# Biomeetria
# 6. praktikum

# Hii-ruut test. Statistiline seos

## Hii-ruut test hüpoteeside kontrollimiseks jaotuse kohta.

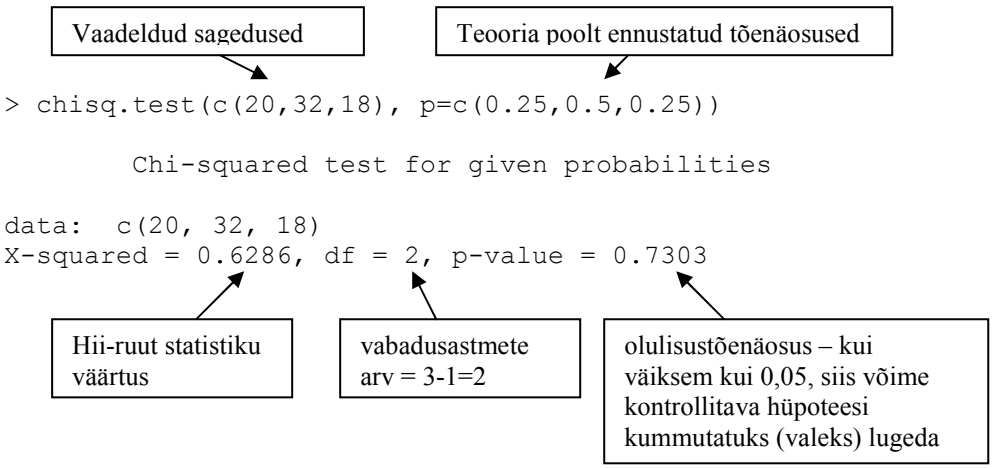
Juhul, kui meid huvitav teooria võimaldas leida uuritava tunnuse väärtuste esinemistõenäosuseid, siis saab selle teooria paikapidavust kontrollida hii-ruut testi abil. Hii-ruut testi saab R-is teha kasutades käsku chisq.test .

**Näide**

Oletame, et soovime kontrollida Mendeli seaduste kehtivust. Heterosügootide ristamisel saadi järgmised tulemused:

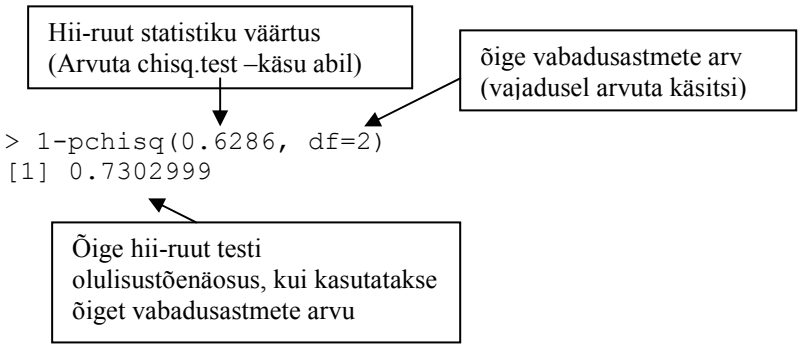
| genotüüp | valim ($n_i$) | väärtuse ennustatud esinemistõenäosus (Mendeli seadus) |
|---|---|---|
| AA | – 20 järglast | 0,25 |
| Aa | – 32 järglast | 0,5 |
| aa | – 18 järglast | 0,25 |

Kas Mendeli seadused kehtivad?




Järeldus: Meie katsetulemused on kooskõlas Mendeli seadustega – nad ei lükka nullhüpoteesi ümber.

Märkus: juhul kui tõenäosuste arvutamisel on kasutatud valimit (vaata loengumaterialist Hardy-Weinbergi tasakaalu kontrollimise näidet), siis R-i (ja ka teiste statistikapakettide) automaatika seda ei tuvasta ning kasutatakse valet vabadusastmete arvu. Sellisel juhul ei saa arvuti poolt vaikimisi esitatud tulemust usaldada – see on lihtsalt vale. Õnneks saab tehtud tööd siiski kasutada – olulisustõenäosust saab leida ka siis, kui oleme ise käsitsi õige vabadusastmete arvu leidnud (hii-ruut statistiku arvutamise võib ikka arvutile jätta). „Käsitsi“ arvutatud vabadusastmete arvu saab kasutada olulisustõenäosuse arvutamisel nii:



**Ülesanne**
Soovime kontrollida, kas kõik uurimisalused aastad (2002-2006) olid kotkaliik-X-le pesitsemiseks samasobivad või oli mõni aasta silmnähtavalt ebasoivam. Igal aastal jälgiti 10-t kotkapesa ja loeti üle täiskasvanuks kasvatatud kotkalaste hulk. Saadi järgmised andmed:

| Aasta | suureks sirgunud kotkaid |
|---|---|
| 2002 | 5 |
| 2003 | 12 |
| 2004 | 8 |
| 2005 | 8 |
| 2006 | 10 |

Kontrolli hii-ruut testi abil, kas pesitsemisaastad olid kotkastele võrdväärselt head? Vihje: kui oleks võrdselt head, siis peaks meie andmestikku sattunud kotkapoeg olema pärit aastast 2002 samasuure tõenäosusega kui aastast 2006.
Kommenteeri testitulemust, arutle, mis võib testitulemust mõjutada (näiteks: kui kontrolliti kogu aeg samu pesi, siis võisid linnud igal aastal järjest vanemaks jääda. Vanade lindude pesitsusedukus ei peagi olema samasugune kui noortel. Tulnuks ehk igal aastal uued peasad otsida)?

**Ülesanne**
Soovitakse teada, kas linna sattuvad suured metsloomad ehk sellepärast, et neid metsas segatakse. Kui see nii on, siis peaks linna sattuma loomi rohkem just nädalavahetustel, kui metsas palju huvilisi käib. Otsustati kontrollida, kas metsloom sattub linna igal nädalapäeval sama tõenäosusega (nullhüpotees – loomade linnasattumine ei sõltu nädalapäevast – loomadel pole ju aimugi 7-päevasest nädalast) või siiski sõltub (kui sõltub, viitab see inimese „süüle“). Koguti kokku 5 aasta jooksul linnadesse/asulatesse sattunud metsloomade andmed. Nädalapäeviti jagunesid inimasulatesse sattumised järgmiselt:

| Nädalapäev | E | T | K | N | R | L | P |
|---|---|---|---|---|---|---|---|
| Metsloomi | 18 | 12 | 8 | 6 | 12 | 14 | 22 |

Kontrolli hii-ruut testi abil, kas on võimalik, et looma linna sattumine ei sõltu nädalapäevast?

## Seosed tunnuste vahel

Käesolevas osas kasutame taas tudengite andmestikku. Alustuseks loemegi andmestiku sisse:

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

Antud küsitluse ankeeti võite näha aadressil
http://www.ms.ut.ee/mart/MC2007/ankeet.pdf

Tahame teada, kas tunnuste *sugu* ja *sport* vahel eksisteerib seos (kas naised on sportlikumad kui mehed? või pole? või ... ?), ja kui eksisteerib, siis milline see on.

Sportimisküsimuse kodeering on järgmine:

Kui tihti tegelete sportimisega?
- 1 – mitte kunagi
- 2 – 1-2 korda nädalas
- 3 – 3-4 korda nädalas
- 4 – üle 5 korra nädalas

Alustame sagedustabelist ja tema visualiseerimisest:

```
> table(sugu, sport)
    sport
sugu   1   2   3   4   5
   1 103 313  80  13   2
   2  21  69  38  19   2
```

Selle põhjal saab koheselt öelda vaid seda, et poisse satub arstiteaduskonda harvem kui tüdrukuid. Hindame tunnuse *sport* jaotuse tüdrukute ja poiste jaoks eraldi:

```
> prop.table(table(sugu, sport),1)
    sport
sugu           1           2           3           4           5
   1 0.201565558 0.612524462 0.156555773 0.025440313 0.003913894
   2 0.140939597 0.463087248 0.255033557 0.127516779 0.013422819
```

Näeme, et naistudengid (sugu==1) spordivad vähem: 20% küsitletud naistudengitest ei spordi kunagi, samas kui nii vähe sporti teevad vaid 14% küsitletud meestudengitest jne.

Võime vaadata ka teistpidi – kui palju on mittesportivate tudengite seas naisi, kui palju rohkelt sportivate tudengite seas jne:

```
> prop.table(table(sugu, sport),2)
    sport
sugu         1         2         3         4         5
   1 0.8306452 0.8193717 0.6779661 0.4062500 0.5000000
   2 0.1693548 0.1806283 0.3220339 0.5937500 0.5000000
```

Näeme, et kui mittesportivatest tudengitest on 83% naised, siis 3-4 korda nädalas sportivatest tudengitest on naistudengeid kõigest 59%.

Esitame saadud tulemuse ka graafiliselt. Selleks on paar erinevat võimalust. Milline neist meeldib Sulle? Kas saad kõigi graafikute korral aru, mida seal kujutatud on?

```
barplot(prop.table(table(sugu, sport),2)*100,
      col=c("orange","skyblue"),
      ylab="%", main="Seos sportimise ja soo vahel",
      legend=c("naised","mehed"), xlim=c(0,7.5))
```

Jätame legendi jaoks veidi ruumi

```
barplot(prop.table(table(sport, sugu),2)*100, col=terrain.colors(5),
        legend= c("ei spordi", "1-2", "3-4", "5 ja enam"),
        names.arg= c("naised","mehed"),
        xlim=c(0,3.5), ylab="%", main="Seos sportimise ja soo vahel")
```

Kas seos eksisteerib tegelikult või mitte? Valimi juhuslikkuse tõttu võime ekslikult arvata, et eri soost tudengite spordivaimustus on erinev. Seda, kas seos ikka tegelikul ka eksisteerib (ka siis alles jääb, kui me lõpmatult palju tudengeid uuriksime) saab kontrollida hii-ruut testi abil. Hii-ruut testi kasutamine kahe tunnuse sõltumatuse kontrolliks näeb antud juhul välja järgmine:

```
> chisq.test(table(sugu, sport) )

        Pearson's Chi-squared test

data:  table(sugu, sport)
X-squared = 39.4783, df = 4, p-value = 5.549e-08

Warning message:
Chi-squared approximation may be incorrect in: chisq.test(table(sugu,
sport))
```

vabadusastmete arv

olulisustõenäosus

Hii-ruut teststatistiku väärtus

Hoiatus – ootuspärased arvud on liiga väiksed!

Vaatame, millepärast anti hoiatus – trükime välja ootused:

```
> chisq.test(table(sugu, sport))$expected
    sport
sugu        1        2        3         4         5
   1 96.00606 295.7606 91.36061 24.775758 3.0969697
   2 27.99394  86.2394 26.63939  7.224242 0.9030303
```

Näeme, et vastusevarianti 5 on valitud liiga harva. Muuseas, milline on vastusevariant 5? Miks sellised vastused üldse on andmestikku tekkinud? Kas me saame seda millegiga kokku panna?

Kaks võimalust – vali neist sinu meelest sobivaim!:

Variant1 – loeme vastusevariandid „5“ lihtsalt vastusevariantideks „4“

```
sport2=sport
sport2[sport2==5]=4
```

Variant2 – loeme vastusevariandid „5“ puuduvateks vastusteks (ehk viskame analüüsist välja)

```
sport2=sport
sport2[sport2==5]=NA
```

Kumb lähenemine tundub sulle õigem? Miks?

Peale tunnuse ümberkodeerimist teeme uuesti hii-ruut testi:

```
> chisq.test(table(sugu, sport2))

        Pearson's Chi-squared test

data:  table(sugu, sport2)
X-squared = 37.9465, df = 3, p-value = 2.901e-08
```

Näeme, et hoiatust enam ei anta. Tunnuste sugu ja sport vahel eksisteerib tõepoolest seos – naised teevad vähem sporti.

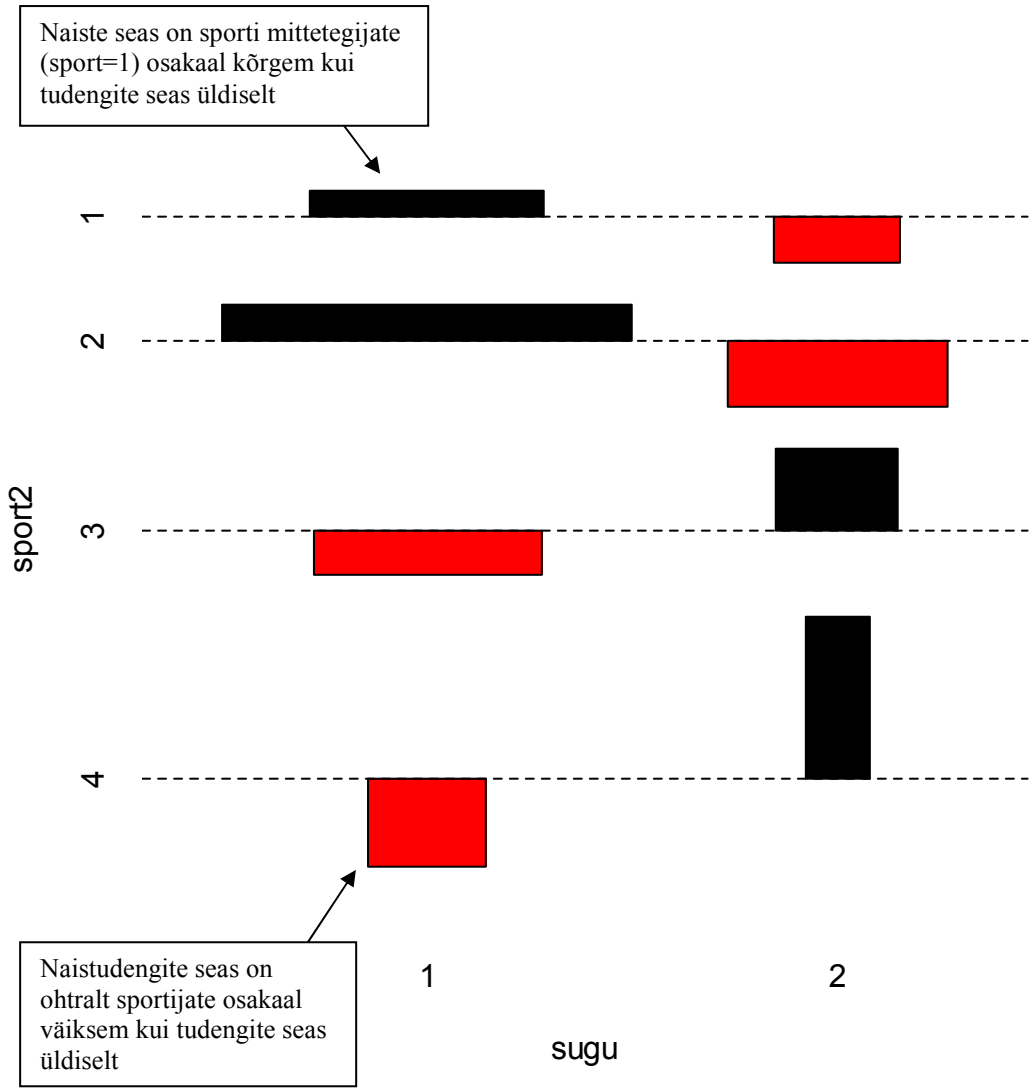
Märkus: Alternatiivina võib lasta R-l arvutada olulisustõenäosuse nn. täpsel meetodil – ilma hii-ruut jaotust kasutamata:

```
chisq.test(table(sugu, sport), simulate.p.value=TRUE, B=100000)
```

Sellisel viisil tehtud hii-ruut testi puhul ei pea kõik ootuspärased sagedused olema 5 või enam (aga soovitavalt peaks siiski paljud seda olema).

Sellest, mida see seos ikka tähendab, võid proovida aru saada ka järgmise graafiku – assocplot-funktsiooni poolt joonistatava graafiku – abil.

```
assocplot(table(sugu, sport2))
```




## Ülesanne:

Vaata tunnuseid *olu* ja *suitsetamine*. Kas nende tunnuste vahel eksisteerib seos? Kui jah, siis milline see on – kirjelda sõnades!

Tunnuste olu ja suitsetamine kodeeringud:

**Suitsetamine**: Milline järgnevatest variantidest kirjeldab kõige paremini Teie praegust suitsetamist?

1 – ei ole kunagi suitsetanud
2 – pean vahet või olen suitsetamisest loobunud
3 – suitsetan harvemini kui üks kord nädalas
4 – suitsetan mitmeid kordi nädalas, kuid mitte iga päev
5 – suitsetan kõige rohkem 9 sigaretti/sigareid/piibutubakat päevas
6 – suitsetan päevas 10-19 sigaretti/sigareid/piibutubakat
7 – suitsetan päevas vähemalt 20 sigaretti (või sama palju sigareid või piibutubakat)

**Õlu**: Kui suurel määral olete tarbinud õlu (keskmiselt) viimase aasta jooksul?

1 – mitte kunagi
2 – vähem kui pudel nädalas (1 pudel = 0,33 l)
3 – 1-4 pudelit nädalas
4 – 5-12 pudelit nädalas
5 – 13-või rohkem pudelit nädalas