From a few to the whole crowd: a practical introduction to sampling and survey data analysis

Annibale Cois, Meng, MPH, PhD
Burden of Disease Research Unit, South African Medical Council &
Division of Epidemiology & Biostatistics,
School of Public Health, University of Cape Town

Email: annibale.cois@mrc.ac.za
ORCID: 0000-0002-7014-6510
WEB: annibalecois.github.io

Cape Town International Convention Centre, 24 September 2024

Data analysis

Survey data:

Clustering --> non-independence

Stratification --> unequal probability of selection

Why should we care?

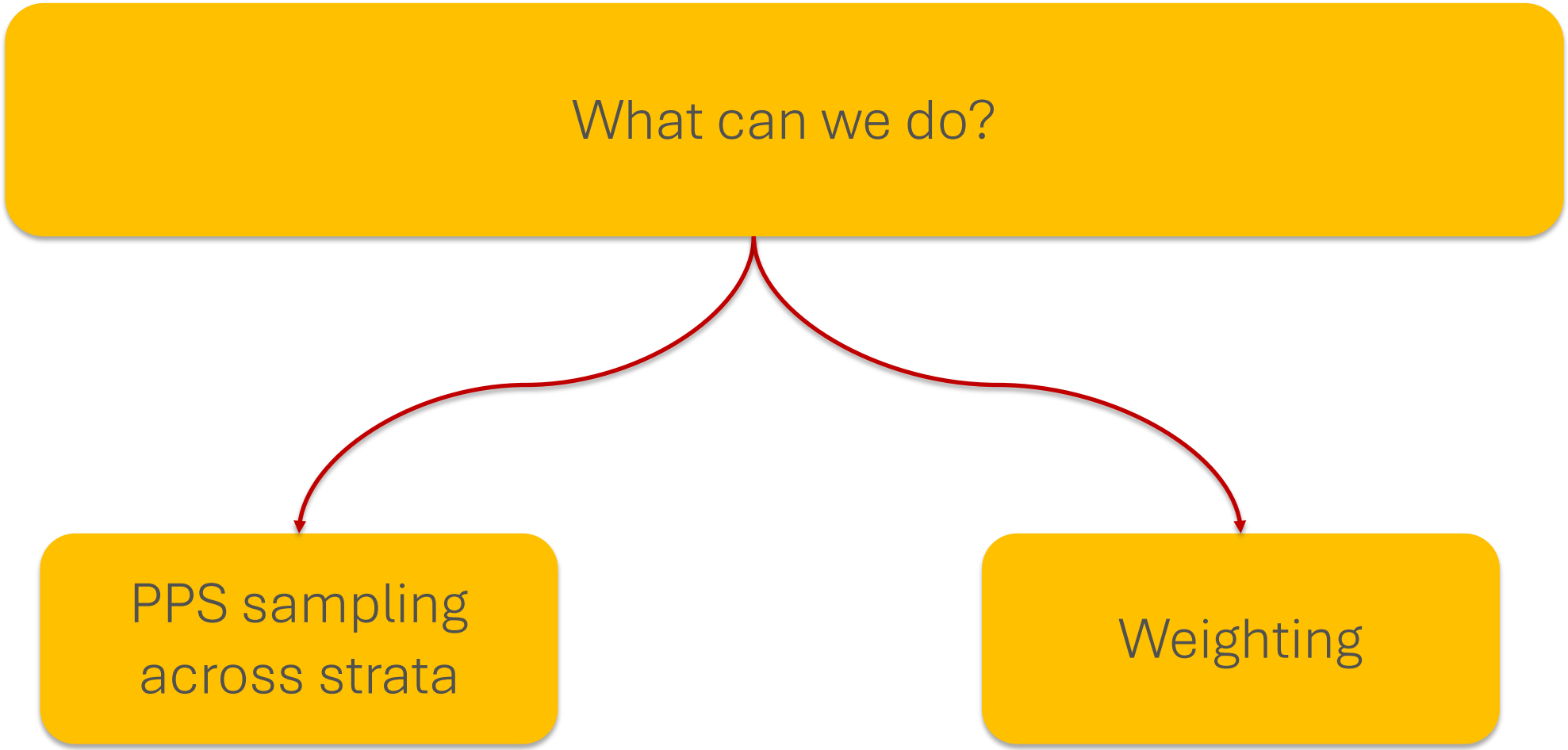
BIASED ESTIMATES

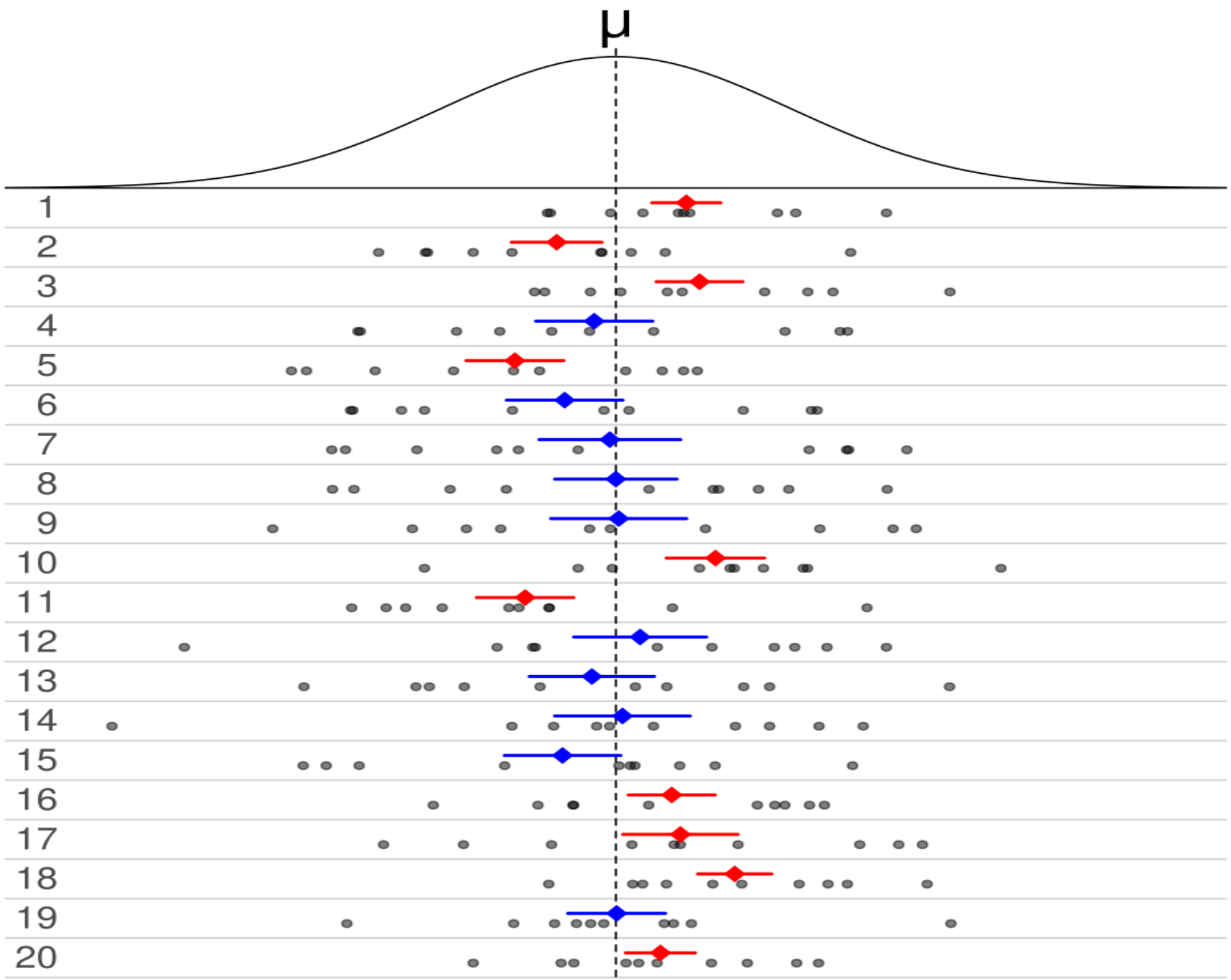
INCORRECT QUANTIFICATION OF SAMPLING ERROR

IID	HID	HR	IR	REGION	TOWN	SUBURB	SBP_1	DISEASE_D	SEX	AGE	Prob
H_1_125_6	H_1_125	Consent	Consent	Beria	Durtis	Lobammu	114	No	Male	36	0.039006
H_1_161_1	H_1_161	Consent	Consent	Beria	Durtis	Fiog	NA	No	Male	32	0.039006
H_10_100_	H_10_100	Consent	Consent	Northern I	Mmabani	Trerruk Ce	138.7	No	Male	40	0.039006
H_10_111_	H_10_111	Consent	Refusal	Northern I	Mmabani	Trerruk Ce	NA	NA	NA	NA	0.039006
H_10_123_	H_10_123	Consent	Consent	Northern I	Mmabani	Trerruk Ce	115.8	No	Male	21	0.039006
H_10_129_	H_10_129	Consent	Refusal	Northern I	Mmabani	Trerruk Ce	NA	NA	NA	NA	0.039006
H_10_138_	H_10_138	Consent	Consent	Northern I	Mmabani	Trerruk Ce	101.4	No	Female	26	0.039006
H_10_166_	H_10_166	Consent	Consent	Northern I	Mmabani	Trerruk Ce	156.3	No	Male	62	0.039006
H_10_172_	H_10_172	Absent		Northern I	Mmabani	Trerruk Ce	NA	NA	NA	NA	0.039006
H_10_177_	H_10_177	Consent	Consent	Northern I	Mmabani	Trerruk Ce	167.8	No	NA	47	0.039006
H_10_177_	H_10_177	Consent	Consent	Northern I	Mmabani	Trerruk Ce	92.6	NA	Male	22	0.039006
H_10_184_	H_10_184	Consent	Refusal	Northern I	Mmabani	Trerruk Ce	NA	NA	NA	NA	0.039006
H_10_186_	H_10_186	Refusal		Northern I	Mmabani	Yummooth	NA	NA	NA	NA	0.039006
H_10_191_	H_10_191	Consent	Consent	Northern I	Mmabani	Yummooth	234	Yes	Female	69	0.039006
H_10_232_	H_10_232	Consent	Consent	Northern I	Mmabani	Yummooth	NA	Yes	Female	78	0.039006
H_10_240_	H_10_240	Consent	Consent	Northern I	Mmabani	Yummooth	133.3	No	Female	19	0.039006
H_10_245_	H_10_245	Consent	Consent	Northern I	Mmabani	Yummooth	178.8	Yes	Female	66	0.039006
H_10_249_	H_10_249	Consent	Consent	Northern I	Mmabani	Yummooth	173.3	No	Female	NA	0.039006
H_10_252_	H_10_252	Consent	Consent	Northern I	Mmabani	Yummooth	137.8	No	Male	19	0.039006
H_10_257_	H_10_257	Consent	Consent	Northern I	Mmabani	Yummooth	57.1	No	Female	23	0.039006

IID	HID	HR	IR	REGION	TOWN	SUBURB	SBP_1	DISEASE_D	SEX	AGE	Prob_1_s	Prob_2_s	Prob_3_s	Prob
H_1_1_1	H_1_1	Consent	Consent	Beria	Durtis	Lobammu	138.3	No	NA	54	1	0.5	0.143266	0.011939
H_1_11_1	H_1_11	Consent	Consent	Beria	Durtis	Lobammu	157.4	No	Female	31	1	0.5	0.143266	0.071633
H_1_437_1	H_1_437	Consent	Consent	Beria	Durtis	Hoban	199.6	No	Female	69	1	0.5	0.143266	0.071633
H_1_448_2	H_1_448	Consent	Refusal	Beria	Durtis	Hoban	NA	NA	NA	NA	1	0.5	0.143266	0.035817
H_11_147_	H_11_147	Consent	Consent	Northern I	Thaye	Chooleb G	153.3	Yes	Female	74	1	0.5	0.255102	0.127551
H_11_148_	H_11_148	Consent	Consent	Northern I	Thaye	Chooleb G	122.4	Yes	Female	69	1	0.5	0.255102	0.063776
H_11_149_	H_11_149	Consent	Consent	Northern I	Thaye	Chooleb G	139.8	No	Male	19	1	0.5	0.255102	0.02551
H_11_153_	H_11_153	Refusal		Northern I	Thaye	Chooleb G	NA	NA	NA	NA	1	0.5	0.255102	0.127551
H_11_158_	H_11_158	Consent	Consent	Northern I	Thaye	Chooleb G	157.6	No	Female	60	1	0.5	0.255102	0.042517
H_21_119_	H_21_119	Consent	Consent	Central Re	Eenranos	Oorolong	124.6	No	Male	22	1	0.333333	0.346021	0.11534
H_21_122_	H_21_122	Consent	Consent	Central Re	Eenranos	Oorolong	NA	No	NA	23	1	0.333333	0.346021	0.028835
H_21_126_	H_21_126	Consent	Consent	Central Re	Eenranos	Oorolong	198.6	Yes	Female	58	1	0.333333	0.346021	0.05767
H_21_127_	H_21_127	Consent	Consent	Central Re	Eenranos	Oorolong	140.1	Yes	Female	45	1	0.333333	0.346021	0.05767
H_21_129_	H_21_129	Consent	Consent	Central Re	Eenranos	Oorolong	109.3	No	Male	30	1	0.333333	0.346021	0.038447
H_21_130_	H_21_130	Consent	Consent	Central Re	Eenranos	Oorolong	150.1	No	Male	31	1	0.333333	0.346021	0.11534
H_21_132_	H_21_132	Consent	Consent	Central Re	Eenranos	Oorolong	105.3	No	Male	59	1	0.333333	0.346021	0.038447
H_21_14_5	H_21_14	Consent	Consent	Central Re	Eenranos	Oorolong	100.6	Yes	Female	39	1	0.333333	0.346021	0.038447
H_21_142_	H_21_142	Consent	Consent	Central Re	Eenranos	Oorolong	141.8	Yes	Female	67	1	0.333333	0.346021	0.05767
H_21_143_	H_21_143	Consent	Consent	Central Re	Eenranos	Oorolong	105.6	Yes	Female	49	1	0.333333	0.346021	0.038447
H_21_147_	H_21_147	Consent	Consent	Central Re	Eenranos	Oorolong	111.9	NA	Female	49	1	0.333333	0.346021	0.11534
H_21_153_	H_21_153	Consent	Consent	Central Re	Eenranos	Oorolong	130.1	NA	Male	22	1	0.333333	0.346021	0.038447
H_21_154_	H_21_154	Absent		Central Re	Eenranos	Oorolong	NA	NA	NA	NA	1	0.333333	0.346021	0.019223
H_21_156_	H_21_156	Consent	Consent	Central Re	Eenranos	Oorolong	119.6	Yes	Female	NA	1	0.333333	0.346021	0.11534
H_21_158_	H_21_158	Consent	Consent	Central Re	Eenranos	Oorolong	80.6	NA	Female	19	1	0.333333	0.346021	0.05767
H_21_16_6	H_21_16	Consent	Consent	Central Re	Eenranos	Oorolong	57.6	No	NA	NA	1	0.333333	0.346021	0.028835


Unequal sampling probabilities
Bias


What can we do?
PPS sampling across strata
Weighting


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

What can we do?

Model the dependence

Conditional models (multilevel)

Marginal models (gee)

Robust estimators

Huber-white (sandwich)

$$\text{globorisk} = \alpha + \beta_1(\text{age}) + \epsilon$$

$$\text{globorisk}_i \sim N\left(\alpha_{j[i]} + \beta_1(\text{age}), \sigma^2\right)$$

$$\alpha_j \sim N\left(\mu_{\alpha_j}, \sigma^2_{\alpha_j}\right), \text{for geo1 j} = 1, \ldots, \text{J}$$

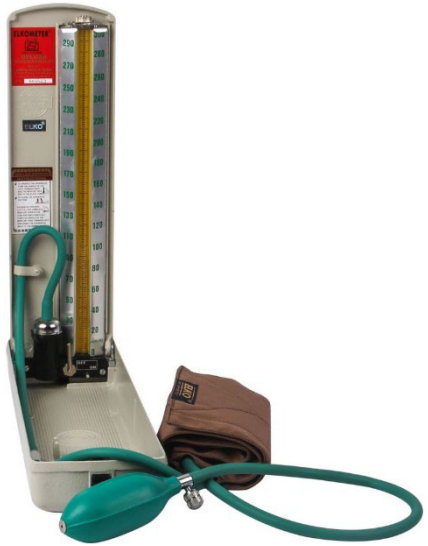
```
library(lme4)
cm <- lmer(
        globorisk_nonlab ~ age + (1 | geo1),
        data = DATA)
summary(cm)
```

$$\mu_i = \beta_0 + \beta_1(\mathrm{age_i})$$
$$Var(globorisk_{nonlab}) = \Sigma$$

```
library(gee)
DATA <- DATA[order(DATA$geo1),]


mm <- gee(globorisk_nonlab ~ age + 1 ,
          id = geo1,
          data = DATA,
          corstr = "exchangeable")
```

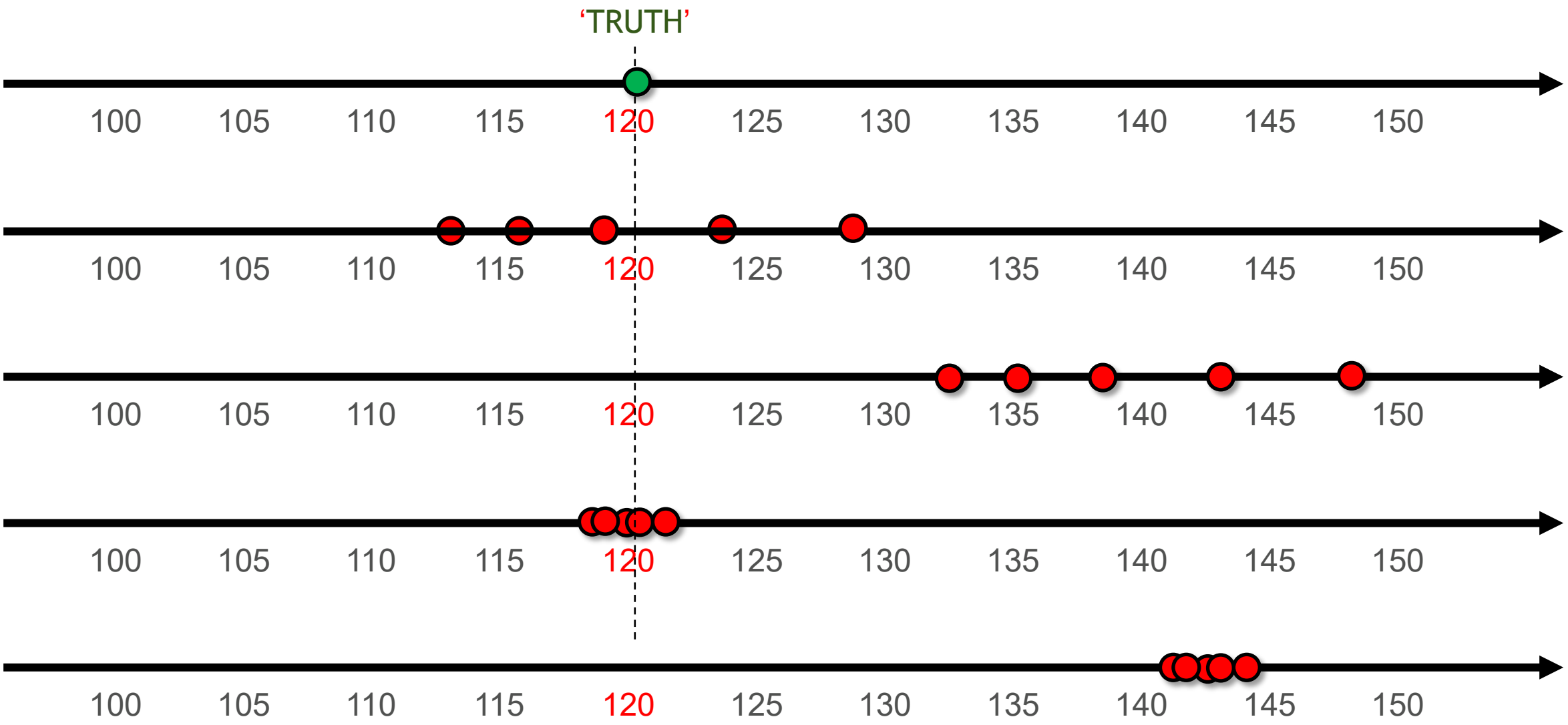
```
library(survey)

SDATA <- svydesign(id = ~geo2, strata = ~geotype, weights = ~weights, nest = TRUE,
                   fpc = NULL, data = DATA)

svymean(~globorisk_nonlab, design = SDATA, na.rm = TRUE)
```

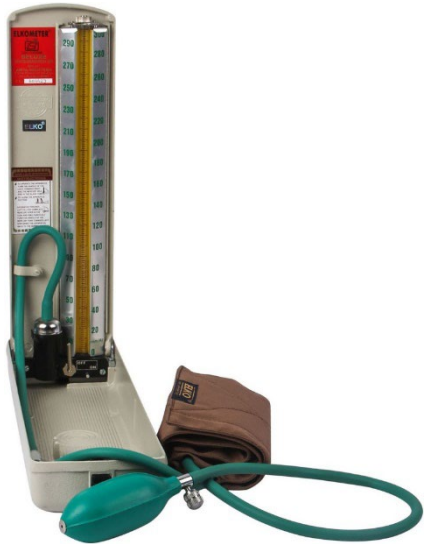
Error



125 mmHg

6%


'TRUTH'
100 105 110 115 120 125 130 135 140 145 150
100 105 110 115 120 125 130 135 140 145 150
100 105 110 115 120 125 130 135 140 145 150
100 105 110 115 120 125 130 135 140 145 150
100 105 110 115 120 125 130 135 140 145 150



125 mmHg

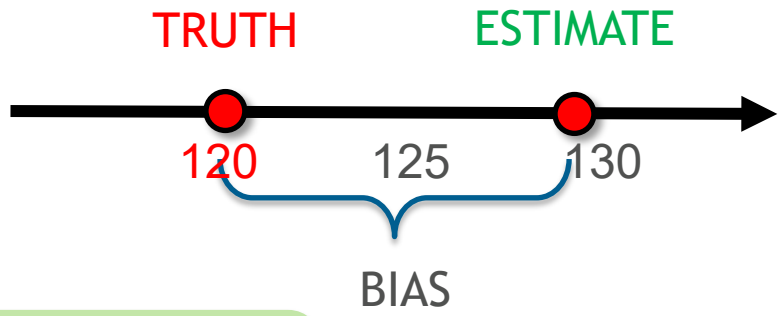
125±5mmHg

125 mmHg
(95% CI: 120 mmHg, 130 mmHg)

6%

6±1%

6%
(95% CI: 4%,6%)



Close to the truth = **accurate**

125 mmHg
(95% CI: 120 mmHg, 130 mmHg)

Small uncertainty = **precise**

References

Freedman, D. A. (2006). "On the so-called “Huber sandwich estimator” and “robust standard errors”". In: *The American Statistician* 60.4, pp. 299-302.

Lumley, T. (2004). "Analysis of Complex Survey Samples". In: *Journal of statistical software* 9, pp. 1-19.

Lumley, T. (2011). *Complex surveys: a guide to analysis using R*. John Wiley & Sons.

MacKinnon, J. G., M. Ø. Nielsen, and M. D. Webb (2023). "Cluster-Robust Inference: A Guide to Empirical Practice". In: *Journal of Econometrics* 232.2, pp. 272-299. ISSN: 0304-4076. DOI: 10.1016/j.jeconom.2022.04.001. (Visited on May. 07, 2023).

Muff, S., L. Held, and L. F. Keller (2016). "Marginal or Conditional Regression Models for Correlated Non-Normal Data?" In: *Methods in Ecology and Evolution* 7.12, pp. 1514-1524. ISSN: 2041-210X. DOI: 10.1111/2041-210X.12623. (Visited on May. 07, 2023).

Take home messages



Thank you

FOR YOUR ATTENTION AND PARTICIPATION

Annibale Cois

Email: annibale.cois@mrc.ac.za

ORCID: 0000-0002-7014-6510

WEB: annibalecois.github.io