Biomeetria bioloogidele
4. praktikum

Usaldusintervall. Hinnangu juhuslikkus.

Vaatame taas Tartu tudengite küsitlemise teel kogutud andmeid (taustainformatsiooni vaata 1. praktikumi materialist):

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

Tunneme huvi tudengite keskmise kaalu kohta. Milline see võiks olla?

```
attach(kokku)
mean(kaal, na.rm=T)
```

Paraku teame, et valimi keskmine ei lange peaaegu kunagi kokku populatsiooni keskmisega ehk keskväärtusega. Sestap on sageli mõistlikum raporteerida valimi keskmise asemel (või temale lisaks) usaldusintervall – näidata väärtuste vahemik, mis üsna kindlasti hõlmab tegeliku keskväärtuse:

```
> t.test(kaal)

        One Sample t-test

data:  kaal
t = 135.7687, df = 656, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 62.18395 64.00905
sample estimates:
mean of x
  63.0965
```

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

95%-usaldusintervall kaalu keskväärtusele

valimi keskmine

Näeme, et Tartu Ülikooli tudengite (täpsemalt öeldes: Tartu Ülikooli arstiteaduskonna 2. aasta tudengite...) tegelik keskmine kaal on kuskil vahemikus 62,1 kg ... 64,1 kg (Vähemalt 95%-kindlusega võime nii öelda). Pane tähele! Usaldusintervalli raporteerides on üldiselt viisakas ümardada usaldusintervall laiemaks!

Proovi ka järgmist käsku, mis leiab 99%-usaldusintervalli keskväärtusele:

```
t.test(kaal, conf.level=0.99)
```

Saadud usaldusintervall on (.....; .....).

Kas 99%-usaldusintervall tuli laiem või kitsam kui 95%-usaldusintervall? Miks?

Tee ise!

Leia 95%-usaldusintervall naistudengite keskmisele kaalule: (.....; .....)

Iga tarkvara esmakordsel kasutamisel võiks proovida, kas arvuti ikka arvutab usaldusintervalli õieti. Proovime korra ka meie.

(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 kõigi tudengite kaalu keskväärtusele:


```
mean(kaal, na.rm=TRUE)+qt(0.025, df=657-1)*sd(kaal, na.rm=T)/sqrt(657)
mean(kaal, na.rm=TRUE)+qt(0.975, df=657-1)*sd(kaal, na.rm=T)/sqrt(657)
```

Kas tuli sama tulemus mis esialgu?

Muuseas, kust tuli valimi suurus 657? Uuritud oli ju 661 tudengit?

```
> length(kaal)
[1] 661
```

Põhjus on lihtne – neli tudengit keeldusid oma kaalu kohta informatsiooni jagamast. Tegelikke, arvutuskõlblikke väärtuseid oli kõigest 657:

```
> sum(is.na(kaal)==FALSE)
[1] 657
```

Arutle!
Kas oma kaalu avaldamisest keelduvate tudengite kaalude jaotus on samasugune, kui vastanute kaalude jaotus? Kui vastuseks on jah, siis võime käsitleda järgijäänud 657-tudengit jätkuvalt kui juhuslikku valimit tudengitest ja teha arvutusi niimoodi, nagu me äsja tegime – justnagu olekski meil kõigest 657 küsitletut olnud.

Mida aga teha siis, kui me ei usu, et vastamisest keeldusid just tüüpilised tudengid? Appi võiks tulla sensitiivsusanalüüs – vaatame, milliste tulemusteni oleksime jõudnud, kui puuduvad kaalud oleksid olnud erakordselt väikesed või erakordselt suured. Näiteks kui kõik puuduvad väärtused oleksid olnud samasuured, kui olemasolevate andmete maksimum/miinimum? Maksimumi ja miinimumi saame leida näiteks nii:

```
> range(kaal, na.rm=T)
[1]  42 117
```

Kui vastamisest keelduksid ekstreem-anorektikud, siis oleksime tulemuseks saanud:

```
> t.test(c(kaal, 42, 42, 42, 42))
95 percent confidence interval:
 62.05324 63.88443
sample estimates:
mean of x
 62.96884
```

Ja kui vastamisest oleksid keeldunud paksukese-poolsed inimesed, siis oleks tulemuseks tulnud:

```
> t.test(c(kaal, 117, 117, 117, 117))
95 percent confidence interval:
 62.46105 64.38434
sample estimates:
mean of x
 63.42269
```

Valimikeskmise võimalike väärtuste piirkonda 62,97..63,42 kutsutakse vahel ka ignorantsuspiirkonnaks (*region of ignorance*). Kui asendame andmestikus olevaid puuduvaid väärtuseid mõeldavate/võimalike väärtustega, siis saame ka erinevaid usaldusintervalle. Võime välja noppida usaldusintervallide alumistest piiridest väikseima ja usaldusintervallide ülemistest piiridest suurima. Saame tulemuseks nn *region of uncertainty*: (62,05...64,39).

Märkus: Kuigi sensitiivsusanalüüse on juba mõnda aega tehtud, pole mõisted „region of ignorance“ ja „region of uncertainty“ siiski praktikute sees kuigi levinud, seega kasuta neid oma artiklis ettevaatlikult – lisa selgitus, mida nad tähendavad või kirjuta lihtsalt, et sensitiivsusanalüüs ei muutnud tulemuste interpretatsiooni vms.

Ülesanne

Leia 95%-usaldusintervall luu mineraalsele tihedusele (*BMD, bone mineral density*):

	ALSPAC (discovery)	
	n = 999	
Age, years	15.4	(0.22)
Men, no. (%)	466	(47%)
Height, cm	169.5	(8.2)
Weight, kg	61.0	(10.7)
Position of cortical section from distal end of tibia	50%	
cortical BA, mm^2	300.5	(48.3)
cortical BMC, mg	330.0	(50.5)
cortical BMD, mg/cm^3	1100.0	(38.1)
cortical Th, mm	5.40	(0.65)
cortical PC, mm	72.6	(6.0)
cortical EC, mm	38.7	(5.8)
Position of trabecular section from distal end of tibia	NA	
trabecular BMD, mg/cm^3	NA	
Total body BMD, g/cm^2	n = 4003	1.03
Femoral neck BMD, g/cm^2	n = 3328	0.98
Lumbar spine BMD, g/cm^2	NA	

Values are mean(SD), unless otherwise stated.
BA = bone area, BMC = bone mineral content, BMD = bone mineral density, Th = th
doi:10.1371/journal.pgen.1001217.t001

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):

```
# Teeme valmis 100 pesa, kuhu kirja panna hinnanguid 100 uuringu kohta
otsus=rep(NA, 100)

# Tsükkel mis teeb 100 uuringut ja hindab nende uuringute tulemuste
# õigsust ehk paikapidavust
for (i in 1:100){

  print(paste("Uuring nr", i))

  # Võtame valimi (n=100) normaaljaotusega populatsioonist, EX=3,2.
  valim = rnorm(100, mean=3.2)

  # Leiame 95%-usaldusintervalli
  abi = t.test(valim)$conf.int
  print(paste("Usaldusintervall:", abi[1], "...", abi[2]))

  # Kui tegelik keskväärtus sattub usaldusintervalli, siis peame
  # uuringu tulemust õigeks, muidu on tulemus vale.
  if ((abi[1]<3.2)&(abi[2]>3.2)) otsus[i]= "õige" else otsus[i]= "väär"
  print(otsus[i])

  # Lihtsalt tühi rida tulemuste vahele
  print("", quote=FALSE)
}

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 R'is ka funtsioon, mis oskab binaarse tunnuse keskväärtusele (või „õnnestumise tõenäosusele") täpset usaldusintervalli leida. Näiteks soovime kirjeldada, kui paljud tudengid on vajanud viimase 2 aasta jooksul kiirabi abi.

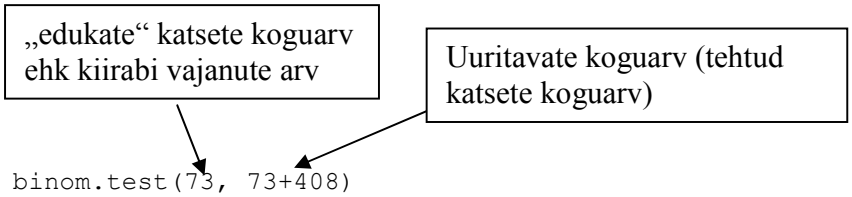
Leiame esmalt ligikaudse usaldusintervalli (mis eeldab, et valimi keskmine käitub juba kui normaaljaotusega juhuslik suurus):

```
> t.test(kiirabi)
95 percent confidence interval:
0.1195883 0.1839460
sample estimates:
mean of x
0.1517672
```

Täpse usaldusintervalli saame leida kasutades binom.test-käsku. Esmalt aga vaatame, kui palju oli abivajajaid:

```
> table(kiirabi)
kiirabi
  0   1
408  73
```

Leiame saadud andmete pealt täpse usaldusintervalli:




```
> binom.test(73, 73+408)

        Exact binomial test

data:  73 and 73 + 408
number of successes = 73, number of trials = 481, p-value < 2.2e-16
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.1208868 0.1870019
sample estimates:
probability of success
             0.1517672
```

Täpne 95%-usaldusintervall 2-aasta jooksul kiirabi abi vajamise tõenäosusele

Kiirabi abi vajavad 2-aasta jooksul hinnanguliselt 15% (12%..19%) tudengitest.

Antud juhul tulid mõlemad usaldusintervallid – ligikaudne normaaljaotust eeldanud usaldusintervall ja täpne usaldusintervall protsendile - siiski suhteliselt sarnased, sest valim oli suur. Väikese valimi korral (näiteks kümnekonna vaatluse puhul) võivad need usaldusintervallid tulla märkimisväärselt erinevad. Sellisel juhul tuleks eelistada täpset usaldusintervalli protsendile.

Iseseisev ülesanne
Loe R´i andmestik "seeme.txt" (tegemist on tekstifailiga, seega peame andmed importima read.table-käsu abil:

```
seeme=read.table("http://www.ms.ut.ee/mart/biomeetria2012/seeme.txt",
            header=TRUE)
```

Soovitakse teada, kas kuivatis kuivatatud seeme annab parema saagi kui põllupeal kuivanud vili (tegemist on Inglismaal kogutud andmetega). Väikesed katsepõllud jagati kaheks, ühele poolele külvati põllul kuivanud vilja, teisele poole aga kuivatis kuivanud vilja. Tunnus HAR iseloomustab põllul kuivanud vilja saagikust (nael/aaker, nael - *lbs* - on u 0,45 kg; aaker on u. 0,4 ha; korruta 1,12 saamaks kg/hektarilt); tunnus KUIV kuivatis kuivatatud vilja saagikust.

Vaata andmeid:

```
seeme
attach(seeme)
boxplot(HAR,KUIV,names=c("Harilik seeme","Kuivatatud"))
```

Leia 95%-usaldusintervalli keskmisele põllul kuivanud vilja saagikusele Inglismaal

.....

Arutle, kummal juhul tuleks usaldusintervall laiem – kas siis, kui kõik katsepõllud paikneksid ühe talu maadel, või siis, kui katsepõllud oleksid juhuslik valim kõigist inglismaa põldudest? Kuidas muutuks tulemuste interpretatsioon?

Arvuta saagikuse erinevus samadel katsepõldudel:
`Vahe=KUIV-HAR`

Leia 95% usaldusintervall leitud vahede keskväärtusele. Kuidas interpreteerid tulemust? Kas 0 kuulub usaldusintervalli? Kui kuulub, siis mida see tähendab? Kui palju muutuks põlluvilja keskmine saagikus (naela/aaker) Inglismaal, kui asendaksime kõigil põldudel põllul kuivanud vilja kuivatis kuivatatud viljaga?

a) Kas võib öelda "95%-tõenäosusega asub populatsiooni tegelik keskväärtus (saagikuste muutuste keskväärtus) vahemikus [-10.7..78.2]"?
b) Mis võiks usaldusintervalliga juhtuda, kui uuritavad põllulapid oleksid olnud veel väiksemad? Miks?
c) Mis arvatavasti juhtuks usaldusintervalliga, kui oleksime teinud mõõtmisi rohkematel põldudel?

Usaldusintervalli kajastamine graafikul

Üks võimalus usaldusintervalli lisamiseks graafikule on kasutada mõne valmiskujul olemasoleva funktsiooni abi (neid on mitmeid erinevaid). Proovime esmalt lisamoodulis gplots paiknevat funktsiooni plotmeans.

Juhul kui antud arvutis pole varem lisamoodulit gplots kasutatud, tuleb vastav lisamoodul esmalt arvutisse paigaldada (tarvis teha vaid ühel korral – uuesti R'i käivitades ei pea seda käsku uuesti andma):

```
install.packages("gplots")
```

Võtame lisamooduli gplots kasutusele:

```
library(gplots)
```

ja joonistame graafiku (kiirabi abi vajanud ja mittevajanud tudengite keskmine pikkus koos 95%-usaldusintervalliga):

```
plotmeans(pikkus~kiirabi)
```

või veidi ilusamal kujul esitatult:

```
plotmeans(pikkus~kiirabi, xaxt="n", xlab="On vajanud kiirabi abi?",
      ylim=c(167, 173))
axis(1, at=c(1,2), c("ei","jah"))
```

Samas võib usaldusintervalli lisada ka mõnele sinu poolt tehtud joonisele. Vaata järgmist näidet ja proovi aru saada, mis toimub.

```
abi1=barplot(prop.table(table(sugu))*100, ylim=c(0, 100), ylab="%",
      names.arg=c("Naistudengid","Meestudengid"))
table(sugu)
binom.test(512,512+149)
binom.test(512,512+149)$conf.int
abi2=binom.test(512,512+149)$conf.int*100
arrows(abi1[1], abi2[1], abi1[1], abi2[2], angle=90, code=3)
```

Kuidas muuta programmi selliselt, et saaksime ka poisslaste protsendile 95%-usaldusintervalli graafikule?