Tartu Ülikool
Matemaatika-informaatikateaduskond
Matemaatilise Statistika Instituut

Natalja Lepik, Imbi Traat

Tõenäosuslik valikuuring I

Tartu, 2016

Sisukord

1 Sissejuhatus

Definitsioon 1.1 ***Üldkogum (ÜK)*** *e.* ***populatsioon*** *(population) - objektide hulk, mille kohta soovitakse vastavalt püstitatud probleemülesandele saada informatsiooni.*

Definitsioon 1.2 ***Osakogum*** *(subpopulation) - üldkogumi alamhulk, mis on fikseeritud tausttunnuse või uuritava tunnuse väärtuste järgi ja mida soovitakse eraldi uurida. Osakogumi objektid on sama tüüpi, mis üldkogumi omad (pered mõlemas, isikud mõlemas vms.)*

Definitsioon 1.3 ***Kõikne uuring*** *ehk* ***loendus*** *(census) tähendab andmete kogumist üldkogumi kõikidelt objektidelt, et saada täpset informatsiooni üldkogumi kohta fikseeritud ajahetkel.*

Definitsioon 1.4 ***Registrid*** *on kirjete kogumid (andmebaasid) mitmesuguste üldkogumite kohta: rahvastikuregister, äriregister, hooneregister, vähiregister jne.*

Definitsioon 1.5 ***Valikuuring*** *(sample survey) on statistiline uuring, milles otsustused üldkogumi kohta tehakse valimi (üldkogumi ühe osa) baasil. Andmeid kogutakse üksnes valimilt.*

Andmekogumismeetodite võrdlus:

- **Kõikne uuring**
  "+": võimaldab täpset infot ÜK kohta fikseeritud ajahetkel;
  "–": töömahukas, kallis;
  "–": mahukuse tõttu kumuleeruvad vead;
  "–": mahukuse tõttu kogutakse andmeid vaid kõige olulisemate tunnuste kohta (piiratud sisu).
- **Register**
  "+": regulaarselt täpsustatud andmed aruannete põhjal;
  "–": registritest saadavad tunnused on fikseeritud registri ülesehitusega;
  "–": ÜK-le vastava ja selle muutusi piisavalt kiiresti kajastava registri loomine on pikaajaline ja töömahukas protsess.
- **Valikuuring**
  "+": väiksem maksumus;
  "+": suurem kiirus;
  "+": operatiivsus (saab korraldada vastavalt vajadustele);
  "+": laiem rakendatavus (saab kasutada intervjuusid, päevikute pidamist jne, mis on sageli raskendatud ÜK uurimisel);

"+": suurem täpsus (väiksema töömahu tõttu on võimalik kasutada kõrgema kvalifikatsiooniga tööjõudu, suurendada andmete fikseerimise täpsust, rakendada suuremat kontrolli andmete kogumisel ja töötlemisel); "−": jääb sisse juhuslik viga, mis on tingitud valimi juhuslikkusest. Seda viga on võimalik tõenäosusliku valiku korral hinnata.

Valikuuringute teooria - teadus andmete kogumisest, töötlemisest, esitamisest ja analüüsimisest. Tegevus algab probleemülesande püstitamisest ja lõpeb tulemuste publitseerimisega.

Esimene VU kursus Eestis - 1993.

Valikuuringu teooria põhieesmärgid:

- leida kõige sobivam valikudisain, mis tagaks võimalikult täpseid hinnanguid parameetritele;
- leida kõige sobivam hinnangufunktsioon parameetri hindamiseks;
- leida selline strateegia, mis minimiseeriks hinnangu dispersiooni (juhuslik viga) ja/või minimiseeriks uuringu maksumuse;
- leida meetodeid hinnangu nihke ja dispersiooni hindamiseks;
- leida meetodeid muude vigade analüüsimiseks (näiteks kadu);
- leida võimalusi lisainformatsiooni integreerimiseks (registrite info) hinnangufunktsioonidesse, et muuta hinnanguid täpsemaks, ka kooskõlaliseks muude teadaolevate näitajatega.

Valikumeetodid võivad olla tõenäosuslikud ja mitte.

Definitsioon 1.6 ***Tõenäosuslikud valikumeetodidh*** *on sellised, kus iga ÜK objekti jaoks on teada tema valimisse sattumise (kaasamise) tõenäosus. Läheb vaja ka kahe objekti koos valimisse sattumise tõenäosust, mis aga iga kord pole täpselt leitav.*

Tõenäosuslikud meetodid jagunevad:

- tagasipanekuta valik – TTA (sampling without replacement); ÜK iga objekt saab olla valitud max 1 kord
- tagasipanekuga valik – TGA (sampling with replacement); ÜK iga objekt saab olla valimis rohkem kui 1 kord

Objekti valimise viisi järgi jagatakse meetodeid ka tõmbe- või loeteluviisi valikuteks.

Definitsioon 1.7 ***Tõmbeviisi valikuks*** *nimetatakse protseduuri, mille korral mingi juhusliku katse tulemus otsustab, milline objekt üldkogumist valimisse tõmmatakse (näiteks objekti tõmbamine urnist). Tulemuseks on fikseeritud mahuga valim.*

Definitsioon 1.8 ***Loeteluviisi valikuks*** *nimetatakse protseduuri, mille korral üldkogumi iga objektio jaoks sooritatakse juhuslik katse, mille tulemus otsustab just selle objekti valimisse võtmise. Tulemuseks on enamasti juhusliku mahuga valim.*

Levinuimad tõenäosuslikud valikumeetodid:

- **Lihtne juhuslik valik (Simple Random Sampling)**. On olemas TTA ja TGA. Kõikidel objektidel on võrdne valimisse sattumise tõenäosus. Veelgi enam, kõik valimid on võrdse esinemistõenäosusega. Tulemuseks on valim etteantud mahuga n. Tõmbeviisi valik, kuid on olemas algoritm, mis teostab loeteluviisi valikut.

- **Poissoni valik**. Iga ÜK objekti valimisse kaasamine otsustatakse sõltumatult Bernoulli juhusliku suuruse abil, kusjuures igal objektil võib olla erinev kaasamistõenäosus, π_i, $0 < \pi_i < 1$. Tulemusena saadud valim on juhusliku valimimahuga, $E\mathbf{n} = \sum_U \pi_i$. Loeteluviisi valik. Juh. valimimaht suurendab hinnangute varieeruvust. Võrdsete kaasamistõenäosuste erijuhul on tegemist **Bernoulli valikuga**.

- **Süstemaatiline valik** (Systematic sampling). Objektide valimine toimub fikseeritud sammu tagant loendist, kusjuures esimene valitav objekt määratakse juhuslikult. Tulemuse täpsus sõltub objektide paigutusest loendis. Tõmbeviisi valik.

- **Suurusega võrdelise tõenäosusega valik** (Sampling with probabilities proportional to size). Objektide kaastamistõenäosused on võrdelised objektide suurustega. Saavutatakse suurte objektide üleesindatus valimis. Hinnangute arvutamisel tuleb andmeid erinevalt kaaluda. Tulemusena tõuseb mõnede hinnangute täpsus. Võib olla nii TTA kui ka TGA. Loeteluviisi valik.

- **Kihtvalik** (Stratified sampling). ÜK jagatakse osadeks ehk kihtideks. Igas kihis rakendatakse sõltumatult mingit tõenäosusliku valikumeetodit. Igas kihis arvutatakse parameetrite hinnanguid, neid sobivalt kombineerides saab leida ÜK hinnanguid. Kihid peavad olema uuritava tunnuse suhtes võimalikult homogeensed.

- **Klastervalik** (Cluster sampling). ÜK koosneb objektgruppidest ehk klastritest. Toimub klastrite juhuslik valik. Parameetrid arvutatakse igas klastris kõiki objekte kasutades.

- **Kaheastmeline valik** (Two-stage sampling). Esimesel sammul toimub klastervalik ja teisel – igas klastris toimub objektide juhuslik valik. Klastrid peavad olema võimalikult heterogeensed.

Empiirilised valikud (mittetõenäosuslikud valikumeetodid) - kaasamistõenäosusi pole teada; eesmärgiks on saada ÜK struktuuriga sarnane valim. Pole võimalik leida arvulisi täpsusnäitajaid.

Empiirilise valiku meetodid:

- **Kvootide meetod** (quota sampling). Valimi struktuur määratakse tausttunnuste järgi (sagedus)
- **Expertvalik** (Expert sampling). Subjektiivset valikut teostab ekspert.
- **Tasakaalustatud valik** (Balanced sampling). Valik on sarnane kvootide valikule, kuid valimi strukuuri määravad mite tausttunnuste osakaalud, vaid muud näitajad, nt keskmine (vanus).

Märkus 1.1 *Viimaste aastate jooksul on välja töötatud mitmeid tasakaalustatud tõenäosuslike valikumeetodeid (the cube method, Deville an Tille (2004); local pivotal method, Grafström et al. (2012, 2014) ja muud). Tulemusena saadakse juhuslik valim, milles objektide kaasamistõenäosused on teada ja mille struktuur on ÜK struktuuriga väga sarnane.*

Valikuuringu teooria areng:

- **Kiaer** (1897). Valim peab esindama üldkogumit, peab olema üldkogumiga sarnane (sobivad juhuslik valik võrdsete kaasamistõenäosustega või tasakaalustatud mittejuhuslik valik).
- **Neyman** (1934). Artikkel, mis võrdleb kiht- ja tasakaalustatud valikut. Näitas ära, et ebavõrdsete tõenäosustega valik (kihtvaliku näitel) võib kaasa tuua heade omadustega hinnanguid.
- Tänapäeva suunad: disaini-põhine teooria (õpime siin) ja mudeli-põhine teooria (on ka teisi variante).

Erinevused klassikalisest statistikast

Klassikaline statistika:

1. ÜK on lõpmatu või abstraktne. Kui ta ongi lõplik, siis valik on tagasipanekuga, mistõttu valimimaht võib ikka lõpmatult kasvada.
2. Juhusliku suuruse Y käitumine on kirjeldatud jaotusega.

3. Juhuslik suurus koos oma jaotusega annab ÜK mudeli: $Y \sim F(\theta)$. Tahame hinnata parameetrit θ.

4. Juhusliku valimi element y_i on juhusliku suuruse Y realisatsioon. Realisatsioonid pärinevad sõltumatult samast jaotusest (ssj)

5. ssj-eeldus lubab leida parameetrile hinnanguid $\hat{\theta} = \hat{\theta}(y_1, y_2, ..., y_n)$ ja uurida $\hat{\theta}$ statistilisi omadusi.

Valikuuringute teooria:

1. ÜK on reaalne, lõplik: $U = 1, 2, ..., N$.

2. TTA valiku korral ei saa valimimaht lõpmatult suureneda.

3. Tunnuse väärtusi $y_1, y_2, ..., y_N$ võib küll vaadelda diskreetse üldkogumijaotusena, kus $p(y_i) = \frac{1}{N}, \forall i$, kuid TTA valiku korral ei toimu valik igal sammul samast ÜK jaotusest.

4. Realiseerunud väärtused y_i pole ssj (va LJV TGA valiku korral).

5. Üldjuhul valim ei peegelda ÜK-t, st hindamisprobleemid vajavad teist lähenemist

6. Hinnangute omadused on määratud valikudisainiga – valimite tõenäosustega.

⇒ **Teatud mõttes on klassikaline statistika vaadeldav VU osana!** LJV TGA valim on nagu klassikalise statistika valim, kus kehtivad klassikalise statistika tulemused.

2 Valikuuringute teooria mõisted ja tähistused

Objekt, element, ühik, indiviid.
Tunnus - uuritav või taust- ehk abitunnus.
Üldkogum, populatsioon.
Osakogum - ÜK alamhulk, mis on fikseeritav tausttunnuse või uuritava tunnuse väärtuste järgi.
Loend, freim (frame) - ÜK elementide loend, mis koosneb ÜK elementidest või nende gruppidest.

Freimi abil peab olema võimalik:

- (1)... valida valimit vastavalt fikseeritud valikudisainile

- (2)... saada kontakti valitud ÜK elementidega

Eristatakse **kahte liiki ÜK-meid**:

1. **sihtkogum** (target population) - objektide hulk, mis tuleb uurida lähtuvalt statistilisest ülesandest
2. **loendile vastav ÜK** (frame population)

Aktuaalne kogum - objektide hulk, mis kuulub nii loendisse kui sihtkogumisse.
Valim - aktuaalse kogumi osahulk, mis määrataks statistilise valikumeetodiga.

Loendi võimalikud vead:

- **ülekaetus** - sisaldab ka ÜK-sse mitte-kuuluvaid elemente
- **alakaetus** - ei sisalda kõiki ÜK-sse kuuluvaid elemente
- **kordumised** - mõni ÜK-i element on kirjeldatud mitmel korral

Näide 2.1 *Ettevõtete uuringus võib juhtuda, et loend sisaldab tegevuse lõpetanud ettevõtteid. Siis on tegemist ülekaetusega. Samal ajal võivad loendist puududa äsja tegevust alustanud ettevõtted. Sel juhul on tegemist alakaetusega.*

Isekaaluv valim - ühesuguse tähtsusega objektidest koosnev valim, iga objekt valimis esindab võrdse arvu ÜK objekte. Hinnangute arvutamisel isekaaluvalt valimilt ei ole objektidele vaja omistada erinevaid kaalusid.
Kadu - valimi osa, mis mingil põhjusel jääb uuringust kõrvale.
Kao määr - kao osakaal valimist.
Vastamismäär - vastanute osakaal valimist.

2.1 Hinnatavad üldkogumi parameetrid

Olgu $U = \{1, 2, ..., N\}$ lõplik üldkogum, N ÜK maht ja y uuritav tunnus.

Sarnaselt klassikalise statistikale tähistame siingi sümboliga θ huvipakkuvat parameetrit üldkogumis. Näiteid:

- uuritava tunnuse **kogusumma** (total): $t = \sum_{i=1}^{N} y_i = \sum_U y_i$. Kui on vaja eristada kahe tunnuse kogusummasid, siis lisame indekseid, t_y või t_x. Mõnes kirjanduses kasutatakse vastavalt tähiseid Y ja X;
- ÜK **keskmine** (mean): $\bar{Y} = t_y/N = t_y \,/\, \sum_U 1$;

- osakogumi (domain) $U_d \subset U$ **osakaal**: $P_d = N_d \,/\, N$, kus N_d on osakogumi U_d maht;
- kahe kogusumma **suhe** (ratio): $R = t_y \,/\, t_x = \sum_U y_i \,/\, \sum_U x_i$.

Kõik eespool nimetatud parameetrid avalduvad summade kaudu!

Osakogumi U_d jaoks defineerime binaarse tunnuse z, kus

$$z_i = \begin{cases} 1 & , \quad \text{kui } i \in U_d; \\ 0 & , \quad \text{muidu.} \end{cases}$$

Osakogumi U_d maht N_d on tunnuse z summa:

$$N_d = t_z = \sum_U z_i = \sum_{U_d} 1,$$

ja U_d osakaal P_d ÜK-s on tunnuse z keskmine:

$$P_d = \bar{Z} = \frac{t_z}{N}.$$

Tähtis! Suurem osa valikuuringute teooriast kontsentreerub uuritava tunnuse summa hindamisele. Kui oskame hinnata t_y, siis oskame hinnata üldjuhul ka teised parameetrid, mis avalduvad summa (või summade) kaudu.

Isegi tunnuse y dispersiooni on võimalik esitada summade funktsioonina,

$$S_y^2 = \frac{1}{N-1}\sum_U (y_i - \bar{Y})^2 = \frac{1}{N-1}\left[\sum_U y_i^2 - N\bar{Y}^2\right] = \frac{1}{N-1}\left[t_{y^2} - \frac{t_y^2}{N}\right].$$

Ülesanne 2.1 *Olgu antud kaks uuritavat tunnust ÜK-s, y ja x. Esitada nendevaheline korrelatsioonikordaja uuritavate tunnuste kogusummade kaudu.*

2.2 Valikudisaini karakteristikud

Olgu $U = 1, 2, ..., N$. VU praktikas võib leida mitmeid viise valimi esitamiseks. Kirjeldame siin kolme nendest.

1) **Hulkvalim**, s - ÜK-i osahulk, $s \in U$, elementide järjestus pole tähtis. Näiteks, $s = \{2, 5, 3\}$ - korduvaid elemente pole! See on kõige levinum valimi esitusviis; kirjanduses kasutatakse enamasti seoses TTA disainidega.

2) **Järjestusvalim**, $js = \{i_1, i_2, ..., i_n\}, i_k \in U$ − valimi elemendid on esitatud elemendi võtmise järjekorras; võivad esineda ka kordused. Näiteks: $js = \{3, 1, 5, 1\}$. Kirjanduses kasutatakse seda viisi TGA disainide korral; praktikas pole TGA disainid väga levinud, kuid teooria on nende jaoks lihtsam.

3) **Vektorvalim**, $k = (k_1, k_2, ..., k_N)$, sama dimensiooniga nagu ÜK, kus k_i on objekti i valikute arv. Juhul, kui $k_i = 0$, siis objekt i pole valimis. Näiteks, $k = (1, 0, 2, 0, ..., 3)$. Valimimaht: $n = \sum_{i=1}^{N} k_i$. Sellist esitust saab kasutada nii TTA kui ka TGA disainide korral.

Definitsioon 2.1 *Juhuslikku vektorit $I = (I_1, I_2, ..., I_N)$ nimetatakse valikuvektoriks, kus I_i (valikuindikaator) näitab objekti i valikute arvu ($i \in U$). Valikuvektori realisatsiooniks on (vektor)valim k.*

Paneme tähele, et $\mathbf{n} = \sum_{i=1}^{N} I_i = \sum_U I_i$ - valimimaht, mis võib olla juhuslik.

Definitsioon 2.2 *Valikudisainiks nim. valikuvektori I jaotust:*

$$I \sim p(k),\ p(k) = P(I = k),\ \sum_k p(k) = 1.$$

Toodud definitsioon on matemaatiliseks aluseks teooria arendamisel. Meie tegeleme selles kursuses nn disainipõhise lähenemisega, kus hinnangute omadused on määratud valikudisainiga.

Praktikas räägitakse valikudisainist ka kui reeglite kogumist, kuidas valimit üldkogumist võtta. Lõppkokkuvõttes määrab aga ka reeglite kogum üheselt võimalike valimite tõenäosused ehk valikudisaini.

Disainikarakteristikud on arvud, mis kirjeldavad jaotust $p(k)$. Kõige tähtsamad jaotuse $p(k)$ karakteristikud on tema momendid:

$E(I_i)$ – esimest järku moment on objekti i oodatav valikute arv
$E(I_i I_j)$ – teist järku moment;
$V(I_i) = E(I_i)^2 - (EI_i)^2$ – valikuindikaatori I_i dispersioon;
$Cov(I_i, I_j) = \Delta_{ij} = E(I_i I_j) - E(I_i)E(I_j)$ –valikuindikaatorite kovariatsioon;
$\Delta_{ij} = 0$ Poissoni disaini korral, muidu $\Delta_{ij} \neq 0$.

Definitsioon 2.3 *Disaini nimetatakse isekaaluvaks, kui $E(I_i) = const\ \forall i$.*

Definitsioon 2.4 *Disaini nimetatakse mõõtuvaks, kui $E(I_i I_j) > 0$.*

Definitsioon 2.5 *Valikudisaini nim. tagasipanekuta disainiks, TTA, kui $I_i \in \{0, 1\}\ \forall i$, muidu tagasipanekuga, TGA.*

Definitsioon 2.6 *ÜK objekti i ($i = 1, ..., N$) kaasamistõenäosuseks π_i nimetatakse tõenäosust, millega see objekt kaasatakse valimisse antud disaini korral:*

$$\pi_i = P(i \in s) = P(I_i \geq 1) = \sum_{k, k_i \geq 1} p(k).$$

Erijuhul, kui tegemist on TTA disainiga, siis $\pi_i = P(I_i = 1)$.

Näide 2.2 *Olgu tegemist väga väikese üldkogumiga, $N = 4$, millest võetakse TTA valim mahuga $n = 2$. Võimalikud realisatsioonid on loetletud järgmises tabelis (valimi saamise tõenäosused $p(k)$ on ette antud):*

U	1.	2.	3.	4.	5.	6.	π_i
1	1	1	1	0	0	0	$\pi_1 = 0,6$
2	1	0	0	1	1	0	$\pi_2 = 0,6$
3	0	1	0	1	0	1	$\pi_3 = 0,4$
4	0	0	1	0	1	1	$\pi_4 = 0,4$
$p(k)$	0,4	0,1	0,1	0,1	0,1	0,2	$\sum_k p(k) = 1$

Valimimaht: $n = \sum_{i=1}^{4} \pi_i = 2$. Kuidas leitakse kaasamistõenäosused π_i, $i = 1, 2, 3, 4$?

Näiteks, tõenäosuse π_1 leidmiseks summeerime kõik sellised $p(k)$, milles esineb 1. objekt:

$$\pi_1 = 0,4 + 0,1 + 0,1 = 0,6.$$

Ülejäänud π_i leitakse analoogiliselt.

Definitsioon 2.7 *Objektide i, j 2. järku kaasamistõenäosuseks nimetatakse tõenäosust, millega need objektid kaasatakse korraga valimisse antud disaini korral:*

$$\pi_{ij} = P(I_i \geq 1, I_j \geq 1).$$

Ülesanne 2.2 *Olgu tegemist TTA disainiga. Tõestada, et $E(I_i I_j) = \pi_{ij}$.*

2.3 Valimimahu karakteristikud

Definitsioon 2.8 *Valikudisaini nim. fikseeritud mahuga n disainiks, kui selle disaini korral $\sum_U I_i \equiv n$.*

Üldjuhul on disaini valimimaht juhuslik suurus, $\mathbf{n} = \sum_U I_i$.

Teoreem 2.1 *Valimimahu $\mathbf{n}$ tähtsamad karakteristikud avalduvad valikudisaini momentide kaudu:*

$$E(\mathbf{n}) = \sum_i E(I_i), \tag{1}$$

$$V(\mathbf{n}) = \sum_i \sum_j \Delta_{ij}. \tag{2}$$

Tõestus. Kuna $\mathbf{n} = \sum_{i=1}^{N} I_i$ ja keskväärtusel on aditiivsuse omadus, siis (1) on tõestatud. Edasi,

$$V(\mathbf{n}) = E\Big[\sum_i I_i - \underbrace{E(\sum_i I_i)}_{\sum_i EI_i}\Big]^2 = E\Big[\sum_i (I_i - EI_i)\Big]^2 =$$
$$E\Big[\sum_i \sum_j (I_i - EI_i)(I_j - EI_j)\Big] = \sum_i \sum_j \Delta_{ij}.$$

■

Teoreem 2.2 *Fikseeritud mahuga n disaini $p(k)$ korral kehtivad seosed:*

$$\sum_i E(I_i) = n, \tag{3}$$

$$\sum_i \sum_j E(I_i I_j) = n^2, \tag{4}$$

$$\sum_i E(I_i I_j) = nE(I_j), \tag{5}$$

$$\sum_i \sum_j \Delta_{ij} = 0, \tag{6}$$

$$\sum_i \Delta_{ij} = \sum_j \Delta_{ij} = 0. \tag{7}$$

Tõestus. Fikseeritud mahuga disaini korral on $\mathbf{n} = \sum_i I_i \equiv n$ konstant, millest tulenevalt on $E(\mathbf{n}) = n$ ja $V(\mathbf{n}) = 0$, ja seega (3) ning (6) on tõestatud eelmise teoreemi põhjal;

(4) $\sum_i \sum_j E(I_i I_j) = E(\underbrace{\sum_i I_i}_{n} \underbrace{\sum_j I_j}_{n}) = n^2;$

(5) $\sum_i E(I_i I_j) = E\sum_i (I_i I_j) = EI_j \underbrace{\sum_i I_i}_{n} = nEI_j;$

(7) $\sum_i \Delta_{ij} = \sum_i E(I_i I_j) - \sum_i (EI_i)(EI_j) = nEI_j - nEI_j = 0.$

■

Ülesanne 2.3 *Olgu tegemist lihtsa juhuvalikuga tagasipanekuta (LJV TTA). Saab näidata, et sel juhul $\pi_i = \frac{n}{N}$ ja $\pi_{ij} = \frac{n}{N}\frac{n-1}{N-1}$, kus n on valimimaht ja N on ÜK maht. Kasutades LJV TTA leia suurused $E\boldsymbol{n}$ ja $V\boldsymbol{n}$ Teoreemi 2.1 valemite (1) ja (2) abil.*

Ülesanne 2.4 *Kasutades LJV TTA tõestada Teoreemi 2.2 seosed (4), (5) ja (7).*

Ülesanne 2.5 *Näidata, et fikseeritud mahuga disainide korral kehtivad järgmised seosed:*

$$\sum_{i\neq j}\sum_{i\in U}\pi_{ij} = n(n-1); \qquad \sum_{j\in U, j\neq i}\pi_{ij} = (n-1)\pi_i.$$

Ülesanne 2.6 *Üldkogum U koosneb järgmistest osakogumitest:*
$U = U_1 \cap U_2 \cap U_3$, *mille mahud on vastavalt* $N_1 = 600$, $N_2 = 300$ *ja* $N_3 = 100$ *(kogu ÜK maht on järelikult* $N = 1000$*). Iga elemendi* $i \in U$ *jaoks teostatakse Bernoulli katset vastavate tõenäosustega* $\pi_i = 0,1\, i \in U_1$, $\pi_i = 0,2\, i \in U_2$ *ja* $\pi_i = 0,8$, $i \in U_3$. *Sellisel viisil saadud valimimaht* $\boldsymbol{n}$ *on juhuslik. Leida oodatav valimimaht* $E(\boldsymbol{n})$ *ja selle dispersioon* $V(\boldsymbol{n})$.

3 TTA disainide karakteristikud ja näited

TTA disainide korral on valikuindikaator I_i Bernoulli juhuslik suurus, $I_i \sim Be(\pi_i)$ tõenäosusfunktsiooniga

$$p(k_i) = P(I_i = k_i) = \pi_i^{k_i}(1-\pi_i)^{1-k_i},\ k_i \in \{0,1\}. \tag{8}$$

Juhusliku suuruse I_i jaotuse karakteristikud on järgmised:

- $E(I_i) = \pi_i$;
- $E(I_iI_j) = P(I_i = 1, I_j = 1) = \pi_{ij}$ –teist järku kaasamistõenäosus;
- $V(I_i) = \pi_i(1-\pi_i)$;
- $\Delta_{ij} = Cov(I_i, I_j) = \pi_{ij} - \pi_i\pi_j$.
- Paneme tähele, et $\pi_{ii} = \pi_i$.

Inglise keeles nimetatakse TTA disaine WOR - Without Replacement designs.
Terve vektori $I = (I_1, I_2, ..., I_N)$ jaotust TTA disainide korral nimetatakse mitmemõõtmeliseks Bernoulli jaotuseks (MB)

Jaotuste keeles on kõik TTA disainid on MB erijuhtumid!

MB jaotust iseloomustab:

- pole üldist funktsionaalset vormi;

- on võimalik ette anda kõikvõimalike tõenäosuste tabelina:

$$P(I = k) = p(k),\ k \in \{0,1\}^N,$$
$$0 \le p(k) \le 1,\ \sum p(k) = 1;$$

- $p(k)$ jaoks on võimalik lõpmata palju erinevaid variante!
- mitmed klassikalised valikudisainid on lihtsa funktsionaalse kujuga MB jaotused.

Näiteid lihtsa funktsionaalse kujuga TTA-disainidest

1. Poissoni disain

$$\left[\begin{array}{lcl} I & = & (I_1, ..., I_N),\ I_i \perp I_j,\ i \neq j; \\ I_i & \sim & Be(\pi_i); \\ I & \sim & p(k) = P(I = k) = \prod_{i=1}^N P(I_i = k_i) \\ & & \quad = \prod_{i=1}^N \pi_i^{k_i}(1-\pi_i)^{1-k_i}. \end{array}\right.$$

Poissoni valimi genereerimise algoritm: (genereerime N korda sõltumatult Bernoulli juh. suuruseid)

$i = 1;$
$$\left[\begin{array}{l} u \sim U(0,1); \\ \text{if } u < \pi_i \text{ then } I_i = 1 \text{ else } I_i = 0; \\ i := i + 1. \end{array}\right.$$

2. Bernoulli disain

$$\left[\begin{array}{l} \text{Poissoni disaini erijuht, kus } \pi_i \equiv \pi; \\ p(k) = \pi^{|k|}(1-\pi)^{N-|k|}, \text{ kus } |k| = \sum_{i=1}^N k_i. \end{array}\right.$$

3. Lihtne juhuslik valik TTA

$$\left[\begin{array}{l} I \sim p(k), I_i \not\perp I_j; \\ p(k) = \left\{\begin{array}{ll} \frac{1}{C_N^n}, & \text{kui } |k| = n,\ C_N^n = \frac{N!}{(N-n)!n!}; \\ 0, & \text{muidu.} \end{array}\right. \end{array}\right.$$

Kõikidel valimitel mahuga n on võrdne tõenäosus olla valitud.

Valimi genereerimise võimalused (palju):

(i) Definitsiooni järgi. Loetleda kõikvõimalikud valimid mahuga n (selliseid võimalusi on C_N^n) ja siis valida üks valim võrdse tõenäosusega, näiteks urnist.

(ii) Tõmbeviis (elemendid on nummerdatud)

$\lceil$ $i = 1 \Rightarrow$ valime tn-ga $1/N$
ja eemaldame ÜK-st;

$i = 2, ..., n \Rightarrow$ valime tn-ga $1/[N - (i - 1)]$
ja eemaldame iga kord ÜK-st. $\rfloor$

Arvutis saab valitava elemendi kätte järgmise eeskirja abil:

$$i\text{.nda el. nr } = \lfloor (N - i + 1) \cdot U(0, 1) + 1 \rfloor,$$

kus $\lfloor a \rfloor$ tähistab arvu a täisosa.

(iii) Loeteluviis (tulemuseks on vektorvalim)

$\forall i = 1, ..., N$ seame vastavusse juh. arvu $u_i \sim U(0, 1)$.

$\lceil$ $i = 1:$ kui $u_1 < n/N \Rightarrow$ 1. el. on valimis

$i = 2, ..., N:$
kui $u_i < \frac{n-n_i}{N-i+1} \Rightarrow i$.s el. on valimis. $\rfloor$

Siin n_i on elementide arv, mis on valitud ÜK-i esimese $i-1$ objekti seast.

(iv) Järjestusvalik

$\lceil$ $\forall i = 1, ..., N$ seame vastavusse juh. arvud $u_1, ..., u_N, u_i \sim U(0, 1)$.
Järjestame ÜK objektid ümber u_i järgi kasvavalt: $u_{(i_1)} < u_{(i_2)} < ... < u_{(i_N)}$
Võtame valimisse esimest n objekti. $\rfloor$

NB! Saab ka kasutada suvalist pidevat jaotust!

Tegelikult on iga n elemeline komplekt, mis on võetud järjestatud failist LJV TTA valim. Seda omadust saab valikuuringutes kasutada vastamiskoormuse reguleerimiseks.

4. Tinglik Poissoni disain

Olgu $I \sim p(k) = \prod_{i=1}^{N} \pi_i^{k_i}(1 - \pi_i)^{1-k_i}$ Poissoni disain.

Tinglik Poissoni disain:

$$I^{TP} \sim P(I = k | \sum_U I_i = n) = \begin{cases} \frac{p(k)}{P(\sum_U I_i = n)}, & \text{kui } \sum_{i=1}^{N} k_i = n; \\ 0, & \text{muidu.} \end{cases}$$

Tingliku Poissoni valimi genereerimine:

$\lceil$ Teostada Poissioni valik nii nagu on kirjeldatud 1. näites.
Kui valimimaht pole n, siis jätta saadud valim kõrvale ja alustada uuesti.
Korrata nii kaua, kuni saavutatakse vajalik valimimaht. $\rfloor$

Ülesanne 3.1 *Olgu tegemist Bernoulli disainiga,* $\pi_i \equiv \pi, \forall i = 1, ..., N$. *Leida* π_{ij}.

Ülesanne 3.2 *Näita, et kui $\pi_i = \pi, \forall i$, siis on tinglik Poissoni disain TTA lihtsa juhusliku valiku disain.*

Ülesanne 3.3 *Olgu tegemist LJV TTA. Leida π_i ja π_{ij}.*

Ülesanne 3.4 *Üldkogum $U = 1, 2, 3$ koosneb kolmest elemendist. Olgu disain $p(k)$ selline, et valimid $s_1 = \{1, 2\}$, $s_2 = \{1, 3\}$, $s_3 = \{2, 3\}$ ja $s_4 = \{1, 2, 3\}$ saavad realiseeruda vastavate tõenäosustega $p(s_1) = 0,4$, $p(s_2) = 0,3$, $p(s_3) = 0,2$ ja $p(s_4) = 0,1$. Teisi võimalusi valimite jaoks pole. Leida kõik π_i, $i = 1, 2, 3$ ning kõik π_{ij}, $i, j = 1, 2, 3$. Leida keskmine valimimaht, $E(\boldsymbol{n})$.*

Ülesanne 3.5 *Üldkogum mahuga 1600 on jagatud 800-ks klastriks, mille suurused on a ($a = 1, 2, 3, 4$) järgmiselt:*

Klastri maht, a	*1*	*2*	*3*	*4*
Klastrite arv, N_a	*250*	*350*	*150*	*50*

Valikudisain on järgmine: kõigepealt valitakse 300 klastrit kasutades lihtsat juhuvalikut (TTA); seejärel valimisse sattunud klastritest küsitletakse kõiki inimesi. Olgu $\boldsymbol{n}$ valimisse sattunud inimeste arv. Leida $E(\boldsymbol{n})$ ja $V(\boldsymbol{n})$.

4 TGA disainide karakteristikud ja näited

Tagasipanekuga disainide korral objektid võivad sattuda valimisse korduvalt, seetõttu on I_i juhuslik suurus, mille realisatsioonid võivad olla hulgast $\{0, 1, 2, ..\}$. Praktikas on väga levinud kahte tüüpi TGA disainid: multinomiaal- ja hüpergeomeetriline disain

1. Multinomiaaldisain:

   - Valikutõenäosused p_i on fikseeritud iga i jaoks, $i \in U$ kogu valikuprotsessis, $\sum_{i=1}^{N} p_i = 1$.
   - Objekt valitakse vastavalt p_i-le, registreeritakse ja seejärel pannakse tagasi ÜK-sse.
   - Protsessi korratakse n korda (kuni valim on käes).

Tähistame: $I \sim M(n; p_1, p_2, ..., p_N)$, mille tõenäosusfunktsioon on järgmine:

$$p(k) = \frac{n!}{\prod_{i=1}^{N} k_i!} \prod_{i=1}^{N} p_i^{k_i}, \text{ kui } |k| = n.$$

Erijuht

Juhul, kui kõik p_i on võrdsed, $p_i \equiv \frac{1}{N}$, siis on tegemist lihtsa juhuvalikuga TGA:

$$p(k) = \frac{n!}{N^n \prod_{i=1}^{N} k_i!}, \text{ kui } |k| = n.$$

• M() disaini korral: $I_i \sim B(n, p_i)$, st et objekt i saab olla valitud $k_i = 1, ..., n$ korda.

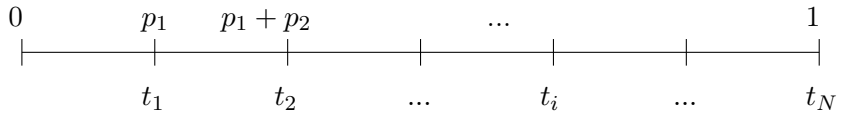
Ülesanne 4.1 *Kirjutada välja $E(I_i)$, $V(I_i)$ ja $Cov(I_i, I_j)$ multinomiaalse disaini korral.*

Multinomiaaldisaini genereerimine

Kasutame nn kumulatiivsete summade meetodit...

Moodustada kum. summad:
$t_i = \sum_{j=1}^{i} p_j, \ i = 1, ..., N.$

Need summad asuvad lõigul $[0, 1]$:



Genereerida $u \leftarrow U(0, 1)$

Kui $u \in (t_{i-1}, t_i,]$ siis element i on valimis.

Korrata protseduuri n korda.

2. Hüpergeomeetriline disain

Iga element saab olla valitud kuni m_i korda:

Olgu $m = \sum_{i=1}^{N} m_i$.

Tähistame $I \sim HG(n; m_1, m_2, ..., m_N)$, kus jaotuse tõenäosusfunktsioon on järgmine:

$$p(k) = P(I = k) = \frac{\prod_{i=1}^{N} C_{m_i}^{k_i}}{C_m^n}, \text{ kui } |k| = n.$$

Erijuht. Kui kõik $m_i \equiv 1$, siis annab HG disain lihtsa juhuvaliku tagasipanekuta:

$$p(k) = \frac{1}{C_N^n}.$$

Ülesanne 4.2 *Olgu $N = 4$, $E\boldsymbol{n} = 3$. Panna kirja kõikvõimalikud valimid ja nende saamise tõenäosused järgmisse tabelisse. Poissoni disaini korral anna ise erinevad kaasamistõenäosused π_i objektidele, nii et $\sum_U \pi_i = 3$:*

k	**LJV TTA**	**LJV TGA**	**Bernoulli**	**Poisson**	**TP**
0000	p(0000)=?				
0001					
...					
0003					
...					
3000					
p(k)					

Ülesanne 4.3 *Lihtsa juhusliku valiku korral tagasipanekuga on kõikide objektide valikutõenäosused võrdsed, $p_i = \frac{1}{N}, \forall i \in U$. Näidata, et 1. järku kaasamistõenäosused on samuti võrdsed ja on esitatavad järgmise valemi abil:*

$$\pi_i = 1 - \left(1 - \frac{1}{N}\right)^n, \forall i \in U$$

kus n on valimimaht.

Ülesanne 4.4 *Eelmise ülesande jätk. Näidata, et lihtsa juhusliku valiku korral TGA teist järku kaasamistõenäosused on*

$$\pi_{ij} = 1 - 2\left(1 - \frac{1}{N}\right)^n + \left(1 - \frac{2}{N}\right)^n, \forall i \neq j \in U.$$

5 Hindamise alused

Olgu $U = \{1, 2, \ldots, N\}$ üldkogum, ja θ üldkogumi parameeter
y_i – uuritav tunnus, mõõdetud objektil $i \in U$.
Oleme huvitatud peamiselt selliste parameetrite θ hindamisest nagu kogusumma $t_y = \sum_U y_i$ või kogusummade suhe $R = \frac{t_y}{t_x}$.

Ka keskmised avalduvad suhetena.

Oleme huvitatud hinnangute $\hat{\theta}$ omadustest, sellistest nagu nihe, dispersioon, dispersiooni hinnang.

Tuletame meelde: Hinnang $\hat{\theta}$ on parameetri θ jaoks nihketa, kui $E\hat{\theta} = \theta$.

Kuna juhuslikkus tekitatakse hinnangusse valikudisaini poolt, siis on keskväärtus defineeritud valikudisaini suhtes:

$$E\hat{\theta} = \sum_k \hat{\theta}(k)p(k),$$

kus summeerimine on üle kõigi võimalike valimite $\mathbf{k}$. Lähenemist, kus hinnangu keskväärtust, ja sellest tulenevalt ka dispersiooni, defineeritakse valikudisaini abil, nimetatakse disainipõhiseks lähenemiseks. On olemas ka mudelipõhine lähenemine. Mõelgem, kuidas intuitiivselt mõista $\hat{\theta}$ disainipõhist keskväärtust ja dispersiooni?

Ülesanne 5.1 *Panna kirja hinnangu $\hat{\theta}$ dispersiooni valem.*

5.1 Valikuuringu andmed

Mida teame enne ja mida pärast valiku teostamist ja andmete kogumist?

- Enne valiku teostamist teame:
  $U = (1, \ldots, N)$ – objektide märgendid (id-kood, nimi, aadress freimis);
  $x = (x_1, \ldots, x_N)$ – abitunnused, mis on teada iga i jaoks või mille kogusummad t_x on teada;
  $I = (I_1, \ldots, I_N) \sim p(k)$ – valikudisain, fikseeritakse planeerimisfaasis, $p(k)$ või selle karakteristikud on teada, I realisatsioon pole teada.

  Ei tea
  $y = (y_1, \ldots, y_N)$ – uuritava tunnuse väärtusi.

  Praktikas on y tegelikult maatriks, milles iga objekti jaoks on veerg paljude tunnuste väärtustega.

$$y_i = \begin{pmatrix} z_i \\ u_i \\ \vdots \end{pmatrix}, \quad i \in U.$$

- Pärast valikut ja andmete kogumist teame:
  $U = (1, \ldots, N)$;
  $x = (x_1, \ldots, x_N)$;
  $I = (I_1, \ldots, I_N)$ realisatsiooni, valimit;
  $y_s = (I_1 y_1, \ldots, I_N y_N)$ objektide mõõtmistulemusi valimis, y_i võib olla vektor, nii nagu ka x_i.

Suurused (I, y_s, x) moodustavad valikuuringute andmete stohhastilise esituse. Siin on ilmutatud kujul näha juhuslikkust põhjustav vektor I, $I \sim p(k)$. Selline esitus on aluseks disainipõhisele valikuteooriale

Mudelipõhine valikuteooria eeldab, et $y = (y_1, \ldots, y_N)$ ise on juhuslik, juba üldkogumis. Tema juhuslikku olemust iseloomustatakse mudeliga, näiteks $y_i \sim N(\mu, \sigma^2)$, sõltumatud.

Kolmiku (I, y_s, x) funktsiooni nimetame statistikuks.

6 Kogusumma nihketa hindamine

Selleks et konstrueerida lihtsaimat hinnangut kogusummale t, vaadelgem lihtsat statistikut, nimelt andmete

$$y_s = (I_1 y_1, \ldots, \underbrace{I_i y_i}_{Y_{si}}, \ldots, I_N y_N)$$

lineaarkombinatsiooni

$$\hat{t} = \sum_U c_i y_{si} = \sum_U c_i I_i y_i,$$

kus c_i on mittejuhuslik konstant y_{si} on objekti i vaatlustulemus.

Tahame, et

$$E\Big[\sum_U c_i I_i y_i\Big] = \sum_U c_i y_i E(I_i) = \sum_U y_i \qquad \text{– nihketus.}$$

Et seos kehtiks, peab olema

$$c_i = \frac{1}{E(I_i)}.$$

Järelikult saame kogusumma nihketa hinnaguks

$$\hat{t} = \sum_U \frac{I_i y_i}{E(I_i)}. \tag{9}$$

Valikudisainide nõue $E(I_i) > 0$ on nüüd selge. Antud hinnangul on kaks tähtsat esitust:

$$\hat{t} = \sum_U I_i \breve{y}_i, \tag{10}$$

kus

$$\breve{y}_i = \frac{y_i}{E(I_i)} - \text{ laiendatud } y_i,$$

ja teiseks,

$$\hat{t} = \sum_U \omega_i y_i, \tag{11}$$

kus

$$\omega_i = \frac{I_i}{E(I_i)} - y_i \text{ valikukaal .}$$

Pangem tähele, et $\omega_i = 0$ mittevalitud objektide jaoks ja ta kaalub üles valitud objekte. Valimiväärtus y_i esindab ω_i objekti üldkogumis (tavaliselt $\omega_i >> 1$).

Näeme valitud objektide erinevat panust hinnangusse. Need, mille valik on oodatavalt suurem ($E(I_i)$ suurem), surutakse maha väiksema kaaluga ω_i.

Pange tähele ja pidage meeles, et kuigi summeerimine toimub üle U, esitavad valemid (9)-(11) ikkagi valimisummasid, ja seega on arvutatavad andmetelt. Realiseerunud valimi korral on need summad tegelikult üle valimi s:

$$\hat{t} = \sum_s I_i \breve{y}_i \quad \text{või} \quad \hat{t} = \sum_s \omega_i y_i,$$

kus I_i ja ω_i on realiseerunud väärtused ja s loendab üldkogumist U valitud erinevaid objekte:

$$s = \{i : i \in U, \text{ mille korral } I_i > 0\}.$$

Teoreetilises käsitluses eelistame selles konspektis summasid üle U.

Hinnangu (9) dispersiooniavaldise saame kasutades tuntud seost tõenäosusteooriast:

$$V\Big[\sum_{i=1}^N c_i X_i\Big] = \sum_{i=1}^N \sum_{j=1}^N c_i c_j Cov(X_i, X_j), \quad X_i - \text{juhuslik suurus.} \tag{12}$$

Seoses on lihtne veenduda, kasutades definitsioone $V(X) = E(X - EX)^2, Cov(X, Y) = E(X - EX)(Y - EY)$.

Rakendades seost (12) meie hinnangule $\hat{t} = \sum_U I_i \breve{y}_i$, saame

$$V(\hat{t}) = \sum_{i=1}^N \sum_{j=1}^N \Delta_{ij} \breve{y}_i \breve{y}_j, \tag{13}$$

kus $\Delta_{ij} = Cov(I_i, I_j)$. Dispersioon (13) on hinangu $\hat{t}$ disainipõhine dispersioon Ta on hinnangu $\hat{t}$ varieeruvuse mõõt antud valikudisaini $p(k)$ korral. $V(\hat{t})$ valemis (13) on teoreetiline avaldis, ta ei ole arvutatav.

Vajame suuruse $V(\hat{t})$ hinnangut.

Hinnangus ei saa kasutada üldkogumiväärtusi $\breve{y}_i$. Saab kasutada valimiväärtusi $I_i \breve{y}_i$.

Kirjutagem (13) valimiväärtuste kaudu ja lisagem tundmatud konstandid c_{ij}:

$$\hat{V}(\hat{t}) = \sum_{i=1}^N \sum_{j=1}^N c_{ij} \Delta_{ij} I_i \breve{y}_i I_j \breve{y}_j.$$

Konstandid c_{ij} määrame nihketuse nõudest:

$$E\left[\hat{V}(\hat{t})\right] = V(\hat{t}), \tag{14}$$

$$E\left[\hat{V}(\hat{t})\right] = \sum \sum\nolimits_U c_{ij} \Delta_{ij} \breve{y}_i \breve{y}_j E(I_i I_j).$$

On selge, et tingimus (14) kehtib, kui valime $c_{ij} = \frac{1}{E(I_i I_j)}$.

Seetõttu on $V(\hat{t})$ nihketa hinnanguks

$$\hat{V}(\hat{t}) = \sum \sum_U \breve{\Delta}_{ij} \breve{y}_i \breve{y}_j I_i I_j,$$

kus $\breve{\Delta}_{ij} = \frac{\Delta_{ij}}{E(I_i I_j)}$ on valikumuutujate laiendatud kovariatsioon. Dispersiooni hinnangu saame alternatiivselt esitada valikukaalude abil

$$\hat{V}(\hat{t}) = \sum \sum_U \breve{\Delta}_{ij} \omega_i y_i \omega_j y_j.$$

Märgime veelkord: kui summa üle U sisaldab valikumuutujaid I_i või valikukaalu ω_i, toimub summeerimine tegelikult üle valimi s, ja seega valem esitab hinnangut.

Saadud tähtsad tulemused on koondatud järgmisse teoreemi.

Teoreem 6.1 (Üldine hindamisteoreem) *Üldkogumi kogusumma $t = \sum_U y_i$ nihketa hinnang on*

$$\hat{t} = \sum_U I_i \breve{y}_i \quad (\textit{või } \hat{t} = \sum_U \omega_i y_i), \tag{15}$$

kus

$$\breve{y}_i = \frac{y_i}{E(I_i)} \textit{ ja } \omega_i = \frac{I_i}{E(I_i)}. \tag{16}$$

Selle disainipõhine dispersioon on

$$V(\hat{t}) = \sum \sum_U \Delta_{ij} \breve{y}_i \breve{y}_j, \tag{17}$$

kus $\Delta_{ij} = Cov(I_i, I_j)$. Dispersiooni nihketa hinnanguks $E(I_i I_j) > 0$ korral on

$$\hat{V}(\hat{t}) = \sum \sum_U \breve{\Delta}_{ij} \breve{y}_i \breve{y}_j I_i I_j \quad (\textit{või } \hat{V}(\hat{t}) = \sum \sum_U \breve{\Delta}_{ij} \omega_i y_i \omega_j y_j), \tag{18}$$

kus

$$\breve{\Delta}_{ij} = \frac{\Delta_{ij}}{E(I_i I_j)}.$$

Märkus 6.1 *Üldine hindamisteoreem kehtib iga valikudisaini korral, nii TTA kui TGA disainide korral. Vaja on vaid teada disainikarakteristikuid*

$$E(I_j),\ E(I_i I_j),\ \Delta_{ij} \quad \textit{for } i = j \textit{ ja } i \neq j.$$

Märkus 6.2 *Maatriksite abil saab kahekordsed summad sageli elegantsemalt esitada. Nii saame dispersiooni jaoks avaldise:*

$$V(\hat{t}) = \breve{y}' \Delta \breve{y}, \tag{19}$$

kus $\Delta = (\Delta_{ij}) : \ N \times N$ *and* $\breve{y} = (\breve{y}_i) : \ N \times 1$. *Sarnaselt saame dispersiooni hinnangu jaoks*

$$\hat{V}(\hat{t}) = \breve{y}_s' \breve{\Delta} \breve{y}_s, \tag{20}$$

kus $\breve{\Delta} = (\breve{\Delta}_{ij}) : \ N \times N$ *ja* $\breve{y}_s = (\breve{y}_i I_i) : \ N \times 1$. *Kuna mittevalitud elemendid vektoris* $\breve{y}_s$ *on nullid, siis saavutavad ruutvormi (20) komponendid väiksema dimensiooni* $\Delta : n \times n$ *ja* $\breve{y}_s : \ n \times 1$, *kus* n *on valimimaht.*

Seda maatriksesitust kasutame hiljem IML programmis.

Ülesanne 6.1 *Tuletada maatriks* Δ *ja* $\breve{\Delta}$ *Bernoulli disaini jaoks.*

Ülesanne 6.2 *Tuletada maatriks* Δ *ja* $\breve{\Delta}$ *lihtsa juhuvaliku tagasipanekuta korral.*

6.1 Dispersiooni hindamine fikseeritud mahuga disainide korral

Tuletame meelde, et hinnangu dispersioon on tema hajuvuse mõõt. Kui valikudisain on fikseeritud, siis hinnangu dispersioon on teatav arvuline konstant (tavaliselt küll tundmatu). Samas selle ühe konstandi jaoks saab konstrueerida mitmeid hinnanguid. Lisaks eespool toodud üldisele hinnangule vaatame siin teist hinnangut, mis kehtib üksnes fikseeritud mahuga disainide korral.

Olgu disain $p(k)$ fikseeritud valimimahuga – $\sum_U I_i \equiv n$.

Teoreem 6.2 *Fikseeritud mahuga disaini* $p(k)$ *korral saab hinnangu* $\hat{t} = \sum_U I_i \breve{y}_i$ *dispersiooni esitada alternatiivsel kujul*

$$V(\hat{t}) = -\frac{1}{2} \sum \sum_U \Delta_{ij} (\breve{y}_i - \breve{y}_j)^2, \tag{21}$$

ja eeldusel, et $E(I_i I_j) > 0 \ \forall i \neq j \in U$, *on dispersiooni* $V(\hat{t})$ *nihketa hinnanguks*

$$\hat{V}(\hat{t}) = -\frac{1}{2} \sum \sum_U I_i I_j \breve{\Delta}_{ij} (\breve{y}_i - \breve{y}_j)^2. \tag{22}$$

Tõestus. Näitame, et (21) on ekvivalentne seosele (16). Avame sulud seoses (21):

$$V(\hat{t}) = -\frac{1}{2} \sum_i \sum_j \Delta_{ij} (\breve{y}_i^2 - 2\breve{y}_i \breve{y}_j + \breve{y}_j^2).$$

Nüüd $\sum_i \sum_j \Delta_{ij} \breve{y}_i^2 = \sum_i \breve{y}_i^2 \underbrace{\sum_j \Delta_{ij}}_{0} = 0.$

Samuti $\sum_i \sum_j \Delta_{ij} \breve{y}_i^2 = 0$, ja saame

$$V(\hat{t}) = \sum_{i=1}^{N} \sum_{j=1}^{N} \Delta_{ij} \breve{y}_i \breve{y}_j,$$

mis on (16) .

On lihtne näha, et (22) on nihketa suuruse (21) jaoks.

■

Märkus 6.3 Avaldis (16) annab hinnangu $\hat{t}$ dispersiooni kõigi valikudisainide jaoks. Avaldis (21) annab selle üksnes fikseeritud mahuga disainide jaoks, millisel juhul ta on võrdne seoses (21) antuga. Aga NB! dispersiooni hinnang (22) ei ole üldjuhul võrdne hinnanguga (17) , isegi mitte fikseeritud mahuga disainide korral.

Märkus 6.4 *Kui valikudisain on fikseeritud mahuga, siis eelistatakse dispersioonihinnangut (22), kuna see on stabiilsem (tal on väiksem varieeruvus üle erinevate valimite) ja üldjuhul ei tule ta negatiivne ($\Delta_{ij} < 0$ enemuse praktikas kasutatavate fikseeritud mahuga disainide korral).*

Märkus 6.5 *Dispersioonihinnangut (22) nimetatakse Sen–Yates–Grundy hinnanguks.*

6.2 Nihketa hinnang kogusummale multinomiaaldisaini korral

Vaatleme tähtsaimat ebavõrdsete tõenäosustega TGA disaini − multinomiaaldisaini:

$$I \sim M(n; p_1, p_2, ..., p_N), \ \sum_U p_i = 1, \ I_i \sim B(n, p_i).$$

Sel juhul

$$\begin{array}{ll} E(I_i) & = np_i; \\ \Delta_{ii} & = V(I_i) = np_i(1-p_i); \\ \Delta_{ij} & = Cov(I_i, I_j) = -np_ip_j; \\ E(I_iI_j) & = n(n-1)p_ip_j; \text{ Näidata!} \\ E(I_i^2) & = np_i(1-p_i+np_i); \text{ Näidata!} \\ w_i = \frac{I_i}{np_i}. & \end{array}$$

Hinnangufunktsioon saab järgmist kuju:

$$\hat{t} = \sum_U \frac{I_i y_i}{np_i}.$$

Seda tüüpi hinnangut nimetatakse Hansen- Hurwitz hinnanguks ja ka $p-$hinnanguks.

Teoreem 6.3 *(Hindamisteoreem multinomiaaldisaini korral) Multinomiaaldisaini korral on nihketa hinnang ÜK kogusummale $t = \sum_U y_i$ järgmine:*

$$\hat{t} = \sum_U \frac{I_i y_i}{np_i}. \tag{23}$$

Hinnangu $\hat{t}$ dispersioon on:

$$V(\hat{t}) = \frac{1}{n}\left[\sum_U \frac{y_i^2}{p_i} - t^2\right]. \tag{24}$$

Dispersiooni kaks nihketa hinnangut on:

$$\hat{V}(\hat{t}) = \frac{1}{n-1}\left[\sum_U \frac{n}{1-p_i+np_i}\left(\frac{I_i y_i}{np_i}\right)^2 - \hat{t}^2\right], \tag{25}$$

$$\hat{V}(\hat{t}) = \frac{1}{n(n-1)}\left[\sum_U I_i \frac{y_i^2}{p_i^2} - n\hat{t}^2\right]. \tag{26}$$

Märkus 6.6 *. Dispersioonihinnang (25) järeldub Üldisest hindamisteoreemist. Dispersioonihinnang (26) järeldub Teoreemist fikseeritud mahuga disainide kohta. Viimast hinnangut nimetatakse Sen-Yates-Grundy (SYG) dispersioonihinnanguks. Ta on lihtsama kujuga ja ta on ka stabiilsem kui dispersioonihinnang (25).*

Ülesanne 6.3 *Tõestada Teoreemi 6.3 kõik väited.*

Märkus 6.7 *Juhul, kui multinomiaaldisaini korral*

$$y_i = cp_i,\ i = 1, ..., N,$$

siis ka $V(\hat{t}) = 0$. St et kui uuritava tunnuse väärtused on võrdelised tõenäosustega p_i, siis hinnang $\hat{t}$ annab täpse ÜK summa.

Praktikas pole võimalik valimi võtmisel kasutada selliseid tõenäosusi p_i, seda enam, et uuritavaid tunnusei on palju. Kui osatakse määrata sellised p_i, mis on ligikaudu võrdelised väärtustega y_i, siis saavutatakse väiksem dispersioon vastavale kogusumma hinnangule.

Üritatakse leida selline tausttunnust x, mis on teada kõigi üldkogumi objektide kohta ja mis on positiivselt tugevasti korrileeritud uuritava tunnusega y. Selle abil määratakse

$$p_i = \frac{x_i}{\sum_U x_i};\ i = 1, ..., N.$$

Valikuuringutes on selliseks tausttunnuseks sageli objekti suuruse tunnus, mistõttu multinomiaaldisaini nimetatakse ka suurusega võrdeliste tõenäosusega disainiks (pps ehk probability proportional-to-size sampling).

Ülesanne 6.4 *Tõestada Märkuse 6.7 väide, et kui* $y_i = cp_i,\ i = 1, ..., N$, *siis* $V(\hat{t}) = 0$.

Ülesanne 6.5 *Näidata, et multinomiaaldisaini korral saame hinnangu dispersiooni avaldada ka järgmisel kujul:*

$$V(\hat{t}) = \frac{1}{n} \sum_U \left(\frac{y_i}{p_i} - t \right)^2 p_i.$$

Ülesanne 6.6 *Näidata, et Sen-Yates-Grundy dispersioonihinnangu alternatiivne kuju on*

$$\hat{V}(\hat{t}) = \frac{1}{n(n-1)} \sum_U \left(\frac{y_i}{p_i} - \hat{t} \right)^2 I_i.$$

Ülesanne 6.7 *Olgu tegemist LJV TTA. Tõesta, et dispersioonide hinnangud ÜHT järgi (18) ja SYG teoreemi järgi (22) langevad sel juhul kokku. (NB! Teiste disainide korral see nii ei pruugi olla.)*

7 Üldkogumi keskmise nihketa hindamine

ÜK keskmine, täpsemalt keskmine objekti kohta, on defineeritud järgmiselt:

$$\bar{Y} = \frac{1}{N} \sum_U y_i = \frac{t_y}{N}.$$

Kui N on teada, saab sellele anda nihketa hinnangu.

(1) N on teada => piisab kogusumma hindamisest:

$$\hat{\bar{Y}} = \frac{\hat{t}_y}{N}. \tag{27}$$

Dispersioon ja dispersioonihinnang järelduvad teadaolevates tulemustest $\hat{t}_y$ kohta:

$$\begin{aligned} V(\hat{\bar{Y}}) &= \frac{1}{N^2} V(\hat{t}_y), \\ \hat{V}(\hat{\bar{Y}}) &= \frac{1}{N^2} \hat{V}(\hat{t}_y). \end{aligned}$$

(2) N pole teada => saab kasutada alternatiivset hinnangut:

$$\hat{\bar{Y}}_{alt} = \frac{\hat{t}_y}{\hat{N}},\ \text{kus}\ \hat{N} = \sum_U w_i,\ w_i = \frac{I_i}{E(I_i)}. \tag{28}$$

Paneme tähele, et nimetajas on disainikaalude summa, mis on nihketa hinnanguks üldkogumi mahule N. Hinnangu $\hat{N}$ omadused tulenevad Üldisest hinadamisteoreemist erijuhul, kui $y_i \equiv 1$. Siis $N = \sum_U 1$ ja $\hat{N} = \sum_U w_i 1$.

Keskmise alternatiivse hinnangu dispersiooni kohe leida ei saa, sest tegemist on kahe juhusliku suuruse suhtega. Dispersioonivalemiteni jõuame hiljem, kui vaatame suhte hindamist üldjuhul.

Märkus 7.1 *Isegi kui N on teada, eelistatakse hinnangut (28) hinnangule (27), kuna üldjuhul on (28) väiksema varieeruvusega. Hinnang (28) annab väikese nihkega tulemuse, kuid see nihe on vähe oluline võrreldes dispersiooniga.*

Märkus 7.2 *Ka kogusumma $\hat{t}_y$ hindamisel eelistatakse sageli järgmist hinnangut, mida nimetatakse kogusumma suhtehinnanguks ja mis on regressioonhinnangu erijuht:*

$$\hat{t}_{y,alt} = \hat{\bar{Y}}_{alt} \cdot N = \frac{N}{\hat{N}} \hat{t}_y.$$

Näeme, et see kogusumma hinnang vajab lisainformatsiooni ehk siin N teadmist.

8 Kahe nihketa hinnangu kovariatsioon

Vaatame kahte kogusummat t_y ja t_x. Olgu vastavad nihketa hinnangud

$$\hat{t}_y = \sum_U I_i \breve{y}_i \text{ ja } \hat{t}_x = \sum_U I_i \breve{x}_i,$$

kus $\breve{y}_i = y_i / EI_i$ ja $\breve{x}_i = x_i / EI_i$. Siis kovariatsiooni definitsiooni järgi:

$$Cov(\hat{t}_y, \hat{t}_x) = E\left[(\hat{t}_y - t_y)(\hat{t}_x - t_x)\right].$$

Paneme tähele, et

$$\hat{t}_y - t_y = \sum_U I_i \breve{y}_i - \sum_U y_i = \sum_U I_i \breve{y}_i - \sum_U E(I_i) \breve{y}_i = \sum_U \left[I_i - E(I_i)\right] \breve{y}_i,$$

samuti

$$\hat{t}_x - t_x = \sum_U \left[I_i - E(I_i)\right] \breve{x}_i.$$

Kasutades seost

$$\left(\sum_{i=1}^N a_i\right)\left(\sum_{i=1}^N b_i\right) = \sum_{i=1}^N \sum_{j=1}^N a_i b_j$$

saame,

$$Cov(\hat{t}_y, \hat{t}_x) = E\left[\sum_i \sum_j (I_i - EI_i)\breve{y}_i(I_j - EI_j)\breve{x}_j\right] = \sum_i \sum_j \underbrace{E(I_i - EI_i)(I_j - EI_j)}_{Cov(I_i, I_j) = \Delta_{ij}} \breve{y}_i \breve{x}_j.$$

Järelikult,

$$Cov(\hat{t}_y, \hat{t}_x) = \sum_{i=1}^{N} \sum_{j=1}^{N} \Delta_{ij} \breve{y}_i \breve{x}_j$$

nihketa hinnanguga:

$$\hat{Cov}(\hat{t}_y, \hat{t}_x) = \sum_{i=1}^{N} \sum_{j=1}^{N} \breve{\Delta}_{ij} \breve{y}_i \breve{x}_j I_i I_j,$$

kus endiselt $\breve{\Delta}_{ij} = \Delta_{ij}/E(I_i I_j)$.

Ülesanne 8.1 *Tuletada $Cov(\hat{t}_y, \hat{t}_x)$ Bernoulli disaini korral.*

Ülesanne 8.2 *Kirjutada välja valem lineaarse korrelatsioonikordaja $Corr(\hat{t}_y, \hat{t}_x)$ jaoks.*

9 Suhte hinnang

Olgu uuritavaks parameetriks kahe tunnuse kogusummade jagatis

$$R = \frac{t_y}{t_x}.$$

Näiteks soovime uurida pere kulutuste osakaalu meelelahutusele. Parameetri R hinnangu leidmiseks peame hindama t_y ja t_x:

$$\hat{R} = \frac{\hat{t}_y}{\hat{t}_x}.$$

See hinnang on kahe juhusliku suuruse mittelineaarne funktsioon ja tema statistiliste omaduste täpne tuletamine ei ole lihtne.

Ligikaudseks tuletamiseks kasutatakse lineariseerimistehnikat Taylori reaks arenduse abil.

9.1 Taylori rida kahe argumendi korral

Olgu X_1, X_2 juhuslikud suurused ja $g(X_1, X_2)$ nedne funkstsioon. Funktsiooni $g(.,.)$ Taylori rea lineaarosa punkti (a_1, a_2) ümbruses on järgmine:

$$g(X_1, X_2) \approx g(a_1, a_2) + \left.\frac{\partial g}{\partial X_1}\right|_{(a_1,a_2)} (X_1 - a_1) + \left.\frac{\partial g}{\partial X_2}\right|_{(a_1,a_2)} (X_2 - a_2),$$

kus $\left.\frac{\partial g}{\partial X_1}\right|_{(a_1,a_2)}$ on g osatuletis punktis (a_1, a_2).

Saadud osatuletised pole enam juhuslikud suurused ja seega saadud avaldis on lineaarne funktsioon X_1 ja X_2 suhtes.

9.2 Suhte hinnangu Taylori rittaarendus

Arendame $\hat{R} = \frac{\hat{t}_y}{\hat{t}_x} = g(\hat{t}_y, \hat{t}_x)$ Taylori ritta punkti (t_y, t_x) ümbruses.

$$\left.\frac{\partial \hat{R}}{\partial \hat{t}_y}\right|_{(t_y,t_x)} = \frac{1}{t_x},$$

$$\left.\frac{\partial \hat{R}}{\partial \hat{t}_x}\right|_{(t_y,t_x)} = -\frac{t_y}{t_x^2}.$$

Seega,

$$\hat{R} \approx \frac{t_y}{t_x} + \frac{1}{t_x}(\hat{t}_y - t_y) - \frac{t_y}{t_x^2}(\hat{t}_x - t_x) = R + \frac{1}{t_x}(\hat{t}_y - R\hat{t}_x).$$

Veendume, et saadud $\hat{R}$ on ligikaudu nihketa:

$$E(\hat{R}) \approx R + \frac{1}{t_x}[\underbrace{E(\hat{t}_y)}_{t_y} - R\underbrace{E(\hat{t}_x)}_{t_x}] = R.$$

Leiame dispersiooni:

$$V(\hat{R}) = \frac{1}{t_x^2} V(\hat{t}_y - R\hat{t}_x) = \frac{1}{t_x^2}\left[V(\hat{t}_y) + R^2 V(\hat{t}_x) - 2RCov(\hat{t}_y, \hat{t}_x)\right],$$

kus kasutame lineaarkombinatsiooni dispersioonivalemit:

$$V(aX + bY) = a^2 V(X) + b^2 V(Y) + 2abCov(X, Y).$$

Selleks, et saada dispersioonile $V(\hat{R})$ hinnangut, võime kasutada nihketa hinnanguid $\hat{t}_x, \hat{t}_y$ ja $\hat{V}(\hat{t}_x), \hat{V}(\hat{t}_y)$ üldisest hindamisteoreemist. Teame ka kovariatsiooni nihketa hinnangut:

$$\hat{V}(\hat{R}) = \frac{1}{\hat{t}_x^2}\left[\hat{V}(\hat{t}_y) + \hat{R}^2\hat{V}(\hat{t}_x) - 2\hat{R}\hat{Cov}(\hat{t}_y, \hat{t}_x)\right].$$

Suhte hinnangu dispersiooni leidmiseks on olemas ka alternatiivne valem. Alternatiivne valem võimaldab suhte hinnangut käsitleda varasema teooria valguses (ÜHT jm tulemused). Vajalikuks osutub sobiva uue tunnuse defineerimine.

Alternatiivseks esituseks kirjutame Taylori rea lineaarliikme veel kord välja:

$$\hat{R} \approx R + \frac{1}{t_x}(\hat{t}_y - R\hat{t}_x) = R + \frac{1}{t_x}\left[\sum_U I_i \breve{y}_i - R\sum_U I_i \breve{x}_i\right] = R + \frac{1}{t_x}\sum_U I_i(\breve{y}_i - R\breve{x}_i). \quad (29)$$

Viimane saadud summa on nihketa hinnang kogusummale

$$\sum_U (y_i - Rx_i)$$

ja me saame kasutada ÜHT dispersioonivalemi saamiseks. Võttes kasutusele uue tunnuse

$$u_i = y_i - Rx_i, \quad (30)$$

saame suhte hinnangu esitada kujul

$$\hat{R} \approx R + \frac{1}{t_x}\sum_U I_i \breve{u}_i,$$

millest saame

$$V(\hat{R}) \approx \frac{1}{t_x^2}\sum_{i=1}^N \sum_{j=1}^N \Delta_{ij}\breve{u}_i\breve{u}_j. \quad (31)$$

Dispersiooni hinnang on vastavalt

$$\hat{V}(\hat{R}) = \frac{1}{\hat{t}_x^2}\sum_{i=1}^N \sum_{j=1}^N \breve{\Delta}_{ij}\breve{\hat{u}}_i\breve{\hat{u}}_j I_i I_j, \quad (32)$$

kus

$$\hat{u}_i = y_i - \hat{R}x_i, \quad (33)$$
$$\breve{\hat{u}}_i = \breve{y}_i - \hat{R}\breve{x}_i. \quad (34)$$

Kaalude abil on dispersioonihinnangu valem:

$$\hat{V}(\hat{R}) = \frac{1}{\hat{t}_x^2}\sum_{i=1}^N \sum_{j=1}^N \breve{\Delta}_{ij} w_i\hat{u}_i w_j\hat{u}_j.$$

Paneme tähele, et $\hat{t}_y - R\hat{t}_x$ valemis (29) hindab nulli, kuna $0 = t_y - \frac{t_y}{t_x}t_x = t_y - Rt_x$. Järelikult hinnang $\hat{t}_y - R\hat{t}_x$ varieerub nulli ümbruses. Kui me jagame selle hinnangu t_x-ga, siis hinnangu varieeruvus muutub veelgi väiksemaks. Seega on suhte R hinnangu dispersioon väike, mistõttu on $\hat{R}$ hea hinnang.

Ülesanne 9.1 *Olgu tegemist Bernoulli disainiga. Hinnatavaks parameetriks on suhe* $R = \frac{t_y}{t_x}$. *Tuletada, et sel juhul*

$$V(\hat{R}) = \frac{1-\pi}{\pi t_x^2} \sum_U (y_i - Rx_i)^2,$$

kus $\pi = n/N$.

10 Hindamine osakogumis

Definitsioon 10.1 ***Osakogumiks*** *nimetatakse üldkogumi* U *alamhulka* U_d, $U_d \subset U$.

Osakogumi mahtu tähistatkse N_d. Osakogumi objektid on sama tüüpi nagu üldkogumi omad. Näiteks, kui üldkogumi objektideks on pered, siis ka osakogum on teatava tunnuse alusel määratud pered (mitte isikud, lapsed vms.)

Mõned osakogumi näited:

1. välisosalusega ettevõtted kõigi ettevõtete hulgas Eestis,
2. töötud riigi tööealiste elanike hulgas,
3. suitsetajad kopsuvähi haigete hulgas.

Osakogumid määratakse mingi tunnuse (osakogumi identifikaatori) järgi. Kui huvipakkuvad osakogumid on enne uuringu läbiviimist teada ja fikseeritud, siis on otstarbekas käsitleda osakogumeid kihtidena ja teostada sobiva kihtvalikut.

Sageli aga tekib olukord, kui huvipakkuvad osakogumid tekkivad hiljem, siis kui valim on juba võetud üldkogumist tervikuna. Nii võib juhtuda, et kui osakogumi maht N_d on väike, siis osakogumist U_d satub valimisse vähe objekte ja arvutatavad osakogumi hinnangud on väga väikese täpsusega.

Öeldakse, et osakogum on väike, kui valimimaht selles on väike (näiteks kuni 10, isegi 0). Selliste osakogumite jaoks on omad hindamismeetodid (*Small area estimation methods*). Need kasutavad uuritava tunnuse modelleerimismeetodeid, et kompenseerida valimi väiksust. Neid meetodeid me valikuteooria baaskursuses ei vaata.

Valimimaht ja osakogumi valimimaht avalduvad valikuindikaatorite abil järgmiste summadena,

$$\mathbf{n} = \sum_U I_i, \quad \mathbf{n}_d = \sum_{U_d} I_i.$$

Näeme, et isegi kui $\mathbf{n}$ on fikseeritud, jääb $\mathbf{n}_d$ juhuslikuks.

Huvi pakuvad järgmised parameetrid:

$$\begin{array}{lll} N_d & - & \text{osakogumi maht,} \\ P_d = \frac{N_d}{N} & - & \text{osakogumi osakaal,} \\ t_d = \sum_{U_d} y_i & - & \text{osakogumi kogusumma,} \\ \bar{Y}_d = \frac{1}{N_d} t_d & - & \text{osakogumi keskmine,} \\ R_d = \frac{\sum_{U_d} y_i}{\sum_{U_d} x_i} & - & \text{suhe osakogumis.} \end{array}$$

Võtame kasutusele indikaatortunnuse z, mis näitab kuuluvust osakogumisse:

$$z_i = \begin{cases} 1, & i \in U_d, \\ 0, & \text{vastasel juhul.} \end{cases}$$

Nüüd saame osakogumi mahu kirja panna ÜK summana,

$$N_d = \sum_U z_i,$$

mistõttu saame rakendada üldist hindamisteoreemi ÜK summade hindamiseks:

$$\hat{N}_d = \sum_U I_i \breve{z}_i, \text{ kus } \breve{z}_i = z_i / E(I_i),$$

ehk

$$\hat{N}_d = \sum_U w_i z_i, \text{ kus } w_i = I_i / E(I_i).$$

Ka $\hat{N}_d$ dispersioon ja dispersioonihinnang tulevad üldisest hindamisteoreemist. Tunnus y_i tuleb vaid asendada indikaatortunnusega z_i.

Hinnang $\hat{N}_d$, kuigi kirja pandud üldkogumi summana, on tegelikult valimisumma. Tähistame valimi (hulkvalimi) seda osa, mis kuulub osakogumisse s_d, st

$$s_d = s \bigcap U_d.$$

Näeme, et nendes tähistustes

$$\hat{N}_d = \sum_U w_i z_i = \sum_s w_i z_i = \sum_{s_d} w_i,$$

ja seega hinnatud osakogumi maht on kaalude summa üle osavalimi s_d.

Osakogumi y-tunnuse summa $t_d = \sum_{U_d} y_i$ hindamiseks loome uue tunnuse y':

$$y'_i = z_i y_i = \begin{cases} y_i, & i \in U_d, \\ 0, & \text{vastasel juhul.} \end{cases}$$

Nüüd saame osakogumi summa kirja panna uue tunnuse ÜK summana:

$$t_d = \sum_U y'_i.$$

Selle summa nihketa hinnang

$$\hat{t}_d = \sum_U w_i y'_i$$

ja tema dispersiooni hinnang tulevad jälle üldisest hindamisteoreemist. Valemites tuleb kasutada vaid uut tunnust y'_i.

Kui teame osakogumi mahtu N_d, siis osakogumi keskmise hinnang on lihtsalt

$$\hat{\bar{Y}}_d = \frac{1}{N_d}\hat{t}_d,$$

mille dispersioon tuleb sellest, et teame $\hat{t}_d$ dispersiooni (ÜHT).

Nägime, et osakogumi mahu ja summa hindamisel saame kasutada üldist hindamisteoreemi.

Kui aga osakogumimaht pole teada, siis saab keskmise hinnang järgmise kuju

$$\hat{\bar{Y}}_d = \frac{\hat{t}_d}{\hat{N}_d}, \tag{35}$$

mis on kahe hinnangu suhe (suhte hinnang) ja seda tüüpi hinnangu dispersiooni leidmist vaatasime eelmises punktis.

Osakogumite korral huvitutakse ka kahe summa suhtest üldisemal kujul,

$$R_d = \frac{\sum_{U_d} y_i}{\sum_{U_d} x_i},$$

mille saab jällegi esitada üldkogumi U tasemel

$$R_d = \frac{\sum_U y_i z_i}{\sum_U x_i z_i} = \frac{\sum_U y'_i}{\sum_U x'_i}.$$

Nüüd saame nii R_d hinnangu kui ka dispersioonihinnangu taandada juba olemasolevatele valemitele. Näiteks,

$$\hat{R}_d = \frac{\sum_U w_i y'_i}{\sum_U w_i x'_i}, \text{ kus } w_i = I_i/E(I_i)$$

Märkus 10.1 *Isegi kui teame osakogumi mahtu N_d, on soovitav kasutada osakogumi keskmise hindamisel hinnangut (35), kuna see on väiksema varieeruvusega. Sellest järelduvalt on kogusumma hindamiseks soovitav kasutada*

$$\hat{t}_d = N_d\hat{\bar{Y}}_d = N_d\frac{\hat{t}_d}{\hat{N}_d}.$$

Ülesanne 10.1 *Olgu $U_d \subset U$. Kirjutada välja ÜHT valemid Bernoulli disaini korral parameetri $t_d = \sum_{U_d} y_i$ jaoks.*

11 Hinnangu täpsus

Olgu θ meid huvitav parameeter ja $\hat{\theta}$ on selle nihketa hinnang.

$\hat{\theta} - \theta$ on viga, mille võivad põhjustada järgmised komponendid:

- **valikuviga** (sampling error): valimi juhuslikusest põhjustatud viga; seda on võimalik hinnata, kui on teada valikudisain $p(k)$ või tema karakteristikud, ja hinnangu $\hat{\theta}$ avaldis
- **muu viga**: kaost/mittevastamisest
  põhjustatud viga, inervjueeriast põhjustatud viga, valesti sõnastatud ankeedist.... Seda laadi viga on raske hinnata, kuid on võimalik määrata vea suundumust (üle/alahinnang).

Uuringu üheks kvaliteedinäitajaks on vastamismäär, $\frac{\text{vastanute arv}}{n}$. Eesti Statistikaameti leibkonnauuringud (Household Budget Surveys) toimuvad regulaarselt kord kuus alates 1995. aastast. Vastamismäär on ligikaudu 50%. Erinevates riikide erinevate uuringute vastamismäär võib olla $20-80\%$. Kaost põhjustatud nihke vähendamiseks kasutatakse tänapäeval mitmesuguseid kalibreerimismeetodeid. Need vajavad lisainformatsiooni oma konstruktsioonis. Kalibreerimishinnanguid antud kursuses ei vaata.

Valikuviga iseloomustavad järgmised suurused:

- $\sqrt{\hat{V}(\hat{\theta})}$ − $\hat{\theta}$ standardhälbe hinnang, standardviga (standard error);
- $\lambda_{\alpha/2}\sqrt{\hat{V}(\hat{\theta})}$ − pool usaldusvahemiku pikkust;
- $CV(\hat{\theta}) = \frac{\sqrt{\hat{V}(\hat{\theta})}}{\hat{\theta}}$ - suhteline viga (relative error, coefficient of variation)
- $\frac{\lambda_{\alpha/2}\sqrt{\hat{V}(\hat{\theta})}}{\hat{\theta}}$ − alternatiivne suhteline viga
- $MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = V(\hat{\theta}) + B^2$, kus $B = E(\hat{\theta}) - \theta$ − keskmine ruutviga.
- $\hat{MSE}(\hat{\theta})$ − keskmise ruutvea hinnang.

Valikuviga ja seega ka hinnangu täpsus sõltub:

- valikudisainist (sealhulgas valimimahust, $n = \sum_U I_i$)
- hinnangufunktsioonist

Valimimahu suurendamine suurendab hinnangu täpsust!

11.1 Valimimahu määramine

Valimimaht n määratakse vastavalt tellija poolt nõutavale hinnangu täpsusele.

Näiteks, (1) hinnangu suhteline viga ei tohi ületada 2%, st

$$\frac{\sqrt{\hat{V}(\hat{t}_y)}}{\hat{t}_y} \leq 0.02 \text{ ehk } \hat{V}(\hat{t}_y) \leq (0.02\hat{t}_y)^2.$$

Nendest võrratustest saab määrata valimimahu, kui teame $\hat{V}(\hat{t}_y)$ ja $\hat{t}_y$ hinnangulisi väärtusi, näiteks eelnevast uuringust või taustuuringust, ja vastavaid valemeid sõltuvalt valimimahust n.

(2) ILO (International Labour Organization) nõuab tööjõu uuringute läbiviimisel, et kasutatav disain oleks selline, et osakogumites, mis moodustavad 5% ÜK-st, ei ületaks hinnangu standardviga 6% hinnangust. Teiste sõnadega, osakogumites, mille maht $N_d = 0.05N$ on nõue hinnangutele järgmine:

$$\sqrt{\hat{V}(\hat{t}_d)} \leq 0.06\hat{t}_d.$$

Selline täpsus nõuab väga suurt valimimahtu.

11.2 Disainiefekt

Peale valimimahtu, sõltub hinnangu täpsus ka valitud disainist. Oluline mõiste valikuuringust on disainiefekt:

- aitab võrrelda disaine hinnangute täpsuse seisukohalt;
- suur praktiline väärtus komplitseeritud disainide korral, mil pole võimalik leida hinnangute dispersioonivalemeid. Sel juhul disainiefekti ligikaudne teadmine (nt. eelnevate uuringute kogemustest) võimaldab hinnata parameeterhinnangute dispersioone.

Definitsioon 11.1 *Valikudisaini $p(s)$* ***disainiefekt*** *on suhe*

$$Deff_{p(s)}(\hat{t}_y) = \frac{V_{p(s)}(\hat{t}_y)}{V_{LJV}(\hat{t}_y)},$$

kus $V_{p(s)}(\hat{t}_y)$ on huvipakkuva hinnangu $\hat{t}_y$ teoreetiline dispersioon vaadeldavas disainis ja $V_{LJV}(\hat{t}_y)$ on teoreetiline dispersioon lihtsa juhuvaliku tagasipanekuta korral.

LJV TTA on võetud disainiks, mille suhtes võrreldakse teisi disaine, kuna see on teoreetiliselt hästi läbitöötatud ja praktikas sageli kasutatav disain.

Ülesanne 11.1 *Lihtsa juhuvaliku TTA korral on tõestatud, et $V(\hat{t}) = \frac{N^2(1-n/N)S_y^2}{n}$. Kumb disainidest on sama N ja n korral efektiivsem - kas LJV TTA ($\pi_i = n/N$) või Bernoulli ($\pi_i = n/N$)? Eeldame, et N on nii suur, et $\frac{N}{N-1} \approx 1$.*

12 Hindamine lihtsa juhuvaliku korral, TTA

Olgu $I = (I_1, .., I_N)$ disaini vektor lihtsa juhuvaliku korral ning valimi maht olgu n. Sellisel juhul:

$$I \sim p(k) = \begin{cases} (C_N^n)^{-1}, & \text{kui } |k| = n; \\ 0, & \text{vastasel juhul.} \end{cases}$$

Tänu oma lihtsusele, LJV ja ka hinnangud selle disaini korral on väga hästi uuritud.

Kasutusvaldkond:

- Sageli on LJV osa mingist keerulisemast disainist (nt. kaheastmeline disain, kus 1.-l astmel valitakse vastavalt LJV-le suurimad objektid ehk klastrid (majad, tänavad, osariigid jne). Ja teisel astmel igas klastris rakendatakse oma (kõige sobivam) disain.
- Valemid, mis on välja töötatud LJV jaoks võivad olla rakendatud lähendina teiste disainide korral, näiteks hinnangu dispersiooni valem süstmaatilise valiku korral.

12.1 LJV disainikarakteristikud

Kõigepealt, esimest järku kaasamistõenäosus: $\pi_i = Pr(I_i = 1) =?$ Sündmus $I_i = 1$ toimub kui vektor $I = (I_1, ..., I_i, ..., I_N)$ saab realisatsiooniks $k_i = 1$, $k = (k_1, ..., 1, ..., k_N)$. Kuna soovime, et valimimaht oleks n, siis selliseid võimalike realisatsioone on C_{N-1}^{n-1}. Seega, saame

$$\pi_i = Pr(I_i = 1) = \sum_{k, k_i=1, |k|=n} p(k) = C_{N-1}^{n-1}(C_N^n)^{-1} = \frac{n}{N}.$$

Suhet n/N nimetatakse sageli valikusuhteks (sampling fraction) ja tähistatakse f-ga.

Analoogiliselt saame avaldise 2.-st järku kaasamistõenäosusele:

$$\pi_{ij} = Pr(I_i = 1, I_j = 1) = \sum_{k, |k|=n, k_i=k_j=1} p(k) = C_{N-2}^{n-2}\,(C_N^n)^{-1} = \frac{n(n-1)}{N(N-1)}.$$

Samuti läheb edaspidi vaja valikuindikaatorite dispersiooni ja kovariatsiooni:

$$\begin{aligned}
\Delta_{ii} &= V(I_i) \overbrace{=}^{def} E(I_i^2) - (EI_i)^2 = ... \\
E(I_i^2) &= 1 \cdot Pr(I_i^2 = 1) + 0 \cdot Pr(I_i^2 = 0) = Pr(I_i = 1) = \pi_i \\
... &= \pi_i - \pi_i^2 = \pi_i(1 - \pi_i) = \frac{n}{N}\left(1 - \frac{n}{N}\right) \\
\Delta_{ij} &= Cov(I_i\,, I_j) \overbrace{=}^{def} \underbrace{E(I_i I_j)}_{\pi_{ij}} - E(I_i)E(I_j) = \pi_{ij} - \pi_i\pi_j = \\
&= \frac{n(n-1)}{N(N-1)} - \left(\frac{n}{N}\right)^2 = \frac{n}{N}\left(\frac{n-1}{N-1} - \frac{n}{N}\right) = \\
&= \frac{n}{N}\frac{Nn - N - Nn + n}{N(N-1)} = -\frac{n}{N}\frac{N-n}{N}\frac{1}{N-1}.
\end{aligned}$$

Kokkuvõttes saame LJV disainikarakteristikuid koondada järgmisesse tabelisse:

$$\begin{aligned}
f = \frac{n}{N} &\quad - \quad \text{valikusuhe;} \\
\pi_i = f &\quad - \quad \text{esimest järku kaasamistõenäosus;} \\
\pi_{ij} = f\frac{n-1}{N-1}, i \neq j &\quad - \quad \text{teist järku kaasamistõenäous;} \\
\Delta_{ii} = f(1-f) &\quad - \quad I_i \text{ dispersioon;} \\
\Delta_{ij} = -f(1-f)\frac{1}{N-1} &\quad - \quad I_i, I_j \text{ kovariatsioon.}
\end{aligned}$$

Ülesanne 12.1 *LJV TTA korral iga n-elemendilise valimi saamise tõenäosus on $1/C_N^n$, kusjuures i.nda objekti kaasamistõenäosus on n/N. Kuid vastupidine väide ei kehti, ehk kui objektidel kaasamistõenäosuseks on n/N, siis see veel ei tähenda, et tegu on LJV TTA. Mõtle välja valimidisain, mis kinnitab seda.*

12.2 Hinnang kogusummale LJV korral

Meid huvitab parameeter $t = \sum_U y_i$. Üldisest hindamisteoreemist saame:

$$\hat{t} = \sum_U \frac{I_i y_i}{EI_i} = \sum_U \frac{I_i y_i}{\pi_i} = \frac{N}{n}\sum_U I_i y_i,$$

millele vastab järgmine tavapärane kuju:

$$\hat{t} = \frac{N}{n}\sum_s y_i = N\bar{y}.$$

Sellel hinnangul on olemas kaks tõlgendust:

$$\hat{t} = \begin{cases} N\bar{y} & - \text{ valimikeskmine esindab kõiki väärtuseid ÜK-st;} \\ \sum_s w_i y_i & - \text{ iga valimiäärtus } y_i \text{ esindab } w_i = N/n \text{ väärtust ÜK-st.} \end{cases}$$

Järgmisena võiksime leida saadud hinnangu disersiooni ja dispersiooni hinnangu kasutades ÜHT. Kuid topeltsummad dispersiooni avaldises võivad osutuda aeganõudvaks suurte andmestikke korral.

Kuna LJV on fikseeritud maguga disain, siis saame kasutada alternatiivset valemit, mida leidsime punktis 3.1.3 (teoreem 5).

$$\begin{aligned} V(\hat{t})_{alt} &= -\frac{1}{2}\sum_{i=1}^{N}\sum_{i=1}^{N}\Delta_{ij}\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \overbrace{=}^{\Delta_{ij}=-f(1-f)\frac{1}{N-1}} \\ &= \frac{1}{2}f(1-f)\frac{1}{f^2}\frac{1}{N-1}\sum_i\sum_j(y_i - y_j)^2. \end{aligned}$$

Lisades $\pm\bar{Y}$ ja avaldades ruutu, paneme tähele, et

$$\sum_i\sum_j(y_i - \bar{Y})^2 = N\sum_{i=1}^{N}(y_i - \bar{Y})^2,$$

$$\sum_{i=1}^{N}(y_i - \bar{Y}) = \sum_{i=1}^{N} y_i - \underbrace{N\bar{Y}}_{\sum_{i=1}^{N} y_i} = 0.$$

Järelikult,

$$\sum_i\sum_j(y_i - y_j)^2 = 2N\sum_{i=1}^{N}(y_i - \bar{Y})^2.$$

Lõplikult saame,

$$V(\hat{t}) = \frac{1}{2}(1-f)\frac{1}{f}2N\underbrace{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2}_{S_y^2} = N^2(1-f)\frac{S_y^2}{n},$$

kus S_y^2 on tunnuse y üldkogumi dispersioon. NB! Pole enam topltsummasid!

Analoogiliselt dispresiooniga saame lihtsustada ka dispersiooni hinnangut:

$$\hat{V}(\hat{t})_{alt} = -\frac{1}{2}\sum\sum_{i,j\in s}\underbrace{\frac{\Delta_{ij}}{\pi_{ij}}}_{-\frac{1-f}{n-1}}\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 = \frac{1-f}{2(n-1)f^2}\sum\sum_{i,j\in s}(y_i - y_j)^2.$$

Nüüd saame kasutada, et

$$\sum \sum_{i,j \in s} (y_i - \bar{y})^2 = n \sum_{i \in s} (y_i - \bar{y})^2,$$

$$\sum_{i \in s} (y_i - \bar{y}) = \sum_s y_i - n\bar{y} = 0.$$

Ja lõplikult saame:

$$\hat{V}(\hat{t}) = \frac{1-f}{2(n-1)f^2} 2n \sum_s (y_i - \bar{y})^2 = \tag{36}$$

$$= N^2(1-f)\frac{1}{n} \underbrace{\frac{1}{n-1} \sum_s (y_i - \bar{y})^2}_{s_y^2} = N^2(1-f)\frac{s_y^2}{n}, \tag{37}$$

kus s_y^2 on tunnuse y valimi dispersioon.

Teoreem 12.1 *Lihtsa juhuvaliku TTA korral nihketa hinnang ÜK summale* $t = \sum_U y_i$ *avaldub järgmiselt:*

$$\hat{t} = \frac{N}{n} \sum_U I_i y_i = \frac{N}{n} \sum_s y_i,$$

ehk alternatiivselt

$$\hat{t} = N\bar{y}.$$

Hinnangu dispersioon on järgmine:

$$V(\hat{t}) = N^2(1-f)\frac{S_y^2}{n}$$

ja dispersiooni hinnang:

$$\hat{V}(\hat{t}) = N^2(1-f)\frac{s_y^2}{n},$$

kus

$f = \frac{n}{N}$ *on valikusuhe,*
$\bar{y} = \frac{1}{n}\sum_s y_i$ *valimikeskmine,*
$S_y^2 = \frac{1}{N-1}\sum_U (y_i - \bar{Y})^2$ *tunnuse* $\underline{y}$ *ÜK dispersioon,*
$s_y^2 = \frac{1}{n-1}\sum_s (y_i - \bar{y})^2$ *tunnuse* $\underline{y}$ *valimidispersioon.*

Järeldus 12.1 *Hinnangud ÜK keskmisele $\bar{Y} = t/N$ on järgmised:*

$$\begin{aligned} \hat{\bar{Y}} &= \bar{y}, \\ V(\hat{\bar{Y}}) &= (1-f)S_y^2/n, \\ \hat{V}(\hat{\bar{Y}}) &= (1-f)s_y^2/n. \end{aligned}$$

Võrdle klassikalise statistika valemitega!
Klassikalises statistikas on y_i sõltumatud sama jaotusega valimis ning

$$V(\hat{\bar{Y}}) = \frac{S_y^2}{n}, \text{ ja } \hat{V}(\hat{\bar{Y}}) = \frac{s_y^2}{n}.$$

LJV TTA korral kui me eemaldame ühe y_i ÜK-st, siis saame hoopis teise ÜK jaotuse. Järgmise elemendi valik toimub juba teise jaotuse järgi. Valikud on omavahel negatiivselt korreleeritud:

$$\Delta_{ij} = -f(1-f)\frac{1}{N-1}.$$

Seepärast ka keskmise hiinangu dispersioonil olemas nn "lõpliku ÜK-i korrigeerimiskordaja" $(1-f)$.
Kui valikusuhe on väike, siis ka valemid on ligikaudselt võrdsed klaasikalise statistika omadega.
Jäta meelde! Klassika statistika tarkvara pakettid töötavad eeldusel, et y_i on sõltumatute samast jaotuseset. Seepärast neid ei saa otseselt kasutada valikuuringute andmetele. Mida keerulisem on disain, seda rohkem erinevad hinnangute dispersioonid.

SUDAAN
WESVAR
CLAN
SAS 8.1
R, pakett SAMPLING } Spetsiaalne tarkvara valikuuringute teostamiseks.

Ülesanne 12.2 *LJV TTA valim mahuga $n = 30$ leibkonda oli võetud asulast A eesmärgiga hinnata väikelaste arvu (y) selles asulas. Asulas elab kokku 284 leibkonda. Valimist sai leitud järgmised näitajad: $\sum_s y_i = 42$ ja $\sum_s y_i^2 = 76$. Arvutada ligikaudne vahemikhinnang I_{t_y} väikelaste koguarvule usaldusnivool 95%.*

Ülesanne 12.3 *LJV TTA korral on vahemikhinnang keskmisele $\bar{Y}$ usaldusnivool $1-\alpha$ järgmine:*

$$I_{\bar{Y}} = \hat{\bar{Y}} \pm 1,96\sqrt{\hat{V}(\hat{\bar{Y}})} = \bar{y} \pm 1,96\sqrt{\frac{(1-f)s_y^2}{n}} = \bar{y} \cdot \left(1 \pm cv_{\bar{y}} \cdot 1,96\sqrt{\frac{1-f}{n}}\right) = \bar{y}(1 \pm A),$$

kus $cv_{\bar{y}} = s_y/\bar{y}$ - variatsioonikordaja (coefficient of variation).
Uuringu tellija soovib, et $A \leq 3\%(0,03)$. Leida uuringu jaoks valimimaht n, kui eelmisest uuringust on teada, et
a) $cv_{\bar{y}} = 0,5$; b) $cv_{\bar{y}} = 1$; c) $cv_{\bar{y}} = 1,5$. Ning $N = 100000$.

12.3 Kovariatsioon kahe hinnangu vahel LJV TTA korral

Olgu meil kaks hinnangut $\hat{t}_y = N\bar{y}$ ja $\hat{t}_x = N\bar{x}$. Üldiselt TTA disainide jaoks avaldub kovariatsioon kahe hinnangu vahel järgmiselt (vt punkt 3.3):

$$Cov(\hat{t}_y, \hat{t}_x) = \sum_i \sum_j \Delta_{ij} \frac{y_i}{\pi_i} \frac{x_j}{\pi_j} = \sum_i \Delta_{ii} \frac{y_i x_i}{\pi_i^2} + \sum \sum_{i \neq j} \frac{y_i}{\pi_i} \frac{x_j}{\pi_j}.$$

LJV TTA korral:

$$\begin{aligned} Cov(\hat{t}_y, \hat{t}_x) &= \frac{N^2}{n^2} \left[f(1-f) \sum_i y_i x_i - f(1-f) \frac{1}{N-1} \sum \sum_{i \neq j} y_i x_j \right] = \\ &= \frac{N}{n}(1-f) \frac{1}{N-1} [(N-1) \sum_i x_i y_i - \underbrace{\sum \sum_{i \neq j} y_i x_j}_{-[(\sum_i y_i)(\sum_i x_i) - \sum_i y_i x_i]}] = \\ &= \frac{N}{n}(1-f) \frac{1}{N-1} \left[N \sum_i y_i x_i - t_y t_x \right] \\ &= N^2(1-f) \underbrace{\frac{1}{N-1} \left[\sum_i y_i x_i - N \bar{Y} \bar{X} \right]}_{S_{yx}} / n. \end{aligned}$$

Lõplikult saame:

$$Cov(\hat{t}_y, \hat{t}_x) = N^2 Cov(\hat{\bar{Y}}, \hat{\bar{X}}) = N^2(1-f) S_{yx}/n,$$

kus

$$S_{yx} = \frac{1}{N-1} \sum_U (y_i - \bar{Y})(x_i - \bar{X})$$

on tunnuste y ja x kovariatsioon ÜK-s.

Punktist 8 saab ka tuletada hinnangu kovariatsioonile:

$$\hat{Cov}(\hat{t}_y, \hat{t}_x) = \sum \sum_{i,j \in s} \frac{\Delta_{ij}}{\pi_{ij}} \frac{y_i}{\pi_i} \frac{x_i}{\pi_i} \overbrace{=}^{LJVTTA} = N^2(1-f) \frac{s_{yx}}{n},$$

kus

$$s_{yx} = \frac{1}{n-1}\sum_s (y_i - \bar{y})(x_i - \bar{x})$$

on valimi kovariatsioon y ja x vahel.

Ülesanne 12.4 *Tuletada* $Cor(\hat{t}_y, \hat{t}_x)$. *Kommenteerida!*

12.4 Suhtehinnang LJV TTA korral

Lihtsa juhuvaliku korral avaldub kahe kogusumma suhte $R = \frac{t_y}{t_x}$ hinnang kujul

$$\hat{R} = \frac{\hat{t}_y}{\hat{t}_x} = \frac{\bar{y}}{\bar{x}} = \frac{\sum_s y_i}{\sum_s x_i}.$$

Hinnangu $\hat{R}$ ligikaudne dispersioon avaldub üldjuhul järgmiselt:

$$AV(\hat{R}) = \frac{1}{t_x^2}\left[V(\hat{t}_y) + R^2 V(\hat{t}_x) - 2RCov(\hat{t}_y, \hat{t}_y)\right].$$

Sellest valemist saame LJV TTA korral

$$AV(\hat{R}) = \frac{1-f}{n\bar{X}^2}\left(S_y^2 + R^2 S_x^2 - 2RS_{xy}\right),$$

kus S_y^2, S_x^2 on ÜK dispersioonid (vt Teoreemi 12.1) ja S_{xy} on üldkogumi kovariatsioon vaadeldud punktis 4.3.

Suhtehinnangu dispersiooni hinnang avaldub lihtsa juhuvaliku korral järgmiselt:

$$\hat{V}(\hat{R}) = \frac{1-f}{n\bar{x}^2}(s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}s_{yx}), \tag{38}$$

kus s_y^2, s_x^2 on valimi dispersioonid ja s_{yx} on valimi kovariatsioon.

Praktilises töös, näiteks statistikaametites armastatakse suhte dispersiooni hindamisel kasutada teist lähenemist, mis seisneb sobivate uute tunnuste moodustamises. Eespool saime lähendava lineaarse valemi suhtele

$$\hat{R} \approx R + \frac{1}{t_x}\sum_U \frac{I_i}{EI_i}u_i, \text{ kus } u_i = y_i - Rx_i.$$

Dispersioon tuleb teisest liidetavast,

$$V(\hat{R}) = \frac{1}{t_x^2} V\left(\sum_U \frac{I_i}{EI_i}u_i\right).$$

Lähtudes üldistest dispersioonivalemitest näitasime eespool, et LJV TTA korral $\hat{t}_y = \sum_U \frac{I_i}{EI_i} y_i$ dispersioon ja dispersiooni nihketa hinnang on

$$V(\hat{t}_y) = N^2(1-f)S_y^2/n \text{ ja } \hat{V}(\hat{t}_y) = N^2(1-f)s_y^2/n.$$

Meie jaoks tulevad need valemid nüüd väljendada tunnuse u_i kaudu. Sõnastame teoreemina.

Teoreem 12.2 *Suhte* $\hat{R} = \frac{\hat{t}_y}{\hat{t}_x} = \frac{\sum_s y_i}{\sum_s x_i}$ *ligikaudne dispersioon avaldub LJV TTA korral valemiga*

$$V(\hat{R}) \approx \frac{N^2(1-f)}{t_x^2} S_u^2/n,$$

ja dispersioonihinnang valemiga

$$\hat{V}(\hat{R}) = \frac{N^2(1-f)}{\hat{t}_x^2} s_u^2/n, \tag{39}$$

kus

$$S_u^2 = \sum_{i=1}^{N} (u_i - \bar{U})^2/(N-1), s_u^2 = \sum_s (u_i - \bar{u})^2/(n-1).$$

Valemiga (39) on dispersioonihinnangut oluliselt lihtsam leida kui valemiga (38)

Ülesanne 12.5 *Avaldada* $\bar{U}$ *ja* $\bar{u}$*.*

12.5 Hindamine osakogumites LJV TTA korral

Kui ÜK on mingi tunnuse väärtuste järgi jagatud osadeks ehk osakogumiteks, siis huvitavad meid nende osakogumite mahud - nii absoluutselt kui suhtelised, kogusummad, keskmised ja suhted osakogumites. Olgu üldkogumiks kõik õpilased ja osakogumiks 1. klassi õpilased. Meid huvitavad järgmised osakogumi näitajad:

Näiteks:

- osakogumi maht ehk kõigi 1. klassi õpilaste arv;
- matemaatika õppimisele kulutatud summaarne aeg;
- keskmine matemaatika õppimisele kuluv aeg õpilase kohta;
- matemaatika õppimisele kuluv aeg osakaaluna kodutöödele kuluvast ajast.

Vaatame kõigepealt osakogumi mahu ja osakaalu hindamist. Osakaalu esitatakse tavaliselt protsentides. Olgu $U_d \subset U$ meid huvitav osakogum, N_d, N on vastavad mahud.

Osakaal on $P_d = N_d/N$. Osavalimi s_d maht on n_d, osakaal valimis on $p_d = n_d/n$.

Iga üldkogumi objektiga seotakse osakogumi indikaator:

$$z_i = \begin{cases} 1, & i \in U_d \\ 0, & \text{vastasel juhul.} \end{cases}$$

Selle tunnuse kogusumma ja keskmine on osakogumi maht N_d ja suhteline maht P_d:

$$N_d = t_z = \sum_U z_i, \quad P_d = \frac{N_d}{N} = \bar{Z} = \frac{t_z}{N}.$$

Nüüd saame rakendada neile Teoreemi 12.1 erijuhul:

$$\begin{aligned} \hat{N}_d &= \hat{t}_z = N\bar{z} = Np_d, \text{ kus } p_d \text{ on valimi } s_d \text{ osakaal;} \\ \hat{P}_d &= \frac{\hat{t}_z}{N} = \bar{z} = p_d. \end{aligned}$$

Hinnangu dispersiooni saamiseks vaatleme esmalt S_z^2:

$$S_z^2 = \frac{1}{N-1}(\sum_U \underbrace{z_i^2}_{z_i^2 = z_i, z_i \in \{0,1\}} - N\bar{Z}^2) = \frac{1}{N-1}(\underbrace{\sum_U z_i}_{NP_d} - NP_d^2) = \frac{N}{N-1}P_d(1-Pd).$$

Järelikult hinnangute $\hat{N}_d$ ja $\hat{P}_d$ dispersioonid ja dispersiooni hinnangud avalduvad järgmiselt (näita!):

$$\begin{aligned} V(\hat{N}_d) &= \frac{N^3}{n}(1-f)\frac{1}{N-1}P_d(1-Pd), \quad V(\hat{P}_d) = \frac{1-f}{n}\frac{N}{N-1}P_d(1-Pd) \\ \hat{V}(\hat{N}_d) &= N^2\frac{1-f}{n-1}p_d(1-p_d), \quad \hat{V}(\hat{P}_d) = \frac{1-f}{n-1}p_d(1-p_d) \end{aligned}$$

Tunnuse y kogusumma osakogumis U_d on $t_d = \sum_{U_d} y_i$. Selleks, et oleks võimalik rakendada ÜHT hinnangute saamiseks, peame esitama t_d ÜK kogusumma kaudu. Selleks kasutame jällegi binaarset tunnust z:

$$t_d = \sum_{U_d} y_i = \sum_U y_i z_i = \sum_U y_i', \text{ kus } y_i' = z_i y_i.$$

Teoreemist 12.1 saame summa hinnangule LJV TTA korral kuju:

$$\hat{t}_d = N\bar{y}' = \frac{N}{n}\sum_s y_i' = \frac{N}{n}\sum_{s_d} y_i.$$

Dispersioonide $V(\hat{t}_d)$ ja $\hat{V}(\hat{t}_d)$ leidmiseks asendame tunnuse y_i tunnusega y_i' Teoreemis 12.1.

Ülesanne 12.6 *Kirjuta välja saadud hinnangu dispersioon ja dispersiooni hinnang.*

Osakogumi keskväärtuse $\bar{Y}_d = t_d/N_d$ hindamiseks saab kasutada Järeldust 12.1:

$$\hat{\bar{Y}}_d = \frac{\hat{t}_d}{N_d} = \frac{N}{N_d}\frac{1}{n}\sum_{s_d} y_i.$$

Pane tähele, et saadud hinnang pole osakogumi valimikeskmine!
Keskväärtuse hindamiseks on tavaliselt alternatiivne hinnang ehk suhtetüüpi hinang parem:

$$\hat{\bar{Y}}_{d,alt} = \frac{\hat{t}_d}{\hat{N}_d}.$$

Vaatame, mis kuju see võtab LJV TTA korral. Paneme tähele, et $\hat{N}_d = N/n\sum_s z_i = (N \cdot n_d)/n$. Siit saame, et

$$\hat{\bar{Y}}_{d,alt} = \frac{N/n\sum_{s_d} y_i}{N \cdot n_d/n} = \frac{1}{n_d}\sum_{s_d} y_i = \bar{y}_d.$$

Seega, suhtetüüpi keskmise hinnang osakogumis on valimi keskmine selles osakogumis. Kui osakogumi maht N_d on teada, siis on parem kogusumma hinnang osakogumis $N_d\hat{\bar{Y}}_{d,alt}$.

Üldjuhul huvitab meid järgmine osakogumi suhe, mille aga saame esitada suhtena kogu üldkogumis uute tunnuste abil:

$$R_d = \frac{\sum_{U_d} y_i}{\sum_{U_d} x_i} = \frac{\sum_U y_i'}{\sum_U x_i'},$$

kus $y_i' = z_i y_i$ ja $x_i' = z_i x_i$. Hinnangu sellele suhtele saame lugeja ja nimetaja nihketa hindamise teel:

$$\hat{R}_d = \frac{N\bar{y}'}{N\bar{x}'} = \frac{\sum_s y_i'}{\sum_s x_i'} = \frac{\sum_{s_d} y_i}{\sum_{s_d} x_i}.$$

Dispersioonivalemid $V(\hat{R}_d)$ ja $\hat{V}(\hat{R}_d)$ järelduvad Teoreemist 12.2, milles tuleb kasutada osakogumitunnuseid y_i' ja x_i'.

13 Hindamine lihtsa juhuvaliku TGA korral

LJV TGA valiku korral tehakse ÜK-s U n valikut tõenäosusega

$$p_i = \frac{1}{N}, \quad \sum_{i=1}^{N} p_i = 1.$$

Iga kord kui objekt on valitud, pannakse see üldkogumisse tagasi.
LJV TGA korral on valikuindikaatorid binoomjaotusega: $I_i \sim Bin(n, \frac{1}{N})$.
Disaini vektori I jaotuseks on multinomiaalne jaotus:

$$p(k) = \Pr(I = k) = \begin{cases} \frac{n!}{N^n \prod_{i=1}^{N} k_i!}, & \text{kui } |k| = n, \\ 0, & \text{vastasel juhul.} \end{cases}$$

Binoom- ja multinoomjaotuse karakteristikud on hästi tuntud:

$$\begin{aligned} E(I_i) &= np_i = \frac{n}{N} \\ V(I_i) &= np_i(1-p_i) = \frac{n}{N}(1 - \frac{1}{N}), \\ Cov(I_i, I_j) &= -np_ip_j = -\frac{n}{N^2}. \end{aligned}$$

Kasutusvaldkonnad:

- valikudisainina ei kasutata;
- valemid, tuletatud LJV TGA jaoks on tavaliselt lihtsa ja ilusa kujuga, neid saab kasutada sageli lähendina teiste disainide juures sobivas olukorras;
- LJV TGA on tähtis nn "taasvaliku" teoorias, kus saadud valimist võetakse korduvalt omakorda valimid kasutades LJV TGA ja selle protseduuri abil leitakse hinnangu dispersiooni hinnang.

Kogusumma hinnangu ja selle dispersiooni tuletamiseks kasutame teoreemi 6.3 punktist 6.2 (hindamisteoreem multinomiaaldisaini korral). Dispersiooni hinnangu saamiseks kasutame alternatiivset valemit, kuna LJV on fikseeritud mahuga disaid ja Sen-Yates-Grundy hinnang on stabiilsem (varieeruvuse mõttes, samuti ei võta ta negatiivseid väärtuseid).

Olgu $\hat{t} = \sum_U \frac{I_i y_i}{E(I_i)}$ on kogusumma $t = \sum_U y_i$ hinnang. Multinomiaaldisaini

korral alternatiivne dispersioonihinnang avaldub järgmiselt:

$$\begin{aligned}\hat{V}(\hat{t}) &= -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} I_i I_j \frac{\Delta_{ij}}{E(I_i I_j)}\left(\frac{y_i}{EI_i}-\frac{y_j}{EI_j}\right)^2 = \\ &= -\frac{1}{2}\sum_{i}\sum_{j} I_i I_j \frac{-np_ip_j}{n(n-1)p_ip_j}\left(\frac{y_i}{np_i}-\frac{y_j}{np_j}\right)^2 \Rightarrow \\ \hat{V}(\hat{t}) &= \frac{1}{2}\frac{1}{n-1}\sum_{i}\sum_{j} I_i I_j \left(\frac{y_i}{np_i}-\frac{y_j}{np_j}\right)^2. \qquad (40)\end{aligned}$$

Avaldame ruutu ja lihtsustame:

$$\sum_i\sum_j I_i I_j \frac{y_i^2}{n^2p_i^2} = \sum_i I_i \frac{y_i^2}{n^2p_i^2}\underbrace{\sum_{j=1}^{2} I_j}_{n} = \frac{1}{n}\sum_i I_i\frac{y_i^2}{p_i^2},$$

$$\sum_i\sum_j I_i I_j \frac{y_i}{np_i}\frac{y_j}{np_j} = \underbrace{\sum_i I_i\frac{y_i}{np_i}}_{\hat{t}}\underbrace{\sum_j I_j\frac{y_j}{np_j}}_{\hat{t}} = \hat{t}^2$$

Lõpuks saame,

$$\begin{aligned}\hat{V}(\hat{t}) &= \frac{1}{2}\frac{1}{n-1}\left[\frac{2}{n}\sum_i I_i\frac{y_i^2}{p_i^2}-2\hat{t}^2\right] = \\ &= \frac{1}{n(n-1)}\left[\sum_{i=1}^{N} I_i\frac{y_i^2}{p_i^2}-n\hat{t}^2\right],\end{aligned}$$

vastava hinnanguga

$$\hat{V}(\hat{t}) = \frac{1}{n(n-1)}\left[\sum_s k_i\frac{y_i^2}{p_i^2}-n\hat{t}^2\right].$$

Teoreem 13.1 *Lihtsa juhusliku valiku tagasipanekuga korral nihketa hinnang ÜK kogusummale $t=\sum_U y_i$ avaldub järgmiselt:*

$$\hat{t} = \frac{N}{n}\sum_U I_i y_i,$$

vastava punktihinnanguga

$$\hat{t} = N\bar{y}.$$

Hinnangu $\hat{t}$ dispersioon on järgmine:

$$V(\hat{t}) = \frac{N(N-1)}{n} S_y^2, \tag{41}$$

ja dispersiooni hinnangufunktsioon:

$$\hat{V}(\hat{t}) = \frac{N^2}{n(n-1)} \left[\sum_{i_1}^{N} I_i y_i^2 - n\bar{y}^2 \right]. \tag{42}$$

Viimasele avaldisele vastav punktihinnang on järgmine:

$$\hat{V}(\hat{t}) = \frac{N^2}{n} s_y^2,$$

kus

$$\begin{aligned} \bar{y} &= \frac{1}{n} \sum_s y_i, \\ \bar{Y} &= \frac{1}{N} \sum_U y_i, \\ S_y^2 &= \frac{1}{N-1} \sum_U (y_i - \bar{Y})^2, \\ s_y^2 &= \frac{1}{n-1} \sum_s (y_i - \bar{y})^2. \end{aligned}$$

Ülesanne 13.1 *Tõestada väide (41).*

Hinnangud keskväärtusele LJV TGA korral avalduvad järgmiselt:

$$\begin{aligned} \hat{\bar{Y}} &= \bar{y}, \\ \hat{V}(\hat{\bar{Y}}) &= \frac{s_y^2}{n}. \end{aligned}$$

Saime klassikalise statistika tulemusi!
Võrdleme SI ja SIR (LJV TTA ja LJV TGA) omavahel:

$$V_{SI}(\hat{t}) = N^2(1-f)\frac{S_y^2}{n}, \text{ kus } f = \frac{n}{N} \text{ ja } \quad V_{SIR}(\hat{t}) = N(N-1)\frac{S_y^2}{n}.$$

Juhul kui $n = N$, $V_{SI}(\hat{t}) = 0$, kuid $V_{SI}(\hat{t}) \neq 0$!
Üldiselt, $V_{SIR}(\hat{t}) \geq V_{SI}(\hat{t})$.Võrdusmärk kehtib kui $n = 1$ ja $n = N - 1$. Lihtne

juhuvalik tagasipanekuga on vähem efektiivne kui tagasipanekuta lihtne juhuvalik.

Disainiefekt:

$$Deff_{SIR} = \frac{V_{SIR}(\hat{t})}{V_{SI}(\hat{t})} = \frac{N(N-1)S_y^2/n}{N^2(1-f)S_y^2/n} \approx \frac{1}{1-f}.$$

Mida suurem on valikusuhe f (st mida lähedasem on ta 1-le), seda vähem on SIR efektiivne.

Ülesanne 13.2 *Olgu $N = 10$ ja $n = 2$. Olgu y väärtused üldkogumis järgmised:* (2, 1, 2, 4, 1, 3, 2, 1, 4, 1)*. Eeldades, et valikudisainiks on lihtne juhuslik valik TGA, leida disaini efekt (täpselt).*

13.1 TGA disaini poolt indutseeritud TTA disain

On tõestatud, et objektide korduvvalikud ei suurenda hinnangu täpsust. Ja tegelikult, kui kordused välja jätta, siis saadud hinnang varieerub vähem.

Definitsioon 13.1 *Indutseeritud tagasipanektuta disain on selline, mis on saadud järgmise algoritmi abil:*

1. *võtta valim üldkogumist lihtsa juhuvaliku tagasipanekuga abil;*
2. *kustutada kordused valimist (jätta objektid ainult ühekordselt).*

Meetodi puuduseks on kaalude ümberarvutamine (vastasel juhul saame nihkega hinnangu) ja valemite keerukus.

Tähistagu $I^{ind} = (I_1^{ind}, I_2^{ind}, ..., I_N^{ind})$ valikuvektorit indutseeritud disaini korral, $I_i^{ind} \in \{0, 1\}$.

Valikuindikaatori keskväärtus on sel juhul:

$$EI_i^{ind} = \pi_i^{ind} = Pr(I_i^{ind} = 1) = Pr(I_i \geq 1) = 1 - Pr(I_i < 1) = 1 - Pr(I_i = 0).$$

Kuna $I_i \sim B(n, \frac{1}{N})$, siis

$$Pr(I_i = 0) = C_n^0 \left(\frac{1}{N}\right)^0 \left(1 - \frac{1}{N}\right)^{n-0} = \left(1 - \frac{1}{N}\right)^n.$$

Kokkuvõttes leiame, et

$$EI_i^{ind} = 1 - \left(1 - \frac{1}{N}\right)^n.$$

Saadud keskväärtuse abil avaldame nihketa hinnangu kogusummale t:

$$\hat{t}^{ind} = \sum_U \frac{I_i^{ind} y_i}{1 - \left(1 - \frac{1}{N}\right)^n}. \tag{43}$$

Hinnangu varieeruvuse $V(\hat{t})$ valemit antud kursuse raames ei tuleta.

Ülesanne 13.3 *Olgu antud väike üldkogum, kus $N = 4$ ja uuritava tunnuse väärtused on $y_1 = 8, y_2 = 10, y_3 = 7, y_4 = 4$. Sellest üldkogumist võetakse valim mahuga $n = 2$ kasutades lihtsat juhuvalikut tagasipanekuga. Seejärel moodustatakse indutseeritud disain. Näidata, et indutseeritud disaini järgi saadud kogusumma hinnangud on tõepoolest nihketa. Vihje: kasutada diskreetse juhusliku suuruse keskväärtuse definitsiooni.*

14 Isekaaluvad disainid

Isekaaluvate disainide korral kehtib:

$$E(I_i) \equiv const.$$

Kui lisaks disain on fikseeritud mahuga n, siis $\sum_{i=1}^N E(I_i) = n$ ja siit järeldub, et

$$E(I_i) \equiv \frac{n}{N}.$$

Sellisel juhul nihketa hinnang ÜK kogusummale on

$$\hat{t} = \sum_U \frac{I_i y_i}{E(I_i)} = \frac{N}{n} \sum_U I_i y_i,$$

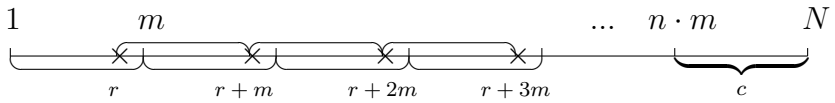
mis tähendab, et hinnang põhineb valimikeskmisel:

$$\hat{t} = N\bar{y}.$$

Ja see omakorda tähendab, et ÜK keskmist hindab valimikeskmine, ÜK osakaalu - valimi osakaal,... $\Leftrightarrow$ valimikarakteristikud esindavad ÜK karakteristikuid. See on aga küllaltki mugav ning võimaldab kasutada hindamisülesannetes tarkvara, mis on mõeldud klassikalise statistika jaoks.
Tähelepanu! Kogusumma hinnangu disperioonid on siiski üldjuhul erinevad eri disainide korral!

15 Süstemaatiline valik

Olgu $U = 1, ..., N$.



Süstemaatilise valiku korral võetakse esimene element valimisse juhuslikult m esimese elemendi hulgast (võrdse tõenäosusega). Valimit moodustavad see esimene element pluss iga m-s element freimist.
Kokku on võimalik saada m erinevat valimit. Iga sellise valimi saamise tõenäosus on $1/m$.
Olgu $I = (I_1, ..., I_N)$ valikuvektor freimis U. Süstemaatilise valiku korral on sellise vektori realisatsiooniks vektor k elementidega 0 ja 1, kus 1 esineb iga m sammu tagant. Seega, on vektori I jaotus järgmine:

$$p(k) = \Pr(I = k) = \begin{cases} 1/m, & \text{kui 1 ilmub esimese } m \text{ hulgas;} \\ 0, & \text{vastasel juhul.} \end{cases}$$

Süstemaatiline disain on TTA disain. Kasutusvaldkonnad:

1. SÜ on lihtsalt teostatav jooksval valikul ja seetõttu ta on vähem tundlik intervjueerijate subjektiivsete vigade suhtes kui LJV või KV (kihtvalik). Eriti kui korralik freim ei ole kättesaadav.

Näiteks, ostjate lihtsa juhusliku valimi mahuga $n = 50$ korjamine tänavanurgal on üsna keerukas. Intervjueerija ei saa otsustada, milliseid ostjaid võtta valimisse, sest ÜK maht N ei ole teada kuni kõik ostjad on ära käinud. Seevastu intervjueerija võib kasutada SÜ-t ja võtta valimisse näiteks iga 20. ostja kuni nõutava mahuga valim on saadud. See protseduur on lihtne isegi kogenematu intervjueerija jaoks.

Need küsitlejad, kes küsitlevad inimesi liikumisel, kasutavadki väga sageli SÜ-t. Nad võivad küsitleda iga 20-nda inimese kassa juures toidu maitse või värvuse kohta. Iga 10-s isik, kes siseneb bussi võib olla küsitletud bussiteeninduse kohta. Samuti metsavahid võivad võtta maatükkide süstemaatilist valimit, ning süstemaatiliselt valida puud, et uurida haigete puude osakaalu. Seetõttu on SÜ väga populaarne valikudisain.

2. SÜ võib anda täpsema informatsiooni kui LJV sama maksmuse korral.

Süstemaatiline valim on reeglina "ühtlasem" ja seega annab rohkem informatsiooni ÜK kohta kui sama mahuga lihtne juhuslik valim. Näiteks, me tahame võtta SÜ valimit mahuga $n = 200$ panga maksekviitungite ÜK-st mahuga $N = 1000$, selleks, et hinnata korrektselt täidetud kviitungite osakaalu. Selleks võtame juhuslikult ühe kviitungi 5-st esimesest määramaks alguspunkti (näiteks, number 3), ja seejärel võtame iga viienda kviitungi.

Oletame, et suurem osa esimesest 500st kviitungist olid täidetud korrektselt, järgmised 500 aga olid kõik täidetud valesti (näiteks, pangateenindaja kogenematuse tõttu). LJV korral võib valimisse ($n = 200$) sattuda liiga palju (võimalik, et kõik) kviitunge esimesest (või teisest) osast kviitungitest. See annab aga ebatäpse hinnangu osakaalule. Seevastu SÜ valib võrdse kviitungite arvu mõlemast gruppist ja annab parema hinnangu valesti täidetud kviitungite osakaalule.

■

SÜ korral igal objektil on olemas võimalus sattuda valimisse, st

$$\pi_i = \Pr(I_i = 1) > 0.$$

Disaini puudus on aga see, et mõned objektid ei saa korraga sattuda valimisse, st mõned 2. järku kaasamistõenäosused on võrdsed 0-ga, $\pi_{ij} = 0$. See aga omakorda tähendab, et pole võimalik leida hinnangute teoreetilist dispersiooni.

Valimimaht, $\mathbf{n}$, on SÜ korral juhuslik ja on määratud sammuga m. SÜ korral kehtib:

$$N = nm + c, \quad 0 \le c < m.$$

Seega, realiseerunud valimimaht, tähistame n_s on:

$$n_s = \begin{cases} n+1, & \text{kui } r \le c; \\ n, & \text{kui } r > c. \end{cases}$$

Leiame ka 1. ja 2. järku kaasamistõenäosused:

$$\pi_i = \Pr(I_i = 1) = \sum_{k, k_i = 1} p(k) = \frac{1}{m},$$

kuna on võimalik ainult üks selline valim k, mis sisaldaks i-ndas positsioonis 1, siis

$$\pi_{ij} = \Pr(I_i = 1, I_j = 1) = \sum_{k, k_i = 1, k_j = 1} p(k) = \begin{cases} 1/m, & \text{kui vahe i ja j vahel on sammu } m \text{ kordne;} \\ 0, & \text{vastasel juhul.} \end{cases}$$

Kuna valimimahud väga ei varieeru, siis on huvitav leida keskmist valimimahtu:

$$E(\mathbf{n}) = E\left[\sum_{i=1}^{N} I_i\right] = \sum_{i=1}^{N} E(I_i) = \sum_{i=1}^{N} \pi_i = \frac{N}{m} = \frac{nm + c}{m} = n + \frac{c}{m}.$$

15.1 Hindamine SÜ korral

Kogusumma $t = \sum_U y_i$ nihketa hinnang ÜHT-i järgi on järgmine:

$$\hat{t} = \sum_U \frac{I_i y_i}{\pi_i} = m \sum_U I_i y_i,$$

millele vastab järgmine punktihinnang:

$$\hat{t} = m \sum_s y_i.$$

Kuna on võimalik saada kokku m erinevat valimit võrdse tõenäosusega (olenevalt alguspunktist r), siis on ka võimalik saada kokku m erinevat hinnangut ÜK kogusummale, tähistame $\hat{t}_1, ..., \hat{t}_m$:

$$\hat{t}_r = m \sum_{s_r} y_i, \quad r = 1, ..., m.$$

Nüüd saame kirja panna $V(\hat{t})$ ilma ÜHT-ta, kasutades diskreetse juhusliku suuruse dispersiooni definitsiooni:

$$V(\hat{t}) = \sum_{r=1}^{m} (\hat{t}_r - \underbrace{E\hat{t}}_{t})^2 \underbrace{\Pr(\hat{t} = \hat{t}_r)}_{1/m} = \frac{1}{m} \sum_{r=1}^{m} (\hat{t}_r - t)^2.$$

Näeme, et $V(\hat{t})$ sõltub sellest, kuidas varieeruvad $\hat{t}_r$ kogusumma t ümber. Dispersioon $V(\hat{t})$ on teoreetiline, seda ei saa välja arvutada, kuna praktikas on meil olemas ainult üks valim s_r ja ainult üks hinnang $\hat{t}_r$. Kuid me saame seda teoreetilist hinnangut kasutada SÜ uurimiseks.

Kõigepealt, lihtsustame situatsiooni ja eeldame, et $c = 0$, st $N = nm$.

Sel juhul $\hat{t}_r = m \sum_{s_r} y_i = \frac{N}{n} \sum_{s_r} y_i = N\bar{y}_r$. Kuna $t = N\bar{Y}$, me saame kirjutada teoreetilise dispersiooni $V(\hat{t})$ ümber järgmiselt:

$$V(\hat{t}) = Nn \sum_{r=1}^{N} (\bar{y}_r - \bar{Y})^2.$$

Seega, varieeruvus $V(\hat{t})$ sõltub valimikeskmiste varieeruvusest. Me soovime, et see varieeruvus oleks väike, see tagaks väikese dispersiooni $V(\hat{t})$. Kuna teisi valimeid pole, siis ka pole võimalik midagi otsustada selle varieeruvuse kohta. Antud olukorras saame kasutada ANOVA lahutust (tunnuse koguvarieeruvus

grupisisese ja gruppidevahelise varieeruvuse kaudu):

$$\begin{aligned} SST &= \sum_U (y_i - \bar{Y})^2 = \sum_{r=1}^{m} \sum_{i \in s_r} (y_i - \bar{y}_r + \bar{y}_r - \bar{Y})^2 = \\ &= \underbrace{\sum_{r=1}^{m} \sum_{i \in s_r} (y_i - \bar{y}_r)^2}_{SSW} + \underbrace{n \sum_{r=1}^{m} (\bar{y}_r - \bar{Y})^2}_{SSB} = SSW + \frac{1}{N} V(\hat{t}). \end{aligned}$$

Fikseeritud ÜK korral on uuritava tunnuse varieeruvus, SST (*Sum of squares total*), samuti fikseeritud. Selleks, et saada väiksema $V(\hat{t})$, SSW (*Sum of squares within groups*) peab olema võimalikult suur. Ja see omakorda tähendab, et tuunuse y varieeruvus valimis s_r peab olema suur $\Rightarrow$ tuunus peab olema valimis s_r võimalikult heterogeenne. Järelikult, dispersiooni $V(\hat{t})$ suurus sõltub objektide järjestusest loendis.

Hea järjestus on järgmine:

- y väärtused, mis asuvad üksteisest kaugusel m peavad olema võimalikult erinevad;
- seda saab saavutada järjestades freimi väärtuseid kas uuritava tunnuse või sellega korreleeruva tunnuse väärtuste järgi.

Halb järjestus:

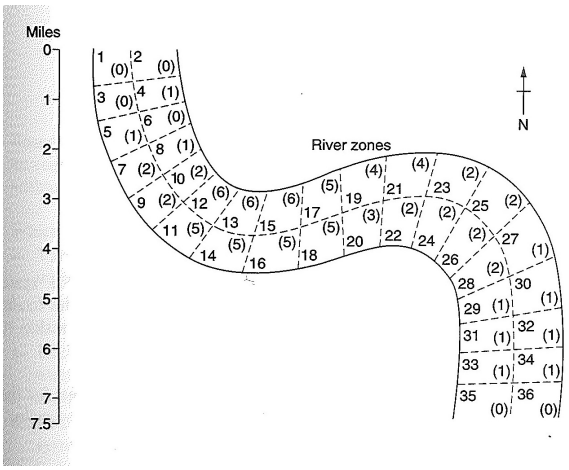
- loendis esineb väärtuste tsüklilisus perioodiga m; sellisel juhul tunnuse varieeruvus valimis on väike.

SÜ korral pole võimalik saada nihketa hinnangut kogusumma hinnangu dispersioonile, $\hat{V}(\hat{t})$. Sel juhul saab kasutada mõnda nihkega hinnangut, näiteks SI hinnangut:

$$\hat{V}(\hat{t}) = N^2(1-f)\frac{s_y^2}{n}.$$

Juhul, kui ÜK on halvasti järjestatud, siis s_y^2 võib osutuda liiga väikeseks ja sel juhul $\hat{V}(\hat{t})$ võib tegeliku dispersiooni alahinnata.

Ülesanne 15.1 *Linnas A on aeg-ajalt tuvastatud jõest ühte pestitsiidi ainet (dieldrin, mis on tunnustatud kantserogeeniks), ja seda territooriumil 7,5 miili. Keskkonnakaitse plaanib läbi viia uuring keskmise pestitsiidi taseme tuvastamiseks jõest. Selleks jaotatakse jõgi 36-ks osaks (vt pilt) ja proovid võetakse 9-st osast kasutades süstemaatilist valikut. Numbrid sulgudes (mikrogramm liitri kohta) näitavad selle osa pestitsiidi taset, mida saab teada alles valikuuringu läbiviimise etapil.*



Kasutades süstemaatilist valikut mahuga $n = 9$ ja juhusliku algusega leida vahemikhinnang usaldusnivool 90% keskmisele pestitsiidi tasemele.

Ülesanne 15.2 *Muuseum plaanib välja panna ühte kallist maalide kollektsiooni 180-ks päevaks ja vaadata, mitu inimest külastab muuseumit antud perioodi jooksul. Kuna igapäevane külaliste monitooring on liiga kallis, siis otsustatakse jälgida igal 10-l päeval. Andmed on toodud järgmises tabelis.*

Päeva nr	*Külastajate arv*
3	*160*
13	*350*
23	*225*
...	*...*
173	*290*
	$\sum_{i=1}^{18} y_i = 4868;\ \sum_{i=1}^{18} y_i^2 = 1321450$

Leida vahemikhinnang külastajate koguarvule 180 päeva jooksul usaldusnivool 90%.

Ülesanne 15.3 *Viiakse läbi sotsioloogiline leibkonna uuring, milles üheks tunnuseks on Kas leibkond uurib elamispinda või mitte. Kui leibkond uurib elamispinda, siis $y_i = 1$; vastasel korral $y_i = 0$. Linnas on kokku $N = 15200$ leibkonda. Kasutades süstemaatilist valikut sammuga 50 on leitud, et $\sum_{i=1}^{304} y_i = 88$. Leida vahemikhinnang rentnike osakaalule usaldusnivool 90 %.*

15.2 SÜ disaini effekt

Eelmises punktis näitsime, et uuritava tunnuse koguvarieeruvuse ÜK-s on võimalik esitada järgmiselt:

$$SST = SSW + SSB = SSW + \frac{1}{N}V(\hat{t}),$$

kus *SSB= Sum Square Between* on varieeruvus gruppide vahel.

$$\Rightarrow V_{SY}(\hat{t}) = N(SST - SSW) = N \cdot SST(1 - \frac{SSW}{SST}) =$$

$$= N(N-1)S_y^2 \left(1 - \frac{SSW}{SST}\right).$$

SÜ disainiefekt:

$$deff(SY) = \frac{V_{SY}(\hat{t})}{V_{SI}(\hat{t})} = \frac{N(N-1)S_y^2\left(1 - \frac{SSW}{SST}\right)}{N^2(1-f)\frac{S_y^2}{n}} =$$

$$= \frac{(N-1)n}{N(1-f)}\left(1 - \frac{SSW}{SST}\right).$$

Süstemaatiline valik on efektiivsem kui lihtne juhuvalik siis, kui $deff(SY) < 1$. See aga tähendab järgmist võrratust:

$$\frac{(N-1)n}{N(1-f)}\left(1 - \frac{SSW}{SST}\right) < 1$$

$$\left(1 - \frac{SSW}{SST}\right) < \frac{N(1-f)}{(N-1)n}$$

$$\frac{SSW}{SST} > \frac{N(n-1)}{(N-1)n}$$

Tähistame

$$S_w^2 = \frac{SSW}{N-m}$$

- valimite sisene hajuvus. Arvestades, et

$$N - m = N - \frac{N}{m} = \frac{N(n-1)}{n}$$

viimasest võrratusest saame, et

$$S_w^2 > S_y^2$$

ehk süstemaatiline valik on lihtsast juhuvalikust efektiivsem, kui tunnuse y valimisisene hajuvus on suur, võrreldes hajuvusega ÜK-s.

Parim hinnang saadakse loendi korral, mis on järjestatud uuritava tunnuse või sellega tugevalt korreleeritud tunnuse väärtuste järgi.

Halva järjestusega loendi puhul võidakse valimisse saada liialt vähe varieeruvad objektid, mille tagajärjeks on ebatäpsed hinnangud alahinnatud usaldusintervalliga.

15.3 SÜ realiseerimine praktikas

1. Mõnikord on SÜ probleemiks, et pole võimalik saavutada täpselt etteantud valimimahtu. Näiteks kui $N = 125$ ja samm $m = 3$, saame $n = \left[\frac{125}{3}\right] = 41$ ehk valimimaht on kas 41 või 42 sõltuvalt juhuslikust stardist. Kui aga $m = 4$, siis $n = 31$, $n + 1 = 32$. Valimimahtusid nt. $33 - 40$ pole võimalik saada. Suurte üldkogumite korral see probleem kaob.

2. Valimimahu reguleerimiseks, kasutatakse teisi SÜ protseduure, millest üks on näiteks ringsüstemaatiline valik. Selle meetodi korral vaadeldakse loendit ringina, kus viimasele elemendile järgneb jälle esimene. Genereeritakse juhuslik arv $1 \leq r \leq N$ ja võetakse talle vastav objekt ning sammu m tagant iga järgnev objekt, kuni soovitud valimimaht on käes.

16 Ebavõrdsete tõenäosustega valik

Andes üldkogumi elementidele erinevaid kaasamistõenäosusi, on võimalik parandada leitavate hinnangute omadusi. Vaatame siinkohal lähemalt üht väga levinud viisi.

Olgu $t = \sum_U y_i$ ja sellele vastav nihketa hinnang $\hat{t} = \sum_U \frac{I_i y_i}{EI_i}$.

Kui valida disain nii, et oodatavad valikute arvud on võrdelised y väärtustele, ehk

$$EI_i \propto y_i \ \ (EI_i = cy_i),$$

siis $y_i/EI_i \equiv 1/c$ ja hinnang saab järgmist kuju:

$$\hat{t} = \frac{1}{c} \sum_U I_i.$$

Kui lisaks disain on fikseeritud mahuga, siis $\sum_U I_i = n$ ja hinnang lihtsustub veelgi rohkem:

$$\hat{t} = n/c.$$

Võtame viimases avaldises mõlemalt pool kesväärtuse ja saame, et $c = n/t$. Järelikult, fikseeritud mahuga disainide korral iga kogusumma hinnang $\hat{t}$ annab meile tulemuseks täpse parameetri t sõltumata realiseerunust valimist.

Juhusliku valimimahuga disainide korral $\sum_U I_i = \mathbf{n}$. Olgu n_s realiseerunud valimimaht. Sel juhul:

$$t = E(\hat{t}) = \frac{1}{c}\, E\left(\sum_U I_i\right) = \frac{1}{c}\, E(\mathbf{n})$$

$$\Rightarrow \quad c = \frac{E(\mathbf{n})}{t},$$

ja kogusumma hinnang:

$$\hat{t} = \frac{t}{E(\mathbf{n})} n_s.$$

Ülalpool kirjeldatud valikut nimetatakse suurusega võrdelise tõenäosusega valikuks (Sampling with Probablities Proportsional to Size, PPS).

Probleemid seotud PPS realiseerimisega:

- kuna y_i pole teada enne valimi võtmist, siis ka pole võimalik leida $EI_i \propto y_i$;
- juhul, kui on võimalik kasutada taustinfot, ütleme tunnust x, mis on teadaolevalt positiivselt seotud uuritava tunnusega, siis saab valida $EI_i \propto x_i$;
- suurtes uuringutes, kus uuritavaid tunuseid on palju, võib juhtuda, et EI_i on võrdelised ainult mõnede tunnustega; sellisel juhul hinnangud teiste tunnuste jaoks tulevad ebatäpsed.

PPS kasutamise **näiteid**...

1. Leibkonna eelarve uuring. Selleks kasutatakse tavaliselt rahvastikuregister (mis sisaldab infot inimeste kohta). Sellest võetakse valim võrdse tõenäosusega iga inimese jaoks. Leibkondadel on sellisel juhul tõenäosus olla valitud võrdeline leibkonna suurusega. Selline valikuviis suurendab hinnangute täpsust, mis on seotud näiteks kulutustega, kuna need tunnused on enamasti tugevalt ja ka positiivselt korreleeritud leibkonna suurusega. Tulud on samuti positiivselt seotud leibkonna suurusega, kuid see seos on nõrgem.

2. Kui tahetakse hinnata vabade töökohtade arvu linnas, siis LJV puhul on valimis enamik väikeettevõtteid (neid on rohkem), aga hinnatav parameeter oleneb just palju suurfirmadest. Seega peaks neil olema suurem võimalus valimisse sattuda.

PPS kasutamise pealmised **põhjused**:

1. hinnangute täpsuse suurendamine;

2. kindlate objektide sattumine valimisse (nt nende edaspidiseks uurimiseks).

Märkus. Miks kutsutakse antud valikut 'ebavõrdsete tõenäosustega valikuks',

kui tingimused on määratud keskväärtustele EI_i. Kus on siin tõenäosused?

Teame, et

$$EI_i = \begin{cases} \pi_i, & \text{TTA disainide korral;} \\ np_i, & \text{TGA disainide korral.} \end{cases}$$

Järelikult, tingimused EI_i jaoks tähendavad ka tingimusi tõenäosuste jaoks.

16.1 Suurusega võrdelise tõenäosusega valik

Eeldame, et enne uuringu teostamist teame tausttunnuse x väärtuseid. Tavaliselt on selliseks tunnuseks mingit suurust iseloomustav tunnus.

Disaini moodustamiseks valime

$$EI_i \propto x_i,$$

mis tähendab, et

$$EI_i = cx_i. \quad \Big| \sum_{i=1}^{N}(...)$$

$$\underbrace{\sum_{i=1}^{N} EI_i}_{E\mathbf{n}} = c \underbrace{\sum_{i=1}^{N} x_i}_{t_x}$$

Järelikult,

$$c = \frac{E\mathbf{n}}{t_x}.$$

Kokkuvõttes võib öelda, et valikuindikaatori keskväärtus peab olema:

$$EI_i = \begin{cases} E(\mathbf{n})x_i/t_x, & \text{juhusliku valimimahuga disainide korral;} \\ nx_i/t_x, & \text{fikseeritud mahuga disainide korral.} \end{cases}$$

Kuigi valemid näevad lihtsad välja, pole siiski lihtne konstrueerida algoritmi fikseeritud mahuga TTA valiku teostamiseks. Üks tuntumaid on nn Sunteri algoritm (vaatame loengul, tahvlil).

TGA valik fikseeritud mahuga ei ole midagi muud kui multinomiaalne disain valikutõenäosustega $p_i = x_i/t_x$ ja valimimahuga n. Teostamisviisi vaatasime varem. Probleemiks - TGA disainid pole kõige eelistatumad disainid praktikas.

Üks lihtsamatest TTA disainidest on Poissoni valik, mis kahjuks annab juhuslikku valimimahtu. Kuid oma lihtsuse tõttu see disain on üsna populaarne praktikas.

16.2 Poissoni valik

Poissoni valiku korral kõik elemendid läbitakse järjest, alates esimesest kuni viimaseni, üks kord. Iga elemendi jaoks saadakse juhusliku valikuindikatori realisatsioon, $I_i \sim Be(\pi_i) = Bin(1, \pi_i)$, I_i on kõik sõltumatud juhuslikud suurused.

PPS valiku korral $\pi_i = nx_i/t_x$, x on tausttunnus. Meeldetuletuseks, Poissoni disaini karakteristikud:

$$\begin{aligned} EI_i &= \pi_i = nx_i/t_x, \\ VI_i &= \pi_i(1-\pi_i), \\ E(I_iI_j) &= \pi_{ij} \overbrace{=}^{I_i \perp I_j} \pi_i\pi_j, \\ Cov(I_i, I_j) &= 0. \end{aligned}$$

Teoreem 16.1 *Poissoni valiku korral, $I \sim Bin(1, \pi_i)$, nihketa hinnang ÜK kogusummale $t = \sum_U y_i$ on järgmine:*

$$\hat{t} = \sum_U \frac{I_i y_i}{\pi_i},$$

vastava punktihinnanguga:

$$\hat{t} = \sum_s \frac{y_i}{\pi_i}.$$

Hinnangu dispersioon on $V(\hat{t}) = \sum_U \frac{1-\pi_i}{\pi_i} y_i^2$
ja dispersiooni hinnang: $\hat{V}(\hat{t}) = \sum_U \frac{1-\pi_i}{\pi_i^2} y_i^2 I_i$
vastava punktihinnanguga

$$\hat{V}(\hat{t}) = \sum_s \frac{1-\pi_i}{\pi_i^2} y_i^2.$$

Ülesanne 16.1 *Tõestada Teoreemi 16.1 väiteid iseseisvalt, arvestades, et $\Delta_{ij} = 0, i \neq j; \Delta_{ii} = VI_i,\ \pi_{ii} = \pi_i$.*

Kuna Poissoni valik on juhusliku valimimahuga disain, siis eelistatakse alternatiivset (suhtetüüpi) hinnangut ÜK kogusummale:

$$\hat{t}_{alt} = \frac{\hat{t}}{\hat{N}} N,$$

kus $\hat{N} = \sum_s 1/\pi_i$.

17 Kihtvalik

Kihtvalik on praktikas enim kasutatav valikudisain, mille korral jaotatakse objektid ÜK-s mõne tausttunnuse (kihistava tunnuse) väärtuse järgi osadesse (kihtidesse). Kihte vaadeldakse üksteisest sõltumatute kogumitena, milledes võib rakendada erinevaid valikumeetodeid.

Kihtvalikut kasutatakse:

- Hinnangu täpsuse tõstmiseks - uuritava tunnuse suhtes homogeensed kihid tagavad valimihinnangu väikese varieeruvuse (disainiefekt < 1).
- Osakogumite hindamiseks - eriti väikeste valimite korral on mõttekas osakogumit esitada eraldi kihtidena, et rakendada seal temale sobivat optimaalset disaini.
- Erinevat käsitlust vajavate kihtide hindamine - kallimalt uuritavate objektide valimit vähendatakse, suure kao korral valimit suurendatakse.
- Uuringu administreerimine - suunamaks valimi paigutust (nt. intervjueerijate keskuste ümber). See võimaldab kokkuhoidu uuringu läbiviimisel.

Kihistava(te) tunnus(t)e valik:

- määratud üldkogumi kõigil objektidel, teada enne uuringu läbiviimist (sugu, vanus, maakond, linn/maa, töötajate arv,...)
- ei määra liiga peent kihistust, mis raskendaks osakogumite hinnangute leidmist.

Disain

Olgu lõplik üldkogum $U = \{1, ..., N\}$ jagatud H kihiks $U_1, ..., U_h, ..., U_H$ vastavate mahtudega $N_1, ..., N_h, ...N_H$ kihtides, kusjuures

$$U = \bigcup_{h=1}^{H} U_h,\ U_h \cap U_g = \emptyset \text{ kui } h \neq g,$$

$$N = \sum_{h=1}^{H} N_h.$$

Tähistame valikuvektorit kihis h: $\boldsymbol{I}_h = (I_r, ..., I_{r+N_h})$, kus r on eelmiste kihtide objektide arv + 1, $r = \sum_{i=1}^{h-1} N_i + 1$. Igas kihis rakendatakse teiste

kihtide omast sõltumatut valikut vastavalt disainile $p_h(\boldsymbol{k}_h) = P(\boldsymbol{I}_h = \boldsymbol{k}_h)$.

Terve valikuvektor $\boldsymbol{I}$ koosneb kihtide alamvektoritest,

$$\boldsymbol{I} = (\boldsymbol{I}_1, ..., \boldsymbol{I}_h, ..., \boldsymbol{I}_H),$$

ning tänu alamvektorite sõltumatusele saab valikudisaini esitada kihtide disainide korrutisena:

$$p(\boldsymbol{k}) = \prod_{h=1}^{H} p_h(\boldsymbol{k}_h),$$

kus $\boldsymbol{k} = (\boldsymbol{k}_1, ..., \boldsymbol{k}_H)$.

17.1 Hindamine kihtvaliku korral

Tähistame:

$t_h = \sum_{U_h} y_i$ - uuritava tunnuse summa kihis U_h,
$\bar{Y}_h = \frac{t_h}{N_h}$ - keskmine kihis U_h.

Soovime hinnata ÜK kogusumma t,

$$t = \sum_{h=1}^{H} t_h,$$

ehk alternatiivselt,

$$t = \sum_{h=1}^{H} N_h \bar{Y}_h.$$

Hinnatavaks parameetriks võib olla ka ÜK keskmine

$$\bar{Y} = \frac{t}{N} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{Y}_h = \sum_{h=1}^{H} W_h \bar{Y}_h,$$

kus W_h on kihi osakaal ÜK-s.

Teoreem 17.1 *(Kihtvalik). Kihtvaliku korral on nihketa hinnang ÜK summale t järgmine:*

$$\hat{t} = \sum_{h=1}^{H} \hat{t}_h,$$

kus $E(\hat{t}_h) = t_h$ *ehk hinnang* $\hat{t}_h$ *on nihketa kihis* U_h. *Hinnangu* $\hat{t}$ *dispersioon on*

$$V(\hat{t}) = \sum_{h=1}^{H} V(\hat{t}_h)$$

ja sellele vastav nihketa hinnang

$$\hat{V}(\hat{t}) = \sum_{h=1}^{H} \hat{V}(\hat{t}_h),$$

kus $E[\hat{V}(\hat{t}_h)] = V(\hat{t}_h)$.

Teoreemi tõestus järeldub hinnangute $\hat{t}_h$ sõltumatusest erikihtides, samuti ka operaatorite E ja V omadustest.

Järeldus 17.1 . *Kihtvaliku korral avaldub hinnang ÜK keskmisele kihikeskmiste hinnangute kaalutud keskmisena,*

$$\hat{\bar{Y}} = \sum_{h=1}^{H} W_h \hat{\bar{Y}}_h,$$

mille dispersioon on

$$V(\hat{\bar{Y}}) = \sum_{h=1}^{H} W_h^2 V(\hat{\bar{Y}}_h).$$

Kui kihtides kasutatakse nihketa hinnanguid dispersioonidele $\hat{V}(\hat{\bar{Y}}_h)$, *siis nihketa hinnang dispersioonile on*

$$\hat{V}(\hat{\bar{Y}}) = \sum_{h=1}^{H} W_h^2 \hat{V}(\hat{\bar{Y}}_h).$$

17.2 Lihtne juhuslik kihtvalik

Kui kõikides kihtides kasutatakse lihtsat juhuvalikut TTA, siis nimetatakse sellist valikumeetodit lihtsaks juhuvalikuks (LJKV). Seejuures võib kihtides kasutada erinevaid valikusuhteid

$$f_h = \frac{n_h}{N_h}, h = 1, ..., H.$$

Paneme tähele, et kuigi ühe kihi piires disain on isekaaluv, pole ta seda terves üldkogumis, mille tõttu valimikeskmine ja osakaal ei ole nihketa hinnanguteks

ÜK keskmisele ja osakaalule.

LJ TTA korral on kihi sees hinnanguks prameetrile t_h

$$\hat{t}_h = \sum_{U_h} \frac{I_i y_i}{\pi_i} = \frac{N_h}{n_h} \sum_{U_h} I_i y_i,$$

või valimi kaudu:

$$\hat{t}_h = N_h \bar{y}_h,$$

kus $\bar{y}_h = \frac{1}{n_h} \sum_{s_h} y_i$ valimikeskmine kihis U_h. Kasutades teoreeme (Kihtvalik ja LJ valik TTA) saame sõnastada teoreemi LJKV jaoks.

Teoreem 17.2 *(Lihtne juhuslik kihtvalik).Lihtsa juhusliku kihtvaliku korral avaldub kogusumma $t = \sum_U y_i$ kujul*

$$\hat{t} = \sum_{h=1}^{H} N_h \bar{y}_h,$$

dispersiooniga

$$V(\hat{t}) = \sum_{h=1}^{H} N_h^2 (1 - f_h) S_{yh}^2 / n_h$$

ja dispersiooni nihketa hinnanguga

$$\hat{V}(\hat{t}) = \sum_{h=1}^{H} N_h^2 (1 - f_h) s_{yh}^2 / n_h,$$

kus

$$S_{yh}^2 = \frac{1}{N_h - 1} \sum_{U_h} (y_i - \bar{Y}_h)^2,$$

$$S_{yh}^2 = \frac{1}{n_h - 1} \sum_{s_h} (y_i - \bar{y}_h)^2.$$

Järeldus 17.2 *Arvestades seost $\bar{Y} = \frac{t}{N}$, avalduvad vastavad avaldised **keskmise hinnangu** puhul järgmiselt:*

$$\hat{\bar{Y}} = \frac{1}{N} \sum_{h=1}^{H} N_h \bar{y}_h,$$

$$V(\hat{\bar{Y}}) = \frac{1}{N^2} \sum_{h=1}^{H} N_h^2 (1 - f_h) S_{yh}^2 / n_h,$$

$$\hat{V}(\hat{\bar{Y}}) = \frac{1}{N^2} \sum_{h=1}^{H} N_h^2 (1 - f_h) s_{yh}^2 / n_h.$$

Näide 17.1 *Reklaamifirmat huvitab, kui palju teha reklaami ühes maakonnas, kuhu kuulub 2 suuremat linna (A ja B) ning maapiirkond. Selleks uuritakse, mitu tundi nädalas inimesed maakonnas keskmiselt televiisorit vaatavad.*

- U_1*: Linn A on ehitatud suure tehase juurde ning enamiku linna elanikkonnast (155 majapidamist) moodustavad tehase töötajad koolieelse lastega.*
- U_2*: Linn B on naabruses asuva suure linna eeslinnaks ning enamik 62 majapidamisest on vanemad inimesed väheste lastega.*
- U_3*: Maapiirkonnas elab 93 majapidamist.*

Raha on 40 majapidamise küsitlemiseks. Otsustatakse, et valimid kihiti on $n_1 = 20$, $n_2 = 8$, $n_3 = 12$. *Igast kihist võetakse LJ valim. Tulemused – TV ees veedetud tunnid nädalas – on toodud allolevas tabelis.*

Kiht 1, Linn A	***Kiht 2, Linn B***	***Kiht 3, Maa***
35 28 26 41 43 29 32 37 36 25 29 31 39 38 40 45 28 27 35 34	*27 4 49 10 15 41 25 30*	*8 15 21 7 14 30 20 11 12 32 34 24*
$n_1 = 20$ $\bar{y}_1 = 33,9$ $s^2_{y1} = 35,358$ $N_1 = 155$	$n_2 = 8$ $\bar{y}_2 = 25,125$ $s^2_{y2} = 232,411$ $N_2 = 62$	$n_3 = 12$ $\bar{y}_3 = 19$ $s^2_{y3} = 87,636$ $N_3 = 93$

Leiame hinnangu TV vaatamisele nädakeskmisele majapidamise kohta koos usalduspiiridega ning suhtelise veaga.

$$\hat{\bar{Y}} = \frac{1}{N}\sum_{h=1}^{H} N_h \bar{y}_h = \frac{1}{310}(155 \cdot 33,9 + 62 \cdot 25,125 + 93 \cdot 19) = 27,7;$$

$$\hat{V}(\hat{\bar{Y}}) = \frac{1}{N^2}\sum_{h=1}^{H} N_h^2(1-f_h)s^2_{yh}/n_h = \frac{1}{310^2}\left[\frac{155^2 \cdot 0,871 \cdot 35,358}{20} + \frac{62^2 \cdot 0,871 \cdot 232,411}{8} + \right.$$
$$\left. + \frac{93^2 \cdot 0,871 \cdot 87,636}{12}\right] = 1,97.$$

Tõenäosusega 95% *saame väita, et keskmiselt vaadatakse TV majapidamises* $27,7 \pm 1,96 \cdot \sqrt{1,97} = 27,7 \pm 2,8$ *tundi nädalas.*

Punktihinnangu suhteline viga, $Suht.v.(\hat{\bar{Y}}) = \frac{\sqrt{\hat{V}(\hat{\bar{Y}})}}{\hat{\bar{Y}}} = \frac{\sqrt{1,97}}{27,7} = 5,1\%$, *mis on väga kõrge näitaja ning järelikult, võib hinnangut usaldada.*

Samas, paneme tähele, et kui leida hinnangud kihiti eraldi ning hinnangu täpsust, siis linnas B tuleb suhteline viga $20,04\%$, *mis on äärmiselt suur.*

Ülesanne 17.1 *Leida linna B hinnang ning hinnangu suhteline viga koos hinnangu usalduspiiridega. Mis aitaks muuta hinnangut täpsemaks?*

Ülesanne 17.2 *Üks firma otsustas hinnata aastase inimtöötundide kadu, mis tekib erinevate õnnetuste tagajärjel. Kuna eri osakondades on juhtumite protsent erinev, siis otsustati lihtsa juhusliku kihtvaliku kasuks. Tabelis on toodud osakondade kogumahud ja valimite andmed (tunnid).*
Hinnata nende põhjal kogu kadu firmas nii punktihinnangu abil kui ka usaldusintervalli abil nivool 95%.

Osakond A	***Osakond B***	***Osakond C***
$N_A = 132$	$N_B = 92$	$N_C = 27$
8 24 0 0 16 32 6 0 16 7 4 4 9 5 8 18 2 0	*4 5 0 24 8 12 3 2 1 8*	*1 8*

Ülesanne 17.3 *Koolis K lõpeb kuues klass matemaatika riigieksamiga. Juhtkond soovib teada prognoosi keskmisele testi tulemusele. Entusiastlikud õpetajad otsustavad sel aastal rakendada kolme erinevat metoodikat matemaatika õpetamisel, mis peaks arvestama individuaalset lähenemist õpilasele. Nii siis, tervet aastat õpivad lapsed matemaatikat metoodikate A, B ja C järgi. Lõpus antakse mõnedele proovitesti. Kuna koolis on kokku 200 õpilast ja on ainult kaks mata õpetajat, kel pole aega kõikide tööde parandamiseks, otsustatakse võtta valim õpilastest mahuga 50 ja katsetada nende põhjal. Direktor, kes on lõpetanud statistika eriala, eeldab, et ühe ja sama metoodika piires punktide varieeruvus peaks olema ühtlasem. Ta otsustab läbiviia lihtsa juhusliku kihtvaliku, kus kihtideks on metoodika.*
Testi tulemused ja õpilaste arvud metoodika järgi on toodud järgmises tabelis.

Metoodika A	***Metoodika B***	***Metoodika C***
$N_A = 55$	$N_B = 80$	$N_C = 65$
80 92 68 85 72 87 85 91 90 81 62 79 61 83	*85 82 48 75 53 73 65 78 49 69 72 81 53 59 68 52 71 61 59 42*	*42 32 36 31 65 29 43 19 53 14 61 31 42 30 39 32*

Leida hinnang keskmisele testi tulemusele koos hinnangu suhtelise veaga.

17.3 Valimi optimaalne paigutus

Kihtvaliku teostamisel on esmaseks ülesandeks kihtide moodustamine üldkogumis. Fikseeritakse tunnused, mille abil objektid jagatakse kihtidesse. Teiseks tähtsaks ülesandeks on valikudisainide määramine kihtides. Kolmandaks oluliseks ülesandeks on valimimahtude määramine kihtides.

Olgu kogu valimimaht n. Temast sõltub hinnangute täpsus. Mida suurem n, seda väiksem dispersioon. Samas mahtu n suurendades kasvab ka uuringu maksumus. Uuringu maksumus on tavaliselt eelarvega fikseeritud, mis paneb kitsendused ka valimimahule. Osutub aga, et valimimahtu n oskuslikult kihtidesse jagades võime nii hinnangute dispersioone kui ka uuringu maksumust vähendada.

Olgu hinnangu $\hat{t}_y$ dispersioon avaldav kujul

$$V = V(\hat{t}_y) = \sum_{h=1}^{H} \frac{A_h}{n_h} + B, \tag{44}$$

kus kihtidesisese hajuvus komponendid A_h ja üldine komponent B ei sõltu valimimahtudest n_h. Olgu uuringu kogumaksumus C avaldatav seosega

$$C = c_0 + \sum_{h=1}^{H} n_h c_h, \tag{45}$$

kus c_0 on üldkulud ja c_h on andmete saamise kulu h-nda kihi objektilt. Suurused c_0 ja c_h on uuringut planeerides teada.

Eesmärgiks on saavutada valimimahtude n_h määramisega kihtides parimaid tulemusi dispersiooni ja maksumuse seisukohalt.

Defnitsioon. Valimimahtude komplekti $n_h, h = 1, ..., H$, nimetatakse optimaalseks, kui kehtib üks järgmistest tingimustest:

1. Etteantud uuringu kogumaksumuse C juures on hinnangu dispersioon $V = V(\hat{t}_y)$ minimaalne.
2. Etteantud hinnangu dispersiooni V juures on uuringu kogumaksumus minimaalne.
3. Etteantud valimimahu n juures on nii dispersioon kui ka maksumus minimaalsed.

Järgnevas tõestame teoreemi, mis annab optimaalsed valimimahud n_h kõigi ülalloetletud eesmärkide saavutamiseks.

Teoreem 17.3 *(valimi optimaalsest paigutusest). Kihtvaliku korral, kus hinnangu dispersioon V ja maksumus C on antud valemitega (44)-(45), saavutatakse valimi optimaalne paigutus, kui*

$$n_h \propto \sqrt{\frac{A_h}{c_h}}, h = 1, ..., H. \tag{46}$$

Tõestus. Ülalloetletud optimaalsuse eesmärkide saavutamiseks tuleb minimiseerida korrutis $V \cdot C$ suuruste n_h suhtes. Jättes kõrvale suurustest n_h mittesõltuvad liikmed, tuleb minimiseerida korrutis

$$K = \left(\sum_{h=1}^{H} \frac{A_h}{n_h}\right)\left(\sum_{h=1}^{H} n_h c_h\right).$$

Kasutame Cauchy-Schwarzi võrratust

$$\sum a_i^2 \sum b_i^2 \geq \left(\sum a_i b_i\right)^2,$$

kus võrdus kehtib parajasti siis, kui $b_i/a_i = const, \forall i$. Saame, et mistahes n_h valiku korral

$$K \geq \left(\sum_{h=1}^{H} \sqrt{A_h c_h}\right)^2.$$

Kuna parem pool ei sõltu suurustest n_h, siis selline n_h valik, mis annab võrduse annab ka K minimaalse väärtuse. Cauchy-Schwarzi teoreemist saame, et võrdus kehtib kui

$$n_h \sqrt{\frac{c_h}{A_h}} = const \text{ ehk } n_h \propto \sqrt{\frac{A_h}{c_h}}.$$

Sellega on teoreem tõestatud.

■

Näeme, et valimi optimaalseks paigutamiseks kihtidesse tuleb rohkem objekte valida sellest kihist, kus kihisisene dispersioonikomponent A_h on suur, aga maksumus c_h väike. Võrdeteguri leidmine sõltub püstitatud optiseerimisülesandest.

Teoreem 17.4 . *Dispersiooni V minimiseerib fikseeritud maksumuse C korral järgmine valimi paigutus*

$$n_h = (C - c_0)\frac{\sqrt{A_h/c_h}}{\sum_{h=1}^{H} \sqrt{A_h c_h}}, h = 1, ..., H, \tag{47}$$

ja minimaalne dispersioon on

$$V_{opt} = \frac{1}{C - c_0}\left(\sum_{h=1}^{H} \sqrt{A_h c_h}\right)^2 + B. \tag{48}$$

Tõestus. Seosest (46) järeldub, et mingi konstandi λ korral kehtib

$$n_h = \lambda\sqrt{\frac{A_h}{c_h}}, h = 1, ..., H.$$

Asendades saadud seose maksumuse avaldisse (45), saame võrdeteguri λ jaoks,

$$\lambda = \frac{C - c_0}{\sum h = 1^H \sqrt{A_h c_h}}.$$

Viimased kaks seost annavadki teoreemi väite (47) n_h kohta. Kasutades optimaalseid valimimahte dispersiooniavaldises (44), saame teoreemi väite (48).

■

Teoreem 17.5 *Maksumuse C minimiseerib fikseeritud dispersiooni V korral järgmine n_h planeering,*

$$n_h = \sqrt{\frac{A_h}{c_h}} \cdot \frac{\sum_{h=1}^{H} \sqrt{A_h c_h}}{V - B}, h = 1, ..., H, \tag{49}$$

ja vastav optimaalne maksumus sel juhul on

$$C_{opt} = c_0 + \frac{1}{V - B}\left(\sum_{h=1}^{H} \sqrt{A_h c_h}\right)^2. \tag{50}$$

Tõestus. Analoogne eelmisele teoreemile.

■

Kogu valimimaht optimaalsete kihisiseste valimimahtude korral on $n = \sum_{h=1}^{H} n_h$.

17.4 Optimaalne valimi paigutus KLJV korral

Kiht lihtne juhuslik valik on praktikas sageli kasutatav disain. Teame, et kogusumma hinnangu dispersioon avaldub sel juhul,

$$V(\hat{t}_y) = \sum_{h=1}^{H} \frac{N_h^2}{n_h}(1 - f_h)S_{yU_h}^2.$$

Valemist näeme, et see dispersioon avaldub just nii nagu meie tulemusteks vaja:

$$A_h = N_h^2 S_{yU_h}^2, B = -\sum_{h=1}^{H} N_h S_{yU_h}^2.$$

Teoreem valimi optimaalsest paigutusest ütleb nüüd, et

$$n_h \propto \frac{N_h S_{yU_h}}{\sqrt{c_h}}. \tag{51}$$

Näeme, et valimi optimaalseks planeerimiseks peame võtma rohkem objekte kihist, mille maht N_h on suurem, milles tunnuse y dispersioon on suurem, aga milles objekti küsitlemine/mõõtmine on odavam. Fikseeritud maksumuse korral on optimaalseks planeeringuks,

$$n_h = \frac{C - c_0}{\sum_{h=1}^{H} N_h S_{yU_h} \sqrt{c_h}} \cdot \frac{N_h S_{yU_h}}{\sqrt{c_h}}. \tag{52}$$

17.5 Alternatiivsed valimi paigutused KLJV korral

Olgu nüüd $c_h = c(const), \forall h$. Tänapäeva praktikas on see enamasti toimiv eeldus. Maksumuse avaldisest (45) saame nüüd, et

$$C - c_0 = c \cdot n. \tag{53}$$

Seega kui uuringu kogumaksumus on ette antud, on sellega fikseeritud ka kogu valimimaht n.

1. Neymani paigutus (1934). Valimimahtude Neymani paigutus on optimaalne paigutus (52) fikseeritud maksumuse korral, kui $c_h = const$. Siis saame valemitest (52)-(53) erijuhu,

$$n_h = n \frac{N_h S_{yU_h}}{\sum_{h=1}^{H} N_h S_{yU_h}}. \tag{54}$$

Paneme tähele, et Neymani paigutus, nagu ka kõik eelnevad valimi paigutused on optimaalsed tunnuse y jaoks. Mõne teise tunnuse z jaoks ei pruugi selline valimi jagamine hea olla.

Näide 17.2 *Koosnegu üldkogum kolmest kihist mahtudega* $N_1 = 150, N_2 = 90, N_3 = 120$. *Eelmistest uuringutest on teada, et* $S_{yU_1} = 100, S_{yU_2} = 200, S_{yU_3} = 300$. *Eeldades konstantset maksumust saaksime optimaalseks paigutuseks kogu valimimahu 12 korral* $n_1 = 2.6, n_2 = 3.1, n_3 = 6.3$, *ehk ümardatult* $n_1 = 3, n_2 = 3, n_3 = 6$.

2. Võrdeline paigutus. Sel juhul on vastavate kihtide osakaalud valimis ja üldkogumis võrdsed:

$$n_h = n \frac{N_h}{N}. \tag{55}$$

Sel juhul on valikusuhted kihtides võrdsed: $f_h = n_h/N_h = n/N = f$. Valemist (54) näeme, et selline paigutus on optimaalne, kui uuritava tunnuse dispersoonid on kõigis kihtides võrdsed, muidugi ka $c_h = const, \forall h$.

Näeme, et võrdeline planeering on tunnuse iseloomu suhtes neutraalne, ühtviisi hea kõikide tunnuste jaoks, aga ei pruugi olla optimaalne ühegi tunnuse jaoks.

Näide 17.3 *Võrdeline paigutus annab eelmise näite andmetel* $n_1 = 5, n_2 = 3, n_3 = 4$.

3. x-optimaalne paigutus. Kuna uuritav tunnus ei ole enne uuringut teada, siis tehakse valimi paigutus kasutades temaga tugevasti korreleeritud teadaolevat x-tunnust.

4. Kogusummaga võrdeline paigutus. Olgu $t_y = \sum_U y_i$ ja $t_{yU_h} = \sum_{U_h} y_i$. Olgu $y_i \geq 0, \forall i$, siis

$$n_h = n\frac{t_{yU_h}}{t_y}.$$

See paigutus on optimaalne, kui variatsioonikordajad on kihiti võrdsed (veendu!):

$$CV_h = \frac{S_{yU_h}}{\bar{Y}_{U_h}} = const, \forall h.$$

Ülesanne 17.4 *Ülesande 17.2 jätk. Paiguta valim mahuga 50 kolmesse kihti kasutades selleks Neymani paigutust. Ülesande 17.2 andmeid saab kasutada kui pilootuuringu andmeid.*

Ülesanne 17.5 *Ülesande 17.3 jätk. Seal on ette antud valimid metoodika A-C järgi. Kas valimipaigutus kihtidesse sobib Neymani või võrdelise paigutusega? Või on hoopis midagi muud?*

Ülesanne 17.6 *Suurfirma, mis toodab tislerpinke, soovib teada saada nende pinkide reitingut skaalal 1..5. Kuna toodangut kasutatakse nii Põhja-Ameerikas, Euroopas kui ka Aasias, on otsustatud kihtvaliku kasuks. Küsitlus viiakse läbi telefoni teel ja seepärast ühe objekti maksumus eri kihtides on erinev. Raha on kokku 26 objekti töötlemiseks. Reitingute varieeruvused eelmistel aastatel, samuti ka 1 objekti küsitlemise maksumus ning kihtide mahud on toodud järgmises tabelis.*
Leida optimaalne valimi paigutus nii maksumuse kui ka dispersiooni suhtes.

Põhja-Ameerika	*Euroopa*	*Aasia*
$c_1 = \$9$	$c_2 = \$25$	$c_3 = \$36$
$s_1^2 = 2,25$	$s_2^2 = 3,24$	$s_3^2 = 3,24$
$N_1 = 112$	$N_2 = 68$	$N_3 = 39$

17.6 LJV ja KLJV võrdlemine

Kogusumma nihketa hinnanguks on TTA disainide korral $\sum_s y_i/\pi_i$. Tahame võrrelda selle hinnangu dispersiooni LJV ja KLJV korral. LJV korral $f_i = n/N$ ja hinnang teiseneb kujule $\hat{t}_y = N\bar{y}$, tema dispersiooniks on

$$V_{LJV}(\hat{t}_y) = \frac{N^2}{n}(1-f)S_y^2. \tag{56}$$

KLJV korral on $\pi_i = n_h/N_h$ kui $i \in U_h$ ja nihketa hinnang saab kuju $\hat{t}_y = \sum_{h=1}^{H} N_h\bar{y}_h$. Selle dispersioon on

$$V_{KLJV}(\hat{t}_y) = \sum_{h=1}^{H} \frac{N_h^2}{n_h}(1-f_h)S_{yU_h}^2. \tag{57}$$

Kumma disaini korral on hinnang on täpsem, kui kogu valimimaht n on sama? Sõltub paljudest asjaoludest. Valimi õige planeerimisega on võimalik saavutada antud kihistuse korral minimaalne dispersioon. Kui maksumus on sama, siis dispersiooni minimiseerib Neymani planeering ja vastav optimaalne dispersioon on

$$V_{opt}(\hat{t}_y) = \frac{N^2}{n}A\sum_{h=1}^{H} W_h(1-f_h)S_{yU_h}, \tag{58}$$

kus $W_h = N_h/N$ on kihi osakaal ja $A = \sum_{h=1}^{H} W_h S_{yU_h}$. Kui dispersioonid kihtides on võrdsed $S_{yU_h} = S_{y0}$ teiseneb dispersioonivalem eriti lihtsale kujule:

$$V_{opt}(\hat{t}_y) = \frac{N^2}{n}(1-f)S_{y0}^2. \tag{59}$$

Kui me ei tea midagi kihi dispersioonidest arvata, kuid kasutame valimi võrdelist paigutust, saame hinnangu dispersioonile valemist (57) kuju:

$$V_{vord}(\hat{t}_y) = N^2\left(\frac{1}{n} - \frac{1}{N}\right)\sum_{h=1}^{H} W_h S_{yU_h}^2. \tag{60}$$

Samas, kui dispersioonid kihtides juhtuvad olema võrdsed, annab see valem sama dispersiooni, mis optimaalne valem (59).

Ülesanne 17.7 *Näidata selle peatüki kõikide väidete kehtivust.*

Valemist (59) näeme taaskord üht kihtvaliku printsiipi, kui objektid on kihtidesse jagatud selliselt, et S_{y0}^2 on väike, on ka hinnangu $\hat{t}_y$ dispersioon väike. Põhimõtteliselt võib kihistamisega saavutada nulldispersiooni.

Üldjuhul, kui valimimaht on sama, kehtivad dispersionide vahel järgmised seosed:

$$V_{opt}(\hat{t}_y) \leq V_{vord}(\hat{t}_y) \leq V_{LJV}(\hat{t}_y).$$

Kokkuvõtteks. KLJV kasutamine LJV asemel on hinnangute täpsuse seisukohalt õigustatud, kui

1. Tunnused on kihtide sees homogeensed (sarnased objektid on samas kihis).

2. Tunnuste keskmised on kihiti erinevad.

Valimimahu võrdeline paigutus on hea, optimaalne paigutus annab väga hea tulemuse kindla uuritava tunnuse korral. Suuremahulistes uuringutes, kus uuritavaid tunnuseid on palju, on mõttekas kasutada võrdelist paigutust, et saada võimalikult hea hinnang kõigi tunnuste korral.

Näide 17.4 *Olgu üldkogum mahuga $N = 6$ jagatud kaheks kihiks, nii et esimesed 3 objekti ühes ja järgmised kolm teises kihis. Seega $N_1 = 3$ ja $N_2 = 3$. Olgu teada ka uuritava tunnuse väärtused $y = (2, 0, 1, 5, 9, 4)$. Olgu $n = 4$. Võrdleme LJV ja võrdelise planeeringuga KLJV, st $n_1 = n_2 = 2$.*

Näeme, et üldkogumis $\bar{Y} = 3,5; \bar{Y}_1 = 1; \bar{Y}_2 = 6$ ja $S_y^2 = 10,7; S_{yU_1}^2 = 1, S_{yU_2}^2 = 7$. Keskväärtuse hinnanguks on LJV korral valimikeskmine $\bar{y} = \sum_s y_i/4$. Võrdelise planeeringuga KLJV korral tuleb selleks samuti tavaline valimikeskmine. Leiame valimikeskmise dispersioonid:

$$V_{LJV}(\hat{\bar{Y}}) = (1-f)S_y^2/n = \frac{6-4}{6} \cdot \frac{10,7}{4} = 0,89$$

$$V_{KLJV}(\hat{\bar{Y}}) = \sum_{h=1}^{2} W_h^2(1-f)S_{yU_h}^2/n_h = \left(\frac{3}{6}\right)^2 \frac{3-2}{3} \cdot \frac{1}{2} + \left(\frac{3}{6}\right)^2 \frac{3-2}{3} \cdot \frac{7}{2} = 0,33.$$

Kommenteeri, mis aitas kaasa dispersiooni vähenemisele.

18 Järelkihistamine

Järelkihistamine on hinnangute täpsuse tõstmise meetod. Seda teostatakse hinnangute arvutamise etapil, st siis kui andmed valimilt on juba kogutud. Seejures valim võib olla võetud tervest üldkogumist (ilma eelneva kihistamiseta) mistahes valikudisainiga.

Järelkihistamisel jagatakse **valimi** objektid gruppidesse – järelkihtidesse. Selleks peab valimi objektidel olema mõõdetud tunnus(ed), mida järelkihistamisel

kasutatakse. Üldkogumi tasemel on vaja teada järelkihtide mahtusid. Kui vajalikud suurused on teada, saab moodustada mitmesuguseid järelkihistusi.

Järelkihistamist kasutatakse hinnangute täpsuse tõstmiseks. Kui õnnestub valim jagada gruppidesse nii, et objektid nendes on võimalikult homogeensed, siis väheneb hinnangute dispersioon.

Järelkihistamist saab kasutada ka kaost põhjustatud nihke vähendamiseks. Selle saavutamiseks jagatakse vastanute valim järelkihtidesse nii, et nendes vastanud on sarnased mittevastanutega.

Järelkihid on oma olemuselt osakogumid. Valimimaht nendes on juhuslik. Üldjuhul pole aga eesmärgiks hinnangute leidmine nendes osakogumites, neid kasutatakse üldkogumihinnangute täpsustamiseks.

Olgu H järelkihti U_h. Need on mittelõikuvad ja ammendavad üldkogumi osad,

$$U = \bigcup_{h=1}^{H} U_h.$$

Olgu N_h järelkihi maht ja $\bar{Y}_h$ järelkihi keskmine:

$$\bar{Y}_h = \frac{\sum_{U_h} y_i}{N_h}.$$

Olgu üldkogumist võetu valim s, mille osa järelkihis U_h tähistame s_h. Olgu disainikaalud $w_i = I_i/EI_i$. Järelkihi keskmise hinnanguks võtame suhte tüüpi hinnangu

$$\tilde{y}_h = \frac{\hat{t}_{yh}}{\hat{N}_h}, \tag{61}$$

kus $\hat{t}_{yh} = \sum_{s_h} w_i y_i$ ja $\hat{N}_h = \sum_{s_h} w_i$. Keskmise (61) baasil moodustatud kogusumma hinnanguks järelkihis on

$$\hat{t}_{yh} = N_h \tilde{y}_h. \tag{62}$$

Näeme, et siin läheb vaja teada järelkihtide mahtusid. Järelkihthinnanguks üldkogumi kogusummale on

$$\hat{t}_{jarel} = \sum_{h=1}^{H} N_h \tilde{y}_h. \tag{63}$$

Järelkihthinnangu dispersiooni leidmine on keeruline, sest liidetavad y_h ei ole sõltumatud, nagu nad olid seda kihtvaliku korral. Hinnang $\hat{t}_{jarel}$ on aga vaadeldav üldisemate regressioon ja kalibreerimishinnangute erijuhuna. Nende dispersiooniavaldised on teada (Särndal jt 1992). Siin toome dispersioonihinnangu valemi,

$$\hat{V}(\hat{t}_{jarel}) = \sum \sum_{s} \breve{\Delta}_{ij} w_i e_i w_j e_j, \text{ kus } e_i = \frac{N_h}{\hat{N}_h}(y_i - \tilde{y}_i), i \in s_h. \tag{64}$$

18.1 Järelkihthinnang LJV korral

LJV korral on $w_i = N/n$, millest

$$\hat{N}_h = \sum_{s_h} w_i = \sum_{s_h} \frac{N}{n} = N\frac{n_h}{n}.$$

Analoogiliselt saame, et

$$\hat{t}_{yh} = \sum_{s_h} w_i y_i = \frac{N}{n} \sum_{s_h} y_i.$$

Avaldisest (61) ja viimasest kahest võrrandist saame kokku LJV korral

$$\tilde{y}_h = \frac{1}{n_h} \sum_{s_h} y_i = \bar{y}_h.$$

Järelkihthinnang LJV korral avaldub valemist (63) järgmiselt:

$$\hat{t}_{jarel} = \sum_{h=1}^{H} N_h \bar{y}_h. \tag{65}$$

Saab näidata, et järelkihistus LJV korral on sama täpne eelkihistusega (KLJV) võrdelise planeeringu korral s.t.

$$V(\hat{t}_{jarel}) = N^2 \frac{1-f}{n} \sum_{h=1}^{H} \frac{N_h - 1}{N-1} S^2_{yU_h}$$

ja

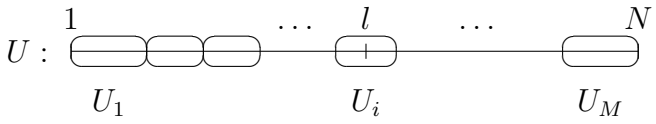
$$\hat{V}(\hat{t}_{jarel}) = N^2 \frac{1-f}{n} \sum_{h=1}^{H} \frac{N_h^2}{n_h} s^2_{ys_h}.$$

19 Klastervalik

Gruppe ehk klastreid on võimalik üldkogumis moodustada peaaegu alati. Näiteks, võib rahvastiku uurimisel kasutada klastrite rollis
· maju, majade plokke;
· tänavaid;
· külasid, asulaid.

Laste üldkogumis kasutatakse klastritena
· koole/lasteaedu;
· klasse/rühmasid.

Olgu üldkogum $U = (1, \ldots, N)$ jagatud klastriteks



U_i – i. klaster;

N_i – klastri U_i suurus;

M – klastrite arv;

$U = \cup_{i=1}^{M} U_i, \quad U_i \cap U_j = \emptyset$;

$N = \sum_{i=1}^{M} N_i$.

Definitsioon 19.1 *Klastervalik on protseduur, mille korral valik toimub klastrite seast vastavalt mingile tõenäosuslikule disainile. Valimisse sattunud klastritest võetakse valimisse kõik selle klastri objektid.*

Valikuühik = klaster.

Olgu I_{ci} klastri U_i valikuindikaator. Siis

$I_c = (I_{c1}, \ldots, I_{cM})$	–	klastervalikule vastav vektor;
$p_c(k_c) = \Pr(I_c = k_c)$	–	klastritele vastav disain;
$E(I_{ci})$	–	i. klastri oodatav valikute arv = i. klastri kõikide objektide oodatav valikute arv;

Disaini $p_c(k_c)$ poolt on määratud ka $V(I_{ci})$ ja $Cov(I_{ci}, I_{cj})$.

$m = \sum_{i=1}^{M} I_{ci}$ – juhuslik suurus, mis vastab klastervalimi mahule;

$n = \sum_{i=1}^{M} N_i I_{ci}$ – juhuslik suurus, mis vastab objektide valimimahule;

Paneme tähele, et TGA disainide korral võivad ühed ja samad klastrid sattuda valimisse >1 arv kordi.

19.1 Hindamine klastervaliku korral

Olgu $Y_i = \sum_{U_i} y_k$ kogusumma klastris U_i.
Siis

$$t_y = \sum_U y_i = \sum_{i=1}^M Y_i \text{ – kogusumma on klastrite kogusummade summa.}$$

Kuna valikuühikuks on klaster ja kõiki objekte klastrist küsitletakse, siis valimisse sattunud klastri U_i jaoks on kogusumma Y_i teada.

Kuna t_y on klasterväärtuste Y_i kogusumma, siis saame selle hindamiseks rakendada varasemaid tulemusi. Teoreemis 6.1 peame vaid asendama y_i Y_i vastu ja I_i vastavalt I_{ci} vastu.

Teoreem 19.1 *(Hindamine klastervaliku korral) Klastervlaiku korral on kogusumma $t_y = \sum_U y_i = \sum_{i=1}^M Y_i$ nihketa hinnanguks*

$$\hat{t}_y = \sum_{i=1}^M \frac{I_{ci} Y_i}{E(I_{ci})} \tag{66}$$

vastava punkthinnanguga

$$\hat{t}_y = \sum_{i \in s_c} w_i Y_i, \text{ kus } w_i = \frac{k_{ci}}{E(I_{ci})}.$$

Hinnangu $\hat{t}_y$ dispersioon on

$$V(\hat{t}_y) = \sum_{i=1}^M \sum_{j=1}^M \Delta_{cij} \frac{Y_i}{E(I_{ci})} \frac{Y_j}{E(I_{cj})},$$

nihketa hinnangufunktsiooniga

$$\hat{V}(\hat{t}_y) = \sum_{i=1}^M \sum_{j=1}^M \frac{\Delta_{cij}}{E(I_{ci} I_{cj})} \frac{Y_i}{E(I_{ci})} \frac{Y_j}{E(I_{cj})} I_{ci} I_{cj}, \tag{67}$$

vastava punkthinnanguga

$$\hat{V}(\hat{t}_y) = \sum_{i \in s_c} \sum_{j \in s_c} \frac{\Delta_{cij}}{E(I_{ci} I_{cj})} w_i Y_i w_j Y_j,$$

kus $\Delta_{cij} = Cov(I_{ci}, I_{cj})$.

Kui klasterdisain on selline, et klastrite valimimaht on fikseeritud,

$$\sum_{i=1}^{M} I_{ci} \equiv m, \text{ constant},$$

siis kogusumma dispersiooni ja selle hinnangut võib leida alternatiivselt

$$V(\hat{t}_y) = -\frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M}\Delta_{cij}\Big(\frac{Y_i}{E(I_{ci})} - \frac{Y_j}{E(I_{cj})}\Big)^2. \tag{68}$$

ja

$$\hat{V}(\hat{t}_y) = -\frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M}\frac{\Delta_{cij}}{E(I_{ci}I_{cj})}\Big(\frac{Y_i}{E(I_{ci})} - \frac{Y_j}{E(I_{cj})}\Big)^2 I_{ci}I_{cj}. \tag{69}$$

Fikseeritud mahuga klasterdisainide korral tuleks eelistada valemit (69) valemile (67), sest see on stabiilsem (väiksema varieeruvusega) ja ei vii negatiivsetele tulemustele.

Uurides valemit (68) on võimalik teha kaks tähtsat järeldust.

1. Kui klastri kogusummad on võrdelised oodatavate valikute arvudega,

   $$Y_i \propto E(I_{ci}),$$

   siis $V(\hat{t}_y) = 0$, ja hinnang $\hat{t}_y$ defineeritud valemiga (66) on täpne.

   Klastri summad Y_i pole teada enne uuringut. Kui aga on teada abitunnuse summad $X_i = \sum_{U_i} x_k$, ja abitunnus x on tugevalt positiivselt korreleeritud uuritava tunnusega y, siis on võimalik hinnangu $\hat{t}_y$ täpsust tõsta kui moodustada valikuindikaatorid nii, et kehtiks

   $$E(I_{ci}) \propto X_i.$$

2. Kui on teada kõikide klastrite suurused N_i, siis on võimalik valida

   $$E(I_{ci}) \propto N_i.$$

   Kasutades $Y_i = N_i\bar{Y}_i$, kus $\bar{Y}_i$ on klastri keskmine, saame esitada vahe valemis (68) järgmiselt

   $$\frac{N_i\bar{Y}_i}{E(I_{ci})} - \frac{N_j\bar{Y}_j}{E(I_{cj})} = const\ (\bar{Y}_i - \bar{Y}_j).$$

   Nüüd me näeme, et mida vähem varieeruvad klastri keskmised $\bar{Y}_i$, seda väiksem tuleb dispersioon (68) $\Rightarrow$ täpsem hinnang $\hat{t}_y$.

Saame kasutada ANOVA lahutus uuritava tunnuse varieeruvusele:

$$\begin{aligned} SST &= \sum_{U} (y_i - \bar{Y})^2 = \sum_{i=1}^{M} \sum_{l \in U_i} (y_l - \bar{Y})^2 = \\ &= \underbrace{\sum_{i=1}^{M} \sum_{U_i} (y_l - \bar{Y}_l)^2}_{SSW} + \underbrace{\sum_{i=1}^{M} \sum_{U_i} (\bar{Y}_l - \bar{Y})^2}_{SSB}. \end{aligned}$$

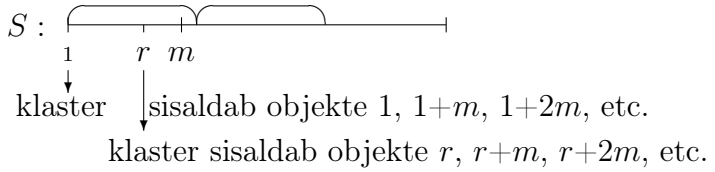
Uuritava tunnuse varieeruvus SST on fikseeritud. Et saada väiksemat SSB, peame suurendama SSW.

Järelikult, klastervalik on efektiivne siis kui

- $E(I_{ci}) \propto N_i$;
- klastrid on võimalikult heterogeensed (sisaldavad erinevaid väärtusi y_i) – viib suurele SSW.

Märkus 19.1 *Kihtvaliku korral me vastupidiselt klastervalikule soovime saada võimalikult homogeenseid (uuritava tunnuse suhtes) kihte.*

Märkus 19.2 *Süstemaatilist valikult on võimalik vaadelda klastervaliku erijuhuna valimimahuga 1.*



Üldkogum U koosneb m klastrist, millest üks võetakse valimisse. Ühe objekti põhjal pole võijmalik hinnata dispersiooni (sellega puutusime kokku süstemaatilise valiku juures).

19.2 Lihtne juhuslik klastervalik

Selle disaini korral valitakse klastreid lihtsa juhuvaliku protseduuri järgi. Kõiki objekte valitud klastritest mõõdetakse. Valikusuhe klastrite jaoks on

$$f_c = \frac{m}{M},$$

kus m valimisse võetud klastrite arv ja M on klastrite arv üldkogumis. Paneme tähele, et $EI_{ci} = f_c$ kaasamistõenäosus iga klastri jaoks.

Kasutades hindamisteoreemi lihtsa juhuvaliku korral tagasipanekuta ja arvestades, et klastrisumma on $Y_i = \sum_{U_i} y_i$ teada valimisse sattunud klastrite jaoks saame panna kirja järgmise teoreemi.

Teoreem 19.2 *(Lihtne juhuslik klastervalik) Lihtsa juhusliku klastervaliku korral on nihketa hinnang üldkogumi summale $t = \sum_U y_i$ järgmine*

$$\hat{t} = \frac{M}{m} \sum_{i=1}^{M} I_{ci} Y_i$$

vastava punkthinnanguga

$$\hat{t} = \frac{M}{m} \sum_{s_c} Y_i = M \bar{Y}_{sc},$$

kus $\bar{Y}_{sc}$ on klastersummade valimikeskmine, $\bar{Y}_{sc} = \frac{1}{m} \sum_{s_c} Y_i$. Hinnangu $\hat{t}$ dispersioon on

$$V(\hat{t}) = M^2(1 - f_c)\frac{S_{cy}^2}{m},$$

mille nihketa hinnanguks on

$$\hat{V}(\hat{t}) = M^2(1 - f_c)\frac{s_{cy}^2}{m},$$

kus

$$S_{cy}^2 = \frac{1}{M-1} \sum_{i=1}^{M} (Y_i - \bar{Y}_c)^2,$$

$$s_{cy}^2 = \frac{1}{m-1} \sum_{s_c} (Y_i - \bar{Y}_{cs})^2,$$

$$\bar{Y}_c = \frac{1}{M} \sum_{i=1}^{M} Y_i.$$

Märkus 19.3 . *Üldiselt ühe ja sama valimimahu korral n on lihtne juhuslik klastervalik vähem efektiivne kui tavaline lihtne juhuslik valik. Seda seetõttu, et hinnangu dispersiooni valemis on objektide valimimahu n asemel klastrite valimimaht m, kuid alati $m << n$. Klastervalikut eelistatakse sageli teistel (praktilistel) põhjustel: freimi on võimalik moodustada ainult klastrite tasemel; mugavus objektide mõõtmisel; rahaline kokkuhoid jne.*

Ülesanne 19.1 *Üks tudeng soovib teada saada tema ühikas elavate tudengite õpingute keskmist tulemust. Selle asemel, et moodustada kõikidest tudengitest freimi ja võtta tavalise juhusliku valimi, paneb ta hoopis tähele, et ühikas on 100 tuba, igas toas elab 4 tudengit. Ta valib juhuslikult 5 tuba ning küsib nendes kõikide tudengite käest õpingute keskmist. Tulemused on järgmised:*

Tudengi Nr	*Tuba (klaster)*				
1	*3,08*	*2,36*	*2,00*	*3,00*	*2,68*
2	*2,60*	*3,04*	*2,56*	*2,88*	*1,92*
3	*3,44*	*3,28*	*2,25*	*3,44*	*3,28*
4	*3,04*	*2,68*	*1,88*	*3,64*	*3,20*
Kokku	*12,16*	*11,36*	*8,96*	*12,96*	*11,08*

Ülesanne 19.2 *Uuritavaks tunnuseks on e-lugerite kasutamine noorte seas (1-on kasutanud käesoleva nädala jooksul, 0-ei kasutanud). Olgu väikelinnas A viis gümnaasiumi (kokku 1000 õpilast), millest LJ klastervalikuga tuli valimisse 2 kooli (U_1 ja U_3). Kõiki õpilasi valitud koolidest küsitleti ja saadi järgmised andmed*

$$U_1 : N_1 = 180; Y_1 = 30$$

$$U_3 : N_3 = 200; Y_3 = 40$$

Leida vahemikhinnang e-lugerite osakaalule $\bar{Y}$ usaldusnivool 0,95.

20 Kahe-astmeline valik

Kahe-astmeline valik on protseduur, kus

1. astmel moodustatakse klaster-valim vastavalt mõnele tõenäosuslikele disainile ja
2. astmel valitud klastritest võetakse omakorda valimid.

Siinjuures 1. ja 2. astme valikudisainid ei sõltu üksteisest (võivad olla samad, võivad olla aga erinevad). Samuti ka eriklastrites võib rakendada erinevat valikudisaini.

Seega, **kihtvalik** on kahe-astmelise disaini erijuht, kui 1. astmel toimub kõikne klastervalik.

Klastervalik on kahe-astmelise valiku erijuhut, kui 2. astmel toimub kõikne valik igas valitud klastris.

Kirjanduses nimetatakse sageli 1. astme valikuühikuid (ehk klastreid) PSU=Primary Sampling Unit; 2. astme ühikuid - SSU=Secondary sampling units.

20.1 Tähistused

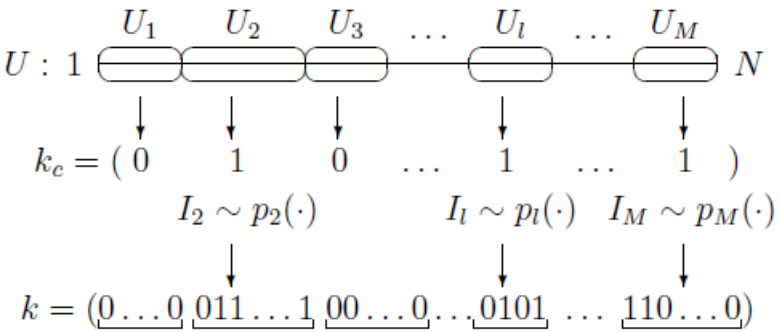
Olgu U_e - PSU, kusjuures $U_e \subset U$, $\bigcup_{e=1}^{M} U_e = U$, $U_e \bigcap U_g = \emptyset$ ja N_e klastri U_e suurus.

Kasutame erinevaid valikuvektoreid:

$I_c = (I_{c1}, ..., I_{ce}, ..., I_{cM})$	valikuvektor klastrite (PSU-de) jaoks;
$I_c \sim p_c(k_c)$	valikudisain PSU jaoks;
$I_e \sim p_e(k_e), e = 1, ..., M$	valikudisain SSU jaoks klastri U_e sees;
$I_e = (I_{\cdot\|e}, ..., I_{i\|e}, ..., I_{\cdot\|e})$	valikuvektor pikkusega N_e SSU jaoks klastris U_e.

Tähtsad eeldused:

- valikud erinevates klastrites (PSU-des) on üksteisest sõltumatud;
- valikud 2. astmel ei sõltu valikust 1. astmel.



Valimimahud:

- $m = \sum_{e=1}^{M} k_{ce}$ valitud klastrite arv = klastervalimi maht;
- $n = \sum_{i=1}^{N} k_i$ lõplik valimimaht.

20.2 Hindamine kahe-astmelise valiku korral

Olgu $Y_e = \sum_{i \in U_e} y_i$ uuritava tunnuse kogusumma klastris U_e. Siis ÜK kogusummat saab esitada kujul

$$t = \sum_{e=1}^{M} Y_e. \tag{70}$$

Kui klastrid on valitud, siis saame nendes leida nihketa hinnangud $\hat{Y}_e$ klastri summadele. Kogu ÜK summat saab hinnata järgmise nihketa hinnangu abil:

$$\hat{t} = \sum_{e=1}^{M} \frac{\hat{Y}_e}{E(I_{ce})} I_{ce}. \tag{71}$$

Edaspidi eeldame, et tegemist on TTA klastervalikuga, ehk $I_{ce} \in \{0, 1\}$.

Klastersumma Y_e hinnangu saame tavalise nihketa hinnangu abil (kasutades ÜHT):

$$\hat{Y}_e = \sum_{i \in U_e} \frac{y_i I_{i|e}}{E(I_{i|e})}, \tag{72}$$

kus $I_{i|e}$ on i-nda objekti valikuindikaator klastris U_e.

Üldisest hindamisteoreemist teame, et hinnangu $\hat{Y}_e$ dispersioon avaldub järgmiselt:

$$V(\hat{Y}_e) = \sum \sum_{i,j \in U_e} \Delta_{ij|e} \frac{y_i}{E(I_{i|e})} \frac{y_j}{E(I_{j|e})} \tag{73}$$

ja tema nihketa hinnang järgmise valemi abil:

$$\hat{V}(\hat{Y}_e) = \sum \sum_{i,j \in U_e} \frac{\Delta_{ij|e}}{E(I_{i|e} I_{j|e})} \frac{y_i}{E(I_{i|e})} \frac{y_j}{E(I_{j|e})} I_{i|e} I_{j|e}. \tag{74}$$

Need dispersioonide valemid kehtivad konkreetse klastri sees. Kogu kaheastmelise protsessi jooksul tekkinud varieeruvust pole nii lihtne leida. Peame arvestama varieeruvuse nii 1. kui ka 2. astmel. Siin saame kasutada tingliku keskväärtuse ja tingliku dispersiooni valemeid:

$$E(\hat{t}) = E_{p_c} E(\hat{t}|I_c), \tag{75}$$

$$V(\hat{t}) = E_{p_c} V(\hat{t}|I_c) + V_{p_c} E(\hat{t}|I_c). \tag{76}$$

Valem (75) tähendab, et esmalt leiame keskväärtuse igas klastris eraldi (II astme valikudisaini suhtes) ja seejärel keskmistame need klastri keskmised omakorda (leiame keskväärtuse I astme disaini suhtes).

Arvestades valemid (71)-(72), kontrollime hinnangu nihketuse omaduse valemi (75) abil:

$$E(\hat{t}) = E_{p_c} E(\hat{t}|I_c) = E_{p_c} E\left(\sum_{e=1}^{M} \frac{\hat{Y}_e}{E(I_{ce})} I_{ce} | I_c \right) = E_{p_c} \left(\sum_{e=1}^{M} \frac{Y_e}{E(I_{ce})} I_{ce} \right) = \sum_{e=1}^{M} Y_e = t, \tag{77}$$

kus $E(\hat{Y}_e|I_c) = Y_e$.

Hinnangu $\hat{t}$ kogu dispersiooni saamiseks leiame kõigepealt valemi (76) 2. liidetava. Selleks paneme tähele, et valemist (77)

$$E(\hat{t}|I_c) = \sum_{e=1}^{M} \frac{Y_e}{E(I_{ce})} I_{ce}, \tag{78}$$

mis on nihketa hinnanguks parameetrile t klasterdisaini $p_c(\cdot)$ suhtes. Sellele hinnangule saame rakendada ÜHT, et leida tema dispersiooni $V_{p_c}[E(\hat{t}|I_c)]$ (ja seda disaini $p_c(\cdot)$ suhtes):

$$V_{p_c}E(\hat{t}|I_c) = V_{p_c}[E(\hat{t}|I_c)] = \sum_{e=1}^{M}\sum_{g=1}^{M}\Delta_{ceg}\frac{Y_e}{E(I_{ce})}\frac{Y_g}{E(I_{cg})} = V_2, \qquad (79)$$

kus $\Delta_{ceg} = Cov(I_{ce}, I_{cg})$.

Nüüd, leiame valemi (76) 1. liidetava,

$$V(\hat{t}|I_c) = V\left(\sum_{e=1}^{M}\frac{\hat{Y}_e}{E(I_{ce})}I_{ce}|I_c\right) =$$

Arvestades, et 2. astmel toimub valik klastritest üksteisest sõltumata, siis saab dispersioonimärgiga V summa sisse minna,

$$= \sum_{e=1}^{M} V(\hat{Y}_e)\frac{I_{ce}^2}{[E(I_{ce})]^2},$$

kus $V(\hat{Y}_e|I_c) = V(\hat{Y}_e)$ on 2. astme valiku dispersioon, mis ei sõltu 1. astme valikust. Lisaks, $I_{ce}^2 = I_{ce}$ TTA disaini jaoks.

Järgmisena,

$$E_{p_c}V(\hat{t}|I_c) = E_{p_c}\left[\sum_{e=1}^{M}V(\hat{Y}_e)\frac{I_{ce}}{(EI_{ce})^2}\right] = \sum_{e=1}^{M}\frac{V(\hat{Y}_e)}{E(I_{ce})} = V_1. \qquad (80)$$

Sõnastame eelnevat teoreemina.

Teoreem 20.1 *(Kahe-astmeline valik, TTA). Kahe-astmelise disaini korral nihketa hinnang ÜK summale on antud valemite (71)-(72) abil. Selle hinnangu dispersioon avaldub kui*

$$V(\hat{t}) = V_1 + V_2,$$

kus V_1 on antud valemis (79) ja V_2 valemis (80).

20.3 Kahe-astmeline lihtne juhuslik valik

Selle valiku korral toimub 1. astmel LJ klastervalik TTA, kus

$$f_c = E(I_{ce}) = \frac{m}{M} \qquad (81)$$

ja teisel astmel igast valitud klastrist võetakse omakorda LJ valik TTA, kus

$$f_e = \frac{n_e}{N_e}. \tag{82}$$

Nüüd, valemitest (71)-(72) saab leida nihketa hinnangu ÜK summale:

$$\hat{t} = \sum_{e=1}^{M} \frac{M}{m} I_{ce} \sum_{i \in U_e} \frac{N_e}{n_e} I_{i|e} y_i. \tag{83}$$

Vastav punktihinnang on

$$\hat{t} = \frac{M}{m} \sum_{e \in s_c} \frac{N_e}{n_e} \sum_{i \in s_e} y_i = \frac{M}{m} \sum_{s_c} N_e \bar{y}_e,$$

kus
s_c on klastervalim;
$i \in s_e$ summa üle valimi klastrist U_e;
$\bar{y}_e$ valimikeskmine valimis s_e.

Hinnangu ÜK summale saab kaalude abil kirja panna järgmiselt:

$$\hat{t} = \sum_{i \in s} w_i y_i,$$

kus $w_i = \frac{M}{m}\frac{N_e}{n_e}$. Paneme tähele, et objektidel erinevatest klastritest on erinevad kaalud!.

Hinnangu dispersiooni $V(\hat{t})$ saab vajaduse korral välja kirjutada Teoreemist 20.1.

20.4 Isekaaluv kahe-astmeline valik

Praktikud eelistavad isekaaluvaid kahe-astmelisi valikuid, kus lõplikelt valikuühikutel on võrdsed kaalud. Sel juhul valimi struktuur vastab ÜK struktuurile ja valimi karakteristikud (keskmine, osakaal) on hinnanguteks vastavatele ÜK parameetritele.

Eeldame, et mõlemal astmel on teostatud TTA valik. Siis

$E(I_{ce}) = \pi_{ce}$ – e-nda klastri kaasamistõenäosus; $E(I_{i|e}) = \pi_{i|e}$ – i-nda objekti kaasamistõenäosus klastris U_e.

Valemitest (71)-(72) hinnang ÜK summale tuleb järgmine:

$$\hat{t} = \sum_{e=1}^{M} \sum_{i \in U_e} \frac{y_i}{\pi_{i|e}\pi_{ce}} I_{i|e} I_{ce}.$$

Disain on isekaaluv, kui

$$\pi_{i|e}\pi_{ce} = c(const), \forall i \in U.$$

Seda on võimalik saavutada kahel viisil:

1. Klastrite kaasamistõenäosused on võrdelised klastri suurustega, $\pi_{ce} = N_e\frac{m}{N}, \forall e$, kusjuures kehtib

$$m = \sum_{e=1}^{M} \pi_{ce}.$$

   2. astmel valitakse iga klastri jaoks võrdne objektide arv valimisse nii, et $\pi_{i|e} = \frac{n_0}{N_e}$. Kokku saame, et

$$\pi_{ce}\pi_{i|e} = N_e\frac{m}{N}\frac{n_0}{N_e} = \frac{mn_0}{N} = \frac{n}{N}, \forall i \in U.$$

   Siin n on lõplik valimimaht.

   Sellise disaini korral kõikidel intervjueeriatel on võrdne arv inimesi küsitlemiseks klastrites.

2. Esimesel astmel valitakse klastrid võrdse tõenäosusega,

$$f_c = \frac{m}{M}, \forall e.$$

   Teisel astmel valitakse objektid võrdse tõenäosusega, st igas klastris on konstantne kaasamistõenäosus,

$$\pi_{i|e} = \frac{n_e}{N_e} = f, (const).$$

   Sellisesl juhul

$$\pi_{ce}\pi_{i|e} = \frac{m}{M}f,$$

   mis on võrne iga objekti jaoks.

   Selliseks disainiks on LJ valik mõlemal astmel, kus 2. astmel toimub valimi võrdeline paigutus valitud klastritesse.

21 Abiinformatsiooni kasutamine hinnangutes

Eeldame, et on moodustatud valim vastavalt mingisugusele valikudisainile, on saadud andmed ning ees ootab hindamine. Siiani on meil kasutuses olnud nihketa hinnang kujul

$$\hat{t} = \sum_{U} \frac{y_i I_i}{EI_i},$$

mille punkthinnang valimi kaudu on järgmine:

$$\hat{t} = \sum_s w_i y_i, \quad w_i = \frac{k_i}{EI_i}.$$

Kaalud w_i selles hinnangus põhinevad pöördväärtusel EI_i, ehk sõltuvad ainult disainist $I \sim p(k)$.

Osutub, et hinnangut summale $t = \sum_U y_i$ on võimalik muuta täpsemaks varieeruvuse mõttes, kui kasutada abiinformatsiooni ja seda just kaalude moodustamise etapil.

Abiinformatsiooniks loetakse

- tunnuseid, mille väärtused on teada kõikide objektide jaoks üldkogumist;
- abitunnuste summasid (näiteks osakogumite kaupa või lihtsalt terves ÜK-s, näiteks meeste arv);
- osakogumite suuruseid üldkogumis (näiteks kihtide mahud).

21.1 Regressioonimudel üldkogumi jaoks

Olgu y_i uuritava tunnuse väärtus objekti i jaoks, $i \in U$. Ja olgu $\boldsymbol{x}_i = (x_{1i}, ..., x_{ji})^T$ on abitunnuste vektor, mis on teada iga objekti $i, i \in U$ jaoks.

Eeldame järgmist abimudelit üldkogumis:

1. väärtus $y_i, i \in U$ on juhusliku suuruse $Y_i \sim \xi$ realisatsioon (jaotusega ξ);
2. jaotuse ξ momendid on järgmiselt defineeritud:
   - $E_\xi Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} = \sum_{j=1}^J \beta_j x_{ji}$,
   - $V_\xi Y_i = \sigma_i^2$;
3. $\boldsymbol{x}_i$ mittejuhuslikud.

Mudel ütleb seda, et võrdsete $\boldsymbol{x}_i$ korral üldkogumis väärtus y_i võib varieeruda, kuid see varieeruvus toimub tema keskväärtuse $\boldsymbol{x}_i^T \boldsymbol{\beta}$ ümbruses (regressioonijoon) dispersiooniga σ_i^2.

Antud juhul on $\boldsymbol{\beta} = (\beta_1, ..., \beta_J)^T$ regressioonikordajate vektor.

Märkame, et regressioonimudel on eeldatud **üldkogumi** väärtustele $y_i, i \in U$.

Kui kõik väärtused oleksid teada, siis saaksime regressioonikordajate leidmiseks kasutada tavalist kaalutud vähimruutude hinnangut kujul

$$\hat{\boldsymbol{\beta}} \overset{tähistame}{=} \mathbf{B} = \left[\sum_U \frac{\boldsymbol{x}_i \boldsymbol{x}_i^T}{\sigma_i^2}\right]^{-1} \sum_U \frac{\boldsymbol{x}_i y_i}{\sigma_i^2}. \tag{84}$$

Prognoositud väärtused y_i-le on $\boldsymbol{x}_i^T \mathbf{B}$ ja jäägid üldkogumi mudeli järgi on

$$E_i = y_i - \boldsymbol{x}_i^T \mathbf{B}, i \in U. \tag{85}$$

Märkame, et suurused E_i ja $\mathbf{B}$ sõltuvad ÜK väärtustest $y_i, i \in U$ ja seega tundmatud. Neid peab hindama valimi põhjal. Paneme samuti tähele, et suurus $\mathbf{B}$ koosneb kahe ÜK summa korrutisest:

$$\sum_U \frac{\boldsymbol{x}_i \boldsymbol{x}_i^T}{\sigma_i^2} \text{ – maatriksite summa, mõõtmetega } J \times J;$$

$$\sum_U \frac{\boldsymbol{x}_i y_i}{\sigma_i^2} \text{ – vektorite summa, mõõtmetega } J \times 1.$$

Neid summasid saame hinnata kasutades ÜHT. Lisaks eeldame edaspidi, et tegemist on TTA disainiga. Siis saame järgmise hinnangu valimist:

$$\hat{\mathbf{B}} = \left[\sum_s \frac{\boldsymbol{x}_i \boldsymbol{x}_i^T}{\sigma_i^2 \pi_i}\right]^{-1} \sum_s \frac{\boldsymbol{x}_i y_i}{\sigma_i^2 \pi_i}. \tag{86}$$

Toetudes sellele hinnangule saab arvutada valimi põhjal leitud prognoosihinnanguid:

$$\hat{y}_i = \boldsymbol{x}_i^T \hat{\mathbf{B}}, i \in U. \tag{87}$$

Mudeli jääkide hinnangud on sel juhul:

$$e_i = y_i - \hat{y}_i, \ i \in s. \tag{88}$$

kus e_i on leitavad üksnes valimis.

21.2 Regressioonihinnang

Et regressioonihinnangut saada, kirjutame üldkogumi summa t ümber:

$$t = \sum_U y_i = \sum_U \hat{y}_i + \sum_U (y_i - \hat{y}_i), \tag{89}$$

kus $\hat{y}_i$ on teada kõikide $i \in U$ korral, ja y_i - üksnes valimis.

Hindame teise liikme avaldises (89) kasutades nihketa hinnangut ÜHT järgi. See viib nn regressioonihinnanguni kujul:

$$\hat{t}_r = \sum_U \hat{y}_i + \sum_s \frac{y_i - \hat{y}_i}{\pi_i}. \tag{90}$$

Näeme, et regressioonihinnang koosneb prognooside summast, millele on liidetud mudel jääkidest koosnev korrigeerimisliige.

Sageli, praktilistel põhjustel esitatakse regressioonihinnang kaalude ja kaalusid korrigeeriva kordaja abil. Selleks kirjutatakse regressioonihinnang ümber:

$$\hat{t}_r = \sum_s \frac{y_i}{\pi_i} + \sum_U \hat{y}_i - \sum_s \frac{\hat{y}_i}{\pi_i}. \tag{91}$$

Nüüd avaldise (87) abil saame

$$\begin{aligned}
\hat{t}_r &= \sum_s \frac{y_i}{\pi_i} + \sum_U \boldsymbol{x}_i^T \hat{\mathbf{B}} - \sum_s \frac{\boldsymbol{x}_i^T \hat{\mathbf{B}}}{\pi_i} = \\
&= \sum_s \frac{y_i}{\pi_i} + \left(\sum_U \boldsymbol{x}_i^T \right) \hat{\mathbf{B}} - \left(\sum_s \frac{\boldsymbol{x}_i^T}{\pi_i} \right) \hat{\mathbf{B}} \\
&= \sum_s \frac{y_i}{\pi_i} + \left(\sum_U \boldsymbol{x}_i^T - \sum_s \frac{\boldsymbol{x}_i^T}{\pi_i} \right) \hat{\mathbf{B}}.
\end{aligned}$$

Edasi rakendame avaldise (86) $\hat{B}$ jaoks:

$$\begin{aligned}
\hat{t}_r &= \sum_s \frac{y_i}{\pi_i} + \underbrace{\left(\sum_U \boldsymbol{x}_i^T - \sum_s \frac{\boldsymbol{x}_i^T}{\pi_i} \right)}_{1\times J} \underbrace{\left[\sum_s \frac{\boldsymbol{x}_i \boldsymbol{x}_i^T}{\sigma_i^2 \pi_i} \right]^{-1}}_{J\times J} \underbrace{\sum_s \frac{\boldsymbol{x}_i y_i}{\sigma_i^2 \pi_i}}_{J\times 1} \\
&= \sum_s \frac{y_i}{\pi_i} \underbrace{\left[1 + \left(\sum_U \boldsymbol{x}_i^T - \sum_s \frac{\boldsymbol{x}_i^T}{\pi_i} \right) \left[\sum_s \frac{\boldsymbol{x}_i \boldsymbol{x}_i^T}{\sigma_i^2 \pi_i} \right]^{-1} \frac{\boldsymbol{x}_i}{\sigma_i^2} \right]}_{g_{is}}
\end{aligned}$$

Lõplikult, regressioonihinnangut saab esitada kujul

$$\hat{t}_r = \sum_s w_i g_{is} y_i, \tag{92}$$

kus

w_i – on valikukaal,

$$g_{is} = 1 + \left(\sum_U \boldsymbol{x}_i^T - \sum_s \frac{\boldsymbol{x}_i^T}{\pi_i} \right) \left[\sum_s \frac{\boldsymbol{x}_i \boldsymbol{x}_i^T}{\sigma_i^2 \pi_i} \right]^{-1} \frac{\boldsymbol{x}_i}{\sigma_i^2}. \tag{93}$$

Valemist (93) näeme, et kui

$$\sum_U x_i^T \approx \sum_s \frac{x_i^T}{\pi_i},$$

ehk x-summad on ligikaudu võrdsed nende hinnangutega, siis $g_{is} \approx 1$, ja

$$\hat{t}_r \approx \sum_s w_i y_i.$$

Regressioonihinnangu dispersiooni tuletuskäiku antud kursuse raames ei vaatle. Siin toome ainult valemi.

Teoreem 21.1 *(Regressioonihinnang). Regressioonihinnang ÜK summale* $t = \sum_U y_i$ *on antud valemiga (90) ja alternatiivse valemiga (92), mille ligikaudne dispaersioon on*

$$V(\hat{t}_r) = \sum \sum_U \Delta_{ij}(w_i E_i)(w_j E_j) \tag{94}$$

ja dispersioonihinnanguga

$$\hat{V}(\hat{t}_r) = \sum \sum_{i,j \in s} \frac{\Delta_{ij}}{\pi_{ij}}(w_i g_{is} e_i)(w_j g_{js} e_j), \tag{95}$$

kus üldkogumi taseme jäägid E_i *on defineeritud valemiga (85), valimist arvutatavad jäägid* e_i *valemiga (88) ja g-kaalud valemiga (93).*

Märkus 21.1 *Teoreemi avaldisest (94) on näha, et mida väiksemad on jäägid* E_i*, seda väiksem on* $V(\hat{t}_r)$*. Jäägid* E_i *näitavad, kui hästi regressioonimudel sobib andmetega. Järelikult, mida parem on regressioonimudel y ja* $\boldsymbol{x}$ *vahel, seda täpsem tuleb hinnang* $\hat{t}_r$*.*

Märkus 21.2 *Valemist (93) näeme, et tunnuste* $\boldsymbol{x}_i$ *üksikuid väärtused peame teadma ainult valimisse sattunud objektide jaoks, üldkogumi tasemel piisab kogusummadest.*

Märkus 21.3 *Paneme tähele, et regressioonihinnangu valem sisaldab suurust* σ_i^2*, mis pole aga teada. Praktikas kasutatakse erijuhte, mis eeldavad spetsiaalseid struktuure dispersiooni* σ_i^2 *jaoks. Tänu nendele probleemi saab vältida.*

.

Ülesanne 21.1 *Tuleta Teoreemi 21.1 väiteid Poissoni valiku jaoks.*

Ülesanne 21.2 *Näidata, et regressioonihinnagu korral kehtib alati:*

$$\sum_s g_{is} \frac{e_i}{\pi_i} = \sum_s \frac{e_i}{\pi_i}.$$

21.2.1 Suhte hinnang

Suhte hinnang on regressioonihinnangu erijuht, mis põhineb ühel abitunnusel ja spetsiaalsel mudelil. Varem vaadeldud suhte tüüpi hinnang on sellega seotud.

Edaspidi eeldame, et tegemist on TTA disainidega. Olgu $\boldsymbol{x}_i = x_i$ (üks abitunnus). Vaatleme järgmist abimudelit:

1. y_i on juhusliku suuruse $Y_i \sim \xi$, $i \in U$ realisatsioon; x_i on mittejuhuslik.
2. Jaotus ξ on selline, et kehtib:

$$E_\xi(Y_i) = \beta x_i; \tag{96}$$

$$V_\xi(Y_i) = \sigma^2 x_i; \tag{97}$$

- regressioonimudel on ilma vabaliikmeta;
- Y kesväärtus on võrdeline tunnusega x;
- Y dispersioon on võrdeline tunnusega x.

Huvitume summa $t = \sum_U y_i$ hinnangust. Leiame kaalud g_{is} valemist (93) arvestades, et x_i on skalaar ja $\sigma_i^2 = \sigma^2 x_i$:

$$\begin{aligned} g_{is} &= 1 + \left(\sum_U x_i - \sum_s \frac{x_i}{\pi_i}\right)\left(\sum_s \frac{x_i^2}{\sigma^2 x_i \pi_i}\right)^{-1} \frac{x_i}{\sigma^2 x_i} = \\ &= 1 + \left(\sum_U x_i - \sum_s \frac{x_i}{\pi_i}\right)\left(\sum_s \frac{x_i}{\pi_i}\right)^{-1} = \\ &= 1 + \frac{\sum_U x_i}{\sum_s w_i x_i} - 1 = \frac{\sum_U x_i}{\sum_s w_i x_i}. \end{aligned}$$

Regressioonihinnang (92) lihtsustub järgmiselt:

$$\hat{t}_r = \sum_U x_i \frac{\sum_s w_i y_i}{\sum_s w_i x_i} = t_x \frac{\hat{t}_y}{\hat{t}_x}, \tag{98}$$

kus $\hat{t}_y$ ja $\hat{t}_x$ on hinnangud vastavalt summadele t_y ja t_x ÜHT järgi.

Kui nüüd võtta $x_i = 1$, siis $\sum_U = N$ - üldkogumi maht. Paneme samuti tähele, et $\sum_s w_i 1 = \hat{N}$ - hinanng üldkogumi mahule. Siis suhte hinnang saab järgmist kuju:

$$\hat{t}_r = N \frac{\sum_s w_i y_i}{\hat{N}}. \tag{99}$$

Teoreem 21.2 *(Suhte hinnang) Suhte hinnang üldkogumi kogusummale* $t = \sum_U y_i$ *on*

$$\hat{t}_r = t_x \frac{\hat{t}_y}{\hat{t}_x}$$

dispersiooniga

$$V(\hat{t}_r) = \sum_U \sum_U \Delta_{ij} \frac{y_i - Rx_i}{\pi_i} \frac{y_j - Rx_j}{\pi_j}$$

nihketa dispersiooni hinnaguga

$$\hat{V}(\hat{t}_r) = \frac{t_x^2}{\hat{t}_x^2} \sum_s \sum_s \frac{\Delta_{ij}}{\pi_{ij}} \frac{y_i - \hat{R}x_i}{\pi_i} \frac{y_j - \hat{R}x_j}{\pi_j},$$

kus $R = \frac{t_y}{t_x}$ *ja* $\hat{R} = \frac{\hat{t}_y}{\hat{t}_x}$.

Ülesanne 21.3 *Tõestada, et lihtsa juhuvaliku korral on Suhte hinnang järgmine:*

$$\hat{t}_r = t_x \frac{\bar{y}}{\bar{x}}$$

dispersiooni- ja dispersiooni nihketa hinnanguga:

$$V(\hat{t}_r) = N^2 \frac{1-f}{n} \frac{\sum_U (y_i - Rx_i)^2}{N-1} \text{ ja } \hat{V}(\hat{t}_r) = \frac{t_x^2}{\bar{x}^2} \frac{1-f}{n} \frac{\sum_s (y_i - \hat{R}x_i)}{n-1}.$$

22 Mittevastamise kompenseerimise meetodeid

Suur kadu valikuuringus mõjub halvasti hinnangute kvaliteedile, põhjustades nihet ja valet dispersiooni (sageli alahinnatud). Mittevastanute keskmine erineb sageli vastanute keskmisest, sest mittevastanud objektid on tihti erilised. Sissetuleku uuringus ei soovi sageli väga rikkad inimesed oma sissetulekut avalikustada ja jäävad uuringust kõrvale.

Probleem: alati ei ole võimalik otsustada, kas vastanute hulk erineb mittevastanute omast.

Eristame edaspidi kahte liiki kadu:

- **tunnuse väärtuse** kadu (*item nonresponse*): sel juhul puudub vaadeldaval objektil mõne tunnuse väärtus (st andmestikus esinevad lüngad). Näiteks jätavad küsitletavad sageli vastamata tundlikele küsimustele (narkootikumide tarbimise, sissetuleku jms kohta)
- **objekti** kadu (*unit nonresponse*): andmestikust puudub terve objekt, puuduvad selle objekti kõikide uuritavate tunnuste väärtused. Näiteks objekti kadu tekib siis, kui inimene keeldub vastamast, inimest pole võimalik kätte saada elukoha muutuse või vale aadressi tõttu.

Järgmises tabelis (× tähendab väärtust ja · puuduvat väärtust) esinev objektidel 2, 3 ja 6 väärtuse kadu ja objektidel 7 ja 8 objekti kadu (st nende kohta on teada vaid registritunnuste väärtuseid).

	Registri tunnused		**Uuritavad tunnused**		
Objekti nr	**1**	**2**	**1**	**2**	**3**
1	×	×	×	×	×
2	×	×	×	×	·
3	×	×	×	·	×
4	×	×	×	×	×
5	×	×	×	×	×
6	×	×	·	×	·
7	×	×	·	·	·
8	×	×	·	·	·

22.1 Tunnuse väärtuse kadu

Selle kaoliigi kompenseerimiseks sobivad nn **omistusmeetodid** (*imputation methods*). Nende meetodite eesmärk on lünkadeta andmestiku saamine, mis on vajalik paljude andmetöötluse programmide ja meetodite kasutamiseks.

Teiseks võimaluseks oleks kõigi selliste objektide eemaldamine andmestikust, millel mõni tunnuseväärtus puudub, kuid selle tagajärjeks on liiga väike valimimaht.

Omistusmeetodite korral asendatakse puuduvad väärtused hinnanguliste väärtustega.

Levinumad omistusmeetodid on järgmised:

- **Üldise keskmise omistus** – omistatakse vastanute hulga keskmine, ei sobi nominaaltunnuste jaoks.
- **Klassi keskmise omistus** – valim jagatakse omistusklassidesse teadaolevate tunnuste väärtuste alusel, omistatakse klassi keskmine.
- **Hot-Deck** ja **Cold-Deck** – keskmisest erinev väärtus. Hot-Deck: sama andmestiku doonori väärtus. Realiseeritakse üldise juhusliku omistuse, klassi juhusliku omistuse, järjestikuse Hot-Decki ja kaugusfunktsioonijärgse omistuse abil. Cold-Deck: muu allikas, näiteks eelmised samalaadsed uuringud.
- **Üldine juhuslik omistus** – omistatakse vastanute seast juhuslikul valitud objekti väärtus.
- **Juhuslik omistus klassis** – moodustatakse nn omistusklassid. Omistatakse samast klassist juhuslikult valitud objekti tunnuse väärtus.

- **Järjestikune Hot-Deck omistus** – objektid järjestatakse ja läbitakse. Omistatakse järjekorras eelneva samasse klassi kuuluva objekti olemasolev väärtus.
- **Kaugusfunktsioonijärgne omistus** – tausttunnuse abil defineeritakse objektidevahelised kaugused. Omistatakse kauguse poolest lähima väärtus.
- **Regressioonijärgne omistus** – vastanud objektide andmeid kasutades moodustatakse regressioonivõrrand, mille abil prognoositakse puuduvad väärtused. Abitunnused peavad olema teada. Arvtunnused-lineaarne regressioon, kvalitatiivne tunnus-logistiline regressioon.
- **Mitmene omistus** – omistatakse mitu (nt m) väärtust. Järjestatakse ja saadakse m andmestikku, analüüsitakse standardsete vahenditega. Saadakse nn kombineeritud hinnang.
- **Objekti asendus** – mittevastanu asemel võetakse uuringusse uus objekt, võimalikult sarnane. Vajalik hea taustinfo ja üldkogum ning valimi põhjalik tundmine.

Omistusmeetodite puudused:

1. Kuna omistatavate väärtuste leidmiseks kasutatakse vastanute andmeid, siis see ei likvideeri nihet hinnangutes (eriti kui mittevastanute hulk erineb väga vastanute hulgast).
2. Hinnangute dispersiooni alahindamine, mille tagajärjeks on liiga kitsad usaldusintervallid. Väiksem dispersioon võib tuleneda:
    - omistusmeetodi eripärast (näiteks keskmise omistus)
    - mittevastanute hulga eripärast (näiteks on rikkamatel inimestel ekstreemsemad väärtused, kuid neid vastanute hulgas pole.)

Populaarsemad omistusmeetodid on kirjeldatud õpikus Traat ja Inno (1997). Näiteid: üldise keskmise omistus, klassi keskmise omistus, regressioonijärgne omistus jne.

22.2 Objekti kadu

Kao kompenseerimiseks objekti tasemel kasutatakse nn kaalumismeetodeid, mis eeldavad tausinfo kasutamise võimalust (registrid, eelmised samalaadsed uuringud jms). Kaalumismeetodi korral leitakse vastanutele uued kaalud, mis parandavad mittevastamisest põhjusetatut nihet hinnangus. Siinuures on levinud kaks lähenemist: kalebreerimismeetodid ja vastamistõenäosuse mudelite meetodid.

22.2.1 Kalibreerimismeetodid

Huvitume nagu tavaliselt kogusumma $t = \sum_U y_i$ hindamisest. Selleks on võetud valim s, kuid kõik valimisse sattunud objektid ei vasta uuringu küsimustele (pole näiteks kättesaadavad).

Tähistame edaspidi vastanute valimi r (*responded*) ja mittevastanute valimi nr (*nonresponded*). Seega, $s = r \cup nr$.

Toetudes ÜHT-le konstrueerime hinnangu kogusummale, kuid meil on võimalik kasutada vaid vastanute hulka:

$$\hat{t}_r = \sum_r w_i y_i.$$

Hinnanang $\hat{t}_r$ on ilmselgelt nihkega hinnanang, mis hindab tegeliku parameetrit alla.

Kalibreerimismeetodi kohaselt otsitakse uue kaalude süsteemi $w_i' = w_i \cdot v_i$, mis laiendaks vastanute hulka kõigepealt valimini s (läbi kaalu v_i) ja seejärel üldkogumini (läbii disainikaalu w_i).

Uue kaalue leidmisel kasutame registrist pärinevat abiinformatsiooni. Praktikas on levinud kolm olukorda:

1. **InfoU**: register võimaldab kasutada abiinformatsiooni **üldkogumi** kohta. Abiinformatsioon võib meile laekuda kahel viisil:
   - **objekti tasemel**, ehk teame iga üldkogumi objekti jaoks ($i = 1, 2, ..., N$) J abitunnuse väärtuseid. Tähistame i. objekti abiinformatsiooni $\boldsymbol{x}_i = (x_{i1}, ..., x_{iJ})^T$.
   - agregeeritud kujul, ehk registrist on kättesaadavad üksnes summad $\boldsymbol{t}_x = \sum_U \boldsymbol{x}_i$. Vastanutelt küsime samu tunnuseid ja seega saame teada ka üksikuid väärtuseid $\boldsymbol{x}_i$ vastanute jaoks.
2. **InfoS**: abiinformatsioon on kättesaadav vaid valimi s kohta. See info võib olla samuti esindatud nii üksikute väärtuste abil kui ka agregeeritud kujul.
3. **InfoUS**: abiinfo on kättesaadav kombineeritud kujul (ühed tunnused valimi ja teised üldkogumi tasemel).

Kasutades abiinformatsiooni (ükskõik, millisel tasemel) otsitakse uued kaalud $w_i' = w_i v_i$ nii, et :

1. nad tagaksid nii väikse nihke hinnangule kui on võimalik;
2. tagaksid hinnangule väikest dispersiooni;

3. rahuldaksid nn **kalibreerimisvõrrandit**: $\sum_r w_i' \boldsymbol{x}_i = \boldsymbol{t}_x$.

Osutub, et neid tingimusi rahuldavad kaalud kujul $w_i' = w_i v_i$, kus (siin kursuses ilma tõestuseta):

$$v_i = 1 + \boldsymbol{\lambda}^T \boldsymbol{x}_i, \tag{100}$$

ja

$$\boldsymbol{\lambda}^T = (\boldsymbol{t}_x - \sum_r w_i \boldsymbol{x}_i)^T (\sum_r w_i \boldsymbol{x}_i \boldsymbol{x}_i^T)^{-1}. \tag{101}$$

Vaatleme selle valemi kahte erijuhtu: abiinformatsioon puudub ja klassifitseeriv abiinfo.

1. Kalibreerimishinnang juhul kui abiinformatsiooni pole

Sel juhul saame alati võtta $\boldsymbol{x}_i \equiv 1 \; \forall i = 1, ..., N$. Leiame vajalikud suurused valemis (101):

$$\boldsymbol{t}_x = \sum_U 1 = N, \;\; \sum_r w_i \boldsymbol{x}_i = \sum_r w_i \text{ ja } \sum_r w_i \boldsymbol{x}_i \boldsymbol{x}_i^T = \sum_r w_i.$$

Seega, valemi (101) ja (100) põhjal:

$$\boldsymbol{\lambda}^T = \lambda = (N - \sum_r w_i) \frac{1}{\sum_r w_i} = \frac{N}{\sum_r w_i} - 1,$$

ja kordaja v_i:

$$v_i = \frac{N}{\sum_r w_i}.$$

Kalibreeritud hinnang on kokkuvõttes

$$\hat{t}_r = \sum_r w_i' y_i = \sum_r w_i \left(\frac{N}{\sum_r w_i} \right) y_i = \frac{N}{\sum_r w_i} \cdot \sum_r w_i y_i.$$

Tähistades $\tilde{y}_r = \frac{1}{\sum_r w_i} \cdot \sum_r w_i y_i$ (kaalutud valimi keskmine), saame esitada kalibreeritud hinnangu lõplikul kujul

$$\hat{t}_r = N \cdot \tilde{y}_r. \tag{102}$$

On lihtne veenduda (ise!), et lihtsa juhuvaliku korral saab hinnang (102) järgmise kuju:

$$\hat{t}_r = N \cdot \bar{y}_r,$$

kus $\bar{y}_r$ on vastanute keskmine valimis.

2. Kalibreerimishinnang klassifitseeriva x-tunnuse korral

Defineerime abitunnust järgmiselt:

$$\boldsymbol{x}_i = (\gamma_{1i}, \gamma_{2i}, ..., \gamma_{Ji})^T,$$

kus

$$\gamma_{ji} = \begin{cases} 1, & \text{kui } i \text{ kuulub gruppi } j \\ 0, & \text{vastasel juhul.} \end{cases}$$

Seega, on vektori $\boldsymbol{x}_i$ elementideks nullid va üks element, mille väärtuseks on 1 (sest iga objekt peab kuuluma mingisugusse gruppi). Suvalise tunnuse baasil on võimalik moodustada sellist klassifitseerivat abitunnust (inglise keeles nimetatakse γ_{ji} *dummy variables*). Arvtunnuse korral jagatakse arvtunnus $\boldsymbol{x}$ intervallideks ning γ_{ji} iseloomustavad kuuluvust konkreetsesse intervalli.

Leiame vajalikud suurused:

$$\boldsymbol{t}_x = \sum_U (\gamma_{1i}, \gamma_{2i}, ..., \gamma_{Ji})^T = (N_1, N_2, ..., N_J)^T,$$

kus $N_j, j = 1, 2, ..., J$ on j. grupi suurus (üldkogumis).

$$\sum_r w_i \boldsymbol{x}_i = \sum_r w_i (\gamma_{1i}, \gamma_{2i}, ..., \gamma_{Ji})^T = (\hat{N}_{1,r}, \hat{N}_{2,r}, ..., \hat{N}_{J,r})^T,$$

kus $\hat{N}_{j,r}$ on hinnang j. grupi suurusele, mis on arvutatud vastanute põhjal.

$$\sum_r w_i \boldsymbol{x}\boldsymbol{x}^T = \sum_r w_i \begin{pmatrix} \gamma_{1i} \\ \gamma_{2i} \\ \vdots \\ \gamma_{Ji} \end{pmatrix} (\gamma_{1i}, \gamma_{2i}, ..., \gamma_{Ji}) = \sum_r w_i \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix},$$

kus element 1 ilmub iga objekti i jaoks peadiagonaalile positsiooni (j, j), sõltuvalt grupist, kuhu ta kuulub. Liites kokku üle vastanute, saame

$$\sum_r w_i \boldsymbol{x}\boldsymbol{x}^T = \begin{pmatrix} \hat{N}_{1,r} & 0 & \cdots & 0 \\ 0 & \hat{N}_{2,r} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{N}_{J,r} \end{pmatrix}$$

Saadud maatriks on diagonaalmaatriks, selle pöördaatriksi saamiseks peame vaid võtma peadiagonaali elementidest pöördväärtuse:

$$\left(\sum_r w_i \boldsymbol{x}\boldsymbol{x}^T \right)^{-1} = \begin{pmatrix} \frac{1}{\hat{N}_{1,r}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\hat{N}_{2,r}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\hat{N}_{J,r}} \end{pmatrix}$$

Nüüd saame leida $\boldsymbol{\lambda}^T$:

$$\boldsymbol{\lambda}^T = (\boldsymbol{t}_x - \sum_r w_i \boldsymbol{x}_i)^T \left(\sum_r w_i \boldsymbol{x}\boldsymbol{x}^T \right)^{-1} = \left[\begin{pmatrix} N_1 \\ N_2 \\ \vdots \\ N_J \end{pmatrix} - \begin{pmatrix} \hat{N}_{1,r} \\ \hat{N}_{2,r} \\ \vdots \\ \hat{N}_{J,r} \end{pmatrix} \right]^T \begin{pmatrix} \frac{1}{\hat{N}_{1,r}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\hat{N}_{2,r}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\hat{N}_{J,r}} \end{pmatrix}$$

$$= \left(\frac{N_1}{\hat{N}_{1,r}}, \frac{N_2}{\hat{N}_{2,r}}, ..., \frac{N_J}{\hat{N}_{J,r}} \right) - (1, 1, ..., 1).$$

Tegur v_i saab väga lihtsa kuju:

$$v_i = 1 + \boldsymbol{\lambda}^T \boldsymbol{x}_i = 1 + \left(\frac{N_j}{\hat{N}_{j,r}} - 1 \right) = \frac{N_j}{\hat{N}_{j,r}}.$$

Hinnangu $\hat{t}_r$ tuletamisel jagame vastanuid samuti gruppidesse (tähistame r_j) γ_j järgi, seega saame J gruppi vastanute jaoks (mõni grupp võib praktikas osutuda tühjaks):

$$\hat{t}_r = \sum_r w_i v_i y_i = \sum_{j=1}^{J} \sum_{i \in r_j} \frac{N_j}{\hat{N}_{j,r}} w_i y_i = \sum_{j=1}^{J} \frac{N_j}{\hat{N}_{j,r}} \sum_{i \in r_j} w_i y_i.$$

Thistades $\tilde{y}_{j,r} = \frac{1}{\hat{N}_{j,r}} \sum_{i \in r_j} w_i y_i$ - kaalutud keskmine grupis j, saame lõplikult

$$\hat{t}_r = \sum_{j=1}^{J} N_j \tilde{y}_{j,r}. \tag{103}$$

Näeme, et hinnangu (103) leidmiseks peame teadma vaid gruppide suuruseid üldkogumis. Objektide kuuluvust gruppidesse peame teadma vaid vastanutel (seda on aga võimalik teada saada uuringu läbiviimisel).

Ülesanne 22.1 *Tuletada hinnang (103) lihtsa juhuvaliku jaoks.*

Ülesanne 22.2 *Lihtsustada kalibreemimishinnang $\hat{t}_r$ ühe pideva abitunnuse korral. Vihje: alustada (100) ja (101) lihtsustamisest.*

Ülesanne 22.3 *Tabelis on toodud valimiandmed arstide koduvisiidide kohta jaanuari kuus asulas A. Kahjuks, mõned asutused ei teatanud oma andmeid, nende kohta saame registrist vaid asutuse poolt pakutava teenuse tüüpi. On teada, et asulas A on 15 perearstikeskust ja 9 õiendusabikeskust. Valikuviis on lihtne juhuvalik mahuga 10 asutust. Hinnata koduviisitide koguarvu linnas A kasutades kalibreerimishinnangut klassifitseeritava abitunnusega.*

Asutuse Nr	*Teenuse tüüp*	*Visiite*
1	*Perearstikeskus*	*138*
2	*Õendusabi*	*64*
3	*Õendusabi*	.
4	*Perearstikeskus*	.
5	*Perearstikeskus*	*162*
6	*Perearstikeskus*	*133*
7	*Õendusabi*	*84*
8	*Perearstikeskus*	.
9	*Perearstikeskus*	*186*
10	*Õendusabi*	*76*

22.2.2 Vastamistõenäosuse mudeli meetodid

Vaatleme siin ainult tagasipanekuta valikudisaine. Sel juhul disainikaal $w_i = \frac{1}{\pi_i}$.

Vastamistõenäsosuse mudel põhineb eeldusel, et vastamine allub mingisugusele seadusele, ehk iga objekti $i \in s$ jaoks leidub kindel vastamistõenäosus $\pi_{i|s}$ (loe: i. objekti vastamistõenäosus tingimusel, et ta on valimis).

Sellisel juhul on objekti i sattumine vastanute hulka r määratud tõenäosusega $\pi_i \pi_{i|s}$.

Saab näidata, et hinnang

$$\hat{t}_r = \sum_r \frac{y_i}{\pi_i \cdot \pi_{i|s}} \tag{104}$$

on nihketa hinnang kogusummale t.

Valemi (104) suureks puuduseks on see, et suurused $\pi_{i|s}$ on tundmatud, ehk keegi ei tea, mis tõenäosusega objekt i vastab uuringus. On välja pakutud mitmeid meetodeid, mis modeleerivad seda vastamistõenäosust. Üks levinumatest on homogeensete vastamisgruppide meetod.

Homogeensete vastamisgruppide meetod

Selle meetodi korral eeldatakse, et konkreetse grupi objektid vastavad sarnaselt (näiteks, abielus inimesed vastavad suurema tõenäosusega kui vallalised; pensionärid vastavad samuti suurema tõenäosusega kui noored jne).

Olgu moodustatud H gruppi, mille kohta on eeldatud, et inimesed nendes gruppides vastavad võrdse tõenäosusega θ_h:

$$\pi_{i|s} = P(i \in r|s) = \theta_h, \ h = 1, ..., H.$$

Vastamistõenäosused θ_h on tundmatud, kuid neid on võimalik üsna lihtsalt hinnata. Olgu m_h - vastanute hulk grupis h. Iga inimene kas vastab või mitte (binaarne juhuslik suurus); inimeste vastamine toimub üksteisest sõltumatult. Seega, kui vaadelda m_h juhusliku suurusena, siis sobib selle kirjeldamiseks binoomjaotus,

$$m_h \sim B(n_h, \theta_h).$$

Siin on n_h - valimimaht grupis h. Suurima tõepära hinnang vastamistõenäosusele on $\hat{\theta}_h = \frac{m_h}{n_h}$.

Kokkuvõttes, saame valemist (104) hinnangu kogusummale homogeensete vastamisgruppide meetodil:

$$\hat{t}_r = \sum_{h=1}^{H} \sum_{i \in r_h} \frac{y_i}{\pi_i} \frac{1}{m_h/n_h} = \sum_{h=1}^{H} \frac{n_h}{m_h} \sum_{r_h} \frac{y_i}{\pi_i}. \tag{105}$$

Ülesanne 22.4 *Leida ülesandes 22.3 hinnang homogeensete vastamisgruppide meetodil.*

Kirjandus

Deville, J.-C., Tille, Y. (2004). Efficient balanced sampling: the cube method. *Biometrika*, **91**, 893-912.

Grafström, A., Lundström, N. L. P., Schelin, L. (2012). Spatially balanced sampling through the pivotal method. *Biometrics*, **68**, 514-520.

Grafström, A., Schelin, L. (2014). How to select representative samples. *Scandinavian Journal of Statistics.* Vol. **41**: 277-290.

Traat, I., Inno, J. (1997). *Tõenäosuslik valikuuring.* Tartu Ülikooli kirjastus.

Särndal, C.-E., Swensson, B., Wretman, J. (1991) *Model Assisted Survey Sampling.* Springer-Verlag.