5 praktikum – Protseduur *SURVEYMEANS* – VU andmete statistiline analüüs

5.1 Protseduuri *SURVEYMEANS* süntaks

See protseduur on spetsiaalselt mõeldud VU andmetele. See arvutab hinnanguid ÜK ja selle osakogumite **keskmisele**, **kogusummadele ja osakaaludele** arvestades valikudisaini. Antakse ka hinnangute täpsushinnangud.

Protseduuri *SURVEYMEANS* abil saab leida: statistikuid:

- ÜK kogusumma hinnangut $\hat{t}_y$ ja selle standardhälvet koos vastava t-testiga,
- ÜK keskmise hinnangut $\hat{\bar{Y}}$ ja selle standardhälvet koos vastava t-testiga,
- osakaalu hinnangut $\hat{P}$ klassifitseeritud tunnuse jaoks koos vastava t-testiga,
- $(1-\alpha)\%$ usalduspiire ÜK kogusumma ja keskmise ning osakaalu hinnangutele,
- andmete üldinfot.

SÜNTAKS:

```
PROC SURVEYMEANS < valikud > < statistikud >;
BY grupitunnused;
CLASS klassifitseeritavad tunnused;
CLASTER klastritunnused;
DOMAIN osakogumitunnused < tunnus * tunnus ...>;
RATIO <'märgend'> tunnus/tunnus;
STRATA kihitunnused < / list >;
VAR uuritavad tunnused;
WEIGHT kaalutunnus;
```

Valikud:

Mida määrab	**Valik**	**Märkus**
sisendandmestik=valim	DATA=<andmestik>	vaikimisi viimati loodud andmestik
olulisusnivoo	ALPHA=α	vaikimisi 0.05
puuduvate väärtuste käsitlemine	MISSING	vaadatakse eraldi väärtusena
valikusuhe f	RATE=f RATE=<andmestik>	mitmeastm. valikul valikusuhe 1. astmel KV korral; andmestikus tunnus _RATE_
ÜK maht N	TOTAL=N TOTAL=<andmestik>	KV korral igas kihis konstantne KV korral; andmestikus tunnus _TOTAL_

Arvutatavad statistikud:

Statistik	**Valik**	**Märkus**
kõik allpool loetletud	ALL	
$100(1-\alpha)\%$ usalduspiirid keskmisele	CLM	
$100(1-\alpha)\%$ usalduspiirid kogusummale	CLSUM	
variatsioonikordaja	CV	
vabadusastmete arv	DF	T-testi jaoks
max. väärtus	MAX	
keskmine	MEAN	vaikimise; kvalit. tunnuse korral osakaal
min. väärtus	MIN	
klastrite arv	NCLUSTERS	
puuduvate vaatluste arv	NMISS	
olemasolevate vaatluste arv	NOBS	
ulatus	RANGE	max-min
kogusumma standardhälve	STD	SUM tellimisel vaikimisi
keskmise standardviga	STDERR	MEAN tellimisel vaikimisi
kaalutud summa $\sum w_i y_i$	SUM	kogusumma hinnang, kui kasut. valikukaale
kaalude summa $\sum w_i$	SUMWGT	
t-väärtus	T	H_0 : ÜK keskmine= 0
keskmise dispersioon	VAR	
kogusumma dispersioon	VARSUM	

5.2 Protseduuri *SURVEYMEANS* näiteid

Näide 1. **LJV TTA**. Kasutame mahuga $N = 50$ andmestiku **Yldkogum**.

1. Võtta sellest `proc surveselect` abil LJ valim mahuga 20 ja panna nimeks **LJ-valim**.

2. Teame, et LJV annab isekaaluva valimi ning seetõttu ei lisa PROC SURVEYSELECT valimisse kaale. Kuid SURVEYMEANS neid kasutab! Seetõttu peame lisama oma valimisse veergu **KAAL**. Mis on selle väärtused?

3. Leiame $\hat{t}$ ja sellega seotud näitajaid antud valimi korral. Uuritavaks tunnuseks olgu **leibkonna suurus** *hhsize*:

```
proc surveymeans
data=LJvalim all /*soovime täisväljundit*/
total=50; /*ÜK maht*/
var hhsize;
weight kaal;
run;
```

4. Vaatame väljundit! Kuidas on leitud hinnang kogusummale? Tõlgime!

Näide 2. **Ebavõrdsete tõenäosustega (PPS) valik**. Vaatame sama **Yldkogum**it ja selles tunnust *totinch* (leibkonna kogu sissetulek).

1. Teostame suurusega võrdelise valiku (pps) mahuga $n = 20$ andmestikust **Yldkogum**. Suuruse tunnuseks valime *totinch*:

```
proc surveyselect data=Yldkogum
method=pps
n=20
out=TotinchPPS;
size totinch;
run;
```

2. Arvutame hinnangud leibkonna keskmisele sissetulekule ja summaarsele sissetulekule. Selleks kasutame järgmist programmi:

```
proc surveymeans data=totinchPPS
total=50 /*50 objekti ÜK-s*/
mean /*hindame tunnuse keskmist*/
sum; /*hindame tunnuse kogusumma*/
var totinch;
weight Samplingweight; /*Valikukaale sisaldav tunnus*/
run;
```

3. Vaatame SURVEYMEANSi väljundit. Milline valem on kasutatud siin keskmise hindamiseks?

4. Millist arvu hindab kaalude summa? Leiame kaalude summa. Milline tuli?

5. Miks summa hinnangu dispersioon on null? Ja miks keskmise hinnangu dispersioon erineb nullist?

6. Antud juhul oleks parem keskmise hinnang sum/50!!

Näide 3. **Hindamine osakogumites**. Soovitakse hinnata nende leibkondade osakaalu, kes rendivad oma elamispinda (tunnus *tenurh*, väärtus 2) ja seda süstemaatilise valiku korral.

1. Võtame SYS valimi mahuga 20 (nimeks paneme **ValimSYS**).

2. Tekitame saadud valimisse uut binaarset tunnust **Rentnik** järgmiselt:

$$\text{Rentnik}_i = \begin{cases} 1, & \text{kui tenurh}_i = 2; \\ 0, & \text{muidu.} \end{cases}$$

3. Lisame valimisse ka kaalutunnuse.

4. Vajalikud näitajad leiame järgmise protseduuri abil:

```
proc surveymeans data=valimSYS1
mean var stderr sum varsum std /*kui muid valikuid pole, siis väljastatakse
keskmine*/
total=50;
var rentnik /*laste keskmine kulu ja gruppide osakaalud*/;
weight kaal;
run;
```

P.S. Kvalitatiivse tunnuse koraal saame samuti osakaale.

Näide 4. **Hindamine kihtides**.

1. Jagame oma **Yldkogum**i kolmeks kihiks geograafiliselt:

Kiht=1 kui 1 $\leq block \leq$24
Kiht=2 kui 25 $\leq block \leq$39
Kiht=1 kui 40 $\leq block \leq$60

Millised tulid kihtide mahud? Kirjutame üles (läheb hiljem vaja).

2. Võtame 1-st kihist 6 objekti, 2-st - 4 ja 3-st 5 objekti valimisse LJV TTA abil. (Kui ÜK-st nii palju pole võtta, siis korrigeeri valimimahtusid.)

3. Uurime, millised tunnused tekkisid juurde.

4. Selleks, et oleks võimalik leida hinnanguid kihtvaliku põhjal, peavad kihtide mahud olema sisestatud eraldi tabelina (asenda **oma** kihtide mahtudega ÜK-s!):

```
data Kokku; /*siia kihtide mahud YK-s*/
input Kiht _total_; /*_total_ on SASi systeemne tunnuse nimi,*/
/*mida ta surveymeans protseduuris total-valiku juures otsima hakkab*/
datalines;
1 25
2 15
3 10
;
run;
```

5. Nüüd leiame hinnangud kogu ÜK keskmisele ja summale kihtvaliku korral järgmiselt:

```
proc surveymeans data=valimKiht
total=Kokku /*Kihtide mahud on andmestikus Kokku*/
mean var stderr sum varsum std;
stratum Kiht; /*kihitunnus*/
var hhsize;
weight SamplingWeight;
run;
```

6. Uurime saadud väljundit.

Kas hinnangud on väljastatud kihtide kaupa või mitte?

7. Täienda oma programm lausega `domain kiht;` Mida see muudab?

ISESEISEV ÜLESANNE

1. Võtta valim mahuga 20 leibkonda LJV TGA abil.

2. Leida saadud valimi põhjal eraldi hinnangud leibkonna keskmisele sissetulekule väikelastega peredes (lapsed alla 5 aasta) ja teistes peredes.

3. Hinnangud ja 95% usaldusintervallid sisesta Moodle Wikisse (5. nädal). Lisaks kirjuta juurde, kas võime olulisuse nivool 0,05 väita, et keskmised sissetulekud kahes grupis on erinevad.