus on 575g?

Vahel soovime 0,95-usaldusintervalli asemel kasutada mõnda teist, näiteks 0,9-usaldusintervalli. Sellist soovi saame R-le edastada kasutades t.testi lisaparameetrit conf.level, näiteks latikate kaalu keskväärtusele saame 0,9-usaldusintervalli käsuga

```
t.test(Weight[Species==1], conf.level=0.9)
```

Usaldusintervalli interpretatsioonist

Valimi pealt leitud 95%-usaldusintervalli kohta pole päris korrektne öelda, et 95%-tõenäosusega asub meid huvitava parameetri tegelik väärtus usaldusintervallis. Iga usaldusintervall on kas õige (tegelik väärtus asub usaldusintervallis) või väär (tegelik väärtus ei asu usaldusintervallis). Pigem saame öelda, et 95%-usaldusintervall on leitud meetodi abil, mis tagab 95% valimite korral õige usaldusintervalli. Proovime seda katseliselt!

Käsk

```
valim=rnorm(100, mean=3.2)
```

võtab juhusliku valimi (n=100) normaaljaotusega populatsioonist, mille keskväärtus on 100. Selle valimi keskmine pole enamasti täpselt 3,2:

```
mean(valim)
```

küll aga peaks 3,2 jääma teist enamikul 95%-usaldusintervalli:

```
t.test(valim)
```

või

```
t.test(valim)$conf.int
```

Kui te võtaksite sellestsamast populatsioonist 100 valimit, ja neist igaühe jaoks leiaksite usaldusintervalli peaks saadud usaldusintervallidest 95% olema sellised, millesse kuulub populatsiooni tegelik keskväärtus (proovi umbeski aru saada, mida alltoodud programm teeb):

```
otsus=rep(NA, 100)
for (i in 1:100){
  valim = rnorm(100, mean=3.2)
  abi = t.test(valim)$conf.int
  if ((abi[1]<3.2)&(abi[2]>3.2)) otsus[i]="õige" else
     otsus[i]="väär"
  print(paste("[",abi[1],"...",abi[2],"]  -- ",otsus[i]))
}

table(otsus)
```

Kui palju õigeid 95%-usaldusintervalle tuli Sinul 100 korraldatud uuringu kohta?

Usaldusintervall binaarsele tunnusele (soole)

Enamasti eeldatakse, et valimi keskmise jaotus on ligilähedaselt normaaljaotus (vähegi suurema valimi korral). Üks suhteliselt halb alguspunkt on, kui esialgne uuritav tunnus on binaarne – kahe võimaliku väärtusega tunnus. Sellisel juhul läheb vaja suhteliselt suurt valimit, enne kui me saame öelda, et valimikeskmise jaotus on ligilähedaselt normaaljaotus. Õnneks on olemas ka funtsioon, mis oskab otse binaarse tunnuse keskväärtusele (või „õnnestumise tõenäosusele“) täpset usaldusintervalli leida. Näiteks kalade soole saab 95%-usaldusintervalli leida järgmiselt (või täpsemalt emase kala saamise tõenäosusele):

```
binom.test(table(Sex))
```

või nii:

```
binom.test(55,72)
```