

Kvalitatiivsete andmete analüüs  
2. praktikum  
**Seose olemasolu testimine**

**Kahe tunnuse sõltuvus**

Loeme sisse ühe näidisandmestiku – kus küsiti inimestelt, kas nad usuvad surmajärgsesse elusse või mitte (andmestik võetud raamatust *An Introduction to Categorical Data Analysis*. Agresti 2007):

```
load(url("http://www-1.ms.ut.ee/mart/kval_2.RData"))
head(elu2)
attach(elu2)
```

Kas uskumine ellu peale surma sõltub inimese soost (kas ühest soost inimesed usuvad sagedamini surmajärgsesse ellu kui teisest soost inimesed?)

Heidame esmalt pilgu peale oma andmetele:

```
table(sugu, usk)
```

Vaatame ka jaotusi sugude kaupa (esimese tunnuse järgi):

```
prop.table(table(sugu, usk), 1)
```

|       | usk       |           |
|-------|-----------|-----------|
| sugu  | ei        | jah       |
| mees  | 0.2632613 | 0.7367387 |
| naine | 0.2525773 | 0.7474227 |

Näeme, et 26,3% meestest ei usu surmajärgsesse elusse ja 25,3% naistest ei usu surmajärgsesse elusse. Aga kas see erinevus eksisteerib ikka päriselt või võib olla tegemist valimi juhuslikkusest tingitud müraga (erinevus vaid selle pärast, et hinnangud osakaaludele pole täpsed)?

Proovime seose olemasolu kontrollida statistiliste testide abil – teeme ligikaudsete testide arvutused läbi ise ja kasutame siis ka valmiskäske oma arvutuste kontrollimiseks.

Kui palju ootaksime oma valimis (n=1091) nägevat mitteusklikke mehi kui uskumine/mitteuskumine ei sõltuks inimese soost? Valimis on kokku 509 meest:

```
table(sugu)
```

| sugu |       |
|------|-------|
| mees | naine |
| 509  | 582   |

mitteusklike osakaal valimis (kõigi peale kokku, sest eeldame, et osakaal ei sõltu inimese soost) on 0,25756... :

```
prop.table(table(usk))
```

| usk |                     |
|-----|---------------------|
|     | ei jah              |
|     | 0.2575619 0.7424381 |

Seega ootaksime sõltumatuse korral mitteuskuvaid mehi nägevat  $509 \cdot 0.2575619 = 131,099$  tükki. Tegelikult nägime aga valimis 134 mitteusklikku meest.

Leiame ootuspärased sagedused (eeldades sõltumatust) kõigi kombinatsioonide jaoks:

```
Ootused=table(sugu) %*%t(prop.table(table(usk)))
Ootused
```

|       | usk     |         |
|-------|---------|---------|
| sugu  | ei      | jah     |
| mees  | 131.099 | 377.901 |
| naine | 149.901 | 432.099 |

Ning leiame hii-ruut teststatistiku väärtuse:

```
tegelik=table(sugu, usk)
statistik=sum((tegelik-Ootused)**2/Ootused)
statistik
```

```
[1] 0.162084
```

Viimaks leiame p-väärtuse – tõenäosuse näha nii suurt või veelgi suuremat teststatistiku väärtust sõltumatuse korral:

```
1-pchisq(statistik, df=1)
[1] 0.6872451
```

Võrdle saadud tulemust R-i enda poolt chisq.test käsu abil saadud tulemusega::

```
chisq.test(table(sugu,usk), correct=FALSE)
```

Loengus tutvusime ka tõepärasuhte testiga. Võime antud testi teststatistiku ka ise arvutada:

```
statistik_LR=sum(2*tegelik*log(tegelik/Ootused))
statistik_LR
1-pchisq(statistik_LR, df=1)
```

Või võime kasutada R-i enda vahendeid. Kui uuritaval tunnusel on kaks võimalikku väärtust:

```
m1=glm(1*(usk=="jah")~sugu, family=binomial)
m2=glm(1*(usk=="jah")~1, family=binomial)
anova(m2,m1)
```

Juhul, kui uuritaval tunnusel on rohkem kui kaks väärtust:

```
library(DescTools)
GTest(table(sugu,usk))
```

Võrdle saadud tulemusi – millise järelduseni jõuad?

Lisaks võime proovida ka täpset testi – näiteks Fisheri täpset testi:

```
fisher.test(table(sugu,usk))
```

Millise tulemuseni jõuad Fisheri täpse testi abil? Kas tulemus on ootuspärane?

## Ülesanne 1

Vaata andmestikku andmed:

```
head (andmed)
```

Tegemist retrospektiivse (juht-kontroll) uuringuga, kus infarkti saanud inimestelt (või nende lähedastelt) küsiti, kas nad olid enne infarkti suitsetajad. Seejärel otsiti iga infarkti saanud inimese kohta ~2 sama vana ja sama soost inimest kes infarkti ei saanud. Ka nende käest uuriti, kas nad suitsetavad või mitte. Uuri, kas tunnused *Suitsetab* ja *Infarkt* on sõltumatud või mitte. Millist testi võiks siin kasutada?

Kui otsustad, et tunnuste *Suitsetab* ja *Infarkt* vahel esineb seos, siis kirjelda seda seost – milles see seisneb?

### Barnardi täpne test

Tehti järgmine kliiniline katse. Sajast patsiendist loositi välja 50 ja neid raviti ravimeetodil A. Ülejäänud patsiente raviti ravimeetodil B. Seejärel vaadati, mis neist patsientidest sai:

|        | Suri | sai terveks |
|--------|------|-------------|
| Ravi A | 10   | 40          |
| Ravi B | 3    | 47          |

Kas suudame näidata, et üks ravimeetod on teisest parem? Või pole ravi ja ravitulemuse vahel statistilist seost?

Võime seose olemasolu testida Fisheri täpse testi abil:

```
fisher.test(rbind(  
  c(10, 40),  
  c( 3, 47)))
```

Aga paraku ei suuda Fisheri täpne test tuvastada ravitulemuse jaotuse muutumist ravi muutmisel:

```
Fisher's Exact Test for Count Data  
p-value = 0.07131
```

Meile võib aga meenuda, et antud juhul saame kasutada ka ühte terasemat testi. Nimelt on meil antud juhul fikseeritud vaid üks sagedustabeli marginaaljaotus (mitu patsienti said ravi A ja mitu ravi B), teine marginaal (mitu patsienti suri ja mitu sai terveks) on juhuslik – surijate arv kordusuuringu puhul võiks tulla teistsugune. Sellise uuringu korral (ja ainult selliste uuringute korral!) võime aga kasutada Barnardi täpset testi. Selle testi jaoks vajalike arvutuste tegemine on aeglane, vastuse leidmine võib aega võtta paar minutit. Ära muretse ja heida meelt, kui arvuti sinuga peale käsu andmist koheselt suhelda ei soovi!

Kasutame Barnardi testi `BarnardTest` lisamoodulist `DescTools` (vajadusel paigalda vastav lisamoodul enne kasutamist):

```
library(DescTools)

BarnardTest(rbind(
  c(10, 40),
  c( 3, 47)), method="csm", fixed=1)
```

Lisaparameter `fixed=1` ütleb arvutile seda, et reasummad (st  $10+40=50$ ) on kordusuuringutes fikseeritud, aga veerusummad on juhuslikud ja võivad muutuda.

Saadud tulemus (Barnardi täpse testi arvutamine võtab veidi aega, ära muretse kui arvuti koheselt soovile ei reageeri):

```
CSM Exact Test

data: 10 out of 50 vs. 3 out of 50
test statistic = NA, first sample size = 50, second sample size = 50,
p-value = 0.04222
alternative hypothesis: true difference in proportion is not equal to 0
sample estimates:
difference in proportion
0.14
```

Barnardi täpne test suudab tõestada, et ravimid pole sama head – ravim B (Treatment II) on tõestatavalt parem.

Veidi kiirema variandi Barnardi testist saad käsuga

```
BarnardTest(rbind( c(10, 40), c( 3, 47)), method="z-pooled", fixed=1)
```

## Ülesanne 2

Uuringusse otsiti 40 kõhuvaluga inimest ja 80 inimest kellel polnud kõhuvalu. Kõigi käest küsiti, kui paljud neist olid eelnevalt kommi söönud. Kõhuvaluga inimestest olid 15 eelnevalt kommiga maiustanud ja kõhuvaluta inimestest oli 12 kommi söönud. Kas kommisöömise ja kõhuvalu esinemise vahel esineb statistiline seos? Testi seose olemasolu Barnardi täpse testi abil!