

Peatükk 8

Korduspüügi meetodid

Kui palju elab Eestis mingit liiki linnukesi? Või kui palju on järves kalakesi?

Mingi huvipakkuva populatsiooni suuruse leidmiseks kasutatakse sageli korduspüügi meetodit (*Capture-Recapture*; *Mark-Recapture*). Selleks püütakse näiteks mõned linnud kinni (kui soovitakse teada lindude arvukust) ja märgistatakse nad. Seejärel lastakse püütud linnud uuesti lahti. Mõne aja pärast ja mõnes teises kohas — näiteks lindude rände ajal — püütakse jälle mõned linnud kinni ja vaadeldakse, kui suur osa kinnipüütud lindudest olid märgistatud. Märgistatud lindude osakaalu kasutades leitakse siis uuritava populatsiooni suurus.

Joonis 8.1: Märgistatud tuvi



8.1 Lincoln-Peterseni meetod

Olgu antud anum, milles on N kuulikest (kui suur on N , seda me ei tea). Võtame anumast välja M kuulikest ja märgistame ära punase värviga. Paneme kuulid anumasse tagasi ja segame hoolega. Nüüd seome silmad kinni ja võtame anumast uuesti välja n kuuli. Neist m kuuli osutusid värvituks. Kui suur on kuulide koguarv, N ? Kui teine valim on juhuslik, siis peaks värvitud kuulide osakaal teises valimis olema samasuur kui üldkogumis, ehk

$$\frac{M}{N} \approx \frac{m}{n}$$

kust võime leida Lincoln-Peterseni (või lihtsalt Peterseni) hinnangu populatsiooni kogusuurusele:

$$N \approx \frac{M}{m} \cdot n$$

Tasub märkida, et antud hinangu kasutamiseks peab teise valimisse sattumine ja märgi kandmine olema sõltumatud sündmused - me eeldame, et teise valimisse sattumise tõenäosus ei sõltu sellest, kas isendil/loomal on märk küljes või mitte. Samas ei pea aga märkide pealekandmine olema juhuslik (märgi saanud isendid ei pea olema juhuslikult valitud kogu populatsioonist) ega ka teine valim ei pea olema ilus ja esindav kogu populatsioonist jaoks — piisab vaid sellest, kui ühte ja teise valimisse sattumine on sõltumatud sündmused.

Näide

Tahetakse teada, kui palju insulte toimus Lõuna-Londonis aastatel 1995-1996. Paluti teavitada insuldihaigetest nii haiglaid kui perearste. Tulemused: haiglad teatasid 169 antud piirkonnast pärit insuldihaigest; perearstide teavitasid, et 80-l nende (vaatlusalusest piirkonnast pärit) patsiendil esines vaadeldud ajavahemikus insult. Patsiente, kes sattusid mõlemasse nimekirja oli 11. Seega toimus hinnanguliselt vaadeldud ajavahemikus $169 \cdot 80 / 11 = 1229$ insuldi juhtu.

Märkus: antud näites kasutatud andmed on pärit artiklist (Tilling, Sterne ja Wolfe, 2001), aga artiklis on kasutatud enam kui kahte allikat insuldihaigete leidmiseks.

8.1.1 Hinnangu täpsus

Kuidas saaksime kirjeldada leitud hinnangu täpsust? Ühte võimalust usalduspiiride leidmiseks tuntakse DeLury usalduspiiride nime all. Nende usal-

duspiiride konstrueerimise loogika on järgmine.

Märgiga isendite arv (teises) valimis on binoomjaotusega juhuslik suurus,

$$m \sim B(p = M/N; n).$$

Seega $E(m) = M/N \cdot n$ ja $D(m) = M/N(1 - M/N) \cdot n$.

Kui valimi suurus n kasvab, siis saab binoomjaotust lähendada normaalkaotuse abil. Seega, ligikaudu (ja suure n korral):

$$m \sim N(M/N \cdot n; M/N(1 - M/N) \cdot n).$$

Kui aga $X \sim N(\mu; \sigma^2)$, siis on 95%-usalduspiirid keskvaärtusele μ leitavad kui $X \pm 1,96\sqrt{\sigma^2}$. Seega võime ka kirja panna 95%-usalduspiirid suurusele $E(m)$:

$$\begin{aligned} 95\% - \text{usaldusintervall} &= m \pm 1,96\sqrt{M/N(1 - M/N) \cdot n} \\ &\approx m \pm 1,96\sqrt{m/n(1 - m/n) \cdot n} \\ &= m \pm 1,96\sqrt{m(1 - m/n)}. \end{aligned}$$

Siit aga

$$\begin{aligned} P(m - 1,96\sqrt{m(1 - m/n)} < M/N < m + 1,96\sqrt{m(1 - m/n)}) &\approx 0,95 \\ P(M/(m - 1,96\sqrt{m(1 - m/n)}) > N > M/(m + 1,96\sqrt{m(1 - m/n)})) &\approx 0,95 \end{aligned}$$

kust saamegi kätte soovitud (ligikaudsed) 95%-usalduspiirid populatsiooni suurusele:

$$\frac{M}{m + 1,96\sqrt{m(1 - m/n)}} \cdots \frac{M}{m - 1,96\sqrt{m(1 - m/n)}}.$$

8.2 Eeldustest

Lincoln-Peterseni hinanngu saamiseks oleme teinud järgmised eeldused:

1. Märgid ei lähe kaduma ja on alati loomal tuvastatavad;
2. populatsioon on kinnine (st loomi ei tule juurde ega lahku);
3. võetavad valimid on teineteisest sõltumatud.

Kui esimene eeldus poleks täidetud, siis ei pea märgiga loomade osakaal (teises) valimis enam olema võrdne märgistatud loomade osakaaluga kogu populatsioonis ja arutluskäigu aluseks olnud valem ei kehtiks. Kui kolmas eeldus ei kehti siis tekib sarnane olukord — märgiga loomade osakaal valimis ei oleks siis sama mis märgistatud loomade osakaal kogu populatsioonis.

Teist eeldust saab vajadusel aga veidi leevendada. Kui loomi peaks märgistamise ja (teise) valimi võtmise vahepeal populatsioonist kaduma (näiteks osad loomad surevad või sõidavad lõunasse talvepuhkusele), siis annab Peterseni hinnang ikkagi korrektse hinanngu populatsiooni suurusele — aga populatsiooni algsele suurusele. Kui samal ajal ei teki populatsiooni juurde uusi isendeid.

Kui aga populatsiooni lisandub vahepeal uusi märgistamata isendeid (aga ei kao vanu märgistatud isendeid) siis annab Peterseni hinanng valeliidse hinnangu populatsiooni lõppsuurusele (kui suur on populatsioon teise valimi võtmise ajal).

Kui kahe valimi võtmise vahepeal osad isendid surevad ja osad sünnivad, siis hakkab vaatlusalune hinnang populatsiooni suurust ülehindama ($\sim$ hinnatakse kõikide vaatlusaluses ajavahemikus antud populatsioonis elanud loomakeste koguarvu — aga populatsiooni enda suurus ei pruukinud kunagi, ühelgi ajahetkel olla tegelikult sedavõrd suur).

Lincoln-Peterseni hinnang

$$\hat{N} = \frac{M}{m}n$$

on üldjuhul lõpmatult suure nihkega:

$$E(\hat{N}) - N = \infty$$

sest juhul kui teise valimisse ei juhtu ühtegi märgistatud isendit $m = 0$ saame hinannguks lõpmatuse (ja seega ka hinanngu keskväärtuseks tuleb lõpmatus). Hinnangu nihke vähendamiseks on välja pakutud mitmeid parandatud hinnanguid, millest mainime siinkohal kahte — Chapmani (tuntud ka kui Seber'i) hinnangut ja Bailey hinnangut. Chapmani hinnang on mõeldud olukorra jaoks, kus teine valim on tagasipanekuta (väljapüütud märgiga kalad süüakse loodusuurijate poolt ära), Bailey hinnang on mõeldud aga tagasipanekuga teise valimi jaoks (sama märgistatud lind võib loodusvaatleja binokli ette mitmeid kordi sattuda).

Chapmani hinang näeb välja järgmine (lisatakse kunstlikult üks märgiga

isend, mis hiljem arvestusest eemaldatakse):

$$\hat{N}_C = \frac{M+1}{m+1}(n+1) - 1$$

$$\hat{D}(\hat{N}_C) = \frac{(M+1)(n+1)(M-m)(n-m)}{(m+1)^2(m+2)}.$$

Bailey hinnang on aga järgmise kujuga:

$$\hat{N}_B = \frac{M}{m+1}(n+1)$$

$$\hat{D}(\hat{N}_B) = \frac{(M)^2(n+1)(n-m)}{(m+1)^2(m+2)}.$$

Näide, baseerub artiklil Nixon, Edwards ja Eberhardt, 1967.

Ühes metsas Ohio osariigis püüti ja märgistati oravaid. Esimesel päeval püüti 22 oravat ($M = 22$), teisel päeval püüti 13 oravat ($n = 13$), nendest 7 olid juba märgistatud ($m = 7$).

Peterseni hinnang oravate arvukusele on:

$$\hat{N} = 13 \cdot 22/7 = 40,9$$

ja antud hinnangu ligikaudsed 95%-usalduspiirid DeLury meetodil oleksid [28...84] ning Leslie meetodil leitult [27...84].

Chapmani hinnang oravate koguarvule tuleb aga

$$\hat{N}_C = \frac{(M+1)(n+1)}{(m+1)} - 1 = 23 \cdot 14/8 - 1 = 39,2 \pm 7,1$$

sest antud hinnangu dispersioon hinnanguks tuleb:

$$\hat{D}(\hat{N}_C) = 23 \cdot 14 \cdot 15 \cdot 6 / (64 \cdot 9) = 50,31.$$

8.3 Sõltumatute valimite eeldusest

Üheks oluliseks eelduseks korduspüügi meetodite korral on see, et teine valim peab olema sõltumatu esimesest — märgiga ja märgita isenditel peab olema sama tõenäosus sattuda teise valimisse. Loodusuurijad üritavad seda

tingimust täita näiteks erinevaid püügiviise kasutades — lõksud või püügivõtted mida kasutatakse esimese valimi saamiseks (loomade märgistamiseks) peavad olema teistsugused kui need püügivõtted, mida pärast märkide olemasolu kontrollimiseks kasutatakse (näiteks võib linde algselt märgistada linnupoegadena pesas, aga hiljem rände ajal püüda võrkude abil). Sest muidu võivad osad märgistatud linnud õppida vältima lõkse kuhu nad juba varem sisse on sattunud.

Sõltumatu teise valimi saamiseks on vahel võimalik kasutada ka ühte teist tehnikat. Nimelt sageli on võimalik võtta juhuslikku valimit mingist suuremast populatsioonist. Näiteks kui tahame leida, kui paljud inimesed Eestis oskavad Bayesi valemit kasutada. Märkide külgepanekuga käib asi lihtsasti — kõik statistika eriala lõpetajad teavad Bayesi valemit ja oskavad seda ka kasutada — seega paneme kõigile antud eriala lõpetajatele mõtteliselt märgid külge (M). Seejärel võime võtta juhusliku valimi kogu populatsioonist ja määrata mingi lihtsa ülesande abil, kui paljud valimisse sattunud oskavad Bayesi valemit kasutada (n). Seejärel võime ka vaadata, kui palju valimisse sattunuist on need, kelle oskustes me juba õppeinfosüsteemis oleva informatsiooni põhjal olime veendunud (m).

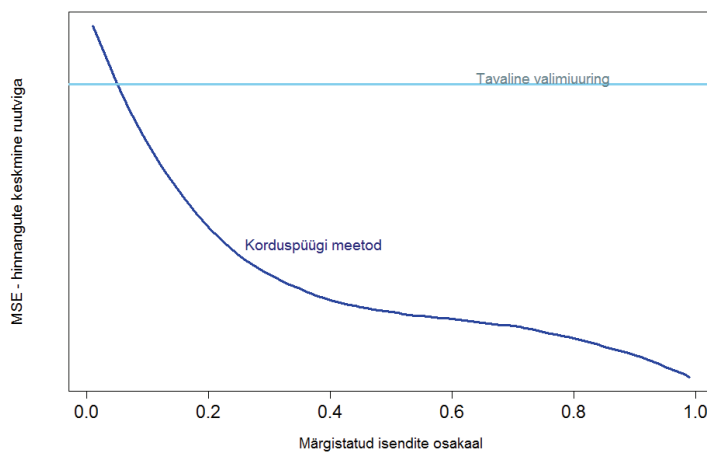
Miks sellises olukorras kasutada korduspüügi meetodit? Saaksime ju Bayesi valemi kasutajate osakaalu Eestis hinnata ka lihtsalt teise valimi põhjal, ja teades Eesti elanike arvu võiksime leida ka Bayesi valemi kasutajate koguarvu? Korduspüügi meetodi kasutamine võib lihtsalt viia täpsemate hinnanguteni. Kui märgistatud on suhteliselt suur osa huvipakkuvast populatsioonist on korduspüügi meetodid tavaliselt märksa täpsemad. Vaata ka joonist 8.2.

8.4 Korduspüügid enam kui ühel päeval

Vahel püütakse ja märgistatakse loomi mitmetel päevadel. Näiteks võisime oravaid püüda ja märgistada neljal päeval:

Päev	püütud loomi n_i	korduspüügid m_i	märgistatud M_i
1	22	0	0
2	13	7	22
3	15	10	28
4	10	5	33

Joonis 8.2: Korduspüügi meetodi kasutamine on enamasti täpsem kui tavaline valimiuring — vähemalt kui märgistatud on palju. Millisest märgistatute osakaalust alates muutub korduspüügi meetod täpsemaks sõltub ka teistest uuringu pisasjadest — näiteks sellest, kui suure osa uuritavast populatsioonist valimisse kaasame.



Kui võiksime valimeid sõltumatuteks lugeda, on meil mitmeid erinevaid võimalusi loomade arvukuse hindamiseks. Näiteks võiksime leida kolm erinevat hinnangut populatsiooni suurusele ja seejärel kombineerida saadud hinnangud näiteks metaanalüüsi abil. Traditsiooniliselt on aga kasutatud pigem nn keskmist Peterseni hinnangut, mis on tegelikult lihtsalt Chapmani hinnangute aritmeetiline keskmine:

$$\begin{aligned}\bar{\hat{N}} &= \frac{1}{3} \sum_{i=2}^4 \frac{(M_i + 1)(n_i + 1)}{(m_i + 1)} \\ &= \frac{1}{3} (23 * 14/8 + 29 * 16/11 + 34 * 11/7) \\ &= 45,3.\end{aligned}$$

Antud hinnangu dispersiooni on lihtne leida kasutades Chapmani hinnangu dispersioone:

$$\begin{aligned}
\hat{D}(\hat{N}) &= \frac{1}{3^2} \sum_{i=2}^4 \frac{(M_i + 1)(n_i + 1)(M_i - m_i)(n_i - m_i)}{(m_i + 1)^2(m_i + 2)} \\
&= \frac{1}{9} \left(23 * 14 * 15 * 6 / (8^2 * 9) + 29 * 16 * 18 * 5 / (11^2 * 12) + 34 * 11 * 28 * 5 / (6^2 * 7) \right) \\
&= 31,87.
\end{aligned}$$

Järjestikuste püüdmispäevade andmete analüüsimiseks on välja mõeldud ka mitmeid teisi meetodeid (Schnabeli meetod näiteks). Järjestikuste püüdmispäevade/püüdmislainete puhul on muidugi raske garanteerida valimite sõltumatust, sest loomad võivad õppida seatud lõksusi vältima.

8.5 Mitu valimit ja valimite sõltumatuse kontroll

Vahel on võimalik kasutada enam kui kahe valimi andmeid. Näiteks võib olla võimalik saada infot patsientide kohta kolmest või enamast allikast (peerarstid, haiglad, kindlustus, ...); rände ajal röntgenstatud loomi on võimalik vaadelda mitmel vaatluspäeval; pikaajalisi linde saab märgistada ja varem märgistatud lindude osakaalu leida igal aastal (näiteks viie aasta jooksul) vms.

Juhul, kui kasutada on enam kui kahe valimi andmed, saab sageli mõnda oletatud eeldustest pehmendada või siis saab kasutada täiendavaid andmeid tehtud eelduste paikapidavuse kontrollimiseks.

Vaatleme järgnevalt, kuidas saaks (saab) kontrollida valimite sõltumatust kasutades kolmanda valimi andmeid.

Olgu meil kasutada 3 valimit. Soovime kontrollida, et tegemist on sõltumatute valimitega. Vaata ka joonist 8.3.

Kui valimid A, B on sõltumatud, siis peaksime kasutades sagedusi d , c ja e -d saama hinnata g -d:

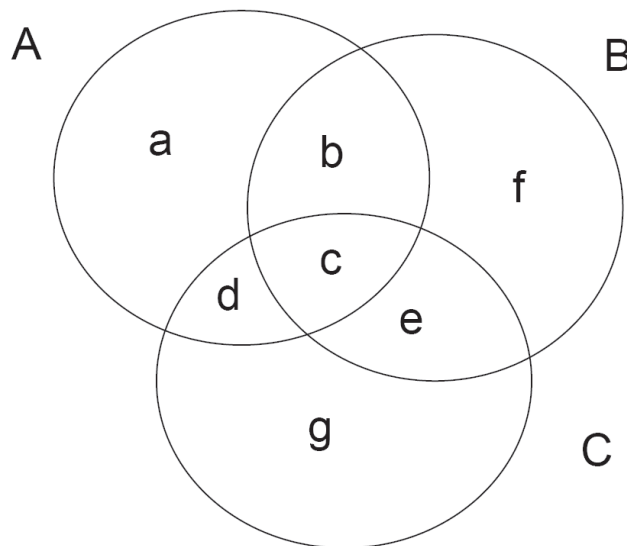
$$g = e/c \cdot d,$$

sest kui valimid oleksid sõltumatud, siis $g/d = e/c$. Siit saame, et sõltumatuse eeldusel $OR = \frac{g/d}{e/c} = 1$.

Kui aga $OR > 1$, siis on valimid positiivselt sõltuvad (valimisse A sattunud isendid kipuvad sattuma ka valimisse B) ja sellisel juhul me neid valimeid kasutades alahindame meid huvitava populatsiooni suurust.

Kui $OR < 1$ siis on valimid negatiivselt sõltuvad (valimisse A sattunud isendid väldivad valimisse B sattumist) ja tulemuseks on ülehinnang populatsiooni suurusele.

Joonis 8.3: Kolm valimit võimaldavad kontrollida valimite sõltumatuse eeldust



Näide

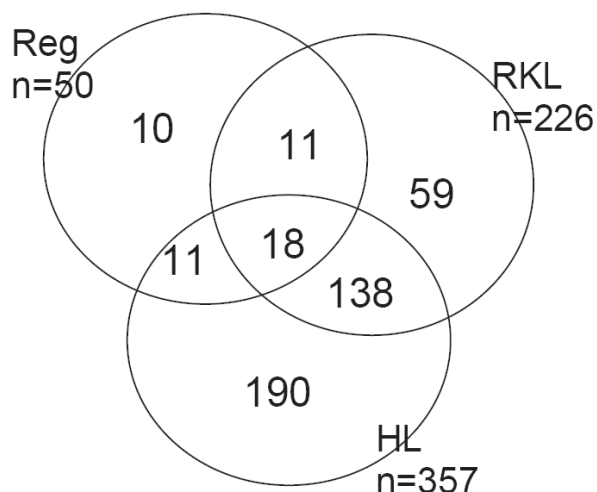
Prantsusmaal peavad arstid alates 1987.a. teavitama igast Legionella juhtumist. Selliseid teateid tuleb umbes 60 haigusjuhu kohta aastas. Sooviti kontrollida, kui efektiivne on loodud süsteem tegelikult haigusjuhtudest teavitamisel.

Andmeid 1995. a. aset leidnud haigusjuhtumite kohta koguti kolmest allikast:

- Riiklik register (Reg), kuhu arstid on kohustatud saatma teabe iga esinenud haigusjuhtumi kohta. Aastal 1995 laekus registrisse informatsioon 50 haigusjuhu kohta;
- Riiklik kesklaboratoorium (RKL), mis ainsana on Prantsusmaal võimaline panema lõplikku (ametlikku) Legionella diagnoosi. Selles laboratooriumis diagnoositi antud aastal 226-l patsiendil legionäärihaigus.
- Haiglate laboratooriumid (HL). Ka haiglate laboratooriumid on võimelised diagnoosima legionäärihaigust, kuid ametlikult tuleks positiivne proov saata kordusanalüüsiks ka riiklikku kesklaboratooriumi. Haiglate laboratooriumides diagnoositi legionäärihaigust 357-l korral.

Detailsemat juhtude jaotust iseloomustab joonis 8.4.

Joonis 8.4: Legionäärihaiguse juhud riiklikus kesklaboratooriumis (RKL), riiklikus registris (Reg) ja haiglates (HL) pandud diagnoosid



Joonisel toodud sageduste põhjal võime leida kolmandasse valimisse satumise šansside suhted, kontrollimaks kasutatavate allikate sõltumatust:

$$OR(Reg; RKL) = 2,4 [1,0 \dots 4,9]$$

$$OR(Reg; HL) = 0,7 [0,3 \dots 1,6]$$

$$OR(HL; RKL) = 1,5 [0,5 \dots 4,6]$$

Näeme, et riiklik register ja riiklik kesklaboratoorium on sõltuvad allikad, kui haigusjuhtum diagnoositakse kesklaboratooriumis siis on suhteliselt suurem tõenäosus, et vastavast haigusjuhtumist teavitatakse ka riiklikku registrit.

Üheks võimalikuks lahenduseks antud juhul oleks nende kahe sõltuva allika kombineerimine üheks allikaks. Sellisel juhul saaksime haigusjuhtude üldarvule järgmise hinnangu:

$$\hat{N} = \frac{M}{m}n = \frac{247}{167}357 = 528$$

ja 95%-usaldusintervalliks DeLury meetodiga saame (475...594).

8.6 Log-lineaarsed mudelid

Paljude valimite korral, kusjuures osad neist valimitest võivad olla sõltuvad — sellises olukorras on sageli abil log-lineaarsetest mudelitest (Poissoni regressioonist).

Tähistame:

n_{10} – isendid, kes esinesid ainult esimeses valimis;

n_{01} – isendid, kes esinesid ainult teises valimis;

n_{11} – isendite arv, kes esinesid nii esimeses kui ka teises valimis;

n_{00} – isendid, kes ei sattunud kumbagi valimisse;

p_1 – tõenäosus sattuda esimesse valimisse

p_2 – tõenäosus sattuda teise valimisse

N – populatsiooni suurus

Kasutades toodud tähistusi võime kirja panna järgmised võrrandid:

$$En_{11} = Np_1p_2$$

$$En_{10} = Np_1(1 - p_2)$$

$$En_{01} = N(1 - p_1)p_2$$

$$En_{00} = N(1 - p_1)(1 - p_2)$$

Kust peale logaritmimeerimist saame järgmised võrrandid:

$$\log(En_{11}) = \log(N) + \log(p_1) + \log(p_2)$$

$$\log(En_{10}) = \log(N) + \log(p_1) + \log(1 - p_2)$$

$$\log(En_{01}) = \log(N) + \log(1 - p_1) + \log(p_2)$$

$$\log(En_{00}) = \log(N) + \log(1 - p_1) + \log(1 - p_2)$$

Kasutades tähistusi

$$\mu := \log(N) + \log(1 - p_1) + \log(1 - p_2)$$

$$\alpha_1 := \log(p_1) - \log(1 - p_1)$$

$$\alpha_2 := \log(p_2) - \log(1 - p_2)$$

Võime varem logaritminimise järel saadud võrrandid esitada log-lineaarse mudelina:

$$\log(E n_{11}) = \mu + \alpha_1 + \alpha_2$$

$$\log(E n_{10}) = \mu + \alpha_1$$

$$\log(E n_{01}) = \mu + \alpha_2$$

$$\log(E n_{00}) = \mu$$

Kui hindame ülaltoodud mudeli parameetrid, siis näitab vabaliikme μ eksponent $\exp(\mu)$ meile nende isendite arvu, kes jäid välja mõlemast valimist.

Vaatame Ohio oravate näite varal ($M = 22$; $n = 13$; $m = 7$) log-lineaarse mudeli kasutamist R-is. Märkusena: mõlemasse valimisse sattus 7 oravat, ainult esimesse valimisse sattus 22-7 oravat, ainult teise valimisse sattus 13-7 oravat:

```
n_ii =c(7, 22-7, 13-7)
valim1=c(1, 1, 0)
valim2=c(1, 0, 1)
mudel=glm(n_ii ~ valim1 + valim2, family=poisson())
# Hinnang mõlemast valimist väljajäänud oravate arvule
exp(coef(mudel)[1])
# (Intercept)
# 12.85714

# Hinnang metsas elavate oravate koguarvule
exp(coef(mudel)[1])+sum(n_ii)
# (Intercept)
# 40.85714

# 95%-usaldusvahemik oravate koguarvule
exp(confint(mudel)[1,])+sum(n_ii)
# 2.5 % 97.5 %
# 31.70873 70.82602
```

Tasub tähele panna, et saadud hinnang oravate koguarvule langeb kokku Peterseni hinnangu abil saaduga, kuid usalduspiirid tulevad antud meetodil veidi teistsugused kui DeLury usalduspiirid. Mõlemad meetodid on usalduspiiride leidmiseks on vaid asümptootiliselt korrektsed (kehtivad vaid suurte valimite korral). Kuigi confint-käsu abil leitud usaldusvahemik on ka vaid

asümptootiliselt korrektne, võiks ta siiski olla enamasti DeLury usalduspiirist täpsem.

Log-lineaarseid mudeleid on kerge üldistada kasutamiseks rohkem kui kahe valimi korral, ühtlasi on võimalik mudelisse lisada ka parameetreid, mis lubavad kirjeldada valimite sõltuvust. Näiteks programm mis hindaks legionärihaiguse juhtude koguarvu prantsusmaal aastal 1995 (eeldades, et kõik kolm allikat ehk valimit on sõltumatud) näeks välja selline:

```
n_ii =c(10, 59, 190, 11, 138, 11, 18)
Reg  =c( 1,  0,   0,  1,   0,  1,  1)
RKL  =c( 0,  1,   0,  1,   1,  0,  1)
HL   =c( 0,  0,   1,  0,   1,  1,  1)

model=glm(n_ii ~ Reg + RKL + HL, family=poisson())
exp(coef(model)[1])+sum(n_ii)
#      520.851
```

ja kui soovime lubada, et Reg ja RKL on sõltuvad allikad või valimid (ühite neist sattumine tõstab või langetab tõenäosust, et sama haige sattuks ka teise registrisse), siis saame seda teha lisades mudelisse koosmõju:

```
model=glm(n_ii ~ Reg + RKL + Reg*RKL + HL, family=poisson())
exp(coef(model)[1])+sum(n_ii)
#      528.018
exp(confint(model)[1,])+sum(n_ii)
#  504.0007 559.6585
```

Noppeid meetodi ajaloost:

- 1662 kasutatud (Graunt) hindamaks Londoni elanike arvu;
- 1783 Laplace kasutas Prantsusmaa elanike arvu hindamiseks; Kasutas sünniregistreid + infot mõne piirkonna elanike koguarvu kohta.
- 1894 Peterson kasutab kalade arvukuse hindamisel (märgistatud kalad);
- 1930 Lincoln kasutas partide arvukuse hindamiseks;
- 1938 kasutati esmakordselt enam kui kahte valimit (Schnabel, kalapüük);
- Eestis kasutusel nii lindude, jõevähekide, konnade kui ka narkomaanide arvukuse hindamisel

Ülesanne

Kaunis Oguni linnas elas linnaisa, kes tahtis oma alamatel elu magusaks teha. Linnaisal oli aga mure – ta tahtis teada, kui paljudele elanikele see otsus ei meeldiks. Teadagi ei meeldi magus elu suhkruhaigetele. Õnneks on Oguni linn tsiviliseeritud paik, neil on seal isegi arvuti ja arvutis suhkruhaigete andmebaas, kust võis näha järgmist informatsiooni: Suhkruhaiged on linnas 234 tükki, neist 110 mehed ja 124 naised. Kokku elab linnas 3320 inimest, neist 1545 on mehed ja 1775 on naised. Muuseas, milline on suhkruhaiguse levimus?

Üks isekas ja tülikas arst arvas, et Oguni linna uhkus, tema suhkruhaigete register, ei kõlba kuhugi. Ta moodustas juhusliku valimi, 200 naist ja 200 meest. Valimisse valitutest nõustus uuringus osalema 122 meest ja 174 naist. Valimis osalejatest oli diabeet 10 mehel ja 16 naisel (Milline tuli iseka arsti hinnang haiguse levimusele?). Skandaali põhjustas aga asjaolu, et neist 10-st haigest mehest 5 polnud registrisse kantud, samuti polnud 7 valimiuuringu käigus avastatud haiget naist endast kunagi registrisse jälge jätnud. Registri loomiseks on kulutatud miljoneid. Sinu rahulik töökoht andmesisestajana on ohus, kui ei suudeta näidata, et registrist on kasu. Ega teised ei aita, leia parem ise registri andmeid kasutades suhkruhaiguse levimusele parem hinnang (st. täpsem hinnang), kui sel tülitekitajast arstil õnnestus saada (ja näita, et saadud tulemus on täpsem kui vaid valimiuuringu andmete põhjal leitud hinnang)!

bibliograafia

- [1] Charles M Nixon, William R Edwards ja Lee Eberhardt. “Estimating squirrel abundance from livetrapping data”. *The Journal of Wildlife Management* (1967), lk. 96–101.
- [2] Kate Tilling, Jonathan AC Sterne ja Charles DA Wolfe. “Estimation of the incidence of stroke using a capture-recapture model including co-variates”. *International journal of epidemiology* 30.6 (2001), lk. 1351–1359.