8 praktikum – Valimimahu määramine, sobiva disaini leidmine

Tänases praktikumis:

- vaatame, kuidas määrata valimimahtu lähtudes konkreetsest disainist ja etteantud suhtelise vea piirist;
- leiame valimimahud kolme valikudisaini korral (LJV TTA, SYS, LJV TGA) ning võrdleme omavavahel;
- viime läbi simulatsiooni, et võrrelda fikseeritud n jaoks hinnangute suhtelised vead eri disainide korral;
- tutvume kahe registriga *StatVillage*s, mida saab kasutada oma projekti 2. osas.

Praktikumis töötame andmestikuga **Yldkogum**, mis sisaldab *Statvillage* andmeid 50 leibkonna kohta. Uuritavaks tunnuseks on juba tuttav leibkonna sissetulek *totinch*. Soovime hinnata valimist leibkonna keskmist sissetulekut, $\hat{\bar{Y}}$.

8.1 LJV TTA, teooria

Olgu tellija poolt antud piirang, et tunnuse *totinch* keskmise hinnangu suhteline viga ei tohi ületada 0,1.
Leiame valimimahu n nii, et hinnang $\hat{\bar{Y}}$ rahuldaks järgmist tingimust:

$$suht.v.(\hat{\bar{Y}}) \leq 0.1.$$

Tuletame meelde, et lihtsa juhusliku valiku korral:

$$\hat{\bar{Y}} = \bar{y}, \tag{1}$$

$$\hat{V}(\hat{\bar{Y}}) = (1-f)\frac{s_y^2}{n}. \tag{2}$$

- Avaldada valimimaht n funktsioonina N-st ja tunnuse variatsioonikordajast $cv(y) = \frac{s_y}{\bar{y}}$.
- Oletame, et eelmise aasta uuringus tuli $cv(y) = 0,6$. Kui suurt valimit peame võtma?
- Fikseeritud ÜK mahu N korral, millest sõltub valimimaht?
- Võta valim mahuga n, mis on leitud viimasest punktist ja arvuta $\hat{\bar{Y}}$ koos suhtelise veaga. Vaata, et kasutaksid arvutamisel kaale!
- Kas leitud hinnang saavutas etteantud täpsuse?

Teatavasti saame leida hinnangu keskväärtusele kasutades alternatiivset lähenemist, $\hat{\bar{Y}}_{alt} = \frac{\hat{t}_y}{\hat{N}}$.

- Leida hinnang $\bar{Y}$-le alternatiivselt. Selleks tekita valimisse uus tunnus X, kus $x_i = 1, i \in s$.
- Arvuta hinnang keskmisele alternatiivse valemi järgi, kasutades SASis RATIO.
- Hinnang ja standardviga tulid samad. Miks?

8.1.1 Simulatsioon, LJV TTA

Ülalpool sai võetud **üks** valim ja leitud **üks** hinnang - täpselt nii nagu reaalses elus toimubki. Hinnangu suhteline viga on leitud valemi järgi, mille tuletasime **analüütiliselt** teoorias. Analüütilise tulemuse kinnitamiseks on hea kasutada simulatsiooni, mis kinnitab/kummutab teoreetilist tulemust.

Plaan

- Võtame 1000 LJV TTA valimit mahuga n antud Yldkogumist.
- Arvutame 1000 hinnangut huvipakkuvale parameetrile, $\hat{\bar{Y}}$.
- Leiame saadud hinnangute keskmise (nimetame Monte-Carlo keskmiseks) ja hinnangute Monte-Carlo dispersiooni, mida kasutame MC suhtelise vea saamisks,

$$E_{MC}(\hat{\bar{Y}}) = \frac{1}{1000}\sum_{i=1}^{1000}\hat{\bar{Y}}_i \quad (3)$$

$$V_{MC}(\hat{\bar{Y}}) = \frac{1}{999}\sum_{i=1}^{1000}(\hat{\bar{Y}}_i - E_{MC}(\hat{\bar{Y}}))^2 \quad (4)$$

Programmid

Korduvvalimite võtmine (igaühel oma valimimaht!):

```
proc surveyselect
data=Yldkogum
method=srs /*teostame lihtsa juhuvalikut*/
rep=1000 /*võtame 1000 valimit*/
n=21 /*mahuga 21 iga valim*/
out=ValimSRSKorduv /*saadud valimid asuvad work kaustas*/;
run;
```

Lisame kaalutunnuse (kaalu väärtust saab *Output* lehelt, igaühel oma!):

```
data ValimSRSKorduv;
set ValimSRSKorduv;
kaal=2.380952;
run;
```

Arvutame 1000 hinnangut:

```
proc sql;
create table hinnangudSRS as /*loob uue tabeli nimega hinnangudSRS*/
select sum(kaal*totinch)/50 as keskmised /*leitakse keskmine kaalutud y-st*/
from valimSRSkorduv /*tabeli nimi, millest on võetud tunnused kaal ja y*/
group by replicate;/*andmeid ühmitatakse korduste järgi ja igas rühmas leitakse
oma keskmine*/
quit;
```

Monte-Carlo keskmist ja dispersiooni saab protseduuri means abil, millest saab juba arvutada Monte-Carlo suhtelise vea. Kas tuli etteantud piirides?

- Sisesta saadud Monte-Carlo standardvea väärtus Moodlesse 8. nädala plokki.

8.2 Süstemaatiline valik

Loengust teame, et süstemaatilise valiku korral dispersiooni hinnangut pole võimalik leida ja selle asemel kastatakse LJV TTA hinnangut. Seega, etteantud täpsuse korral saame sama valimimahtu, kui LJV TTA jaoks.

Siiski, saame hinnangu täpsust parandada (küll mitte hinnata) uuritava tunnuse väärtuste järjestamisega.

- Kas antud $N = 50$ on võimalik süstemaatilise valikuga saavutada valimimahtu $n = ..$? Kui ei, siis millist on võimalik?
- Protseduuri surveyselect meetod SYS kasutab valimi võtmiseks reaalset intervalli N/n, mitte täisarvulist nagu oleme seda õppinud loengul. Seepärast SYS abil on võimalik saavutada suvalist valimimahtu.
- Hinnangu täpsuse parandamiseks järjestada Yldkogum tunnuse *hhsize* järgi. Miks see hea on?
- Võtta valim SYS meetodiga, leida $\hat{\bar{Y}}$ koos suhteliste veaga.
- Hinnangu standardviga arvutatakse SYS valiku korral hoopis LJV TTA valemi järgi! Seega **tegelikku varieeruvust** arvuliselt teada **ei saa**.
- Teosta simulatsioon nii nagu see oli tehtud LJV TTA korral ja selgita välja, millega võrdub MC keskmine, selle dispersioon ja suhteline viga.
- Kas uuritava tunnuse järjestamine *hhsize* järgi viis väiksema suhtelise veani? Kui ei, siis millist tunnust saaks veel võtta?
- Sisesta saadud Monte-Carlo standardvea väärtus Moodlesse 8. nädala plokki.

8.3 LJV TGA

LJV TGA korral kehtivad järgmised tulemused:

$$\hat{\bar{Y}} = \bar{y}, \tag{5}$$

$$\hat{V}(\hat{\bar{Y}}) = \frac{s_y^2}{n}. \tag{6}$$

- Arvuta selle disaini korral valimimaht n, kui täpsus ei tohi ületada $0,1$.
- Kas valimimaht n tuli suurem või väiksem võrreldes LJV TTA?
- Vii läbi simulatsioon valimimahuga, mis sai leitud LJV TTA jaoks, et võrrelda LJV TTA ja TGA omavahel.

- Pane tähele, et kuna valimisse tekkivad kordused, siis ka kaalu tunnust peab korrutama korduste arvuga, ehk `Kaal=2.380952*numberhits;`
- Sisesta saadud Monte-Carlo standardvea väärtus Moodlesse 8. nädala plokki.