Kvalitatiivsete andmete analüüs

Kaur Lumiste, 2021

1 Sissejuhatus

Käesolevas kursuses vaatleme meetodeid, mis sobivad kvalitatiivsetele tunnustele. Vaatleme ühis-, marginaalseid ja tinglikke jaotusi. Vaatleme kahe ja enama tunnuse sagedustabeleid. Enamasti on eesmärgiks sõltuvuse kindlaks tegemine tunnuste vahel, samuti sõltuvuse tugevuse ja iseloomu uurimine. Vaatleme mitmeid statistilisi teste ja ka kvalitatiivse tunnuse modelleerimist.

1.1 Kvalitatiivne tunnus

Kvalitatiivne tunnus (categorical variable) on mittearvuline tunnus. Tema väärtusi nimetatakse sageli kategooriateks või tasemeteks.

Näide. *Tunnus "poliitiline vaade"*
väärtused (kategooriad/tasemed): liberaal, mõõdukas, konservatiiv, ei oska öelda.

Suur osa, kui mitte enamus, statistikas käsitletavaist tunnustest on kvalitatiivsed. Näiteid sotsiaalteadustest, meditsiinist, arvamus- ja turu-uuringutest, töökõu-uuringust, kvaliteedikontrollist jne. Võimalikud tunnuseväärtused pannakse paika **skaalaga**, millel tunnust mõõdetakse. Sama tunnust võib mõõta erinevatel skaaladel.

Näide. *Tunnus "juuste värv": skaala hele/tume või skaala heleblond, blond, ...brünett, must.*

Ka arvulise tunnuse võib skaala muutusega teha kvalitatiivseks.

Näide. *Tunnused "alkoholi sisaldus õlles (puhast alkoholi mahuühiku kohta, %)":*
"0-0.05%"(mitte-alkohoolne), "0.05-1%"(madala alkohooli sisaldusega),
"1-4%"(hele/kerge), "4-8%"(keskmise kangusega), "8-13%"(kanged).

1.2 Tunnuste liigitamine

Kvalitatiivsed tunnused liigituvad

- **Nominaaltunnused –** puudub loomulik järjestus tunnuse väärtuste vahel;

  Näide. *"usuline kuuluvus" – katoliiklane, protestant, judaism, islam, muu.*
  "liiklusvahend tööle sõiduks".

- **Järjestustunnused**(ordinaaltunnused) – on olemas järjestus, skaala või järgitakse kindlat hierarhiat.

Näide. *"sotsiaalne klass" – alam, kesk, kõrg.*
"Usaldus õigussüsteemi vastu" – 0 - "Üldse ei usalda", ... , 10 - "Usaldan täielikult".

Järjestustunnuse väärtused on küll järjestatavad, kuid kaugused väärtuste vahel pole määratavad. Kas alam- ja keskklass asuvad teineteisest sama kaugel, kui kesk- ja kõrgklass, ja kui suur see kaugus on? Probleem tekib paljude statistiliste meetodite rakendamisel. Proovige graafikul palga regressioonseost kujutada sõltuvana sotsiaalsest klassist (alam, kesk, kõrg).

- **Interval variable** – kaugused tunnuse väärtuste vahel on määratud. Kõik arvulised tunnused (ükskõik kas diskreetsed või pidevad) on intervalltunnused.

  Näide. *"Temperatuur (Celcius)", "Happelisus (pH)".*

- **Erijuhtum - Binaarne tunnus** – tunnus, millel on 2 väärtust (taset).

  Näide. *"Sugu", "HIV viiruse kandja", "Rämpspost".*

  Binaarne tunnus võib olla nii arvuline, järjestus, kui ka nominaalne tunnus. Ka nominaalse tunnusena on ta formaalselt järjestatav, kuigi mitte sisuliselt.

Hierarhia statistiliste meetodite rakendatavuses on järgmine:

nominaal $->$ järjestus $->$ intervall

Kõik nominaaltunnuste jaoks välja töötatud meetodid kõlbavad ka ülalpool aga mitte vastupidi. Samas on madalama taseme meetodid ülemise taseme jaoks vähem võimsamad, kui ülemise taseme omad meetodid.

Intervalltunnuse saab skaala muutusega teha järjestustunnuseks või nominaalseks. Kui aga andmed on kogutud madalama taseme skaalal, siis kõrgema taseme tunnust teha ei saa. Näiteks "haridus".

Ülesanne 1.1. *Mis liiki tunnusega on tegu?*

- *patsiendi elukestus (kuudes);*
- *lemmik toidupood (Konsum, Rimi, Selver, Maxima);*
- *vähi seisund pärast keemiaravi (paranenud, vähenenud, sama, suurenenud);*
- *perekonnaseis;*
- *Kas omad Toyotat?*

Antud kursuses vaatleme meetodeid, mis on välja töötatud nominaal- ja järjestustunnuste jaoks. Meetodeid saab rakendada ka väheste väärtustega intervalltunnustele. Skaalamuutusega saab intervalltunnuse väärtusi alati vähendada.

Tunnuseid liigitatakse ka nende koha järgi analüüsis ja tähistame vastavalt Y ja X:

Y	X
uuritav tunnus	seletav tunnus
sõltuv tunnus	sõltumatu tunnus
funktsioontunnus	argumenttunnus
response variable	*explanatory variable*

1.3 Uuringute liigitamine

Uuringu liigi teadmine on äärmiselt tähtis küsimus statistilise analüüsi tegemisel. Uuringu liigist sõltuvalt võivad olla mõned tunnused juhuslikud, mõned aga fikseeritud (mittejuhuslikud). Statistilisi otsustusi saab teha juhusliku tunnuse kohta. Fikseeritud tunnuseid saab kasutada seletavate tunnuste rollis.

1. **Prospektiivne ehk etteulatuv uuring**. Üldkogumist moodustatakse valim või eraldatakse teatava tunnuse alusel kõik objektid. Edasi on 2 erijuhtu:

   - **kohortuuring**. Eraldatud kogumit jälgitakse teatud ajaperioodi vältel ja registreeritakse sündmused, mis nendega toimuvad.
   - **kliiniline katse**. Valitud objektid allutatakse juhuslikult (randomiseeritakse) gruppidesse. Grupid saavad erinevat ravi või muud mõjutust. Gruppe jälgitakse teatud aja jooksul.

   Öeldakse, et kliiniline katse on ainus *eksperimentaalne* statistiline uuring.

   Huvipakkuv tunnus Y saab oma väärtuse objekti jälgimisperioodil, ta on juhuslik. Tunnuse X alusel valisime jälgitavad objektid või määrame ravigrupi, see on mittejuhuslik.

   Näide. *Võtame uurimise alla inimesed kes suitsetavad (X suitsetamine) ja jälgime neid, kas neil tekib kopsuvähk (Y) 2 aasta jooksul.*

   Näide. *Vaktsineerime inimesed COVID-19 vastu ära (X vaktsineerimine) ja jälgime kas nad haigestuvad uuesti või kas neil püsivad antikehad (Y) päras 1 aasta möödumist.*

2. **Retrospektiivne ehk tagasivaatav uuring**. Objektid valitakse uuritava tunnuse põhjal ja seejärel saadakse nende seletavad tunnused. Peamine uuringutüüp siin on

   - **juht-kontroll uuring** (case-control study). Ühes grupis on "juhud", need, kellel esineb meid huvitav omadus, teises grupis on "kontrollid", need, kellel seda omadust ei esine.

   Näide. *Y infarkt, X suitsetamine kauem kui 3 aastat. Juhtgruppi võetakse haigla infarktihaiged. Kontrollgruppi terved või selle haigla muud haiged. Küsitletakse suitsetamise kohta.*

3. **Ristlõikeline uuring** (cross-sectional study). Objektid valitakse juhuslikult. Objektide tunnused registreeritakse fikseeritud ajamomendil (ajamoment võib ka pikk olla, näiteks aasta). Ka tulemused käivad selle ajamomendi kohta. Paljudes olukordades

ei vali me objekte ise, vaid need söödab meile ette mingi juhuslik protsess.

Seda tüüpi uuringus on kõik tunnused Y ja X juhuslikud.

Näide. *Leibkonnauuring, tööjõu-uuring on juhusliku valimiga elanike üldkogumist. Aga kindlustuskompaniile laekuvad kahjunõuded teataval kuul on juhusliku protsessi tulemus.*

Märkus valikuteooria kohata. Valikuteoorias saadakse juhuslik valim lõplikust üldkogumist keerulise valikudisainiga, mida tuleb hinnangute ja nende omaduste leidmisel arvesse võtta. See on eraldi statistikaharu. Ka antud aines võib tegu olla lõplike üldkogumitega, kuid ristlõikelise valimi kohta eeldatakse tavaliselt lihtsat juhuslikku valikut (kas TTA või TGA).

Kordame veelkord, et statistilise analüüsi aluseks on juhuslikkuse mõiste. Juhuslikkust arvestades, saame rääkida hinnangute nihketusest, nende dispersioonist, usalduspiiridest tundmatule näitajale. Ka modelleerida saame juhuslikku uuritavat tunnust.

Kui tunnus ei ole juhuslik, siis tema näitajad (keskmine, osakaal, dispersioon) kehtivad ainult valimis, üldkogumit nad ei iseloomusta. Samuti ei ole mõtet standardvigadel, usalduspiiridel ega olulisustõenäosustel.

1.4 Jaotused kvalitatiivse tunnuse korral

Kvalitatiivsed tunnused on diskreetsed, st võimalike tasemete (väärtuste) arv on lõplik ja tavaliselt väike.

Kvalitatiivse **tunnuse jaotuseks** nimetame tasemete tõenäosusi. Need saab anda tabelina. Vaata näiteks Tabel 1. Suurused π_i on nn üldkogumi tõenäosused, mis on tundmatud:

Tabel 1: Tunnuse "kahjunõude tüüp" jaotus

tulekahju	veekahju	vandalism	muu
π_1	π_2	π_3	π_4

$\pi_1 = P(tulekahju)$.

Kuidas hinnata suurusi π_i, kui valimiväärtusteks on kategooriad:
tulekahju, tulekahju, muu, veekahju, ...

Sellistelt andmetelt ei saa arvutada enamust statistikas kasutatavaid valimifunktsioone (keskväärtust, dispersiooni, mediaani). Õnneks saab üle minna arvudele, milleks on väärtuste esinemissagedused.

Juhusliku tunnuse esinemissagedused on juhuslikud (teises valimis teistsugused). Sageduste juhuslikku olemust on loomulik kirjeldada jaotusega, mille parameetril on sisuline tähendus ja mille hindamine pakub meile huvi.

Antud näites olgu $n_1 =$ tulekahjude arv vaatluskuul. Selleks, et n_1 abil hinnata tõenäosust π_1, on vaja teada eeldusi n_1 kujunemisel: kas igakuised vaatlused on sõltumatud, kas tõenäosus π_1 on sama igal kuul. Teame, et sellisel juhul on tõenäosuse π_1 nihketa hinnanguks osakaal n_1/n. Hinnangu teised omadused, sealhulgas dispersioon, sõltuvad aga n_1 jaotusest. On mitmesuguseid kandidaatjaotusi.

1.5 Sageduse jaotus

Olgu Y I tasemega kvalitatiivne tunnus, kus tasemete koodid on $1, 2, \ldots, I$. Olgu n_i taseme i sagedus valimis. Tabelis 2 on sagedused.

Tabel 2: Kvalitatiivse tunnuse sagedustabel

Y	1	...	i	...	I
sagedus	n_1	...	n_i	...	n_I

Konkreetses valimis on n_i fikseeritud arv, mittejuhuslik. Meid huvitab aga n_i juhuslik olemus. Juhuslikku olemust kirjeldab jaotus. Tabelis 3 kirjeldame klassikalised jaotused, mida kasutatakse n_i juhusliku olemuse kirjeldamiseks.

Tabel 3: Jaotusmudelid sageduse jaoks

i tunnuse Y tase, n_i (siin juhuslik) väärtuse i esinemissagedus, π tõenäosus i esinemiseks ühes katses

Jaotus	$P(n_i = x),\ x = 0, 1, ...$	Keskv.	Dispersioon	Tekkemehhanism
Binoom $B(n, \pi)$				
Poisson $Po(\mu)$				
Neg.bin. $NB(r, \pi)$				

On palju teisi diskreetseid jaotusi, mida saab kasutada sageduse kui diskreetse juhusliku suuruse kirjeldamiseks. Uuematest on juttu artiklis Bakouch et. al (2014). A new discrete distribution. *Statistics* , vol. 48, 200-240. Väljapakutud uus jaotus on

$$p(x) = \frac{p^x}{1+\theta}[\theta(1-2p) + (1-p)(1+\theta x)],$$

kus $p = \exp(-\theta),\ \theta > 0,\ x = 0, 1, \ldots$. Keskväärtuse ja dispersiooni valemid ei ole nii lihtsad. Näiteks

$$EX = \frac{p(2\theta - \theta p + 1 - p)}{(1+\theta)(1-p)^2}.$$

Sageduse n_i jaotuse valikul:

- vaadatakse n_i tekkemehhanismi (uuringu tüüp)
- proovitakse sobitada sobiv jaotus (kui on pikemajaliselt sageduse andmed saadaval)
- sageli eelistatakse lihtsamat (Poissoni jaotus)

Poissoni jaotuse sobitamiseks andmetele on kitsendavaks asjaoluks keskväärtuse ja dispersiooni võrdsus. Selle tõttu ei sobi Poissoni jaotus igasuguse sündmuse esinemissagedust kirjeldama. Selle tõttu räägime ala- ja üledispersioonist.

Tabelist 3 näeme, et

Binoom jaotus	$B(n,\pi)$	$Dn_i < En_i$
Poissoni jaotus	$Po(\mu)$	$Dn_i = En_i$
Poissoni jaotus	$NB(r,\pi)$	$Dn_i > En_i$

Millist jaotusmudelit andmetel kasutada? Kui andmed viitavad, et

$$En_i = Dn_i \Rightarrow \textbf{Poissoni jaotus}$$

Kuidas veenduda, et Poissoni mudel (jaotus) sobib sageduse n_i jaoks? Teooria kohaselt

Kui $Dn_i = En_i$	$\Rightarrow$ Sobib Poissoni jaotus
Kui $Dn_i > En_i$	$\Rightarrow$ Üledispersioon, peaks sobima Neg.Binoom-jaotus
Kui $Dn_i < En_i$	$\Rightarrow$ Aladispersioon, peaks sobima Binoom-jaotus

Aga reaalses elus ei ole Dn_i ja En_i teada. Võimalik hinnata kui oleks n_i-d mitmest valimist (enamasti on aga meil ainult üks sageduse vaatlus).

Näide 1.1. *Olgu Y = kahjunõude tüüp, fikseerime ühe konkreetse tüübi i - tulekahju ning vaatleme sagedust n_i - tulekahjude arv aastas. Kui on vaadeldud mitu aastat (k aastat), siis on meil sageduse n_i valim:*

$$n_{i1}, n_{i2}, n_{i3}, \ldots, n_{ik}.$$

Siit saame arvutada valimi keskmise ja valimi dispersiooni.

Üledispersiooni olemasolu saame aimata isegi valimit vaatamata. Kui katsetingimused ei ole samad. Näiteks väga külm talv põhjustab suuremat tulekahjude arvu kui Poissoni jaotusele kohane oleks.

Näide 1.2. *Olgu m putukate arv, kes jäävad ellu fikseeritud mürgidoosi korral. Kui n on kogu putukate arv, siis*

$$m \sim B(n,\pi),$$

kus π on ellujäämise tõenäosus ühel putukal fikseeritud tingimuste korral.

Kui nüüd n putukat allutatakse samale doosile kuid tingimused on erinevad, siis $Dn_i > n\pi(1-\pi)$.

1.6 Sageduste vektori jaotus

Olgu Y I tasemega kvalitatiivne tunnus. Olgu $(n_1, n_2, \ldots, n_I)$ tasemete vaadeldud sagedused. Eelmises peatükis kontsentreerusime ühele sagedusele n_i ja vaatasime jaotusi, mis sobivad tema juhusliku olemuse kirjeldamiseks (Tabel 3). Siin huvitume sageduste vektori $(n_1, n_2, \ldots, n_I)$ ühisjaotusest. Teisisõnu küsime tõenäosust

$$p(n_1, n_2, \ldots, n_I) = P(\text{saada väärtuste vektorit } (n_1, n_2, \ldots, n_I)).$$

Kui sagedused kui juhuslikud suurused on sõltumatud, siis on ülesanne lihtne. Tähistades $p(n_i) = P(\text{saada väärtust } n_i)$, saame

$$p(n_1, n_2, \ldots, n_I) = \prod_{i=1}^{I} p(n_i). \tag{1}$$

Näiteks kui $n_i \sim Po(\mu_i)$ on sõltumatud, siis

$$p(n_1, n_2, \ldots, n_I) = \prod_{i=1}^{I} [e^{-\mu_i} \frac{\mu_i^{n_i}}{n_i!}]. \tag{2}$$

Sõltumatute n_i korral on kogu valimimaht $n = \sum_{i=1}^{I} n_i$ juhuslik suurus. Selline olukord on kvalitatiivse tunnuse vaatlemisel juhuslikus protsessis (näiteks tunnuse "kahjunõude tüüp" korral on nii ühes kuus esinevate kahjude sagedused n_i kui ka koguarv n juhuslikud suurused).

Kui aga valimimaht n on fikseeritud, siis sagedused n_i ei ole sõltumatud. Nad ju summeeruvad fikseeritud arvuks. Sel juhul tuleb sageduste vektori jaotuse kirjeldajana mängu uus jaotus – **multinomiaaljaotus**.

Multinomiaaljaotuse tekkemehhanism on järgmine. Olgu kvalitatiivse tunnuse Y üldkogumijaotus antud tabeliga 4.

Tabel 4: Kvalitatiivse tunnuse üldkogumijaotus

Y	1	...	i	...	I
$P(Y=i)$	π_1	...	π_i	...	π_I

Üldkogumist võetakse juhuslikult ja sõltumatult n objekti, igal objektil mõõdetakse tunnust Y. Tõenäosus on π_i, et mõõtmistulemuseks osutub tase i. Lõpuks koostatakse andmetelt sageduste vektor $(n_1, n_2, \ldots, n_I)$. Selle vektori juhuslikku olemust kirjeldabki multinomiaaljaotus. Multinomiaaljaotust tähistame, näidates jaotuse parameetrid:

$$(n_1, n_2, \ldots, n_I) \sim M(n; \pi_1, \pi_2, \ldots, \pi_I).$$

Multinomiaaljaotuse tõenäosused arvutatakse valemiga:

$$p(n_1, n_2, \ldots, n_I) = \frac{n!}{\prod_{i=1}^{I} n_i!} \prod_{i=1}^{I} \pi_i^{n_i}, \tag{3}$$

kus $\sum_{i=1}^{I} \pi_i = 1$ ja $\sum_{i=1}^{I} n_i = n$. Multinomiaaljaotuse omadused:

- iga 1-dimensionaalne marginaaljaotus on binoomjaotus, $n_i \sim B(n, \pi_i)$ (tuleneb tekkemehhanismist);
- $E(n_i) = n\pi_i$, $D(n_i) = n\pi_i(1 - \pi_i)$ (binoomjaotuse omadused);
- $Cov(n_i, n_j) = -n\pi_i\pi_j$ (miinusmärk tuleneb fikseeritud summast n).

Ülesanne 1.2. On 100 valikvastustega küsimust. On 4 valikut ja üks valik igas küsimuses on õige. Tudeng ei ole õppinud ja pakub vastuseid juhuslikult. Olgu m õigete vastuste arv.

- Mis on m jaotus?
- Leia oodatav õigete vastuste arv $E(m)$ ja leia $D(m)$.
- Kas oleks üllatav, kui tudeng annaks vähemalt 50 õiget vastust?
- Olgu n_i i-nda variandi valikute arv. Mis jaotusega on (n_1, n_2, n_3, n_4)?
- Leia tõenäosus, et tudeng valib kõiki vastusevariante ühepalju.

Realiseerunud sagedused $(n_1, n_2, \ldots, n_I)$ on teada, aga tõenäosused $(\pi_1, \pi_2, \ldots, \pi_N)$ on üldjuhul tundmatud. Kuidas neid sageduste valimilt hinnata? Leiame hinnangud suurima tõepära meetodil.

1.7 Multinomiaaljaotuse parameetrite hinnangud

Paneme tähele, et multinomiaaljaotuse parameetrid on tõenäosused π_i, mis kvalitatiivse tunnuse korral iseloomustavad tasemete jaotust. Näiteks silmade värvi (sinised, pruunid, rohelised, mustad) jaotust. Kasutame suurima tõepära meetodit, et saada tõenäosustele valimhinnangud.

Olgu meil sageduste valim $(n_1, \ldots, n_I)$, $n = \sum_{i=1}^{I} n_i$. Teame, et nad pärinevad multinomiaaljaotusest (vt tekkemehhanism). Siis tõenäosus saada sellist sageduste valimit, $p(n_1, \ldots, n_I)$, avaldub seosega (3). Väärtused π_i, mis maksimeerivad selle tõenäosuse ongi suurima tõepära hinnangud:

$$(\hat{\pi}_1, \ldots, \hat{\pi}_I) = \arg \max_{\pi_1, \ldots, \pi_I} p(n_1, \ldots, n_I).$$

Maksimeerimiseks π_i järgi ei ole vaja vaadata tegureid, mis neist ei sõltu, siin $n!/\prod_{i=1}^{I} n_i!$. Seega on vaja maksimeerida

$$L = \prod_{i=1}^{I} \pi_i^{n_i}$$

kitsenduste $\sum_{i=1}^{I} \pi_i = 1,\ \pi_i > 0\ \forall i$ olemasolul. Käsikirjaline materjal

Ülesanne 1.3. Olgu sageduste vektor sõltumatute komponentidega, kusjuures $n_i \sim Po(\mu)\ \forall i$. Leida μ suurima tõepära hinnang.

Ülesanne 1.4. Ilmnegu genotüübid (AA, Aa, aa) tõenäosustega $(\theta^2, 2\theta(1-\theta), (1-\theta)^2)$. Olgu meil multinomiaalne valim suurusega n ja genotüüpide sagedustega (n_1, n_2, n_3). Pane kirja tõepärafunktsioon, logaritmiline tõepärafunktsioon ja näita, et $\hat{\theta} = (2n_1 + n_2)/(2n_1 + 2n_2 + 2n_3)$

2 Kvalitatiivsete tunnuste ühisjaotus

Eelmises peatükis vaatlesime ühte kvalitatiivset tunnust Y, tema üldkogumijaotust, tema vaatlemisel tekkivaid sagedusi ning nende jaotusi. Selles peatükis vaatleme kahe kvalitatiivse tunnuse üldkogumijaotust ja defineerime seonduvad mõisted.

Olgu X tasemete arvuga I ja Y tasemete arvuga J kaks kvalitatiivset tunnust (sobib ka väheste väärtustega arvuline tunnus). Paari (X, Y) võimalike väärtuste arv on $I \times J$. Võimalike väärtuste tõenäosused $P(X = i, Y = j) = \pi_{ij}$ esitavadki (X,Y) ühisjaotust (vt Tabel 5).

Tabel 5: Kahe tunnuse ühisjaotus

$X \backslash Y$	1	...	j	...	J	X jaotus
1	π_{11}	...	π_{1j}	...	π_{1J}	π_{1+}
...	...	...	...	...	...	
i	π_{i1}	...	π_{ij}	...	π_{iJ}	π_{i+}
...	...	...	...	...	...	
I	π_{I1}	...	π_{Ij}	...	π_{IJ}	π_{I+}
Y jaotus	π_{+1}	...	π_{+j}	...	π_{+J}	1

Ühisjaotusega $\{\pi_{ij}\}$ on määratud palju muid jaotusi.

a) Marginaaljaotused (tabeli viimases veerus ja reas):

$$\begin{aligned} P(X = i) &= \pi_{i+} = \sum_{j=1}^{J} \pi_{ij},\ i = 1, \ldots, I, \\ P(Y = j) &= \pi_{+j} = \sum_{i=1}^{I} \pi_{ij},\ j = 1, \ldots, J \\ 1 &= \sum_{i} \pi_{i+} = \sum_{j} \pi_{+j} = \sum_{i,j} \pi_{ij}. \end{aligned}$$

b) Tinglikud jaotused. Tunnusel Y on I tinglikku jaotust, üks jaotus iga fikseeritud X väärtuse korral. Y tinglikud jaotused paiknevad tabeli ridades ja neid nimetatakse ka rea tinglikeks jaotusteks. X tinglikud jaotused paiknevad tabeli veergudes, neid on J tükki ja neid nimetatakse veerutinglikeks jaotusteks. Muidugi võib tunnuseid X ja Y vahetada, nii et meid huvitavad tinglikud jaotused paikneksid alati ridades.

Toome tähised ja seosed Y tingliku jaotuse jaoks:

$$P(Y = j|X = i) = \pi_{j|i},\ j = 1, \ldots, I.$$

Tinglikud tõenäosused avalduvad ühistõenäosustest,

$$\pi_{j|i} = \frac{\pi_{ij}}{\pi_{i+}}$$

ja need moodustavad ridapidi tõepoolest jaotuse

$$\sum_j \pi_{j|i} = \sum_j \frac{\pi_{ij}}{\pi_{i+}} = \frac{1}{\pi_{i+}} \sum_j \pi_{ij} = 1$$

Üldkogumijaotuse tõenäosused on tundmatud. Meile pakub huvi hinnata ja kontrollida

Tabel 6: Tunnuse Y tinglikud jaotused

$X \backslash Y$	1	...	j	...	J	summa
1	$\pi_{1\|1}$	...	$\pi_{j\|1}$	...	$\pi_{J\|1}$	1
...	...	...	...	...	...	
i	$\pi_{1\|i}$	...	$\pi_{j\|i}$	...	$\pi_{J\|i}$	1
...	...	...	...	...	...	
I	$\pi_{1\|I}$	...	$\pi_{j\|I}$	...	$\pi_{J\|I}$	1

hüpoteese tõenäosuste $\pi_{ij}, \pi_{j|i}, \pi_{i+}, \pi_{+j}$ kohta. Eriti suurt huvi pakub sõltuvuse kontrollimine. Kas Y (kopsuvähk) sõltub tunnusest X (suitsetatud sigarettide arv)? Kui Y tinglik jaotus X tasemetel on sama, siis Y ei sõltu X-st. Tinglikud jaotused ise ei ole teada, saame kasutada hinnanguid. NB! Vastavalt uuringu tüübile saab hinnata ainult teatavaid tõenäosusi.

N. Reaalselt eksisteeriv üldkogumijaotus. Kui üldkogumiks on lõplik hulk, siis tunnuste vaatlemisel kogu üldkogumis on võimalik üldkogumijaotus reaalselt kirja panna (vt Tabel 7).

Tabel 7: Kahe tunnuse üldkogumijaotus N-riigis
X-sugu, Y-värvipimedus. Sulgudes tinglik jaotus.

$X \backslash Y$	1 (värvipime)	2 (ei)	$P(X = i)$
1 (mees)	0.05	0.45	0.5
	(0.1)	(0.9)	(1)
2 (naine)	0.01	0.49	0.5
	(0.02)	(0.98)	(1)
$P(Y = j)$	0.06	0.94	1

N. Fiktiivne ehk ettekujutatav üldkogumijaotus. Selleks on tõenäosused uuritavas juhuslikus protsessis. Näiteks Y - tekitatava kahju suurus liiklusõnnetuses (väike, keskmine, suur), X - auto mark. Jälgida saame vaid selle protsessi realisatsioone.

2.1 Sagedustabel

Sagedustabel moodustub valimist. Selle abil saab uurija hinnata ühisjaotust π_{ij} või tinglikke jaotusi $\pi_{j|i}$. NB! Mitte igasuguselt tabelilt ei saa hinnata mõlemaid.

Olgu n valimimaht, kõigi vaadeldud objektide arv ja $n_{ij} = \#\{X = i, Y = j\}$ väärtuspaaride (i, j) arv valimis.

Tabel 8: Kahe tunnuse sagedustabel

$X\backslash Y$	1	...	j	...	J	X sagedused
1	n_{11}	...	n_{1j}	...	n_{1J}	n_{1+}
...	...	...	...	...	...	
i	n_{i1}	...	n_{ij}	...	n_{iJ}	n_{i+}
...	...	...	...	...	...	
I	n_{I1}	...	n_{Ij}	...	n_{IJ}	n_{I+}
Y sagedused	n_{+1}	...	n_{+j}	...	n_{+J}	n

Siin n_{i+} ja n_{+j} on marginaalsagedused ehk

$$n_{i+} = \sum_{j=1}^{J} n_{ij} = \#\{X = i\} \text{ on vaatluste arv kus tunnus } X \text{ omandab taseme } i$$

$$n_{+j} = \sum_{i=1}^{I} n_{ij} = \#\{Y = i\} \text{ on vaatluste arv kus tunnus } Y \text{ omandab taseme } j$$

Pane tähele, et

$$n = \sum_i \sum_j n_{ij} = \sum_i n_{i+} = \sum_j n_{j+}.$$

2.2 Juhuslikkus sagedustabelis ja selle kirjeldamine

Olgu vaadeldud kahte tunnust X ja Y tasemetega I ja J ning üldkogumijaotusega $\{\pi_{ij}\}$. Olgu saadud sagedustabel $\{n_{ij}\}$. Loogelistes sulgudes on kõigi tõenäosuste/sageduste tabel, $i = 1, \ldots, I$ ja $j = 1, \ldots, J$. Olgu marginaaljaotused $\{\pi_{i+}\}$ ja $\{\pi_{+j}\}$ ja vaadeldud marginaalsagedused $\{n_{i+}\}$ ja $\{n_{+j}\}$, need on vektorid, vastavalt $i = 1, \ldots, I$ ja $j = 1, \ldots, J$. Koguvalimimaht olgu n.

Sõltuvalt uuringu tüübist on mõned nendest sagedustest mittejuhuslikud, meie endi poolt fikseeritud arvud. Sellest tulenevalt muutub ka $\{n_{ij}\}$ juhuslik loomus.

Ristlõikelise uuringu korral, kus n on fikseeritud ja objektid valitakse juhuslikult (lihtne juhuvalik tagasipanekuga mõttes), on

$$\begin{aligned} \{n_{ij}\} &\sim M(n, \{\pi_{ij}\}), \\ \{n_{i+}\} &\sim M(n, \{\pi_{i+}\}), \\ \{n_{+j}\} &\sim M(n, \{\pi_{+j}\}), \\ n_{ij} &\sim B(n, \pi_{ij}). \end{aligned}$$

Kõik tõenäosused nii ühis-, marginaalsed on hinnatavad osakaaludega. Näiteks $\hat{\pi}_{ij} = n_{ij}/n$ ja $\hat{\pi}_{i+} = n_{i+}/n$. Need on suurima tõepära hinnangud ja on ka nihketa hinnangud.

Vaatame tingliku tõenäosuse hindamist sellise sagedustabeli korral. Tinglik tõenäosus on funktsioon ühis- ja marginaaltõenäosustest:

$$\pi_{j|i} = \frac{\pi_{ij}}{\pi_{i+}}.$$

Kasutame **suurima tõepära hinnangu invariantsuse omadust** – kui on tegu parameetrite funktsiooniga, siis asendades tundmatud parameetrid suurima tõepära hinnangutega, saame funktsioonile suurima tõepära hinnangu.

Kirjutame välja suurima tõepära hinnangu tinglikule tõenäosusele:

$$\hat{\pi}_{j|i} = \frac{\hat{\pi}_{ij}}{\hat{\pi}_{i+}} = \frac{n_{ij}/n}{n_{i+}/n} = \frac{n_{ij}}{n_{i+}},$$

saime, et ka see on sobivalt leitud osakaal.

N. Sotsiaaluuringus (vt Agresti (1996, lk.17)) küsitleti $n = 1091$ isikut: X="sugu", Y="teine elu", kus "teine elu" tähendab usku elusse pärast surma. Saadi sagedustabel

Tabel 9: Tunnuste "sugu" ja "teine elu" sagedustabel

$X \backslash Y$	1 (jah)	2 (ei)	summa
1 (naine)	435	147	582
2 (mees)	375	134	509
summa	810	281	1091

Sotsiaaluuring on tavaliselt ristlõikeline uuring.

Ülesanne 2.1. Leia hinnangud ja interpreteeri: $\hat{\pi}_{11}$, $\hat{\pi}_{21}$, $\hat{\pi}_{1|1}$, $\hat{\pi}_{1|2}$, $\hat{\pi}_{+1}$, $\hat{\pi}_{1+}$.

Ülesanne 2.2. Olgu üldkogumijaotuseks soo-värvipimeduse jaotus tabelis 7. Üldkogumist on juhuslikult valitud $n = 20$ isikut. Pane kirja tabeli sageduste ühisjaotus. Leia oodatavad sagedused? Leia tõenäosus, et $n_{11} = 2$, $n_{12} = 8$, $n_{21} = 1$, $n_{22} = 9$. Pane kirja n_{11} kui juhusliku suuruse jaotus.

Prospektiivse uuringu korral on uuritav tunnus Y juhuslik, aga seletav X sageli mittejuhuslik. St oleme ise fikseerinud objektide arvud $\{n_{i+}\}$ tunnuse X tasemetel. Need ei ole juhuslikud. Nende abil ei saa hinnata π_{i+}, samuti ei saa $\{n_{ij}\}$ abil hinnata ühistõenäosusi $\{\pi_{ij}\}$. Küll aga on uuritav tunnus juhuslik igal X tasemel ja sagedused ridades alluvad jaotusele:

$$\{n_{ij},\ \text{reas } i\} \quad \sim \quad M(n_{i+}, \{\pi_{j|i}\}).$$

Hinnatavad on tinglikud tõenäosused ridades, vastavate suhteliste sagedustega.

N. Viidi läbi arstide 5-aastane terviseuuring (vt Agresti (1996, lk.20)). Uuriti aspiriini mõju südame infarktile (myocardical infarction). Arstid randomiseeriti kahte gruppi, üks grupp sai regulaarselt aspiriini ja teine platseebot. Kas aspiriini võtmine vähendab suremust südame infarkti?

Ülesanne 2.3. *Missuguseid tõenäosusi on mõttekas hinnata? Leia* $\hat{\pi}_{1|1}$, $\hat{\pi}_{1|2}$.

Tabel 10: Tunnuste X="aspiriin" ja Y="südame infarkt" sagedustabel

$X\backslash Y$	1 (jah)	2 (ei)	summa
1 (platseebo)	189	10 845	11 034
2 (aspiriin)	104	10 933	11 037

Retrospektiivse uuringu korral on uuritav tunnus Y fikseeritud. Grupid on moodustatud uuritava tunnuse järgi ja sagedused $\{n_{+j}\}$ on katse läbiviijate poolt fikseeritud. Nüüd on seletav tunnus X juhuslik, ja

$$\{n_{ij},\ \text{veerus } j\} \quad \sim \quad M(n_{+j}, \{\pi_{i|j}\}),$$

kus $\{\pi_{i|j}\}$ on j-nda veeru tinglikud tõenäosused. Hinnatavad on tinglikud tõenäosused veergudes.

N. Juhtkontrolluuringus võeti südameinfarkti haigete kõrvale kontrollgrupp mitte-haigeid ja küsiti, kas nad on olnud kunagi suitsetajad (vt Agresti (1996, lk.24)).

Tabel 11: Tunnuste X="suitsetaja" ja Y="südame infarkt" sagedustabel

$X\backslash Y$	1 (infarkt)	2 (ei)
1 (suitsetaja)	172	173
2 (ei)	90	346

On hinnatavad tinglikud tõenäosused veergudes, aga mitte teised tõenäosused.

Ülesanne 2.4. Leia hinnangud tõenäosustele $P(X = \text{suitsetaja}|\text{infarkt})$, $P(X = \text{suitsetaja}|\text{ei infarkt})$.

Kui multinomiaaljaotus kirjeldab sageduste vektorit tabeli ridades (fikseeritud $\{n_{i+}\}$), siis read on üksteisest sõltumatud ja kogu tabeli tõenäosusjaotus saadakse reatõenäosusfunktsioonide korrutamise teel.

Ülesanne 2.5. Pane kirja $p(n_{11}, n_{12}, \ldots, n_{IJ})$ sõltumatute multinomiaaljaotusega ridade korral.

Poissoni jaotus sagedustabelile. Kui koguvalimimaht n ei ole fikseeritud, vaid on juhuslik, siis on tabeli sagedused n_{ij} sõltumatud ja kogu tabeli tõenäosusjaotus on üksikute tõenäosusjaotuste korrutis. Sageli on siin jaotusmudeliks Poissoni jaotus $n_{ij} \sim Po(\mu_{ij})$. Selline olukord esineb juhuslikult toimuvate sündmuste korral (vt Tabel 12). Kõik mõõdetavad tunnused juhuslikult toimuva sündmuse korral on ka juhuslikud (ristlõikeline uuring). Poissoni parameetri μ_{ij} nihketa hinnanguks on n_{ij}.
Sageli on ka multinomiaalsete tõenäosuste arvutamisel lihtsam kasutada lähendina Poissoni jaotust. Seos parameetrite vahel antakse ooteväärtuse kaudu $\mu_{ij} = n\pi_{ij}$ (oodatav sagedus tabeli ruudus (ij))

Tabel 12: Autoavariid Floridas 1988. aastal (Agresti 1996, lk 47)
Tunnused X="turvavöö" ja Y="tagajärg".

$X \backslash Y$	1 (fataalne)	2 (ei)	summa
1 (ei)	1 601	162 527	164 128
2 (jah)	510	412 368	412 878
summa	2 111	574 895	577 006

Ülesanne 2.6. Olgu üldkogumijatuseks soo-värvipimeduse jaotus tabelis 7. Kasutame Poissoni mudelit lähendina sagedustabelile, st sagedused n_{ij} on sõltumatud ja $n = 20$. Pane kirja vastav sageduste ühisjaotus. Leia tõenäosus, et $n_{11} = 2$, $n_{12} = 8$, $n_{21} = 1$, $n_{22} = 9$. Võrdle varemarvutatud multinomiaalse tõenäosusega.

3 Ühisjaotuse karakteristikud mõõtmaks sõltuvust

Oluline küsimus kahe tunnuse korral on sõltuvuse küsimus. Et mõista sõltuvust, tuleb kõigepealt defineerida sõltumatus. Teame, et sõltumatuse korral kehtib:

$$\pi_{ij} = \pi_{i+}\pi_{+j}, \forall i, j, \tag{4}$$

kus $\pi_{ij} = P(X = i, Y = j)$, $\pi_{i+} = P(X = i)$, $\pi_{+j} = P(Y = j)$. Uurime rea tinglikke jaotusi sõltumatuse korral. Vaatame rida i:

$$\pi_{j|i} = \frac{\pi_{ij}}{\pi_{i+}} = \frac{\pi_{i+}\pi_{+j}}{\pi_{i+}} = \pi_{+j}, \forall j, \tag{5}$$

kus $\pi_{j|i} = P(Y = j|X = i)$. Kuna rea indeks i oli suvaline, saime et sõltumatuse korral on tinglikud jaotused kõigis ridades omavahel võrdsed ja nimelt langevad kokku tunnuse Y marginaaljaotusega.

Ülesanne 3.1. Näita, et Y ja X sõltumatuse korral veeru tinglikud jaotused langevad kokku tunnuse X marginaaljaotusega.

Kui tinglikud jaotused eri ridades (või veergudes) on erinevad, siis on tegu sõltuvate tunnustega Y ja X.

Enamasti tahame lisaks sõltuvuse tuvastamisele mõõta ka selle tugevust. Tähtsad sõltuvuse tugevuse näitajad on defineeritud 2×2 jaotustabelite korral.

3.1 Sõltuvuse tugevuse näitajad 2×2 jaotustabeli korral

Olgu vaatluse all kaks tunnust X ja Y, mõlemad kahe tasemega. Vastav ühisjaotus kujutab endast 2×2 jaotustabelit.

Olgu Y uuritav tunnus ja X seletav. Olgu tegu ristlõikelise või prospektiivse uuringuga, nii, et Y on juhuslik. Nimetame väärtust $Y = 1$ eduks ja $Y = 2$ ebaeduks. Meid huvitavad edu tõenäosused X eri tasemetel, $\pi_{1|1}$ ja $\pi_{1|2}$. Kui need on erinevad, siis Y sõltub tunnusest X. Nende baasil on defineeritud rida sõltuvuse tugevuse näitajaid. Sõltuvuse tugevuse

Tabel 13: X ja Y ühisjaotus ja rea tinglikud jaotused

$X\backslash Y$	1	2	summa
1	π_{11}	π_{12}	π_{1+}
	$\pi_{1\|1}$	$\pi_{2\|1}$	1
2	π_{21}	π_{22}	π_{2+}
	$\pi_{1\|2}$	$\pi_{2\|2}$	1
summa	π_{+1}	π_{+1}	1

näitajad defineerime teoreetiliste tõenäosuste baasil. Valimist saadakse nende näitajate hinnangud.

A. Riskide vahe (tinglike tõenäosuste vahe, osakaalude vahe).

Meid huvitava sündmuse tõenäosust nimetatakse sageli riskiks.

$$P = \pi_{1|1} - \pi_{1|2},\ P \in [-1, 1],\ P = 0 \text{ on sõltumatus.}$$

Näitab mitme võrra on edu tõenäosus $X = 1$ korral suurem $X = 2$ omast. Mida suurem erinevus, seda suurem sõltuvus. Näitena vaatame südame infarkti ja aspiriini vahelist seost (tabel 10). Leiame tinglikud jaotused:

Tabel 14: Y=”südame infarkt” tinglikud jaotused

$X\backslash Y$	1 (jah)	2 (ei)	summa
1 (platseebo)	0.017	0.983	1.000
2 (aspiriin)	0.009	0.991	1.000

(Täida lüngad) Tabelist näeme, et aspiriini mittetarvitamine suurendab südame infarkti riski hinnanguliselt võrra. Ilusamgi on öelda, et aspiriini tarvitamine vähendab südame infarkti riski võrra.

Absoluutne erinevus on väike, sest mõlemad tõenäosused on väikesed. Sellisel korral on seose tugevuse mõõtmiseks õigem vaadata suhtelist erinevust.

B. Suhteline risk (tinglike tõenäosuste või osakaalude suhe)

$$R = \frac{\pi_{1|1}}{\pi_{1|2}},\ R \in [0, \infty),\ R = 1 \text{ sõltumatus}\ .$$

Näitab, mitu korda on edu tõenäosus $X = 1$ korral suurem $X = 2$ omast. Parema sõnastuse saamiseks $R < 1$ korral on hea kasutada seost, et suurendamine R korda tähendab vähendamist $1/R$ korda. Näiteks $R = 0.2$ korral on parem öelda, et edu tõenäosus, võrreldes tasemega $X = 2$ väheneb 5 (1/0.2) korda. Tabelist 14 näeme, et aspiriini mittetarvitamine suurendab südameinfarkti riski hinnaguliselt korda. Aspiriini tarvitamine aga

vähendab korda.

C. Šansid (i.k odds).

Edu šansid $X = 1$ korral on:

$$\Omega_1 = \frac{\pi_{1|1}}{\pi_{2|1}},\ \Omega_1 \in [0, \infty).$$

Pangem tähele, et edu tõenäosus $\pi_{1|1}$ ja edu šansid Ω_1 on eri näitajad.

Edu šansid $X = 2$ korral on:

$$\Omega_2 = \frac{\pi_{1|2}}{\pi_{2|2}},\ \Omega_2 \in [0, \infty).$$

Sõltumatuse korral $\Omega_1 = \Omega_2$, sest lugejad ja nimetajad on võrdsed.

Meie näite korral on südameinfarkti šansid aspiriini mittetarvitamisel hinnanguliselt ja aspiriini tarvitamisel . Šansid on erinevad, mis viitab infarkti sõltuvusele aspiriinist. Sõltuvuse tugevus mõõdab šansside suhe.

D. Šansside suhe.

On väga tähtis sõltuvuse tugevuse näitaja, sest teda saab arvutada igat tüüpi uuringute korral. Valem on

$$\theta = \frac{\Omega_1}{\Omega_2},\ \theta \in [0, \infty),\ \theta = 1 \text{ sõltumatus},$$

ehk

$$\theta = \frac{\pi_{1|1}\ \pi_{2|2}}{\pi_{2|1}\ \pi_{1|2}}.$$

Näitab, mitu korda on edu šansid $X = 1$ korral suuremad kui $X = 2$ korral. Parema sõnastuse otsimisel, eriti kui $\theta < 1$ võime moodustada suhte vastupidi ja öelda, mitu korda on edu šansid $X = 2$ korral suuremad kui $X = 1$ korral. Meie näites Tabelist (14) saame hinnanguks

$$\hat{\theta} = \frac{0.017 \cdot 0.991}{0.983 \cdot 0.009} = 1.90,$$

st aspiriini mittetarvitamine suurendab südameinfarkti šansse hinnanguliselt 1.9 korda ehk tarvitamine vähendab 1.9 korda.

Arvutustes kasutasime prospektiivse uuringu sagedustabelit, kus uuritav tunnus oli juhuslik ja seega kõik vajalikud tinglikud tõenäosused hinnatavad osakaaludega. Kui Y on mittejuhuslik (retrospektiivne uuring), siis neid tinglikke tõenäosusi osakaaludega hinnata ei saa. Näitame, et sellele vaatamata saab suurust θ ikka hinnata. Kõigepealt veendume, et

$$\theta = \frac{\pi_{1|1}\ \pi_{2|2}}{\pi_{2|1}\ \pi_{1|2}} = \frac{\pi_{11}\ \pi_{22}}{\pi_{12}\ \pi_{21}}.$$

Kui nüüd X on juhuslik ja Y fikseeritud, siis saame hinnata tinglikke tõenäosusi veergudes $P(X = 1|Y = 1),\ P(X = 2|Y = 1),\ P(X = 1|Y = 2),\ P(X = 2|Y = 2)$. Nende tinglike tõenäosuste abil saame kirja panna šansid $X = 1$ jaoks esimeses veerus ja teises veerus

ning moodustada šansside suhte näitamaks, mitu korda on $X = 1$ šansid esimeses veerus suuremad kui teises. Paneme kirja ja teisendame

$$\frac{P(X=1|Y=1)/P(X=2|Y=1)}{P(X=1|Y=2)/P(X=2|Y=2)} = \frac{\pi_{11}\ \pi_{22}}{\pi_{12}\ \pi_{21}} = \theta.$$

Saime sama avaldise nagu juhusliku Y-tunnuse jaoks. Seega šansside suhe ei sõltu sellest, kumb tunnus on juhuslik või kas mõlemad on juhulikud. Sobib sõltuvuse mõõduks igat tüüpi uuringute korral. Võime väljenduda uuritava tunnuse keskselt, isegi kui see ei ole juhuslik.

D. Šansside suhte hinnang.

Nägime, et Šansside suhte θ väärtus ei muutu, kui lähtume rea tinglikest tõenäosustest või veeru tinglikest tõenäosustest. Kui rea tinglikud tõenäosused on osakaaludega hinnatavad, saame

$$\hat{\theta} = \frac{\hat{\pi}_{1|1}/\hat{\pi}_{2|1}}{\hat{\pi}_{1|2}/\hat{\pi}_{2|2}} = \ldots = \frac{n_{11} \cdot n_{22}}{n_{12} \cdot n_{21}}.$$

Kui veeru tinglikud tõenäosused on osakaaludega hinnatavad, saame sama avaldise.

Veendu viimases.

Järeldus. Šansside suhte hinnanguks on sõltumata uuringu tüübist sagedustabeli ristkorrutis. Või, kui on antud suhtelised sagedused, siis nende ristkorrutis. Tegemist on mõjusate hinnangutega, sest tinglike tõenäosuste hinnagud vastavate osakaaludega on mõjusad (lähenevad teoreetilistele tinglikele tõenäosustele).

E. šansside suhte logaritm.

$$\ln\theta \in (-\infty, +\infty),\ \ln\theta = 0 \text{ sõltumatus.}$$

Paneme tähele, et $\theta \in [0, \infty)$ on ebasümmeetriline sõltumatuse olukorra suhtes: θ ja $1/\theta$ näitavad sama tugevat sõltuvust eri suundades, kuid kaugus väärtusest 1 ei ole sama. Logaritm $\ln\theta$ on sellest puudusest vaba: $-\ln\theta$ ja $\ln\theta$ on 0-punktist sama kaugel ja näitavad sama tugevat sõltuvust eri suundades.

F. Seos Šansside suhte ja suhtelise riski vahel.

Käsikirjaline materjal

G. Šansside suhe $I \times J$ tabeli korral.

Suurema tabeli korral defineeritakse rohkem kui üks šansside suhet. Vaadeldakse kas järjestikuseid ridade veergude paare või fikseeritakse baasrida, -veerg ja defineeritakse šansside suhted nende suhtes. Samuti saab teha teiste seosetugevuse nätajatega (suhtelise riski, riskide vahe ja teistega).

Ülesanne 3.2. Vaadakem andmestikku tabelis 12. Leida ja sõnastada kõik sõltuvuse näitajad, kui huvi all on fataalse lõpuga autoõnnetus sõltuvalt turvavöö kasutamisest.

3.2 Tinglike tõenäosuste statistiline võrdlemine

Kui tinglikud tõenäosused $\pi_{1|1}$ ja $\pi_{1|2}$ on erinevad, siis tunnus Y sõltub tunnusest X. Sagedustabelist saame hinnangud neile tõenäosustele. Tekib küsimus, kas hinnangute erinevus võib ikka tähendada $\pi_{1|1}$ ja $\pi_{1|2}$ erinevust, ja kui erinevus on, kumb on suurem. Siin aitavad otsusele jõuda usalduspiirid.

Vaatame riskide vahet $\pi = \pi_{1|1} - \pi_{1|2}$, ja selle hinnangut

$$\hat{\pi} = \hat{\pi}_{1|1} - \hat{\pi}_{1|2} = n_{11}/n_{1+} - n_{21}/n_{2+}.$$

Asümptootilised ehk suure valimi usaldusvahemik avaldub kujul

$$UI(\pi) = \hat{\pi} \pm \lambda_{\alpha/2}\sqrt{\hat{D}\hat{\pi}},$$

kus $\lambda_{\alpha/2}$ on $N(0,1)$ täiendkvantiil. Dispersiooni on lihtne leida, kui n_{1+} ja n_{2+} on fikseeritud konstandid (prospektiivne uuring). Siis

$$n_{11} \sim Bin(n_{1+}, \pi_{1|1}),\ n_{21} \sim Bin(n_{2+}, \pi_{1|2}) \text{ ja sõltumatud}$$

ja

$$D(\hat{\pi}) = \frac{\pi_{1|1}(1-\pi_{1|1})}{n_{1+}} + \frac{\pi_{1|2}(1-\pi_{1|2})}{n_{2+}}.$$

Ülesanne 3.3. Leia dispersiooni hinnang, avalda usaldusvahemik ja leia see südameinfarkti korral (Tabel 14). Interpreteeri.

Saab kasutada ka juba tuttavat statistilist testi (Matemaatiline statistika 1 kursusest) binoomjaotuse parameetrite vahe hindamiseks. Olgu

$$H_0 :\ \pi_{1|1} - \pi_{1|2} = 0.$$

Eeldame nüüd, et H_0 kehtib, ja moodustame tsentreeritud ja normeeritud statistiku, mis on ligikaudu normaaljaotusega

$$T = \frac{(\hat{\pi}_{1|1} - \hat{\pi}_{1|2}) - 0}{\sqrt{\hat{D}(\pi)}} \sim N(0,1).$$

Kasutades normaaljaotust leitakse väärtuse T olulisustõenäosus. Testimisel saab kasutada ka teadmist, et

$$T^2 \sim \chi^2(1).$$

Kui n_{1+} ja n_{2+} ei ole fikseeritud, saab kasutada hii-ruut testi sõltumatuse tuvastamiseks, mida vaatame hiljem.

3.3 Tinglikud tõenäosused klassifitseerimisel

Sageli on tegu klassifitseerimisülesandega, kus on vaja otsustada, kas patsiendil on oletatav haigus, kui vereproovis teatud näitaja on positiivne, või kas sportlane on kasutanud dopingut, kui vastav test on näitab seda. Sellises ülesandes on tinglikel tõenäosustel omad tähendused.

Olgu X haiguse olemasolu (jah/ei) ja Y diagnostiline test (positiivne/negatiivne)

Tabel 15: Tinglikud tõenäosused klassifitseerimisel

$X\backslash Y$	positiivne	negatiivne	summa
jah	$\pi_{1\|1}$	$\pi_{2\|1}$	1
ei	$\pi_{1\|2}$	$\pi_{2\|2}$	1

Tundlikkus (sensitivity) $-$ $\pi_{1|1}$ tõenäosus, et haigestunu test on positiivne.

Spetsiifilisus/eristavus (specificity) $-$ $\pi_{2|2}$ tõenäosus, et terve inimese test on negatiivne.

Valepositiivsus $-$ $\pi_{1|2}$ $-$ tõenäosus, et terve inimese test on positiivne.

Valenegatiivsus $-$ $\pi_{2|1}$ tõenäosus, et haigestunu test on negatiivne.

Need tõenäosused on omavahel seotud, kuna reasummad on 1. Hea test on selline, kus tundlikkus ja eristavus on suured (1 lähedased), siis on valepositiivsus ja valenegatiivsus väikesed.

Üleüldine õigesti klassifitseerimise tõenäosus saadakse järgmiselt:

$$\text{P(õige klassifitseerimine)} = \pi_{11} + \pi_{22} = \pi_{1|1}\pi_{1+} + \pi_{2|2}\pi_{2+}.$$

Vaatame mammograafilise testi näitajaid.

Tabel 16: Hinnangulised tinglikud jaotused rinnavähi diagnoosimisel
Agresti (2002, lk.38)

Rinnavähk\ Test	positiivne	negatiivne	summa
jah	0.82	0.18	1.0
ei	0.01	0.99	1.0

Näeme, et mammograafiline uuring tuvastab rinnavähi 82% rinnavähiga naistel, ja 99% naistest, kellel pole rinnavähki diagnoositakse õigesti.

Ülesanne 3.4. Sõnasta sisukeskselt valenegatiivsused ja valepositiivsused järgmiste tuntumate testide korral tabelis 17 (K. Fischer, Postimees 06.04.2013, algallikas Wikipedia, PubMedi teadusartiklite andmebaas)

Tabel 17: Mõningate testide omadused

Test\ Omadused	Tundlikkus	Spetsiifilisus	Valenegatiivsus	Valepositiivsus
Apteegis müüdav rasedustest	65-97%	$\approx 100\%$		
Apteegis müüdav tsöliaakia test	90%	95%		
HIV test	99.7%	99.99%		
Eesnäärmevähi PSA test	20%	94%		

Ülesanne 3.5. Kirjuta tabelisse 18 10000 sportlase hüpoteetilise dopingutesti tulemused, kui dopingutarvitajaid on 2%, testi tundlikkus on 25% ja spetsiifilisus on 99.9%. Arvuta saadud tabeli pealt veerutinglike tõenäosuste hinnangud: 1) kui suur osa positiivse testitulemuse saanud sprtlastest on süütud, 2) kui suur osa negatiivse testitulemuse saanud sportlastest on dopingut kasutanud.

Tabel 18: Sportlaste jagunemine hüpoteetilises olukorras

Dopingukasutaja\ Testitulemus	positiivne	negatiivne	kokku
jah			
ei			

3.4 Seos kolmandate tunnustega

Oleme varasemalt märkinud, et kui ei suuda katse keskkonda kontrollida, siis ei pruugi tulemused kokku langeda meie ootustega. Näiteks kindlustuskompaniile laekuvate tulekahjude nõuete arv kuus ei pruugi olla Poissoni jaotusega, kus tulekahjude arvu ooteväärtus langeb kokku dispersiooniga. Muutuvad ja mitte arvesse võetud ilmastikuolud võivad tekitada üledispersiooni.

Ka kahe tunnuse X ja Y sõltuvuse uurimisel tuleb mõelda kolmandatele tunnustele, mis arvesse võetuna võiksid sõltuvuse iseloomu muuta. Eriti tuleb olla ettevaatlik **põhjuslike** järelduste tegemisel.

N. Nähes oma sagedustabelist, et kopsuvähk esineb sagedamini suitsetajate hulgas, ei pruugi suitsetamine veel olla põhjuseks kopsuvähi tekkele. Võib leiduda kolmas tunnus, mis põhjustab nii suitsetamist kui kopsuvähki. Kuna viimast pole analüüsis, näeme vale põhjuslikku seost (statistiline seos on olemas).

Näiteks olgu kolmas tunnus depressioon, mis mõjutab nii suitsetamist kui ka kopsuvähki.

Peaksime depressiooni fikseerima, st uurima suitsetamise ja kopsuvähi vahelist seost eraldi

depressiooniga isikute ja tervete hulgas. Kui mõlemal depressiooni tasemel näeme seost suitsetamise ja kopsuvähi vahel, siis sõltumata depressioonist põhjustab suitsetamine kopsuvähki (kui pole kolmandaid, neljandaid jne segavaid tunnuseid). Sõnaga "põhjustab" tuleb alati ettevaatlik olla. Ohutum on öelda näiteks, et suitsetamine suurendab kopsuvähi riski.

Üldiselt uuringutes, kus tahetakse põhjuslikke seoseid uurida 2 tunnuse vahel, tuleks fikseerida kõik tunnused, mis võiksid neid kahte mõjutada ja teha sõltuvusanalüüs läbi kõigil fikseeritud tunnuste tasemete kombinatsioonidel.

N. Surmanuhtlus kui karistus sõltuvalt süüaluse rassist (Agresti 2002, lk. 48)

Olgu Y – surmanuhtlus (jah/ei);
X – süüaluse rass (valge/must);
Z – ohvri rass(valge/must).

Kas mustad või valged süüalused saavad sagedamini surmanuhtluse? Mida näeme siis, kui ohvri rass on fikseeritud.

Tabel 19: surmanuhtlus süüaluse ja ohvri rassi kaupa

Ohvri rass	Süüaluse rass	Surmanuhtlus jah	Surmanuhtlus ei	protsent "jah"
valge	valge	53	414	11.3
	must	11	37	22.9
must	valge	0	16	0.0
	must	4	139	2.8
kokku	valge	53	430	11.0
	must	15	176	7.9

Tabelis 19 on kolme tunnuse X, Y, Z sagedustabel. Igal Z tasemel näeme tunnuste X, Y 2×2 sagedustabelit ja kahes viimases reas pealkirjaga "kokku" näeme tunnuste X, Y marginaalset sagedustabelit, kus sagedused on summeeritud üle tunnuse Z (veendu!). Tabeli viimases veerus on surmanuhtluse osakaal vaadeldava rea kõigi kohtuotsuste hulgas (protsentides).

Vaadeldes marginaalset sagedustabelit, kus tunnust Z esindatud ei ole, näeme, et valge süüalune on saanud sagedamini surmanuhtluse kui must süüalune. Samas võttes mängu kolmanda tunnuse Z=ohvri rass, näeme, et nii valgete ohvrite kui ka mustade ohvrite hulgas on mustad süüalused saanud sagedamini surmanuhtluse. Z tunnuse arvesse võtmisel on eelmine tulemus pea peale pööratud.

Osutub, et siin on tunnus Z=ohvri rass surmanuhtlusele nn "põhjustav tunnus". Ta on tugevalt seotud nii X- kui Y-tunnusega. X ja Y vaheline seos, kus Z ei ole analüüsis, on küll üks statistiline seos, aga mitte põhjuslik.

Ülesanne 3.6. Veendugem, et ohvri rass on tugevalt seotud nii süüaluse rassiga, kui ka

surmanuhtluse tunnusega. Kirjuta suurest tabelist välja vastavad 2×2 sagedustabelid ja leia seosetugevuse mõõduna shansside suhted.

Fikseerides ohvri rassi, elimineerime me selle tunnuse mõju ja kui rohkem segavaid/seotud tunnuseid ei ole, näeme nüüd X põhjustatud mõju Yle.

Ülesanne 3.7. Leia suhtelised riskid ($\hat{R}$), mis võrdlevad surmanuhtluse tõenäosust süüaluse rassi kaupa fikseeritud ohvri rassi korral. Sõnasta.

Sellist nähtust, kus tinglikud seosed X, Y vahel fikseeritud Z väärtuste korral on vastupidised marginaalsetele seostele nimetatakse Simpsoni paradoksiks. Sellised näited tuuakse ohunäitena, kui uurijad hakkavad põhjuslikke seoseid väitma.

Kui marginaalses tabelis on väga tugev seos X, Y vahel, siis ta ei saa vastupidiseks muutuda tinglikes tabelites.

Ülesanne 3.8. Kontrolli suhtelise riski ($\hat{R}$) abil, et siin on marginaalses tabelis nõrk seos X (süüaluse rass), Y (surmanuhtlus) vahel.

3.5 Suuremad kui 2×2 jaotustabelid

Olgu nüüd X I-tasemega ja Y J-tasemega. Jaotustabel siasldab nüüd $I \times J$ tõenäosust π_{ij}. Sõltumatuse korral kehtib $\pi_{ij} = \pi_{i+}\pi_{+j},\ \forall i, j$. Samuti näitasime, et sel korral on rea tinglikud jaotused omavahel võrdsed ja võrdsed ka Y marginaaljaotusega. Analoogiline omadus kehtib veeru tinglike jaotuste kohta.

Kui aga X ja Y vahel on sõltuvus, siis kuidas selle tugevust mõõta ühe arvuga?

3.5.1 Seosekordajad nominaaltunnuste korral

Nominaaltunnuste korral puudub tasemete järjestus. Raske on kirjeldada seose tugevust ühe numbriga ja see pole ka alati kasulik. Mõtiskleme kõigepealt varieeruvuse mõiste üle. Vaatame Y jaotust:

Y	1	...	j	...	J
$P(Y=j)$	π_{+1}	...	π_{+j}	...	π_{+J}

Olukorrad:
$\pi_{+k} = 1,\ \pi_{+j} = 0,\ j \neq k$ – konstantne tunnus, ainult väärtus k on võimalik, varieeruvus on 0;

π_{+k} suur, teised tõenäosused väikesed – vähe varieeruv tunnus (valimis näeme valdavalt väärtust k);

$\pi_{+j} = 1/J,\ \forall j$ – maksimaalse varieeruvusega tunnus.

Arvuliselt saab nominaalse tunnuse varieeruvust mõõta mitmeti. Üks võimalus on entroopia mõiste kaudu.

Definitsioon. Diskreetse juhusliku suuruse entroopia on arvuline suurus:

$$H(Y) = -\sum_{j=1}^{J} P(Y=j) \ln P(Y=j).$$

Entroopial on järgmised omadused:
$H(Y) \geq 0$;
$H(Y) = 0$, kui $\exists k$ nii et $P(Y=k) = 1$;
$H(Y) = \ln J$ (maksimum), kui $P(Y=k) = \frac{1}{J},\ \forall k$.

Kasutades entroopia mõistet on tunnuse Y dispersioon

$$D(Y) = -\sum_{j=1}^{J} \pi_{+j} \ln \pi_{+j} \tag{6}$$

ja tunnuse Y tinglik dispersioon tunnuse X tasemel i

$$D(Y|X=i) = -\sum_{j=1}^{J} \pi_{j|i} \ln \pi_{j|i}. \tag{7}$$

Järgmine tabel illustreerib maksimaalse sõltuvuse juhtu tunnuste X ja Y vahel. Tabeli 20

Tabel 20: Tunnuse Y tinglikud jaotused maksimaalse sõltuvuse korral

$X \backslash Y$	1	...	j	...	J	summa
1	0	...	0	...	1	1
...	...	...	...	...	...	...
i	1	0	...	...	...	1
...	...	...	...	...	...	...
I	0	...	1	0	...	1
$P(Y=j)$	π_{+1}	...	π_{+j}	...	π_{+J}	1

ridades ja veergudes on ainult üks 1, st tinglikel jaotustel varieerivus puudub. Niipea, kui X tase on fikseeritud, on Y väärtus täpselt teada (tõenäosusega 1). See on maksimaalse sõltuvuse juht. Samas on marginaaljaotuselt tinglikele üleminekul toimunud dispersiooni vähenemine, antud juhul kõige ekstreemsemal viisil. Dispersiooni vähenemisele ongi rajatud sõltuvuse mõõtmine nominaaltunnuste korral.

Definitsioon. Nominaaltunnuste sõltuvusekordaja mõõdab suhtelist dispersiooni vähenemist tinglikustamisel, ja see on

$$U = \frac{D(Y) - E_X[D(Y|X)]}{D(Y)}. \tag{8}$$

Näeme, et $U = 0$ kui $D(Y|X) = D(Y)$ (dispersiooni vähenemist ei toimu, Y ei sõltu X-st); $U = 1$, kui $D(Y|X) = 0$ (maksimaalne sõltuvus). Väljendame U, kasutades meie dispersiooni definitsiooni. Leiame kõigepealt:

$$E_X[D(Y|X)] = \sum_{i=1}^{I} D(Y|X=i)P(X=i) = -\sum_{i} \sum_{j} \pi_{j|i} \ln \pi_{j|i} \cdot \pi_{i+} \tag{9}$$

Ülesanne 3.9. Näita kasutades (6)-(9), et seosekordaja U saab viia kujule:

$$U = -\frac{\sum_i \sum_j \pi_{ij} \ln \frac{\pi_{ij}}{\pi_{i+}\pi_{+j}}}{\sum_j \pi_{+j} \ln(\pi_{+j})}. \tag{10}$$

Suurust U nimetatakse ka "uncertainty coefficient". Teda saab hinnata ristlõikelise uuringu korral. SAS väljundis tähistatakse seda $U(C|R)$ – veerutunnuse (C) varieeruvuse osa, mis on kirjeldatud reatunnuse (R) poolt. Saab leida ka $U(R|C)$.

Märkus. Nominaaltunnuse varieeruvust saab mõõta ka Gini kordajaga: $G = \sum_{j=1}^{J} \pi_j(1 - \pi_j)$. Millised on selle väärtused suurima ja vähima varieeruvuse korral?

3.5.2 Seosekordajad järjestustunnuste korral

Järjestustunnuse korral ei ole kaugused tasemete vahel määratud, seega vahet ei saa leida ja seega arvuliste tunnuste korrelatsioonikordaja ei ole siin rakendatav. Saab vaadelda **monotoonse trendi** olemasolu: kas X taseme kasvamisele vastab valdavalt ka Y taseme kasvamine või hoopis kahanemine.

Kordaja defineerimiseks tuuakse sisse samasuunalisuse ja vastasuunalisuse mõiste. Vaatame kahte objekti a ja b ja tunnsvektori (X, Y) mõõtmistulemusi neil. Tähistame mõõtmistulemusi (X_a, Y_a) ja (X_b, Y_b).

Definitsioon. Objekte a ja b nimetatakse:

samasuunalisteks, kui $X_b > X_a \Rightarrow Y_b > Y_a$ või kui $X_b < X_a \Rightarrow Y_b < Y_a$;

vastassuunalisteks, kui $X_b > X_a \Rightarrow Y_b < Y_a$ või kui $X_b < X_a \Rightarrow Y_b > Y_a$;

seotud objektideks, kui $X_a = X_b$ ja/või $Y_a = Y_b$.

Tuletame X ja Y vahelise seose tugevuse mõõtmiseks kordaja teoreetilise avaldise. See kasutab üldkogumijaotust $\pi_{ij} = P(X = i, Y = j)$. Hiljem leiame kordaja hinnangu valemi ja rakendame seda tabelile 16 (Agresti, lk 57).

Tabel 21: Sissetulek ja tööga rahulolu

	Tööga rahulolu			
Sissetulek (dollarites)	Äärmiselt rahulolematu	Pigem rahulolematu	Pigem rahul	Väga rahul
< 15000	1	3	10	6
$15000 - 25000$	2	3	10	7
$25000 - 40000$	1	6	14	12
> 40000	0	1	9	11

Käsikirjaline materjal.

Ülesanne 3.10. Leia seosetugevuse mõõdud $U(C|R)$ ja Goodman-Kruskali γ sotsiaaluuringu sagedustabelis Tabel 21. Interpreteeri.

Tabel 22: Suhtumine abielueelsesse ja homoseksuaalsesse seksi (Agresti, lk 65)

	Homoseksuaalne seks			
Abielueelne seks	Alati vale	Peaaegu alati vale	Vahel vale	Üldse mitte vale
Alati vale	300	4	4	17
Peaaegu alati vale	78	15	3	14
Vahel vale	107	16	46	54
Üldse mitte vale	234	32	35	336

Ülesanne 3.11. Tee tabelist 22 2×2 tabel tasemetega "ei ole lubatud", "on lubatud". Leia šansside suhe ja Yule'i Q. Interpreteeri.

4 Hüpoteesid ja testid sagedustabeli korral

Oleme vaadanud, kuidas sagedustabelilt hinnata kvalitatiivsete tunnuste üldkogumijaotuse tõenäosusi või tinglikke tõenäosusi:

$$\hat{\pi}_{ij} = \frac{n_{ij}}{n},\ \hat{\pi}_{j|i} = \frac{n_{ij}}{n_{i+}},\ \hat{\pi}_{i+} = \frac{n_{i+}}{n},\ \hat{\pi}_{+j} = \frac{n_{+j}}{n}.$$

Ristlõikelise uuringu korral on kõik toodud hinnangud nihketa. Prospektiivse või retrospektiivse uuringu korral saab sagedustabelilt hinnata vaid juhusliku tunnuse tinglikke ja marginaalseid tõenäosusi, seda samuti ülaltoodud valemitega, saadakse nihketa hinnangud. Nägime, et eeldades sageduste ühisjaotuse kuulumist teatavasse perre (multinomiaaljaotus) osutusid antud hinnangud suurima tõepära hinnanguteks.

4.1 Hüpoteesid üldkogumi jaotuse sobivuse kohta (goodness-of-fit tests)

Kui on tegemist kahe tunnus X ja Y n sõltumatu vaatlusega samast üldkogumijaotusest $\{\pi_{ij}\}$, siis $n_{ij} \sim B(n, \pi_{ij})$. Kui sõltumatud vaatlused on rea tinglikust jaotusest, siis $n_{ij} \sim B(n_{i+}, \pi_{j|i})$. Selles paragrahvis püstitame hüpoteesid tõenäosuste π_{ij} või $\pi_{j|i}$ väärtuste kohta ja vaatame, kas andmed kinnitavad meie hüpoteesi sobivust.

Esituse lihtsustamiseks läheme üle ühele indeksile, st vaatame N tasemega kvalitatiivset tunnust Y, kus on tehtud n sõltumatut vaatlust ja saadud sagedused $n_i = \#\{Y = i\}$, $\sum_{i=1}^{N} n_i = n$. Tõenäosused $\pi_i = P(Y = i)$ on tundmatud. Teame, et

$$n_i \sim B(n, \pi_i).$$

Püstitame nullhüpoteesi:

$$H_0 :\ \pi_i = \pi_i^0,\ i = 1, 2, \ldots, N,$$

kus $\sum_{i=1}^{N} \pi_i = \sum_{i=1}^{N} \pi_i^0 = 1$. Karl Pearson (1900) vaatles esmakordselt antud hüpoteesi kontrollimist just kvalitatiivsete andmete korral, nimelt ruletiratta näitel, soovides testida, kas võimalikud tulemused on võrdtõenäosed.

Hüpoteesi H_0 eeldusel on üldkogumi jaotuse tõenäosusteks $P(Y = i) = \pi_i^0$ ja seega

$$n_i \sim B(n, \pi_i^0),$$

millest oodatavad sagedused valimis oleksid

$$m_i = E(n_i) = n\pi_i^0.$$

Pearsoni hii-ruut statistik võrdleb realiseerunud sagedusi ja H_0 õigsuse korral oodatavaid sagedusi:

$$H = \sum_{i=1}^{N} \frac{(n_i - m_i)^2}{m_i}.$$

Pearson tõestas, et suure n ja H_0 õigsuse korral on juhuslik suurus H hii-ruut jaotusega vabadusastmete arvuga $N - 1$:

$$H \sim \chi^2(N - 1).$$

Paneme tähele, et vabadusastmete arv on 1 võrra väiksem kui liidetavate arv summas. See tuleneb ühe kitsenduse olemasolust juhuslikele suurustele n_i H avaldises, $\sum_{i=1}^{N} n_i = n$. Seega vabalt saavad muutuda vaid $N - 1$ tükki n_i-dest. Siit ka nimi "vabadusastmete arv" hii-ruudu parameetrile (Fisher 1922).

Kui realiseerunud sagedused erinevad palju oodatavatest H_0 õigsuse korral, siis on H väärtus suur, mis vihjab H_0 sobimatusele. Lõplikuks otsustamiseks leiame olulisustõenäosuse. Olgu statistiku H väärtus antud valimi korral h. Olulisustõenäosus

$$p = P(H \geq h)$$

leitakse, kasutades teadmist, et $H \sim \chi^2(N - 1)$. Kui p on väike ($p < 0.05$), siis järelikult on h lubamatult suur statistiku H väärtus nullhüpoteesi õigsuse korral. Sel korral H_0 kummutatakse, andmed ei kinnita hüpoteesi üldkogumijaotuse kohta. Väärtus p on eksimistõenäosus H_0 kummutamisel, sest selle väikese tõenäosusega võib ka H_0 õigsuse korral ette tulla valim, mis annab H väärtuseks h või temast suurema väärtuse.

Ülesanne 4.1. Tunne hii-ruutu. Olgu $H \sim \chi^2(\nu)$. Kuidas saad aru, et H väärtus on suur või väike, ilma olulisustõenäosust p arvutamata.

Klassikaline näide Pearsoni testi rakendamisest on Mendeli teooria testimine.

Näide. Mendel. Mendel ristas herneid, puhast rohelist liiki puhta kollase liigiga ja vaatles tunnust "värv" teises põlvkonnas. Ta püstitas hüpoteesi värvi tõenäosuste jaoks:

$$H_0 : \ \pi_{roh} = 0.25, \ \pi_{kol} = 0.75.$$

Tema esimese eksperimendi saagis oli $n = 8023$ hernest, kus

$$n_{roh} = 2001, \ n_{kol} = 6022.$$

Oodatavad esinemissagedused H_0 õigsuse korral oleksid

$$m_{roh} = 8023 \cdot 0.25 = 2005.75,\ m_{kol} = 8023 \cdot 0.75 = 6017.25.$$

Arvutades leiame statistiku H väärtuse $h = 0.015$. Siin $H \sim \chi^2(1)$. Kasutades hii-ruut jaotuse tabeleid või sobivat tarkvara (näiteks R), leiame

$$p = P(H \geq 0.015) = 0.88.$$

Sellest järeldame, et erinevus oodatavate ja realiseerunud sageduste vahel on mitteoluline ja jääme H_0 juurde.

Näide jätkub looga, et Mendel tegi seda liiki katseid palju. Fisher uuris kõiki neid eksperimente 1936. aastal. Ta testis hüpoteesi H_0 üle kõigi eksperimentide kasutades hii-ruudu aditiivsuse omadust.

Aditiivsus: olgu $H_i \sim \chi^2(N_i)$ ja sõltumatud, $i = 1, 2, \ldots, k$. Siis $\sum_{i=1}^k H_i \sim \chi^2(\sum_{i=1}^k N_i)$.

Fisher leidis, et $k = 84$, $H_{sum} = \sum_{i=1}^k H_i$ andis väärtuse 42, $\sum_{i=1}^k N_i = 84$. Osutus, et H_{sum} väärtus 42 näitab liiga head kooskõla nullhüpoteesi ja katseandmete vahel. Kasutades $H_{sum} \sim \chi^2(84)$ leiti olulisustõenäosus

$$p = P(H_{sum} \geq 42) = 0.99996.$$

Teisisõnu, väärtus 42 oli üliharuldane H_0 õigsuse eeldusel. Väärtusi vahemikust 0 kuni 42 võiks saada üksnes neljal juhul 100 000-st: $P(H_{sum} < 42) = 0.00004$. Fisheri järeldus olukorra selgitamiseks oli, et aednik pettis Mendelit andmete ülestähendamisel. Püüdes Mendelile meele järgi olla, näitas ta katsetulemusi oodatavatele lähemal.

Märkus. Enamuses ülesannetes kasutatakse hii-ruut testi olulise erinevuse avastamiseks. Sellele viitab suur statistiku väärtus ja väike olulisustõenäosus. Leidub aga ka ülesandeid, kus liiga hea kooskõla nullhüpoteesiga seab samuti selle kehtivuse kahtluse alla. Niisugune olukord esineb näiteks juhuslike arvude generaatorite kontrollimisel. Genereeritud arvude liiga hea kooskõla jaotusega viitab sellele, et arvud ei käitu nii nagu juhuslikud arvud. Ülesannetes, kus nii liiga suur erinevus kui ka liiga hea kooskõla nullhüpoteesiga seavad selle kehtimise kahtluse alla, tehakse test olulisustõenäosusega järgmiselt: kui

$$\alpha < p < 1 - \alpha,$$

kus α on väike tõenäosus, siis jääme nullhüpoteesi juurde, vastasel korral kummutame selle. Eksimistõenäosus nullhüpoteesi kummutamisel on 2α.

Hüpoteesi H_0 fikseerimiseks peab uurijal olema teadmine tõenäosuste π_i^0 kohta. Mendelil olid arvud 0.25 ja 0.75. Kui teadmist pole, võib valimist infot saada. Samas ei tohi kõiki hüpoteesis H_0 antud parameetreid valimist hinnata, nagu näitab järgmine arutelu.

Kirjutame hii-ruut statistiku teisiti:

$$H = \sum_{i=1}^N \frac{(n_i - n\pi_i^0)^2}{n\pi_i^0} = \sum_{i=1}^N \frac{n(p_i - \pi_i^0)^2}{\pi_i^0},$$

kus $p_i = n_i/n$. Kui kõigi π_i^0 asemel kasutada valimhinnanguid p_i, siis $H = 0$ konstantselt ja pole üldse juhuslik suurus. Test on aga rajatud hii-ruut jaotusega juhuslikule suurusele. Testimise eesmärgiks on tuvastada või ümber lükata andmete ja nullhüpoteesis väidetud jaotuse sobivus. Siin aga valisime nullhüpoteesi jaoks täpselt andmetega sobiva jaotuse. Testimine kaotab mõtte.

Teatud juhtudel, mida kirjeldame allpool, saab valimist siiski abi otsida nullhüpoteesi püstitamisel ja hii-ruut statistiku koostamisel. Eriliselt tähelepanelik tuleb siis olla hii-ruut jaotuse vabadusastmete arvu määramisel.

Sõltugu meie üldkogumijaotuse tõenäosused väiksemast arvust parameetritest:

$$\begin{aligned} \pi_i &= f_i(\theta),\ i = 1, 2, \ldots, N \\ \theta &= (\theta_1, \theta_2, \ldots, \theta_t),\ t < N. \end{aligned}$$

Valimist leiame suurima tõepära hinnangu $\hat{\theta}$ vektorile θ. Invariantsuse printsiibist lähtuvalt on suurima tõepära hinnangu funktsioonid $\hat{\pi}_i = f_i(\hat{\theta})$ ka suurima tõepära hinnangud, antud juhul tõenäosustele π_i. Saadud hinnanguid kasutame nullhüpoteesis:

$$H_0 :\ \pi_i = \hat{\pi}_i = f_i(\hat{\theta}),\ i = 1, 2, \ldots N.$$

Vastav hii-ruut statistik saab kuju,

$$H = \sum_{i=1}^{N} \frac{n(p_i - \hat{\pi}_i)^2}{\hat{\pi}_i}.$$

Ka nüüd on H nullhüpoteesi õigsuse korral asümptootiliselt hii-ruut jaotusega, kuid väiksema vabadusastmete arvuga. Maha tuleb lahutada valimist hinnatud parameetrite arv t:

$$H \sim \chi^2(N - 1 - t).$$

Üldine reegel vabadusastmete määramiseks on: juhuslike liidetavate arv summas miinus kitsenduste arv neile, miinus valimist hinnatud tõenäosuste arv.

Näide. Vasikate nakatumine kopsupõletikku.

Käsikirjaline materjal

Tabel 23: Kopsupõletikku nakatumine

	Teisene nakatumine	
Esmane nakatumine	Jah	Ei
Jah	30 (38.1)	63 (39.0)
Ei	0(−)	63 (78.9)

4.2 Hüpoteesid sõltumatuse kohta

Uurime kahte kvalitatiivset tunnust X ja Y eesmärgiga teha kindlaks sõltuvus. Nullhüpotees väidab sõltumatust

$$H_0: \ \pi_{ij} = \pi_{i+}\pi_{+j}, \ i = 1, 2, \ldots, I, \ j = 1, 2, \ldots, J.$$

Kas valim, st sagedustabel $\{n_{ij}\}$ on kooskõlas nullhüpoteesiga? Testimiseks kasutame Pearsoni hii-ruut statistikut. Eeldame, et tegu on ristlõikelise uuringuga fikseeritud valimimahuga n, siis $n_{ij} \sim B(n, \pi_{ij})$. Eelkõige leiame oodatavad sagedused nullhüpoteesi eeldusel:

$$m_{ij} = E(n_{ij}) = n\pi_{ij} = n\pi_{i+}\pi_{+j}.$$

Kuna ka marginaalsed tõenäosused pole tavaliselt teada, hindame need valimist:

$$\hat{\pi}_{i+} = \frac{n_{i+}}{n}, \ \hat{\pi}_{+j} = \frac{n_{+j}}{n}.$$

Nüüd saame oodatavate sageduste hinnangud:

$$\hat{m}_{ij} = n\hat{\pi}_{i+}\hat{\pi}_{+j} = \frac{n_{i+}n_{+j}}{n}.$$

Hii-ruut test. Hii-ruut statistik, mis võrdleb realiseerunud sagedusi oodatavatega (nullhüpoteesi eeldusel) on järgmine:

$$H = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}}.$$

Statistiku H asümptootiline jaotus nullhüpoteesi eeldusel on

$$H \sim \chi^2(\nu),$$

kus vabadusastmete arv kujuneb järgmiselt:

$$\nu = IJ - 1 - (I - 1) - (J - 1) = (I - 1)(J - 1).$$

Siin IJ on juhuslike suuruste n_{ij} arv summas H, neil on üks kitsendus (summeeruvus n-ks). Valimist hinnatakse $I - 1$ tunnuse X marginaalset tõenäosust π_{i+}, üks π_{i+} on määratud teistega, sest summa on 1. Valimist hinnatakse ka $J - 1$ teise tunnuse marginaalset tõenäosust. Kui H_0 kehtib, st tunnused X ja Y on sõltumatud, siis on realiseerunud n_{ij} ja oodatavad $\hat{m}_{ij}$ lähedased ja statistiku H väärtus väike. Suur H viitab sõltuvusele. Test tehakse olulisustõenäosuse abil või jaotuse täiendkvantiili abil.

Meeldetuletuseks, jaotuse $H \sim \chi^2(\nu)$ α-täienkvantiiliks on väärtus q_α jaotuse määramispiirkonnas, mille korral $P(H > q_\alpha) = \alpha$. Testimisel võrreldakse statistiku H väärtust h täiendkvantiiliga. Kui $h > q_\alpha$, siis H_0 kummutatakse ja väidetakse, et tunnused X ja Y on sõltuvad. Eksimistõenäosuseks selle otsuse juures on α.

Kui marginaalsed tõenäosused ei ole valimist hinnatud vaid on püstitatud uurija poolt valimist sõltumatult, siis kasutatakse testimisel statistikut

$$H = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - m_{ij})^2}{m_{ij}} \sim \chi^2(\nu),$$

kus vabadusastmete arvuks on nüüd $\nu = IJ - 1$.

Näide. Religioosne enesemääratlus. Eesmärgiks on testida sõltumatust, kas religioosne enesemääratlus noorena ja hetkel on sõltuvad?

Tabel 24: Religioosne enesemääratlust hetkel ja 16 aastaselt

Enesemääratlus 16 aastaselt	Enesemääratlus hetkel				
	Protestant	Katoliiklane	Juut	Muu	Kokku
Protestant	918	27	1	70	1016
Katoliiklane	30	351	0	37	418
Juut	1	1	28	1	31
Muu	29	5	0	25	59
Kokku	978	384	29	133	1524

Ül. Leia tabelisse oodatavad sagedused sõltumatuse korral. Leia hii-ruut statistiku väärtus. Mis on vabadusastmed? Tee test. Kommenteeri.

Tõepärasuhte test on teine võimalus kontrollida X ja Y sõltumatust sagedustabeli baasil. Teststatistikuks on suhe, kus nimetajas maksimiseeritakse tõepärafunktsiooni üle kogu parameeterruumi, lugejas aga üle kitsendatud ruumi vastavalt nullhüpoteesis seatud tingimustele.

Meie valimiks on sagedused $\{n_{ij}\}$ ja nende ühisjaotuseks fikseeritud n korral eeldame multinomiaaljaotust (loogelistes sulgudes on kogu tabel):

$$\{n_{ij}\} \sim M(n; \{\pi_{ij}\}).$$

Tõepärafunktsiooniks on valimi saamise tõenäosus, mis vastavalt multinomiaaljaotuse valemile (3) on

$$L(\pi_{ij}) = c \cdot \prod_{i=1}^{I} \prod_{j=1}^{J} \pi_{ij}^{n_{ij}}. \tag{11}$$

Ühistõenäosuse π_{ij} suurima tõepära hinnanguks on $\hat{\pi}_{ij} = \frac{n_{ij}}{n}$. Need maksimiseerivad nimetaja. Asendades need valemisse (11), saame teststatistiku nimetaja.

Hüpoteesi H_0 eeldusel $\pi_{ij} = \pi_i \pi_j$, mille suurima tõepära hinnanguks on $\hat{\pi}_{ij} = \frac{n_{i+}n_{+j}}{n^2}$. Tehes asendused valemisse (11), saame teststatistiku lugeja. Seega tõepärasuhte teststatistik on:

$$\Lambda = \frac{\prod_{i=1}^{I} \prod_{j=1}^{J} (n_{i+}n_{+j}/n^2)^{n_{ij}}}{\prod_{i=1}^{I} \prod_{j=1}^{J} (n_{ij}/n)^{n_{ij}}}.$$

Teststatistiku töö põhimõttest aru saamiseks tuletame meelde, et suurima tõepära hinnang maksimiseerib tõepärafunktsiooni. Seega kui H_0 on väär, saavutab tõepärafunktsioon maksimumi Λ nimetajas ja $\Lambda \leq 1$. Mida väiksem, seda põhjendatum on H_0 kummutamine.

Viies Λ lugeja ja nimetaja ühise korrutismärgi alla, saame

$$\Lambda = \prod_{i=1}^{I}\prod_{j=1}^{J}\left(\frac{n_{i+}n_{+j}/n}{n_{ij}}\right)^{n_{ij}} = \prod_{i=1}^{I}\prod_{j=1}^{J}\left(\frac{\hat{m}_{ij}}{n_{ij}}\right)^{n_{ij}}, \tag{12}$$

kus $\hat{m}_{ij}$ on hinnatud oodatav ruudusagedus sõltumatuse eeldusel. On näidatud, et suhte (12) logaritmitud kuju, mida nimetatakse hälbimuseks, on nullhüpoteesi õigsuse korral asümptootiliselt hii-ruut jaotusega:

$$G = -2\ln\Lambda = 2\sum_{i}\sum_{j} n_{ij}\ln(n_{ij}/\hat{m}_{ij}), \tag{13}$$

$$G \sim \chi^2((I-1)(J-1)). \tag{14}$$

Mida suurem on G väärtus, seda põhjendatum on H_0 kummutamine.

Statistikutel H ja G on sama piirjaotus. Nad lähenevad oma hii-ruut jaotusele, kui n kasvab. Viimasega kaasneb ka oodatavate ruudusageduste $m_{ij} = n\pi_{ij}$ kasv. Üks üldlevinud kriteerium hii-ruudul põhinevate testide lubatavuse kohta on $m_{ij} \geq 5, \forall i, j$. Ühest olukorda siiski ei ole. Halvad juhud on hõredad tabelid, ka need, kus mõned ruudusagedused on väga väikesed ja mõned väga suured. Testi tulemuse kahtluse korral tuleb kasutada teisigi kontrolli võimalusi, ka näiteks täpseid teste.

Sõltuvuse kindlaks tegemine ei pruugi uurijale olla lõppeesmärgiks. Paljude tasemetega kvalitatiivsete tunnuste korral soovitakse sageli uurida sõltuvuse struktuuri, st missugused tasemed panustavad sõltuvusse kõige rohkem. Vaadatakse, missugused ruudusagedused on oodatavast palju väiksemad, missugused aga suuremad. Oodatav on sõltumatuse olukorra ruudusagedus. Selleks uuritakse jääke, kas Pearsoni jääke,

$$e_{ij} = \frac{n_{ij} - \hat{m}_{ij}}{\sqrt{\hat{m}_{ij}}},$$

või siis normeeritud jääke,

$$r_{ij} = \frac{n_{ij} - \hat{m}_{ij}}{\sqrt{\hat{m}_{ij}(1-p_{i+})(1-p_{+j})}}.$$

Viimaste väljalangemine ± 2 piiridest vihjab sellele, et tunnuste X ja Y tasemed i ja j panustavad sõltuvusse, mida tuleks siis ka interpreteerida.

Ülesanne 4.2. Näita, et $\hat{D}(n_{ij} - \hat{m}_{ij}) = \hat{m}_{ij}(1-p_{i+})(1-p_{+j})$.

Ülesanne 4.3. Testi, kas on seos hariduse ja jumalasse uskumise vahel.

Tabel 25: Haridus ja usk jumalasse (Sotsiaaluuring, Agresti 2013, lk 77)

	Usk jumalasse						
haridus	ei usu	pole võimalik tõestada	mingi kõrgem võim	vahel usub	usub kuid kahtleb	jumal on olemas	kokku
alla keskkooli	9	8	27	8	47	236	335
keskkool	23	39	88	49	179	706	1084
bakalaureus	28	48	89	19	104	293	581
kokku	60	95	204	76	330	1235	2000

4.3 Väikese valimi testid

Eespool toodud testid töötavad suure valimi korral, sest siis on teststatistik ligilähedaselt hii-ruut jaotusega, mis tema jaoks on tuletatud. Tänapäeval on tänu arvutitele võimalik teostada ka täpseid teste, mis töötavad väikeste valimite korral.

4.3.1 Fisheri täpne test sõltumatuse testimiseks

Fisher vaatles spetsiaalset katseplaani, kus 2×2 sagedustabeli marginaalsagedused n_{i+} ja n_{+j} on fikseeritud: Selles tabelis on ainult üks ruudusagedus juhuslik. Teised on temaga ja

$X \backslash Y$	1	2	summa
1	n_{11}	n_{12}	n_{1+}
2	n_{21}	n_{22}	n_{2+}
summa	n_{+1}	n_{+2}	n

ääresagedustega määratud. Olgu juhuslikuks n_{11}. Tahame kontrollida hüpoteesi $H_0 : X$ ja Y on sõltumatud, mis on samaväärne:

$$H_0 : \ \theta = 1.$$

Statistikuna vaatleme θ punkthinnangut,

$$\hat{\theta} = \frac{n_{11}\ n_{22}}{n_{12}\ n_{21}}.$$

Näeme, et alternatiivhüpoteesile $\theta \neq 1$ vihjab nii suur kui ka väike n_{11}. Kui suur on aga vea tõenäosus H_0 kummutamisel? Selle määramiseks on vaja teada $\hat{\theta}$ jaotust, siin siis n_{11} jaotust. Osutub, et sõltumatuse eeldusel on n_{11} jaotuseks hüpergeomeetriline jaotus:

$$p(t) = P(n_{11} = t) = \frac{C_{n_{1+}}^{t}\ C_{n_{2+}}^{n_{+1}-t}}{C_{n}^{n_{+1}}}. \tag{15}$$

Sagedus n_{11} saab omada täisarvulisi väärtusi, kitsenduste tõttu on olemas aga piirid (näita!):

$$n_{11} \leq min\{n_{1+}, n_{+1}\},$$
$$n_{11} \geq max\{0, n_{1+} + n_{+1} - n\}.$$

Hüpoteesi H_0 testime n_{11} väärtusega. Olgu antud valimil $n_{11} = t_0$, mille tagajärjel $\hat{\theta} > 1$. Võttes vastu $H_1 : \ \theta > 1$, saame vea tõenäosuse (p-value) arvutada saba summana valemi (15) abil:

$$p = P(n_{11} \geq t_0) = p(t_0) + p(t_0 + 1) + \dots .$$

Alternatiivi $H_1 : \ \theta < 1$ korral arvutame vasakpoolse saba tõenäosuse $p = P(n_{11} \leq t_0)$.

Näide. Fisheri teejooja (1935). Fisheri kolleeg, inglise leedi, väidab, et saab aru, kas piim või tee on valatud tassi esimesena, kui talle pakutakse piimaga teed. Fisher tegi katse: 8 tassi teed, 4-le piim esimesena. Leedi teadis, et 4 tassi on ühtmoodi ja 4 teistmoodi. Ta pidi andma 4 ennustust ühte liiki ja 4 teist. Tassid anti ette juhuslikus järjekorras.
$H_0 : \ \theta = 1$,
$H_1 : \ \theta > 1$, leedi väide on õige, on olema seos selle vahel, mis oli tegelikult ja mida pakkus leedi. Leiame, kui tõenäone on saada 3 või 4 õiget vastust H_0 eeldusel:

Tegelik	Arvamus I komponendi kohta		
I komponent	piim	tee	summa
piim	3	1	4
tee	1	3	4
summa	4	4	8

$$P(n_{11} = 3) = \frac{C_4^3 \, C_4^1}{C_8^4} = 0.229, \quad P(n_{11} = 4) = \frac{C_4^4 \, C_4^0}{C_8^4} = 0.014.$$

Seega olulisustõenäosus on $p = 0.229 + 0.014 = 0.243$. Ei saa kummutada H_0. Ka sõltumatuse eeldusel on suur tõenäosus äraarvamiseks. Kui oleks kõik 4 korda õigesti pakkunud, siis saaks inglise leedi võimeid uskuda, eksimistõenäosus oleks väike.

Suuremate kui 2×2 tabelite korral jätab ääresageduste fikseerimine rohkem vabu ruudutõenäosusi. Nende jaotuseks on sõltumatuse eeldusel mitmemõõtmeline hüpergeomeetriline jaotus. Arvuti on võimeline olulisustõenäosuse arvutama.

Märkus. Harilikult ei ole mõlemad marginaalsagedused fikseeritud. Näiteks Poissoni valiku korral (juhuslik objekt ajas või ruumis) ei ole ükski sagedus fikseeritud. Siis kehtib ülalkirjeldatud täpne test tinglikult, st tingimusel, et rea- ja veerusagedused on just need nagu antud katses realiseerusid.

4.4 Testid kolme kvalitatiivse tunnuse korral

Kolme tunnuse korral tekib rida praktilist tähtsust omavaid ülesandeid. Sageli on vaja uurida, kas 2 tunnuse vahel on seos, kui kolmas on fikseeritud. Kui tugev see seos on? Kas seos on ühetugevune kolmanda tunnuse eri tasemetel?

On rida uuringuid, kus ühe tunnuse tasemed on fikseeritud ja juhuslikud on 2 ülejäänut. Siis saabki teha tinglikku analüüsi. Mõnikord saab siiski teha ühise järelduse, üle kõigi fikseeritud tunnuse tasemete. Kirjeldatud 3 tunnuse analüüsi skeemi mahub näiteks metaanalüüs, kus sama nähtust on uuritud mitmetes uurimisgruppides ja nende tulemuste baasil tahetakse mingit ühist järeldust teha.

4.4.1 Tingliku sõltumatuse test logistilise modelleerimisega

Vaadatakse $2 \times 2 \times K$, $K \geq 2$ sagedustabelit. Seega tegu on 3 tunnusega, kus uuritav $Y \in \{0, 1\}$ on binaarne ja seletv $X \in \{x_1, x_2\}$ on samuti binaarne, $x_1 = 1, x_2 = 0$. Kolmandal tunnusel Z on K taset. Soovitakse uurida X mõju tunnusele Y fikseeritud Z korral.

Järgnevate meetodite illustreerimiseks kasutame sagedustabelit (Agresti 2013, lk 226).

Tabel 26: Kliiniline katse, mis uurib ravi ja tulemuse vahelist seost 8 keskuses

		Tulemus		shansside		
Keskus	Ravi	Edu	Ebaedu	suhe	μ_{11k}	$D(n_{11k})$
1	rohi	11	25	1.19	10.36	3.79
	place	10	27			
2	rohi	16	4	1.82	14.62	2.47
	place	22	10			
3	rohi	14	5	4.80	10.50	2.41
	place	7	12			
4	rohi	2	14	2.29	1.45	0.70
	place	1	16			
5	rohi	6	11	∞	3.52	1.20
	place	0	12			
6	rohi	1	10	∞	0.52	0.25
	place	0	10			
7	rohi	1	4	2.0	0.71	0.42
	place	1	8			
8	rohi	4	2	0.33	4.62	0.62
	place	6	1			

Olgu edu tõenäosus fikseeritud x_i ja k korral

$$\pi_{ik} = P(Y = 1|X = x_i, Z = k). \tag{16}$$

Tähistame edu log-shansse tavapäraselt

$$logit(\pi_{ik}) = \ln \frac{P(Y = 1|X = x_i, Z = k)}{1 - P(Y = 1|X = x_i, Z = k)}.$$

Eeldame edu log-shanssidele mudelit

$$logit(\pi_{ik}) = \alpha + \beta x_i + \beta_k^Z, \ i = 1, 2, \ k = 1, \ldots, K, \tag{17}$$

mis väljendab eeldust, et shansside suhe on igas osatabelis k sama, nimelt

$$\theta = e^{\beta}.$$

Tingliku sõltumatuse testimiseks on nüüd vaja kontrollida hüpoteesi

$$H_0 : \ \beta = 0.$$

Parameetrit β hinnatakse logistilises regressioonis suurima tõepära meetodi abil, kusjuures kasutatakse kõiki osatabeleid. Eelduseks on, et β on sama kõigis osatabelites. Meetod annab ka hinnangu standardvea SE ja Wald'i statistik nullhüpoteesi kontrolliks on $(\hat{\beta}/SE)^2$, mis on hii-ruut jaotusega vabadusasmete arvuga 1. Saab moodustada ka tõepärasuhte teststatistiku.

Vaadates tabelit 26, näeme et shansside suhted on kõigis osatabelites v.a. viimases samasuunalised. Seega on mõtet kontrollida hüpoteesi $H_0: \ \beta = 0$. Hinnanguks saadi $\hat{\beta} = 0.777$ standardveaga $SE = 0.307$. Waldi statistiku väärtus on 6.42 (p=0.011), mis räägib nullhüpoteesi vastu, ehk kinnitab seose olemasolu rohu võtmise ja ravitulemuse vahel. Kuna $\hat{\beta} > 0$, siis seos on positiivne.

4.4.2 Cochran-Mantel-Haenzeli test tinglikuks sõltumatuseks

Test on mõeldud kolmele tunnusele, kus X ja Y on tabuleeritud 2×2 tabelisse igal Z tunnuse tasemel. Viimasel on K taset. See test ei baseeru mudelile. Nullhüpoteesiks on

$$H_0: \ X \perp Y, \text{ kui } Z \text{ on fikseeritud.}$$

Testil on sarnaseid jooni Fisheri testiga, sest igas 2×2 tabelis eeldatakse fikseeritud marginaalsagedusi ja nii jääb ainult 1 ruudusagedus juhuslikuks. Retrospektiivse või prospektiivse uuringu korral on ühed marginaalsagedused katseplaani kohaselt fikseeritud. Kui marginaalsagedused ei ole fikseeritud, rakendatakse testi siiski, kuid teoreetiliselt taandub sõltumatuse kontroll siis realiseerunud marginaalsagedustega tabelitele. Praktikas üldistatakse tulemust ka juhuslike marginaalsagedustega tabelitele, kuid siis p-value on vaid ligikaudne.

Kolme tunnuse korral on sagedusel 3 indeksit, n_{ijk}. Olgu n_{11k} juhuslik ruudusagedus tabelis k ($Z = k$ korral). Selle tabeli marginaalsagedused n_{1+k}, n_{2+k} ja n_{+1k}, n_{+2k} on fikseeritud, siis H_0 kehtivuse korral on n_{11k} hüpergeomeetrilise jaotusega, kusjuures

$$\begin{aligned} \mu_{11k} &= E(n_{11k}) = n_{1+k}n_{+1k}/n_{++k} \\ D(n_{11k}) &= n_{1+k}n_{2+k}n_{+1k}n_{+2k}/[n_{++k}^2(n_{++k}-1)] \end{aligned}$$

Teststatistik summeerib n_{11k}, mis on sõltumatud juhuslikud suurused $\forall k$, mistõttu summa on ligikaudu normaaljaotusega. Summa normeeritakse ja tsentreeritakse, kasutades nullhüpoteesi ooteväärtust:

$$CMH = \frac{[\sum_k (n_{11k} - \mu_{11k})]^2}{\sum_k D(n_{11k})}.$$

Suure valimi korral on CMH ligikaudu hii-ruutjaotusega vabadusastmete arvuga 1. Kui $\theta_{XYk} > 1$, siis oodatavalt on realisatsioonid $n_{11k} - \mu_{11k} > 0$. Kui kõik $\theta_{XYk} > 1$, siis tuleb CMH väärtus suur ja sama toimub kui kõik $\theta_{XYk} < 1$, mis viitab sõltuvusele X ja Y vahel fikseeritud Z korral.

Paneme, tähele, et osatabelid, kus pole ühtegi edu või ühtegi ebaedu, ei anna informatsiooni sõltuvuse kohta. CMH-statistikusse need osatabelid ei panusta, sest siis $n_{11k} = \mu_{11k}, \ D(n_{11k} = 0)$.
Vaadates tabelit 26, näeme et shansside suhted on kõigis osatabelites v.a. viimases samasuunalised. Seega on mõtet kombineerida tabelid CMH arvutamiseks. Saame $CMH = 6.38, \ df = 1$, mis on küllaldane nullhüpoteesi kummutamiseks, p=0.012.

R'is teeb seda testi mantelhaen.test(). CMH statistikut on üldistatud ka $I \times J \times K$ tabelitele.

Ülesanne 4.4. Uurides tabelit (14), kas on sõltuvus süüaluse rassi ja surmanuhtluse kui karistuse vahe, kui ohvri rass on fikseeritud.

4.4.3 Hinnang ühisele shansside suhtele

Seosetugevuse näitaja on informatiivsem suurus, kui hüpoteeside testimisega sõltuvuse tuvastamine.

Kui tinglikes osatabelites on seos stabiilne, st üldiselt $n_{11k} > \mu_{11k}$ (või üldiselt vastupidi), siis Mantel-Haenzel (1959) pakkusid välja ühisele shansside suhtele hinnangu,

$$\hat{\theta}_{MH} = \frac{\sum_{k=1}^{K} n_{11k}\, n_{22k}/n_{++k}}{\sum_{k=1}^{K} n_{12k} n_{21k}/n_{++k}}.$$

Kuna sõltumatuse korral osatabelites $\theta = 1$ ja hinnang peaks olema 1 lähedal, siis vastavalt $n_{11k}\, n_{22k} \approx n_{12k}\, n_{21k}$, mistõttu ka $\hat{\theta}_{MH} \approx 1$.

4.4.4 Meta-analüüs 2×2 tabelitele

Meta-analüüs on statistiline analüüs, mis kombineerib informatsiooni mitmetest uuringutest. Kui uurimustes võrreldakse binaarset ravitulemust kahes erinevas grupis, on tegu $2 \times 2 \times K$ tabelitega.

Eeldatakse, et parameetri väärtus, mille uurimisele kõik K uuringut püüdlevad on sama. Kuigi see on lihtsustav eeldus, osutub, et meetod töötab ka siis, kui parameetriväärtused erinevad veidi.

Tingliku sõltumatuse kontrolliks sobivad eelpool kirjeldatud meetodid (logistiline modelleerimine ja CMH-test). Ühise sõltuvusparameetri β või $\theta = e^{\beta}$ hindamiseks sobib nii suurima tõepära meetod kui ka ühise shansside suhte hinnang $\hat{\theta}_{MH}$.

Tõenäosuste vahe on ka tähtis seosetugevuse näitaja

$$\delta = \pi_{1k} - \pi_{2k},$$

kus $\pi_{ik} = P(Y = 1 | X = x_i, Z = k)$. Eeldades, et δ on sama kõigis osatabelites, saame vaadata mudelit (kontrolli, et δ on tõenäosuste vahe)

$$\pi_{ik} = \alpha + \delta x_i + \beta_k^Z,\ i = 1, 2,\ k = 1, \ldots, K. \tag{18}$$

Mudelit kasutades, saame standardse tehnikaga suurima tõepära hinnangu $\hat{\delta}$.

Kirjutame veel välja Mantel-Haenszel tüüpi hinnangu ühisele parameetrile δ. Vaatame osatabelit k ja osakaalude vahet selles,

$$\hat{\delta}_k = \frac{n_{11k}}{n_{1+k}} - \frac{n_{21k}}{n_{2+k}}.$$

Hinnang ühisele parameetrile δ moodustatakse kaalutud keskmisena,

$$\hat{\delta} = \frac{\sum_k w_k \hat{\delta}_k}{\sum_k w_k},$$

kus kaalud on $w_k = n_{1+k} n_{2+k} / (n_{1+k} + n_{2+k})$. Hinnangule on olemas ka dispersioonihinnang (Agresti 2013, lk 231).

Ülesanne 4.5. Veendu, et kaalud summeeruvad 1-ks.

Märkus. Seosekordajate arvutamine koos usalduspiiridega annab rohkem informatsiooni sõltuvuse kohta kui lihtsalt testimine.

4.5 Suure valimi vahemikhinnangud parameetritele

Kahe kvalitatiivse tunnuse korral on meid seni huvitanud:

$$\pi_{ij},\ \pi_{j|i},\ \pi_{i+},\ \Omega_i,\ R,\ \theta,\ U,\ \gamma.$$

Oleme vaadelnud punkhinnanguid neile parameetritele ja selgitanud, missugustelt sagedustabelitelt mida hinnata saame. Selles peatükis kontseentreerume vahemikhinnangutele, et väljendada hinnangute usaldusväärsust.

Suure valimi vahemikhinnang baseerub hinnangu asümptootilisele jaotusele. Tavaliselt on selleks normaaljaotus. Veelgi enam, hinnangu funktsiooni jaotus on enamasti samuti normaaljaotus.

Olgu θ tundmatu parameeter ja T_n selle hinnang, nii et

$$\sqrt{n}(T_n - \theta) \ \xrightarrow{\mathcal{D}}\ N(0, \sigma^2). \tag{19}$$

Avaldis ütleb, et suure valimi korral on T_n normaaljaotusega, kusjuures keskväärtus ja dispersioon on

$$E(T_n) = \theta, \quad D(T_n) = \frac{\sigma^2}{n}$$

ning vahemikhinnangu θ-le saame normaaljaotuse kvantiile kasutades:

$$UI(\theta) = T_n \pm \lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}},$$

mille usaldusnivooks on

$$P\left(UI(\theta) \ni \theta\right) \approx 1 - \alpha.$$

Enamasti tuleb ka σ asendada tema hinnanguga valimist.

Kui soovime hinnata funktsiooni $g(\theta)$, siis saame seda teha hinnanguga $g(T_n)$, mis on samuti normaaljaotusega, kui vaid g on 2 korda differentseeruv θ ümbruses. Väide tuleneb Taylori rittaarendusest,

$$g(T_n) = g(\theta) + g'(\theta)(T_n - \theta) + R,$$

millest

$$\sqrt{n}(g(T_n) - g(\theta)) = \sqrt{n}g'(\theta)(T_n - \theta) + R.$$

Vasakul pool oleva suuruse käitumise määrab paremal olev lineaarliige, jäägi R osatähtsus kahaneb kiiresti n kasvades. Seega suure valimi korral

$$Eg(T_n) = g(\theta), \quad Dg(T_n) = [g'(\theta)]^2 D(T_n) = \frac{[g'(\theta)]^2\sigma^2}{n}, \tag{20}$$

ja (19) tõttu

$$\sqrt{n}(g(T_n) - g(\theta)) \ \xrightarrow{\mathcal{D}}\ N(0, [g'(\theta)]^2\sigma^2). \tag{21}$$

Usaldusintervalli ligikaudsel usaldusnivool $1 - \alpha$ saame normaaljaotuse kvantiilide abil:

$$UI[g(\theta)] = g(T_n) \pm \lambda_{\alpha/2}\frac{g'(\theta)\sigma}{\sqrt{n}}.$$

Tulemust (21) nimetatakse delta meetodiks asümptootilise jaotuse saamisel. Paneme tähele, et praegusel kujul sisaldab usaldusintervall tundmatuid suurusi $g(\theta)$ ja ka σ võib sõltuda θ-st. Asendades need hinnangutega, saame nn Wald'i usaldusintervalli:

$$g(T_n) \pm \lambda_{\alpha/2} \frac{g'(T_n)\sigma(T_n)}{\sqrt{n}}.$$

Ülesanne 4.6. Olgu tegu ristlõikelise uuringuga. Pane kirja punkthinnang ja vahemikhinnang parameetrile π_{ij}.

Ülesanne 4.7. Olgu tegu prospektiivse uuringuga. Pane kirja punkthinnang ja vahemikhinnang parameetrile $\pi_{j|i}$.

Paljud meie huvialused parameetrid nõuavad mitmemõõtmelist delta meetodit, sest nad on mitme argumendi funktsioonid, nagu näiteks $\theta = \pi_{11}\pi_{22}/[\pi_{12}\pi_{21}]$. Sel juhul läheb vaja osatuletiste vektorit. Sõnastame asümptootilise jaotuse multinomiaalsete sageduste funktsioonide jaoks.

Olgu

$$(n_1, \ldots, n_N) \sim M(n; \pi_1, \ldots, \pi_N), \ \sum n_i = n, \ \sum \pi_i = 1.$$

Teame,

$$\begin{aligned} E(n_i) &= n\pi_i, \\ D(n_i) &= n\pi_i(1-\pi_i), \\ Cov(n_i, n_j) &= -n\pi_i\pi_j. \end{aligned}$$

Tõenäosuste hinnanguteks on multinomiaalse mudeli korral suhtelised sagedused $p_i = \frac{n_i}{n}$. Nende jaoks järeldub ülaltoodud valemitest:

$$E(p_i) = \pi_i, \tag{22}$$
$$D(p_i) = \pi_i(1-\pi_i)/n, \tag{23}$$
$$Cov(p_i, p_j) = -\pi_i\pi_j/n. \tag{24}$$

Läheme maatrikskujule:

$$\pi = (\pi_1, \ldots, \pi_N)'$$
$$p = (p_1, \ldots, p_N)'.$$

Siis seostest (22)-(24) järeldub:

$$\begin{aligned} E(p) &= \pi, \quad N \times 1, \\ D(p) &= [diag(\pi) - \pi\pi']/n, \quad N \times N. \end{aligned}$$

Olgu meid huvitav suhteliste sageduste funktsioon $g(p): \ R^N \to R^1$.

Teoreem (C.R. Rao 1973). Multinomiaalse mudeli korral sagedustele, kehtib:

$$\sqrt{n}(p - \pi) \ \xrightarrow{\mathcal{D}} \ N(0, \Sigma), \text{ kus } \Sigma = diag(\pi) - \pi\pi',$$

samuti kehtib funktsioonide $g(p)$ jaoks

$$\sqrt{n}(g(p) - g(\pi)) \xrightarrow{\mathcal{D}} N(0, \sigma^2), \text{ kus } \sigma^2 = \phi'\Sigma\phi.$$

Vektor $\phi = (\phi_1, \ldots, \phi_N)'$ on osatuletiste vektor

$$\phi_i = \frac{\partial g(p)}{\partial p_i}|_{p=\pi}, \ \forall i.$$

Järeldus. Alternatiivne kuju dispersioonile σ^2 on

$$\sigma^2 = \sum_{i=1}^{N} \pi_i \phi_i^2 - (\sum_{i=1}^{N} \pi_i \phi_i)^2. \tag{25}$$

Tõestus. Rao teoreemist:

$$\begin{aligned} \sigma^2 &= \phi'\Sigma\phi \\ &= \phi'[diag(\pi) - \pi\pi']\phi \\ &= \phi' \begin{bmatrix} \pi_i & & 0 \\ & \ddots & \\ 0 & & \pi_N \end{bmatrix} \phi - \underbrace{\phi'\pi}_{\text{skalaar}} \pi'\phi \end{aligned}$$

Kuna $\phi'\pi$ on skalaar, siis summa teine liige tuleb $(\phi'\pi)^2$. Kirjutades maatrikstehte lahti summadena saamegi soovitud tulemuse:

$$\sigma^2 = \sum_{i=1}^{N} \pi_i \phi_i^2 - (\sum_{i=1}^{N} \pi_i \phi_i)^2.$$

□

4.5.1 Vahemikhinnang shansside suhtele

Tuletame meelde shansside suhte. Olgu vaadeldud kahte tunnust X ja Y, mõlemad 2 tasemega - $I = 2$ ja $J = 2$ ning üldkogumijaotusega $\{\pi_{ij}\}$:

Tabel 27: X ja Y ühisjaotus

$X \backslash Y$	1	2
1	π_{11}	π_{12}
2	π_{21}	π_{22}

Shansside suhte arvutame järgmiselt:

$$\theta = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}; \qquad \theta \in [0, \infty); \qquad \text{kui } \theta = 1 \Rightarrow X \perp Y \tag{26}$$

Reaalsete andmete puhul saame leida shansside suhte hinnangu kasutades sagedusi:

$$\hat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{p_{11}p_{22}}{p_{12}p_{21}}. \tag{27}$$

Hinnangu $\hat{\theta}$ saame leida igasuguse uuringu jaoks aga $\hat{\theta}$ jaotus on ebasümmeetriline, paremale poole pikema sabaga. Selle tõttu on koondumine normaaljaotuseks aeglane. Vaatame aga hoopis log-shansside suhet, see on sümmeetriline ja koondub normaal jaotuseks kiiremini.

Teoreem
Log-shansside suhte osas kehtib:

$$\sqrt{n}(\ln\hat{\theta} - \ln\theta) \xrightarrow{\mathcal{D}} N(0, \sigma^2), \tag{28}$$

kus

$$\sigma^2 = \frac{1}{\pi_{11}} + \frac{1}{\pi_{12}} + \frac{1}{\pi_{21}} + \frac{1}{\pi_{22}}. \tag{29}$$

Tõestus. Teoreemi tõestuseks kasutame Rao teoreemi. Näitame kuidas saame $g(\pi)$ ja $g(p)$:

$$\begin{aligned} \ln\theta &\underset{(26)}{=} \ln\pi_{11} + \ln\pi_{22} - \ln\pi_{12} - \ln\pi_{21} = g(\pi) \\ \ln\hat{\theta} &\underset{(27)}{=} \ln p_{11} + \ln p_{22} - \ln p_{12} - \ln p_{21} = g(p), \end{aligned}$$

kus $\pi = (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})$ ja $p = (p_{11}, p_{12}, p_{21}, p_{22})$. Vastavalt Rao teoreemile kehtib (28), kus (NB! nüüd tuleb 2 indeksit)

$$\sigma^2 = \sum_i \sum_j \pi_{ij}\phi_{ij}^2 - \Big(\sum_i \sum_j \pi_{ij}\phi_{ij}\Big)^2.$$

Leiame osatuletised

$$\phi_{ij} = \frac{\partial g(p)}{\partial p_{ij}}|_{p=\pi}$$

$$\phi_{11} = \frac{1}{\pi_{11}}; \quad \phi_{12} = -\frac{1}{\pi_{12}}; \quad \phi_{21} = -\frac{1}{\pi_{21}}; \quad \phi_{22} = \frac{1}{\pi_{22}}.$$

Nüüd $\sum_i \sum_j \pi_{ij}\phi_{ij} = 0$, seega

$$\sigma^2 = \sum_i \sum_j \pi_{ij}\phi_{ij}^2 = \sum_i \sum_j \frac{1}{\pi_{ij}}$$

□

Järeldus. Suure valimi korral:

$$\ln\hat{\theta} \sim N\left(\ln\theta, \frac{\sigma^2}{n}\right),$$

kus σ^2 on antud valemiga (29). Järelikult

$$\hat{D}\big[\ln\hat{\theta}\big] = \frac{\hat{\sigma}^2}{n} = \frac{1}{n}\Big[\frac{1}{\hat{\pi}_{11}} + \cdots + \frac{1}{\hat{\pi}_{22}}\Big] = \frac{1}{n}\Big[\frac{n}{n_{11}} + \cdots + \frac{n}{n_{22}}\Big] = \frac{1}{n_{11}} + \cdots + \frac{1}{n_{22}}.$$

Seega on usaldusvahemik $\ln\theta$-le suure valimi korral:

$$UI[\ln\theta] = \ln\hat{\theta} \pm \lambda_{\alpha/2}\Big[\sum_i \sum_j \frac{1}{n_{ij}}\Big]^{1/2}.$$

Siit saab tuletada usaldusvahemiku θ jaoks järgmise aruteluga. Kui

$$P(a < \ln\theta < b) = 1 - \alpha,$$

kus

$$\begin{aligned} a &= \ln\hat{\theta} - \lambda_{\alpha/2}\Big[\sum_i \sum_j \frac{1}{n_{ij}}\Big]^{1/2} \\ b &= \ln\hat{\theta} + \lambda_{\alpha/2}\Big[\sum_i \sum_j \frac{1}{n_{ij}}\Big]^{1/2}, \end{aligned}$$

siis tänu faktile, et eksponentsiaalne funktsioon on monotoonselt kasvav funktsioon, siis

$$P(e^a < \theta < e^b) = 1 - \alpha.$$

Ehk $UI(\theta) = (e^a; e^b)$.

Probleem: Kui mõni $n_{ij} = 0$, siis pole $\ln\hat{\theta}$ või $\ln\hat{\theta}$ standardviga määratud (see probleem kaob kui n kasvab). Selle tõttu kasutatakse praktikas alternatiivset shansside suhte hinnangut (nö. pidevuse korrektsiooniga hinnangut):

$$\tilde{\theta} = \frac{(n_{11} + 0.5)(n_{22} + 0.5)}{(n_{12} + 0.5)(n_{21} + 0.5)}$$

. Suure n korral $\hat{\theta} \approx \tilde{\theta}$ ja alternatiivse hinnangu $\tilde{\theta}$ asümptootiline jaotus on sama mis $\hat{\theta}$ jaotus. Seega praktikas

$$UI[\ln\theta] = \ln\tilde{\theta} \pm \lambda_{\alpha/2}\Big[\sum_i \sum_j \frac{1}{n_{ij} + 0.5}\Big]^{1/2}.$$

Ülesanne 4.8. On olnud retrospektiivne uuring ja andmed on tabelis 28. Leia shansside suhte hinnang ja ka vahemikhinnang. Interpreteeri.

Tabel 28: kopsuvähk ja suitsetamine

Suits.\Haigus	kopsuvähk	kontrollgr
ei (<5)	62	190
jaa (≥ 5)	1295	1197

4.5.2 Vahemikhinnang suhtelisele riskile

Vaatame suhtelist riski (tee ise tinglike tõenäosuste tabel)),

$$R = \frac{\pi_{1|1}}{\pi_{1|2}},\ R = 1,\ \text{kui } X \perp Y.$$

Järgnev tuletuskäikk on tehtud eeldusel, et n_{1+} ja n_{+1} on fikseeritud (prospektiivne uuring, ristlõikelise uuringu korral kasuta tuletamisel Rao teoreemi). Hinnang $\hat{R}$ on valimi osakaalude suhe (millist tüüpi uuringute korral?)

$$\hat{R} = \frac{p_{1|1}}{p_{1|2}}, \text{ kus } p_{1|1} = \frac{n_{11}}{n_{1+}}.$$

Kuna logaritmimine teeb $\hat{R}$ jaotust sümmeetrilisemaks, mis koondub normaaljatuseks kiiremini, siis vaatame seda:

$$\ln \hat{R} = \ln p_{1|1} - \ln p_{1|2}. \tag{30}$$

Vaatame liidetavaid eraldi. Vaja on leida $D(\ln p_{1|1})$. Tähistades $\theta = \pi_{1|1}$, $g(\theta) = \ln \pi_{1|1}$ ja $g(T_n) = \ln p_{1|1}$, saame valemist (20)

$$D(\ln p_{1|1}) = g'(\theta)^2 D(p_{1|1}) = \frac{1}{\pi_{1|1}^2} D(p_{1|1}).$$

Kuna $n_{11} \sim B(n_{1+}, \pi_{1|1})$, siis

$$D(p_{1|1}) = \frac{D(n_{11})}{n_{1+}^2} = \frac{\pi_{1|1}(1-\pi_{1|1})}{n_{1+}}.$$

Lõppkokkuvõttes

$$D(\ln p_{1|1}) = \frac{1-\pi_{1|1}}{n_{1+}\pi_{1|1}} (= D_1).$$

Samamoodi saame teise liidetava jaoks valemis (30)

$$D(\ln p_{1|2}) = \frac{1-\pi_{1|2}}{n_{2+}\pi_{1|2}} (= D_2).$$

Kuna liidetavad on sõltumatud, siis $\ln \hat{R}$ jaoks dispersioonid liituvad. Oleme saanud $\ln \hat{R}$ asümptootilise jaotuse

$$\ln \hat{R} \sim N(\ln R,\ D_1 + D_2).$$

Ülesanne 4.9. Näita, et

$$\hat{D}_1 + \hat{D}_2 = \frac{1}{n_{11}} + \frac{1}{n_{21}} - \frac{1}{n_{1+}} - \frac{1}{n_{2+}}.$$

Pane kirja $UI(\ln R)$ ja eksponentteisendusega $UI(R)$.

Märkus. Kui mõni sagedustabeli element on 0, kasutatakse pidevusparandust $+0.5$ kõigile sagedustele.

4.5.3 Vahemikhinnang tinglike tõenäosuste vahele

Vaatame tinglike tõenäosuste vahet $\pi_{1|1} - \pi_{1|2}$. Selle hinnanguks on osakaalude vahe $p_{1|1} - p_{1|2}$ (millist tüüpi uuringute korral?). Liidetavad on siin sõltumatud juhuslikud suurused.

Ülesanne 4.10. Tuleta usaldusvahemik vahele $\pi_{1|1} - \pi_{1|2}$.

Kui see on tervenisti positiivsel poolel, on $\pi_{1|1} > \pi_{1|2}$ tõenäosusega $1-\alpha$. Saame isegi väita, kui palju on esimene tõenäosus suurem teisest.

4.6 Statistilised otsustused paarisvaatluste korral

Seni vaatlesime olukorda, kus kahe tunnuse X, Y sõltuvuse väljaselgitamiseks uurisime nende ühisjaotust, võrdlesime tinglikke tõenäosusi ja nende baasil defineeritud sõltuvuskordajaid (shansside suhet, suhtelist riski jt). Teostasime testimist ja leidsime vahemikhinnanguid. Mõne ülesande korral on vaja aga hoopis võrrelda X, Y marginaaljaotusi.

Vajadus marginaaljaotuste võrdlemiseks tekib kui mõõdetakse sama tunnust paarisobjektil

- paarisobjektiks on
  - ◇ sama objekt kahel ajahetkel (vererõhk enne ja pärast ravimit);
  - ◇ sama objekti kaks kohta (vasaku ja parema silma nägemisteravus);
  - ◇ kaks hinnangu andjat samale objektile (veini maitseomaduste hindamine);
  - ◇ lapse vanem, (tunnus haridus);

Kui põhiline eeldus klassikalises statistikas on sõltumatus objektide vahel, siis siin on see rikutud, paarisobjektid on sõltuvad. Viimast tuleb arvesse võtta statistiliste otsustuste (testid, usalduspiirid) tegemisel.

Kuna mõõdetakse sama tunnust, siis tasemete arv on sama $I = J$. Sagedustabel on ruudukujuline ja marginaaljaotused on sama dimensiooniga. On paari määratlev tunnus Z (ravim, asukoht, hindaja) kahe väärtusega 1, 2. Olgu X uuritav tunnus $Z = 1$ korral ja Y uuritav tunnus $Z = 2$ korral. Eesmärgiks sõltuvuse tuvastamine tunnusest Z.

Uuritakse tunnuste X, Y marginaaljaotuste homogeensust. Marginaaljaotusi nimetatakse homogeenseteks, kui teoreetilised marginaaltõenäosuste vektorid langevad kokku. Kokkulangemise korral ei ole tunnusel Z uuritavale tunnusele mõju.

Vaatame juhtu $I = 2$. Soovime võrrelda X, Y marginaaljaotusi, st tõenäosuste vektoreid (π_{1+}, π_{2+}) ja (π_{+1}, π_{+2}). See taandub küsimusele, kas $\pi_{1+} = \pi_{+1}$. Sõltuvuse tuvastamiseks soovime testida nullhüpoteesi

$$H_0 : \ \pi_{1+} - \pi_{+1} = 0.$$

Nullhüpotees väidab, et tunnusel Z ei ole mõju. Näeme, et nullhüpoteesi kehtimisel on ühisjaotuse tabel sümmeetriline, $\pi_{12} = \pi_{21}$, Seega võiksime ka ühisjaotuse sümmeetriat testida valimi põhjal.

Homogeensuse testimiseks kasutame sagedustabelit ja eeldame sellele multinomiaalset mudelit

$$\{n_{ij}\} \sim M(n; \{\pi_{ij}\}).$$

Hüpoteesis H_0 toodud tõenäosuste vahet hindame osakaalude vahega:

$$d = p_{1+} - p_{+1}.$$

H_0 saame testida usaldusvahemikuga või teststatistikuga, mis on konstrueeritud d abil. Selleks on vaja leida dispersioon $D(d)$:

$$D(d) = D(p_{1+} - p_{+1}) = \frac{1}{n^2} D(n_{1+} - n_{+1}) = \frac{1}{n^2} D(n_{12} - n_{21}).$$

Kuna sageduste $\{n_{ij}\}$ ühisjaotuseks on multinomiaaljaotus, siis $n_{12} \sim B(n, \pi_{12})$ ja $n_{21} \sim B(n, \pi_{21})$. Nüüd saame dispersioonile avaldise:

$$D(d) = \frac{1}{n^2}[D(n_{12}) + D(n_{21}) - 2cov(n_{12}, n_{21})] = \frac{1}{n^2}[n\pi_{12}(1-\pi_{12}) + n\pi_{21}(1-\pi_{21}) - 2n\pi_{12}\pi_{21}].$$

Pärast sulgude avamist saame

$$D(d) = \frac{1}{n}[\pi_{12} + \pi_{21} - (\pi_{12} - \pi_{21})^2].$$

Asendades tõenäosused nende hinnangutega, st suhteliste sagedustega, saame $\hat{D}(d)$. Nüüd on käes suure valimi usalduspiirid tõenäosuste vahele

$$UI(\pi_{1+} - \pi_{+1}) = p_{1+} - p_{+1} \pm \lambda_{\alpha/2}\sqrt{\hat{D}(d)}.$$

Kui vahemik katab 0, ei saa kummutada marginaaljaotuste homogeensuse hüpoteesi H_0. Kui kummutame, saame lisaks öelda, kui palju on üks tõenäosus teisest suurem.

Leiame nüüd teststatistiku H_0 kontrollimiseks. Meil suure valimi korral

$$d \sim N(\pi_{1+} - \pi_{+1}, \hat{D}(d)),$$

mis H_0 kehtimise korral saab kuju

$$d \sim N(0, \hat{D}(d)).$$

Normeerides, saame teststatistiku H_0 kontrollimiseks,

$$Z = \frac{d}{\sqrt{\hat{D}(d)}} \sim N(0, 1).$$

Statistiku väärtust võrdleme normaaljaotuse kvantiiliga, või leiame statistiku väärtuse olulisuse tõenäosuse.

Osutub, et teststatistik Z lihtsustub veelgi. H_0 eeldusel teiseneb dispersioon $D(d)$ kujule

$$D(d) = \frac{1}{n}[\pi_{12} + \pi_{21}],$$

mistõttu tema hinnang saab kuju

$$\hat{D}(d) = \frac{1}{n^2}[n_{12} + n_{21}].$$

Statistik Z saab nüüd kuju

$$Z = \frac{d}{\sqrt{\hat{D}(d)}} = \frac{p_{1+} - p_{+1}}{\sqrt{n_{12} + n_{21}}/n} = \frac{n_{12} - n_{21}}{(n_{12} + n_{21})^{1/2}}.$$

Võttes ruutu, saame hii-ruut jaotusega teststatistiku

$$Z^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}} \sim \chi^2(1).$$

H_0 testimist selle statistiku abil nimetatakse McNemari testiks (1947). SAS-protseduur freq annab selle testi valikuga /agree. R'is teeb seda mcnemar.test().

Näeme, et peadiagonaali elemendid on ebaolulised testimaks marginaaljaotuste homogeensust. Kui kõrvaldiagonaali elemendid erinevad palju, tuleb statistiku väärtus suur. Kui piisavalt palju suur, kummutatakse H_0.

5 Kvalitatiivse tunnuse modelleerimine

Kui uuritavaks tunnuseks on kvalitativne tunnus, siis kuidas modelleerida mittearvulist tunnust? Üks võimalus on modelleerida väärtuse esinemise tõenäosust. Uuritakse, kas tõenäosus kasvab või kahaneb seletavate tunnuste muutudes. Teine võimalus on modelleerida uuritava tunnuse oodatavat sagedust n vaatluse hulgas sõltuvana argumenttunnustest.

Kasutatavad mudelid kuuluvad üldistatud lineaarsete mudelite (ÜLM) klassi. ÜLM kohta on eraldi kursus. Siin toome ÜLM mõiste lühidalt ja vaatame kvalitatiivse tunnuse jaoks sobivaid konkreetseid mudeleid.

Nii nagu tavaliselt eeldatakse mudeli püstituses, et uuritav tunnus on juhuslik ja tema jaotus kuulub mingisse jaotuste perre. Abitunnused ehk seletavad tunnused eeldatakse olevat mittejuhuslikud. Uuritava ja abitunnuste vahel defineeritakse sobival viisil seos.

Olgu meil $k+1$ seletavat tunnust $X_0, X_1, \ldots, X_k$. Tavaliselt X_0 on konstantselt 1, see tekitab vabaliikme mudelisse. Olgu meie vaatlused,

$$(y_i, \mathbf{x}_i),\ i = 1, 2, \ldots, n,$$

kus $\mathbf{x}_i = (x_{i0}, x_{i1}, \ldots x_{ik})$ on seletavate tunnuste väärtused objektil i. Väärtus y_i on juhusliku suuruse Y_i realisatsioon objektil i. Täpsemini on tegemist tingliku juhusliku suurusega, mis sõltub seletavate tunnuste vektorist, $Y_i = Y|\mathbf{x}_i$.

ÜLM eeldused:

- Uuritava tunnuse jaotus on eksponentsiaalsest perest:

$$Y_i \sim f(y; \theta_i),\ \text{kus}\ f(y, \theta_i) = a(\theta_i) b(y) \exp[y\, Q(\theta_i);]$$

- keskväärtuse mingi funktsioon $g(\cdot)$ on lineaarselt seotud seletavate tunnustega:

$$g[E(Y_i)] = \sum_{j=0}^{k} \beta_j x_{ij}.$$

Seonduvad mõisted:
y – tihedusfunktsiooni argument;
θ_i – parameeter ($\forall Y_i$ jaoks oma, st sõltub $\mathbf{x}_i$-st);
$Q(\theta_i)$ – parameetri funktsioon (loomulik parameeter);
$a(\cdot),\ b(\cdot)$ – mingid funktsioonid;
$g(\cdot)$ – linkfunktsioon, seosefunktsioon.
$\beta_j, \forall j$ – kordajad, mille hindamine meid huvitab (iseloomustavad tunnuste X_j mõju uuritavale tunnusele).

Klassikaline ÜLM erijuht on:

$$\begin{aligned} Y_i &\sim N(EY_i, \sigma_i^2),\ \text{normaaljaotus on eksponentsiaalse pere liige;} \\ g(EY_i) &= EY_i = \sum_{j=0}^{k} \beta_j x_{ij},\ \text{samasusteisendus, ühiklink.} \end{aligned}$$

Sõltuvalt seletavate tunnuste tüübist on tegu:

- mitmese lineaarse regressiooniga (arvulised paljude väärtustega);
- dispersioonanalüüsiga (kvalitatiivsed, väheste väärtustega arvulised);
- üldise lineaarse mudeliga (igat tüüpi seletavad tunnused).

5.1 Logit-mudel ehk logistiline regressioon

Siin on uuritav tunnus binaarne. Väärtusi kodeerime $\{0,1\}$, kus 1 olgu "edu". Seega

$$Y_i = Y|\mathbf{x}_i \sim B(1, \pi(\mathbf{x}_i)).$$

Meid huvitab "edu" tõenäosuse sõltuvus abitunnustest:

$$\pi(\mathbf{x}_i) = P(Y = 1|\mathbf{x}_i).$$

Paneme tähele, et Bernoulli jaotuse korral on keskväärtus võrdne "edu" tõenäosusega:

$$EY_i = \pi(\mathbf{x}_i).$$

Samuti paneme tähele, et Bernoulli jaotus kuulub eksponentsiaalsesse perre. Diskreetse juhusliku suuruse korral asendub tihedus kohal y tõenäosusega saada väärtust y. Seega:

$$f(y, \pi(\mathbf{x}_i)) = P(Y_i = y) = \pi(\mathbf{x}_i)^y (1 - \pi(\mathbf{x}_i))^{1-y} \tag{31}$$

$$= (1 - \pi(\mathbf{x}_i)) \left(\frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)} \right)^y = (1 - \pi(\mathbf{x}_i)) \exp[y \ln \frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)}]. \tag{32}$$

Viimases avaldises tunneme ära eksponentsiaalse pere kuju. Tõenäosuse funktsiooni selles avaldises nimetatakse logit'iks, tähistame $l(\mathbf{x}_i)$:

$$l(\mathbf{x}_i) = \ln \frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)}.$$

Osutub, et logit-funktsioon sobib sidumaks argument-tunnuseid lineaarselt. Saame ÜLM erijuhu, mida nimetatakse logit-mudelik ehk logistiliseks regressioonikst:

$$\begin{aligned} Y_i = Y|\mathbf{x}_i &\sim B(1, \pi(\mathbf{x}_i)); \\ l(\mathbf{x}_i) &= \sum_{j=0}^{k} \beta_j x_{ij} = \beta' \mathbf{x}_i, \end{aligned} \tag{33}$$

kus $\beta' = (\beta_0, \ldots, \beta_k)$ on parameetrite vektor. Paneme tähele, et otsene lineaarne seos $\pi(\mathbf{x}_i) = \sum_{j=0}^{k} \beta_j x_{ij}$ ei sobi, sest $\pi(\mathbf{x}_i) \in [0,1]$, aga parem pool ei kuulu tingimata sellesse vahemikku.

Varasemast teame, et kui $\pi(\mathbf{x}_i)$ on "edu" tõenäosus, siis on

$$\Omega(\mathbf{x}_i) = \frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)} - \text{"edu" shansid;} \tag{34}$$

$$l(\mathbf{x}_i) = \ln \frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)} - \text{"edu" log-shansid.} \tag{35}$$

Seose(35) abil saame nii shansid kui tõenäosuse avaldada $l(\mathbf{x}_i)$ kaudu:

$$\Omega(\mathbf{x}_i) = e^{l(\mathbf{x}_i)}, \tag{36}$$

$$\pi(\mathbf{x}_i) = \frac{e^{l(\mathbf{x}_i)}}{1+e^{l(\mathbf{x}_i)}}, \tag{37}$$

kus $l(\mathbf{x}_i)$ on antud seoses (33).

Seosest (33) näeme, et kui $\beta_j = 0$, siis tunnus X_j ei mõjuta "edu" log-shansse $l(\mathbf{x}_i)$. Seega seoste (36) ja (37) põhjal ei mõjuta vastav tunnus ka "edu" shansse ega tõenäosust. Meenutame, et tõenäosuses $P(Y = 1|\mathbf{x}_i)$ tähendab $Y = 1$ "edu".

Kui $\beta_j > 0$, siis on tunnuse X_j ja "edu" log-shansside vahel samasuunaline seos. Hoides teised tunnused fikseeritud, siis X_j kasvades "edu" log-shansid kasvavad; (36) ja (37) tõttu kasvavad ka "edu" shansid ja tõenäosus. Samasuunaline seos on ka tunnuse X_j ja binaarse tunnuse enda vahel, mida suurem X_j, seda sagedamini esineb väärtust "edu".

Kui $\beta_j < 0$, siis on "edu" log-shansside, shansside ja tõenäosuse vahel vastassuunaline seos: X_j kasvades teised suurused vähenevad.

Vaatame abitunnuse X_m muutust ühiku võrra:

$$x'_{im} = x_{im} + 1.$$

Olgu teiste tunnuste väärtused, $x_{ij},\ j \neq m$, fikseeritud. Vastavat vektorit tähistame $\mathbf{x}'_i$. Seosest (33) saame,

$$l(\mathbf{x}'_i) = \sum_{j\neq m}^{k} \beta_j x_{ij} + \beta_m(x_{im}+1) = l(\mathbf{x}_i) + \beta_m. \tag{38}$$

Seega "edu" log-shansid muutuvad β_m võrra. Tulemust (38) kasutades, saame

$$\Omega(\mathbf{x}'_i) = e^{l(\mathbf{x}'_i)} = e^{l(\mathbf{x}_i)}e^{\beta_m},$$

mis ütleb et "edu" shansid kasvavad argument-tunnuse X_m ühikulisel muutusel e^{β_m} korda. Teiste sõnadega e^{β_m} on shansside suhe: shansid punktis $\mathbf{x}'_i$, kus tunnus x_m on teinud ühikulise muutuse, jagatud shanssidega punktis $\mathbf{x}_i$

Tõenäosuse muutumine argument-tunnuse muutudes on keerulisem, see jääb sõltuma argument-tunnuse väärtusest. Kui meid huvitab tõenäosuste erinevus punktides $\mathbf{x}'_i$ ja $\mathbf{x}_i$, siis võime leida lihtsalt vahe

$$\pi(\mathbf{x}'_i) - \pi(\mathbf{x}_i),$$

mida saab teha, kui mudel on hinnatud. Teiseltpoolt teame, et funktsiooni muutumise kiirust iseloomustab tema tuletis argumendi järgi. Tuletis on puutuja tõus, mis on erinev erinevate argumendiväärtuste korral. Tuletis annab ühikulisel argumendi muutusel funktsiooni muudu, kui selles piirkonas käitub funktsioon lineaarselt.

Uurime tõenäosuse tuletist argumendi x_{im} järgi. Diferentseerides avaldist

$$\pi(\mathbf{x}_i) = \frac{e^{l(\mathbf{x}_i)}}{1+e^{l(\mathbf{x}_i)}},$$

kasutame jagatise tuletise ja liitfunktsiooni tuletise reegleid. Saame,

$$\frac{\partial \pi(\mathbf{x}_i)}{\partial x_{im}} = \frac{[1+e^{l(\mathbf{x}_i)}]e^{l(\mathbf{x}_i)}\frac{\partial l(\mathbf{x}_i)}{\partial x_{im}} - e^{l(\mathbf{x}_i)}e^{l(\mathbf{x}_i)}\frac{\partial l(\mathbf{x}_i)}{\partial x_{im}}}{[1+e^{l(\mathbf{x}_i)}]^2}.$$

Seose (33) tõttu $\frac{\partial l(\mathbf{x}_i)}{\partial x_{im}} = \beta_m$, mida kasutades saame,

$$\frac{\partial \pi(\mathbf{x}_i)}{\partial x_{im}} = \beta_m \frac{e^{l(\mathbf{x}_i)}}{[1+e^{l(\mathbf{x}_i)}]^2}.$$

Näeme, et tuletis, st tõenäosuse muutumise kiirus, sõltub seletavate tunnuste vektorist $\mathbf{x}_i$ ja kordajast β_m.

Kasutades tõenäosuse avaldist ja lisaks avaldist

$$1 - \pi(\mathbf{x}_i) = \frac{1}{1+e^{l(\mathbf{x}_i)}},$$

saame alternatiivse seose tuletisele,

$$\frac{\partial \pi(\mathbf{x}_i)}{\partial x_{im}} = \beta_m \pi(\mathbf{x}_i)(1-\pi(\mathbf{x}_i)).$$

Siit näeme, et tõenäosus muutub kõige kiiremini sellises punktis $\mathbf{x}_i$, kus $\pi(\mathbf{x}_i) = 1/2$. Kui selle punkti ühikulises ümbruses on funktsioon ligikaudu lineaarne, saame öelda, et tunnuse X_m ühikuline muutus selles punktis toob kaasa "edu" tõenäosuse muutuse $\beta_m/4$ võrra.

5.2 Parameetrite interpreteerimine kvalitatiivse seletava tunnuse korral

Lihtsuse mõttes olgu 1 kvalitatiivne seletav tunnus x, mis on I tasemega. Kvalitatiivset seletavat tunnust nimetatakse faktoriks. Edu tõenäosuseks on $\pi_i = P(Y = 1|x = i)$. Kuna x pole arvuline, siis ei saa teda ennast mudelisse panna, pannakse tema mõju. Seega logit-mudel saab kuju:

$$\ln \frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_i, \ i = 1, \ldots, I. \tag{39}$$

Siin on I võrrandit, aga $i+1$ parameetrit vaja leida. Et saaks üheselt leida on vaja lisada üks kitsendus. Klassikaliselt kasutatakse ühte kahest kitsendusest:

- $\sum_{i=1}^{I} \beta_i = 0$. Siis on β_0 edu log-shansside üldkeskmine, β_i kas suurendab või vähendab seda;
- $\beta_I = 0$. Siis on vabaliige β_0 edu log-shansid faktori x tasemel I, β_i on muutus I-nda taseme suhtes. Näita!

Teise punkti puhul on seega e^{β_i} shansside suhe. Näitab, mitu korda on edu shansid $x = i$ korral suuremad shanssides $x = I$ korral. Näita.
Kui $\beta_i = 0, \ \forall i$, siis kvalitatiivsel tunnusel ei ole mõju uuritavale tunnusel Y, teisisõnu Y ei sõltu tunnusest x.

5.3 Logistiline regressioon retrospektiivse uuringu korral

Juht-kontrolluuringus on juhud ja kontrollid võetud uurija poolt, mistõttu uuritav tunnus juht/mitte ei ole juhuslik ja tekib õigustatud küsimus tema modelleerimiseks seletavate kaudu. Osutub, et logistilist regressiooni võib ka selliste andmete puhul teha. Kordajad, mis iseloomustavad X-tunnuste mõju on õiged, ainult vabaliige ei ole õige ja vajab korrigeerimist järgmiselt (vt ka Agresti 2013, lk 168):

$$\beta_0* = \beta_0 + \ln\frac{\pi}{1-\pi} - \ln\frac{\tilde{\pi}}{1-\tilde{\pi}},$$

kus π on juhtude osakaal üldkogumis ja $\tilde{\pi}$ on juhtude osakaal uuringus.

5.4 Mudeli parameetrite hindamine logistilises regressioonis

Siin ei ole võimalik analüütilisel kujul hinnangute valemeid saada. Küll aga saab hinnangud leida numbriliselt. Kas maksimiseerides tõepärafunktsiooni (Fisheri skoorimeetod) või kasutades kaalutud vähimruutude meetodid (viimane on vaikimisi). Mõlemad meetodid on iteratiivsed ja teostavad arvutusi seni, kuni kaks järjestikust lahendit on teineteisele piisavalt lähedal.

Tähistame vaatluste vektori $\mathbf{y} = (y_1, \ldots, y_n)$ ja tõenäosuste vektori $\pi = (\pi_1, \ldots, \pi_n)$, kus $\pi_i = \pi(\mathbf{x}_i)$, siis tõepärafunktsiooniks on

$$L(\mathbf{y}, \pi) = \prod_{i=1}^{n} f(y_i, \pi(\mathbf{x}_i)) = \prod_{i=1}^{n} \pi(\mathbf{x}_i)^{y_i}(1-\pi(\mathbf{x}_i))^{1-y_i}. \tag{40}$$

Kuna $\pi(\mathbf{x}_i)$ sõltub parameetritest $\beta = (\beta_0, \ldots, \beta_k)$ (vt (37) ja (33)), siis sõltub ka tõepärafunktsioon nendest parameetritest ja on maksimiseeritav β suhtes. Maksimumpunkt $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_k)$ ongi tundmatu parameetervektori β suurima tõepära hinnanguks.

5.5 Mudeli sobivuse kontroll

Olgu mudeli parameetrid hinnatud, st on saadud $\hat{\beta}$. Sellega on ka tõenäosused π hinnantud:

$$\hat{\pi}_i = \hat{\pi}(\mathbf{x}_i) = \frac{e^{\hat{\beta}'\mathbf{x}_i}}{1+e^{\hat{\beta}'\mathbf{x}_i}}. \tag{41}$$

Nüüd on saadavad mudeli poolt prognoositud väärtused, need on tinglikud keskväärtused, mis Bernoulli jaotuse korral langevad kokku tõenäosustega:

$$\hat{y}_i = \hat{E}(Y|\mathbf{x}_i) = \hat{\pi}(\mathbf{x}_i).$$

Mudel sobib andmetega hästi kui y_i ja $\hat{y}_i$ on lähedased. Mudel sobib täpselt, kui $y_i = \hat{y}_i$, $\forall i$. Täielikku sobimist on alati võimalik saavutada parameetrite arvu suurendamisega. Kui nõuda, et tinglik keskväärtus on täpselt võrdne vastava vaatlusega,

$$\hat{\pi}(\mathbf{x}_i) = y_i,$$

saame täpselt sobiva mudeli. Sellise mudeli parameetriteks on kõik vaatlused, teda nimetatakse küllastunud mudeliks. Küllastunud mudel sobib küll hästi antud andmestikuga, kuid ta ei ole kasulik. Ta ei üldista andmetes peituvat sõnumit, ei näita üldisi suundumusi,

ja teise andmestiku korral ta ei sobi. Küllastunud mudelil on oma koht teiste mudelite võrdlemisel.

Kui vaadelda kõiki võimalikke mudeleid antud andmestikule $\mathbf{y} = (y_1, \ldots, y_n)$, siis tõepärafunktsioon saavutab maksimumi just küllastunud mudeli korral (maksimaalne sobivus andmetega).

Ülesanne 5.1. Leia tõepärafunktsiooni (40) väärtus küllastunud mudeli korral, st $L(\mathbf{y}, \mathbf{y})$.

Üldistatud lineaarsete mudelite korral kasutatakse mudeli sobivuse kontrolliks hälbimuse (deviance) mõistet. Hälbimus mõõdab erinevust meie mudeli log-tõepära ja küllastunud mudeli log-tõepära vahel.

Tõepärafunktsiooni väärtus meie hinnatud mudeli korral on $L(\mathbf{y}, \hat{\pi})$, kus $\hat{\pi}$ elemendid on avaldises (41). Kehtib $L(\mathbf{y}, \hat{\pi}) \leq L(\mathbf{y}, \mathbf{y})$.

Definitsioon. Hälbimuseks nimetatakse suurust

$$D(\hat{\pi}, \mathbf{y}) = -2(\ln L(\hat{\pi}, \mathbf{y}) - \ln L(\mathbf{y}, \mathbf{y})). \tag{42}$$

Kui n on suur, siis eeldusel, et sobib meie mudel, on $D(\hat{\pi}, \mathbf{y}) \sim \chi^2(df)$, kus $df = n-(k+1)$. Kui $D(\hat{\pi}, \mathbf{y})$ on suur, siis meie mudel ei sobi. Sobib siis, kui ta on võrreldav oma vabadusastmete arvuga. Meie mudeli eeliseks on parametriseeritud kuju ja märksa väiksem parameetrite arv kui küllastunud mudelil. Tähtis omadus on ka see, et vastandina küllastunud mudelile, sobib meie mudel kirjeldama ka uusi antud probleemiga seonduvaid andmeid.

Hälbimust kasutatakse ka 2 mudeli võrdlemisel. Olgu mudel M_0 hinnatud tõenäosustega $\hat{\pi}_0$ ja mudel M_1 hinnatud tõenäosustega $\hat{\pi}_1$. Olgu M_0 lihtsam mudel, mis sisaldub mudelis M_1 (osa parameetreid β_j on nullid). Siis kehtib,

$$L(\hat{\pi}_0, \mathbf{y}) \leq L(\hat{\pi}_1, \mathbf{y}),$$

sest maksimum üle väiksema hulga parameetrite ei saa olla suurem kui maksimum üle kõigi parameetrite. Kummagi mudeli hälbimuste jaoks saame valemist (42),

$$D(\hat{\pi}_1, \mathbf{y}) \leq D(\hat{\pi}_0, \mathbf{y}),$$

mis ütleb, et lihtsamatel mudelitel on suuremad hälbimused. Ometi, kui erinevus ei ole suur, võiks kasutada lihtsamat mudelit. Otsustamine taandatakse jaotusele, suure n korral on eeldusel, et kehtib lihtsam mudel,

$$D(\hat{\pi}_0, \mathbf{y}) - D(\hat{\pi}_1, \mathbf{y}) \sim \chi(df),$$

kus df on kahe mudeli parameetrite arvu vahe. Kumbki hälbimus on samuti hii-ruut jaotusega, vt (42).

Tarkvara annab sageli välja 0-mudeli hälbimuse (selles kasutatakse ainult vabaliiget, st on ainult 1 parameeter) ja meie sobitatava mudeli hälbimuse. Soovime, et hälbimuses on toimunud vähenemine.

Näide Osa 1 seletava tunnusega logistilise regressiooni väljundist.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.3508     2.6287  -4.698 2.62e-06 ***
      Width   0.4972     0.1017   4.887 1.02e-06 ***
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " "  1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 194.45  on 171  degrees of
```

Pärast mudeli kui terviku sobivuse kontrolli, tuleb uurida ka iga seletava tunnuse olulisust mudelis. Kui hinnangud $\hat{\beta}_j$ on saadud, tuleb selgitada, kas nad on oluliselt erinevad nullist. Tarkvara annab välja Wald'i teststatistiku, $z_j = \hat{\beta}_j / \sqrt{\hat{D}(\hat{\beta}_j)}$, mis on asümptootiliselt $N(0,1)$ eeldusel $H_0 : \ \beta_j = 0$.

Mudeli sobivuse illustratsioonideks on graafikud. Neis võrreldakse mudel-prognoose andmetega või andmete pealt leitud näitajatega.

Ülesanne 5.2. Olgu tegu tabuleeritud andmetega, kus abitunnuste vektoril $\mathbf{x}_i$ on I taset (komponentide tasemete kombinatsioonid). Olgu n_{i1} edude arv tasemel i ja n_{i+} kõigi objektide arv tasemel i. Siis $n_{i1} \sim B(n_{i+}, \pi_i(\mathbf{x}_i))$. Pane kirja tõepärafunktsioon (40) tabuleeritud andmete korral. Mis on uuritava tunnuse andmestik antud juhul? Mis on prognoos ehk ooteväärtused? Missugune oleks küllastunud mudel? Mis on tõepärafunktsiooni väärtus küllastunud mudeli korral?

5.6 Teisi linkfunktsioone

Binaarse uuritava tunnuse korral modelleerime edu tõenäosust $\pi(\mathbf{x})$ sõltuvalt argumenttunnuste vektorist $\mathbf{x}$. Üldistatud lineaarne mudel nõuab, et selle tõenäosuse mingi funktsioon oleks lineaarselt seotud vektoriga $\mathbf{x}$:

$$g(\pi(\mathbf{x})) = \beta'\mathbf{x} = \beta_0 + \beta_1 x_1 + \ldots \beta_k x_k.$$

Vaatasime juhtu, kus $g()$ oli logit-funktsioon. On aga teisigi linkfunktsioone. Enamlevinud on probit ja täiend-log-log link. Probit-link kasutab standardnormaaljaotuse jaotusfunktsiooni:

$$\Phi^{-1}(\pi(\mathbf{x})) = \beta'\mathbf{x}.$$

Siis tõenäosuse jaoks saame

$$\pi(\mathbf{x}) = \Phi(\beta'\mathbf{x}).$$

Täiend-log-log (complementary-log-log) lingi korral,

$$\ln[-\ln(1-\pi(\mathbf{x}))] = \beta'\mathbf{x}.$$

Nimi on selline, sest log-log rakendub täientõenäosusele. Antud juhul avaldub tõenäosus:

$$\pi(\mathbf{x}) = 1 - \exp[-\exp(\beta'\mathbf{x})].$$

Parameetrite tähendus on sama, mis varem: $\beta_j = 0$, siis x_j ei mõjuta edu tõenäosust; $\beta_j > 0$ või $\beta_j < 0$, siis x_j kasvades edu tõenäosus vastavalt kasvab või kahaneb. Argumenttunnuse ühikuline muutus ei ole nii hästi interpreteeritav, kui logit mudeli korral.

Nii logit-, kui probit-mudelid on sümmeetrilised tõenäosuse $\pi(\mathbf{x}) = 1/2$ suhtes ja tõenäosuse graafikud lähenevad ühele ja nullile sama kiirusega. Probit-mudel võimaldab tõenäosuse kiiremat muutumist $\pi(\mathbf{x}) = 1/2$ ümbruses. Täiend-log-log lingiga mudel võimaldab modelleerida andmeid, kus edu tõenäosus muutub aeglaselt $\pi(\mathbf{x}) = 0$ ümbruses ja kiiresti $\pi(\mathbf{x}) = 1$ ümbruses.

Mudeli parameetrite hindamine toimub numbriliste meetoditega. Hinnatud parameetreid kasutades saab leida prognoosväärtused. Mudeli sovivuse kontroll toimub hälbimuse abil samadel alustel nagu logit-mudeli korral.

5.7 Prognoosivõime kirjeldamine

Binaarne uuritav tunnus on klassi tähis (1-haige, 0-terve). Olles hinnanud logistilise regressiooni mudeli olemasolevatelt andmetelt, saame selle abil prognoosida uue objekti jaoks klassikuuluvuse. Tuletame meelde, et

$$\pi(x_i) = P(Y = 1|x_i),$$

mistõttu on loomulik klassifitseerida objekt klassi "Y=1", kui $\pi(x_i)$ on suur. Otsustamiseks on vaja anda lävi π_0. Nii, et objekt, mille tunnusvektor on x saab prognoosi

$$\hat{y} = 1, \text{ kui } \pi(x) > \pi_0 \text{ ja } \hat{y} = 0, \text{ kui } \pi(x) \leq \pi_0.$$

Tehes nüüd $(y_i, \hat{y}_i)$ sagedustabeli, saame valesti klassifitseerimiste osakaalu. Jääb π_0 valik, et võimalikult vähe vigu teha. Võimalusteks on

- $\pi_0 = 0.5$,
- "1" - de osakaal valimis,
- ristvalideerimine "1-välja"meetodil.

Erinevate mudelite prognoosivõime võrdlemiseks lastakse lävel π_0 muutuda. Graafikule kantakse iga π_0 korral "valepositiivsuse määr"ja "õigepositiivsuse määr". Saadakse nn ROC-kõver "Receiver Operating Characteristic".

Näide. Krabide klassifitseeritud andmestiku korral hindasime tõenäosuse omada satelliite järgmiselt:

$$\pi(laius_i) = \frac{\exp(-12.008 + 0.484 * laius)}{1 + \exp(-12.008 + 0.484 * laius)}.$$

Sagedustabelites all on toodud tegelikkus koos prognoosidega erinevate lävede korral.

```
      Progn, lävi=0.2         Progn, lävi=0.4         Progn, lävi=0.5
Tegelik onsat  ei        Tegelik onsat  ei        Tegelik onsat ei
  onsat   111   0          onsat   104   7          onsat    95 16
  ei       61   1          ei       46  16          ei       35 27
```

Ülesanne 5.3. 1. Leia klassifitseerimisviga (valesti klassifitseerimise osakaal) iga tabeli korral.
2. Leia tundlikkus, spetsiifilisus, valepositiivsuse määr, valenegatiivsuse määr iga tabeli korral.
3. Pane antud tabelitest ROC-kõvera graafikule 3 punkti (x-teljel "valepositiivsuse määr", y-teljel "õigepositiivsuse määr").

Tahame, et klassifitseerimismeetod oleks selline, et kõvera-alune pindala AUC oleks ≈ 1. Kui klassifitseerimine toimub täiesti juhuslikult, siis on ROC-kõveraks nurgapoolitaja ja vastav pindala 0.5.

5.8 Poissoni log-lineaarne mudel

Kui uuritava tunnuse väärtusteks on arvukused (sagedused) ja soovime uurida nende sõltuvust seletavatest tunnustest, sobib Poissoni log-lineaarne mudel ehk Poissoni regressioon.

Näide krabid. Uuritav "satelliitide arv emakrabil", seletavateks "värvus", "kaal", "kilbi laius" jt.

Arvukus on juhuslik suurus, mille jaotuseks sobib Poissoni jaotus. Poissoni jaotus kuulub eksponentsiaalsesse jaotuste perre:

$$Y \sim f(y,\mu) = \frac{e^{-\mu}\mu^y}{y!} = e^{-\mu}\frac{1}{y!}\exp[y(\ln\mu)],\ y = 0,1,\ldots. \tag{43}$$

Jaotuse keskväärtuseks on $EY = \mu$. Keskväärtuse funktsioon, mis seotakse lineaarselt seletavate tunnustega on $\ln\mu$:

$$\ln\mu(\mathbf{x}) = \sum_{j=0}^{k}\beta_j x_j = \beta'\mathbf{x}. \tag{44}$$

Seosed (43) ja (44) annavad üldistatud lineaarse mudeli.

Vaatame parameetrite interpreteerimist. Keskväärtuse jaoks saame avaldise:

$$\mu(\mathbf{x}) = \exp(\beta'\mathbf{x}) = \exp(\beta_0 + \beta_1 x_1 + \ldots \beta_k x_k) = e^{\beta_0}(e^{\beta_1})^{x_1}\cdots(e^{\beta_k})^{x_k}. \tag{45}$$

Valemist (45) näeme, et sõltuvalt β_j märgist keskväärtus kas kasvab või kahaneb tunnuse x_j kasvades, $\beta_j = 0$ korral tunnusel x_j ei ole mõju. Ühikuline muutus, $x_j' = x_j + 1$, toob kaasa keskväärtuse $\mu(\mathbf{x})$ e^{β_j}-kordse muutuse:

$$\mu(\mathbf{x}') = e^{\beta_j}\mu(\mathbf{x}).$$

Hea on interpreteerida suhtelist muutust

$$\frac{\mu(\mathbf{x}') - \mu(\mathbf{x})}{\mu(\mathbf{x})} = e^{\beta_j} - 1.$$

Väärtuse $(e^{\beta_j} - 1)100\%$ abil, näeme sõltuvalt märgist, mitme protsendi võrra keskmine kasvab või kahaneb.

Mudeli sobivuse kontroll on analoogiline logit-mudelile. Nüüd on prognoosväärtuseks $E(Y|\mathbf{x}) = \mu(\mathbf{x})$. Kui mudel on hinnatud, st $\hat{\beta}_j$ leitud, on prognoosväärtused leitavad valemist (45). Hälbimus küllastunud mudeli suhtes ja kahe mudeli hälbimuste vahe on kasutatavad mudeli sobivuse üle otsustamisel. Parameetrite hinnangud annnab tarkvara koos usalduspiiridega.

Ülesanne 5.4. Pane kirja Poissoni log-lineaarse mudeli tõepärafunktsioon, kui andmed on $(y_i, \mathbf{x}_i')$, $i = 1,\ldots,n$.

5.9 Otsustuspuud

Kasutame materjali James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An Introduction to Statistical Learning. Wiley. Sealt on ka järgnevad joonised.

Ka siin on meil uuritav tunnus y ja seletavate tunnuste vektor $\mathbf{x}$. Eesmärgiks on $\mathbf{x}$ abil prognoosida y väärtust. Jaotuseeldusi y jaoks ei tehta.

Meetodi olemus seisneb sellles, et teatud kriteeriumeid kasutades jagatakse objektid erinevatesse lihtsatesse piirkondadesse $\mathbf{x}$ väärtuste järgi. Jagamist saab esitada puu kujul. Lõplikke piirkondi R_j, $j = 1, 2, \ldots$ nimetatakse lehtedeks. Igas lehes prognoositakse y-väärtus järgmiselt:

- arvulise y korral lehe keskmisena $\hat{y}_{R_j} = \sum_{i \in R_j} y_i / |R_j|$, kus $|R_j|$ on objektide arv piirkonnas R_j.
- kvalitatiivse y korral lehe sagedaima väärtusega piirkonnas R_j.

Otsustuspuid, kus uuritav tunnus on arvuline nimetatakse regressioonipuudeks, kui aga kvalitatiivne, siis klassifitseerimispuudeks.

5.9.1 Regressioonipuud

Näide. Olgu tegu lihtsa andmestikuga (väljavõte R-paketi ISLR pesapallimängijate andmestikust Hitters):

Palk	Aastad	Tabamused
475	14	81
750	11	169
.....		

Siin y =Palk (aastapalk tuhandetes dollarites), $\mathbf{x}$=(Aastad, tabamused), (mängustaaž, eelmise aasta tabamused). Soovime prognoosida palka sõltuvalt aastatest ja tabamustest.

Kasutatud on teisendatud suurust log(Palk), et saada lahti rasketest sabadest (et uuritava tunnuse jaotus sarnaneks rohkem normaaljaotusele).

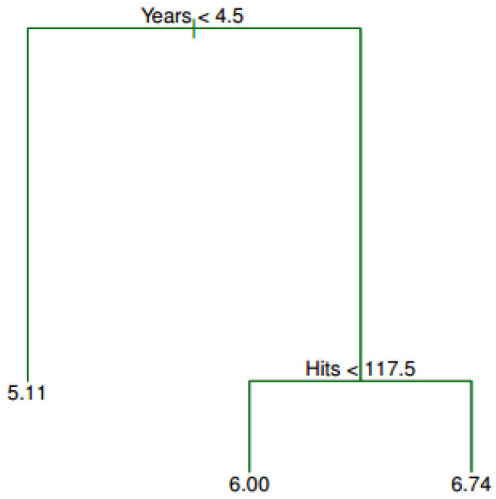
Joonisel 1 on andmetelt leitud otsustuspuu ja tulemus ütleb, et kui mängitud on vähem kui 4.5 aastat, siis keskmine log(Palk)=5.107 ehk keskmine palk on $e^{5.107} = 165.7$ tuhat dollarit.

Piirkonnad seletavate tunnuste ruumis on R_1, R_2, R_3 (Joonis 2), kus $R_2 = \{\mathbf{x} | Aastad \geq 4.5, Tabamused \leq 117.5\}$. Pane kirja R_1 ja R_2.

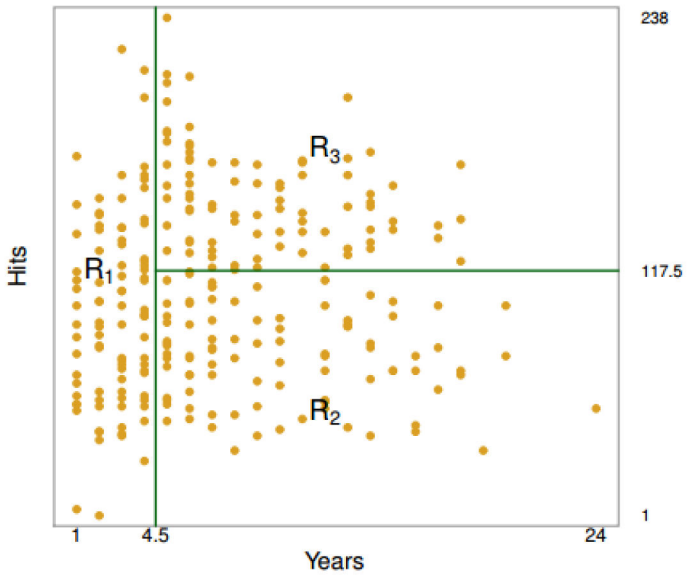
Ülesanne 5.5. Pane kirja R_1, R_3 ja prognoositud Palk piirkondades. Interpreteeri.

Antud puul on 3 lõpetavat tippu (lehte) ja 2 sisemist tippu (viimastes toimub jagunemine). Jooned, mis tippe ühendavad on oksad/harud.
Otsuspuud intuitiivselt hästi interpreteeritavad ja neil on hea graafiline esitus.

Joonis 1: Otsustuspuu.



Joonis 2: Vastavad piirkonnad seletavate tunnuste ruumis.



5.9.2 Puu konstrueerimine

Olgu $\mathbf{x} = (x_1, \ldots, x_p)$ seletavate tunnuste vektor. Selle võimalike väärtuste ruumi nimetame $\mathbf{x}$-ruumiks. Sammud

1. $\mathbf{x}$-ruum jagatakse mittelõikuvateks ja ammendavateks piirkondadeks $R_1, \ldots, R_J$.
2. Objektid, mis langevad piirkonda R_j saavad sama prognoosi $\hat{y}_j$, milleks on uuritava tunnuse keskmine selles piirkonnas.

Andmestikku, mille põhjal puu ehitatakse nimetatakse treeningandmestikuks.

Kui tuleb uus objekt väljaspool treeningandmestikku oma $\mathbf{x}$-vektoriga, $\mathbf{x} = \mathbf{x}_0$, kuid mille y-tunnus ei ole teada siis vastavaks prognoosiks saab $\hat{y}_j$, kui $\mathbf{x}_0 \in R_j$.

Piirkondadeks jagamise algoritm. Leida ristkülikud (lihtsuse mõttes) $R_1, \ldots, R_J$ nii, et et jääkide ruutude summa RSS oleks minimaalne:

$$RSS = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2. \tag{46}$$

Tehakse nn rekursiivset binaarset tükeldamist. Tükeldamine on ülalt alla, st algselt on kõik objektid ühes piirkonnas. Iga tükeldus tekitab 2 uut oksa allapoole. Tükelduse tegemiseks valitakse tunnus x_j ja lõikepunkt s nii, et $\mathbf{x}$-ruumi jagamine kaheks piirkonnaks $R_1(j,s) = \{\mathbf{x} | x_j < s\}$ ja $R_2(j,s)\{\mathbf{x} | x_j \geq s\}$ annab RSS suurima vähenemise.

Seega vaadatakse läbi kõik tunnused ja kõikvõimalikud lõikepunktid ning leitakse selline j ja s, mis minimiseerib

$$\sum_{i:x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2. \tag{47}$$

Seejärel korratakse protsessi järgmise parima tükelduse jaoks valides ühe saadud piirkondadest. Tulemuseks on 3 piirkonda. Edasiseks tükelduseks valitakse jälle 1. Nii kuni lõpetamise tingimuseni, näiteks et piirkonnas on alla 5 objekti.

Pügamine. Algoritmi puuduseks on ülesobitatud puu. Liiga keeruline puu. Pole vaja saada ülihead andmetega sobivust treeningandmete korral, pigem on vaja tagada head prognoosid ka testandmete korral. Lihtsam puu on parem.

Parem tulemus testandmete jaoks saadakse, kui kasvatatakse väga suur puu T_0, mis seejärel pügatakse tagasi, nii et saadakse alampuu T. Pügamisel minimiseeritakse avaldis (48):

$$\sum_{m=1}^{|T|} \sum_{i:x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha |T|, \tag{48}$$

üle vaadeldavate alampuude T. Kui $\alpha = 0$, on lahendiks T_0, kui $\alpha > 0$, siis on lahendiks mingi alampuu. Mida suurem α, seda lihtasam alampuu. Sobiv α on selline, millele vastav puu prognoosib testandmeid kõige paremini. Leitakse nn ristvalideerimisega.

5.9.3 Klassifitseerimispuu

Mõisted ja konstruktsioon on sarnased eelpookirjeldatud regressioonipuule. Erinevus seisneb uuritavas tunnuses, mis siin on kvalitatiivne ja prognoosis.

Kui $\mathbf{x}$-ruum on puu kasvatamisega piirkondadeks jagatud, siis y-tunnuse prognoosiks piirkonnas R_m on kõige sagedamini esinev y-väärtus selles piirkonnas. Kui selleks on väärtus k, siis $\hat{y}_{R_m} = k$. Tuleb meeles pidada, et puu ehitatakse treeningandmete pealt ja seal on y-väärtused piirkonnas R_m olemas ning prognoos on leitav. Hiljem saab seda prognoosi aga kasutada uue objekti jaoks, millel on teada vaid $\mathbf{x}$-vektor, kuid mitte y-väärtus.

Lisaks prognoosile huvitab meid sageli ka tõenäosuse hinnang ehk väärtuse k osakaal piirkonnas R_m:

$$\hat{p}_{mk} = \frac{\#\{y = k | \mathbf{x} \in R_m\}}{|R_m|}.$$

Puu kasvatamisel kasutatakse endiselt rekursiivset binaarset tükeldamist, aga suuruse RSS minimiseerimine ei ole võimalik, sest y_i on kvalitatiivne, vt (46). Võimalik on minimiseerida

- klassifitseerimisviga, mis piirkonna siseselt on

$$E = 1 - \max_k \hat{p}_{mk}, \tag{49}$$

- Gini indeksit, mis mõõdab koguvarieeruvust üle K klassi

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}), \tag{50}$$

- summarset entroopiat (cross-entropy), mis sarnaneb Gini indeksile

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}. \tag{51}$$

Puu kasvatamisel ei ole klassifitseerimisviga piisavalt tundlik näitaja ja kasutatakse kahte ülejäänud näitajat. Valemist (50) näeme, et G on väike, kui suurused $\hat{p}_{mk}$ on piirkonnas m lähedal 0-le või 1-le. Seega on G piirkonna puhtuse (purity) näitaja – kõik y-väärtused on ühetaolised. Puu ehitamine G minimiseerimise kaudu püüdleb ühetaoliste ehk puhaste lehtede poole. Sama roll on ka suurusel D.

Puu pügamisel võib kasutada kõiki kolme näitajat, aga klassifitseerimisviga E on eelistatud näitaja, kui lõpliku pügatud puu ennustustäpsus on oluline.

Näide. Südamehaigus. Andmestikus on 303 patsienti. Tegu on kvalitatiivse uuritava tunnusega südamehaigus (HD: jah/ei). On 13 seletavat tunnust.

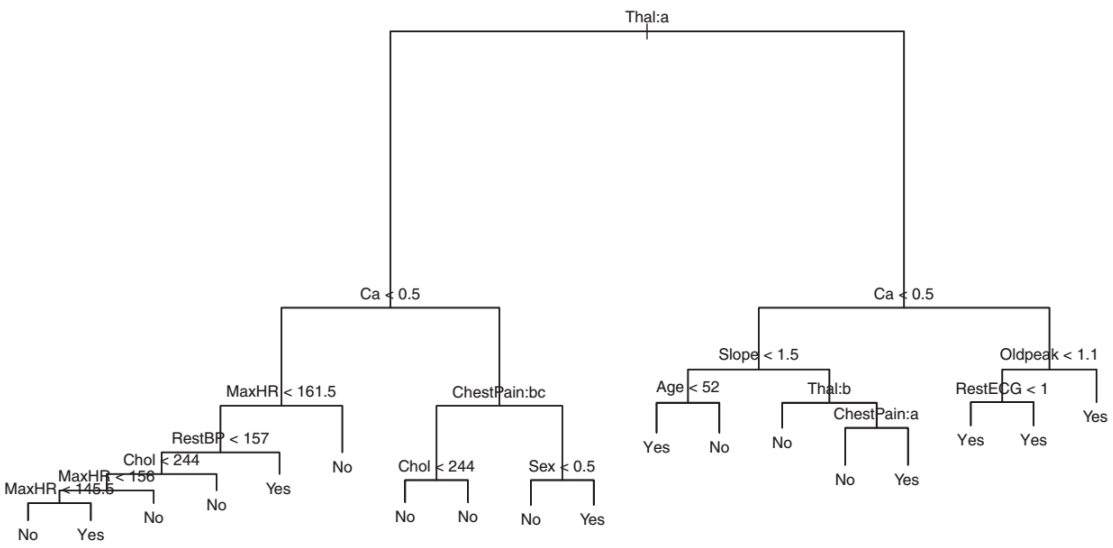
Age	Sex	ChPain	RestBP	Chol	Fbs	RestECG	MaxHR	ExAng	Oldpeak	Slope	Ca	Thal	HD
63	1	typical	145	233	1	2	150	0	2.3	3	0	fixed	No
67	1	asympt	160	286	0	2	108	1	1.5	2	3	normal	Yes
37	1	nonanginal	130	250	0	0	187	0	3.5	3	0	normal	No
.....													

Kvalitatiivse tunnuse järgi tükeldamisel võetakse ühte rühma kõik teatud y-väärtusega objektid ja teise ülejäänud. Vaata ja interpreteeri jooniseid.

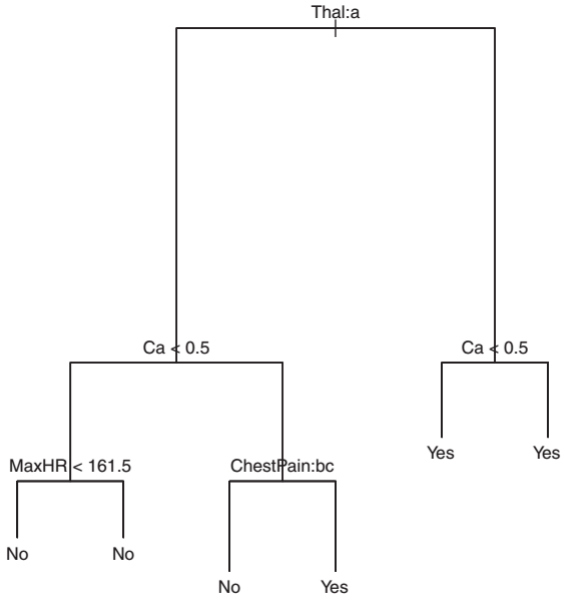
Miks on osades jagunemispunktides tulemuseks sama prognoos? Näiteks Joonisel 3 on jagunemiskohas $RestECG < 1$ mõlema lehe tulemuseks "jah". Miks siis selline jagunemine üldse tehti?

Jagunemine toimus, sest see tagas parema lehe "puhtuse" ehk parempoolsemas lehes on kõik 9 vaatluse y tunnnuse väärtuseks "jah". Vasakpoolses lehes on 7/11st väärtusega "jah".

Joonis 3: Pügamata puu.



Joonis 4: Pügatud puu vastavalt minimaalsele ristvalideerimisveale.



Milleks see vajalik võib olla?

Näiteks kui meil on uus vaatlus, siis parempoolsesse lehte langedes saame kindlad olla, et prognoosi väärtuseks on ”jah”. Vasakpoolsemas lehes on prognoosiks tõenäoliselt ”jah”. Isegi kui jagunemiskoht $RestECG < 1$ ei vähenda klassifitseerimisviga, siis paraneb Gini indeks ja entroopia, mis on lehe puhtuse osas tundlikumad.

5.10 Otsustuspuude eelised ja puudused

Nii regressiooni- kui klassifitseerimispuudel on klassikaliste mudelite ees mitmeid eeliseid

- Nad on lihtsasti selgitatavad ja inimestele rakendustest arusaadavad;
- Öeldakse, et nad peegeldavad inimese mõtlemist otsustamisel;
- Puud saab kujutada graafiliselt (ka väga paljude seletavate tunnuste korral);
- Puu meetodid saavad kergesti hakkama kvalitatiivsete seletatavate tunnustega (ei ole vaja luua indikaatortunnuseid);
- Mittelineaarse seose korral uuritava tunnuse ja seletavate vahel saab puude meetod klassifitseerimisega paremini hakkama, kui lineaarsed meetodid (vt jooniseid).

Puuduseks on see, et prognoositäpsuse osas on otsustuspuudest paremaid meetodeid. Täpsuse tõstmiseks on otsustuspuude meetodeid edasi arendatud. Appi tuleb "bagging", "boosting", "juhuslik mets".

Joonis 5: Puud vs lineaarne mudel