Simulācijas eksperimentu izmantošana apsekojumu izmaksu efektivitātes novērtēšanā

Mārtiņš Liberts

LR Centrālā statistikas pārvalde

2014. gada 17. marts

Šis darbs izstrādāts ar Eiropas Sociālā fonda atbalstu projektā **«Atbalsts doktora studijām Latvijas Universitātē-2»**.

Saturs

Saturs

- Mārtiņš Liberts
- LR CSP Matemātiskā nodrošinājuma daļas vadītāja vietnieks
- Statistikā kopš 1999. gada
- 2013. gada maijā – doktora grāds matemātikā (varbūtību teorijas un matemātiskās statistikas apakšnozarē)
  - Promocijas darba tēma: "Izlases dizaina optimizācija"
  - Darba zinātniskais vadītājs: Dr. habil. math., profesors Aleksandrs Šostaks
  - Darba zinātniskais konsultants: Dr. math. Jānis Lapiņš

- Prezentācijas mērķis – iepazīstināt ar simulācijas eksperimentu pielietojumiem izlases apsekojumu tematikā

Saturs

Lies, damned lies, and statistics

Lies, damned lies, and statistics

Meli, lieli meli un statistika

θ	populācijas parametrs
$\hat{\theta}$	populācijas parametra izlases novērtējums
$\hat{\theta} \neq \theta$	
$\hat{\theta} - \theta$	izlases kļūda

Nozare	**Novērt.**	**Kļ. rob.**
Pavisam	900.2	±15.6
(A) Lauksaimniecība, mežsaimniecība un zivsaimniecība	71.6	±8.6
(C) Apstrādes rūpniecība	123.5	±11.4
(D) Elektroenerģija, gāzes apgāde, siltumapgāde un gaisa kondicionēšana	15.0	±3.9
(E) Ūdens apgāde; notekūdeņu, atkritumu apsaimniekošana un sanācija	5.8	±2.0
(F) Būvniecība	71.7	±8.6
(H) Transports un uzglabāšana	82.6	±9.7
...		

Tabula : NB061. Nodarbinātie pēc saimnieciskās darbības veida pa ceturkšņiem (`http://data.csb.gov.lv`)

- Kļūdas robeža (*angl. Margin of Error*)

$$\mathrm{MoE}\left(\hat{\theta}\right)$$

$$\mathrm{p}\left(\left|\hat{\theta} - \theta\right| < \mathrm{MoE}\left(\hat{\theta}\right)\right) = 0.95$$

Nozare	**Novērt.**	**Kļ. rob.**
Pavisam	900.2	±15.6
(J) Informācijas un komunikācijas pakalpojumi	23.4	±5.3
(L) Operācijas ar nekustamo īpašumu	22.7	±6.1
...		

Tabula : NB061. Nodarbinātie pēc saimnieciskās darbības veida pa ceturkšņiem (`http://data.csb.gov.lv`)

Saturs

Izmaksu efektivitāte

- Izlases apsekojumu plānošana
- Fiksēts budžets
- Uzdevums ir atrast tādu izlases dizainu, kas minimizē izlases kļūdas fiksēta budžeta ietvaros

Definīcija

Parametra θ novērtēšanai ar fiksētām apsekojuma izmaksām γ izlases dizains $p\left(s\right)$ ir izmaksu ziņā efektīvāks par izlases dizainu $q\left(s\right)$, ja $\mathrm{Var}_p\left(\hat{\theta}_p \middle| C_p \approx \gamma\right) < \mathrm{Var}_q\left(\hat{\theta}_q \middle| C_q \approx \gamma\right)$.

$$\mathrm{MoE}\left(\hat{\theta}\right) = 1.96\sqrt{\mathrm{Var}\left(\hat{\theta}\right)}$$

Izlases apsekojuma plānošana

- Vai izlases apsekojumi ir efektīvs izmaksu ziņā?
- Vienkāršs vai sarežģīts izlases dizains
- Izlases apsekojuma plānošanas fāze:
  - Kāda ir sagaidāmā populācijas parametru novērtējumu precizitāte?
  - Kādas ir sagaidāmās apsekojuma izmaksas?
  - Kādu izlases dizainu izmantot apsekojuma veikšanai, lai minimizētu izlases kļūdas pie fiksētām apsekojuma izmaksām?

Literatūra I

- Mahalanobis (1940) un Jessen (1942)
- Hansen, Hurwitz un Madow (1953) un Kish (1965)
- Aproksimācija $C\sqrt{n}$ (Beardwood, Halton un Hammersley, 1959)

Ordinarily the sampler has no precise data on cost factors, and must base his decisions on estimates or guesses. Often he can make good enough guesses to eliminate designs that would be obviously uneconomical. (Kish, 1965)

Literatūra II

- Groves (1989)

  Since it is likely that closed form-solutions to such problems will not exist with complex cost and error models, ***simulation approaches*** *will be useful to measure the sensitivity of results to changes in various design, cost, or error parameters. (Groves, 1989)*

Literatūra III

- Izlases apsekojuma operāciju pētīšana – jauns statistikas zinātņu nozares virziens
- Chen (2008) un Cox (2012)
- "*Survey Cost Workshop*" (2006)
- "*Workshop on Microsimulation Models for Surveys*" (2011)

Simulāciju eksperimenti

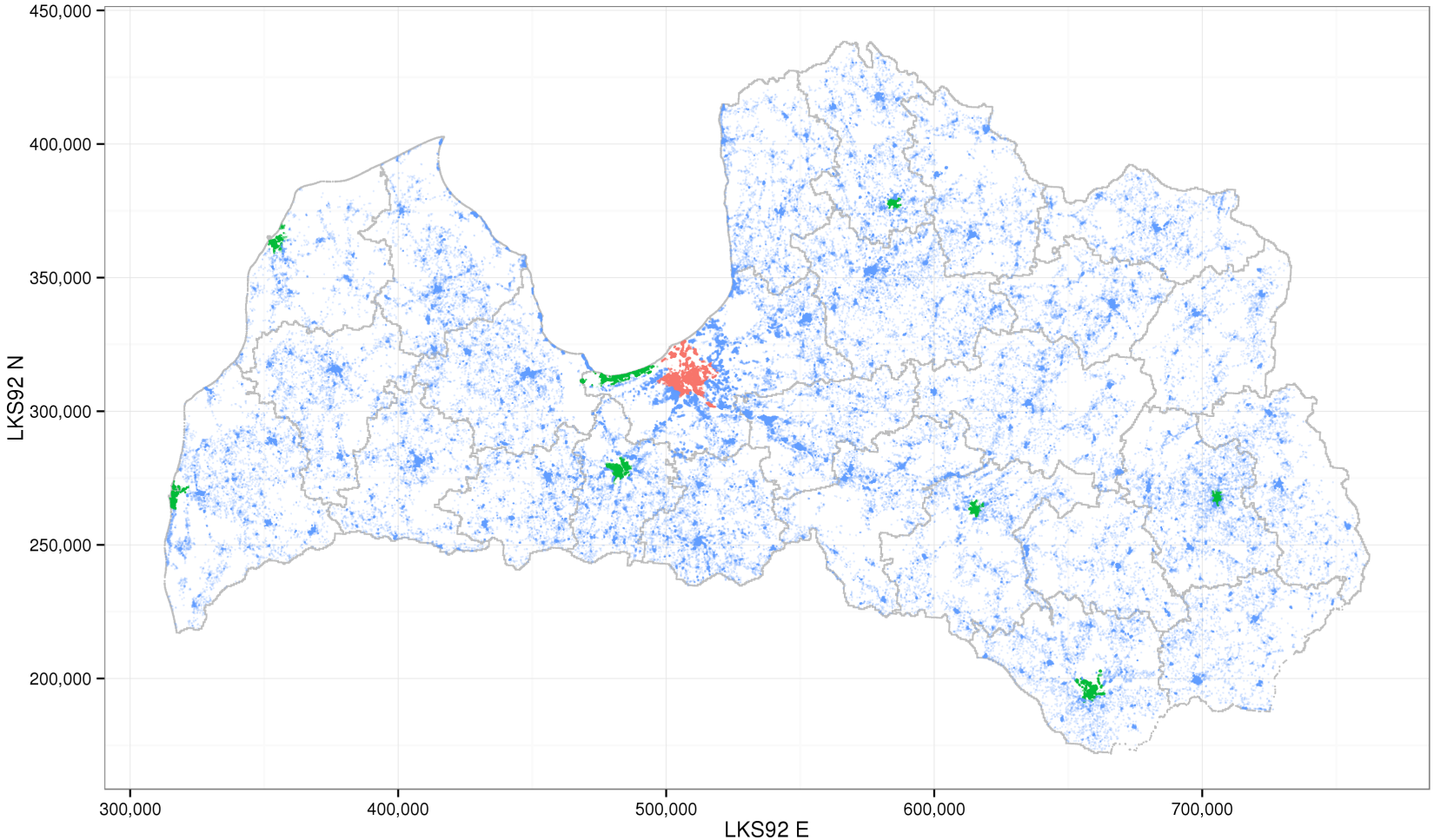
- Simulāciju eksperimenti – metode izlases apsekojumu plānošanas procesā
- Atbildētības kritums fiksēta vai samazināta budžeta situācijā
- Metodes izplatīšanās ir saistīta ar mūsdienās pieejamo "lēto" datoru jaudu

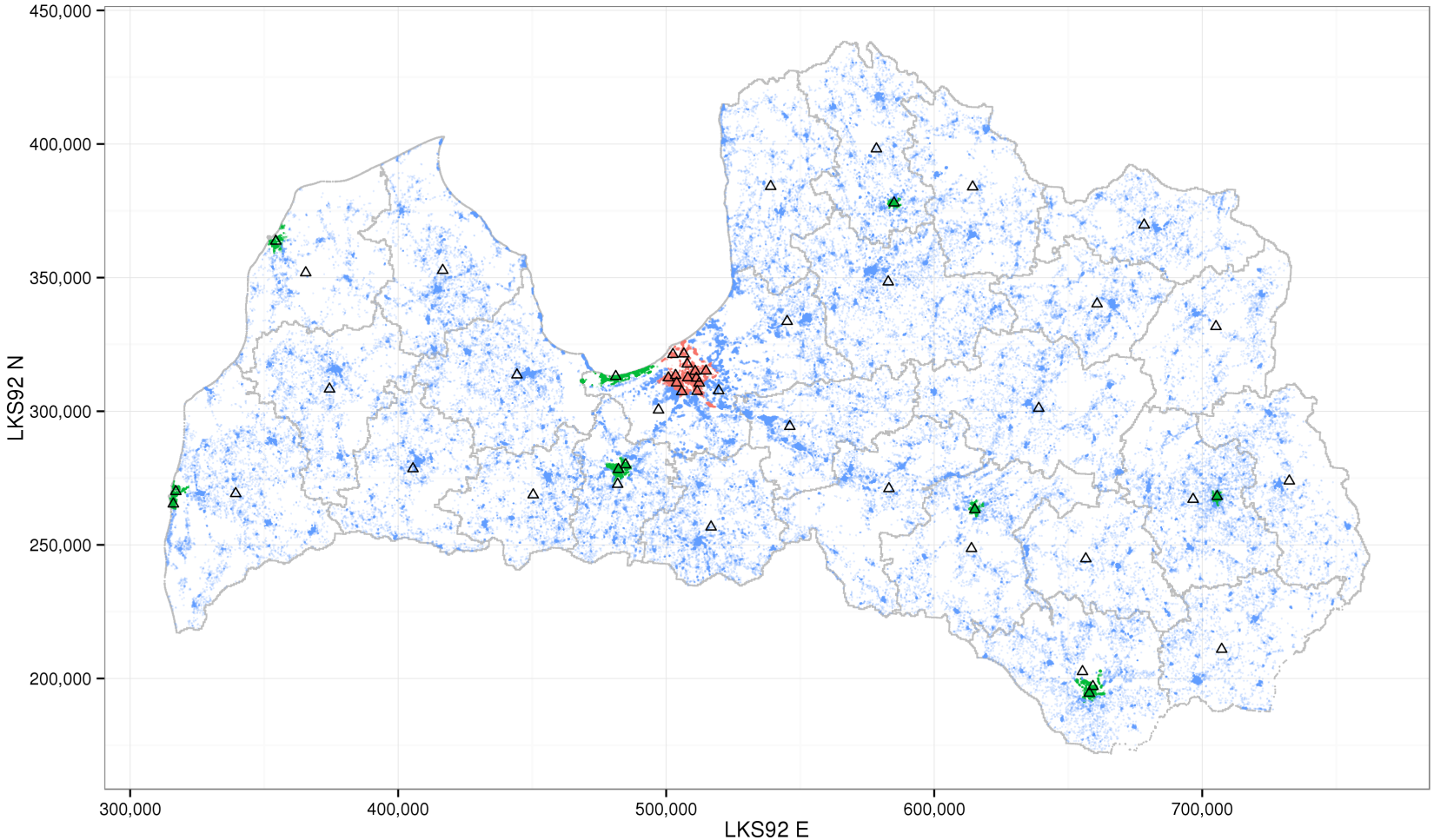
Izmaksu modelis

Izmaksu komponentes:

- ceļa izdevumi – maksa par kilometru
- intervēšanas izdevumi – maksa par anketām

$$c\left(s\right) = c_1\left(s\right) + c_2\left(s\right) = K_f C_f K_d \sum_{g=1}^{G} d_g + mC_h + nC_p$$


450,000
400,000
350,000
300,000
250,000
200,000
LKS92 N
300,000
400,000
500,000
600,000
700,000
LKS92 E


450,000
400,000
350,000
300,000
250,000
200,000
LKS92 N
300,000
400,000
500,000
600,000
700,000
LKS92 E

Saturs

Izlases apsekojuma analīzes aparāts

Ievades informācija:

1. Populācijas parametri
2. Izmaksu modelis
3. Izlases dizainu izvēle
4. Populācijas dati
5. Apsekojuma budžeta dati

Populācijas parametri

- Jādefinē populācijas parametru kopa
- Darbam izvēlēti 90 populācijas parametri
- Summārais populācijas parametrs:

$$Y = \sum_U y_i$$

- Divu populācijas summāro attiecība:

$$R = \frac{Y}{Z} = \frac{\sum_U y_i}{\sum_U z_i}$$

- Kopā Latvijā un dalījumā pa populācijas domēniem (teritorija, vecuma grupas)

Izmaksu modelis

Izmaksu komponentes:

- ceļa izdevumi – maksa par kilometru
- intervēšanas izdevumi – maksa par anketām (mājsaimniecības un personu)

$$c(s) = c_1(s) + c_2(s) = K_f C_f K_d \sum_{g=1}^{G} d_g + mC_h + nC_p$$

Izlases dizainu izvēle

References dizains:

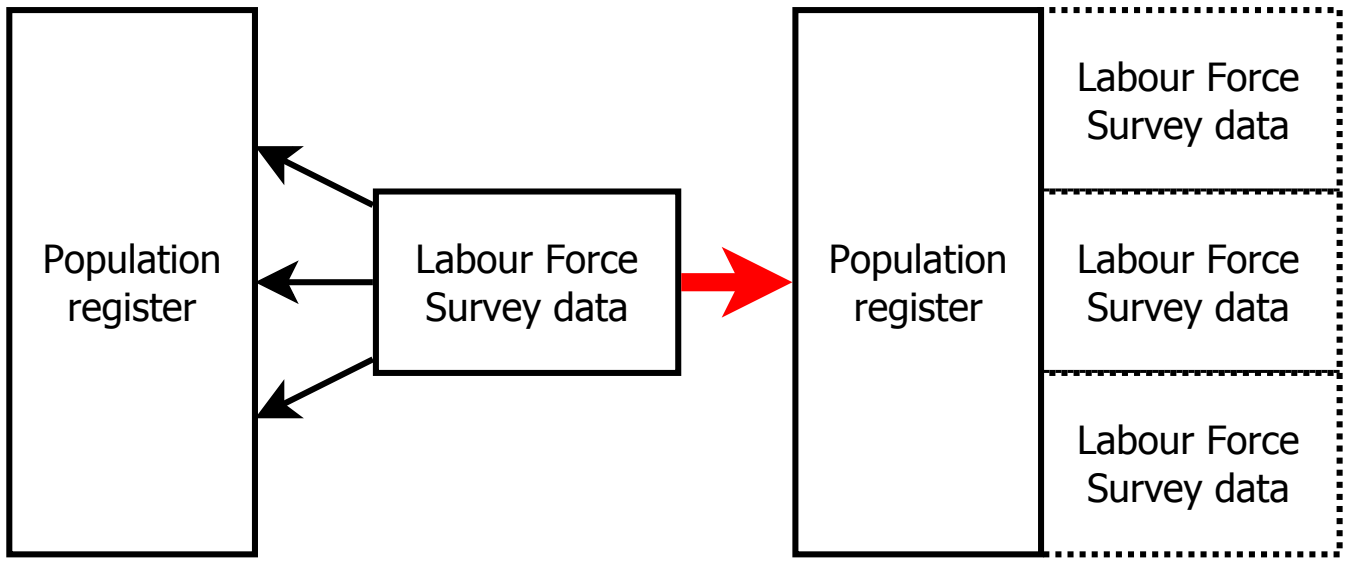
- Divpakāpju izlases dizains:
  1. Stratificētas sistemātiskas izlases dizains teritorijām ar izlasē iekļūšanas varbūtībām proporcionālām teritoriju lielumam
  2. Vienkāršā gadījuma izlase mājokļiem

Alternatīvie izlases dizaini:

- Stratificēta vienkāršā gadījuma izlase personām
- Stratificēta vienkāršā gadījuma izlase mājokļiem

Populācijas dati

- Tautas skaitīšanas dati
- Statistiskā mājokļu reģistra dati kombinēti ar Darbaspēka apsekojuma datiem

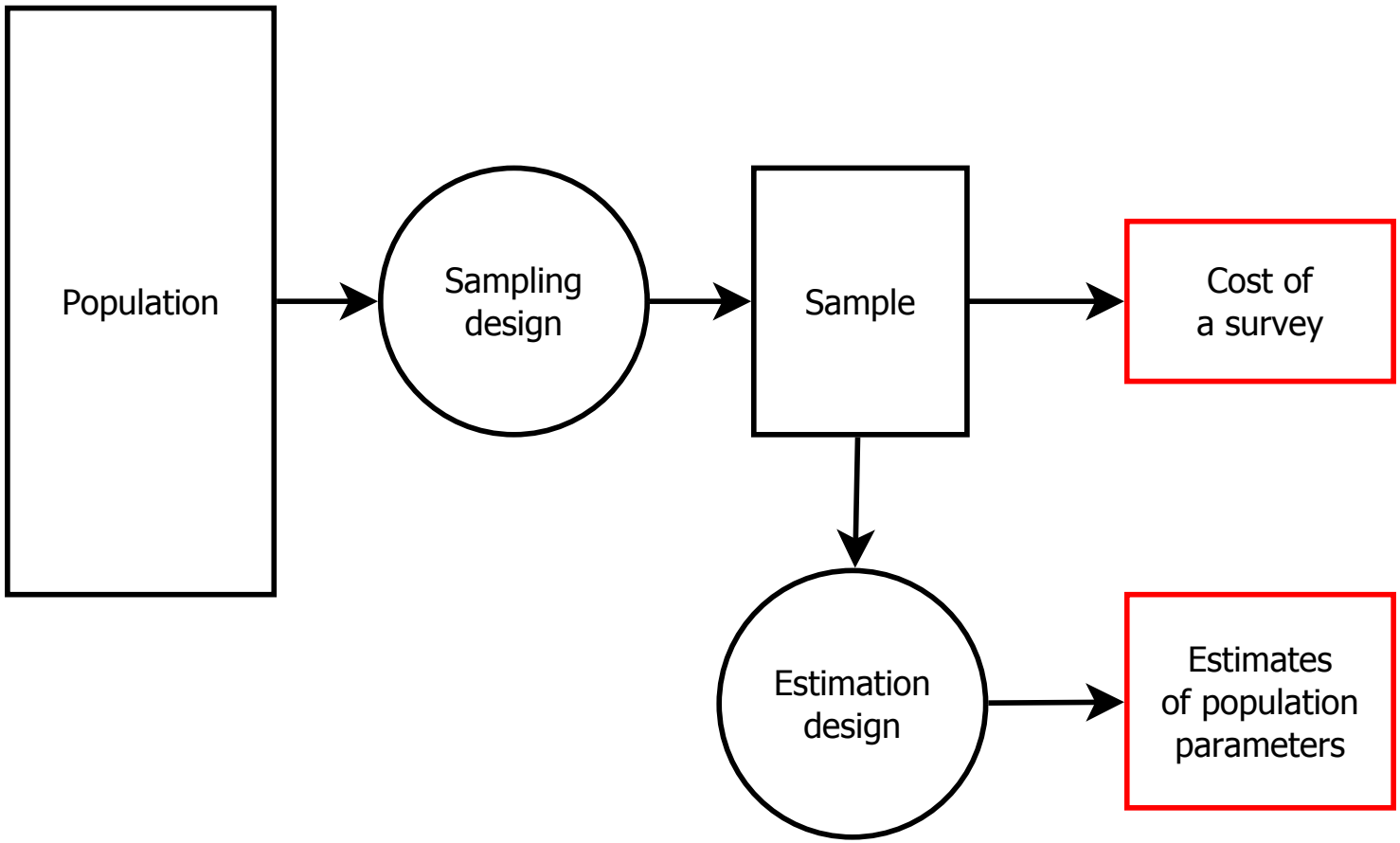


Apsekojuma budžets

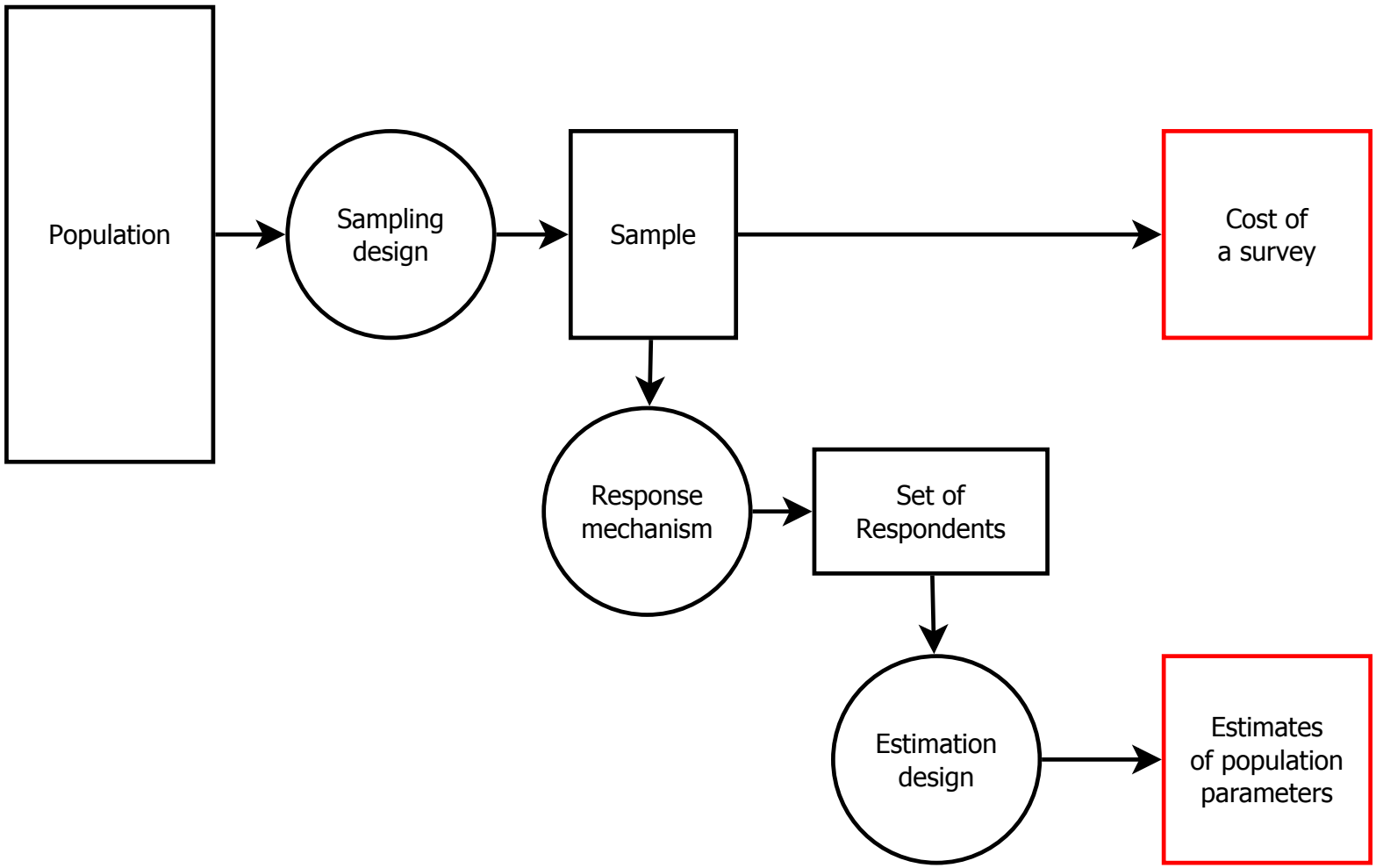
- Darbaspēka apsekojuma tuvinātas izmaksas 2010. gadā

Saturs

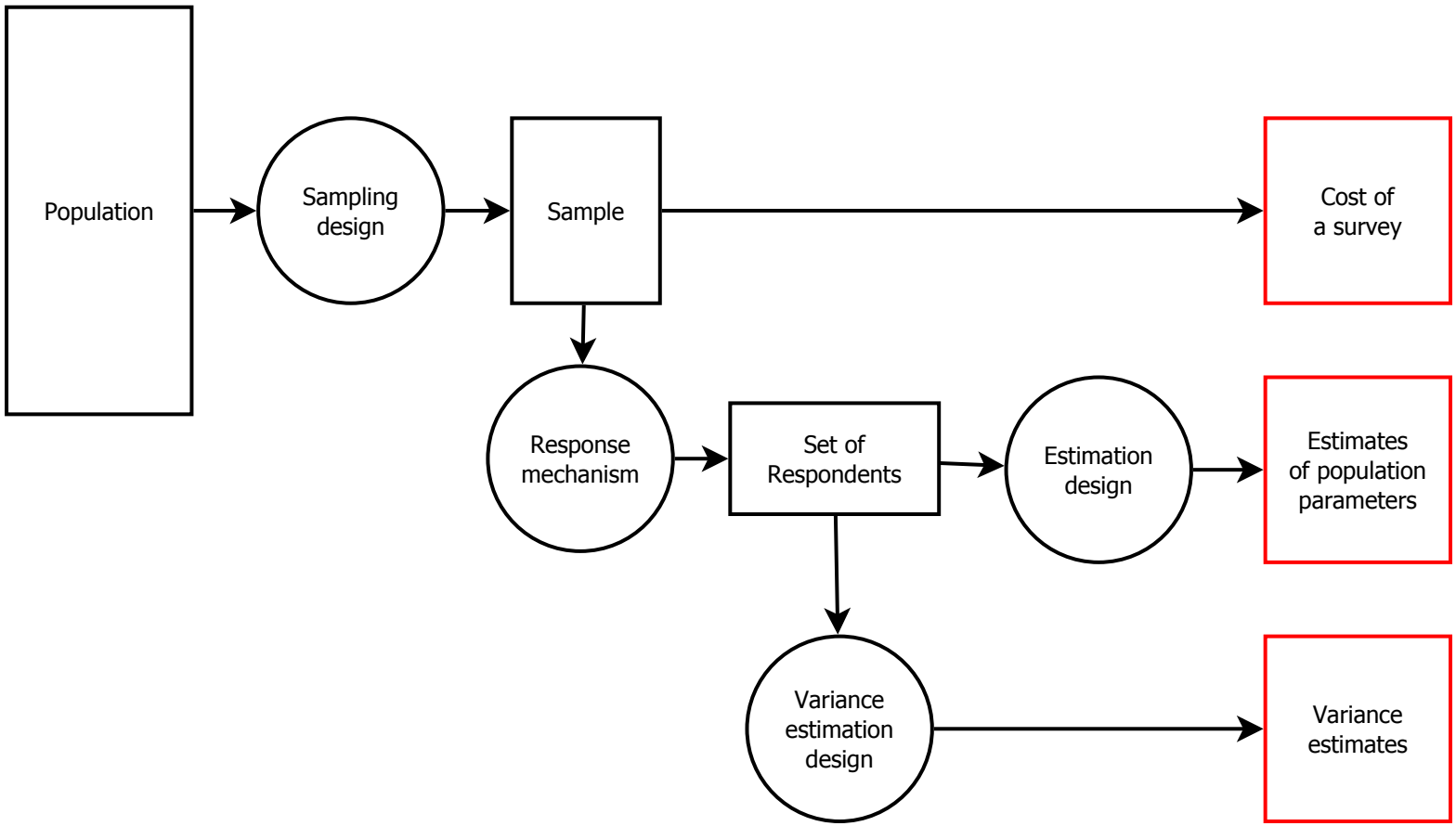
Simulācijas modeļu piemēri



Simulācijas modeļu piemēri



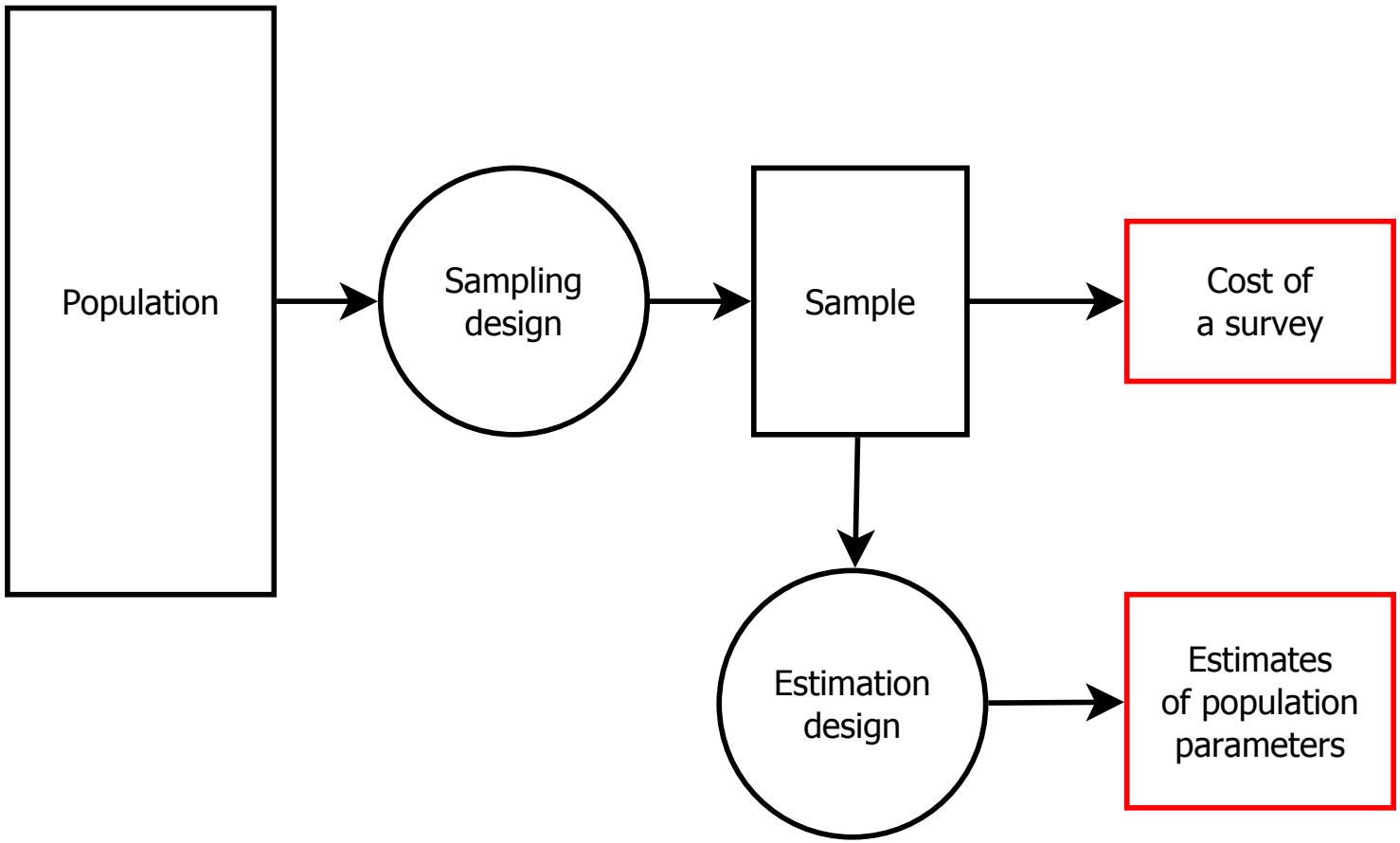
Simulācijas modeļu piemēri



Divu soļu simulācija

1. Izlases dizaina parametru noteikšana, lai sasniegtu līdzīgas sagaidāmās izmaksas
2. Precizitātes noteikšana populācijas parametru novērtējumiem

Simulācijas modeļu piemēri



Pirmais solis

- Izvēlas izlases apjomu režģi alternatīvajiem izlases dizainiem
- Ar simulāciju palīdzību novērtē sagaidāmās izmaksas katrā režģa punktā
- Aproksimē izmaksu funkcijas atkarību no izlases apjoma
- Nosaka intervālu, kas ietver atrisinājumu
- Intervālu sadala režģa punktos ar mazu soli
- Ar simulāciju palīdzību novērtē sagaidāmās izmaksas katrā režģa punktā
- Nosaka atrisinājumu

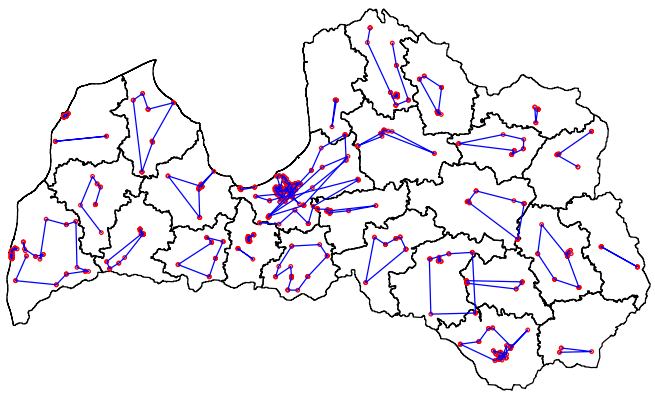
Pirmā soļa atrisinājums

stratum	design	n.h	n.i	c.travel	c.interview	c.total
Riga	SSRSi	.	1 261	403	5 037	5 440
Riga	SSRSh	1 001	2 105	352	5 108	5 460
Riga	TSSh	1 040	2 185	91	5 305	5 395
Cities	SSRSi	.	1 781	660	7 099	7 759
Cities	SSRSh	1 404	2 963	582	7 175	7 757
Cities	TSSh	1 456	3 073	279	7 441	7 720
Other	SSRSi	.	2 834	11 631	11 302	22 933
Other	SSRSh	2 340	5 554	10 357	12 574	22 930
Other	TSSh	3 536	8 318	3 964	18 926	22 890
Total	SSRSi	.	5 876	12 694	23 438	36 132
Total	SSRSh	4 745	10 622	11 291	24 857	36 147
Total	TSSh	6 032	13 576	4 334	31 672	36 005

Vienpakāpju un divpakāpju izlases

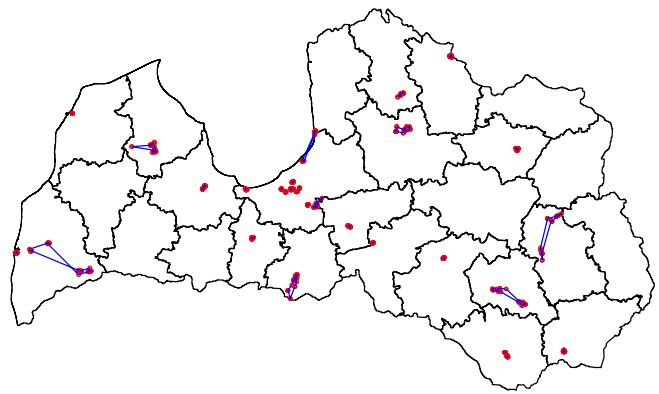
Cluster Sampling of persons

Sample size = 464; Trip = 3242 (km)

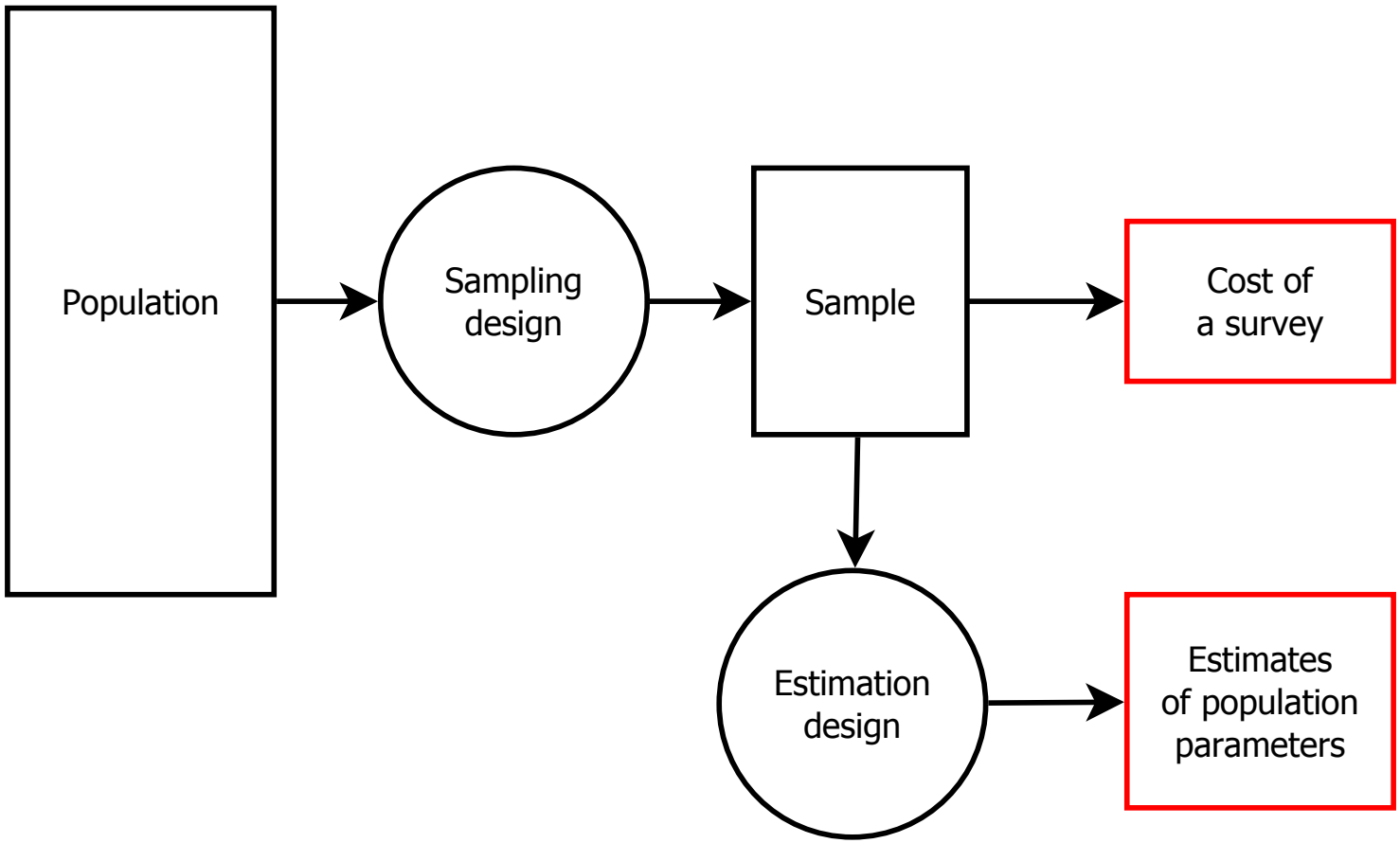


Two Stage Sampling of Dwellings

Sample size = 464; Trip = 487 (km)



Simulācijas modeļu piemēri



Otrā soļa rezultāti

par.	age	value	σ_1	σ_2	σ_3	p-val.	des
empl	15–74	330 855	7 381	8 272	8 329	1.000	SSRSi
unem	15–74	47 160	4 284	3 569	3 504	0.000	TSSh
inact	15–74	160 949	6 938	6 062	6 009	0.040	SSRSh
empl	15–24	31 245	3 543	2 903	2 960	1.000	SSRSh
unem	15–24	8 152	1 851	1 452	1 435	0.011	SSRSh
inact	15–24	40 138	3 980	3 300	3 301	0.508	SSRSh
empl	25–74	299 610	7 533	7 509	7 430	0.017	SSRSh
unem	25–74	39 007	3 928	3 222	3 184	0.008	TSSh
inact	25–74	120 810	6 322	5 329	5 250	0.001	TSSh

Tabula : Precizitātes novērtējumi Rīgā

Otrā soļa rezultāti

par.	age	value	σ_1	σ_2	σ_3	p-val.	des
empl	15–74	330 855	7 381	8 272	8 329	1.000	SSRSi
unem	15–74	47 160	4 284	3 569	3 504	0.000	TSSh
inact	15–74	160 949	6 938	6 062	6 009	0.040	SSRSh
empl	15–24	31 245	3 543	2 903	2 960	1.000	SSRSh
unem	15–24	8 152	1 851	1 452	1 435	0.011	SSRSh
inact	15–24	40 138	3 980	3 300	3 301	0.508	SSRSh
empl	25–74	299 610	7 533	7 509	7 430	0.017	SSRSh
unem	25–74	39 007	3 928	3 222	3 184	0.008	TSSh
inact	25–74	120 810	6 322	5 329	5 250	0.001	TSSh

Tabula : Precizitātes novērtējumi Rīgā

Otrā soļa rezultāti

par.	age	value	σ_1	σ_2	σ_3	p-val.	des
empl	15–74	278 650	7 004	7 761	5 583	0.000	TSSh
unem	15–74	36 859	3 103	2 405	2 085	0.000	TSSh
inact	15–74	177 540	6 129	5 698	4 516	0.000	TSSh
empl	15–24	34 396	3 001	2 401	2 043	0.000	TSSh
unem	15–24	9 078	1 568	1 165	1 023	0.000	TSSh
inact	15–24	56 926	3 802	3 252	3 013	0.000	TSSh
empl	25–74	244 254	6 779	6 787	4 821	0.000	TSSh
unem	25–74	27 781	2 710	2 043	1 754	0.000	TSSh
inact	25–74	120 615	5 285	4 461	3 473	0.000	TSSh

Tabula : Precizitātes novērtējumi lauku teritorijās

Otrā soļa rezultāti

Domain	SSRSi	SSRSh	TSSh	Total
Latvia	1	0	17	18
Riga	1	6	11	18
Cities	1	4	13	18
Towns	0	0	18	18
Rural areas	0	0	18	18
Total	3	10	77	90

- TSSh tika izvēlēts kā efektīvākais izlases dizains 77 gadījumos no 90

Saturs

Rezultāti

- Mākslīgās populācijas datu ģenerēšanas metodoloģija
- Izstrādāts izlases dizainu efektivitātes analīzes aparāts (izmantojot, simulācijas eksperimentu metodes)
- Aparāta praktiskai realizācijai ir izstrādāts programmas R kods
- Izstrādātais aparāts ir pielietots trīs izlases dizainu analīzei
- Ir secināts, ka tagadējais divpakāpju izlases dizains ir efektīvāks, salīdzinot ar vienpakāpju izlases dizainiem

Secinājumi

- Piedāvātā pieeja ir noderīga apsekojuma plānošanas un izvērtēšanas stadijā
- Ar simulāciju palīdzību var vērtēt dažādu izmaiņu ietekmi uz apsekojuma izmaksām un precizitāti
- Sākumā ir jāiegulda darbs populācijas datu sagatavošanai un modeļu izstrādei
- Jāseko līdzi modeļu atbilstībai reālajai situācijai

Izmantotā literatūra I

Beardwood, J., Halton, J. H. un Hammersley, J. M. (1959). The shortest path through many points. *Mathematical Proceedings of the Cambridge Philosophical Society*, *55*, 299327.

Calinescu, M., Bhulai, S. un Schouten, B. (2013). Optimal resource allocation in survey designs. *European Journal of Operational Research*, *226*(1), 115–121.

Chen, B.-C. (2008). *Stochastic simulation of field operations in surveys* (pētnieciskais ziņojums). Washington: U. S. Census Bureau. Pieejams: `https://www.census.gov/srd/www/byyear.html`

Izmantotā literatūra II

Cox, L. (2012). The case for simulation models of federal surveys. *Research conference papers of federal committee on statistical methodology research conference 2012.* Washington. Pieejams: `http://www.fcsm.gov/events/papers2012.html`

Groves, R. M. (1989). *Survey errors and survey costs*. New Jersey: Wiley.

Hansen, M. H., Hurwitz, W. N. un Madow, W. G. (1953). *Sample survey methods and theory* (sēj. I). New-York: Wiley.

Jessen, R. J. (1942). *Statistical investigation of a sample survey for obtaining farm facts* (Research Bulletin Nr. 304). Iowa State College of Agriculture and Mechanic Arts.

Izmantotā literatūra III

Kish, L. (1965). *Survey sampling*. New-York: John Wiley & Sons.

Mahalanobis, P. C. (1940). A sample survey of the acreage under jute in Bengal. *Sankhyā: The Indian Journal of Statistics*, *4*(4), 511–530.

Resource allocation is a relatively new research area in survey designs and has not been fully addressed in the literature. Recently, the declining participation rates and increasing survey costs have steered research interests towards resource planning. Survey organizations across the world are considering the development of new mathematical models in order to improve the quality of survey results while taking into account optimal resource planning. (Calinescu, Bhulai un Schouten, 2013)