Lineaarsed mudelid

Praktikum

Testi võimsus

Vaatame loengu ülesande näitel kolme erinevat võimalust testi võimsust arvutada ning arutleme, millal eelistada ühte ja millal mõnda teist meetodit.

Näide 1

Kavatseme uuringusse kaasata 30 inimest Lõuna-Eestis, 30 inimest Kesk-Eestis, 30 inimest Põhja-Eestis. Oletame, et uuritava tunnuse keskväärtus on Lõuna-Eestis 100, Kesk-Eestis 110, Põhja-Eestis 120. Uuritava tunnuse standardhälve Lõuna-Eestis on 20, oletame et ka muudes piirkondades võiks uuritava tunnuse standardhälve olla sarnane.

Kui tõenäoliselt suudaksime tõestada, et vaadeldud kolmes piirkonnas pole uuritava tunnuse keskväärtus samasugune (kui kasutame andmete analüüsimisel dispersioonanalüüsi)?

Variant 1
Dispersioonanalüüsi võimsust saab leida R'i käsu power.anova.test abil:

```
power.anova.test(3, n=30, between.var=var(c(100,110, 120)), within.var=400)
```

Või praktilisem näide – graafiku koostamine, kust saab erinevate valimi mahtude korral lugeda välja saavutatavat võimsust. Graafikul on eraldi ära märgitud võimsuse 0,8 saavutamiseks vajalik valimi suurus, sest tavaks on uuringutelt nõuda vähemalt 80% võimsust – madalama võimsuse korral sageli uuringut korraldama ei hakatagi:

```
n_grupi_kohta=2:50
v6imsus=power.anova.test(3, n=n_grupi_kohta,
      between.var=var(c(100,110, 120)), within.var=400)$power
plot(n_grupi_kohta*3,  v6imsus, type="l", xlab="n (valimi suurus)",
      ylab="Võimsus", ylim=c(0,1))

# Võimsuse 0,8 saavutamiseks vajalik valimi suurus
v6imsus80=power.anova.test(3, power=0.8,
      between.var=var(c(100,110, 120)), within.var=400)$n*3
arrows(0, 0.8, v6imsus80, 0.8, length=0, col="gray70", lwd=2)
arrows(v6imsus80, 0.8, v6imsus80, 0, length=0.1, col="gray70", lwd=2)
```

Variant 2

Esimest võimalust –sisseehitatud R'i käsku – on küll mugav kasutada, kuid ta on mõeldud eelkõige ühe väga kindla katseplaani jaoks (üks faktortunnus, igas grupis sama arv mõõtmiseid jne). Kui soovime leida võimsust veidi ebatüüpilisemate/elulisemate katseplaanide jaoks siis peame oma arvutused ikka ise tegema:

```
# Leiame F-statistiku kriitilise väärtuse
      kriit=qf(0.95, 2, 90-3)

# Leiame tõenäosuse, et tegelikult (meie oletuste kehtides) f-statistiku väärtus tuleb
# suurem kui kriitiline väärtus
      1-pf(kriit, 2, 90-3, ncp=((100-110)**2*30 + (110-110)**2*30+ (120-
      110)**2*30)/400)
```

See iseleitud võimsus tuleb praegu täpselt sama kui anova.power.test-i abil leitud – sest antud näite korral saab mõlemal viisil võimsust leida. Kui aga midagi katseplaanis muuta jääb isearvutus ainsaks lahenduseks. Näiteks kui võtaksime Lõuna- ja Põhja-Eestist kaks korda suurema valimi (mõlemast 36) kui Kesk-Eestist (18 inimest, kokku n=90), siis milline tuleks testi võimsus?

```
      # Mittetsentraalsuse parameetri leidmine
      n_grupi_kohta= c(36,18,36)
      grupp =rep(1:3, n_grupi_kohta)
      keskvaartusvektor =rep(c(100, 110, 120), n_grupi_kohta)
      p1=predict(lm(keskvaartusvektor~factor(grupp)))
      p2=predict(lm(keskvaartusvektor~1))
      ncp=sum((p1-p2)**2)/400

      # Kriitilise väärtuse leidmine
      kriit=qf(0.95, 2, 90-3)

      # Testi võimsus
      1-pf(kriit, 2, 90-3, ncp=ncp)
```

Näeme, et antud juhul on testi võimsus suurem kui siis, kui võtaksime igast maakonnast sama arvu uuritavaid.

Vaata ka näiteks järgmist joonist:

```
valimimaht=2:50
v6imsus=power.anova.test(3, n=valimimaht,
      between.var=var(c(100,110, 120)), within.var=400)$power
plot(valimimaht*3,  v6imsus, type="l", xlab="n (kokku)",
      ylab="Võimsus", ylim=c(0,1), col="skyblue", lwd=3)

valimimaht=seq(5, 150, 5)
v6imsus2=rep(NA, length(valimimaht))

for (i in 1:length(valimimaht)){
  n=valimimaht[i]
  npergrupp= n*c(0.4, 0.2, 0.4)
  grupp=rep(1:3, npergrupp)
  keskv1=rep(c(100, 110, 120), npergrupp)
  p1=predict(lm(keskv1~factor(grupp)))
  p2=predict(lm(keskv1~1))
  ncp=sum((p1-p2)**2)/400
  kriit=qf(0.95, 2, n-3)
  v6imsus2[i] = 1-pf(kriit, 2, n-3, ncp=ncp)
}
lines(valimimaht, v6imsus2, lwd=2, col="blue")

legend("bottomright", c("grupi suurused 2:1:2",
      "grupi suurused 1:1:1"), lwd=3, col=c("blue","skyblue"))
```

Ülesanne 1.
Kavatseme ikkagi võtta igast piirkonnast (Lõuna-Eesti; Kesk-Eesti; Põhja-Eesti) sama arvu uuritavaid (30 uuritavat igast piirkonnast). Aga seekord kavatseme kodeerida maakonna pideva tunnusena (1: Lõuna-Eesti; 2: Kesk-Eesti; 3: Põhja-Eesti) ja dispersioonanalüüsi asemel kavatseme hinnata regressioonanalüüsi mudeli. Kui testime sirge tõusu olulisust siis milline tuleks testi võimsus? Leia lähenemist 2 kasutades õige vastus!

Vihjeks: Õige vastus on suurem kui 0,94 ja väiksem kui 0,98!

Variant 3
Mõnikord on kõige mugavam leida testi võimsust simulatsiooni abil. Näiteks kui võtaksime juhusliku valimi Eestis elavatest karuküttidest (valimi suurus n=90) ja teaksime, et 1/3 karuküttidest elab Põhja-Eestis, 1/3 Kesk-Eestis ja 1/3 Lõuna-Eestis, siis milline tuleks testi võimsus (kui karuküttidel uuritava tunnuse keskväärtused piirkonniti oleks 120; 110 ja 100)? Eelnenud arvutused eeldasid, et võtame igast piirkonnast kindla arvu uuritavaid. Juhusliku valimi korral me aga ei tea ette, mitu inimest Põhja-Eestis, mitu Kesk-Eestist ja mitu Lõuna-Eestist valimisse sattub. Siiski saame testi võimsust ka sellises situatsioonis mugavalt leida simulatsiooni abil (tõsi küll, ligikaudselt):

```
korduseid=10000
pvalue=rep(NA, korduseid)
for (i in 1:korduseid){
  grupp=sample(1:3, 90, replace=TRUE)
  y=c(100, 110, 120)[grupp]+rnorm(90, 0, sqrt(400))
  m=lm(y~factor(grupp))
  pvalue[i] = anova(m)[1,5]
}
# Testi võimsus (koos arvutustäpsust näitava usaldusintervalliga):
t.test(pvalue<0.05)
```

Alternatiivina simulatsioonile võiksime proovida testi võimsust leida täistõenäosuse valemit kasutades:

$$P(võimsus) = \sum_{\{n1+n2+n3=90\}} P(võimsus|n1; n2; n3)P(n1; n2; n3),$$

ehk leiame testi tingliku võimsuse mistahes võimalike valimi suuruste korral ja seejärel leiaksime täistõenäosuse valemi abil lihtsalt testi võimsuse. Aga vastav arvutus muutub valimimahu kasvades kiiresti ebapraktiliselt keeruliseks.

Näide 2
Sageli sisaldab meie poolt kasutatav mudel paljusid erinevaid tunnuseid lisaks sellele tunnusele, mille mõju vastu me tegelikult huvi tunneme. Mudelis võib olla sees inimese vanus ja sugu lisaks meid huvitavale tunnusele (töötlus). Kuidas sellisel juhul töötluse mõju tuvastamise tõenäosust leida?

Võib muidugi ette anda filigraanselt välja mõeldud keskväärtuste vektori iga vaatluse jaoks ja teha selle baasilt korrektne arvutus. Sellise täpse keskväärtuste vektori väljamõtlemine võib olla päris vaevarikas (peaksime ju välja mõtlema lisaks töötluse mõjule ka soo ja vanuse mõjud). Samas tasub tähele panna, et võime alati keskväärtuste vektorile juurde lisada midagi kujul $X_0\beta_0^*$ ja sellest meie poolt leitav mittetsentraalsuse parameeter ei muutu – seega võime alati näiteks soo ja vanuse mõju arvutustes võtta nulliks – leitud testi võimsus sellest ei muutu. Näiteks soovime hinnata mudelit y~factor(faktorA)+factor(faktorB) ja soovime leida, millise tõenäosusega võiksime tuvastada tunnuse faktorA mõju. Ütleme et igal faktorite A ja B kombinatsiooni korral soovime teha 20 vaatlust, jääkide dispersioon on 25. Järgmised kaks programmi jõuavad täpselt sama tulemuseni:

Variant 1 (keerukam)

```
#Keskmised iga faktorite A ja B kombinatsiooni kohta
grupi_keskmised=c( 10, 12, 20, 22, 30, 32)

# Mitu vaatlust sellise kombinatsiooni korral tehti
n_grupi_kohta  =c( 20, 20, 20, 20, 20, 20)

faktorA = rep(  c(  1,  2,  1,  2,  1,  2), n_grupi_kohta)
faktorB = rep(  c(  1,  1,  2,  2,  3,  3), n_grupi_kohta)

# Kõigi vaatluste keskväärtuste vektor
ykeskmised=rep(grupi_keskmised, n_grupi_kohta)

m1=lm(ykeskmised~factor(faktorA)+factor(faktorB))
m2=lm(ykeskmised~factor(faktorB))

# Leiame mittetsentraalsuse parameetri (jääkide dispersioon 25)
ncp=sum((predict(m1)-predict(m2))^2)/25
ncp

# Leiame F-statistiku kriitilise väärtuse
kriit=qf(0.95, 3, 120-4)

# Leiame tõenäosuse, et tegelikult (meie oletuste kehtides)
# f-statistiku väärtus tuleb suurem kui kriitiline väärtus:
1-pf(kriit, 3, 120-4, ncp=ncp)
```

Variant 2 (lihtsam – aga sama tulemus)

```
grupi_keskmised = c(  0,  2)
n_grupi_kohta   = c( 60, 60)

faktorA = rep(1:2, n_grupi_kohta)

ykeskmised=rep(grupi_keskmised, n_grupi_kohta)

m1=lm(ykeskmised~factor(faktorB))
m0=lm(ykeskmised~1)

ncp=sum((predict(m1)-predict(m0))^2)/25
ncp

# Leiame F-statistiku kriitilise väärtuse
kriit=qf(0.95, 3, 120-4)
# Leiame tõenäosuse, et tegelikult (meie oletuste kehtides)
# f-statistiku väärtus tuleb suurem kui kriitiline väärtus:
1-pf(kriit, 3, 120-4, ncp=ncp)
```

Ülesanne 2

Vanarahvas teadis: „need lapsed, kes hommikuti putru söövad kasvavad suureks“. Kavatseme teha uuringu selle väite tõesuse kontrolliks. Jagame juhuslikult lastekodusse sattunud lapsed kaheks grupiks. Üks gruppidest saab hommikuti putru, teine teistsugust hommikusööki. Kui katses osalenud lapsed on ükskord täiskasvanuks saanud (näiteks saanud 20 aastat vanaks) siis mõõdame nende pikkused ära. Kui palju katsealuseid vajaksime tõestamaks pudrusöömise mõju pikkusele (näiteks kui suurt valimit vajame võimsuse 0,8 saavutamiseks)? Leia vajalik valimi suurus!

Kas pudru mõju testimisel peaksime mudelisse lisama ka katsealuse soo? Põhjenda oma otsust!