Izlases kļūdu novērtēšana ar programmas *R* bibliotēku `vardpoor`

Mārtiņš Liberts, Juris Breidaks

Centrālā statistikas pārvalde

2016. gada 22. novembrī

Centrālā statistikas pārvalde

Saturs

Saturs

Izlases apsekojumi

- Datu vākšana izlases veidā
- Populācija (ģenerālkopa)
- Varbūtiskās izlases
- Vienkāršā gadījuma izlase
- Izlases kļūda

Piemērs #1

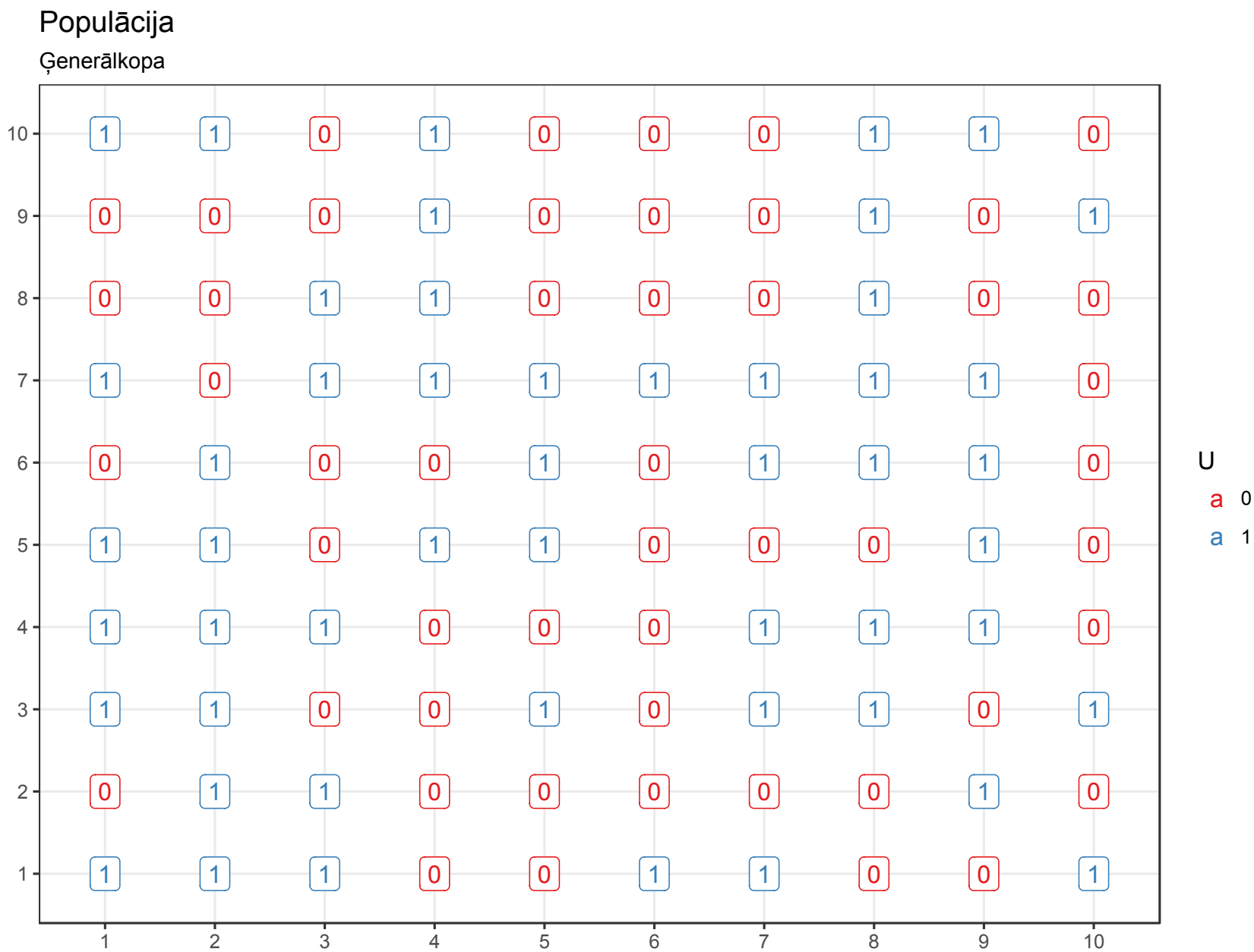
- Populācijas apjoms $N = 100$
- Binārs izpētes mainīgais ar vērtībām 0 vai 1

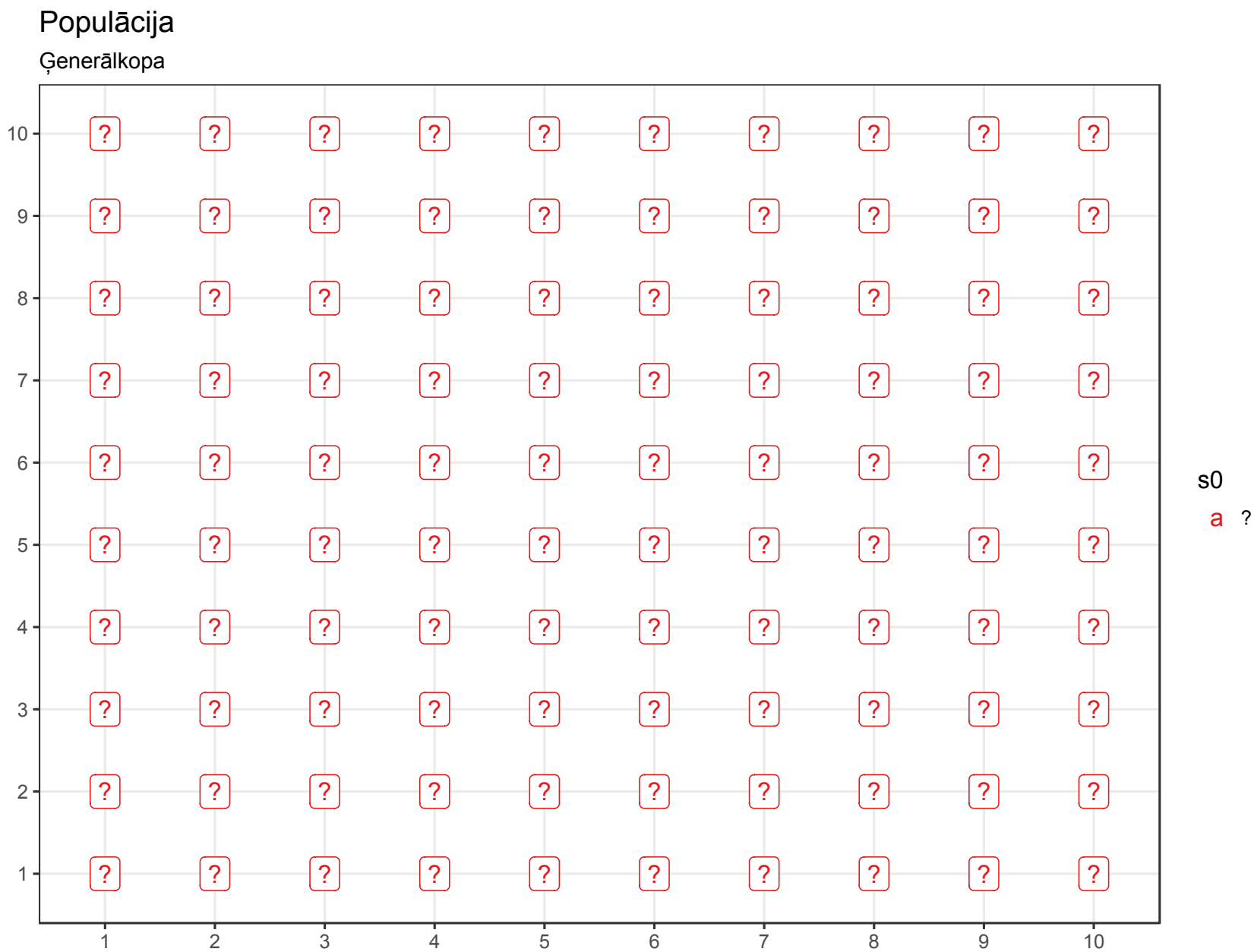
$$y_i = \begin{cases} 0 \\ 1 \end{cases}$$

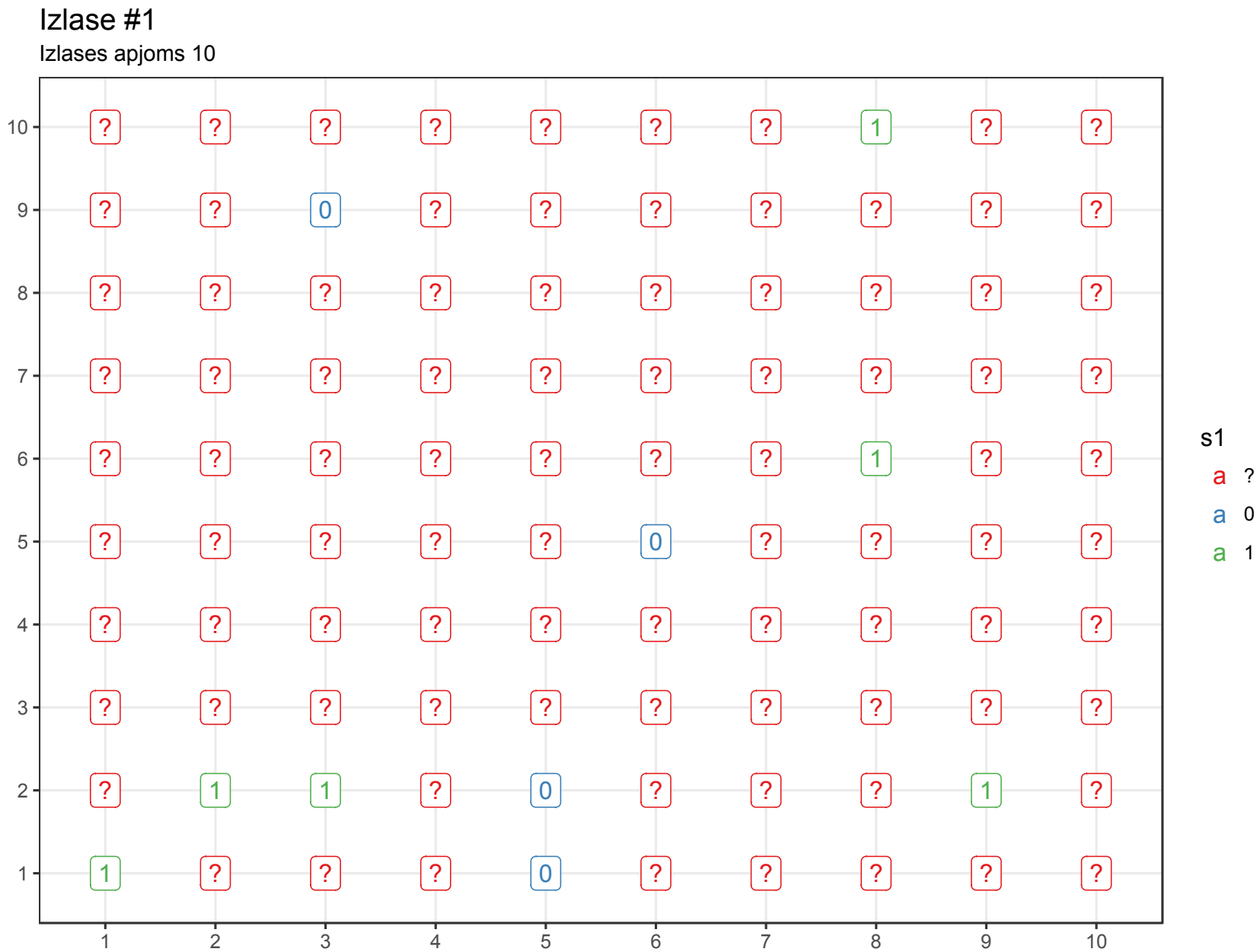
- Populācijas parametrs: proporcija

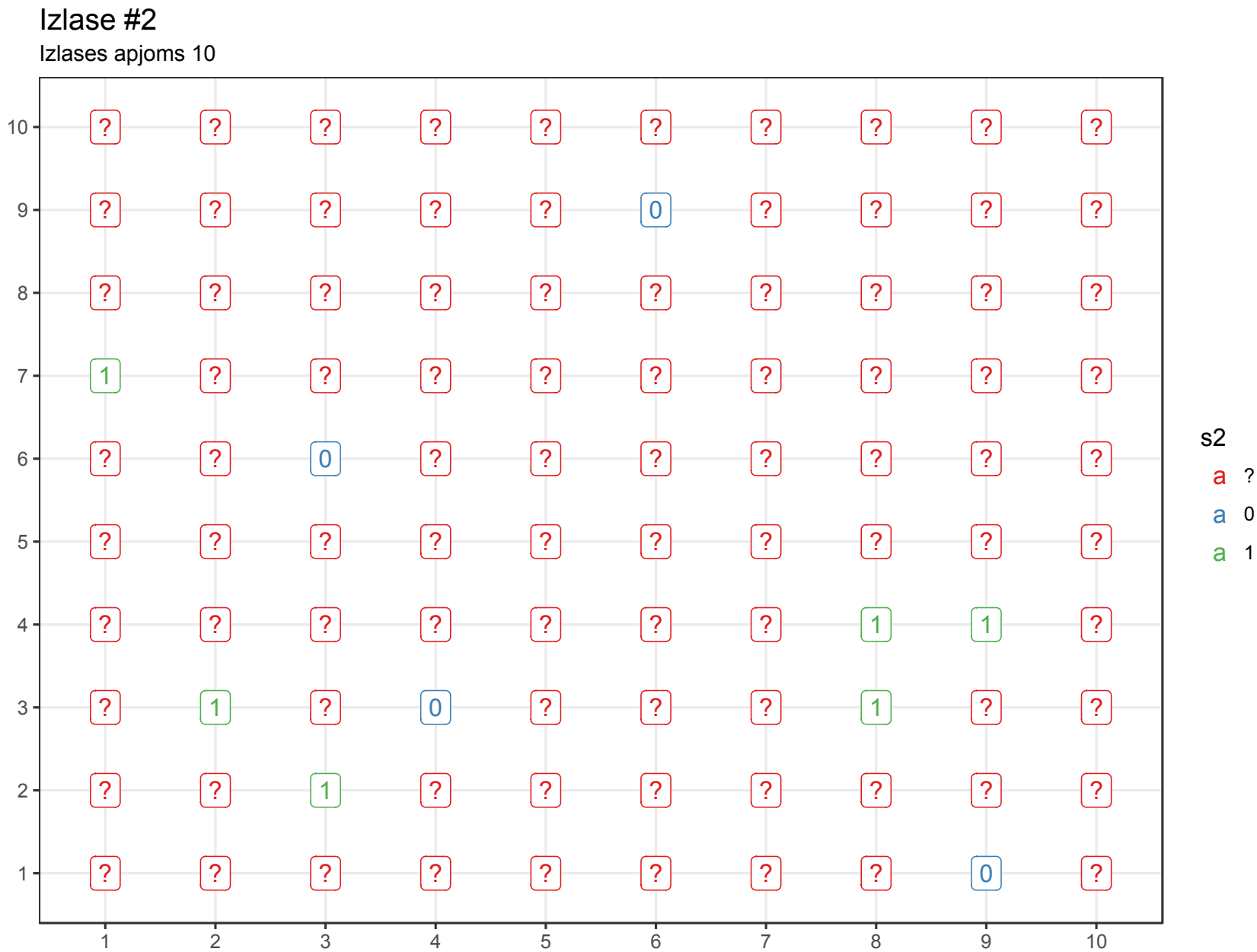
$$P = \frac{1}{N} \sum_U y_i = 0{,}5$$

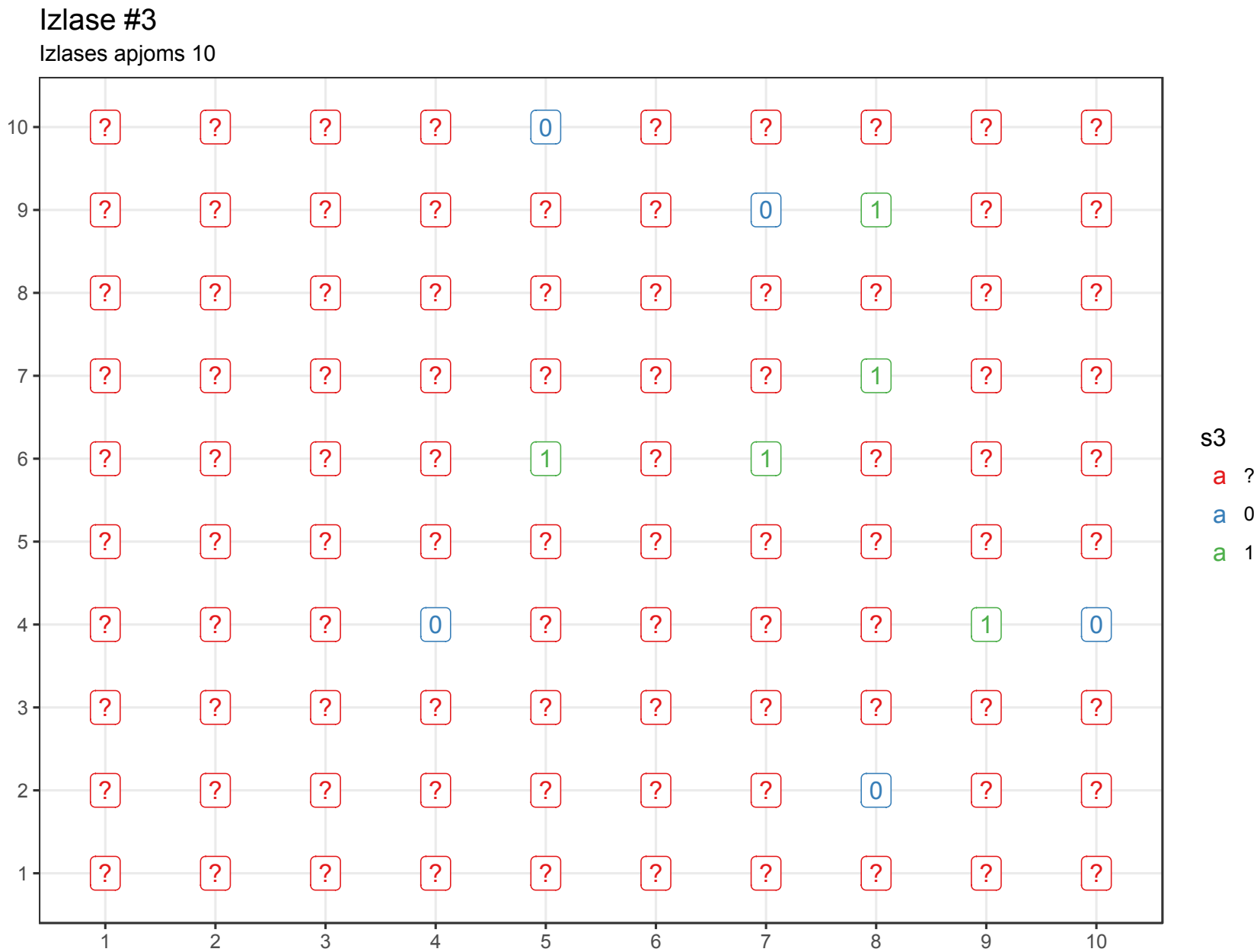
- Izlases apjoms $n = 10$
- Vienkāršā gadījuma izlase


Populācija
Ģenerālkopa
10 1 1 0 1 0 0 0 1 1 0
9 0 0 0 1 0 0 0 1 0 1
8 0 0 1 1 0 0 0 1 0 0
7 1 0 1 1 1 1 1 1 1 0
6 0 1 0 0 1 0 1 1 1 0
5 1 1 0 1 1 0 0 0 1 0
4 1 1 1 0 0 0 1 1 1 0
3 1 1 0 0 1 0 1 1 0 1
2 0 1 1 0 0 0 0 0 1 0
1 1 1 1 0 0 1 1 0 0 1
1 2 3 4 5 6 7 8 9 10
U
a 0
a 1


Populācija
Ģenerālkopa
10
9
8
7
6
5
4
3
2
1
1
2
3
4
5
6
7
8
9
10
s0
a ?


Izlase #1
Izlases apjoms 10
10
9
8
7
6
5
4
3
2
1
1
2
3
4
5
6
7
8
9
10
s1
a ?
a 0
a 1


Izlase #2
Izlases apjoms 10
10
9
8
7
6
5
4
3
2
1
1
2
3
4
5
6
7
8
9
10
s2
a ?
a 0
a 1


Izlase #3
Izlases apjoms 10
10
9
8
7
6
5
4
3
2
1
1
2
3
4
5
6
7
8
9
10
s3
a ?
a 0
a 1

Piemērs #1

- Izlases proporcija p ir populācijas proporcijas P novērtējums

$$\hat{P} = p = \frac{1}{n} \sum_{s} y_i$$

  - **Ja tiek lietota vienkāršā gadījuma izlase**
- Izlases kļūda $\hat{P} - P$
- Trīs izlases:
  - $\hat{P}_1 = 0{,}6$; izlases kļūda ir $0{,}1$
  - $\hat{P}_2 = 0{,}6$; izlases kļūda ir $0{,}1$
  - $\hat{P}_3 = 0{,}5$; izlases kļūda ir $0{,}0$

Piemērs #2

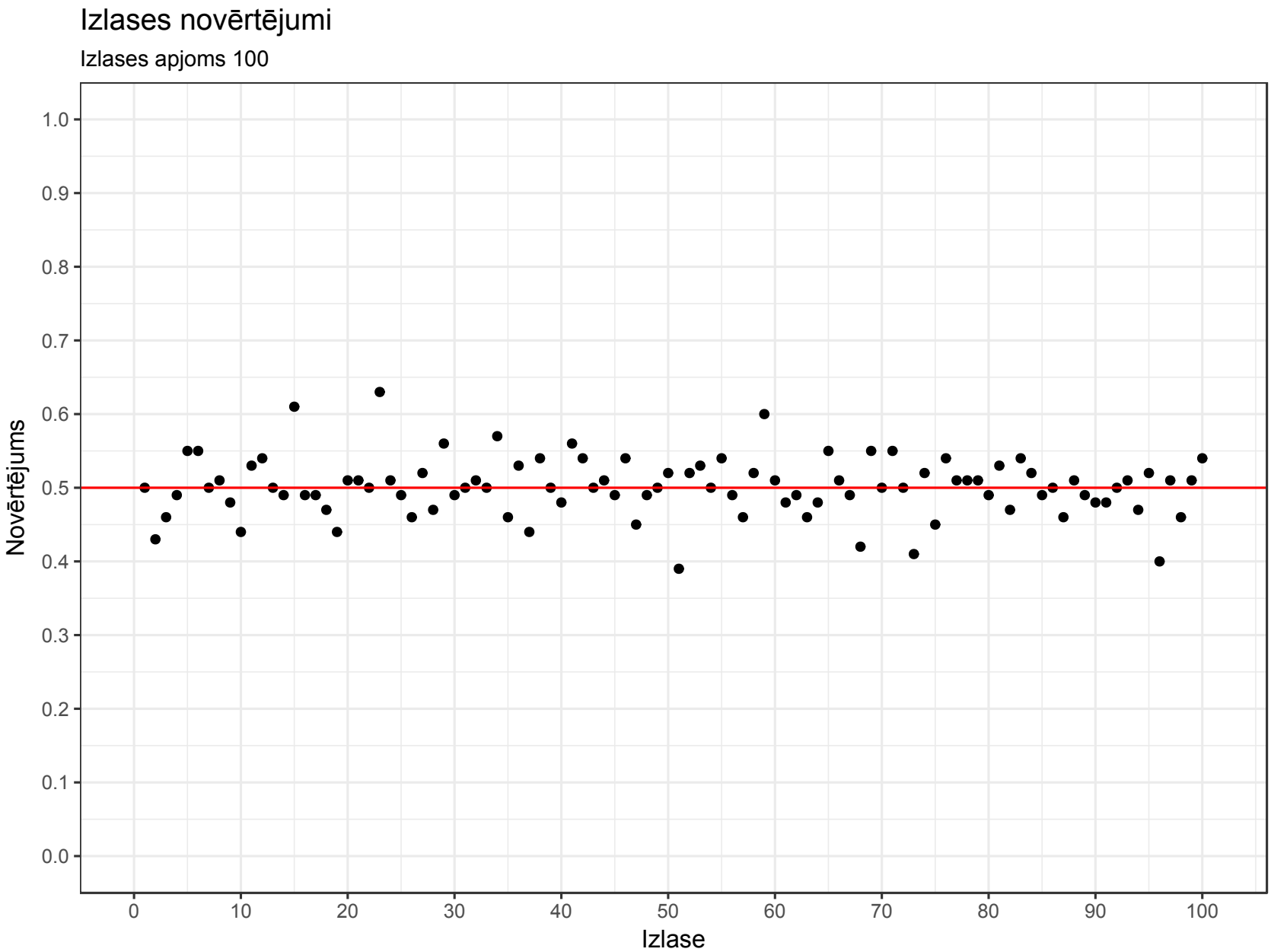
- Populācijas apjoms $N = 10\,000$
- Binārs izpētes mainīgais ar vērtībām 0 vai 1

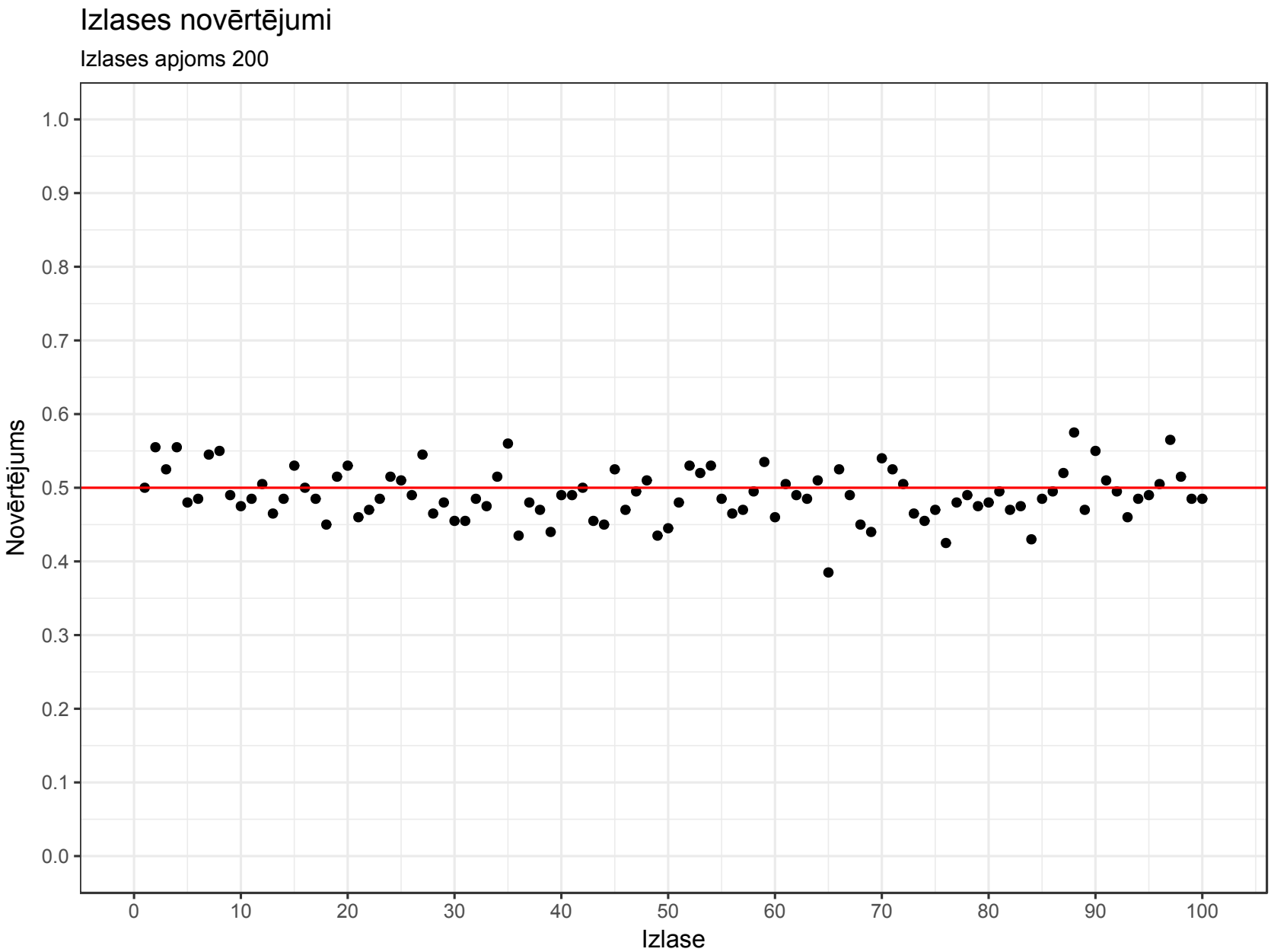
$$y_i = \begin{cases} 0 \\ 1 \end{cases}$$

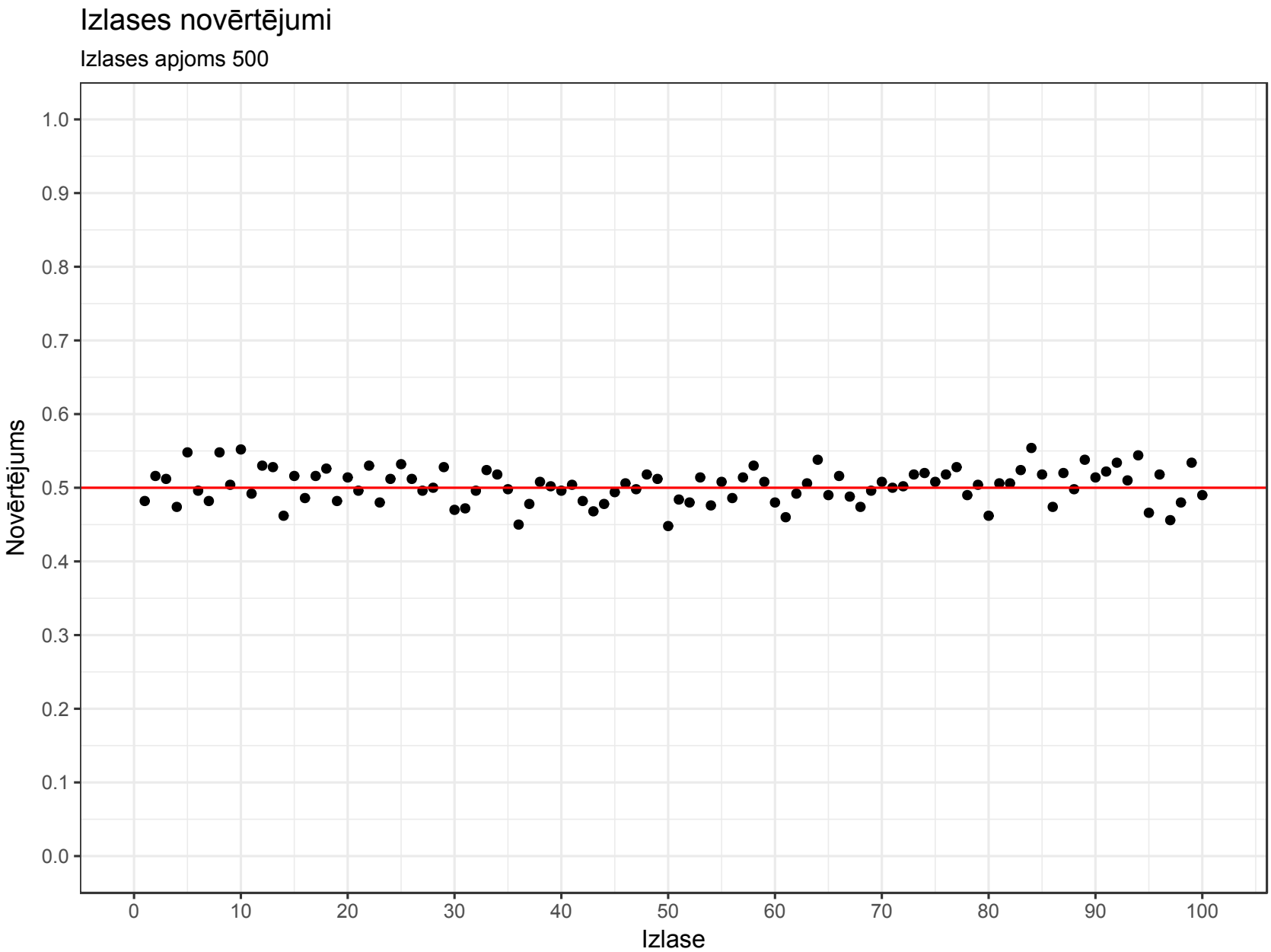
- Populācijas parametrs: proporcija

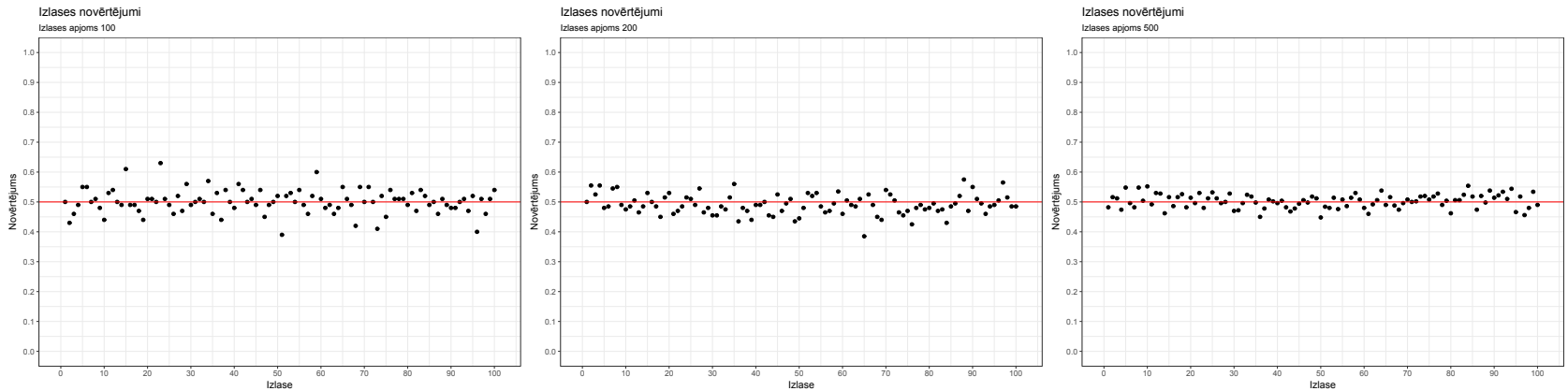
$$P = \frac{1}{N} \sum_U y_i = 0{,}5$$

- Izlases apjomi $n = \{100; 200; 500\}$
- Vienkāršā gadījuma izlase

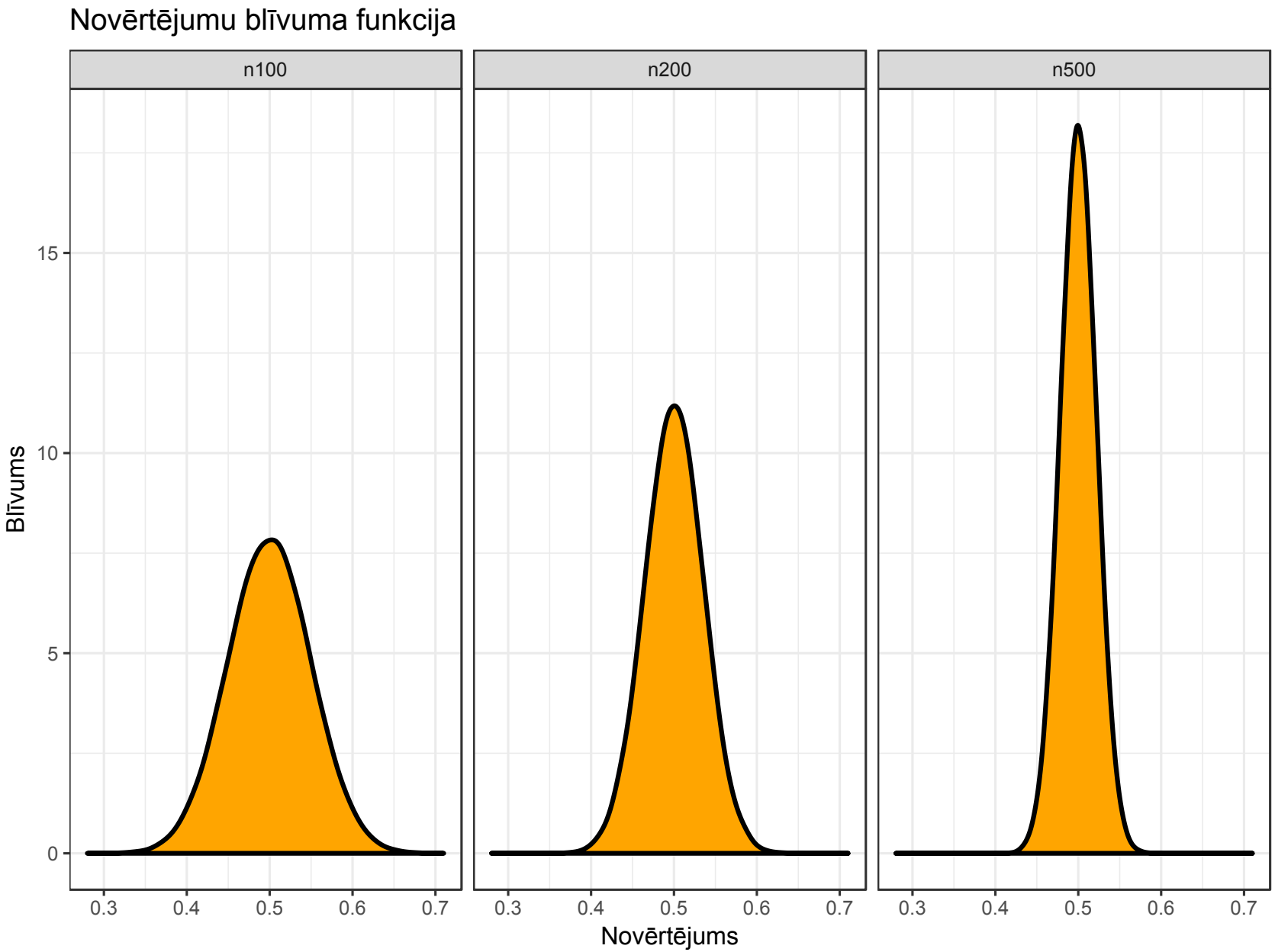

Izlases novērtējumi
Izlases apjoms 100
Novērtējums
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
10
20
30
40
50
60
70
80
90
100
Izlase


Izlases novērtējumi
Izlases apjoms 200
Novērtējums
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
10
20
30
40
50
60
70
80
90
100
Izlase


Izlases novērtējumi
Izlases apjoms 500
Novērtējums
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
10
20
30
40
50
60
70
80
90
100
Izlase


Izlases novērtējumi
Izlases apjoms 100
Izlases novērtējumi
Izlases apjoms 200
Izlases novērtējumi
Izlases apjoms 500
Novērtējums
Izlase

- Vai izlases kļūdu $\hat{P} - P$ var aprēķināt?
- Izlases kļūdu precīzi nekad nevar aprēķināt
- Var aprēķināt (novērtēt) **kļūdas robežu**


Novērtējumu blīvuma funkcija
n100
n200
n500
Blīvums
0
5
10
15
0.3
0.4
0.5
0.6
0.7
Novērtējums

- Datu dispersija

$$S_y^2 = \frac{1}{N-1} \sum_U \left(y_i - \bar{Y} \right)^2$$

- Proporcijas gadījumā

$$S_P^2 = \frac{N}{N-1} P\left(1-P\right) \approx P\left(1-P\right)$$

- Proporcijas novērtējuma dispersija

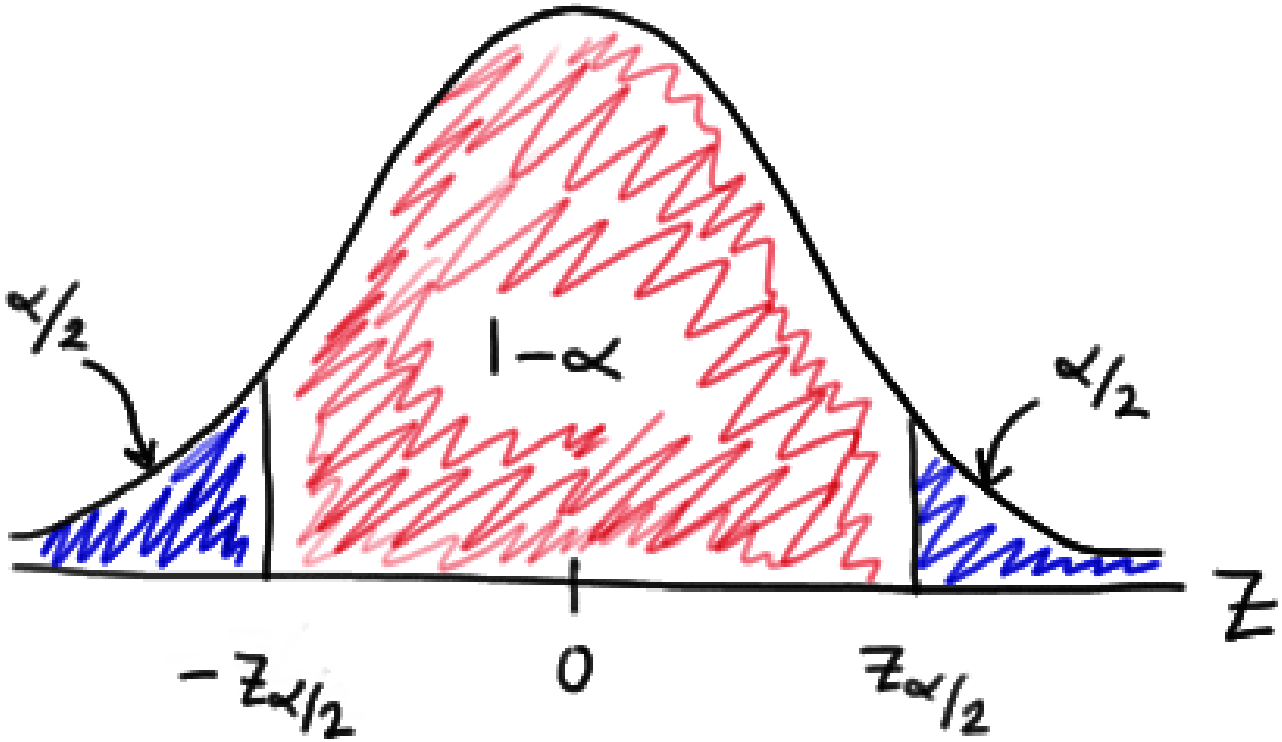
$$\text{Var}\left(\hat{P}\right) = \frac{1 - \frac{n}{N}}{n} S_P^2 \approx \frac{S_P^2}{n} \approx \frac{P\left(1 - P\right)}{n}$$

- Proporcijas novērtējuma standartkļūda

$$\text{SE}\left(\hat{P}\right) = \sqrt{\text{Var}\left(\hat{P}\right)}$$

- Proporcijas novērtējuma kļūdas robeža

$$\underset{1-\alpha}{\text{MoE}}\left(\hat{P}\right) = Z_{\frac{\alpha}{2}} \sqrt{\text{Var}\left(\hat{P}\right)}$$



https://onlinecourses.science.psu.edu/stat414/book/export/html/194

- $\alpha = 0{,}05$
- $1 - \alpha = 0{,}95$
- $Z_{\frac{\alpha}{2}} = 1{,}96$
- Proporcijas novērtējuma kļūdas robeža

$$\underset{1-\alpha}{\mathsf{MoE}}\left(\hat{P}\right) = Z_{\frac{\alpha}{2}}\sqrt{\mathsf{Var}\left(\hat{P}\right)}$$

$$\underset{0{,}95}{\mathsf{MoE}}\left(\hat{P}\right) = 1{,}96\sqrt{\mathsf{Var}\left(\hat{P}\right)}$$

- Varbūtība, ka absolūtā kļūda ir lielāka par kļūdas robežu

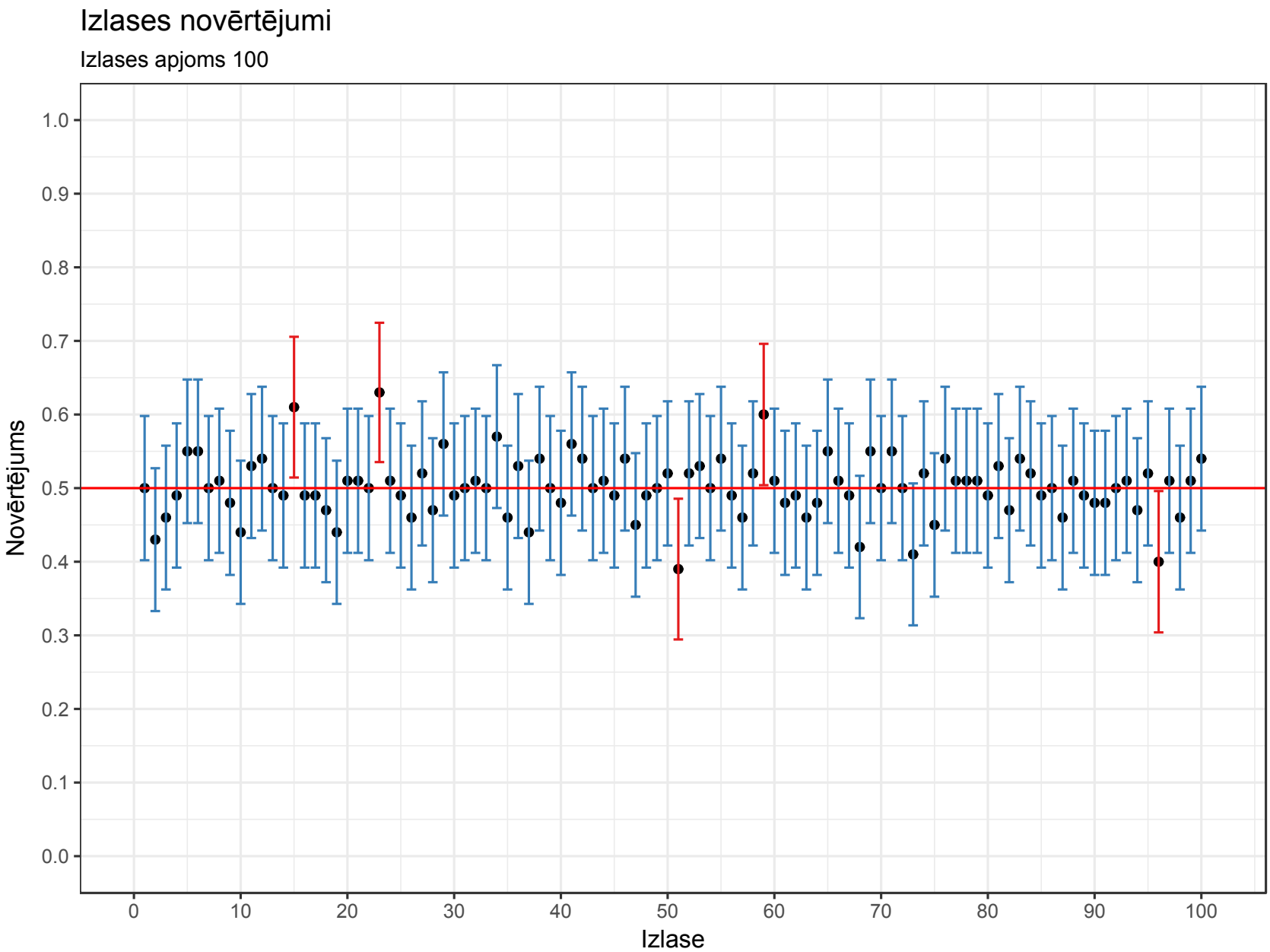
$$\text{Prob}\left[\left|\hat{P}-P\right| > \underset{1-\alpha}{\text{MoE}}\left(\hat{P}\right)\right] = \alpha$$

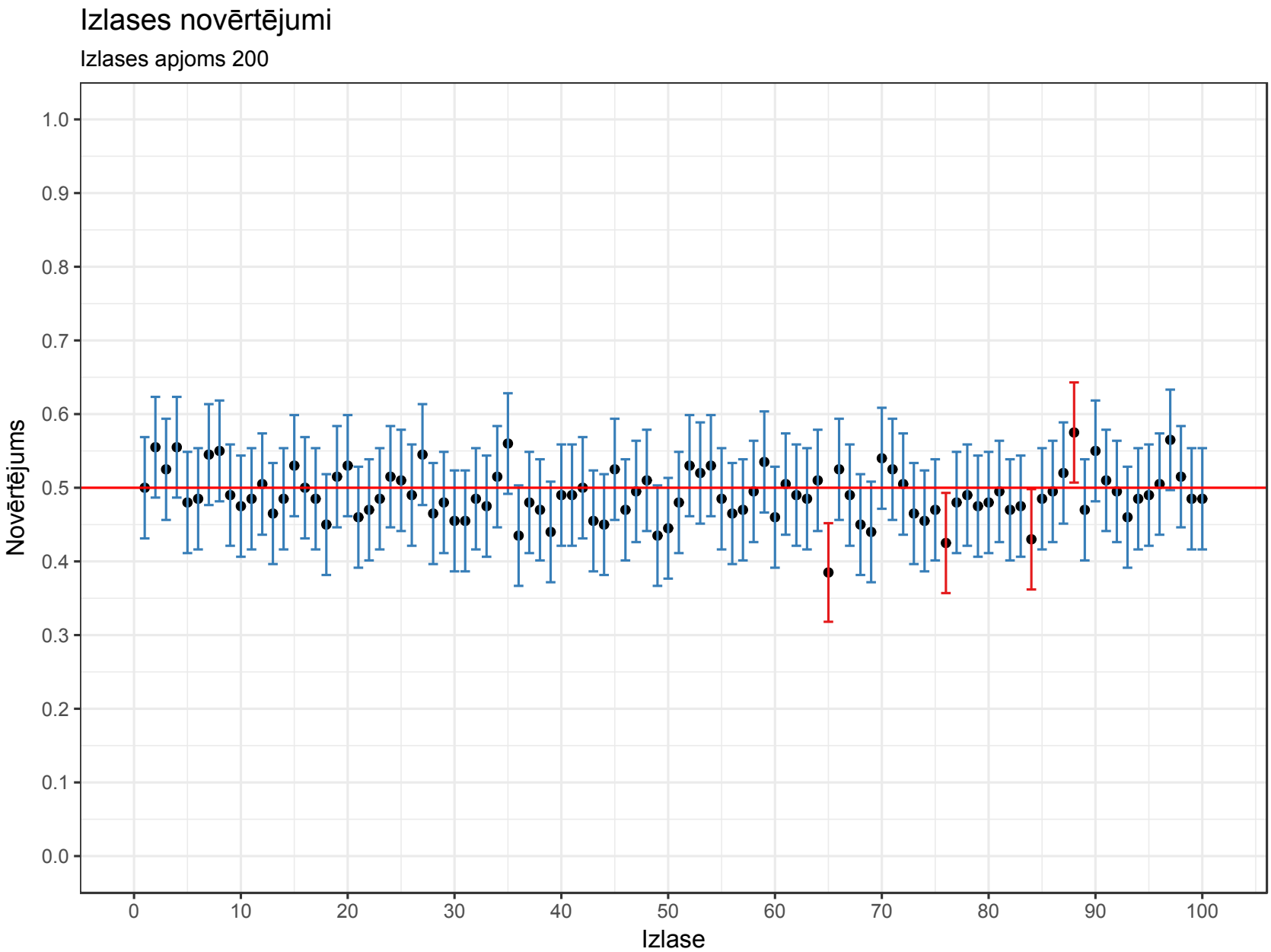
$$\text{Prob}\left[\left|\hat{P}-P\right| > \underset{0,95}{\text{MoE}}\left(\hat{P}\right)\right] = 0{,}05$$

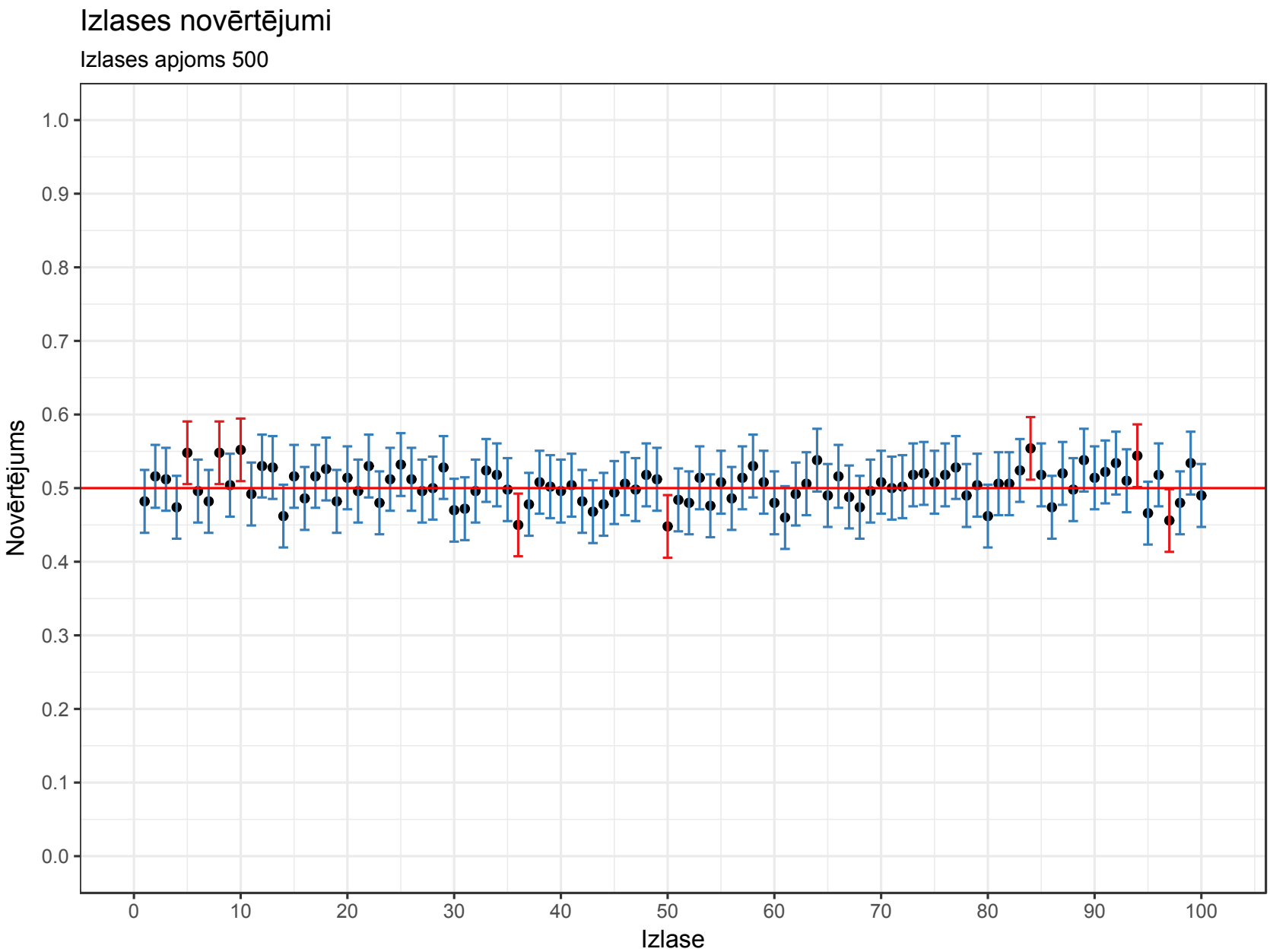
- Ticamības intervāls

$$\underset{1-\alpha}{\text{CI}}\left(\hat{P}\right) = \hat{P} \pm \underset{1-\alpha}{\text{MoE}}\left(\hat{P}\right)$$

Ticamības intervāls ar varbūtību $1-\alpha$ ietver populācijas parametru (šajā gadījumā P)


Izlases novērtējumi
Izlases apjoms 100
Novērtējums
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
10
20
30
40
50
60
70
80
90
100
Izlase


Izlases novērtējumi
Izlases apjoms 200
Novērtējums
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
10
20
30
40
50
60
70
80
90
100
Izlase


Izlases novērtējumi
Izlases apjoms 500
Novērtējums
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
10
20
30
40
50
60
70
80
90
100
Izlase

Secinājumi

- Izlases kļūdu $\hat{P} - P$ precīzi nevar aprēķināt
- Var aprēķināt (novērtēt) **kļūdas robežu** $\text{MoE}_{1-\alpha}\left(\hat{P}\right)$ ar ticamības līmeni $1 - \alpha$
- Izlases kļūdas absolūtā vērtība ar varbūtību $1 - \alpha$ nebūs lielāka par kļūdas robežu

Saturs

Piemērs #2

- Populācijas apjoms $N = 10\,000$
- Binārs izpētes mainīgais ar vērtībām 0 vai 1

$$y_i = \begin{cases} 0 \\ 1 \end{cases}$$

- Populācijas parametrs: **proporcija**

$$P = \frac{1}{N} \sum_U y_i = 0{,}5$$

- Izlases apjomi $n = \{100; 200; 500\}$
- **Vienkāršā gadījuma izlase**

Prakse

- Sarežģītāki izlases plānojumi (*sampling design*)
- Sarežģitāki populācijas parametri (nelineāras statistikas)
- Novērtējumi populācijas domēniem
- Neatbildētība
- Svaru kalibrācija

vardpoor

- Ideja no Osier (2012) raksta (prezentēts Net-SILC2 seminārā)
- `vardpoor` (Breidaks, Liberts, & Ivanova, 2016) ir R programmas (R Core Team, 2016) bibliotēka
- `https://cran.r-project.org/package=vardpoor`
- `https://github.com/CSBLatvia/vardpoor`

Izlases plānojums

- Dispersija tiek novērtēta ar galīgo klāsteru metodi (*Ultimate Cluster Method*) (Hansen, Hurwitz, & Madow, 1953)
- Nepieciešamā informācija no izlases plānojuma:
  - Stratifikācijas mainīgais
  - Primāro izlases vienību mainīgais
  - Svari
  - Primāro izlases vienību skaits populācijā sadalījumā pa stratām

Populācijas parametri

- Populācijas parametrs kā funkcija $\theta = f\left(T_1, T_2, \ldots\right)$
- Ja f ir nelineāra funkcija attiecībā pret $T_1, T_2, \ldots$
- Piemēram, divu summāro attiecība $\theta = \frac{Y}{X}$
- Divas pieejas (Hulliger et al., 2011):
  - Linearizācija
  - *Resampling* metodes (*Jackknife*, *Bootstrap*)
- `vardpoor` tiek izmantota linearizācija

Populācijas domēni

- Populācijas domēns ir patvaļiga populācijas elementu apakškopa
- Katram domēnam – papildus mainīgais (Särndal, Swensson, & Wretman, 1992)

$$y_{di} = \begin{cases} y_i & \text{, ja} \quad i \in U_d \\ 0 & \text{, ja} \quad i \notin U_d \end{cases}$$

- Nepieciešamā informācija no izlases plānojuma:
  - Mainīgais, kas sadala izlases elementus pa domēniem

Neatbildētība

- Ir dažādi neatbildētības modeļi:
  - pilnībā nejauša iztrūkšana (*missing completely at random*)
  - nejauša iztrūkšana (*missing at random*)
  - nenejauša iztrūkšana (not missing at random)
- `vardpoor` tiek pieņemts nejaušas iztrūkšanas modelis atkarībā no stratas
- Nepieciešamā informācija:
  - Respondējušās izlases vienības
  - Stratifikācijas mainīgais

Svaru kalibrācija

- Svaru kalibrācija (Särndal & Lundström, 2005) izmanto palīginformāciju (*auxiliary information*), lai
  - uzlabotu novērtējumu precizitāti
  - nodrošinātu saskaņotību
- Lineārās regresijas modelis $y_i \sim \boldsymbol{x}_i$
- Regresijas atlikumu novērtēšana
- Regresijas atlikumi tiek izmantoti dispersijas novērtēšanai
- Nepieciešamā informācija:
  - Palīginformācijas matrica
  - Svari pirms un pēc kalibrācijas
  - Mainīgais ar kuru var modelēt y_i dispersiju

Saturs

Funkcijas $Domain$ pirmais piemērs

```
library(vardpoor)
Y ← as.matrix(1:10)
colnames(Y) ← "Y1"
D ← as.matrix(rep(1, 10))
colnames(D) ← "D1"
domain(Y, D)
```

```
Y1__D1.1
       1
       2
       3
       4
       5
       6
       7
       8
       9
      10
```

Funkcijas $Domain$ otrais piemērs

```
library(vardpoor)
Y ← matrix(1:20, 10, 2)
colnames(Y) ← paste("Y", 1:2, sep="")
D ← matrix(rep(1:2, each = 5), 10, 1)
colnames(D) ← "D"
domain(Y, D)
```

Y1__D1.1	Y1__D1.2	Y2__D1.1	Y2__D1.2
1	0	11	0
2	0	12	0
3	0	13	0
4	0	14	0
5	0	15	0
0	6	0	16
0	7	0	17
0	8	0	18
0	9	0	19
0	10	0	20

Funkcijas $lin.ratio$ piemērs

```
library(vardpoor)
Y ← data.table(rchisq(10, 3))
Z ← data.table(rchisq(10, 3))
weights ← rep(2, 10)
data.table(Y=Y, Z=Z,
lin=lin.ratio(Y, Z, weights))
```

Y	Z	lin
3.6432589	2.4821452	0.0326281022
2.5108335	3.3109947	0.0097398454
2.5108335	3.3109947	0.0097398454
0.7405179	0.7047231	0.0050372698
2.0297970	6.7144124	-0.0242858630
0.9128694	3.3111987	-0.0132438881
1.1209361	4.4016804	-0.0189368667
2.0698536	2.9931873	0.0059288776
5.6912921	2.4708737	0.0621729309
0.1870044	0.4260161	-0.0007035935
0.3465608	7.9501632	-0.0583368147

Funkcijas $residual\_est$ piemērs

```
library(vardpoor)
Y ← data.table(Y=rchisq(10, 3))
X ← matrix(rchisq(20, 3), 10, 2)
w ← rep(2, 10)
q ← rep(1, 10)
data.table(Y=Y, X=X,
res=residual_est(Y, X, w, q))
```

Y	X.V1	X.V2	res
1.2926472	3.336664	3.4861403	0.13738231
1.2025586	4.434408	3.1212594	-0.01204974
2.9325728	1.212557	7.0899417	1.27736769
2.2783905	5.149664	2.7034016	1.06332727
1.0264867	12.280625	6.1377484	-1.80553595
1.3319923	5.342985	0.4191577	0.57745338
7.2254940	1.856086	0.3305869	6.92413344
0.2293984	2.367635	4.9886706	-1.12391325
1.0413855	1.827882	2.2901352	0.32781560
1.1539753	1.721365	2.0512345	0.50435842

Funkcijas $variance\_est$ piemērs

```
library(vardpoor)
n ← 10
Ys ← rchisq(10, 3)
Ys
10.301873 2.385022 1.121394 6.219925 2.599457
2.324564 5.696204 6.631298 1.819270 6.661898
w ← rep(2, n)
PSU ← 1:n
H ← rep("Strata_1", n)
variance_est(Y=Ys, H=H, PSU=PSU, w_final=w)
Y
174.7292
```

Funkcijas var_srs piemērs

```
library(vardpoor)
Ys ← rchisq(10, 3)
w ← c(rep(2, 5), rep(3, 5))
Ys
2.1122151 1.7542117 1.1983224 6.1953243 0.2399501
1.8848531 4.4761515 3.0871110 4.5247349 5.4728425
PSU ← 1:n
H ← rep("Strata_1", n)
var_srs(Ys, ws)
Y
174.729
```

Funkcijas $vardom$ piemērs

```
library(vardpoor)
data(eusilc)
dataset ← data.frame(1:nrow(eusilc), eusilc)
colnames(dataset)[1] ← "Idd"
aa ← vardom(Y="eqIncome", id="Idd", H="db040",
PSU="db030", w_final="rb050", Dom = "db040", period =
NULL, N_h=NULL, Z = NULL, X = NULL, g = NULL,
dataset = dataset)
```

Funkcijas $vardom$ piemērs

```
> aa
$lin_out
NULL

$res_out
NULL

$all_result
   variable         db040 respondent_count n_nonzero pop_size       estim          var        se        rse       cv absolute_margin_of_error
1: eqIncome    Burgenland              549       549   260564  5537191902 8.976677e+16 299611036 0.05410884 5.410884                 587226841
2: eqIncome     Carinthia             1078      1078   563648 11051269480 1.370555e+17 370210100 0.03349933 3.349933                 725598464
3: eqIncome Lower Austria             2804      2804  1555709 31185109934 4.543803e+17 674077371 0.02161536 2.161536                1321167369
4: eqIncome      Salzburg              924       924   535451 10297003708 1.626315e+17 403275920 0.03916439 3.916439                 790406280
5: eqIncome        Styria             2295      2292  1167045 22263233916 2.947571e+17 542915353 0.02438619 2.438619                1064094538
6: eqIncome         Tyrol             1317      1317   701899 12977922220 1.702527e+17 412616931 0.03179376 3.179376                 808714325
7: eqIncome Upper Austria             2805      2805  1421620 29065619630 4.071691e+17 638098001 0.02195370 2.195370                1250649101
8: eqIncome        Vienna             2322      2322  1598931 32725907650 4.324358e+17 657598545 0.02009413 2.009413                1288869465
9: eqIncome    Vorarlberg              733       733   377355  7647739630 1.040589e+17 322581558 0.04217999 4.217999                 632248237
   relative_margin_of_error    CI_lower    CI_upper   var_srs_HT   var_cur_HT   var_srs_ca  deff_sam deff_est      deff     n_eff
1:                10.605138  4949965061  6124418743 9.440463e+16 8.976677e+16 9.440463e+16 0.9508726        1 0.9508726  577.3644
2:                 6.565748 10325671017 11776867944 1.412077e+17 1.370555e+17 1.412077e+17 0.9705949        1 0.9705949 1110.6590
3:                 4.236533 29863942565 32506277304 3.632952e+17 4.543803e+17 3.632952e+17 1.2507194        1 1.2507194 2241.9098
4:                 7.676080  9506597429 11087409988 1.268048e+17 1.626315e+17 1.268048e+17 1.2825340        1 1.2825340  720.4487
5:                 4.779605 21199139378 23327328454 2.518172e+17 2.947571e+17 2.518172e+17 1.1705201        1 1.1705201 1960.6669
6:                 6.231462 12169207895 13786636545 1.593310e+17 1.702527e+17 1.593310e+17 1.0685475        1 1.0685475 1232.5142
7:                 4.302847 27814970529 30316268731 3.585008e+17 4.071691e+17 3.585008e+17 1.1357550        1 1.1357550 2469.7228
8:                 3.938377 31437038185 34014777116 4.071656e+17 4.324358e+17 4.071656e+17 1.0620638        1 1.0620638 2186.3094
9:                 8.267126  7015491394  8279987867 1.084978e+17 1.040589e+17 1.084978e+17 0.9590874        1 0.9590874  764.2682

>
```

Paldies par uzmanību!

- `https://cran.r-project.org/package=vardpoor`
- `https://github.com/CSBLatvia/vardpoor`

Literatūras saraksts I

Breidaks, J., Liberts, M., & Ivanova, S. (2016). vardpoor: Estimation of indicators on social exclusion and poverty and its linearization, variance estimation [Computer software]. Riga, Latvia. (R package version 0.7.6)

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

Hulliger, B., Alfons, A., Bruch, C., Filzmoser, P., Graf, M., Kolb, J.-P., … Zins, S. (2011). *Report on the simulation results* (Research Project Report Nos. WP7 – D7.1). FP7-SSH-2007-217322 AMELI. Retrieved from `http://ameli.surveystatistics.net`

Literatūras saraksts II

Osier, G. (2012). *The linearisation approach implemented by Eurostat for the first wave of EU-SILC: what could be done from second wave onwards?* (Tech. Rep.). Luxembourg: Institut National de la Statistique et des Etudes Economiques (STATEC Luxembourg).

R Core Team. (2016). R: A language and environment for statistical computing [Computer software]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Särndal, C.-E., & Lundström, S. (2005). *Estimation in surveys with nonresponse*. West Sussex, England: John Wiley & Sons.

Särndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. New-York: Springer.