Biomeetria
4. praktikum

Usaldusintervall ja prognoosiintervall

Loeme sisse andmestiku fishcatch.dat:

andmed=read.table("http://www.ms.ut.ee/mart/biomeetria2007/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.

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:

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

Ülesanne:

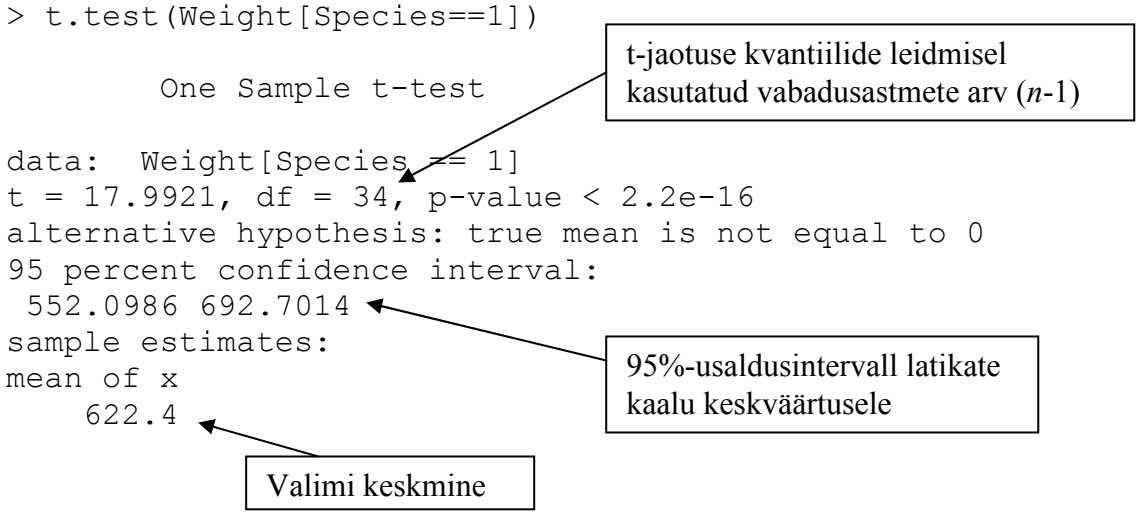
1. Leia 90% usaldusintervall järvest püütud kalade keskväärtusele. Kas see tuleb laiem/kitsam kui 95%-usaldusintervall?
2. Enamasti sellist pool-käsitsi arvutamist ei kasutata – kasutatakse mõnda valmisprotseduuri, R-is funktsiooni t.test. Aga vahel pole meil algandmeid käepärast ja peame usaldusintervalli 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?

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
```



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)
```

Aga milline on järgmise kinnipüütava latika kaal? Kas oskad leida 95%-prognoosiinteralli?

Esmalt tuleks kontrollida, kas latikate kaalud on ikka normaaljaotusega (kui usaldusintervalli võime suure valimi korral leida ülaltoodud viisil sõltumata sellest, millise jaotusega on uuritav tunnus, siis prognoosiintervalli valem kehtib vaid siis, kui uuritava tunnuse jaotuseks on tõepoolest normaaljaotus). Kui on, siis võime prognoosiintervalli leidmiseks kasutada kas valemit

$$\left[\mu + \sigma z_{\alpha/2} \ldots \mu + \sigma z_{1-\alpha/2}\right]$$

kus $z_{a/2}$ on standardse normaaljaotuse $a/2$-kvantiil ja $z_{1-a/2}$ on (1-a/2)-kvantiil (R'is leitavad käsu `qnorm` abil). Parem veel oleks kasutada järgmist valemit (mis võtab arvesse, et me ei tea keskväärtust ega standardhälvet täpselt:

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

Milline tuleb 95%-prognoosiintervall latika kaalule? Kas see tuleb kitsam või laiem kui 95%-usaldusintervall? Miks?

Esimeses praktikumis uurisime Tartu Ülikooli (meditsiini)tudengite andemstikku. Loe see sama andmestik uuesti sisse:

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

Ja leia nais- (sugu=1) ja meestudengite (sugu=2) pikkuste jaoks 95%-usaldusintervallid ja 95%-prognoosiintervallid järgmise nais- ja meestudengi pikkusele. Arutle samuti, kas toodud arvutusvalemeid üldse võib kasutada – kas usaldusintervalli ja prognoosiintervalli leidmise valemite eeldused on täidetud.