2 praktikum – Valimi võtmine protseduuri *Surveyselect* abil

Tänase praktikumi eesmärk on tutvuda SASi protseduuriga SURVEYSELECT, mis võimaldab teostada mitmeid tõenäosuslike valikuid.

2.1 Protseduuri süntaks

```
PROC SURVEYSELECT
     DATA=SAS-andmestik
     OUT=
     METHOD=
     SEED=
     SAMPSIZE=
     SAMPRATE=;
     STRATA kihitunnused;
     SIZE suurustunnus ;
     ID väljundtunnused ;
RUN;
```

Tähtsad valikud on toodud allpool.

Mida määrab	**Valik**	**Märkus**
Sisendandmestik = freim	DATA=<andmestik>	vaikimisi viimati loodud andmestik
Väljundandmestik	OUT=<andmestik>	vaikimisi nimeks DATA
Väljatrüki keelamine	NOPRINT	
Valikumeetod	METHOD = PPS SRS SYS URS PPS_WR PPS_BREWER nüüd ka BERNOULLI, POISSON	vaikimisi SRS (=LJV, kui puudub SIZE) või PPS (= suurusega võrdeline valik, kui SIZE olemas);
Valimimaht n	SAMPSIZE= n	KV => igas kihis konstantne
	SAMPSIZE=<$n_1, n_2, , n_H$> <andmestik>	KV korral; Andmestikus tunnus _NSIZE_
Valikusuhe $f = n/N$	SAMPRATE = f	SRS, URS, SYS KV => igas kihis konstantne
	SAMPRATE= <$f_1, f_2, ..., f_H$> =<andmestik>	KV korral; Andmestikus tunnus _RATE_
Juhuslik seeme	SEED=nr	vaikimisi kellaaeg
Väljundandmete lisamine	JTPROBS - ühistõenäosused	PPS, PPS WR
	OUTSIZE - info valiku kohta $(n, f, ...)$	
	STATS - valikutõenäosused, kaalud	SRS, URS, SYS; teistel vaikimisi

Kindlasti peab olema määratud üks kahest valikust: SAMPSIZE / SAMPRATE

2.2 Kasutamise näiteid

Viimase praktikumi iseseisvas ülesandes sai moodustatud valim 2-astmelise disaini abil lõpliku mahuga 50. Samuti said hangitud andmed ja salvestatud need SAS andmestikuna.

Järgnevad näited põhinevad nendel andmetel.

- Loo andmefailist koopia ja nimeta see ümber **Yldkogum**. Edaspidi käsitleme neid andmeid üldkogumina, millest hakkame võtma valimeid protseduuri SURVEYSELECT abil.
- Loo otsetee antud failile nimega **Valik1**.

2.2.1 Bernoulli valik

Kirjeldada Bernoulli valikut teoreetiliselt kasutades valikuindikaatorit I_i.

```
/*Bernoulli valik*/
PROC SURVEYSELECT data=valik1.Yldkogum
       method=BERNOULLI
       samprate=0.2
       out=ValimBERN; /*Valim suunatakse kasutasse WORK*/
run;
```

- Uurida tabelit aknas SASViewer.
- Milline valimimaht realiseerus?
- Mis on iga objekti kaaluks ja kaasamistõenäosuseks?
- Mida võiks tähendada Adjusted Sampling Weight?

2.2.2 Lihtne juhuvalik, *SRS - Simple Random Sampling*

```
 /*LJV*/
PROC SURVEYSELECT data=valik1.Yldkogum
       method=srs
       n=10
       out=ValimLJV; /*Valim suunatakse kasutasse WORK*/
run;
```

Kui aga soovime andmeid aknasse *SASViewer*, siis saab kirjutada juurde:

```
PROC PRINT data=valimLJV;
       title1 'Valim 10-st leibkonnas, LJV';
```

2.2.3 Lihtne juhuslik kihtvalik, *Stratified SRS*

Kihtvaliku korral peab ÜK sisaldama tunnust, mille järgi saab iga objekti kohta ÜK-st identifitseerida, mis kihti see objekt kuulub. Selle tunnuse (või isegi mitme tunnuse) järgi tuleb freimi (ÜK) kõigepealt sorteerida.

Olgu sellisteks tunnuseks leibkonna perepea sugu ja emakeel (tunnused *hmsex* ja *hmmtn*, tunnuste väärtuste seletused on *Codebook*'es).

Vaatame esmalt kihtide mahtusid ÜK-is ja seejärel otsustame valimimahtude kohta.

```
PROC SORT data=valik1.Yldkogum;
      by hmsex hmmtn;
run;

PROC FREQ data=valik1.Yldkogum;
      tables hmsex*hmmtn;
run;
```

Kui ÜK kihtide mahud võimaldavad, siis võtame igast kihist võrdse mahuga $n = 2$ valimi järgmise protseduuri abil.

```
 /*KLJV*/
PROC SURVEYSELECT data=valik1.Yldkogum
       method=srs
       n=2
       out=ValimKiht1; /*Valim suunatakse kasutasse WORK*/
STRATA hmsex hmmtn;
run;
```

- Uuri kindlasti saadud andmestikku!
- Leia valimitabelis kaalude veerg ja uuri, miks kaalud on erinevad.

Kihtide mahud ÜK-s tugevasti varieerusid, seetõttu pole ilmselt õiglane võtta igast kihist samapalju objekte valimisse.

- Uuri kihtide mahtusid ÜK-s veel kord.
- Olgu lõplik valimimaht 20. Arvuta proportsionaalselt valimimahtusid kihtides nii, et sumaarselt tuleks 20. (Suurem kiht => suurem valimimaht selles kihis).

Sellist valikut saab teosatada analoogilise programmi abil:

```
/*KLJV ebavõrdsete valimimahtudega kihtides*/
PROC SURVEYSELECT data=valik1.Yldkogum
       method=srs
       n=(10 6 3 1)
       out=ValimKiht2; /*Valim suunatakse kasutasse WORK*/
STRATA hmsex hmmtn;
run;
```

Viimast varianti oleks mugavam kasutada, kui annaksime ette valimimahud protsentuaalselt. Võtame igast kihist $20/50 * 100\% = 40\%$ väärtustest. Kasutame seekord süstemaatilist valikut kihtides.

```
/*Kihtvalik ebavõrdsete valimimahtudega kihtides*/
PROC SURVEYSELECT data=valik1.Yldkogum
       method=sys
       rate=0.4
       out=ValimKiht3; /*Valim suunatakse kasutasse WORK*/
STRATA hmsex hmmtn;
run;
```

2.2.4 Tõenäosustega võrdeline valik, *PPS- Probability Proportional to Size sampling*

Selle valikuviisi korral on kõik kaasamistõenäosused võrdelised mingisuguse ÜK-s teadaoleva tunnuse x väärtustega, $\pi_i = \frac{E(n)x_i}{t_x}$. Valik võib olla nii TTA kui ka TGA, juhusliku või fikseeritud mahuga. SAS pakub mitmeid meetodeid pps-valiku teostamiseks (vt. *Help*'is). Kuid kõik need meetodid nõuavad x tunnust, mis on teada terves ÜK-s.

Oletame, et leibkonna suurust teame iga $i \in U$, st $x = hhsize$.

- Uuri, kas järgmine programm teostab fikseeritud või juhusliku mahuga, ning kas TTA või TGA valiku.
- Vaata, mis tunnused lisanduvad valimisse. Kuidas nad on arvutatud? Ka *Viewer*'it ei tohi unustada!

```
 PROC SORT data=valik1.Yldkogum;
      by hhsize;
/*Tõenäosustega võrdeline valik*/
PROC SURVEYSELECT data=valik1.Yldkogum
       method=pps
       n=15
       out=ValimPPS1; /*Valim suunatakse kasutasse WORK*/
 SIZE hhsize;
run;
```

Täienda eelmine programm argumendiga `JTPROBS`. Mis muutub valimis?

Järgmine protseduur võtab igast kihist 2 väärtust pps-valiku abil *hhsize* põhjal.

- Uuri, mida tähendab *seed*.
- Millised on siin kihid?
- Miks pole valikuks valimimahtu?

```
 PROC SORT data=valik1.Yldkogum;
      by tenurh;
/*Tõenäosustega võrdeline valik, n_h=2*/
PROC SURVEYSELECT data=valik1.Yldkogum
       method=pps_brewer
       seed=48702
       out=ValimPPS2; /*Valim suunatakse kasutasse WORK*/
```

```
 SIZE hhsize;
 STRATA tenurh;
run;
```

2.3 Iseseisev töö

1. Kirjuta protseduur, mis võtab ÜK-st valimi mahuga 25 *URS* meetodi abil. Uuri *Help*'i abil, missugusele loengust teadaolevale meetodile see vastab. Abiks on ka saadud valimi viimane veerg.

2. Moodusta kaks kihti: pered, kus elab alla 5-aastaseid lapsi (kiht 1) ja teised pered (kiht 2). Kirjuta protseduur, mis võtab esimesest kihist valimisse 3 peret ja teisest kihist 4 peret süstemaatilise valiku abil.

3. StatVillage korral on teada, et põhja pool elavad rikkamad pered kui lõuna pool. Jaga sissetuleku alusel ÜK kolmeks kihiks ja võta igast kihist 20% väärtustest valimisse PPS valiku abil, mis põhineb tunnusel "Magamistubade arv".