

Kvalitatiivsete andmete analüüs  
5. praktikum  
**Cochran-Mantel-Haenszeli test**  
logistiline regressioon

Kui katsetame enda jaoks uusi statistilisi meetodeid, on sageli mõistlik esmalt proovida uut meetodit simuleeritud andmete peal – veendumaks, et meie poolt kasutatav statistiline meetod ikka leiab andmetest üles just sellised seosed, mille oleme sinna salakavalalt peitnud ja mitte midagi muud (st ei valeta meile).

Tekitame selle katsetamiseks ühe näiteandmestiku:

```
# Uuringusse kaasatud inimeste arv:
n=10000

# Pooled neist olgu vanad ja pooled noored:
vanus=rep(c("vana", "noor"), each=n/2)

# Olgu noortel väiksem tõenäosus olla vaktsineeritud kui vanadel
# noortest 10% on vaktsineeritud, vanadest 80%,
# tähistused: 1-vaktsineeritud, 0-vaktsineerimata :
vaktsiin=rbinom(n, 1, p=0.1*(vanus=="noor")+0.8*(vanus=="vana"))

# Pilk genereeritud andmetele:
table(vanus, vaktsiin)
prop.table(table(vanus, vaktsiin), 1)

# Tekitame ka tunnuse, mis näitab kas inimene suri
# 5 aasta jooksul või mitte:
surm=rbinom(n, 1, p=0.001*(vanus=="noor")+0.1*(vanus=="vana"))
table(surm, vaktsiin)

# Näeme, et vaktsineeritudel on suurem tõenäosus surra:
prop.table(table(surm, vaktsiin), 2)

# Ka fisheri täpne test näitab seose olemasolu:
fisher.test(table(surm, vaktsiin))
```

Praegu on segavaks tunnuseks muidugi inimese vanus – vanemaid inimesi on vaktsineeritud rohkem ja neil on ka suremistõenäosus suurem. Proovime analüüsis arvestada segava tunnuse mõjuga, kasutades Mantel-Haenszel'i testi:

```
# Teeme eraldi tabelid noortele ja vanadele:
table(surm, vaktsiin, vanus)
```

ja vaatame šansside suhteid eraldi vanadel ja noortel ning kombineerime siis saadud tulemused:

```
mantelhaen.test(table(surm, vaktsiin, vanus))
```

Kas näed vaktsiini mõju? Kas see peaks olema olema või mitte?

NB! Segav faktor peab olema tunnuste loetelus viimasel kohal!

## Ülesanne 1

Muuda andmete genereerimist selliselt, et vaktsiin vähendab inimese suremise šansse 2x (OR=0,5). Korda analüüsi – mida Mantel-Haenzeli test nüüd arvab vaktsiini mõju kohta?

## Ülesanne 2 – krabid.

Paaritumisperioodil koguneb emakrabi ümber mitmeid isakrabisid. Mõni emakrabi jääb aga üksikuks. Milllest see küll sõltub?

Loe sisse emakrabide populaarsust kirjeldav andmestik:

```
krabid = read.table(url("http://www-1.ms.ut.ee/mart/kvalitatiivsed/crabs.txt"),
  dec = ".", header = T)
head(krabid)
dim(krabid)
```

Tunnuste tähendus:

- color (1 heledaim, 4 tumedaim),
- spine (selgroog, 1 mõlemad segmendid head, 2 üks kulunud/murtud, 3 mõlemad kulunud/murtud),
- width (kilbi laius cm),
- weight (kg),
- satell (satelliitide arv ehk emakrabi ümber tiirlevate isaskrabide arv)

Tekita uus tunnus onsat mis näitab, kas emakrabi ümber tiirleb isaseid krabisid või mitte (0 - ei ühtki; 1 - vähemalt 1):

```
krabid$onsat = 1 * (krabid$satell > 0)
```

Üsna usutav on see, et emakrabi värvil on mõju sellele, kas isased teda märkavad või mitte (kontrolli seda väidet siiski, tee sobiv test)

Aga kas ka selgroo vigastuste olemasolu mõjutab isakrabide eelistusi? Antud küsimusele vastamisel peab muidugi arvestama, et emakrabi värv on arvatavasti segavaks faktoriks (oskad põhjendada miks?). Kontrolli sobival viisil, kas emakrabi selgroovigastused (*spine*) omab mõju isakrabide eelistustele?

Vahel tahame teha analüüsi juba avaldatud andmetele (mis esitatud sagedustabelina). Vaatame järgnevalt, kuidas seda teha!

Ka loengus tehti läbi näide vaktsineerimise mõju uurimise kohta.

Sisestame loengus tutvustatud tabeli R-i:

```
and=data.frame(  
  vanus = rep(c("vana", "noor"), each=4),  
  vaktsiin = rep(c("vaktsineeritud", "vaktsineerimata"), 4),  
  surm = rep(c("suri", "elus", "suri", "elus"), each=2),  
  arv = c(200, 20, 99800, 980, 5, 500, 100000, 1000000)  
)  
and
```

Seos vaktsineerimise ja COVID-suremuse vahel eksisteerib:

```
tabel=xtabs(arv~vaktsiin+surm, data=and)  
tabel  
chisq.test(tabel)
```

Ja näeme ka seda, et hinnanguliselt on vaktsineeritud 2x suuremad šansid surra COVID-i tõttu:

```
library(epitools)  
oddsratio(tabel)
```

Nagu loengus kuulsime, võib selline seos tekkida sellepärast, et epideemia teataval ajahetkel tähendas vaktsineeritud olemine seda, et sa olid vana (ja vanadel inimestel oligi suurem risk surra covid-i tõttu). Võime teha analüüsi ka vanusegruppide kaupa:

```
tabel_noor=xtabs(arv~vaktsiin+surm,  
                 data=and[and$vanus=="noor",])  
tabel_noor  
oddsratio(tabel_noor)  
  
tabel_vana=xtabs(arv~vaktsiin+surm,  
                 data=and[and$vanus=="vana",])  
tabel_vana  
oddsratio(tabel_vana)
```

Märka, kuidas šansside suhte hinnang on muutunud – kui sa oled vana, siis vaktsineerimine aitab. Kui sa oled noor, siis ka vaktsineerimine aitab.

```
tabel2=xtabs(arv~vaktsiin+surm+vanus, data=and)  
tabel2  
summary(tabel2)
```

Sellisel viisil saab testida, kas antud kolm tunnust on sõltumatud või mitte

Teeme Cochran-Mantel-Haenszeli testi ( $H_0$ : vaktsineerimine ja suremus on tinglikult sõltumatud tingimusel et inimese vanus on antud)

```
mantelhaen.test(tabel2)

# Loengus näidatud viisil saame teha arvutusi nii
mantelhaen.test(tabel2, correct = FALSE)
```

Ühist hinnangut šansside suhtele võime saada ka logistilist regressiooni kasutades. Selleks vajame aga veidi teisel kujul andmestikku:

```
and2=data.frame(
  vanus = c("vana", "vana", "noor", "noor"),
  vaktsiin = c("vakts", "vaktsineerimata", "vakts",
               "vaktsineerimata"),
  surmi = c(200, 20, 5, 500),
  elus=c(99800, 980, 100000, 1000000)
)
and2

mudel1=glm(cbind(surmi,elus)~factor(vaktsiin)+factor(vanus),
            family=binomial(), data=and2)
summary(mudel1)
coef(mudel1)
```

Vaktsiini mõju –  $\log(OR)$ :

```
coef(mudel1)[2]
```

Vaktsiini mõju –  $OR$ :

```
exp(coef(mudel1)[2])
```

Eelnenud šansside suhe näitas seda, kui mitu korda on vaktsineerimata inimese suremisšansid suuremad vaktsineeritud inimese suremisšanssides. Kui tahaksime leida teistpidi šansside suhet (kui mitu korda on vaktsineeritud inimese suremisšansid suuremad vaktsineerimata inimese suremisšanssides, siis peame muutma referentsnivood ehk taset kelle suhtes võrdlusi tehakse):

```
mudel1=glm(cbind(surmi,elus)~relevel(factor(vaktsiin),
    ref="vaktsineerimata")+factor(vanus), family=binomial(),
    data=and2)
summary(mudel1)
coef(mudel1)[2]
exp(coef(mudel1)[2])
```