10 praktikum – Kihtvalik teostamine SAS-is. Töö registritega

Tänases praktikumis:

- uurime registreid HOUSEREG ja POPREG ning nende võimalusi kihtvaliku moodustamisel;
- kirjutame SAS koodi kihtvaliku teostamiseks ja hinnangute leidmiseks;
- võrdleme LJ KIHTvalikut tavalise LJ valikuga ja seda kahe valimipaigutuse korral.

Praktikumis kasutame projektitöö käigus loodud andmestikku + registrid. Uuritavaks tunnuseks on juba tuttav leibkonna sissetulek *totinch*. Soovime hinnata valimist leibkonna keskmist sissetulekut, $\hat{\bar{Y}}$.

10.1 Lisainformatsioon valiku teostamisel (registrid)

Tänapäeval on olemas mitmesugused registrid (rahvastiku-, ettevõttete, maksumaksjate, kariloomade, ehituse register jne), milles leidub rohkesti lisainformatsiooni. Seda lisainformatsiooni saab kasutada nii valiku kui ka hindamissammul.

StatVillage jaoks on olemas **Eluasemeregister** (**HOUSEREG**), milles on informatsiooni eluaseme vanuse ja rentimise kohta.

BUILTH on kodeeritud järgmiselt: **1** – kuni 1920; **2** – 1921-1945; **3** – 1946-1960; **4** – 1961-1970;...; **8** – 1991 ja uuemad.

Tutvume siinkohal veel ühe registriga (**rahvastik**) - **POPREG**:

- SUGU (1– mees; 2 – naine; 3 – laps)
- VANUSKL (0 – 0..4 aastat; 5 – 5..17 aastat jne vastab *Codebook*'ile *StatVillage*'s)
- NIMI (väljamõeldud tunnus).

Paneme tähele, et andmed on esitatud inimeste tasemel, meil on aga leibkonnad. SQL-i abil saab aga registis olevat gruppeeringut muuta. Loome uue registri, mis sisaldab leibkondade suuruseid terves *StatVillage*s:

```
proc sql;
create table leibkonna_suurus as
select block, unit, count(vanuskl) as hhsize from popreg
group by block, unit;
quit;
```

10.2 Kihtvaliku teostamine

Eluasemeregistrist pärit informatsiooniga saaksime üldkogumit kihistada

- geograafiliselt (kasutades aadressitunnust BLOCK - lõuna ja põhja plokid);
- maja vanuse järgi (BUILTH);

- tunnuse omanik/rentnik järgi (TENURH);
- kahe, kolme tunnuse ristklassifitseerimisega.

Rahvastikuregistrit saab kasutada järgmiseks kihistamiseks:

- leibkonna suuruse järgi;
- lastega vs lasteta pered;
- väikelastega pered jne.

Kihistamine aitab tõsta hinnangute täpsust, kui uuritavad tunnused on võimalikult homogeensed kihtides. Ka uuritavat osakogumit püütakse võimaluse korral kasutada kihina. **Mille järgi oleks otstarbekas moodustada kihte kui uuritavaks tunnuseks on leibkonna sissetulek?**

Iseseisvalt (võib teha koos oma projektirühmaga):

- Mõtle välja 3-4 kihti tunnuse *totinch* uurimiseks. Kihistava(-te) tunns(t)e väärtused peavad olema kättesaadavad registrist ning tunnuse *totinch* väärtused peavad olema kihiti homogeensed. Oma projektitöö andmestiku saab kasutada pilootuuringuna kontrollimaks homogeensust.

Teame, et LJ kihtvaliku korral annab minimaalse dispersiooni hinnangule **Neymani paigutus**, mille korral

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

mis tähendab seda, et selles kihis, kus tunnuse varieeruvus (läbi S_{yU_h}) on suurem, võetakse valimisse rohkem objekte. Samuti, ka suurest kihist (läbi N_h) võetakse samuti rohkem objekte.

Märkus. Kui uuringus on palju uuritavaid tunnuseid, siis Neymani paigutus võib osutuda väga heaks ainult ühe konkreetse uuritava tunnuse jaoks, teiste jaoks aga väga halvaks. Neymani paigutust eelistatakse siis, kui kõik uuritavad tunnused on omavahel positiivselt korreleeritud.

Praktikas on väga levinud ka nn **võrdeline paigutus** (pole üldjuhul optimaalne):

$$n_h = n \cdot \frac{N_h}{N}.$$

Kui kihtides on dispersioonid ($S^2_{yU_h}$) võrdsed, siis langevad Neymani ja võrdeline paigutused kokku.

Iseseisvalt (võib teha koos oma projektirühmaga):

- Arvutada Neymani paigutus tunnusele *totinch*, lähtudes kogu valimimahust $n = 100$. Dispersioone S_{yU_h} võib hinnata oma projektitöö andmestiku põhjal.
- Täita järmine tabel.

Valik	N, N_h	n, n_h
LJ kihtvalik võrdelise paigutusega		
LJ kihtvalik Neymani paigutusega		

- Valida üks paigutusest ja moodustada registrifailist Housereg_str kihtvalim leitud mahtudega $n_1,\ n_2, ...$ Abiks on järgmine kood:

```
proc surveyselect data=Housereg_str
n=(n1  n2)
method=srs
out=LJKvalim;
strata str;
run;
```

- Hankida *Statvillage* andmeid kodulehelt ja salvesta need faili *LJKandmed*.

10.3 Hindamine kihtvaliku korral

Meeldetuletuseks:
Kihtvaliku korral on nihketa hinnang ÜK summale:

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

kus hinnang $\hat{t}_h$ on nihketa kihis U_h.
Hinnangu $\hat{t}$ dispersioon ja dispersioonihinnang on

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

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

Sellele toetudes, leiame hinnangud kogu sissetulekule (*totinch*) kihtides eraldi. Selleks tuleks valim lahku lüüa alamvalimiteks. Abiks on järgmine kood

```
data LJKandmed1;
set LJKandmed;
if str=1 then output;
run;
data LJKandmed2;
set LJKandmed;
if str=2 then output;
run;
```

Iseseisvalt:

- Leida kõikide alamvalimite põhjal hinnangud kogusissetulekule koos hinnangute dispersiooniga (*proc surveymeans*). Kirjutada välja:

- $\hat{t}_1 =$ $\hat{V}(\hat{t}_1) =$
- $\hat{t}_2 =$ $\hat{V}(\hat{t}_2) =$
- $\hat{t}_3 =$ $\hat{V}(\hat{t}_3) =$
- $\hat{t}_4 =$ $\hat{V}(\hat{t}_4) =$

- Kombineerides saadud hinnanguid leida hinnang leibkonna keskmisele sissetulekule koos standard- ja suhtelise veaga. Kirjutada need siia:
- $\hat{\bar{Y}} = \frac{\hat{t}}{N} = \frac{\hat{t}_1+\hat{t}_2+\hat{t}_3+\hat{t}_4}{N} =$ $stand.v.(\hat{\bar{Y}}) = \frac{1}{N}\sqrt{\hat{V}(\hat{t})} =$
- $suht.v.(\hat{\bar{Y}}) = \sqrt{\hat{V}(\hat{\bar{Y}})}/\hat{\bar{Y}} =$
- Leida oma projektitöös kasutatud valimi põhjal samu näitajaid (ilma kihistamata) ja kirjutada välja:
- $\hat{\bar{Y}}_{LJ} =$ $stand.v.(\hat{\bar{Y}}_{LJ}) =$ $suht.v.(\hat{\bar{Y}}_{LJ}) =$
- Kas lihtne juhuslik KIHTvalik andis parema tulemuse kui tavaline lihtne juhuslik valik?

Lihtsa juhusliku kihtvaliku korral saab hindamiseks kasutada tervet valimit korraga järgmiselt. Kõigepelat moodustame kihimahtudest N_d koosneva faili ja seejärel kasutame seda faili *procsurveymeans*-protseduuris:

```
data Kokku;
input str _total_; /* total on SASi systeemne tunnuse nimi*/
datalines;
1 N1
2 N2
;
run;

proc surveymeans data=LJKandmed
total=Kokku
mean;
stratum str;
var totinch;
weight SamplingWeight;
run;
```

Pane tähele, et viimast programmi saab kasutada ainult LIHTSA juhusliku kihtvaliku korral. Kui aga on ühes kihis rakendatud mingid muud valikuviisi, kasutab SAS dispersioonide leidmiseks ikkagi lihtsa juhusliku valiku dispersiooni.