TARTU ÜLIKOOL

Matemaatika-informaatikateaduskond

Matemaatilise statistika instituut

OLGA GORJAJEVA

SISESTUSVIGADE LEIDMINE STATISTILISTE MEETODITEGA

MAGISTRITÖÖ

Juhendaja: Märt Möls

Tartu 2005

Sisukord

Sissejuhatus

Kaasaegses maailmas tehakse suurem osa tööst andmetega arvutite abil. Vaatamata sellele, et andmeid hoitakse arvutis, jääb sisestajaks enamasti ikkagi inimene või inimeste rühm. Inimene võib aga kergesti eksida ning seetõttu võib suurematesse andmebaasidesse sattuda andmete kogumisel või sisestamisel tekkinud vigu.

Suuri andmebaase kasutavad analüüside ja aruannete tegemiseks paljud inimesed. Nende oskused ja võimed avastada ning vältida andmebaasis esinevaid vigu võib olla erinev. Mitte kõik ei pruugi tunda statistikas kasutatavaid sobivaid meetodeid (näiteks erindite analüüsi) ja samas ei tarvitse üksikkasutajal olla õigust kasutada kõiki andmestikus sisalduvaid tunnuseid, mis aitaksid vigu avastada. Seega oleks soovitav, et andmebaasi looja või haldaja ise tagaks andmebaasis sisalduvate andmete kõrge kvaliteedi.

Andmebaasi haldaja aga ei pruugi teada, milliseid analüüse või aruandeid soovitakse olemasolevate andmete pealt teha, seepärast peab olema kindel, et kasutaja saab võimalikult täpse vastuse igasuguse analüüsi korral. Paljud statistikas tuntud meetodid sisestamisvigade avastamiseks on aga eelkõige kasutatavad mingi konkreetse analüüsi korral (näiteks ühe konkreetse regressioonmudeli jääkide analüüs).

Andmete kvaliteedi kontrollimine koosneb mitmest etapist. Kõige esimene ja olulisem nendest on andmete "puhastamine". Sõna "puhastamine" all mõistetakse selles töös valede väärtuste ehk vigade leidmist andmebaasis ja nende parandamist või eemaldamist. Lisaks andmete puhastamisele peaks toimuma ka andmete täielikkuse kontroll. Võimalusi andmete puhastamiseks vaadeldakse käesoleva töö teises peatükis.

Tuleb silmas pidada, et ükski meetod ei tööta ideaalselt ja mõned valed väärtused võivad ka hoolikaimal kontrollimisel andmestikku jääda. Sestap pakub huvi ka vigastest algandmetest tuleneva ebatäpsuse kirjeldamine.

Käesolevas magistritöös on välja pakutud meetodeid, mis võimaldavad sisestusvigu sisaldavate andmete põhjal hinnata uuritava tunnuse tegelikku jaotust.

Töö kolmandas osas on teises peatükis esitatud meetodeid katsetatud tegelike andmete

peal. Erinevate meetodite võimet leida üles sisestusvigu on katsetatud TÜ Tervishoiu-instituudis sisestatud ankeetküsitluse andmetel.

1. Erindi ja vea mõiste

1.1 Erind

Kirjanduses võib leida erinevaid erindi definitsioone. Toome ära ja kasutame ühte neist:

Definitsioon 1. *Erindiks antud andmestikus nimetame objekti (või objektide hulka), mis paistab olevat vastuolus selle andmestiku teiste vaatlustega* [1].

Definitsiooni fraas '*paistab olevat vastuolus'* on siin otsustav. See on uurija subjektiivne otsus.

Erindiks võib olla tunnuse tegelik, õigesti mõõdetud ja korrektselt sisestatud väärtus kui ka vigaselt sisestatud väärtus. Traditsiooniliselt pööratakse erinditele kõrgendatud tähelepanu kahel põhjusel:

a) praktikas on sisestusvigade protsent erindite seas suurem kui andmestikus tervikuna, seega võimaldab erindite kontroll efektiivselt avastada tegelikke vigu;

b) vigane vaatlustulemus, mis ühtlasi on erind, võib andmestiku põhjal arvutatud statistikute väärtused täielikult rikkuda (näiteks üks vigaselt sisestatud palganumber võib Eesti keskmise palga kahekordistada).

Samas võivad muret valmistada ka sellised vead, mis ei osutu erinditeks. Näiteks võivad andmete sisestamisel Antsu ja Jaani andmed vahetusse minna. Statistilise analüüsi tegemise seisukohast pole tegemist õnnetusega, küll aga võib eksitus mainitud isikutele kõvasti tüli tekitada.

Erindite leidmiseks on välja pakutud mitmeid erinevaid meetodeid. Antud töö raames pööratakse eelkõige tähelepanu nende võimele avastada jämedaid vigu.

Käesolev magistritöö

- leidub mis paneb mõtlema, et andmestikus on erind,
- kuidas see tekib,

- milliseid meetodeid on erindi tabamiseks
- ja mida edasi teha, kui erind on kindlaks määratud?

Toome näiteks olukorra, kus tekivad kahtlused vaatluse väärtuse õigsuse kohta. Barnett ja Lewis [1] toovad järgmise näite. Analüüsiti temperatuuride andmestikku, milles oli mitme aasta jooksul mõõdetud temperatuuri iga tunni aja järel. Algselt oli temperatuur mõõdetud Fahrenheiti kraadides. Ajavahemikus 31.12.1960. aasta hilja õhtust kuni 01.01.1961. aasta varahommikuni mõõdetud temperatuuride reas oli kaks üllatuslikult suurt väärtust. Hiljem selgus, et kesköö oli meteoroloogiaamet muutnud mõõtühiku Fahrenheiti kraadidest 0.1 Celsiuse kraadideks. Kui oli tehtud vastav teisendus, siis varasemad suured väärtused ei paistnud enam teistest märkamisväärselt erinevad. Selles näites toodud kahe vaatluse tulemused ei olnud kooskõlas teiste tulemustega ning osutusid peale kontrollimist vigadeks.

Siit järeldub, et kui uurijal tekib tunne, et mingi vaatlus ei ole teistega kooskõlas, tuleb hoolikalt kontrollida algandmeid. Toodud näites hakati otsima viga siis, kui vaatluse osad olid "liiga" kaugel teiste vaatluste keskmisest või mediaanist. Mitmed erindi konkreetsed definitsioonid lähtuvad sellest põhimõttest.

Erindi leidmiseks võib kasutada erinevaid meetodeid. Ühe- ja kahemõõtmeliste andmete korral edukalt kasutatavad erinevad graafilised meetodid, näiteks karpdiagramm või hajuvusgraafik.

Kui vaadeldava tunnuse jaotusel on pikad sabad ja enamik väärtustest asub jaotuse keskel (näiteks normaaljaotus), siis erindid kipuvad olema sabade otstes. Sellel juhul võib erindite leidmiseks kasutada meetodeid, mis mõõdavad vaatluse kaugust mediaanist või keskväärtusest.

Arvestades seda asjalolu võib näiteks tuua järgmise erindi definitsiooni ühemõõtmelise jaotuse jaoks [2].

Definitsioon 2. *Kui kvartiilide vahe tähistada Q, siis erindiks loetakse vaatlus, mis asub alumisest või ülemisest kvartiilist kaugemal kui 1.5Q.*

Tunnuse väärtused on oma loomult hajuvad. Pidevate tunnuste puhul võib seda loomulikku, "lubatud" hajuvust kirjeldada näiteks regressioonanalüüsi abil. Muret teevad need tunnuste väärtused, mis on märgatavalt erinevad isegi peale "lubatud" erinevuste arvesse võtmist.

Ühe tunnuse väärtuste seast erindite leidmiseks on välja töötatud mitmeid erinevaid meetodeid. Suhteliselt head meetodid erindite leidmiseks on välja pakutud ka juhuks, kui üheagselt uuritakse kahte tunnust. Mitmemõõtmeliste andmete jaoks on vähe hästitöötavaid meetodeid.

1.2 Vead

Kuna andmeid koguvad ja sisestavad inimesed, seepärast on loomulik, et tehakse vigu. Võimalik on eksida mõõtühikuga, vaatlused võivad nihkesse sattuda, vigu võib põhjustada hooletus. Kui andmebaasi haldaja ei kasuta erakorralisi meetmeid vigade ohjamiseks, siis umbes 1-5% tüüpilises andmebaasis sisalduvatest kirjetest on vigased [3].

Andmete kogumisel ja sisestamisel tehtavaid vigu saab jagada mitmeks erinevaks liigiks. Osad vead on mõõtmisvead. Mõõtmisvead võivad juhtuda andmete kogumisel ebatäpse mõõtmise (inimeste pikkus mõõdetud vaid sentimeetri täpsusega) või informatsiooni puudulikkuse tõttu (laps, kelle kodu asub Tartu linna piiril, ei pruugi osata küsitlejale vastata, kas ta elab Tartu linnas või Tartumaal).

Mõõtmisvead on ebasoovitavad ja oleks parem, kui kõik tunnused oleksid mõõdetud absoluutse täpsusega või vähemalt võimalikult täpselt. Samas ei pruugi huvipakkuvate tunnuste täpsem mõõtmine olla praktiline või võimalik. Sobivaid statistilisi meetodeid kasutades (näiteks struktuurivõrrandeid) võib mõõtmisvigu sisaldavaid andmeid analüüsides jõuda korrektsete tulemusteni. Näiteid mõõtmisvigadega andmestiku analüüsiks võib leida raamatus [4]. Mitmed enam-kasutatavad statistilised meetodid on samuti suhteliselt robustsed väiksemate mõõtmisvigade suhtes.

Veidi teistsugused on jämedad ehk mitteinformatiivsed vead, mis võivad tekkida nii

andmete sisestamisel (eksimine komakohaga, mõõtmistulemuste kirjutamine valele reale vms.) kui ka muudes situatsioonides (vead andmete kopeerimisel või ülekandmisel, küsitleja ei kuulnud küsitletava vastust õieti vms).

Mitte iga hooletus ei pruugi lõppeda vale väärtuse sattumisega andmebaasi, ka juhuslikult löödud number (teeme sisestamisvea) võib kokku langeda õige väärtusega. Teiste sõnadega, sisestusvea tagajärjel võib andmebaasis olev väärtus olla vigane, kuid alati ei tarvitse sisestusviga põhjustada vale väärtuse sattumist andmebaasi. Antud töös nimetame kõiki selliseid eksimusi sisestusvigadeks.

Definitsioon 3. *Sisestusviga – andmete kogumisel või sisestamisel tehtud viga, mille tagajärjel andmebaasis olev väärtus võib (kuid ei pruugi) olla vigane.*

Matemaatiliselt võib sisestusviga kirjeldada järgmiselt: olgu meid huvitav tunnus X *jaotusega* F, $X \sim F$. Andmete kogumisel või sisestamisel asendatakse sisestusvea toimumisel tunnuse X õige väärtus x ühe teise juhusliku suuruse $X_{viga} \sim G$ realisatsiooniga.

Teisisõnu: vaadeldud väärtuseid sisaldava tunnuse $X_{vaadeldud}$ võib kirja panna kui

$$X_{vaadeldud} = (1 - I_X)X + I_X X_{viga},$$

kus I_X on indikaatortunnus, mis näitab sisestusvea tegemist tunnuse X sisestumisel. Eeldades, et sisestusvea toimumine ei sõltu tunnuse tegelikust väärtusest, $I_X \perp X$, on vaadeldud väärtuste $X_{vaadeldud}$ jaotuseks

$$F_{X_{vaadeldud}} = (1 - \lambda)F + \lambda G,$$

kus λ on tõenäosus teha „sisestusviga“.

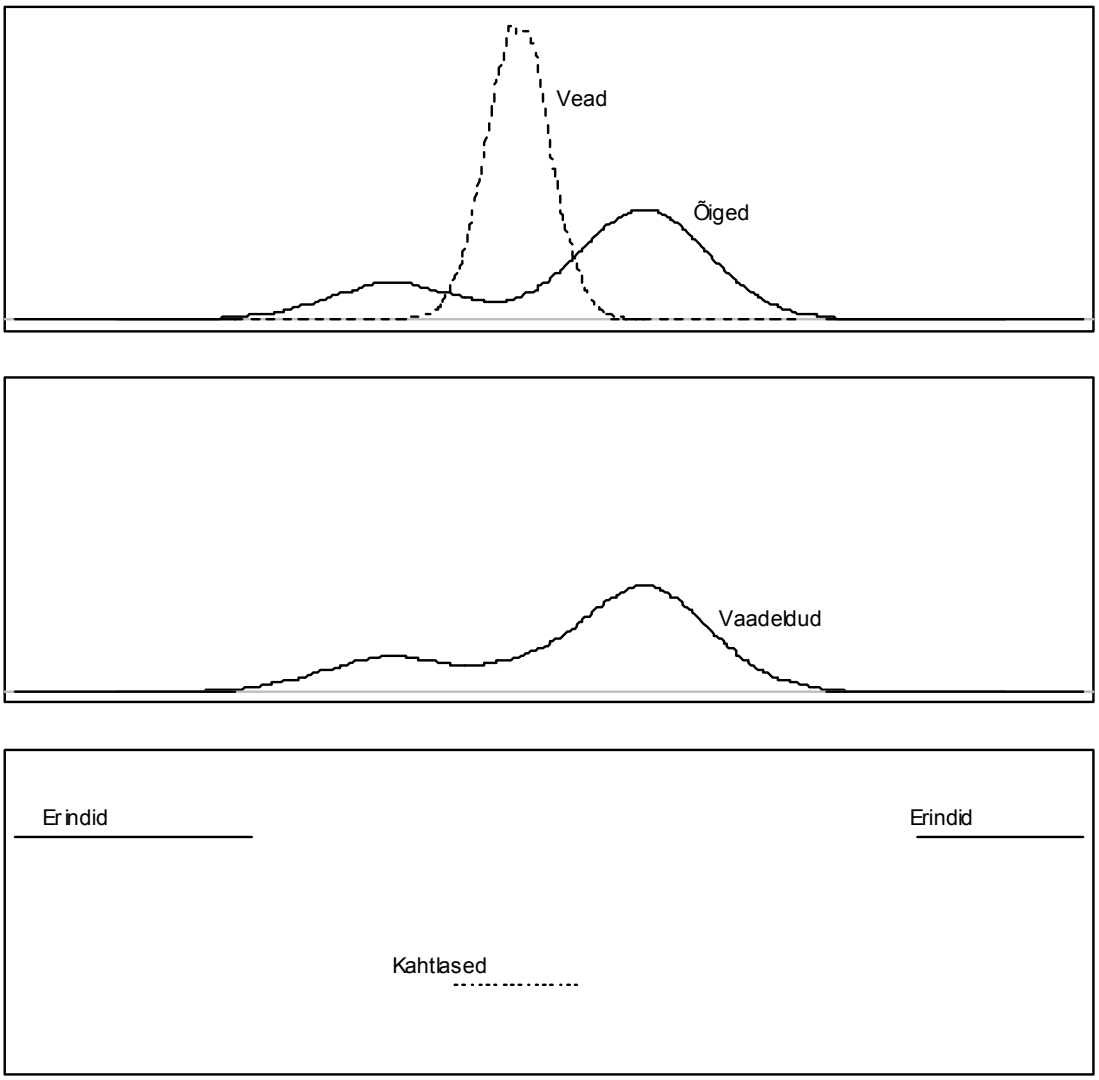
Definitsioon 4. *Kahtlane väärtus – väärtus, mis suure tõenäosusega on sisestusviga.*

Võib esineda olukordi, kus erind on ühtlasi kahtlaseks väärtuseks ja vastupidi, kuid võib ette tulla ka situatsioone, kus erinditeks osutuvad ühed ja kahtlasteks väärtusteks täiesti

teised vaatlused. Seda väidet illustreerime järgmiste näidetega.

Näide 1

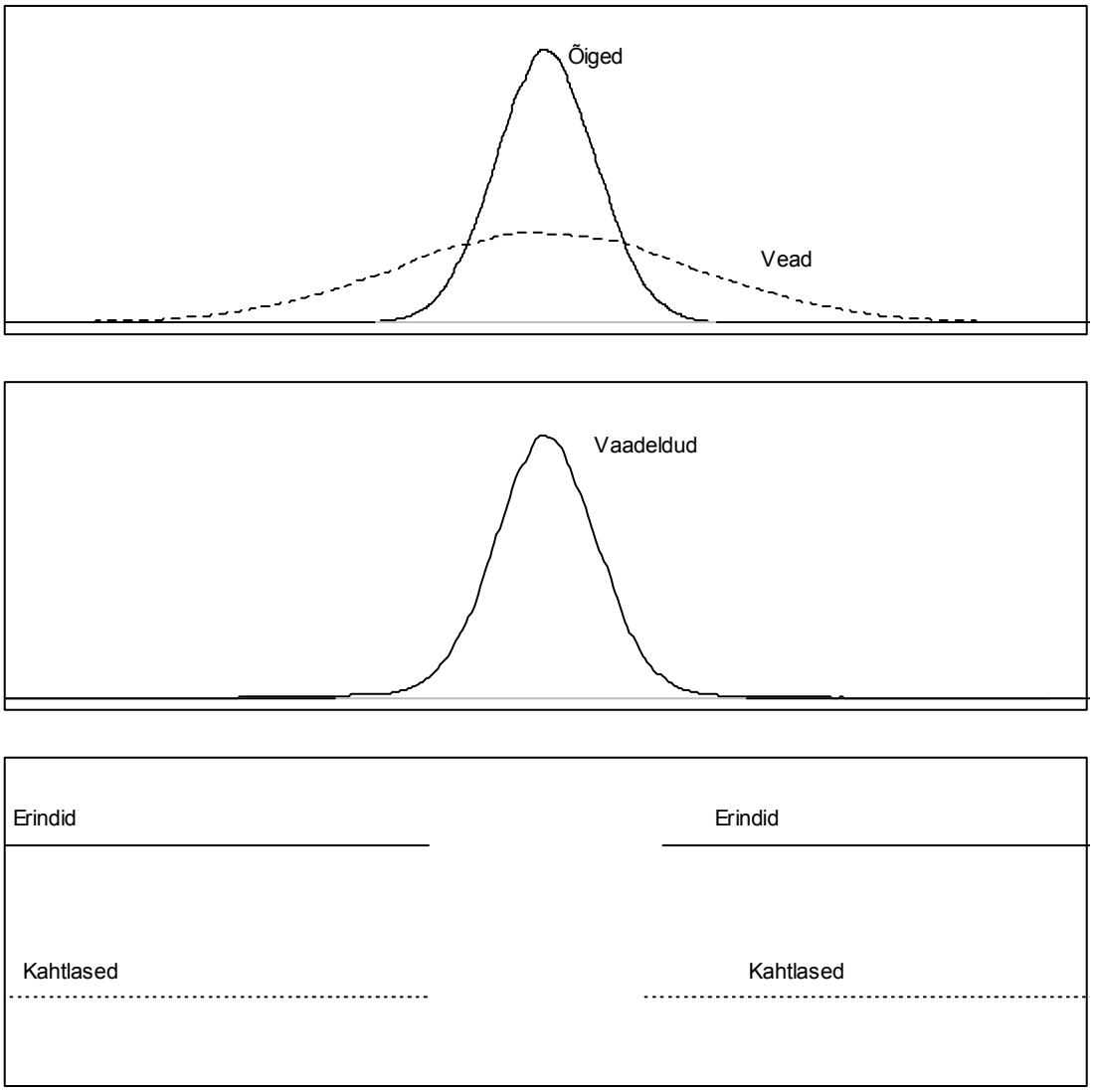
Vaatleme situatsiooni, kus erinditeks ja kahtlasteks väärtusteks osutuvad vaatlused ei lange kokku. Erindite leidmisel kasutasime definitsiooni 2.



Joonis 1. Erindid ja kahtlased väärtused

Näide 2

Selles näites on meil situatsioon, kus erinditeks ja kahtlasteks väärtusteks osutuvad samad vaatlused. Selline olukord võib esineda näiteks siis, kui sisestusvigade hajuvus on suurem tunnuse tegelike väärtuste hajuvusest.



Joonis 2. Erindid ja kahtlased väärtused

Definitsioon 4, mis määratleb kahtlase väärtuse mõiste, võib praktiliseks kasutamiseks osutuda liiga ebamääraseks. Järgnevas definitsioonis üritame pakkuda välja praktiliseks kasutamiseks paremini sobivat definitsiooni.

Definitsioon 5. *Sisaldagu andmestik tunnuseid X, Y, ..., W. Tunnuse X vaatlust x_k loeme kahtlaseks (olulisuse nivool c), kui*

$$P(x_k \text{ sisestamisel tehti sisestusviga} \mid X=x_k, Y=y, \ldots, W=w) > c. \qquad (1)$$

Kui valemis (1) esinevat tinglikku tõenäosust pole võimalik otseselt arvutada, siis tuleb ta asendada hinnanguga $\hat{P}(x_k \text{ sisestamisel tehti sisestusviga} \mid X=x_k, Y=y, \ldots, W=w)$.

2. Kahtlase väärtuse leidmine

Kahtlase väärtuse definitsioonis (Definitsioon 5) toodud tinglikku tõenäosust pole üldjuhul võimalik ilma lisaeelduseid tegemata või täiendavat informatsiooni hankimata leida. Sõltuvalt tehtud lisaeeldustest tuleb valida ka sobiv metoodika kahtlaste väärtuste leidmiseks.

Järgnevalt esitame erinevatel eeldustel baseeruvaid meetodeid tingliku tõenäosuse (1) leidmiseks.

2.1 Vigade jaotus on teada

Esiteks vaatleme kõige lihtsamat juhtu, kui meil on teada nii sisestusvigade jaotus $X_{viga} \sim F_{X_{viga}}$, kui ka vaadeldud väärtuste jaotus $X_{vaadeldud} \sim F_X$ (näiteks kui samad andmed on kättesaadavad mitmetest infoallikatest). Siis on sisestusvea toimumise $\{I_X = 1\}$ tinglik tõenäosus leitav valemiga:

$$P(I_X = 1 \mid X_{vaadeldud} = x) = \frac{P(X_{vaadeldud} = x \mid I_X = 1)P(I_X = 1)}{P(X_{vaadeldud} = x)}. \qquad (2)$$

Juhul, kui meie andmestikus on veel tunnuseid, mida saab esitada tunnusvektorina $Y=(Y_1, Y_2, ..., Y_k)$, kuid viga saab tekkida vaid tunnuse X sisestamisel, siis võib valemit (2) üldistada järgmiselt:

$$P(I_X = 1 \mid X_{vaadeldud} = x, Y = (y_1, y_2, ..., y_k)) = \frac{P(X_{vaadeldud} = x, Y = (y_1, y_2, ..., y_k) \mid I_X = 1)P(I_X = 1)}{P(X_{vaadeldud} = x, Y = (y_1, y_2, ..., y_k))}.$$

Juhul, kui teame tunnuse X õigete väärtuste jaotust, on samuti võimalik leida tinglikku tõenäosust (1):

$$\begin{aligned} P(I_X = 1 \mid x, y_1, y_2, ..., y_k) &= 1 - P(I_X = 0 \mid x, y_1, y_2, ..., y_k) \\ &= 1 - \frac{P(x, y_1, y_2, ..., y_k \mid I_X = 0)}{P(x, y_1, y_2, ..., y_k)} \end{aligned}$$

Näide 3

Eesti rahvastiku vanuseline koosseis 1. jaanuaril 2000 on teada kahest allikast – on olemas rahvastikuregistri baasil saadud tulemus ja rahvaloenduse andmed [5]. Loeme siin näites rahvaloenduse põhjal saadud vanuselise kooseisu tunnuse „vanus“ õigeks jaotuseks ja rahvastikuregistri põhjal (korrigeerimata) tulemuse tunnuse „vanus“ vaadeldud väärtuseks.

Kasutades mõlemat jaotust, on võimalik leida, millises vanuses inimeste andmed on rahvastikuregistris „kõige kahtlasemad“. Alljärgnevas tabelis on toodud tõenäosused, et vastavasse vanusegruppi kuuluva inimese kirje on vigane. Konstandiga c on tähistatud vigade osakaalu andmestikus $c := P(I_X = 1)$:

	Vaadeldud väärtuste jaotus, %	Tegelike väärtuste jaotus, %	Tinglik tõenäosus, et vaadeldud väärtus on viga
0	0.86	0.88	1.02*c - 0.02
1–4	3.53	3.60	1.02*c - 0.02
5–9	5.85	5.96	1.02*c - 0.02
10–14	7.73	7.83	1.01*c - 0.01
15–19	7.49	7.54	1.01*c - 0.01
20–24	7.15	6.90	0.97*c + 0.03
25–29	7.26	6.90	0.95*c + 0.05
30–34	6.92	6.50	0.94*c + 0.06
35–39	7.25	7.11	0.98*c + 0.02
40–44	7.32	7.26	0.99*c + 0.01
45–49	6.91	6.91	c
50–54	6.08	6.15	1.01*c - 0.01
55–59	5.34	5.52	1.03*c - 0.03
60–64	5.80	6.00	1.03*c - 0.03
65–69	4.91	5.11	1.04*c - 0.04
70–74	4.27	4.41	1.03*c - 0.03
75–79	2.74	2.81	1.03*c - 0.03
80–84	1.31	1.35	1.03*c - 0.03
85+	1.26	1.29	1.02*c - 0.02

Kuna $c<1$, siis näeme, et vanuses 25-34 inimeste andmed on rahvastikuregistris kirjas märksa madalama kvaliteediga kui ülejäänud vanusegruppide andmed.

2.2 Andmete topeltsisestamine

Andmete topeltsisestamist kasutatakse sageli seal, kus sisestusvigade esinemist on tarvis minimiseerida. Seni teadaolevatest meetoditest on andmete topeltsisestamine parim kvaliteetsete andmete saamiseks [6]. Näiteks kasutavad kaks kolmandikku Põhja-

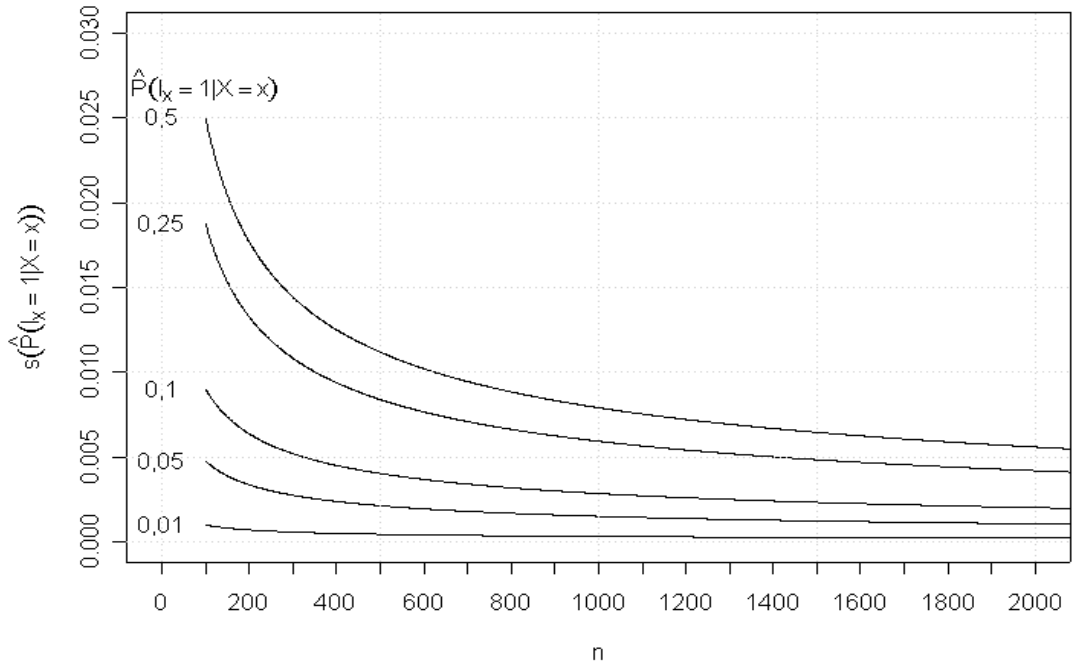
Ameerikas asuvaist biostatistikaga tegelevatest üksustest andmete sisestamisel topeltsisestamise meetodit [7].

Topeltsisestamise protsess on lihtne. Kaks inimest sisestavad sama osa andmestikust. Pärast sisestamist tehakse failide võrdlemine. Kui sisestatud väärtused langevad kokku siis loetakse sisestatud väärtus õigeks. Lahknevuste korral kontrollitakse vastav väärtus üle ja korrigeeritakse. Võimalik on talletada nii tehtud vigu (sisestusvigade jaotuse uurimiseks) kui ka hinnata sisestusvigade osakaalu [8].

Suurte andmekogumite korral kogu andmestiku topeltsisestamine on aga mahukas ja palju aega nõudev töö. Samal ajal vigade jaotuse hindamiseks on piisav ainult osa andmestiku topeltsisestamine (näiteks 1000 kirjet).

Vaatluste lahterdamine sisestusvigadeks (kahtlaseks väärtusteks) ja "õigeteks", kasutades topeltsisestamise teel saadud valimit, on tüüpiline klassifitseerimisülesanne. Vaatluste klassifitseerimisel saab kasutada näiteks logistilist regressiooni.

Valimimahu kasvuga kahaneb erinevus tegelike tinglike tõenäosuste ja topeltsisestamisemeetodil saadud hinnangute vahe. Allpool oleval graafikul on näha hinnatud tingliku tõenäosuse vea käitumist sõltuvalt vea protsendist ja vastava kategooria $X = x$ esinemissagedusest topeltsisestatud valimis (n). Mida väiksem on vea tekkimise tõenäosus ja mida suurem on topeltsisestamiseks kasutatud valimimaht, seda väiksem on tingliku tõenäosuse hinnangu viga.



Näide 4[1]

Selles näites vaatame situatsiooni, kui mingi diskreetse või kodeeritud tunnuse väärtuste sisestamisel inimene sisestab vale numbri. Samal ajal see number kuulub tunnuse tegelike väärtuste hulka. Näiteks, kui sisestavaks tunnuseks on inimese sugu, siis võib juhtuda, et kogemata ühe (mees) asemel sisestatakse kaks (naine) ja vastupidi.

Vaatame nelja võimaliku väärtusega diskreetset tunnust X. Olgu see inimese haridustase tunnus, kus 1=alg-, 2=põhi-, 3=kesk- ja 4=kõrgharidus. Sisestusvea toimumisel sisestatakse õige väärtuse asemel juhuslikult valitud haridustase.

Genereerime tunnuse X väärtused järgmisest jaotusest:

X	1	2	3	4
$P(X=x)$	0.1	0.3	0.4	0.2

[1] Näidetes kasutatud programmid on leitavad tööle lisatud CD pealt.

Seejärel genereerime tunnuse X_{viga} väärtused järgmisest jaotusest:

X	1	2	3	4
$P(X=x)$	0.25	0.25	0.25	0.25

Sisestusvea toimumist näitava indikaatortunnuse I_X väärtused genereerime Bernoulli jaotusest $I_X \sim B(0.2)$.

Moodustame tunnuse $X_{vaadeldud}$ väärtused järgmise eeskirja kohaselt:

$$X_{vaadeldud} = \begin{cases} X, & I_X = 0 \\ X_{viga}, & I_X = 1. \end{cases}$$

Teades tegelike väärtuste (X) ja sisestusvigade (X_{viga}) täpseid jaotusi, arvutasime välja tinglikud tõenäosused, et konkreetse vaadeldud väärtuse x puhul on tegemist sisestusveaga, $P(I_X = 1 \mid X_{vaadeldud} = x)$:

$X_{vaadeldud}$	1	2	3	4
$P(I_X = 1 \mid X_{vaadeldud} = x)$	0.3828849	0.1684824	0.1357812	0.2438549

Praktikas pole tinglike tõenäosuste leidmine toodud viisil võimalik, sest tunnuste X ja X_{viga} jaotus pole teada.

Juhul, kui topeltsisestamise teel n vaadeldud väärtuste jaoks on teada, kas tegemist on sisestusveaga või mitte (tunnuse I_X väärtused on teada), siis on võimalik hinnata tinglikku tõenäosust (2) iga x väärtuse korral.

$X_{vaadeldud}$	Vigade osakaal	Hinnatud tõenäosus $\hat{P}(I_X = 1 \mid X_{vaadeldud} = x)$
1	0.3828849	0.339985
2	0.1684824	0.163265
3	0.1357812	0.130117
4	0.2438549	0.252725

Me võrdlesime tulemusi ka juhul, kui topeltsisestamiseks on kasutatud erinev arv kirjeid (nimetame seda valimi mahuks):

Valimimaht	Hinnatud tõenäosuse $\hat{P}(I_X = 1 \mid X_{vaadeldud} = x)$ keskmine standardviga
500	0.0015
1000	0.0011
3000	0.0006
5000	0.0004
10000	0.0002

Tinglikke tõenäosuseid $\hat{P}(I_X = 1 \mid X_{vaadeldud} = x)$ saab hinnata näiteks logistilise regressiooni abil. Logistiline regressioonanalüüs võimaldab tingliku tõenäosuse hindamisel kasutada ka teisi tunnuseid peale tunnuse *X*.

Pidevad tunnused.

Kui *X* on pidev tunnus, siis $P(X_{vaadeldud} = x)=0$ iga *x* korral. Tõenäosuste asemel tihedusfunktsiooni kasutades saame tinglikku tõenäosust $P(I_X = 1 \mid X_{vaadeldud} = x)$ leida järgmise valemi abil:

$$P(I_X = 1 \mid X_{vaadeldud} = x) = \frac{f_{Xviga}(x)P(I_X = 1)}{f_{Xvaadeldud}(x)},$$

kus f_{Xviga} on sisestusvigade tihedusfunktsioon ja $f_{Xvaadeldud}$ vaadeldud väärtuste tihedus-funktsioon.

Kui me tahame teada tõenäosust, kas mingi väärtus on vigane, siis tuleb kasutada tingliku tõenäosuse valemit:

$$P(I_X = 1 \mid X_{vaadeldud} \le x) = \frac{P(X_{vaadeldud} \le x \mid I_X = 1)P(I_X = 1)}{P(X_{vaadeldud} \le x)}.$$

Tõenäosus, et pidev juhuslik suurus $X_{vaadeldud}$ on väiksem või võrdne mingist väärtusest *x*, tingimusel, et ontoimunud sisestusviga, leitakse valemiga:

$$P(X_{vaadeldud} \le x \mid I_X = 1) = P(X_{viga} \le x) = F_{X_{viga}}(x).$$

Seetõttu võime kirjutada:

$$P(I_X = 1 \mid X_{vaadeldud} \le x) = \frac{P(X_{vaadeldud} \le x \mid I_X = 1)P(I_X = 1)}{P(X_{vaadeldud} \le x)} = \frac{F_{X_{viga}}(x)P(I_X = 1)}{F_{X_{vaadeldud}}(x)}, \quad (3)$$

kus $F_{X_{viga}}(x)$ ja $F_{X_{vaadeldud}}(x)$ on vastavalt tunnuste X_{viga} ja $X_{vaadeldud}$ jaotusfunktsioonid kohal x.

Valemi (3) abil saab leida ainult vasakpoolsed tõenäosused. Selleks, et leida parempoolsed tõenäosused, tuleb kasutada järgmist valemit:

$$P(I_X = 1 \mid X_{vaadeldud} > x) = \frac{P(X_{vaadeldud} > x \mid I_X = 1)P(I_X = 1)}{P(X_{vaadeldud} > x)} = \frac{(1 - P(X_{vaadeldud} \le x \mid I_X = 1))P(I_X = 1)}{1 - P(X \le x)} = \frac{F_{X_{viga}}(x)P(I_X = 1)}{1 - F_{X_{vaadeldud}}(x)}. \quad (4)$$

Alguses oli mainitud, et kuna $X_{vaadeldud}$ on pidev tunnus, siis $P(X_{vaadeldud} = x)=0$ iga x jaoks. Kui me tahame leida sellist tõenäosust igas konkreetses punktis, siis tuleb kasutada teist valemit.

Näide 5

Näites kasutati simuleeritud andmeid. Õiged ja vigased väärtused olid sama keskväärtusega. Tunnuste jaotuseks oli võetud normaaljaotus, õigete väärtuste jaotus oli $X \sim N(10, 10)$ ja vigade väärtuste jaotuseks oli $Y \sim N(10, 25)$. Vigade indikaatorfunktsioon $I \sim B(0.05)$. Kokku oli genereeritud kümme tuhat vaatlust.

Eeldasime, et topeltsisestamiseks oli kasutatud ainult esimesed tuhat vaatlust (valim X_1, $X_{vaadeldud1}$ ja I_1), samal ajal kasutasime kõik $X_{vaadeldud}$ väärtuseid.

Võrdluseks võtsime ka logistilise regressiooni kõige lihtsama mudeli: $P(I_1=1) = X_{vaadeldud} + X^2_{vaadeldud}$. Meetodite töö võrdlemiseks kasutasime ROC-kõverad (*receiver operating*

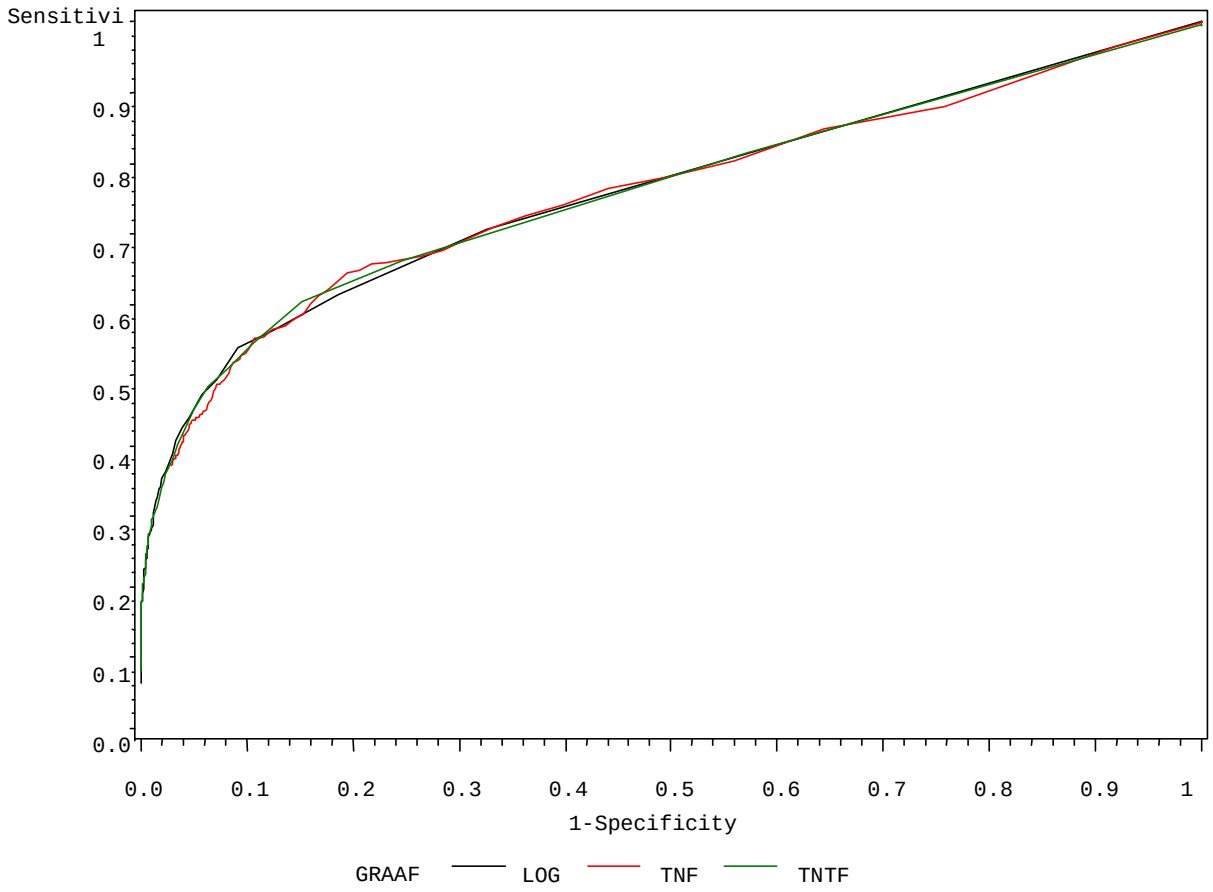
characteristic curves). Kontrollivate vaatluste arvu ja seega ka avastatud vigade arvu saab muuta muutes valitud olulisuse nivood. ROC-kõverate punktide määramiseks kasutati SAS makrot, mille abil arvutati vajalikud murdepunktid ja funktsioonide väärtused nendes punktides.

Murdepunktiks võtame tunnuse $X_{vaadeldud}$ iga punkti ja arvutame tõenäosusfunktsiooni väärtuse nii vasakul kui ka paremal poolt seda punkti. Selleks kasutasime valemeid (3) ja (4). Peale tulemuste leidmist valisime edaspidiseks kasutamiseks neist suurema, tulemused salvestati TNF nime alla.

Vaadeldud väärtuste jaotusena $F_{X_{vaadeldud}}$ kasutasime empiirilist jaotust. Kuna X ja X_{viga} korral oli teada, et tunnused on normaaljaotusega, siis jaotusfunktsiooni väärtus kohal x oli võimalik leida kaustades SAS funktsiooni *probnorm*, ja valimi põhjal hinnatud tunnuste X_1 ja X_{viga1} keskväärtust ja standardhälvet.

Kolmanda mudeli TNTF jaoks kasutasime valemit (5). Tunnuste X_{viga1} ja $X_{vaadeldud}$ tihedusfunktsioonid olid leitud empiiriliselt, kasutades SAS protseduuri *KDE*.

Järgmisel graafikul on esitatud ROC-kõverad kirjeldatud kolme meetodi jaoks.



Joonis 4. ROC kõverad.

Mudelite töö võrdlemiseks kasutati kõvera alla jäävat pindala (AUC - Area Under the Curve) - mida suurem on kõvera alune pindala, seda parema meetodiga on tegemist.

Saadud tulemused - kõveraalused pindalad - on esitatud alljärgnevas tabelis.

Tabel 1. Tõenäosus- ja logistilise regressiooni mudelite tulemuste võrdlemine.

Valimi maht	TNF	TNTF	LOG
1000	0.726	0.737	0.734
2000	0.727	0.738	0.737
3000	0.730	0.735	0.741
5000	0.729	0.734	0.741
10000	0.730	0.732	0.738

2.3 Eeldused vigade tekkemehhanismi kohta

Eelnevalt vaatasime olukorda, kus tinglikku tõenäosust (1) hinnatakse kasutades andmete topeltsisestamist. Andmete topeltsisestamiseks ei pruugi paraku alati leiduda piisavalt aega või raha. Alljärgnevalt vaatleme, kuidas saaks sisestusvigu eraldada õigetest väärtustest tehes eelduseid sisestusvigade tekkemehhanismi kohta.

2.3.1 Sisestusvead ja tegelikud väärtused on sama marginaaljaotusega

Selle meetodi eeldusteks on:

- tunnus X, mille sisestamisel võib tekkida sisestusvigu, on statistiliselt sõltuv teistest andmestikus esinevatest tunnustest; sisestusvead on aga ülejäänud tunnustest sõltumatud;
- meil on teada sisestusvigade esinemissagedus;
- teised andmestiku esinevad tunnused võib lugeda veavabaks;
- sisestusvigade X_{viga} jaotus on sama, mis õigete väärtuste jaotus.

Vaatame, millal sellised eeldused võiksid olla täidetud.

Enamik kogutavaid tunnuseid pole täiesti sõltumatud (kaal ja pikkus, emakeel ja rahvus jne), seega võiks esimene eeldus olla rahuldatud.

Teine ja kolmas eeldus võiksid olla rahuldatud, kui samade tunnustega on läbi viidud eelnevaid uuringuid (näiteks ühe tunnuse väärtused on korduvalt sisestatud). Sellel juhul on võimalik hinnata vigade osakaalu. Kui osa uuringus kasutatud tunnustest oli võetud mingist andmebaasist või andmebaasi osast, mis oli varem juba kontrollitud (vead olid välja korjatud või asendatud tegelike väärtustega), siis võib oletada, et antud tunnused on vigadest puhtad ja meil on võimalik eksida ainult uue tunnuse sisestamisel.

Viimane eeldus sama jaotuse kohta võib olla rahuldatud, kui andmete kokkuviimisel

toimus andmete ümbertõstmine (sarnane nimi vms), või kui uue kirje sisestamisel sisestusvea puhul sisestatakse eelneva inimese sama tunnuse väärtus.

Tinglike tõenäosuste hindamine

Olgu antud kahe tunnuse, X ja Y, ($m \times n$)-mõõtmeline sagedustabel. Soovime leida tinglikku tõenäosust, et $X = x_i$ ja $Y = y_j$ puhul on tunnus X valesti sisestatud.

Tõenäosuse $P(X = x_i)$ hinnangut tähistame $p_{i\cdot}$, tõenäosuse $P(Y = y_j)$ hinnangut tähistame $p_{\cdot j}$ ja tähistame $P(X = x_i,\ Y = y_j)$ hinnangut sümboliga p_{ij}. Tunnuste X ja Y ühisjaotus avaldub seega järgmisel kujul:

	y_1	...	y_n	
x_1	p_{11}	…	p_{1n}	$p_{1\cdot}$
x_2	p_{21}	…	p_{2n}	$p_{2\cdot}$
.	.	…	.	.
.	.	…	.	.
.	.	…	.	.
x_m	p_{m1}	…	p_{mn}	$p_{m\cdot}$
	$p_{\cdot 1}$	…	$p_{\cdot n}$	1

Kui tegemist on sisestusveaga ($I_X = 1$), siis X ja Y on sõltumatud ja

$$p^*_{ij} = p_{i\cdot}\, p_{\cdot j}.$$

Olgu üldkogumi mahuks N. Selliste sisestusvigade arv, mille puhul $X_{viga} = x_i$ ja $Y = y_j$, on

$$\hat{n}^*_{ij} = N p^*_{ij} P(I_X = 1).$$

Kui iga lahtri tegelik sagedus on n_{ij}, siis veaväärtuse tõenäosuse teoreetilise hinnangu ($\widehat{P}(I = 1 \mid X = x, Y = y)$) leiame järgmise valemi abil:

$$\widehat{P}(I = 1 \mid X = x, Y = y) = \frac{\hat{n}^*_{ij}}{n_{ij}}. \tag{5}$$

Näide 6

Vaatame situatsiooni, kus andmestikus on kaks binaarset tunnust ja sisestusvigu on esineb vaid ühes neist tunnustest. Olgu näiteks üheks tunnuseks rahvus (1 - eesti, 2 -

vene) ja teiseks vastus küsimusele „kas teie emakeeleks on eesti keel?“ (0 = ei ole, 1 = on). Oletame, et esimese tunnuse väärtused on saadud mingist andmebaasist (rahvastikuregister) ja olid varem kontrollitud. Teise tunnuse väärtused aga sisestatakse käsitsi. Sellel juhul võib oletada, et kui viga tekib, siis see tekib uute andmete sisestamisel.

X (õiged väärtused) ja Y jaotustabel:

	$Y = 1$	$Y = 2$
$X = 0$	0.4	0.2
$X = 1$	0.1	0.3

Vaadeldud väärtuste genereerimiseks kasutatakse sisestusvigade indikaatortunnust I, $I \sim B(0.2)$. Kui $I = 1$, siis tehakse sisestusviga ning tunnuse X tegelik väärtus asendatakse juhusliku suurusega $X_{viga} \sim B(0.4)$.

Kasutades valemit (5), hindame vea tegemise tinglikud tõenäosused.

Jagades omavahel oodatud ja tegelikud sagedused ning korrutades saadud tulemuse 0.2-ga, saame teoreetilistele tõenäosustele $P(I_X = 1 \mid X = x_i, Y = y_j)$ hinnangud. Allpool toodud tabelis on antud hinnatud tõenäosused ning sulgudes on tegelikud vigade esinemissagedused.

	$Y = 1$	$Y = 2$
$X_{vaadeldud} = 0$	0.1579 (0.1586)	0.2734 (0.2727)
$X_{vaadeldud} = 1$	0.3324 (0.3353)	0.1427 (0.1465)

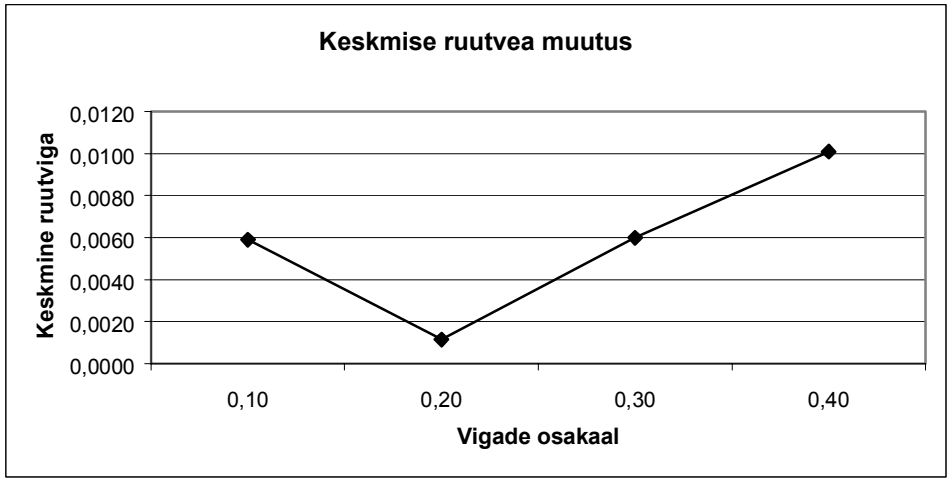
Ja lõpuks leiame tõenäosused $P(I_X = 1 \mid X = x_i, Y = y_j)$, kasutades Bayesi valemit:

	$Y = 1$	$Y = 2$
$X_{vaadeldud} = 0$	0.1579	0.2727
$X_{vaadeldud} = 1$	0.3351	0.1425

Hinnangu täpsuse kirjeldamiseks leidsime vigade suhtelised esinemissagedused ja hinnatud tõenäosuste vahe absoluutväärtuste keskmise. Tegelike suhteliste sageduste n_{ij}^* / n_{ij} ja valemi (5) abil hinnatud tõenäosuste (p^*) vahe absoluutväärtuste keskmine oli

0.002, ning n_{ij}^{*} / n_{ij} ja Bayesi hinnangu vahe oli 0.001.

Vigade tegelikku osakaalu pole enamasti võimalik täpselt ära arvata, seetõttu huvitab meid, kuivõrd tundlik võiks antud meetod olla tõenäosuse vale spetsifikatsiooni suhtes .



Graafikust on näha, et mõneprotsendiline eksimus vigade osakaalu määramisel ei tekita veel märkamisväärset viga tinglike tõenäosuste $P(I_X = 1 \mid X = x_i, Y = y_j)$ hindamisel.

2.3.2 Naaberklahvi sisestamine

Vaatame veel üht diskreetse tunnuse juhtu. Seekord on aga eeldatav sisestusvea tekkemehhanism teistsugune.

Eeldame, et sisestusviga tehes vajutakse õige klahvi asemel naaberklahvile. Näiteks 2 asemel sisestatakse 1 või 3 ja 3 asemel sisestatakse 2 või 4.

Vaatamata sellele, et sisestusviga võib osutuda tunnuse lubatud väärtuseks (kui 2 asemel sisestatakse 3 ja vastupidi), võivad sellised sisestusvead mõjutada hilisemat andmete analüüsi.

Sellise vigade tekkemehhanismi korral jäävad sisestusvead sõltuvaks teistest tunnustest. Eelmise osa eeldused ei kehti ja me vajame teist meetodit tõenäosuse $P(I_X = 1 \mid X = x)$ hindamiseks.

Tinglike tõenäosuste hindamine

Olgu meil diskreetne tunnus *X*, millel on *m* kategooriat. Meil on teada, et vale väärtuse sisestamine toimub tõenäosusega 0.5 ja sisestusveaks saab olla kategooria naaberväärtus. Näiteks kui õigeks väärtuseks on 2, siis tõenäosusega 0.5 sisestatakse kas 1 või 3. Kui teaksime vigade osakaalu andmestikus ja õigete väärtuste jaotust, saaksime innata tõenäosust, et vaadeldud väärtus pole viga:

$$P(I = 0 \mid X_{vaadeldud} = x, Y) = \frac{P(X_{vaadeldud} = x, Y \mid I = 0)P(I = 0)}{P(X_{vaadeldud} = x, Y)}.$$

Kui me teame seda tõenäosust, siis saame leida ka tõenäosuse, et vaadeldud *x* on viga:

$$P(I = 1 \mid X_{vaadeldud} = x, Y) = 1 - P(I = 0 \mid X_{vaadeldud} = x, Y) \quad (6)$$

Õigete väärtuste jaotuse hindamiseks loome võrrandite süsteemi.

Meil on teada, kuidas tekib vaadeldud väärtuste jaotus, kui $X_{vaadeldud} = 1$:

$$F_{X_{vaadeldud}} = \frac{P(I = 1)F_X(x_2)}{2}.$$

Siit leiame, et

$$F_X(x_2) = \frac{2F_{X_{vaadeldud}}(x_1)}{P(I = 1)}.$$

Analoogselt leitakse

$$F_{X_{vaadeldud}} = \frac{P(I = 1)F_X(x_{m-1})}{2}$$

ja

$$F_X(x_{m-1}) = \frac{2F_{X\,vaadeldud}(x_m)}{P(I = 1)}.$$

Ülejäänutel juhtudel:

$$F_{X_{vaadeldud}}(x_i) = P(I=0)F_X(x_i) + \frac{P(I=1)(F_X(x_{i\text{-}1}) + F_X(x_{i+1}))}{2},$$

$$F_X(x_i) = \frac{2F_{X_{vaadeldud}}(x_i) - P(I=1)(F_X(x_{i\text{-}1}) + F_X(x_{i+1}))}{2P(I=0)}.$$

Sellel viisil saame leida kõik õigete väärtuste tõenäosused kuna *m* teadmata väärtuste leidmiseks tekib meil *m*+2 võrrandit.

Kui saksime kasutada tunnuse $X_{vaadeldud}$ tegelikku jaotust, oleks esitatud võrrandsüsteem kooskõlaline. Kasutades tõenäosuste asemel aga tõenäosuste hinnanguid võime saada vastuolulise süsteemi. Jõudmaks lahendini, mis võimalikult hästi sobiks üldtoodud võrrandisüsteemiga, kasutame vähemruutude meetodit.

Hinnatavad tõenäosused peavad olema piiratud alt ning ülevalt. Need ei saa olla väiksemad nullist ja suuremad kui $F_{Xvaadeldud}(x_i)/P(I=0)$. Ülemine piir leiti järgmisel moel:

$$F_X(x_i) = \frac{2F_{X_{vaadeldud}}(x_i) - P(I=1)(F_X(x_{i\text{-}1}) + F_X(x_{i+1}))}{2P(I=0)}.$$

Eeldades, et

$$P(I=1)(F_X(x_{i\text{-}1}) + F_X(x_{i+1})) \geq 0,$$

saame ülemise piiri:

$$F_X(x_i) < \frac{F_{X_{vaadeldud}}(x_i)}{P(I=0)}. \qquad (7)$$

Viimane tinglik tõenäosus $P(X = x_n \mid Y)$ arvutatakse valemist:

$$P(X = x_n \mid Y) = 1 - \sum_{i=1}^{n-1} P(X = x_i \mid Y).$$

Näide 7

Olgu meil kaks tunnust *X* ja *Y*. Nende ühisjaotus on järgmine:

	X = 2	*X = 3*	*X = 4*	*X = 5*	*X = 6*	*X = 7*
Y = 0	0.1	0.1	0.1	0.1	0	0
Y = 1	0.15	0	0.2	0	0.05	0.2

Ning vigade ja õigete ühisjaotus on

	$X_{viga} = 1$	$X_{viga} = 2$	$X_{viga} = 3$	$X_{viga} = 4$	$X_{viga} = 5$	$X_{viga} = 6$	$X_{viga} = 7$	$X_{viga} = 8$
X = 2	0.1	0	0.15	0	0	0	0	0
X = 3	0	0.05	0	0.05	0	0	0	0
X = 4	0	0	0.15	0	0.15	0	0	0
X = 5	0	0	0	0.05	0	0.05	0	0
X = 6	0	0	0	0	0.025	0	0.025	0
X = 7	0	0	0	0	0	0.1	0	0.1

Veaindikaatorfunktsioon on $I \sim B(0.2)$.

Tingliku tõenäosuse hindamine

Ülaltoodud võrrandsüsteemi lahndamiseks kasutati R'i funktsiooni *constrOptim*. Selles funktsioonis saab anda algväärtusi, ning piirata iga tõenäosuse hinnangut alt ja ülevalt. Alumiseks piiriks oli null ja ülemine piir oli leitud valemi (7) abil. Tulemused on toodud allolevas tabelis. Saadud lahendeid kasutati tinglike tõenäosuste (6) hindamiseks.

	Tegelik	Bayes	Hinnatud	Tegelik - Bayes	Tegelik – Hinnatud
$Y = 0, X_{vaadeldud} = 1$	1.0000	1.0000	1.0000	0	0.0000
$Y = 0, X_{vaadeldud} = 2$	0.1173	0.1116	0.1089	0.00567	0.0084
$Y = 0, X_{vaadeldud} = 3$	0.2061	0.1996	0.2012	0.0065	0.0049
$Y = 0, X_{vaadeldud} = 4$	0.2029	0.1977	0.1996	0.00518	0.0033
$Y = 0, X_{vaadeldud} = 5$	0.1097	0.1115	0.1111	-0.00183	-0.0014
$Y = 0, X_{vaadeldud} = 6$	1.0000	1.0000	0.9720	0	0.0280
$Y = 0, X_{vaadeldud} = 7$	-	-	-	-	-
$Y = 0, X_{vaadeldud} = 8$	-	-	-	-	-
$Y = 1, X_{vaadeldud} = 1$	1.0000	1.0000	1.0000	0	0.0000
$Y = 1, X_{vaadeldud} = 2$	0.0000	0.0000	0.0259	0	-0.0259
$Y = 1, X_{vaadeldud} = 3$	1.0000	1.0000	0.9890	0	0.0110
$Y = 1, X_{vaadeldud} = 4$	0.0000	0.0000	0.0004	0	-0.0004
$Y = 1, X_{vaadeldud} = 5$	1.0000	1.0000	0.9674	0	0.0326
$Y = 1, X_{vaadeldud} = 6$	0.3366	0.3336	0.3248	0.003	0.0118
$Y = 1, X_{vaadeldud} = 7$	0.0305	0.3076	0.0307	-0.27711	-0.0002
$Y = 1, X_{vaadeldud} = 8$	1.0000	1.0000	1.0000	0	0.0000
Keskmine				-0.0185	0.0051

Keskmine erinevus tegeliku ja teoreetilise tõenäosuse vahel on 0.005.

3. Meetodite rakendamine praktikas

Vaadeldud meetodite rakendatavust praktikas katsetati andmestiku peal, mis kirjeldas immuunsuspuudulikkuse viirus (HIV) ning B- ja C-hepatiidi (HBV, HCV) levikut süstivate narkomaanide seas. Reaalse andmestiku peal sooritatud katse eesmärgiks oli tuvastada, kas mõni väljapakutud meetoditest suudab efektiivselt eristada ankeetide sisestamisel tehtud vigu korrektselt sisestatud andmetest.

Kasutatud andmestiku kirjeldus

Andmestik sisaldab narkomaanide poolt kasutatud süstadel B- ja C-hepatiidi ning inimese immuunsuspuudulikkuse viiruse markerite määramise andmeid. Uuringu on läbi viidud Tartu Ülikooli, New Yorgi Riikliku Ülikooli ja Yale Ülikooli teadlaste poolt.

Andmebaas sisaldas andmeid erinevate haiguste esinemise kohta näiteks HIV, HCV ja HBV. Samuti oli uuritud mitmeid sotsiaalseid, demograafilisi ja käitumisega seotud tunnuseid.

Andmete sisestamine toimus tabeltöötlemise programmi Excel, korduvsisestamist ei toimunud.

Andmete kogumine

Küsitlus oli läbi viidud nõelte vahetamise programmi raames (AIDSi keskuses Tallinnas) nende hulgas, kes:

- on 18 aastased või vanemad;
- vahetavad nõelu iseenda tarbeks;
- räägivad eesti või vene keelt;
- nõustuvad osalema antud küsitluses.

Kui inimene oli nõus uuringus osalema, võeti temalt verd, selleks et määrata haiguste olemasolu (HIV, HBV, HCV) ning viidi läbi intervjuu. Üles märgiti ka toodud süstalde

arv.

Küsitlusankeedis olid järgmised küsimused:

- üldised demograafilised näitajad;
- narkootikumide tarvitamisega seonduvad riskifaktorid;
- HIV, HBV ning HCV riskifaktorid (seksuaalne aktiivsus, meditsiinilised protseduurid jms);
- teistele inimestele infektsiooni üleandmisega seotud tegurid (vereandmine jne).

Narkootikumide tarvitamisega seonduvaid riskifaktoreid uuriti järgmiste küsimuste abil:

- vanus, millal hakati narkootikume tarvitama;
- narkootikumide tarvitamise sagedus;
- tarvitatavate narkootikumide tüübid;
- nõelte kättesaadavus ning nende jagamine (kas kasutati sama nõela koos teiste inimestega).

Küsitletuid oli 162, mõõdetavaid tunnuseid oli kokku 237, nendest 230 olid kategooriatunnused (sugu, rahvus, tööhõive jne). Väljapakutud meetodite kontrollimisel kasutati kokku 123 tunnust. Ülejäänud tunnused sisaldasid kas liiga palju puuduvaid väärtuseid või ei sobinud muudel põhjustel analüüsis kaasamiseks. Lisas 1[2] on toodud osa näidisandmestikust. Lisas 2 on esitatud kasutatud tunnuste tähendused. Lisas 3 on toodud programmid, mida kasutati meetodite töö kontrollimiseks.

3.4.1 Topeltsisestamine

Topeltsisestamiseks kasutati kaheksa küsimuse (20 - 27) andmeid kõikidest olemasolevatest ankeetidest (kokku 1134 väärtust).

[2] Kõik tabelid saab leida tööle lisatud CD-l.

Tulemused olid järgmised: 1134-st väärtusest esinevad lahknevused 46 väärtuse korral, neist

- tegi teine sisestaja (kontrollsisestaja) 2 viga;
- ühe väärtuse puhul pole ankeet korrektselt täidetud (pole arusaadav, milline vastus on märgitud);
- kahe järjestikuse ankeedi korral toimus andmete ümbervahetamine nii, et teise ankeedi andmed sattusid eespool oleva objekti andmeteks ja vastupidi (kokku 13 väärtust);
- 3 juhul toimus naaberväärtuse sisestamine;
- ülejäänud 27 väärtust ei sattunud mingil põhjusel andmefaili, kuigi vastused olid märgitud.

Käsitledes topeltsisestatud osa põhjal võime öelda, et sisestusvea tegemise tõenäosus on ligikaudu 3.9%. Ehk teisisõnu $P(I_X = 1) = 0.039$.

3.4.2 Vigade jaotus on teada

Eeldasime, et vigade jaotus on ühtlane jaotus – kõik tunnuse lubatavad väärtused lugesime võrdtõenäolisteks. Peatükis 2.1 kirjeldatud meetodi töö kontrollimiseks iga sobiva tunnuse jaoks oli leitud tõenäosus, et tunnuse väärtuse sisestamisel oli tehtud viga. Tõenäosuse väärtused järjestati kahanevas järjekorras ning võeti 72 kõige suuremale tõenäosusele vastavat sisestatud väärtust (lisa 4). Kuna sama tunnuse väärtus esines mitmes ankeedis, siis kokku kontrolliti 137 erinevat sisestatud väärtust. Osa tabelist, mida kasutati andmete kontrollimiseks, on toodud lisas 5.

Kontrollitud vaatluste seas oli 13 sisestusveaga ehk 9.5% kontrollitud väärtustest osutusid sisestusvigadeks (95% usaldusintervall sisestusvigade protsendile on 5% kuni 16%).

3.4.3 Sisestusvead ja tegelikud väärtused on sama marginaaljaotusega

Selle meetodi eeldusteks oli seose olemasolu tunnuste vahel (lisaks teised eeldused, mis olid toodud peatükis 2.3.1).

Iga vaatlusaluse uuritava tunnuse korral leiti temaga statistiliselt seotud tunnus.

Valitud tunnusepaari jaoks leiti tinglik tõenäosus, et tunnuse vaadeldav väärtus on vigane (lisa 7). Seejärel liikudes suuremast tõenäosuse väärtusest väiksema poole, oli leitud ankeetide numbrid, mis sisaldasid kahtlasemaid väärtuste kombinatsioone (kontrolliti 74, osa neist esitatud lisas 8).

Kontrollitud vaatluste seas esines kaks sisestusviga ehk 2.7% kontrollitud väärtustest osutusid sisestusvigadeks (95% usaldusintervall sisestusvigade protsendile on 0.5% kuni 10.3%).

Vaatamata sellele tundub, et meetod töötab hästi, eriti statistiliselt tugevalt sõltuvate tunnuste korral. Eriti tugev seos oli erinevate haiguste diagnooside puhul, mis on seletatav andmete spetsiifilisusega: kui inimene põeb üht uuritud haigust, siis suure tõenäosusega ta põeb ka teist.

3.4.4 Naaberklahvi sisestamine

Antud meetod eeldas, et sisestusviga tekib sisestamisel naaberklahvile vajutamisel.

Nagu ka teiste meetodite puhul, oli välja arvutatud tõenäosuste pingerida (lisa 9) ja seejärel leitud ankeetide numbrid, mis sisaldasid meile vajalikke tunnuste väärtused. Kokku kontrolliti 107 väärtust (lisa 10).

Kontrollitud väärtustest 10 osutusid sisestusveaks ehk 9% kontrollitud väärtustest osutusid sisestusvigadeks (95% usaldusintervall sisestusvigade protsendile on 4.8% kuni 16.9%).

Lisaks esines üks juht, kus käekiri ankeedis polnud loetav.

3.4.4.1 Modifitseeritud naaberklahvi juht

Andmete topeltsisestamisel selgus, et 27 vaatluse puhul oli õige väärtuse asemel andmestikku sattunud tühik. Seetõtu otsustati kolmandat meetodit modifitseerida nii et lubatav oleks ka õige väärtuse asemel tühiku sisestamine.

Korduvalt sisestatud osa põhjal hinnati tinglik tõenäosus, et sisestusvea tekkimisel on tunnuse tegelik väärtus asendatud tühikuga:

$$\widehat{P} \text{ (õige väärtuse asemel on tühik | toimus sisestusviga)} = \frac{27}{30} = 0.9.$$

Ja jagasime 30-ga, sest ainult nende väärtuste puhul toimus kas naabri või vaikimisi sisestatava väärtuse sisestamine.

Järgmisena muutsime programmis olevat tõenäosuste üleminekumaatriksit, et iga võimaliku väärtuse "naabriks" osutus ka tühik. Põhimõtteliselt on kasutatud programmile võimalik ette anda suvalist üleminekutõenäosuste maatriksit .

Nagu ka eelmiste meetodi korral, leidsime tinglike tõenäosuste pingerida (lisa 11) ja korjasime välja kontrollimiseks 137 kirjet (lõik lisas 12).

Kontrollitud vaatluste seast oli 34 sisestusviga, ehk 24.8% kontrollitud väärtustest osutusid sisestusvigadeks (95% usaldusintervall sisestusvigade protsendile on 18% kuni 33%). Seega osutus antud meetod peale modifitseerimist kõige paremaks sisestusvigade eristamisel.

Kokkuvõte

Antud magistritöös vaadeldi erinevad statistilisi meetodeid kahtlaste väärtuste ja sisestusvigade tuvastamiseks.

Töö esimese osas antakse erindi ja sisestusvea definitsioonid ning illustreeritakse toodud definitsioone näidete abil.

Teises osas käsitleti erinevatel eeldustel baseeruvaid meetodeid sisestusvigade leidmiseks. Vaadeldi võimalusi otsida sisestusvigu kui:

a) sisestusvigade jaotus on teada;

b) osa andmestikust on topeltsisestatud;

c) kui on võimalik teha eelduseid vigade tekkemehhanismi kohta.

Kolmandas peatükis vaadeldi kolme väljapakutud meetodi võimet leida sisestusvigu reaalsest andmestikust. Samuti sisestati osa andmestikust uuesti leidmaks hinnanguid sisestusvigade protsentidele andmestikus.

Sisestusvigade leidmisel osutus efektiivsemaks modifitseeritu naaberklahvi juhu meetod. Andmestikust kontrollimiseks välja selekteeritud osas oli sisestusvigu 24.8%, ehk umbes kuus korda rohkem, kui andmestikus tervikuna.

Kõik meetodid andsid erineva hinnangu vigade osakaalule kogu andmestikus. Omavahel olid sarnased esimese ja kolmanda ning teise ja topeltsisestamise meetodite hinnangud.

Modifitseeritud naabriklahvi juhu meetod töötas edukamalt teistest ja tema poolt välja valitud väärtuste seas oli sisestusvigade osakaal kõige suurem.

Antud töö tulemusena on välja pakutud mitu meetodit sisestusvigade leidmiseks. Väljapakutud meetodeid on rakendatud reaalsete andmete peal. Tänu magistritööle lisatud autori enda poolt kirjutatud programmidele saab väljatöötatud meetodeid rakendada praktikas.

Statistical methods for detecting data entry errors

Olga Goryayeva

Summary

In our days most part of work with data is done by computers. Although, the information gathering and data entry processe is still in human hands. Humans can make mistakes and even if they do not, no one can be sure, that gathered information quality is high enough for the good analysis. That is why quality of data is always a big problem for data users and analysts.

Big databases are used by many people to make different kind of reports and analysis. Their ability and possibility for detect and avoid errors in data can be at different level. Not everyone must know corresponding statistical methods (for example, outlier analysis) and at the same time, not every database user has rights to use all the variables, what can help to detect errors. That is why it is recommended, that data creator or keeper assures the data high quality. Even if data keeper doesn't know what kind of analysis or information user can ask, he must be sure, that the answer will be of high accuracy.

The data quality analysis consists of many steps. The first one is cleaning. In this work if we say "cleaning", we mean error detection and correction. After that comes control of data for:

- accessibly (can we pass to original information?),
- comparability (comparison to the same type of data),
- completeness (coverage analysis).

Big data consists of many variables. Cleaning of multidimensional data is tricky and has more possibilities than data with one variable. It can happen, that one row entry apart taken variable values are real and perfectly understandable but all together they form unlikely combination. If database was combined from different kind of other data (in big

offices from different compartments, in state office from different departments), then it can be very useful to know what value of that entry is mostly suspicious.

We must keep in mind that no method will work ideally and give all answers to our questions. In this work there are represented statistical methods that can facilitate error detection in multidimensional data and their localizing (in other words, what attribute of the data we must check first).

In the first part of this work definitions of outlier and error are brought, because some outlier detection methods can help in error detection process. First we looked outlier example, definitions and different ways of there detection. Secondly, error and suspicious value definitions were brought to attention.

In many cases errors and outliers can be very close. Outlier is a data value, which is unusual with respect to the group of data in which it is found. It may be a single isolated value far away from all the others, or a value, which does not follow the general pattern of the rest. By Barnett and Lewis: "*outliers are data that appear inconsistent with respect to the remainder of the database"* [1]. So can be an error, but the reason while it is different is other. For the outlier very often it is very high or very low point of the data, for the error it can be anywhere.

So how can we recognize the error and correct it?

If the data is two-dimensional we can try different graphical methods like box-plot or scatter plot. With multidimensional data it is not possible and error is not so easily recognizable any more, because there are too many variables. Even it is OK with one particular column, there still can be problems in rows and vice versa.

If the data is combined from different small datasets, what came from different places, then can be problems with double notes, different codes, different name spelling and so on. It would be very useful to have a contact person who can always explain original data and predict from where this or that error comes from, but unfortunately it is not possible in every situation.

In the second part we decrypted some error and error location detection methods what were based on probability formulas.

In part 2.1 we presented the simplest situation when all required probabilities are known (this can happened if data is simulated). Example with real data was brought. Statistical Office of Estonia published two population numbers in year 2000. That happened because of census, after witch population number was corrected.

In part 2.2 we looked at the situation then we can estimate all the needed distributions using duplicate performance method.

The duplicate performance method provides for double processing of all n items is the database by two individuals (or machines). Then one file is compared to the other file, and any differences are noted in a third file, which records "true" if the entries are the same and "false" if they are not. The third person then checks each false against the actual census manuscript, and determines which was right. All false indicators on the so-called "master" file are changed to the correct data and we have as error-free a product as humans can devise [5]. This method was presented for desecrate and continues variables.

In part 2.3 we constructed two methods for the situations, where we set up some conditions. First one was designed for two dependent variables, when only one can contain errors and is independent from other data variables. Second method assumed that person, who enter data from keyboard upper number row can enter neighbour value instead of right one, for example 1 or 3 instead of 2.

In part 3 was brought example with real data to demonstrate work of four described methods.

Data was collected from individuals arriving to exchange needles.

The study was carried out by scientists from Tartu University, Yale University and SUNY at Albany (State University of New York).

Dataset had 162 rows (participants) and contained 237 columns (variables - diagnoses of illnesses and answers to questionnaire). 230 variables were categorical like sex, nationality, economical status etc. Finaly we used 123 variables, other contained too much missing values or did not suit for another reasons.

Data was controlled by all methods. Firstly was found probability that variable can contain errors and all information was sorted in decreasing order. Then moving from up to down we took about 100 values for comparison with real (paper) data. All results were recorded and saved.

We discovered:

- 46 errors among 1134 checked values with duplicate performance method;
- 13 errors among 137 checked values with second method, there we assumed that error distribution is uniform;
- 2 errors among 74 checked values with third method, there we assumed that variables are depended;
- 10 errors among 107 checked values with neighbour key method;
- 34 errors among 137 checked values with modified neighbour key method, where instead of right value user could enter blank.

For the duplicated presentation we used 8 questionnaire answers (from 20 to 27) from all paper questionnaires, it make 1134 value. We found out 46 errors among them. In 27 cases in place of real value was recorded blank character. This circumstance leaded to little transformation of third method, where we added probability of entering default value (in our case blank). This time we captured 37 errors among 137 values.

The main result of this work is in comparison of different methods for detecting data entry errors. Different methods were presented and illustrated with real data example. They can be used for data cleaning and data analysis.

Kasutatud kirjandus

1. V. Barnett, T. Lewis, “Outliers in Statistical Data”, 3rd ed. Wiley, Chichester, 1994

2. Dr. Dang Quang A, Dr. Bui The Hong, “Statistical Data Analysis”, Chapter 3, http://www.netnam.vn/unescocourse/statistics/statistics.htm

3. Thomas C. Redman, “The Impact of Poor Data Quality on The Typical Enterprise”, CACM, vol. 40, nr.5, lk. 79-82

4. Kenneth A. Bollen, “Structural Equations with Latent Variables”, Wiley, 1989, lk. 151-178

5. “Rahvastik 2000 Population”, Statistikaamet, Tallinn 2001, lk. 28-29.

6. Joyce C. Niland, Tamara L. Odom-Maryon, Jennifer Lee, Barbara C. Tilley, „A Survey of Bio statistical Consulting Units throughout North America“, The American Statistician, Vol. 49, No. 2 (May, 1995), lk. 183-189.

7. Rod Anderson, “Error Detection & Verification Procedures”, http://www.fsu.edu/~guadalaj/english/guides/error.htm

8. Mike West, Robert L. Winkler, “Data Base Error Trapping and Prediction”, Journal of the American Statistical Association, Vol. 86, No. 416 (Dec., 1991), lk. 987-996.

Lisa 1. Reaalandmed (lõik)

ID	HIV	HBs	HCV	HBVc	SY	SHIV	SHBs
1001	1	2	1	1	2	1	2
1002	1	2	1	1	2	1	2
1003							
1004	2	2	1	2	2	1	2
1005	2	2	1	1	2	1	2
1006	1	2	1	1	1	2	2
1007	1	2	1	1	2	1	1
1008	2	1	1	2	2	2	2
1009	2	2	1	2	2	2	2
1010	1	2	1	1	2	1	2
1011	1	2	1	1	2	1	2
1012	1	1	1	1	2	1	1
1013	2	1	1	1	2	2	1
1014	2	2	1	2	2	2	2
1015	1		1		2	1	2
1016	1		1		2	1	2
1017	1	1	1	1	2	1	2
1018	1	2	1	1	2	1	2
1019	1	2	1	1	2	2	2
1020	1	2	1	1	2	1	2
1021	1	2	1	1	2	1	2
1022	1	2	1	2	2	1	2
1023	1	2	1	1	2	1	2
1024	1	2	1	1	2	1	2
1025	1	1	1	1	2	1	2
1026	1	2	1	1	2	1	2
1027	1		1		2	1	1
1028	1	1	1	1	2	1	1
1029	1	2	1	1	2	1	2
1030	1	2	1	1	2	2	2
1031	1	2	1	1	2	1	2
1032	2	2	1	1	2	2	2
1033	2	2	1	1	2	2	2
1034	2	2	1	1	2	2	2
1035	2	2	1	1	2	1	2
1036	1	1	2	1	2	1	2
1037	1	1	1		2	1	2
1038	2	2	1	1	2	2	2
1039	2	2	1	2	2	2	2
1040	2	2	1	1	2	2	2
1041	1	2	1	1	2	1	2
1042	1	1	1	1	2	1	1
1043	1	2	1	1	2	1	1
1044	2	1	1	1	2	2	2

Lisa 2. Töös kasutatud tunnuste tähendused

Tõlge vene keelest

Tunnuse nimi	Seletus
ID	ankeedi järjekorra number
HIV	HIV viiruse diagnoos (1 - põeb, 0 - ei põe)
HBs	B-hepatiidi diagnoos (1 - põeb, 0 - ei põe)
HCV	C-hepatiidi diagnoos (1 - põeb, 0 - ei põe)
HBVc	
SY	
SHIV	
SHBs	
SHCV	
SHBV	
K1	Rspondendi sugu
K3	Kas olete sündinud Eestis? (jah, ei)
K5	Rahvus (1 - eestlane, 2 - venelane, 3 - muu)
K6	Mis linnas te elate praegu?(1- Tallinn, 2 - Tartu, 3 - Narva, 4 - muu linn)
K9	Tööhõive (1- täiskoormus, 2 - osaline, 3 - töötu)
K10	Kas te õpite? (1 - põhiõppe, 2 - kaugõppe, 3 - ei õppi)
K11	Mis on teie perekonnaseis? (1- abiellus, 2 - eraldi, 3 - lahutatud, 4 - lesk, 5- pole kunagi abielus olnud)
K13	Mitu inimest, teie kaasaarvates, elab selle sissetuleku peal? (vaba vastus)
SM15	Kas te olete oma elu jooksul siutsetanud 100 sigaretti? (1- ei, 2 - jah)
SM16	Kui vana te olite kui suitsestasite esimese sigaretti? (vaba vastus)
SM17	Mitu sigareti olete suitsetanud viimase 90 päeva jooksul? (vaba vastus)
A19	Kui tihti te tarbite alkohoolseid jooke? (1 - ei tarbi, 2 - 1-2 korda kuus, 3 - 1 kord nädalas, 4 - mitu korda nädalas, 5 - 1-2 korda päevas, 6 - 3 või rohkem korda päevas)
A20	Kui te joote, siis mitu jooki te tarbite päeva jooksul? (1 - 1, 2 - 2, 3 - 3-4, 5 - 5-6, 6 - 7 või rohkem)
A21	Kas te olete mõelnud alkohooli tarvitamise vähendamisele? (1 - ei, 2 - jah)
A22	Kas teid häirivad inimesed, kes kritiseerivad teied joomise pärast ? (1 - ei, 2 - jah)
A23	Kas te kunagi tundsite ennast halvasti või süüdi oma joomise pärast? (1 - ei, 2 - jah)
A24	Kas oli teil kunagi vaja juua juba hommikul selleks et rahustada närve või selleks et pead parandada? (1 - ei, 2 - jah)
D25	Kas te olete kunagi marihuaanat suitsetanund? (1 - ei, 2 - jah)
D27	Mitu korda te olete suitsetanud marihuaanat viimase 90 päeva jooksul? (1 - mitte kunagi, 2 - 1-3 korda kuus, 3 - üks kord nädalas, 4 - mitu korda nädalas, 5 - 1-2 korda päevas, 6 - 3 või rohkem korda päevas)
D28	Kas te olete proovinud kokaiini? (1 - ei, 2 - jah)
D30	Mitu korda olete tarvitanud kokaiini viimase 90 päeva jooksul? (1 - mitte kunagi, 2 - 1-3 korda kuus, 3 - üks kord nädalas, 4 - mitu korda nädalas, 5 - 1-2 korda päevas, 6 - 3 või rohkem korda päevas)
D31	Kas te olite proovinud heroiini? (1 - ei, 2 - jah)

D33	Mitu korda olete tarvinud heroiini viimase 90 päeva jooksul? (1 - mitte kunagi, 2 - 1-3 korda kuus, 3 - üks kord nädalas, 4 - mitu korda nädalas, 5 - 1-2 korda päevas, 6 - 3 või rohkem korda päevas)
D34	Mitu korda olete tarvitanud ebaseaduslikku või tänavalt ostetud metadooni viimase 90 päeva jooksul? (1 - mitte kunagi, 2 - 1-3 korda kuus, 3 - üks kord nädalas, 4 - mitu korda nädalas, 5 - 1-2 korda päevas, 6 - 3 või rohkem korda päevas)
D36	Kas te olite kunagi proovinud omakasvatud oopiumi? (1 - ei, 2 - jah)
D38	Mitu korda olete tarvitanud omakasvatatud oopiumi viimase 90 päeva jooksul? (1 - mitte kunagi, 2 - 1-3 korda kuus, 3 - üks kord nädalas, 4 - mitu korda nädalas, 5 - 1-2 korda päevas, 6 - 3 või rohkem korda päevas)
D39	Kas te olite tarvitanud muid narkootikume, mida siin mainitud? (1 - ei, 2 - jah)
I41	Kas olete ise süstinud endale narkootikume või ravimeid viimase 6 kuu jooksul? (1 - ei, 2 - jah)
I42	Kas olete kasutanud nõela narkootikumite süstimiseks viimase 90 päeva jooksul? (1 - ei, 2 - jah)
I44	Kas te jagasite oma nõelu kellegi viimase 90 päeva jooksul? (1 - ei, 2 - jah)
I47	Kas viimane kasutatud nõel oli puhas, steriilne? (1 - ei, 2 - jah)
I48	Kas te teadsite või kahtlusite viimase kasutatud nõela kohta, et see oli juba varem kasutatud kellegi teise poolt? (1 - ei, 2 - jah)
I50	Kui te kasutasite nõela viimast korda, kas te kasutasite vett, vatti või soojendit, mille kohta te teadsite või kahtlusite, et see oli juba varem kasutatud kellegi teise poolt? (1 - ei, 2 - jah)
I51	Kui te kasutasite nõelat viimast korda, kas keegi kasutas seda pärast teid? (1 - ei, 2 - jah)
I52	Kui te kasutasite nõela viimast korda, kas keegi kasutas vett, vatti või soojendit pärast teid? (1 - ei, 2 - jah)
I53	Kui te kasutasite nõela viimast korda, kas keegi kasutas oma süstalt narkootikumi pumpamiseks teie süstla siise? (1 - ei, 2 - jah)
I54	Kui te kasutasite nõelat viimast korda, kas te kasutasite oma süstalit narkootikumi pumpamiseks kellegi teise süstla sisse? (1 - ei, 2 - jah)
I55_1	Kas viimati ostetud süstitav narkootikum tuli koos süstlaga? (1 - ei, 2 - jah)
I55_2	Kas viimati ostetud süstitav narkootikum pärines otse diileri konteinerist? (1 - ei, 2 - jah)
I56_R	Kas te olete kunagi süste saanud koos venelastega? (1 - ei, 2 - jah)
I56_R1	Kui jah, siis mitme inimestega? (vaba vastus)
I56_R2	Kui tihti viimase 6 kuu jooksul? (vaba vastus)
I56_E	Kas te olete kunagi süste saanud koos eestlastega? (1 - ei, 2 - jah)
I56_E1	Kui jah, siis mitu inimestega? (vaba vastus)
I56_E2	Kui tihti viimase 6 kuu jooksul? (vaba vastus)
S62	Kas olete elanud suguelu viimase aasta jooksul? (1 - ei, 2 - jah)
S63	Mitme inimesega on teil olnud suguühe viimase aasta jooksul? (vaba vastus)
S64_1	Mitu korda on teil aset leidnud suguühe narkomaaniga viimase 90 päeva jooksul?(vaba vastus)
S64_2	Mitu korda on teil aset leidnud suguühe narkootikumite mitte tarvitava inimesega viimase 90 päeva jooksul?(vaba vastus)
S65	Kellega te seksite? (1- ainult mestega, 2 - ainult naistega, 3 - nii meste kui ka naistega)

S66R	Kas te seksisite venelastega? (1- ei, 2 - jah)
S66R_1	Mittu neid oli?(vaba vastus)
S66R_2	Kui tihti viimase 6 kuu jooksul? (vaba vastus)
S66E	Kas te seksisite eestlastega? (1- ei, 2 - jah)
S66E_1	Mittu neid oli?(vaba vastus)
S66E_2	Kui tihti viimase 6 kuu jooksul? (vaba vastus)
S660	Kas te seksisite muust rahvusest inimestega? (1- ei, 2 - jah)
S67	Kas te olete kasutanud kaitsevahendeid seksi ajal viimase 90 päeva jooksul? (1 - ei, 2 - jah)
S69_1	Mis põhjusel te seda kasutasite? (1 - raseduse vältimiseks, 2 - haiguste eest kaitsmiseks, 3 - mõlemad, 4 - muul põhjusel)
S70	Kas keegi pakkus teile raha selle eest, et seksiksite temaga? (1- ei, 2 - jah)
S71	Kas te olete maksnud kellelegi seksi eest raha? (1- ei, 2 - jah)
S72	Kas arst on rääkinud teile, et teil on mingi suguhaigus? (1- ei, 2 - jah)
S74	Kas teil oli mingi allpool toodud sümptomitest? (1- ei, 2 - jah)
S75	Kas te pöördusite arsti poole, kui avastasite sellist sümptomit? (1- ei, 2 - jah)
H76	Kuidas te hindaksite oma tervist? (1 - suurepärane, 2 - hea, 3 - keskmine, 4 - nõrk, 5 - vastamata)
H77	Kas teil on haigekassakaart? (1 - ei, 2 - jah)
H78	Kas teil oli haigekassakaart eelmisel aastal? (1 - ei, 2 - jah)
H79	Mitu korda te pöördusite arsti poole viimase aasta jooksul? (vaba vastus)
H80	Kas teile tehti mingeid operatsioone viimase aasta jooksul? (1 - ei, 2 - jah)
H81	Kas te läbisite muid protseduure viimase aasta jooksul? (1 - ei, 2 - jah)
H83_1	Kas arsti vastuvõttu ajal? (1 - ei, 2 - jah)
H83_2	Kas kiirabis? (1 - ei, 2 - jah)
H83_3	Kas haiglas? (1 - ei, 2 - jah)
H84	Kas te olite kunagi olnud vere doonoriks? (1 - ei, 2 - jah)
H86	Kas üheks põhjuseks vere andmiseks oli HIV-viiruse testimine? (1 - ei, 2 - jah)
H87	Kas keegi soovitas teile kunagi teha HIV-testi?(1 - ei, 2 - jah)
H89	Kas teile tehti kunagi muid HIV-teste, väljaarvatatud neid mis tehakse vere või palsma andmisel? (1 - ei, 2 - jah)
H92	Kas teil oli kunagi positiivne HIV-testi tulemus? (1 - ei, 2 - jah)
H93	Kui suur on tõenäosus, et te teete HIV-testi järgmise aasta jooksul? (1 - kindlasti, 2 - arvatavasti jah, 3 - arvatavasti ei, 4 - vähe tõenäoline)
H94	Kas te olite võitnud osa teistes AIDSi Tugekeskuse uuringutes? (1- ei, 2 - jah)
B95_1	Kui te esimest korda tarvitasite narkootikumi, milline alltoodud sündmustest toimus teie elus. Kas lahutus elukaaslasega? (1 - ei, 2 - jah)
B95_2	Vangistus (1 - ei, 2 - jah)
B95_3	Lähisugulase surm (mitte aabikaasa) (1 - ei, 2 - jah)
B95_4	Raske isiklik trauma või haigus (1 - ei, 2 - jah)
B95_5	Abielu või kooselu alustamine (1 - ei, 2 - jah)
B95_6	Töökoha kaotamine (1 - ei, 2 - jah)
B95_7	Ametikoha vahetamine samal töökohal või äris (1 - ei, 2 - jah)
B95_8	Materiaalsed muutused (1 - ei, 2 - jah)
B95_9	Lähisõbra surm (mitte perekonnaliige) (1 - ei, 2 - jah)
B95_10	Suur isiklik saavutus (1 - ei, 2 - jah)

B95_11	Õppe algus või lõpp (1 - ei, 2 - jah)
B95_12	Eluaseme tingimuste muutmine (1 - ei, 2 - jah)
B95_13	Isikliku harjumuste muutmine (1 - ei, 2 - jah)
B95_14	Tööaja või tingimuste muutmine (1 - ei, 2 - jah)
B95_15	Kolimine (1 - ei, 2 - jah)
B95_16	Sotsiaaltegevuse muutmine(1 - ei, 2 - jah)
B95_17	Pere kokkutuleku sageduse muutmine (1 - ei, 2 - jah)
B95_18	Väike seaduse rikumine (1 - ei, 2 - jah)
B96_1	Kas te elate samas elukohas, kus 1991. aastal? (1 - ei, 2 - jah)
B96_2	Kas te elate paremas elukohas, kui 1991. aastal? (1 - ei, 2 - jah)
B96_3	Kas te elate halvemas elukohas, kui 1991. aastal? (1 - ei, 2 - jah)
T97	Perekonna ssisetulek võrreldes 1991. aastaga (1 - jäi samaks, 2 - natukene kasvanud, 3 - suurel määral kasvas, 4 - natukene vähenes, 5 - suurel määral vähenes)
T98	Mitu korda te käisite reisil eelmisel aastal? (vaba vastus)
SYR100	Mitu süstalt te tagastate?
SYR101	Mitu neist te kasutasite ISE?
SYR102	Kas te jagasite mõnd neist süstaldest kellegagi?
SYR103	Mitme inimesega?
SYR104	Mitu süstalt te kasutasite ise ja ei andnud teistele inimestele?
SYR105	Kuidas te suhtute sellist tüüpi uuringutesse? (1 - väga hästi, 2 - pole midagi, 3 - halvasti, 4 - ei ole arvamust selle kohta)

Lisa 3. Kasutatud programmid

```
# Meetod 1. Vead on ühtlase jaotusega.

# Funktsiooni argumendid:

# tunnus - vaadeldav tunnus

# p - vigade osakaal

# Funktsiooni tulemus: vigade tõenäosus

# Kommentaarid:

# error - vigade ühe kategooria sagedus

# actual - vaadeldud väärtuste jaotus

# uus - puuduvate väärtuste mahavõtmine

vtn=function(tunnus, p){

 uus = factor(as.vector(tunnus), exclude=".")

 error=1/(length(table(uus)))

 actual=table(uus)/length(uus)

 tn=(p*error)/actual

 return(tn)}

# Meetod 2. Vead ja õiged on omavahel seotud

# Funktsiooni argumendid:

# tunnus1 ja tunnus2 - omavahel seotud tunnused

# p - vigade osakaal
```

```
# Funktsiooni tulemus: vigade tõenäosus

vtn2=function(tunnus1, tunnus2, p){

 uus1 = factor(as.vector(tunnus1), exclude=".")

 uus2 = factor(as.vector(tunnus2), exclude=".")

 tn=p*(chisq.test(table(uus1, uus2))$expected/(chisq.test(table(uus1, uus2))$observed))

return(tn)}

# Meetod 3. Naaberväärtus

#Vea tekkimise tõenäosust hindav funktsioon.

# Funktsiooni argumendid:

# tunnus - andmestiku tunnus

# p - vigade osakaal

# Funktsiooni tulemus: vigade tõenäosus

# Kommentaarid:

# X1 - algtunnus ilma puuduvate väärtuseta

#Abifunktsioon.

fnSisem=function(TransMat, X, p){

Xalg=X%*%solve(TransMat)

Xalg[Xalg<0]=0

abi=(X-(1-p)*as.vector(Xalg))/X

return(list(vaadeldud=X, oige=Xalg, TingToen=abi))}
```

```
# Põhifunktsioon

fnValim=function(tunnus, p){

X1=factor(as.vector(tunnus), exclude=".")

X=table(X1)/length(X1)

n=length(X)

mat=diag(rep(1, n))

for (i in 1:n){

if ((i-1)>0) {mat[i, i-1]=0.5*p; mat[i,i]=mat[i,i]-0.5*p}

if (i<n) {mat[i, i+1]=0.5*p; mat[i,i]=mat[i,i]-0.5*p}}

fnSisem(mat, X, p)}

# Meetod 3_1. Naaberväärtus (vaikimisi sisestava väärtuse (tühik) tõenäosusega)

#Abifunktsioon.

fnSisem=function(TransMat, X, p){

Xalg=X%*%solve(TransMat)

Xalg[Xalg<0]=0

abi=(X-(1-p)*as.vector(Xalg))/X

return(TingToen=abi)}

#Vea tekkimise tõenäosust hindav funktsioon.

# Funktsiooni argumendid:

# tunnus - andmestiku tunnus
```

```
# p - vigade osakaal

# Funktsiooni tulemus: vigade tõenäosus

# Kommentaarid:

# X1 - algtunnus ilma puuduvate väärtuseta

fnValim=function(tunnus, p){

X1=factor(tunnus, levels=(unique.default(c(".",names(table(tunnus))))))

X=table(X1)/length(X1)

n=length(X)

mat=diag(rep(1, n))

for (i in 2:n){

if ((i-1)>1) {mat[i, i-1]=0.05*p; mat[i,i]=mat[i,i]-0.05*p}

if (i<n) {mat[i, i+1]=0.05*p; mat[i,i]=mat[i,i]-0.05*p

mat[i,1]=0.9*p; mat[i,i]-0.9*p}}

fnSisem(mat, X, p)}
```

Lisa 4. Esimese meetodi pingerida

Tunnus.väärtus	Tunnuse nimi	Tinglik tõenäosus
I41.1	I41	0,810
S68_1.1	S68_1	0,810
S69_4.uudihimu	S69_4	0,810
K6.3	K6	0,540
H80.0	H80	0,540
D30.4	D30	0,405
D30.6	D30	0,405
S69_1.4	S69_1	0,405
B96_1.4	B96_1	0,405
K11.4	K11	0,324
A20.4	A20	0,324
SYR101.15	SYR101	0,324
SYR101.25	SYR101	0,324
SYR101.30	SYR101	0,324
I42.1	I42	0,270
I56_E2.6	I56_E2	0,270
S66E_1.1	S66E_1	0,270
SYR100.15	SYR100	0,270
SYR100.30	SYR100	0,270
SYR100.70	SYR100	0,270
SYR103.20	SYR103	0,231
SYR103.4	SYR103	0,231
SYR103.5	SYR103	0,231
SY.1	SY	0,203
D34.3	D34	0,203
D38.4	D38	0,203
T98.10	T98	0,203
T98.6	T98	0,203
T98.7	T98	0,203
SYR104.225	SYR104	0,203
SYR104.25	SYR104	0,203
SYR104.3	SYR104	0,203
SYR104.5	SYR104	0,203
S66E_2.3	S66E_2	0,180
S66E_2.30	S66E_2	0,180
S66E_2.50	S66E_2	0,180
S66E_2.6	S66E_2	0,180
I56_E1.1	I56_E1	0,180
K13.0	K13	0,162
SM15.1	SM15	0,162
D31.1	D31	0,162
S66R.1	S66R	0,162
S69_4.1	S69_4	0,162
HCV.2	HCV	0,135

A19.6	A19	0,135
D25.1	D25	0,135
D27.6	D27	0,135
D34.4	D34	0,135
D38.3	D38	0,135
I55_2.2	I55_2	0,135
I56_R2.12	I56_R2	0,135
I56_R2.15	I56_R2	0,135
I56_R2.9	I56_R2	0,135
I56_E2.5	I56_E2	0,135
S66R_1.3	S66R_1	0,135
SYR100.5	SYR100	0,135
H79.12	H79	0,116
H79.45	H79	0,116
H79.6	H79	0,116
H79.7	H79	0,116
SYR103.3	SYR103	0,116
A20.5	A20	0,108
K6.4	K6	0,108
SM17.3	SM17	0,108
I56_R.1	I56_R	0,101
T98.5	T98	0,101
SYR104.10	SYR104	0,101
S63.200	S63	0,095
S63.300	S63	0,095
S63.50	S63	0,095
S63.60	S63	0,095
S63.8	S63	0,095

Lisa 5. Ankeetide andmetega võrdlemiseks kasutatud andmed, esimene meetod (lõik)

ID	Tunnuse väärtus	Tunnuse nimi
1001	4	D30
1001	1	SM15
1001	6	D27
1002	4	SYR103
1002	5	I56_E2
1002	7	H79
1003	1	I42
1003	1	I56_R
1004	7	T98
1004	1	D25
1005	4	K11
1006	1	SY
1006	3	SYR103
1007	1	S68_1
1007	4	S69_1
1007	1	S69_4
1007	4	D34
1008	3	D34
1008	1	D25
1008	2	I55_2
1009	4	S69_1
1009	1	D31
1009	1	S69_4
1010	3	D34
1011	1	I42
1011	5	SYR103
1014	1	I41
1016	4	K6
1018	1	S69_4
1020	4	D38
1020	1	SM15
1021	0	H80
1023	1	S69_4
1023	3	SM17
1024	4	K6
1025	3	SM17
1026	3	SM17
1026	60	S63
1034	5	A20
1036	1	S68_1

Lisa 6. Meetod 2, korrelatsioonimaatriks (lõik)

	HIV	*HBVc*	*SHCV*	*K1*	*K5*	*A21*	*A23*	*D25*	*D30*
SHIV	0,808	0,319							
H92	0,455	0,168	0,108	0,161	0,039	0,001	0,025	0,058	0,036
S66E_1	0,237	0,332	0,044	0,237	0,404	0,110	0,034	0,000	0,064
H86	0,197	0,171	0,068	0,344	0,312	0,495	0,387	0,221	0,139
S660	0,177	0,080	0,071	0,013	0,107	0,146	0,013	0,073	0,075
B95_4	0,161	0,068	0,059	0,043	0,111	0,034	0,004	0,097	0,160
B95_12	0,145	0,119	0,020	0,072	0,021	0,016	0,051	0,027	0,102
S69_1	0,138	0,455	0,389	0,174	0,036	0,304	0,201	0,063	0,443
I52	0,127	0,113	0,057	0,005	0,236	0,084	0,076	0,031	0,016
B95_9	0,123	0,071	0,083	0,012	0,094	0,037	0,110	0,112	0,050
H93	0,116	0,054	0,085	0,101	0,244	0,009	0,055	0,045	0,214
SYR104	0,114	0,037	0,071	0,041	0,005	0,092	0,074	0,017	0,040
H84	0,104	0,121	0,039	0,092	0,124	0,245	0,258	0,050	0,107
SHBV	0,103	0,194	0,453						
D30	0,096	0,047	0,019	0,060	0,016	0,090	0,075	0,481	1,000
S66E	0,094	0,006	0,246	0,108	0,321	0,015	0,047	0,146	0,027
B95_14	0,088	0,003	0,013	0,158	0,080	0,024	0,003	0,092	0,074
B95_11	0,079	0,100	0,075	0,112	0,036	0,055	0,026	0,063	0,034
SYR103	0,078	0,057	0,134	0,071	0,050	0,032	0,127	0,018	0,015
S67	0,074	0,077	0,092	0,051	0,066	0,164	0,021	0,051	0,020
A22	0,068	0,075	0,048	0,127	0,067	0,448			
S66E_2	0,060	0,314	0,076	0,436	0,131	0,231	0,284	0,000	0,627
A24	0,056	0,072	0,148	0,193	0,008	0,223	0,404		
B95_15	0,049	0,003	0,046	0,062	0,038	0,031	0,010	0,024	0,145
S64_2	0,047	0,149	0,024	0,241	0,029	0,143	0,011	0,041	0,132
S74	0,042	0,020	0,121	0,314	0,080	0,108	0,211	0,038	0,172
B95_10	0,035	0,088	0,207	0,007	0,035	0,017	0,155	0,069	0,034
B95_7	0,026	0,093	0,023	0,069	0,005	0,001	0,050	0,090	0,017
SYR102	0,021	0,016	0,022	0,011	0,090	0,008	0,040	0,078	0,052
S75	0,020	0,057	0,234	0,404	0,081	0,257	0,215	0,086	0,139
H83_2	0,018	0,191	0,088	0,163	0,182	0,068	0,093	0,037	0,057
S70	0,017	0,160	0,012	0,649	0,073	0,177	0,304	0,174	0,087
S64_1	0,012	0,155	0,041	0,469	0,154	0,139	0,212	0,067	0,008
H79	0,012	0,045	0,101	0,227	0,037	0,098	0,040	0,015	0,141
S66R_2	0,011	0,177	0,101	0,256	0,078	0,110	0,098	0,052	0,176

Lisa 7. Teise meetodi pingerida (lõik)

Tunnus 1 nimi	Tunnus 1 väärtus	Tunnus 2 nimi	Tunnus 2 väärtus	Tõenäosus
S67	2	S69_1	2	0,098
S63	1	S64_1	2	0,086
S63	3	S66R_2	2	0,064
SYR101	4	SYR104	3	0,064
SYR103	2	SYR104	1	0,058
HIV	1	SHIV	2	0,055
SYR101	2	SYR104	5	0,054
S62	2	S67	3	0,050
HIV	2	SHIV	1	0,050
I56_R2	2	H86	2	0,047
H83_1	3	H83_2	3	0,045
I56_R2	6	H86	3	0,043
S66R_1	3	S660	1	0,043
S63	9	S75	2	0,041
A21	2	A22	3	0,040
K5	1	S66E_1	1	0,040
D33	2	H86	2	0,040
K1	3	S64_1	1	0,040
S62	3	S67	1	0,038
SHCV	2	SHVB	1	0,038
S63	2	S64_1	1	0,037
D33	6	H86	3	0,037
H84	1	H86	3	0,036
S63	3	SYR102	3	0,035
S66R_1	2	S66E	3	0,031
K1	3	S70	2	0,030
S64_2	4	S66E_2	1	0,030
S65	4	S66E_2	1	0,030
S66E_1	3	S75	2	0,030
I51	2	I52	1	0,030
S64_2	2	S66R_2	8	0,030
D25	1	D30	2	0,029
S63	2	S75	1	0,029
S63	7	S75	1	0,029
B95_12	1	B95_15	2	0,028
K1	2	S70	3	0,028
S66R_1	2	S66E	1	0,027
S63	9	S64_2	2	0,027
S66R_1	3	S66E	2	0,026
H83_1	2	H83_2	2	0,026
H86	2	B95_12	2	0,024
B95_6	1	B95_7	3	0,023
S63	6	S64_2	1	0,023
S63	6	S64_1	1	0,023
I50	3	I52	1	0,023

S66R_1	4	S66E	2	0,023
S63	10	S66E_2	1	0,023
S64_1	24	S66E_2	1	0,023
S64_2	14	S66E_2	1	0,023
S66R_2	4	S66E_2	1	0,023
S66R_1	1	S66E	3	0,022
S63	6	S64_2	2	0,022
H86	3	B95_12	1	0,022
S66R_1	2	S660	1	0,021
I56_R4	7	H88	4	0,021
S66E_2	3	S75	2	0,021
S64_2	3	S66R_2	1	0,021
I51	1	I52	2	0,020
S69_1	2	H84	2	0,020

Lisa 8. Ankeetide andmetega võrdlemiseks kasutatud andmed, teine meetod (lõik)

ID	Tunnus 1 nimi	Tunnus 1 väärtus	Tunnus 2 nimi	Tunnus 2 väärtus
1001	HIV	1	SHIV	2
1001	H83_1	3	H83_2	3
1002	HIV	1	SHIV	2
1004	H83_2	3	H83_3	3
1006	HIV	2	SHIV	1
1006	I56_R2	2	H86	2
1007	HIV	1	SHIV	2
1010	HIV	1	SHIV	2
1011	HIV	1	SHIV	2
1012	HIV	1	SHIV	2
1013	I56_R2	2	H86	2
1014	I56_R2	2	H86	2
1015	HIV	1	SHIV	2
1016	HIV	1	SHIV	2
1016	H83_3	3	H83_4	3
1017	HIV	1	SHIV	2
1018	HIV	1	SHIV	2
1019	HIV	2	SHIV	1
1019	H83_4	3	H83_5	3
1020	HIV	1	SHIV	2
1020	H83_5	3	H83_6	3
1021	HIV	1	SHIV	2
1021	H83_6	3	H83_7	3
1022	HIV	1	SHIV	2
1023	HIV	1	SHIV	2
1023	H83_7	3	H83_8	3
1024	HIV	1	SHIV	2
1024	H83_8	3	H83_9	3
1025	HIV	1	SHIV	2
1025	H83_9	3	H83_10	3
1026	HIV	1	SHIV	2
1026	H83_10	3	H83_11	3
1027	HIV	1	SHIV	2

Lisa 9. Kolmanda meetodi pingerida

Tunnus.väärtus	Tunnuse nimi	Tinglik tõenäosus
I41.1	I41	0,810
K6.3	K6	0,809
SYR100.15	SYR100	0,779
SYR101.15	SYR101	0,799
K11.4	K11	0,663
H80.0	H80	0,507
SYR104.10	SYR104	0,343
I42.1	I42	0,270
T98.10	T98	0,226
SY.1	SY	0,199
SM15.1	SM15	0,162
D31.1	D31	0,162
B96_1.4	B96_1	0,160
S66R.1	S66R	0,154
I56_R2.15	I56_R2	0,141
H79.45	H79	0,136
SYR104.0	SYR104	0,136
D25.1	D25	0,135
HCV.2	HCV	0,133
SM17.2	SM17	0,126
K13.0	K13	0,123
H79.6	H79	0,111
I55_2.2	I55_2	0,108
A20.4	A20	0,105
I56_R.1	I56_R	0,101
SYR103.1	SYR103	0,093
D34.2	D34	0,081
S66R_2.10	S66R_2	0,079
H79.15	H79	0,076
I56_R2.10	I56_R2	0,075
S63.200	S63	0,075
S66R_1.3	S66R_1	0,072
SYR102.2	SYR102	0,070
S66E_1.1	S66E_1	0,070
H79.20	H79	0,069

Lisa 10. Ankeetide andmetega võrdlemiseks kasutatud andmed, kolmas meetod (lõik)

ID	Tunnuse väärtus	Tunnuse nimi
1001	1	SM15
1001	1	SYR103
1001	2	SYR102
1002	2	SYR102
1003	1	I42
1003	1	I56_R
1003	2	D34
1004	1	D25
1005	4	K11
1006	1	SY
1006	2	SYR102
1008	1	D25
1008	0	K13
1008	1	SYR103
1009	1	D31
1009	10	I56_R2
1011	1	I42
1011	0	SYR104
1011	2	SYR102
1014	1	I41
1015	2	HCV
1019	2	HCV
1020	1	SM15
1027	2	D34
1027	15	H79
1029	1	SYR103
1032	2	SYR102
1034	0	SYR104
1035	2	D34
1036	2	HCV
1040	3	K6
1040	10	T98
1042	1	S66R

Lisa 11. Meetodi 3.1 pingerida

Tunnuse nimi.väärtus	Tinglik tõenäosus
K1..	1.000000000
I48..	1.000000000
I50..	1.000000000
I53..	1.000000000
B95_2..	1.000000000
B95_5..	1.000000000
B95_7..	1.000000000
B95_14..	1.000000000
B96_2..	1.000000000
B96_3..	1.000000000
H77..	0.883336081
B95_13..	0.856579324
SYR100..	0.718347231
H79..	0.713887770
T97..	0.687119865
B95_10..	0.669643243
H78..	0.615768514
K9..	0.579936743
B95_16..	0.575994595
B96_1..	0.479304588
SM16..	0.452527024
SYR105..	0.452524013
S74..	0.431925811
S70..	0.402196081
H92..	0.381385270
SYR101..	0.359717496
D34..	0.355261381
S72..	0.351655541
S62..	0.348200946
S71..	0.342736622
D27..	0.286214444
I56_E..	0.272751351
D33..	0.252377871
SM17..	0.234235692
H83_1..	0.208952838
S67..	0.138292405
H87..	0.125589189
S65..	0.103935445
S660..	0.098452006
SYR104..	0.096724000
I41.1	0.089279279
K6.3	0.088823235
SYR100.15	0.085850759
SYR101.15	0.085850511
S66E..	0.085344189

Lisa 12. Ankeetide andmetega võrdlemiseks kasutatud andmed, meetod 3.1 (lõik)

ID	Tunnuse väärtus	Tunnuse nimi
1001	.	H79
1001	.	SM17
1002	.	D34
1003	.	SYR105
1003	.	SM17
1003	.	S67
1003	.	S65
1003	.	S660
1003	.	SYR104
1003	.	S66E
1004	.	D27
1005	.	B95_14
1005	.	B95_13
1005	.	H83_1
1005	.	H87
1005	.	SYR104
1007	.	S70
1007	.	S72
1007	.	S71
1008	.	D27
1008	.	SM17
1008	.	S67
1008	.	S65
1008	.	S660
1008	.	S66E
1009	.	B95_2
1009	.	B95_5
1009	.	B95_10
1009	.	S74
1009	.	S70
1009	.	H92
1009	.	S72
1009	.	S71
1009	.	D27
1009	.	I56_E
1009	.	D33
1009	.	S660
1009	.	S66E
1010	.	B95_10