Biomeetria

5. praktikum

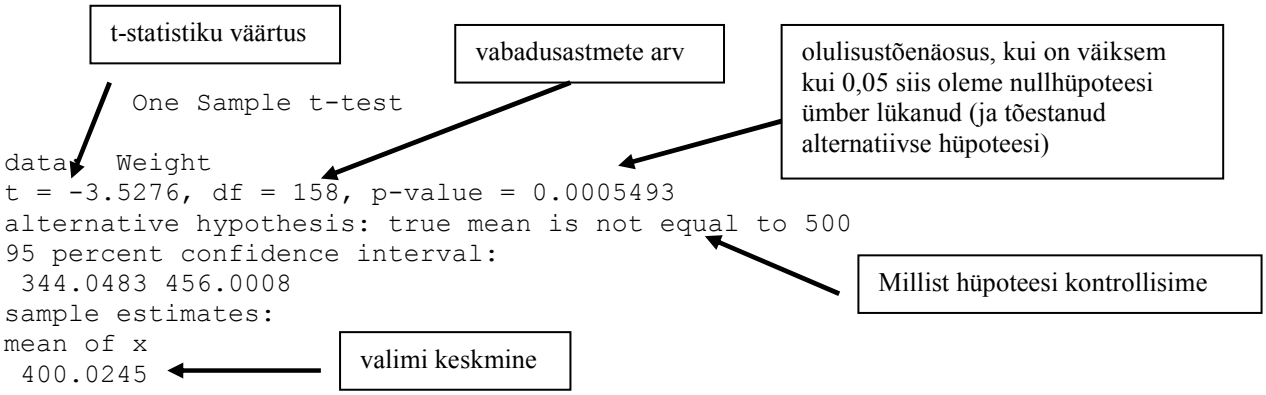
1. Loe R´i andmestik "fishcatch.dat". Selleks anna käsk:
   andmed=read.table("http://www.ms.ut.ee/mart/biomeetria2007/fishcatch.dat", header=TRUE)
2. Lühikeste tunnuste nimede kasutamiseks anna käsk
   `attach(andmed)`

T-test hüpoteeside kontrollimiseks (ühe) populatsiooni keskväärtuse kohta

Meenutamaks eelmist praktikumi kontrolli järgmiseid hüpoteese

a) Antud järvest püütavate kalade kaalude keskväärtus on pool kilo
   H0: EX=500
   H1: EX<>500
Antud hüpoteesipaari saab kontrollida funktsiooni t.test abil:

```
> t.test(Weight,mu=500)
```



Antud juhul peame tõdema, et väide, nagu oleks järvest püütud kalade kaalude keskväärtus pool kilo on väheusutav.

b) Kontrolli, kas antud järvest püütud siigade (tunnuse *Species* väärtus 2) kaalude keskväärtus on pool kilo?

Erinevad t-testid

Algset t-testi on mitut moodi kohandatud, kolm neist variantidest on väga sageli kasutatud:

- T-test hüpoteeside kontrollimiseks keskväärtuse kohta – kontrollib mingit väidet ühe populatsiooni (ühe tunnuse) keskväärtuse kohta. Kontrollitavaks hüpoteesiks on näiteks väide H_0: EX=10; H_1: EX<>10.
- T-test keskväärtuste võrdlemiseks, sõltumatud valimid. Sobib kasutamiseks, kui soovime võrrelda, kas kahe uuritava populatsiooni keskväärtused on võrdsed või mitte. Näiteks tahame teada saada, kas naiste ja meeste keskmine palk on sama või mitte. Või kui tahame uurida, kas genotüübiga AA isenditel on keskmine kolesteroolitase samasugune kui genotüübiga aa isenditel (või mitte). Kontrollitavaks hüpoteesideks on väited H_0:EX=EY; H_1: EX<>EY
- T-test keskväärtuste võrdlemiseks, sõltuvad valimid. Kontrollitakse ikka väidet, kas kahes grupis on uuritava tunnuse keskväärtus sama või mitte, näiteks uuritakse, kas alkoholi tarbinute keskmine reaktsioonikiirus on sama, mis alkoholi mittetarbinutel. Eelmisest olukorrast on erinev vaid see, et samad objektid kuuluvad mõlemasse gruppi – samal inimesel mõõdetakse reaktsioonikiirust enne ja pärast alkoholi tarbimist. Või mõõdetakse mingi ensüümi aktiivsust igal loomal 2 ja 12 päeva peale sündi ja tahetakse teada, kas antud ensüümi (keskmine) aktiivsustase on muutunud - ka

sellisel juhul on tegemist sõltuvate valimitega (2. päeva mõõtmised ja 12. päeva mõõtmised on tehtud samadel loomadel).

Vaatamegi alljärgnevalt, kuidas teha t-testi kahe populatsiooni keskväärtuste võrdlemiseks (nii sõltuvate kui ka sõltumatute valimite korral)

T-test sõltumatute valimite jaoks

Võimaldab kontrollida, kas kahe populatsiooni keskväärtused on võrdsed või mitte:

H0: EX=EY
H1: EX<>EY

Mainitud hüpoteese saab kontrollida samuti kasutades käsku t.test, näiteks võime kontrollida, kas siigade (*Species* = 2) kaalude keskväärtus erineb särgede (*Species* = 3) kaalude keskväärtusest:


```
> t.test(Weight[Species==2], Weight[Species==3])

        Welch Two Sample t-test

data:  Weight[Species == 2] and Weight[Species == 3]
t = 2.924, df = 5.212, p-value = 0.03127
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  49.11858 697.78142
sample estimates:
mean of x mean of y
   531.00    157.55
```

Taas saame tulemuseks olulisustõenäosuse, mis on väiksem kui 0,05. Järelikult on siigade ja särgede kaalude keskväärtused erinevad. Seega oleme kaudselt leidnud kinnitust oma kahtlusele, et tegemist võivad olla eri liiki kaladega.

Vasta järgmistele küsimustele:
Kas latikate (Species = 1) ja siigade (Species = 2) kaalude keskväärtused on erinevad?
Aga kaalude keskmised?

T-test sõltuvate valimite jaoks

Loe R´i andmestik "seeme.txt". Selleks:

seeme=read.table("http://www.ms.ut.ee/mart/biomeetria2007/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 (lbs/aaker, nael on u 0,45 kg; aaker on u. 0,4 ha; korruta 1,12 saamaks kg/hektarilt); tunnus KUIV kuivatis kuivatatud vilja saagikust.

Moodustame uue tunnuse, mis näitab erinevalt kuivatatud seemnete viljakuse erinevust samal põllulapil:

```
VAHE = KUIV-HAR
```

Kui viljakus ei sõltu seemnete kuivatamise viisist, siis saagikus peaks olema ligikaudu samasuur. Erinevused saagikuses oleksid siis tingitud pigem sellest, et sama põllu kaks poolt ei pruugi olla täiesti samaväärsed kasvutingimuste poolest, samuti võib ühele põllupoolele ka muidu mõni hakkajam taim sattuda, mis samuti tasakaalu rikub. Aga üle kõigi erinevate juhuslikult poolitatud põldude peaks keskmine erinevus tulema 0, ehk tunnuse VAHE keskväärtus peaks sellisel juhul olema 0. Matemaatiliselt kirja pandult:

H0: E(VAHE)=0 (kuivatamisviis ei mõjuta keskmist saagikust)
H1: E(VAHE)<>0 (kuivatamisviis mõjutab keskmist saagikust)

Selliseid hüpoteese saab kontrollida t-testi abil:

```
> t.test(VAHE)

One Sample t-test

data:  VAHE
t = 1.6905, df = 10, p-value = 0.1218
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -10.72710  78.18164
sample estimates:
mean of x
 33.72727
```

Kus *t = 1.6905* on t-statistiku väärtus, df=10 näitab kasutatud vabadusastmete arvu, p-value on olulisustõenäosus.

Kas olulisuse nivool 0,05 saab järeldada, et põllu viljakus sõltub seemnete kuivatamise viisist?

Kas võib tõestatuks lugeda, et viljakus ei sõltu seemnete kuivatamise viisist?

Üks lugupeetud teadlane on ütelnud, et kuivatis kuivatatud seemnete saagikus on keskmiselt 100 naela/aakri kohta suurem kui põllul kuivanud vilja viljakus. Kas meie andmed kinnitavad selle maineka teadlase arvamust? Püstitame hüpoteesid:

H0: E(VAHE)=100 (keskmine erinevus saagikuses on 100naela/aaker)
H1: E(VAHE)<>100

Kontrollime neid hüpoteese:

```
> t.test(VAHE, mu=100)
```

Kas saame tõestada, et teadlasel oli õigus?

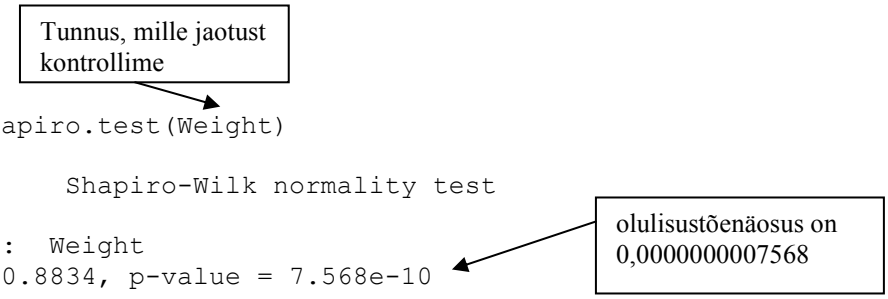
Kas saame tõestada, et teadlane eksis?

R' is koos lisamoodulitega on võimalik kasutada üle 100 erineva statistilise testi, millistest paljud on välja mõeldud üsna spetsiifiliste hüpoteeside kontrollimiseks. Alljärgnevalt mainime paari testi nende sajakonna seast:

Shapiro-Wilk'i test (normaaljaotuse eelduse testimine)

Kui valim on väike, siis tagab t-testi õiged arvutustulemused vaid siis, kui uuritav tunnus on normaaljaotusega. Järvest on aga välja püütud vaid 6 siiga, tegemist on selgelt väikese valimiga. Peaksime kontrollima, kas siigade kaalude jaotus võiks olla normaaljaotusega. Võisime seda teha graafiliselt, aga vastavaid hüpoteese on võimalik kontrollida ka statistilise hüpoteeside testimise protseduuri abil. Shapiro-Wilk'i testi võib kasutada kontrollimaks, kas pideva tunnuse jaotuseks võiks olla mõni normaaljaotus. Kontrollime näiteks, kas kalade kaalude jaotus on normaaljaotus:

H_0: Tunnuse *Weight* jaotus on (mingi) normaaljaotus, $Weigth \sim N(\mu; \sigma^2)$
H_1: Tunnuse *Weight* jaotus pole normaaljaotus




```
> shapiro.test(Weight)

        Shapiro-Wilk normality test

data:  Weight
W = 0.8834, p-value = 7.568e-10
```

Antud juhul näeme, et olulisustõenäosus tuleb väiksem kui 0,05, seega oleme nullhüpoteesi kummutanud (nullhüpotees ei saa paika pidada – antud tunnuse jaotuseks pole küll normaaljaotus).

Märkus: tasub teada, et eriti suuremate valimite korral võib Shapiro-Wilk'i test olla mõttetult range. Halvimal juhul võib ta normaaljaotuse eelduse kummutada (suure valimi korral) imeväikeste probleemide pärast (näiteks mõõtmistulemuste ümmardamine). Sellised imepisikesed probleemid enamasti ei muuda mistahes normaaljaotust eeldava meetodi tulemusi kuigivõrd. Sestap vahel soovitatakse eelistada suuremate valimite korral ikka normaaljaotuse graafilist kontrollimist – inimsilmal on nagunii raske selliseid tähtsusetuid kõrvalekaldeid normaaljaotusest tuvastada ja see ongi hea. Enamasti on normaaljaotust eeldavad meetodid robustsed – kui vaatlused on peaaegu normaaljaotusega, siis kõlbab neid meetodeid ikka veel kasutada. Matemaatilist täpsust pole antud eelduse kontrollimisel niiväga vajagi ☺.

Ülesanne
Kontrolli, kas siigade (Species = 2) kaalude jaotus võiks olla normaaljaotus?

Uuritava tunnuse hajuvust kontrollivad testid

Uurivad tunnuse hajuvust võrdlevad testid võiksid näiteks kasulikud olla sama liigi asurkondade uurimisel – kui antud loomad ühes asurkonnas on väga sarnased võrreldes teiste asurkondadega, siis võib põhjuseks olla hiljutine „pudelikaela" läbimine, mis geneetilist varieeruvust on tugevalt vähendanud. Samuti eeldavad mõned statistilised meetodid, et uuritava tunnuse hajuvus kõigis uuritavates populatsioonides on samasuur – ka selle eelduse paikapidavust saab kontrollida mainitud testide abil.

Bartlett´ test
bartlett.test – (F-testi edasiarendus) – kontrollib, kas uuritava tunnuse hajuvus on vaadeldavates populatsioonides samasugune või mitte:

H_0: $\sigma_1^2 = \sigma_2^2$
H_1: $\sigma_1^2 \neq \sigma_2^2$

Eeldab, et uuritavad tunnuse jaotus (kõigis populatsioonides) on normaaljaotus. Kasutusnäide:

```
bartlett.test(Height, Species)
```

Kontrollime eelduse täidetust (kas kõigi liikide puhul võiks uuritava tunnuse jaotus olla normaaljaotus):

```
par(mfrow=c(3,3))
by(Height, Species, qqnorm)
```

Fligner-Killeen'i test
fligner.test - kontrollib, kas uuritava tunnuse hajuvus on vaadeldavates populatsioonides samasugune või mitte. Ei nõua, et uuritava tunnuse jaotus oleks normaaljaotus. Kasutusnäide (kalade andmestiku puhul):

```
fligner.test(Weight, Species)
```

Sündmuse toimumise tõenäosuse kohta käivate hüpoteeside kontrollimine

Suure valimi korral võib hüpoteeside kontrollimisel tõenäosuse(te) kohta kasutada ka t-testi. Väikese valimi korral seda teha ei tohi (uuritava tunnuse jaotus pole ju normaaljaotus). Sestap on olemas spetsiaalne test hüpoteeside kontrollimiseks tõenäosuste kohta, mida võib kasutada väikeste valimite korral ja mis annab täpsemaid tulemusi kui t-test ka suurte valimite korral. Selle testi nimeks on binom.test. Näide binom.test-i abiinfost:

```
## Conover (1971), p. 97f.
## Under (the assumption of) simple Mendelian inheritance, a cross between plants of two particular
##  genotypes produces progeny 1/4 of which are "dwarf" and 3/4 of which are "giant", respectively.
##  In an experiment to determine if this assumption is reasonable, a cross results in progeny having
##  243 dwarf and 682 giant plants. If "giant" is taken as success, the null hypothesis is that
##  p =3/4 and the alternative that p != 3/4.
binom.test(c(682, 243), p = 3/4)
binom.test(682, 682 + 243, p = 3/4)   # The same.
## => Data are in agreement with the null hypothesis.
```

Sama test võrdluseks t-testi abil tehtuna:

```
t.test( rep(c(1,0),c(682, 243)) , mu = 3/4)
```