Master thesis

Enhancements to the REVIVALS package

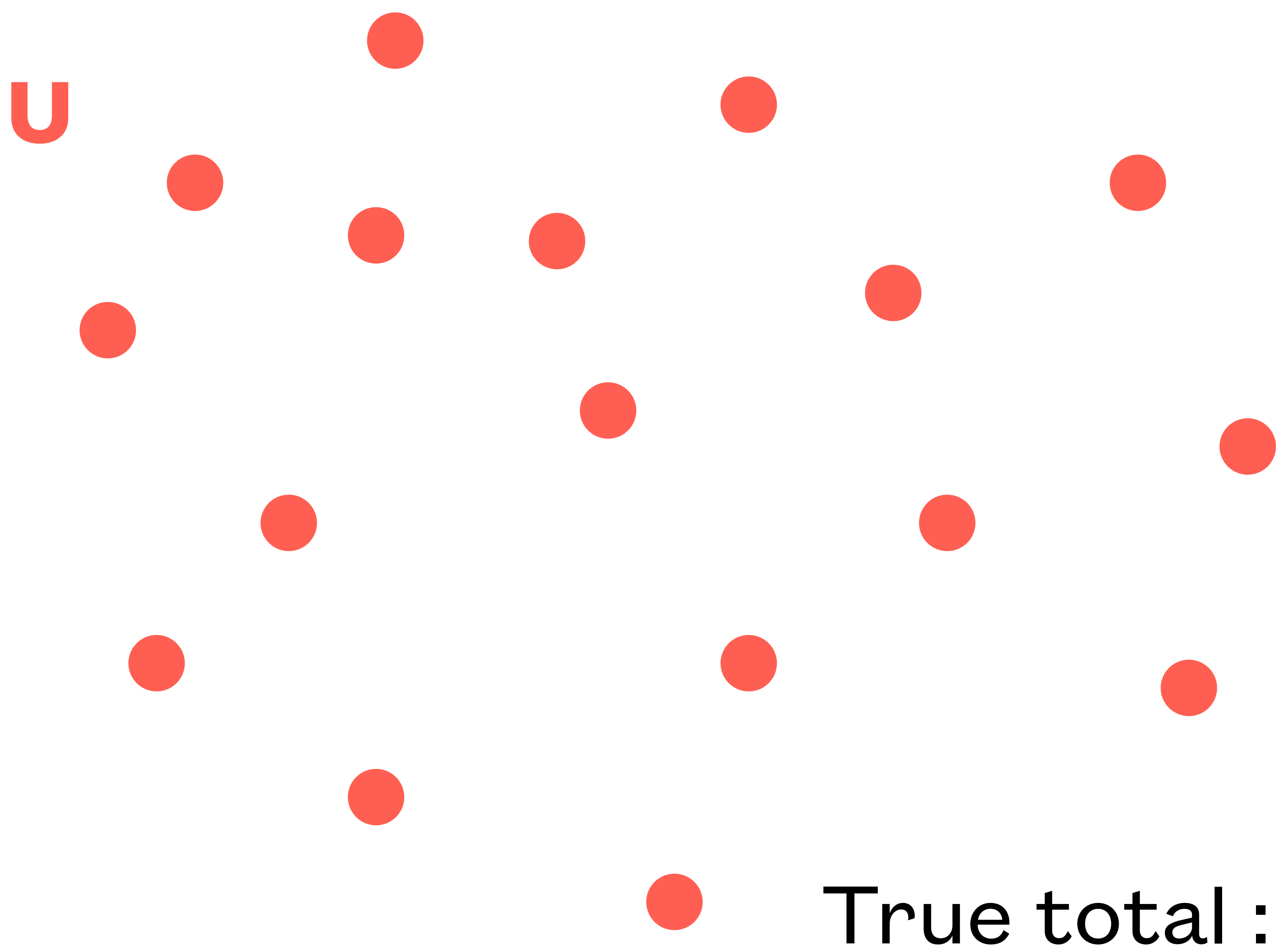
Guillem FORTÓ

SEPTEMBER 2020

Introduction

Aim

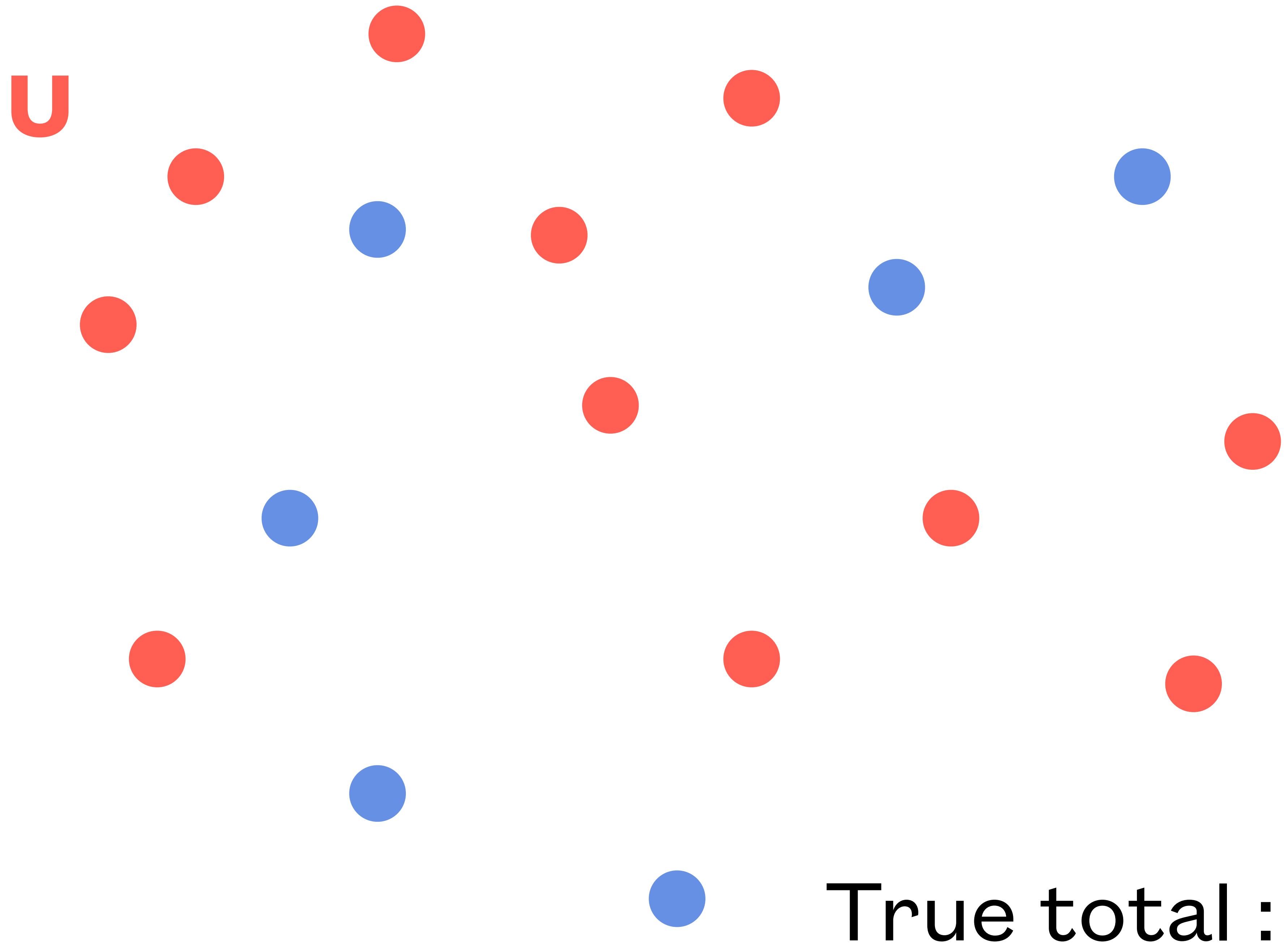
1. We are trying to estimate a total

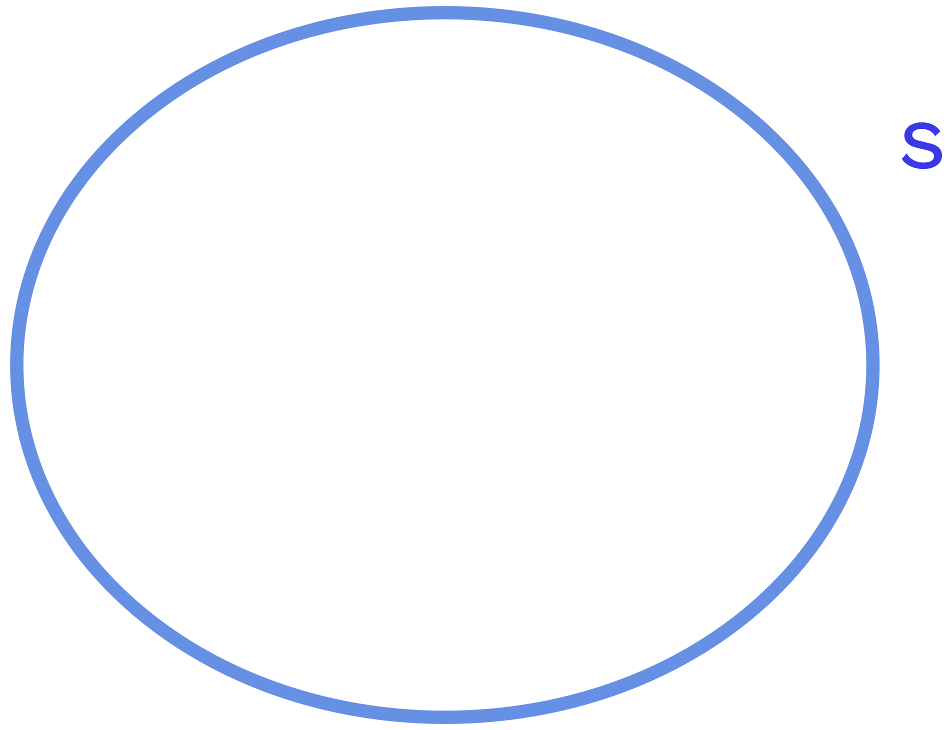


True total : $\theta = \sum_{i \in U} y_i$

Aim

1. We are trying to estimate a total





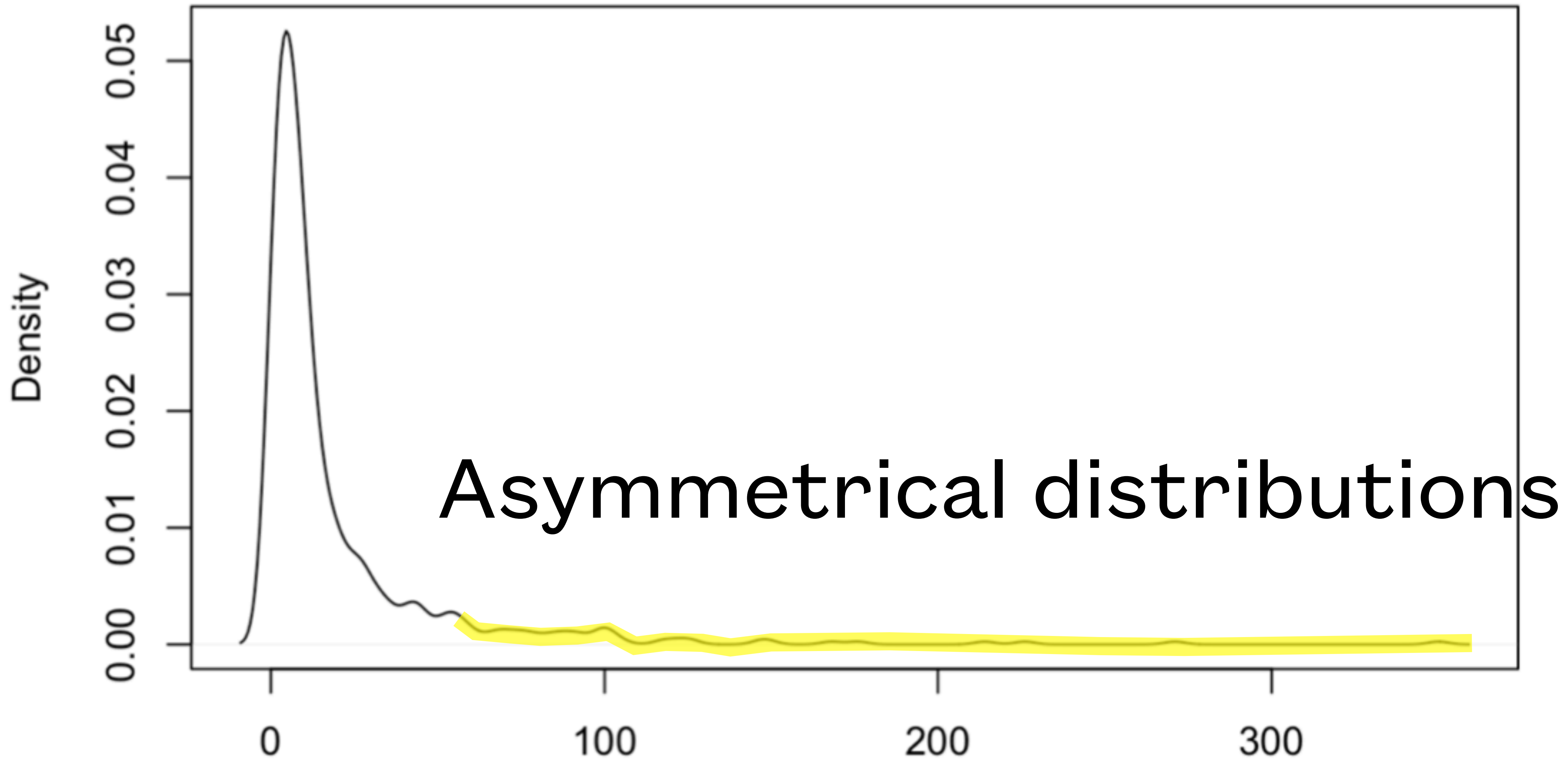
True total : $\theta = \sum_{i \in U} y_i$

Horvitz-Thompson? $\hat{\theta} = \sum_{i \in s} d_i y_i$

Aim

1. We are trying to estimate a total

2. Issues



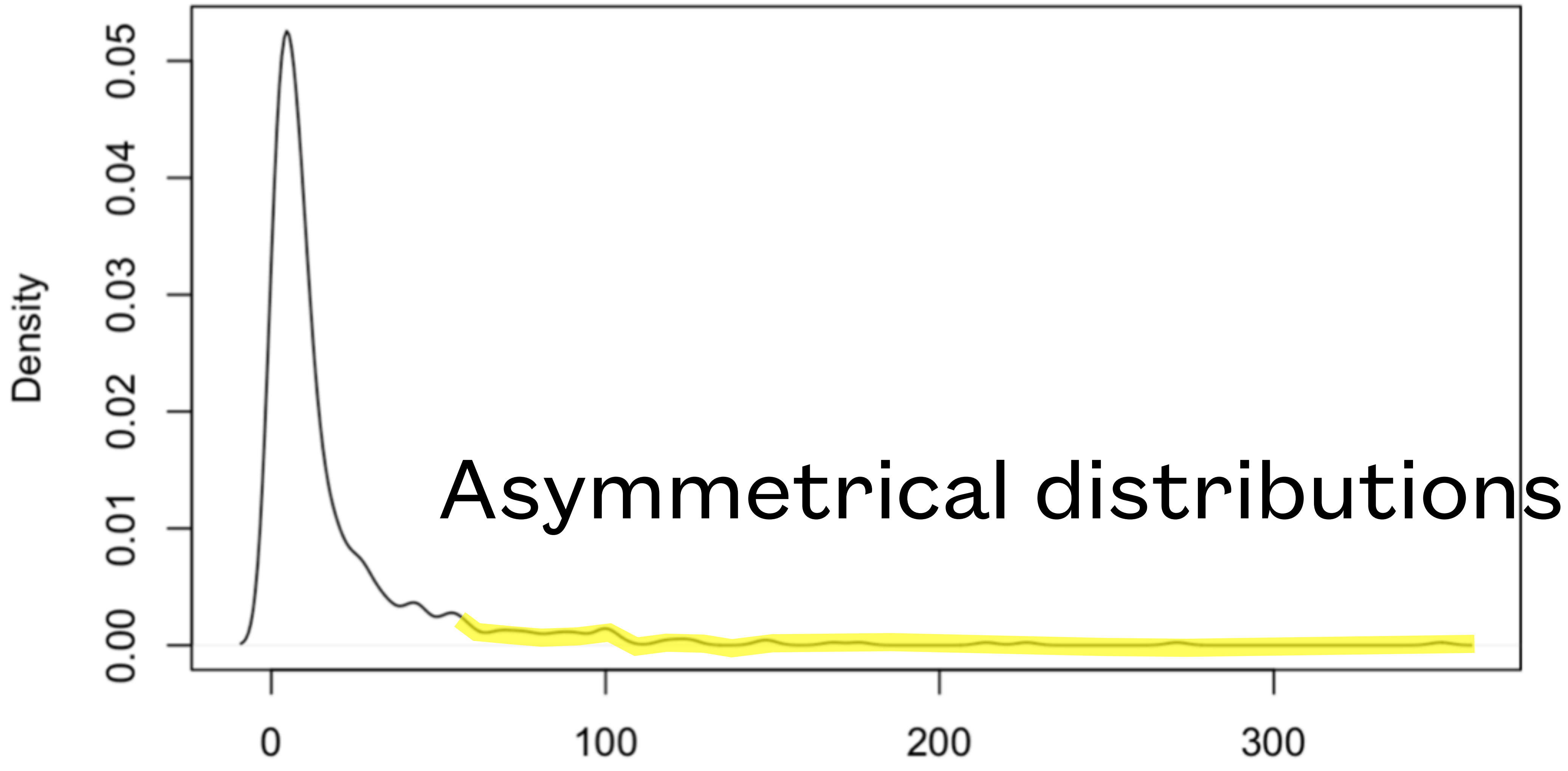
+

Editing stage has already been done

Aim

1. We are trying to estimate a total

2. Issues



Asymmetrical distributions

\+ Editing stage has already been done

3. We need robust estimators

Work done

- Beaumont et. al. (2013) proposed to use the conditional bias (CB) as a measure of influence

→ Robust version of the Horvitz-Thompson estimator

- C. Favre-Martinoz started to build a package (`revivals`)

GENERAL GOAL

Enhancing the `revivals` package

Work done

1.

`wrapper`

1. Conditional Bias estimation
2. Robust HT estimation
3. Associated tuning constant
4. Robust estimator under winsorised form

2. Robustness enhancements

3. Article

SUMMARY

- Introduction
- Dissecting the revivals package
- Application with wrapper function

Literature review
Sampling designs

- Finite population of N units: $U = \{1, \ldots, N\}$

Literature review
Sampling designs

- Finite population of N units: $U = \{1, \ldots, N\}$
- Variable of interest: y_i , $\forall i \in U$

Literature review
Sampling designs

- Finite population of N units: $U = \{1, \ldots, N\}$
- Variable of interest: y_i , $\forall i \in U$
- Sample: $s = \left(I_1, \ldots, I_i, \ldots, I_N\right)^T$ where $I_i = \begin{cases} 1 & \text{if } i \in s \\ 0 & \text{otherwise.} \end{cases}$

Literature review
Sampling designs

- Finite population of N units: $U = \{1, \ldots, N\}$
- Variable of interest: y_i , $\forall i \in U$
- Sample: $s = (I_1, \ldots, I_i, \ldots, I_N)^T$ where $I_i = \begin{cases} 1 & \text{if } i \in s \\ 0 & \text{otherwise.} \end{cases}$
- 1st-order inclusion probabilities: $\pi_i = \mathbb{P}(i \in s)$

Literature review
Sampling designs

- Finite population of N units: $U = \{1, \ldots, N\}$
- Variable of interest: y_i , $\forall i \in U$
- Sample: $s = \left(I_1, \ldots, I_i, \ldots, I_N\right)^T$ where $I_i = \begin{cases} 1 & \text{if } i \in s \\ 0 & \text{otherwise.} \end{cases}$
- 1st-order inclusion probabilities: $\pi_i = \mathbb{P}(i \in s)$
- Design weight: $d_i = \dfrac{1}{\pi_i}$

Literature review
Sampling designs

Design	Description	π_i
`si`	We draw a sample of size n with equiprobability among all possible samples.	$\frac{n}{N}$
`poisson`	We fix $\pi = (\pi_1, \pi_2, ..., \pi_N)$ and randomly draw each observation independently. The sample size is random	(fixed)
`rejective`	We fix $\pi = (\pi_1, \pi_2, ..., \pi_N)$ and follow a Poisson design, rejecting all samples until we obtain one of size n.	(fixed)

Table 1: Non-stratified sampling designs used in `REVIVALS`

Literature review
Conditional bias

- For a parameter θ and an estimator $\hat{\theta}$ (Moreno-Rebollo et al. - 1999):

$$B_i^{\hat{\theta}}(I_i = 1) = \mathbb{E}(\hat{\theta} - \theta \mid I_i = 1),$$

- In the case of the HT estimator of a total, it becomes:

$$B_i^{HT}(I_i = 1) = \sum_{j \in U} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) y_j$$

Notation: $B_i^{HT}(I_i = 1)$ will simply be denoted B_i^{HT}

Literature review
Conditional bias

Revivals function

`HTcondbiasest`

SI

- Theoretical $$B_i^{HT} = \frac{N}{N-1}\left(\frac{N}{n} - 1\right)(y_i - \bar{Y}_U),\ \forall i \in U,$$ where $\bar{Y}_U = \frac{1}{N}\sum_{i \in U} y_i$

- Estimated $$\hat{B}_i^{HT} = \frac{n}{n-1}\left(\frac{N}{n} - 1\right)(y_i - \bar{y}),\ \forall i \in U,$$ where $\bar{y} = \frac{1}{n}\sum_{i \in S} y_i$

Conditions to check : $n > 1$ / N is known

Literature review
Conditional bias

Revivals function
HTcondbiasest

POISSON

- Theoretical $$B_i^{HT} = \left(\frac{1}{\pi_i} - 1 \right) y_i$$
- No need to estimate it

Literature review
Robust estimator

Revivals function
robustest

How to use the conditional bias in order to obtain robust estimators in a design-based framework?

=> Robust Horvitz-Thompson (**RHT**) estimator (Beaumont et al. - 2013) :

$$\hat{t}_y^{RHT} = \hat{t}_y^{HT} - \frac{1}{2}\left(\hat{B}_{min}^{HT} + \hat{B}_{max}^{HT}\right)$$

Literature review
Associated constant c

Revivals function
`tuningconst`

General form

$$\hat{t}_y^{RHT} = \hat{t}_y^{HT} - \sum_{i \in S} \hat{B}_i^{HT}(I_i = 1) + \sum_{i \in S} \psi(\hat{B}_i^{HT}(I_i = 1))$$

where ψ is the Huber function, defined as: $\psi(x) = \text{sign}(x) \times \min(|x|, c)$ where $c > 0$.

Literature review
Associated constant c

Revivals function
`tuningconst`

General form

$$\hat{t}_y^{RHT} = \hat{t}_y^{HT} - \sum_{i \in S} \hat{B}_i^{HT}(I_i = 1) + \sum_{i \in S} \psi(\hat{B}_i^{HT}(I_i = 1))$$

where ψ is the Huber function, defined as: $\psi(x) = \text{sign}(x) \times \min(|x|, c)$ where $c > 0$.

Optimal constant c ?

$$\min_c \max\{|\hat{B}_i^{RHT}(c)| \; i \in s\}$$

Literature review
Winsorised form

Revivals function
`robustweights`

Idea: if $d_i y_i > K$ then $y_i \to \tilde{y}_i$. $\to \hat{t} = \sum_{i \in s} d_i \tilde{y}_i$

Literature review
Winsorised form

Revivals function
`robustweights`

Idea: if $d_i y_i > K$ then $y_i \to \tilde{y}_i$. $\to \hat{t} = \sum_{i \in s} d_i \tilde{y}_i \Leftrightarrow \hat{t} = \sum_{i \in s} \tilde{d}_i y_i$

Literature review
Winsorised form

Revivals function
robustweights

Idea: if $d_i y_i > K$ then $y_i \to \tilde{y}_i$. $\to \hat{t} = \sum_{i \in s} d_i \tilde{y}_i \Leftrightarrow \hat{t} = \sum_{i \in s} \tilde{d}_i y_i$

2 winsorisation forms

- Standard $\tilde{d}_i^s = d_i \dfrac{min\left(y_i, \frac{K}{d_i}\right)}{y_i}$
- Dalén-Tambay $\tilde{d}_i^{DT} = 1 + (d_i - 1)\dfrac{min\left(y_i, \frac{K}{d_i}\right)}{y_i}$

Literature review
Winsorisation constant

Revivals function
determinconstws
determinconstwDT

Optimal constant K_{opt} ?

- We want: $\hat{t}_y^{BHR} = \hat{t}_y^{std} = \hat{t}_y^{DT}$
- We know that: $\hat{t}_y^{RHT} = \hat{t}_y^{HT} - \frac{1}{2}\left(\hat{B}^{HT}_{min} + \hat{B}^{HT}_{max}\right)$
- Same form: $\hat{t}_y^{HT} + \Delta(K)$

$$\min_K \ \max\{\,|\,\hat{B}_i^{RHT}(K)\,|\ i \in s\}$$

$$\Leftrightarrow \Delta(K) = -\frac{1}{2}\left(\hat{B}^{HT}_{min} + \hat{B}^{HT}_{max}\right)$$

$$\Leftrightarrow \sum_{j\in s} a_j \max\left(0,\ d_j y_j - K\right) = \frac{\hat{B}^{HT}_{min} + \hat{B}^{HT}_{max}}{2}$$

$a_j = 1$ in the case of the std form, and $a_j = (d_j - 1)/d_j$ in the case of the DT form

Revivals
Structure

1. Conditional Bias estimation
2. Robust HT estimation
3. Associated tuning / winsorisation constants
4. Robust weights computation

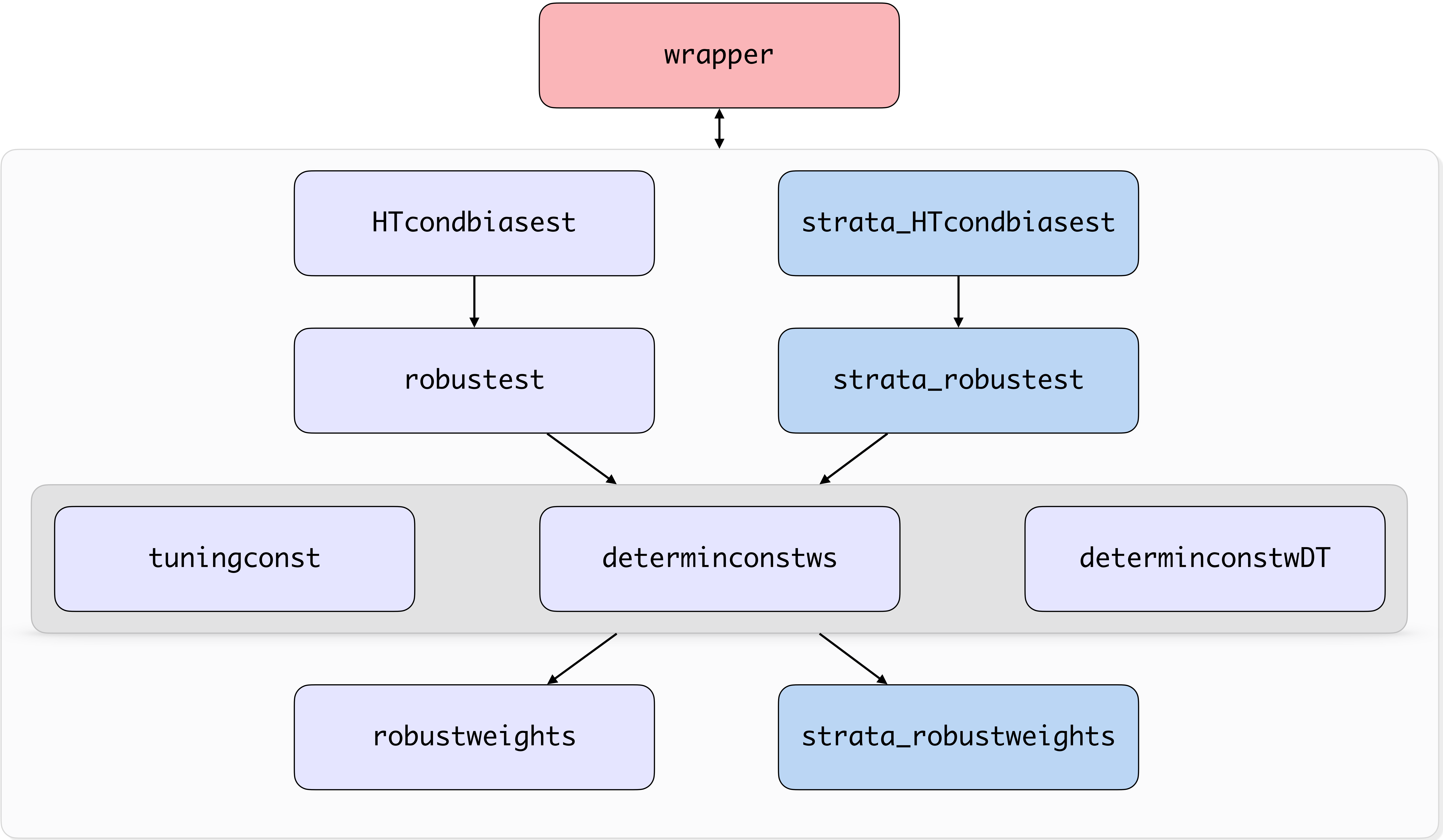


Illustration with the `wrapper` function

Illustration
Data presentation

rec99htegne database:

- $N = 554$ observations
- $p = 7$ variables regarding the French communes in the Haute-Garonne department

	CODE_N	COMMUNE	BVQ_N	POPSDC99	LOG	LOGVAC	STRATLOG
1	31014	ARGUENOS	31020	57	94	1	1
2	31131	CAZAUNOUS	31020	47	56	4	1
3	31348	MONCAUP	31020	26	57	2	1
4	31447	RAZECUEILLE	31020	37	89	6	1
5	31140	CHEIN-DESSUS	31020	184	174	28	2

Illustration
Data presentation

rec99htegne database:

- $N = 554$ observations
- $p = 7$ variables regarding the French communes in the Haute-Garonne department

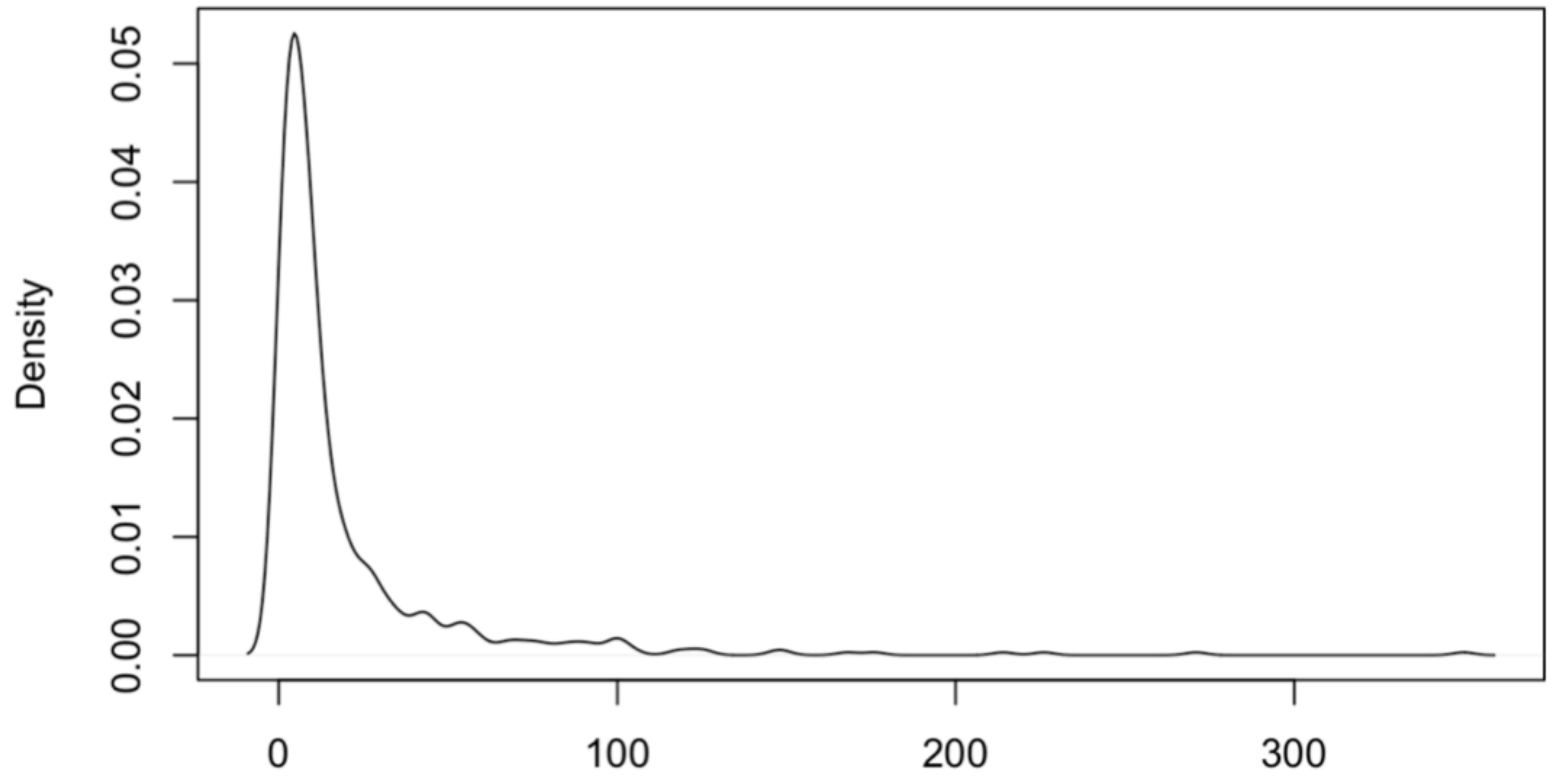
	CODE_N	COMMUNE	BVQ_N	POPSDC99	LOG	LOGVAC	STRATLOG
1	31014	ARGUENOS	31020	57	94	1	1
2	31131	CAZAUNOUS	31020	47	56	4	1
3	31348	MONCAUP	31020	26	57	2	1
4	31447	RAZECUEILLE	31020	37	89	6	1
5	31140	CHEIN-DESSUS	31020	184	174	28	2

Illustration
Data presentation

- Basic descriptive statistics for LOGVAC

Min	Q1	Median	Mean	Q3	Max	$\mathbb{V}ar$	Kurtosis	Skewness
0.00	4.00	8.00	19.44	20.00	350.00	1 104.50	29.450	4.533

- LOGVAC density plot



any drawn sample will potentially contain some very influential values

Illustration
Sampling designs

- Sample size: $n = 80$ units

Illustration
Sampling designs

- Sample size: $n = 80$ units

- First order incl. probabilities

  - SRSWOR: $$\pi_i = \frac{80}{554} \simeq 0.144$$

Illustration
Sampling designs

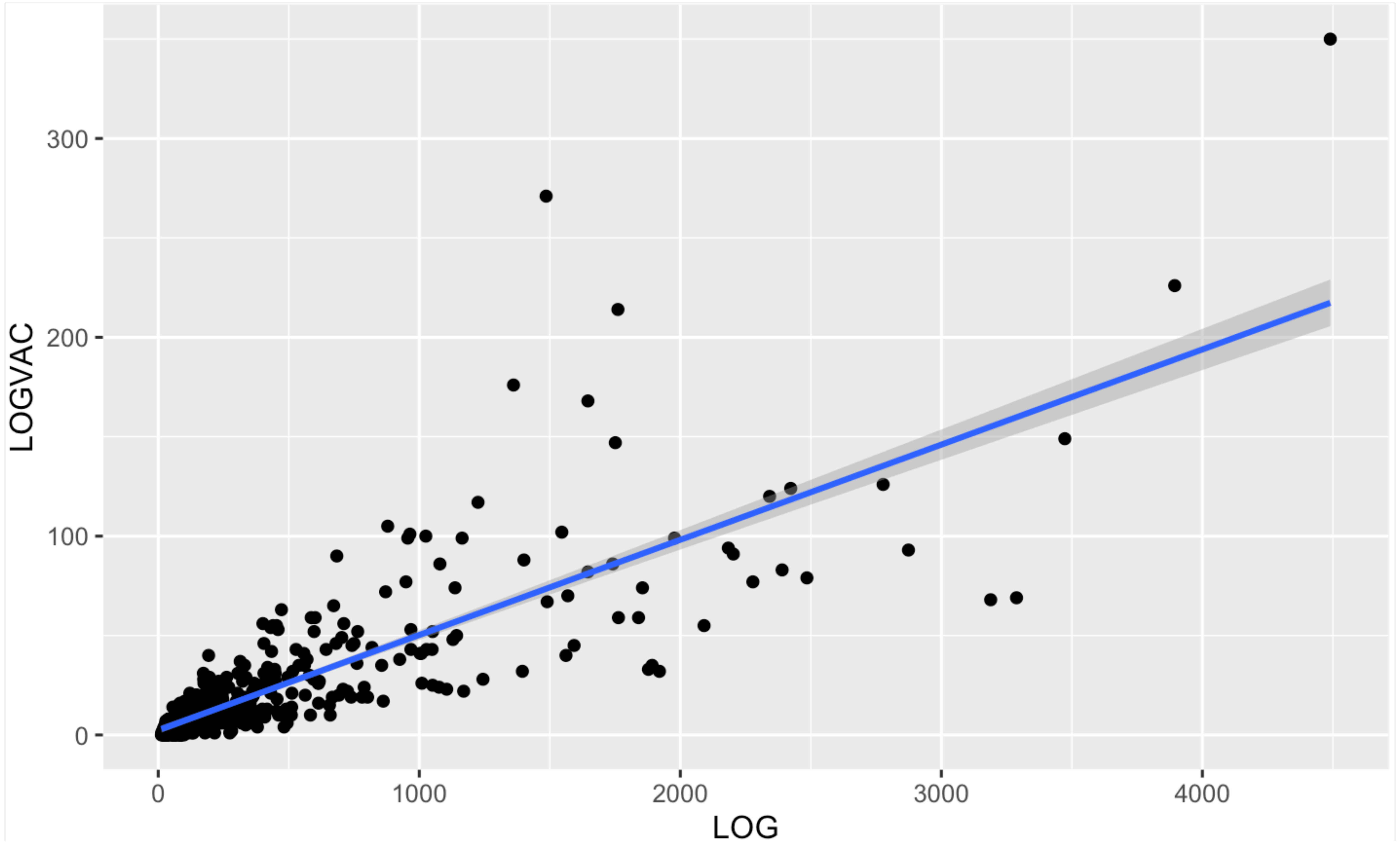
- Sample size: $n = 80$ units

- First order incl. probabilities

  - SRSWOR: $\pi_i = \dfrac{80}{554} \simeq 0.144$

  - Poisson and Rejective: use an auxiliary variable LOG

Illustration
Sampling designs



- Linear relationship

Confirms the validity of LOG as an auxiliary variable.

- $\rho(LOG, LOGVAC) \simeq 0.8189$

Illustration
Sampling designs

- Sample size: $n = 80$ units
- First order incl. probabilities

  - SRSWOR: $\pi_i = \dfrac{80}{554} \simeq 0.144$

  - Poisson and Rejective: $\pi_i = \dfrac{log_i \times n}{LOG} = \dfrac{log_i \times 80}{197\,314}$

Illustration
Syntax

```
wrapper(data = ech,
        varname = c("LOGVAC"),
        gn = N,
        est_type = c("BHR", "standard", "DT"),
        method = "si",
        pii = ech$piks,
        id = "CODE_N")
```

Illustration
Syntax

```
wrapper(data = ech,
        varname = c("LOGVAC"),
        gn = N,
        est_type = c("BHR", "standard", "DT"),
        method = "si",
        pii = ech$piks,
        id = "CODE_N")
```

Illustration
Syntax

```
wrapper(data = ech,
        varname = c("LOGVAC"),
        gn = N,
        est_type = c("BHR", "standard", "DT"),
        method = "si",
        pii = ech$piks,
        id = "CODE_N")
```

Illustration
Syntax

```
wrapper(data = ech,
        varname = c("LOGVAC"),
        gn = N,
        est_type = c("BHR", "standard", "DT"),
        method = "si",
        pii = ech$piks,
        id = "CODE_N")
```

Illustration
Syntax

```
wrapper(data = ech,
        varname = c("LOGVAC"),
        gn = N,
        est_type = c("BHR", "standard", "DT"),
        method = "si",
        pii = ech$piks,
        id = "CODE_N")
```

Illustration
Syntax

```
wrapper(data = ech,
        varname = c("LOGVAC"),
        gn = N,
        est_type = c("BHR", "standard", "DT"),
        method = "si",
        pii = ech$piks,
        id = "CODE_N")
```

Illustration
Syntax

```
wrapper(data = ech,
        varname = c("LOGVAC"),
        gn = N,
        est_type = c("BHR", "standard", "DT"),
        method = "si",
        pii = ech$piks,
        id = "CODE_N")
```

Illustration
Syntax

```
wrapper(data = ech,
        varname = c("LOGVAC"),
        gn = N,
        est_type = c("BHR", "standard", "DT"),
        method = "si",
        pii = ech$piks,
        id = "CODE_N")
```

Illustration
Summary tables

	est_type	var	RHT	tuning_const	HT	rel_diff	nb_modif_weights
1	BHR	LOGVAC	11 776.00	642.00	12 284.95	-4.14	1
2	standard	LOGVAC	11 776.00	973.00	12 284.95	-4.14	1
3	DT	LOGVAC	11 776.00	887.10	12 284.95	-4.14	1

Table 5: Summary table for SRSWOR

	est_type	var	RHT	tuning_const	HT	rel_diff	nb_modif_weights
1	BHR	LOGVAC	11 174.90	239.78	11 323.94	-1.32	5
2	standard	LOGVAC	11 174.90	292.60	11 323.94	-1.32	5
3	DT	LOGVAC	11 174.90	271.66	11 323.94	-1.32	6

Table 6: Summary table for Poisson sampling

Illustration
Detailed tables

CODE_N	init_weight	LOGVAC	condbias LOGVAC	new_weights LOGVAC_BHR	modifed LOGVAC_BHR	new_weights LOGVAC_standard	modifed LOGVAC_standard	new_weights LOGVAC_DT	modifed LOGVAC_DT
31342	6.92	10.00	-73.05	6.92	FALSE	6.92	FALSE	6.92	FALSE
31020	6.92	90.00	406.95	6.92	FALSE	6.92	FALSE	6.92	FALSE
31086	6.92	5.00	-103.05	6.92	FALSE	6.92	FALSE	6.92	FALSE
31002	6.92	7.00	-91.05	6.92	FALSE	6.92	FALSE	6.92	FALSE
31264	6.92	11.00	-67.05	6.92	FALSE	6.92	FALSE	6.92	FALSE

Table 7: Detailed table for SRSWOR

CODE_N	init_weight	LOGVAC	condbias LOGVAC	new_weights LOGVAC_BHR	modifed LOGVAC_BHR	new_weights LOGVAC_standard	modifed LOGVAC_standard	new_weights LOGVAC_DT	modifed LOGVAC_DT
31342	15.81	10.00	148.12	15.81	FALSE	15.81	FALSE	15.81	FALSE
31083	19.60	16.00	297.64	15.99	TRUE	18.29	TRUE	17.11	TRUE
31027	34.37	4.00	133.50	34.37	FALSE	34.37	FALSE	34.37	FALSE
31052	4.21	20.00	64.26	4.21	FALSE	4.21	FALSE	4.21	FALSE

Table 8: Detailed table for Poisson sampling

Revivals
Robustness - Warnings

- Notification that d_i is redundant and only π_i is being used ;
- For rejective sampling design, D must be large enough and N/D bounded ;
- Message when the estimation method is wrongly specified or multiple methods are specified. In both cases the method is set to ‘si’ by default.
- Warning when one of the interest variables contains negative values : it could be that the $\hat{B}^{HT}_{\min} + \hat{B}^{HT}_{\max}$ from the robust estimator equation $\hat{t}_y^{RHT} = \hat{t}_y^{HT} - \frac{1}{2}\left(\hat{B}^{HT}_{min} + \hat{B}^{HT}_{max}\right)$ is negative.

Revivals
Robustness - Stops

- We make sure that none of the interest variables contains any missing value ;
- We stop the code when π_i (or d_i) has a different number of rows than the data file ;
- We check that $\frac{1}{2}(\hat{B}_{min} + \hat{B}_{max}) \geq 0 \Leftrightarrow \hat{B}_{max} \geq -\hat{B}_{min}$. Functions *determinconsts* and *determinconstwDT* giving the winsorisation constants do not work otherwise.

Conclusion

Bibliography

- J. Dalén. Practical estimators of a population total which reduce the impact of large observations. Statistiska centralbyrån, 1987.
- J.-L. Tambay. “An Integrated Approach for the Treatment of Outliers in Sub-Annual Economic Surveys”. In: Proceedings of the Section on Survey Research Methods: American Statistical Association. 1988, pp. 229–234.
- P.N. Kokic and P.A. Bell. “Optimal winsorizing cutoffs for a stratified finite population estimator”. In: Journal of Official Statistics 10.4 (1994), p. 419.
- J.-L. Moreno-Rebollo, A. Muñoz-Reyes, and J. Muñoz-Pichardo. “Miscellanea. influence diagnostic in survey sampling: conditional bias”. In: Biometrika 86.4 (1999), pp. 923–928.
- J.-F. Beaumont, D. Haziza, and A. Ruiz-Gazen. “A unified approach to robust estimation in finite population sampling”. In: Biometrika 100.3 (2013), pp. 555–569.
- C. Favre-Martinoz. “Estimation robuste en population finie et infinie”. PhD thesis. 2015.

Bibliography

- C. Favre-Martinoz, A. Ruiz-Gazen, J.-F. Beaumont, and D. Haziza. “Robustness in survey sampling using the conditional bias approach with R implementation”. In: Convegno della Società Italiana di Statistica. Springer. 2016, pp. 3–13.
- Y. Tillé and A. Matei. “Package ‘sampling’”. In: Survey Sampling. Kasutatud 23 (2016), p. 2017. url: http://cran.r-%20project.org/src/contrib/Descriptions/sampling. html.
- J.-M. Béguin and O. Haag. “Méthodologie de la statistique annuelle d’entreprises. description du système «ésane»”. In: (2017).
- T. Deroyon and C. Favre-Martinoz. “Comparison of the conditional bias and Kokic and Bell methods for Poisson and stratified sampling”. In: Survey Methodology 44.2 (2018), pp. 309–338.