From a few to the whole crowd

Dr Annibale Cois, PhD

Burden of Disease Research Unit, South African Medical Research Council &
School of Public Health, University of Cape Town

Dr Makandwe Nyirenda, PhD
Burden of Disease Research Unit, South African Medical Research Council &
School of Nursing and Public Health, University of KwaZulu-Natal

Cape Town, SAMRC Conference Centre, 2-3 October 2025

The South African Medical Research Council

recognizes the catastrophic and persisting consequences of colonialism and apartheid, including land dispossession and the intentional imposition of educational and health inequities.

Acknowledging the SAMRC's historical role and silence during apartheid, we commit our capacities and resources to the continued promotion of justice and dignity in health research in South Africa.

transformation

WORKING TOGETHER FOR EXCELLENCE

Program

Day 1

- Fundamentals: concepts and terminology
- Sampling strategies
- The *SurveyLab* environment
- Conducting a survey
- Discussion & recap

Day 2

- Analysing survey data: principles and approaches
- Survey data analysis with R
- Answering our research questions
- From questions, to sampling to estimates: a summary
- Conclusions

The COPD survey

Lessons, doubts, questions...

Analysing survey data

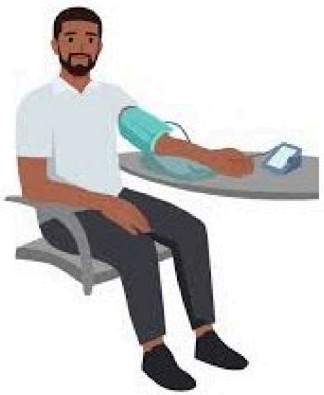
principles and approaches

The problem:

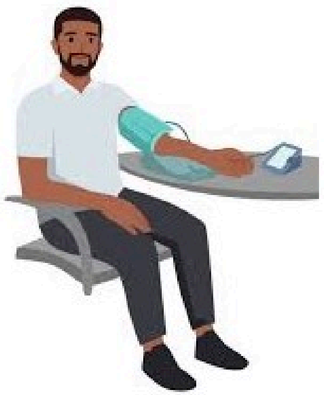
~~i.i.d.~~

$$sbp_{pop} \sim \mathrm{N}(\mu, \sigma)$$

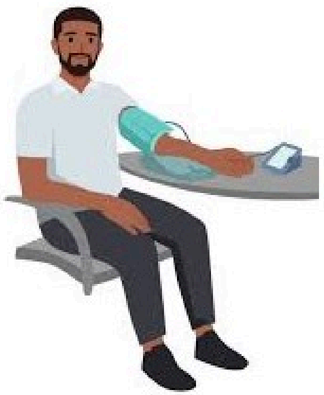
$sbp_1 \sim \mathrm{N}(\mu, \sigma)$



$sbp_2 \sim \mathrm{N}(\mu, \sigma)$

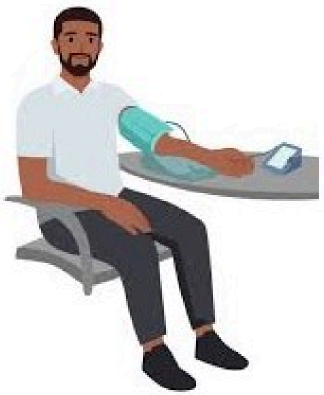


$sbp_3 \sim \mathrm{N}(\mu, \sigma)$



• • •

$sbp_n \sim \mathrm{N}(\mu, \sigma)$



Why should we care?

①

Stratification --> unequal probability of selection --> Bias

②

Suboptimal response rates --> unequal probability of selection --> Bias

③

Clustering --> non-independence --> incorrect quantification of precision

A research question

What is the prevalence of active TB in the adult population (15+)?

Simple random sampling

Current sample

Towns: 110
Households: 3878
Individuals: : 4000

Strategy

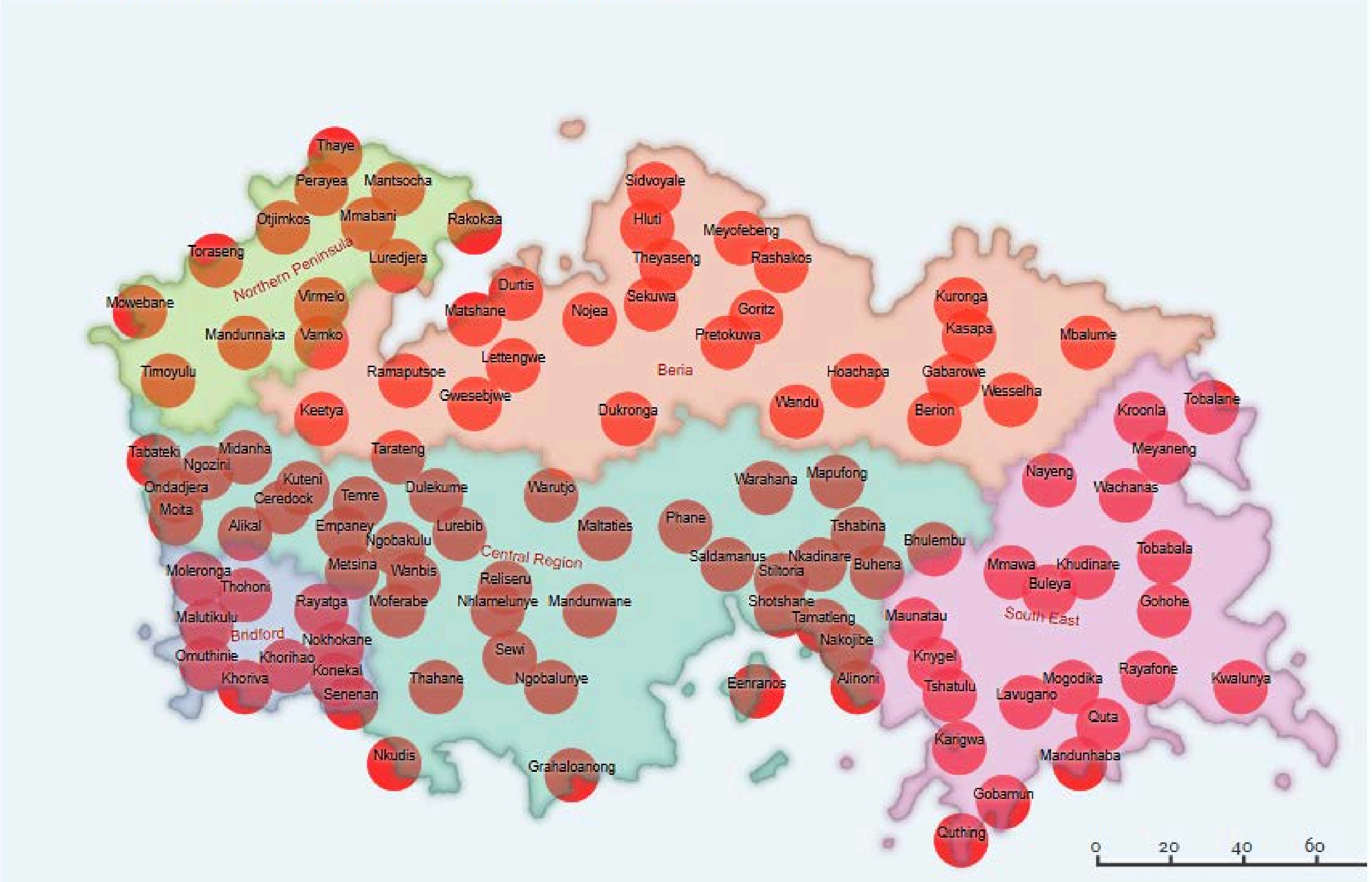
Type: simple
Strata: none
Clusters: none

Sample

NEW

RANDOM

Target population size: 167617, Maximum sampling ratio: 2.4%



```r
summary(s1[,c(5:6,8:11)])
```

```
                REGION            TOWN           SEX            AGE          DISEASE_TB      Probs
 Beria             : 956   Sewi      :  60   Female:2010   Min.   : 15.00   No :3892   Min.   :5.966e-06
 Bridford          : 342   Mandunhaba:  59   Male  :1990   1st Qu.: 25.00   Yes: 108   1st Qu.:5.966e-06
 Central Region    :1421   Gobamun   :  56                 Median : 36.00              Median :5.966e-06
 Northern Peninsula: 417   Warahana  :  52                 Mean   : 38.51              Mean   :5.966e-06
 South East        : 864   Dukronga  :  51                 3rd Qu.: 49.25              3rd Qu.:5.966e-06
                           Durtis    :  51                 Max.   :103.00              Max.   :5.966e-06
                           (Other)   :3671
```

```r
s1[1:18,c(1,2,5:11)]
```

```
# A tibble: 18 × 9
   IID       HID     REGION TOWN    SUBURB     SEX      AGE DISEASE_TB      Probs
   <chr>     <chr>   <fct>  <fct>   <fct>      <fct>  <dbl> <fct>           <dbl>
 1 H_1_116_1 H_1_116 Beria  Durtis Lobammunye Male      67 No         0.00000597
 2 H_1_128_2 H_1_128 Beria  Durtis Lobammunye Female    23 No         0.00000597
 3 H_1_134_1 H_1_134 Beria  Durtis Lobammunye Male      32 No         0.00000597
 4 H_1_138_2 H_1_138 Beria  Durtis Lobammunye Male      31 No         0.00000597
 5 H_1_148_2 H_1_148 Beria  Durtis Lobammunye Female    23 No         0.00000597
 6 H_1_181_1 H_1_181 Beria  Durtis Fiog       Male      26 No         0.00000597
 7 H_1_195_1 H_1_195 Beria  Durtis Fiog       Male      38 No         0.00000597
 8 H_1_205_4 H_1_205 Beria  Durtis Fiog       Male      24 No         0.00000597
 9 H_1_208_1 H_1_208 Beria  Durtis Fiog       Male      35 No         0.00000597
10 H_1_212_1 H_1_212 Beria  Durtis Fiog       Male      41 No         0.00000597
11 H_1_233_2 H_1_233 Beria  Durtis Fiog       Female    35 No         0.00000597
12 H_1_245_4 H_1_245 Beria  Durtis Fiog       Female    17 No         0.00000597
13 H_1_253_1 H_1_253 Beria  Durtis Fiog       Male      38 No         0.00000597
14 H_1_27_1  H_1_27  Beria  Durtis Lobammunye Male      25 No         0.00000597
15 H_1_306_2 H_1_306 Beria  Durtis Hoban      Female    41 No         0.00000597
16 H_1_328_1 H_1_328 Beria  Durtis Hoban      Male      39 No         0.00000597
17 H_1_347_8 H_1_347 Beria  Durtis Hoban      Female    19 No         0.00000597
18 H_1_350_1 H_1_350 Beria  Durtis Hoban      Male      65 No         0.00000597
```

```r
d <- table(s1$DISEASE_TB)
print(d)
```

```

  No  Yes 
3892  108 
```

```r
p <- (d[2]/sum(d))
round(p*100,2)
```

```
Yes 
2.7 
```

```r
se <- sqrt((p * (1- p))/sum(d)) |> print()
```

```
        Yes 
0.002562762 
```

```r
cv <- qnorm((1-0.05/2)) |> print()
```

```
[1] 1.959964
```

```r
ci <- c(p  - cv*se, p + cv*se)
round(ci*100,2)
```

```
Yes Yes 
2.2 3.2 
```

```r
round((ci[2]-ci[1])*100,2)
```

```
Yes 
  1 
```

Estimates

Sample estimates

Group	No	Yes
All respondents	97.3	2.7

Population true values

Group	No	Yes
Target population	97.18	2.82

Estimates

Sample estimates

Group	No	Yes
Beria	98.64	1.36
Bridford	91.81	8.19
Central Region	96.97	3.03
Northern Peninsula	97.84	2.16
South East	98.26	1.74

Population true values

1	No	Yes
Beria	97.54	2.46
Bridford	92.12	7.88
Central Region	97.34	2.66
Northern Peninsula	98.18	1.82
South East	98.05	1.95

Stratified random sampling

Current sample

Towns: 110
Households: 3829
Individuals: : 4000

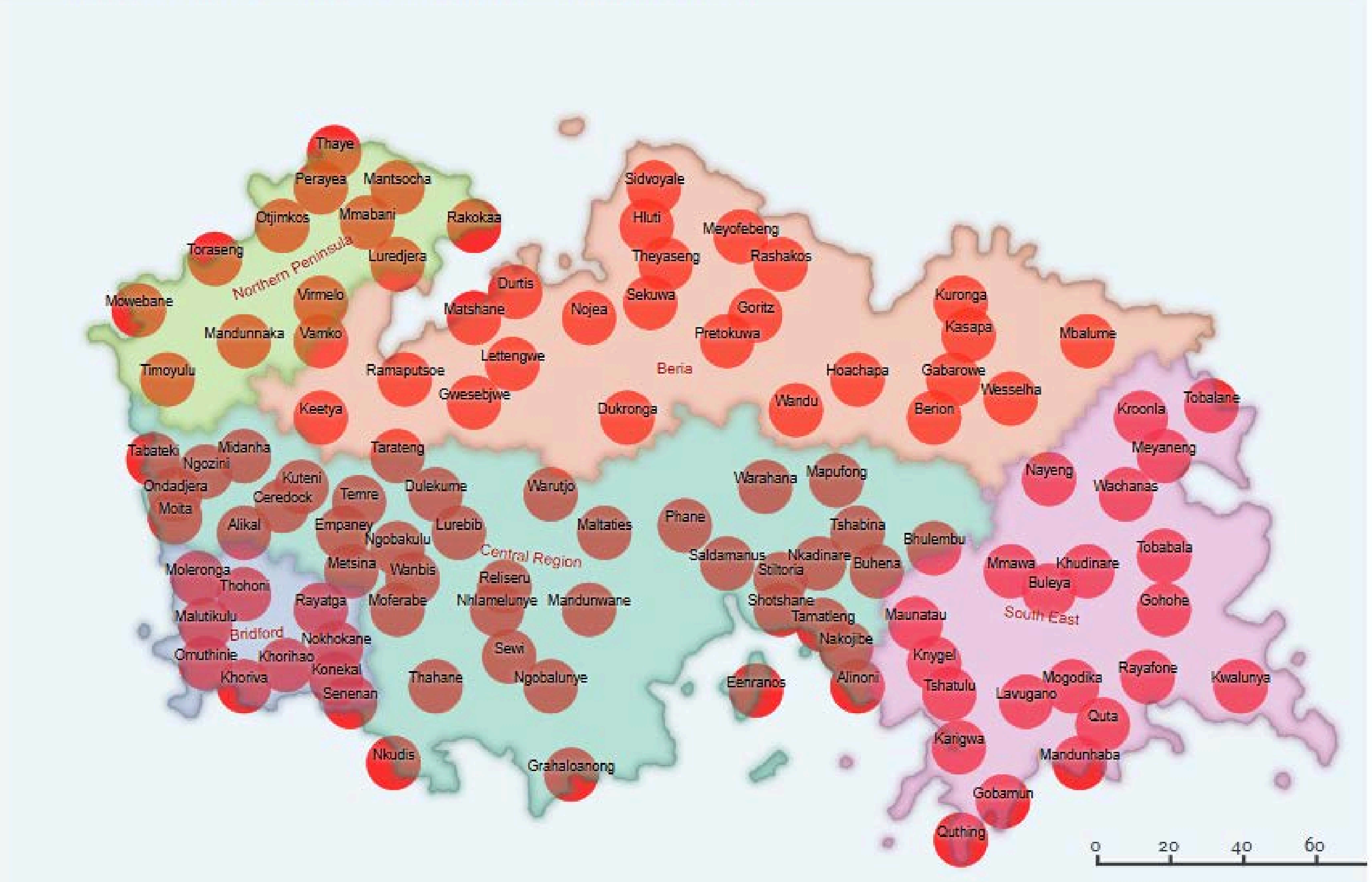
Strategy

Type: stratified
Strata: region
Clusters: none

Sample

NEW RANDOM

Target population size: 167617, Maximum sampling ratio: 5.6%



```r
summary(s2[,c(5:6,8:11)])
```

```
                REGION            TOWN          SEX            AGE          DISEASE_TB     Probs          
 Beria             :800   Khorihao  : 116   Female:1973   Min.   : 15.00   No :3867   Min.   :1.691e-05  
 Bridford          :800   Nokhokane : 104   Male  :2027   1st Qu.: 25.00   Yes: 133   1st Qu.:2.516e-05  
 Central Region    :800   Malutikulu:  98                 Median : 35.00              Median :2.716e-05  
 Northern Peninsula:800   Omuthinie :  95                 Mean   : 38.05              Mean   :3.916e-05  
 South East        :800   Khoriva   :  91                 3rd Qu.: 49.00              3rd Qu.:5.695e-05  
                          Mmabani   :  88                 Max.   :102.00              Max.   :6.961e-05  
                          (Other)   :3408                                                                
```

```r
s2[1:18,c(1,2,5:11)]
```

```
# A tibble: 18 × 9
   IID       HID     REGION TOWN   SUBURB           SEX      AGE DISEASE_TB     Probs
   <chr>     <chr>   <fct>  <fct>  <fct>            <fct>  <dbl> <fct>          <dbl>
 1 H_1_133_1 H_1_133 Beria  Durtis Lobammunye       Male      40 No         0.0000252
 2 H_1_137_6 H_1_137 Beria  Durtis Lobammunye       Female    61 No         0.0000252
 3 H_1_20_1  H_1_20  Beria  Durtis Lobammunye       Male      62 No         0.0000252
 4 H_1_222_2 H_1_222 Beria  Durtis Fiog             Female    23 No         0.0000252
 5 H_1_229_4 H_1_229 Beria  Durtis Fiog             Female    33 No         0.0000252
 6 H_1_237_5 H_1_237 Beria  Durtis Fiog             Female    25 No         0.0000252
 7 H_1_239_4 H_1_239 Beria  Durtis Fiog             Female    25 No         0.0000252
 8 H_1_258_1 H_1_258 Beria  Durtis Fiog             Female    53 No         0.0000252
 9 H_1_266_6 H_1_266 Beria  Durtis Fiog             Male      37 No         0.0000252
10 H_1_275_3 H_1_275 Beria  Durtis Hoban            Male      46 No         0.0000252
11 H_1_282_1 H_1_282 Beria  Durtis Hoban            Male      27 No         0.0000252
12 H_1_294_4 H_1_294 Beria  Durtis Hoban            Male      35 No         0.0000252
13 H_1_33_2  H_1_33  Beria  Durtis Lobammunye       Male      38 No         0.0000252
14 H_1_41_6  H_1_41  Beria  Durtis Lobammunye       Female    61 No         0.0000252
15 H_1_471_4 H_1_471 Beria  Durtis Hoban            Female    20 No         0.0000252
16 H_1_496_5 H_1_496 Beria  Durtis Hoban            Male      23 No         0.0000252
17 H_1_497_1 H_1_497 Beria  Durtis Hoban            Male      41 No         0.0000252
18 H_1_584_1 H_1_584 Beria  Durtis Plegak Boulevard Male      35 No         0.0000252
```

```r
d <- table(s2$DISEASE_TB)
p <- (d[2]/sum(d))
round(p*100,2)
```

```
 Yes
3.33
```

```r
se <- sqrt((p * (1- p))/sum(d))
cv <- qnorm((1-0.05/2))
ci <- c(p  - cv*se, p + cv*se)
round(ci*100,2)
```

```
 Yes  Yes
2.77 3.88
```

```r
round((ci[2]-ci[1])*100,2)
```

```
 Yes
1.11
```

Adjusting for sampling design:

Stratified

```
[1] "Point Estimate: 3"

[1] "95% Confidence Interval: 2.49 - 3.64"

[1] "95% Confidence Interval width: 1.15"

[1] "Design effect: 1.3"
```

Simple

```
[1] "Point Estimate: 2.7"

[1] "95% Confidence Interval: 2.2 - 3.2"

[1] "95% Confidence Interval width: 1"

[1] "Design effect: 1"
```

Cluster random sampling

Current sample

Towns: 110
Households: 3829
Individuals: : 4000

Strategy

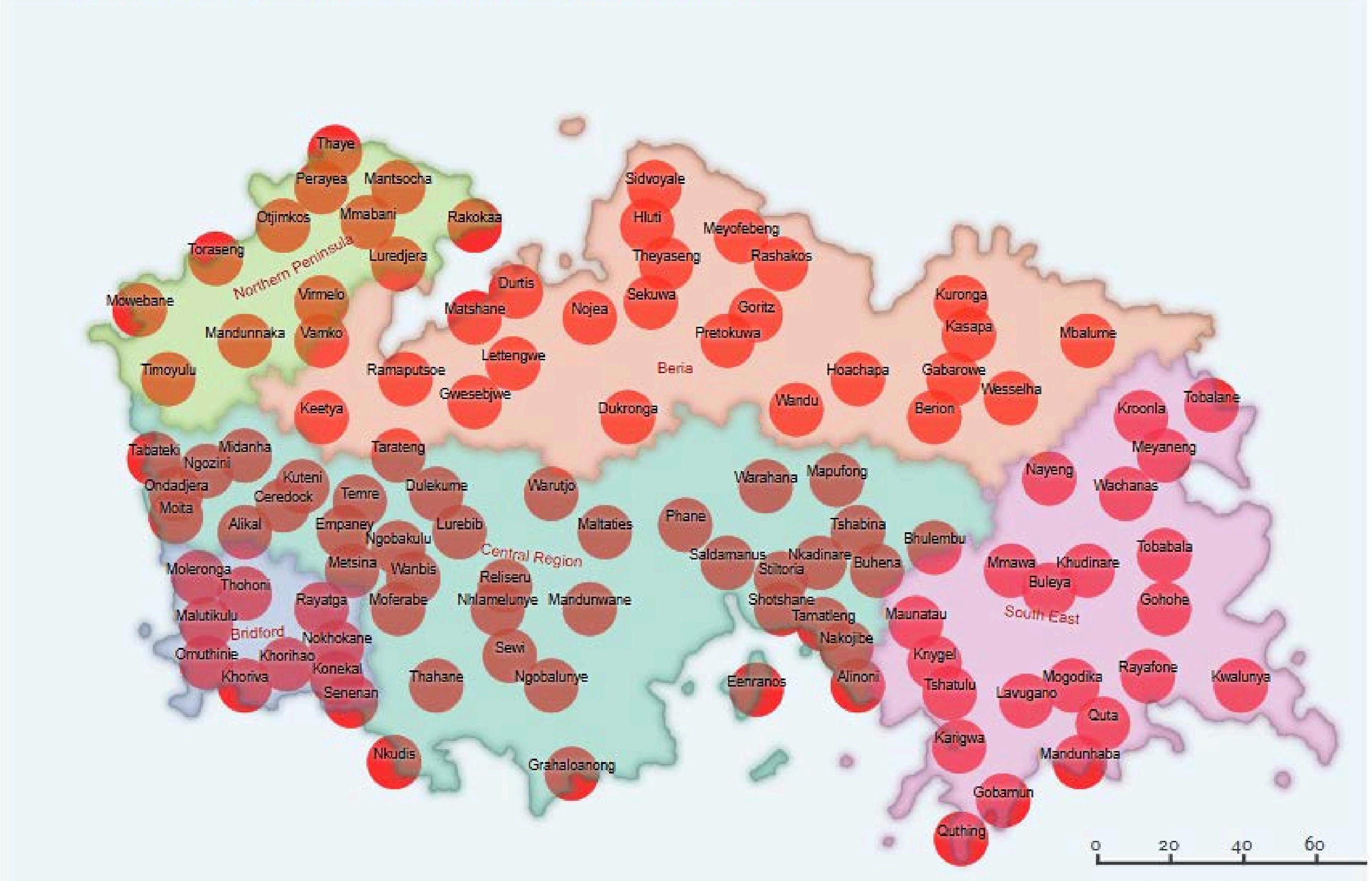
Type: stratified
Strata: region
Clusters: none

Sample

NEW

RANDOM

Target population size: 167617, Maximum sampling ratio: 5.6%



```r
d <- table(s3$DISEASE_TB)
p <- (d[2]/sum(d))
round(p*100,2)
```

```
 Yes
1.45
```

```r
se <- sqrt((p * (1- p))/sum(d))
cv <- qnorm((1-0.05/2))
ci <- c(p  - cv*se, p + cv*se)
round(ci*100,2)
```

```
 Yes  Yes
1.08 1.82
```

```r
round((ci[2]-ci[1])*100,2)
```

```
 Yes
0.74
```

Adjusting for sampling design:

Cluster

```
[1] "Point Estimate: 2"
[1] "95% Confidence Interval: 0.37 - 7.51"
[1] "95% Confidence Interval width: 7.14"
[1] "Design effect: 50.9"
```

Simple

```
[1] "Point Estimate: 2.7"
[1] "95% Confidence Interval: 2.2 - 3.2"
[1] "95% Confidence Interval width: 1"
[1] "Design effect: 1"
```

Response rates

Current sample

Towns: 20
Households: 4000
Individuals: : 4000

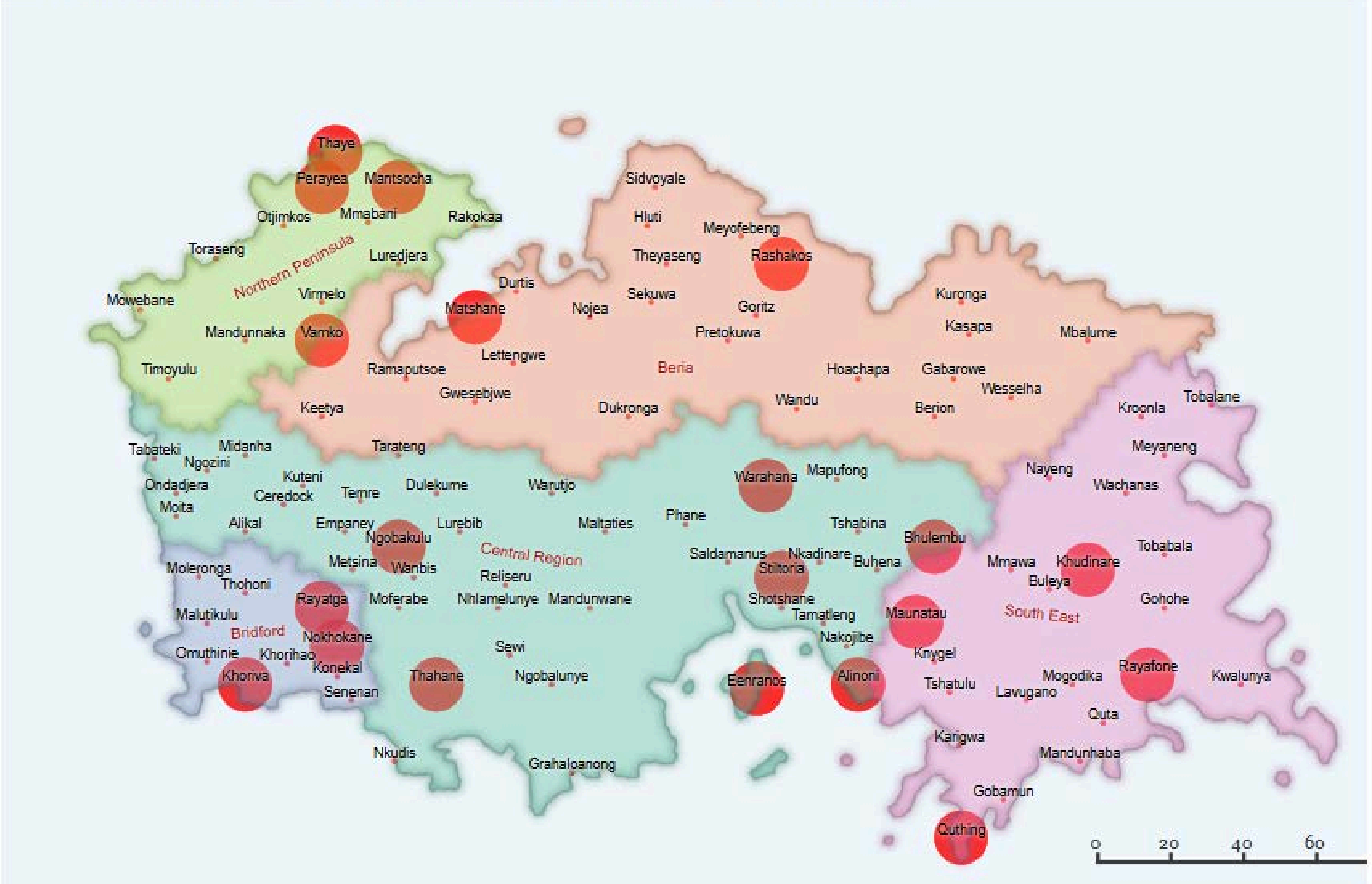
Strategy

Type:
Strata: none
Clusters: town, hid

Sample

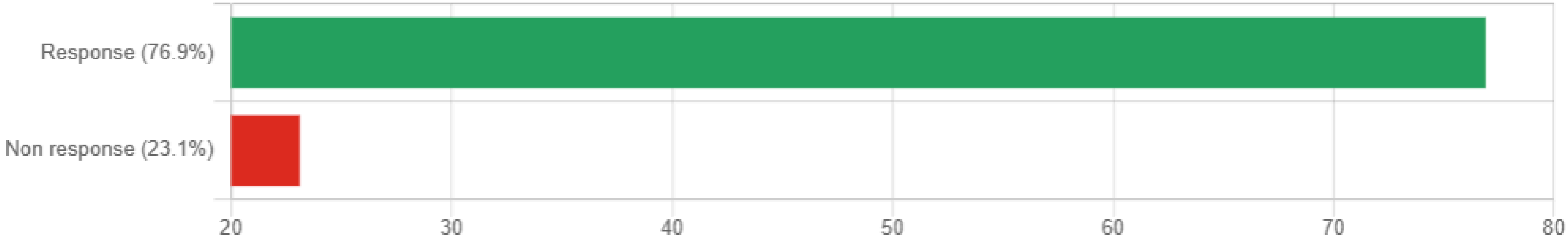
NEW RANDOM

Target population size: 167617, Maximum sampling ratio: 18.2% / 69.2% / 100%

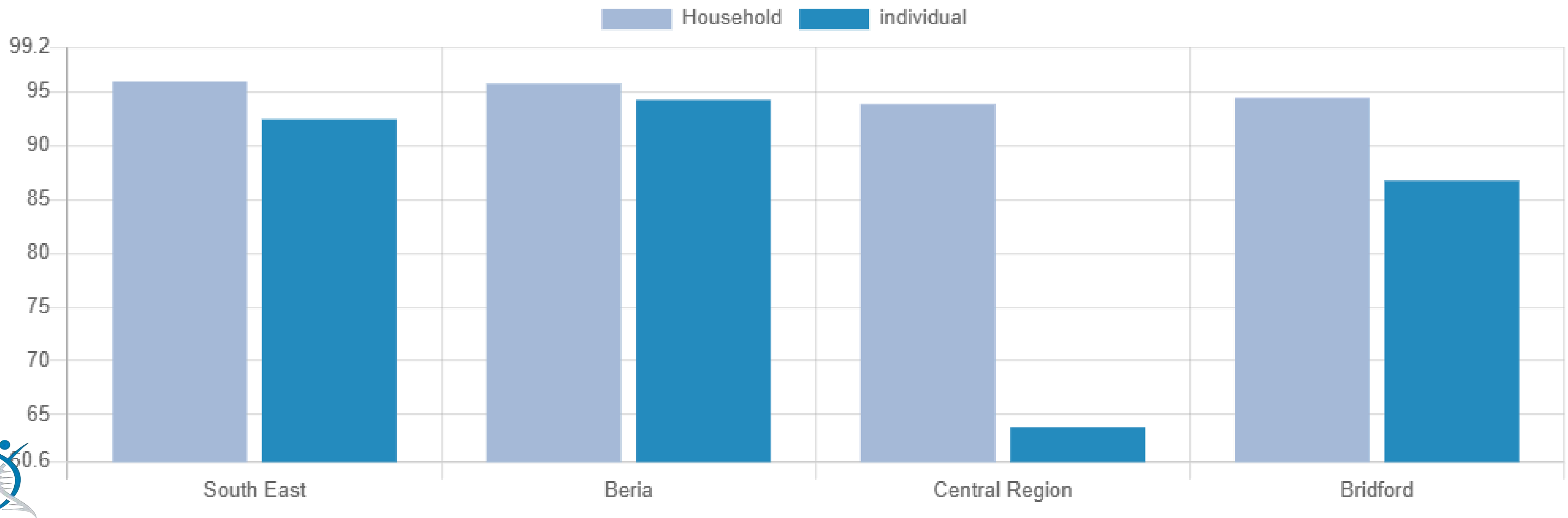


Response rates

Household: 95% Individual: 80.9% Total: 76.9%



Response rates by region



```r
summary(s4[,c(5:6,8:11)])
```

```
                 REGION              TOWN          SEX            AGE         DISEASE_TB       Probs
 Beria             : 956   Gabarowe  :  54   Female:1391   Min.   :15.00   No  :2711   Min.   :5.966e-06
 Bridford          : 335   Sewi      :  54   Male  :1393   1st Qu.:25.00   Yes :  73   1st Qu.:5.966e-06
 Central Region    :1422   Buhena    :  53   NA's  :1216   Median :36.00   NA's:1216   Median :5.966e-06
 Northern Peninsula: 405   Reliseru  :  51                 Mean   :38.76               Mean   :5.966e-06
 South East        : 882   Alinoni   :  50                 3rd Qu.:50.00               3rd Qu.:5.966e-06
                           Malutikulu:  50                 Max.   :94.00               Max.   :5.966e-06
                           (Other)   :3688                 NA's   :1216
```

```r
s4[1:18,c(1,2,5,6,8:11)]
```

```
# A tibble: 18 × 8
   IID         HID       REGION              TOWN       SEX     AGE DISEASE_TB      Probs
   <chr>       <chr>     <fct>               <fct>      <fct> <dbl> <fct>           <dbl>
 1 H_102_180_2 H_102_180 South East          Mandunhaba Male     28 Yes        0.00000597
 2 H_102_198_1 H_102_198 South East          Mandunhaba Male     42 Yes        0.00000597
 3 H_108_120_2 H_108_120 South East          Tobabala   Male     28 Yes        0.00000597
 4 H_108_172_1 H_108_172 South East          Tobabala   Male     25 Yes        0.00000597
 5 H_108_71_1  H_108_71  South East          Tobabala   Male     39 Yes        0.00000597
 6 H_108_756_1 H_108_756 South East          Tobabala   Male     30 Yes        0.00000597
 7 H_108_774_1 H_108_774 South East          Tobabala   Male     53 Yes        0.00000597
 8 H_11_32_3   H_11_32   Northern Peninsula Thaye      Male     51 Yes        0.00000597
 9 H_16_136_1  H_16_136  Central Region      Tabateki   Male     32 Yes        0.00000597
10 H_16_145_6  H_16_145  Central Region      Tabateki   Male     27 Yes        0.00000597
11 H_16_175_4  H_16_175  Central Region      Tabateki   Male     27 Yes        0.00000597
12 H_17_138_2  H_17_138  Central Region      Sewi       Male     20 Yes        0.00000597
13 H_17_167_1  H_17_167  Central Region      Sewi       Male     25 Yes        0.00000597
14 H_17_222_5  H_17_222  Central Region      Sewi       Male     19 Yes        0.00000597
15 H_17_578_1  H_17_578  Central Region      Sewi       Male     28 Yes        0.00000597
16 H_17_597_2  H_17_597  Central Region      Sewi       Male     32 Yes        0.00000597
17 H_17_600_2  H_17_600  Central Region      Sewi       Male     39 Yes        0.00000597
18 H_19_142_1  H_19_142  Central Region      Mapufong   Male     28 Yes        0.00000597
```

```
d <- table(s4$DISEASE_TB)
p <- (d[2]/sum(d))
round(p*100,2)
```

```
 Yes
2.62
```

```
se <- sqrt((p * (1- p))/sum(d))
cv <- qnorm((1-0.05/2))
ci <- c(p  - cv*se, p + cv*se)
round(ci*100,2)
```

```
 Yes  Yes
2.03 3.22
```

```
round((ci[2]-ci[1])*100,2)
```

```
 Yes
1.19
```

Adjusting for sampling design:

Cluster

```
[1] "Point Estimate: 2.62"

[1] "95% Confidence Interval: 2.03 - 3.22"

[1] "95% Confidence Interval width: 1.19"

[1] "Design effect: 1"
```

Simple

```
[1] "Point Estimate: 2.7"

[1] "95% Confidence Interval: 2.2 - 3.2"

[1] "95% Confidence Interval width: 1"

[1] "Design effect: 1"
```

Estimates

Sample estimates

Group	No	Yes
Beria	98.03	1.97
Bridford	93.24	6.76
Central Region	97.92	2.08
Northern Peninsula	97.67	2.33
South East	98.24	1.76

Population true values

1	No	Yes
Beria	97.54	2.46
Bridford	92.12	7.88
Central Region	97.34	2.66
Northern Peninsula	98.18	1.82
South East	98.05	1.95

SAMRC advancinglife

Sample	Tool	Cost	Unit cost
Towns: 110 **Households: 3878** **Individuals: : 4000**	**Items: 2** **Measurements: 1** **Time: 6 [min]**	**Set up: 346** **Collection: 22** **Total: 368**	**Estimated: 0.09** **Realised: 0.09**
Towns: 110 **Households: 3829** **Individuals: : 4000**	**Items: 2** **Measurements: 1** **Time: 6 [min]**	**Set up: 342** **Collection: 21** **Total: 363**	**Estimated: 0.09** **Realised: 0.09**
Towns: 20 **Households: 4000** **Individuals: : 4000**	**Items: 2** **Measurements: 1** **Time: 6 [min]**	**Set up: 70.5** **Collection: 10.1** **Total: 80.6**	**Estimated: 0.02** **Realised: 0.02**
Towns: 110 **Households: 3871** **Individuals: : 4000**	**Items: 2** **Measurements: 1** **Time: 6 [min]**	**Set up: 346** **Collection: 21.8** **Total: 367.8**	**Estimated: 0.09** **Realised: 0.13**

What to do?

UNEQUAL SAMPLING PROBABILITY

SELF_WEIGHTING SAMPLE

```
          Beria          Bridford    Central Region Northern Peninsula        South East
          39749             14366             59128             17559             36815
```

```
round(POP/sum(POP),2)
```

```
          Beria          Bridford    Central Region Northern Peninsula        South East
           0.24              0.09              0.35              0.10              0.22
```

```
round(POP/sum(POP)*1000,0)
```

```
          Beria          Bridford    Central Region Northern Peninsula        South East
            237                86               353               105               220
```

WEIGHTING

```r
P <- 800/POP |> print()    # sampling probability
```

```
          Beria          Bridford    Central Region Northern Peninsula          South East
          39749             14366             59128              17559               36815
```

```r
w <- 1/(800/POP) |> print()    # inverse of sampling probability = sampling weight
```

```
          Beria          Bridford    Central Region Northern Peninsula          South East
     0.02012629        0.05568704        0.01352997         0.04556068          0.02173027
```

Estimates

Sample estimates

Group	No	Yes
Beria	96.75	3.25
Bridford	93.25	6.75
Central Region	97.25	2.75
Northern Peninsula	98.62	1.38
South East	97.5	2.5

Population true values

1	No	Yes
Beria	97.54	2.46
Bridford	92.12	7.88
Central Region	97.34	2.66
Northern Peninsula	98.18	1.82
South East	98.05	1.95

```r
pr <- c(3.25,6.75,2.75,1.38,2.5)
pr %*% w/sum(w)
```

```
         [,1]
[1,] 3.012974
```

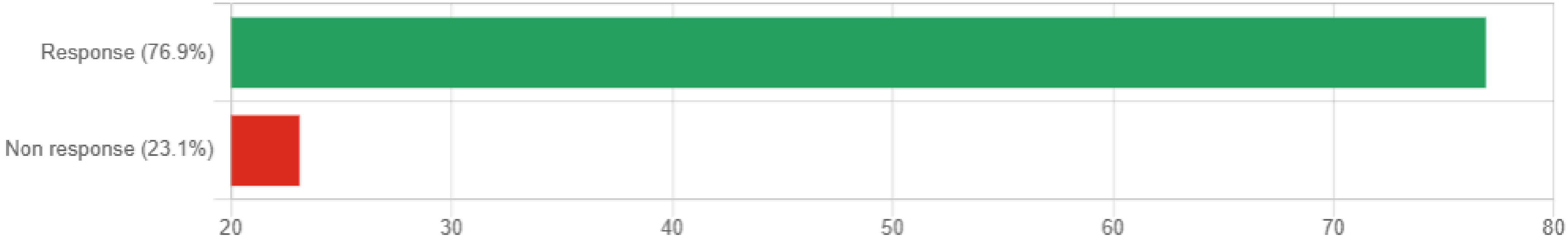
```r
pp <- c(2.46,7.88,2.66,1.82,1.96)
pp %*% w/sum(w)
```

```
         [,1]
[1,] 2.818222
```

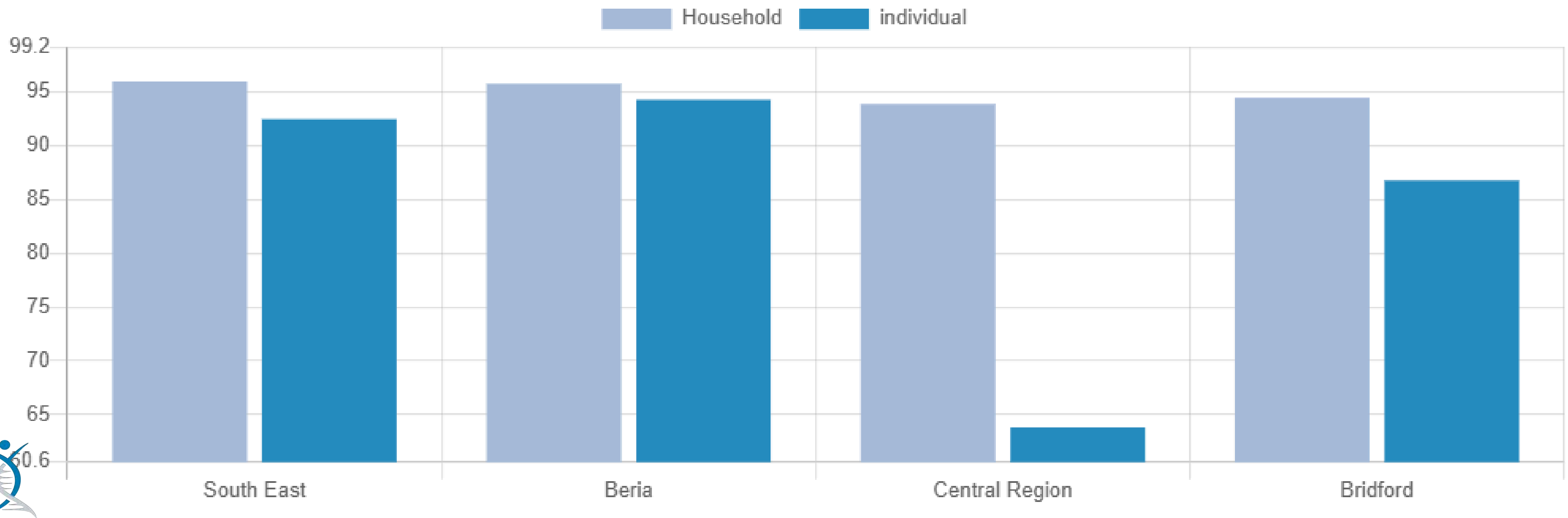
DESIGN WEIGHTS

Response rates

Household: 95% Individual: 80.9% Total: 76.9%



Response rates by region



Estimates

Sample estimates

Group	No	Yes
All respondents	97.74	2.26

Population true values

Group	No	Yes
Target population	97.18	2.82

Estimates

Sample estimates

Group	No	Yes
Beria	98.03	1.97
Bridford	93.24	6.76
Central Region	97.92	2.08
Northern Peninsula	97.67	2.33
South East	98.24	1.76

Population true values

1	No	Yes
Beria	97.54	2.46
Bridford	92.12	7.88
Central Region	97.34	2.66
Northern Peninsula	98.18	1.82
South East	98.05	1.95

```r
pr1 <- c(1.97,6.76,2.08,2.33,1.76)

rr <- c(0.69,0.85,0.71,0.71,0.44)
w1 <- 1/rr
pr1 %*% w1/sum(w1)
```

```
         [,1]
[1,] 2.724337
```

POST-STRATIFIED/CALIBRATED WEIGHTS

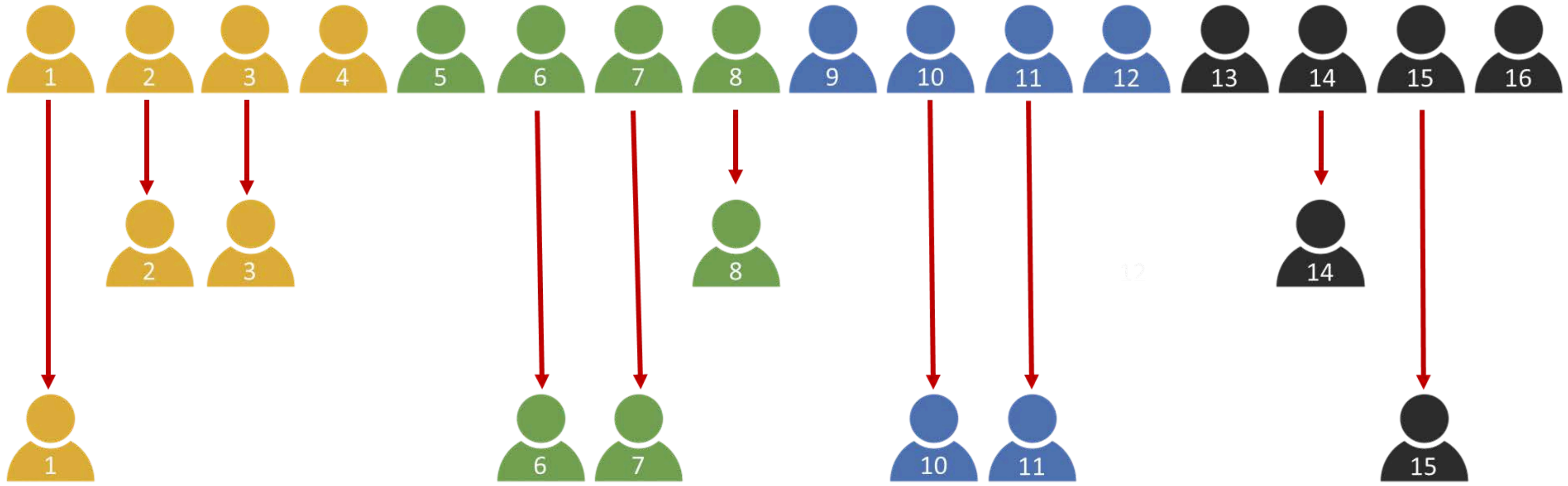
Design weights that have been further adjusted (calibrated) to match known population totals for different demographic groups (like age, race, or sex). This ensures the sample accurately reflects the proportions of these groups in the population

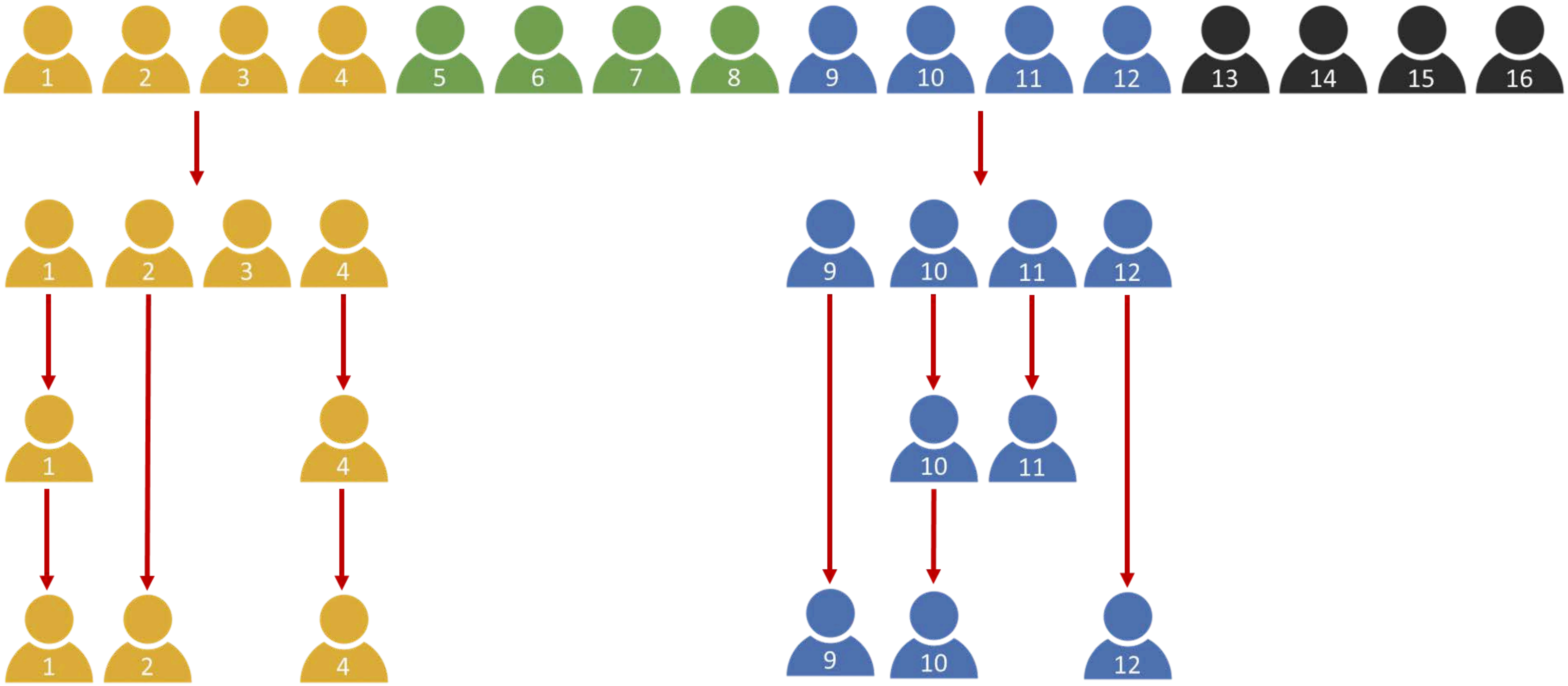
Wittenberg (2009); Wittenberg and Branson (2021)

```
# A tibble: 19 × 8
   w1_hhid w1_prov2001 w1_geo2001  w1_dc2001 w1_cluster      w1_wgt     w1_dwgt    w1_dtwgt
     <dbl>  <hvn_lbll> <hvn_lbll> <hvn_lbll> <hvn_lbll> <hvn_lbll> <hvn_lbll> <hvn_lbll>
 1  101012           2          2         15       2033    1436.426    2596.722    2601.198
 2  101013           2          2         15       2033    3869.675    2596.722    2601.198
 3  101014           2          2         15       2033    3013.344    2596.722    2601.198
 4  101015           2          2         15       2033    2352.057    2596.722    2601.198
 5  101016           2          2         15       2034    2251.118    2585.365    2589.822
 6  101017           2          2         15       2033    4595.201    2596.722    2601.198
 7  101018           2          2         15       2033    5870.196    2596.722    2601.198
 8  101020           2          2         15       2033    2417.754    2596.722    2601.198
 9  101021           2          2         15       2033    2633.920    2596.722    2601.198
10  101022           2          2         15       2033    2923.825    2596.722    2601.198
11  101023           2          2         15       2033    2826.571    2596.722    2601.198
12  101024           2          2         15       2033    2275.558    2596.722    2601.198
13  101025           2          2         15       2033    3635.668    2596.722    2601.198
14  101027           2          2         15       2033    3365.485    2596.722    2601.198
15  101028           2          2         15       2033    2931.683    2596.722    2601.198
16  101029           2          2         15       2033    2830.776    2596.722    2601.198
17  101030           2          2         15       2033    2648.129    2596.722    2601.198
18  101033           2          2         15       2034    2956.331    2585.365    2589.822
19  101034           2          2         15       2034    3092.255    2585.365    2589.822
```

NON-INDEPENDENCE (CLUSTER SAMPLING)

Why it decreases precision?


1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
2 3 8 14
1 6 7 10 11 15


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
1
2
3
4
9
10
11
12
1
4
10
11
1
2
4
9
10
12

Clustering:

Dividing the population into groups and sampling from a random subset of these groups will **decrease precision for a given sample size**

The reduction in precision increases as the units within each cluster become **more similar to each other**.

A measure of the level of "similarity" of the units in each cluster (compared to units in different clusters) is given by the **intraclass correlation coefficient**:

$$ICC = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_e^2}$$

A measure of the effect of ICC on the precision of the estimates is given by the **design effect**, which is (an estimate of) the increase of the standar error of the estimates when using cluster sampling rather than simple random sampling:

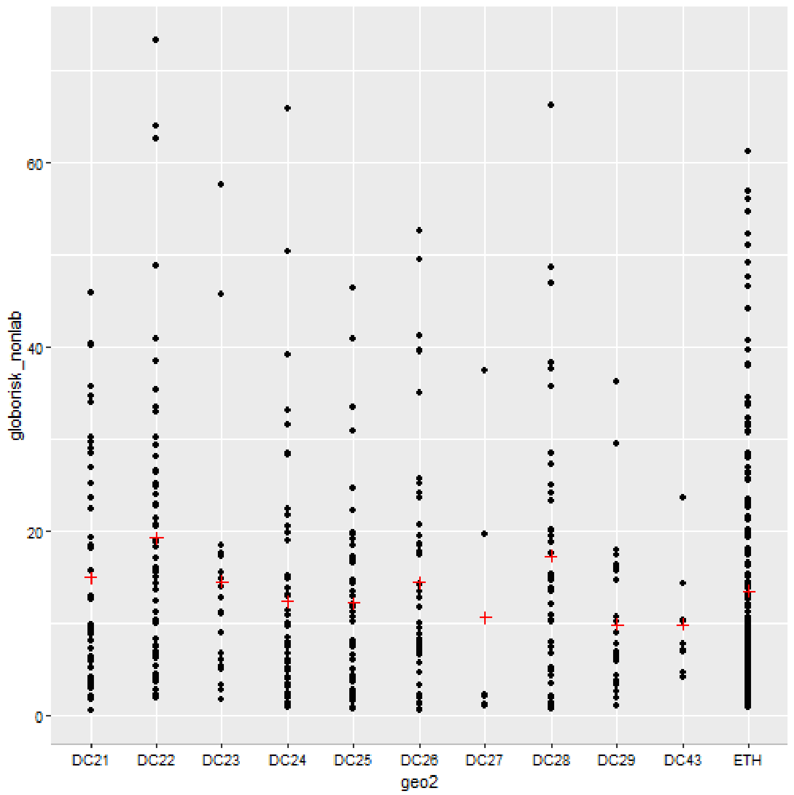
$$DEFF = \sqrt{\frac{\sigma^2_{CS}}{\sigma^2_{SRS}}}$$

The **effective sample size** in survey data is (an estimate of) the sample that we would have needed in the hypothesiss of simple random sampling:

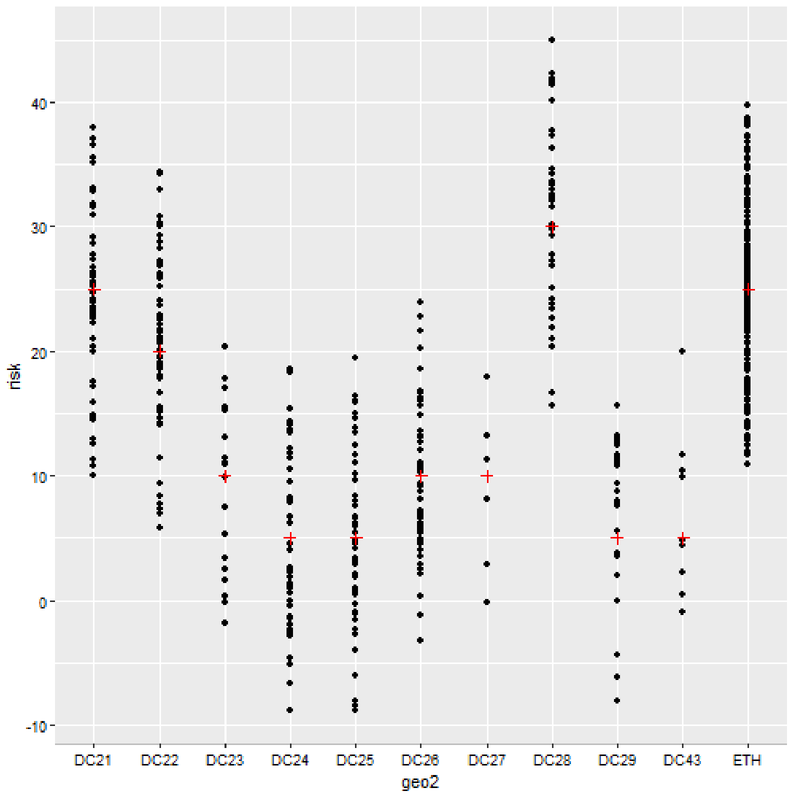
$$n_{eff} = \frac{n}{DEFF}$$

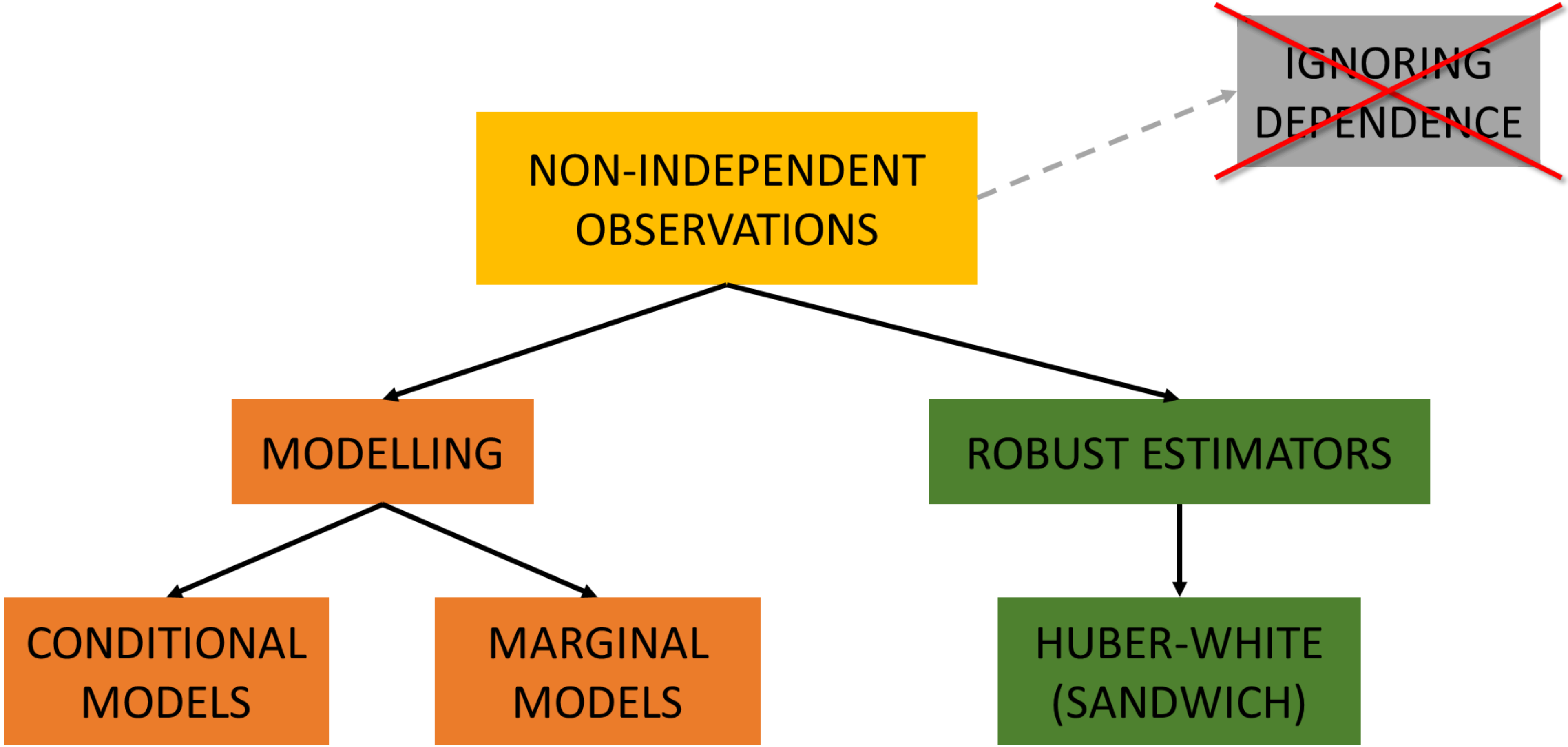
An example

ICC = 0.01, DEFF = 1.62, NEFF = 339



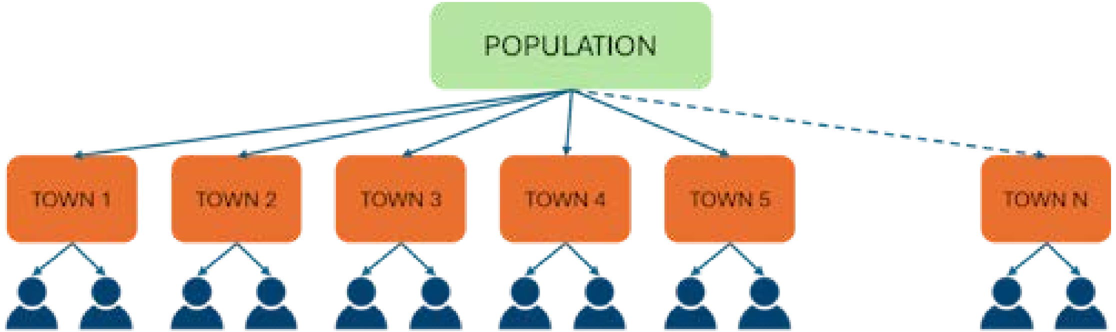
ICC = 0.64, DEFF = 31.58, NEFF = 17





MacKinnon, Nielsen, and Webb (2023)

Conditional Models



$$
\begin{aligned}
sbp_i &= \beta_0 + \epsilon_i \\
\epsilon_i &\sim N(\mu, \sigma) \\
\beta_0 = \beta_{0,j} &\sim N(\mu', \sigma')\text{ , for town j} = 1, \ldots,\text{N}
\end{aligned}
$$

$$
\begin{aligned}
sbp_i &= \beta_0 + \beta_1 \cdot age_i + \epsilon_i \\
\epsilon_i &\sim N(\mu, \sigma) \\
\beta_0 = \beta_{0,j} &\sim N(\mu', \sigma')\text{ , for region j} = 1, \ldots,\text{N}
\end{aligned}
$$

```r
model_unadjusted <- lm(
  SBP_1 ~ 1,
  data = s5)

summary(model_unadjusted)$coefficients
```

```
            Estimate Std. Error  t value Pr(>|t|)
(Intercept) 110.7501  0.5772193 191.8684        0
```

--

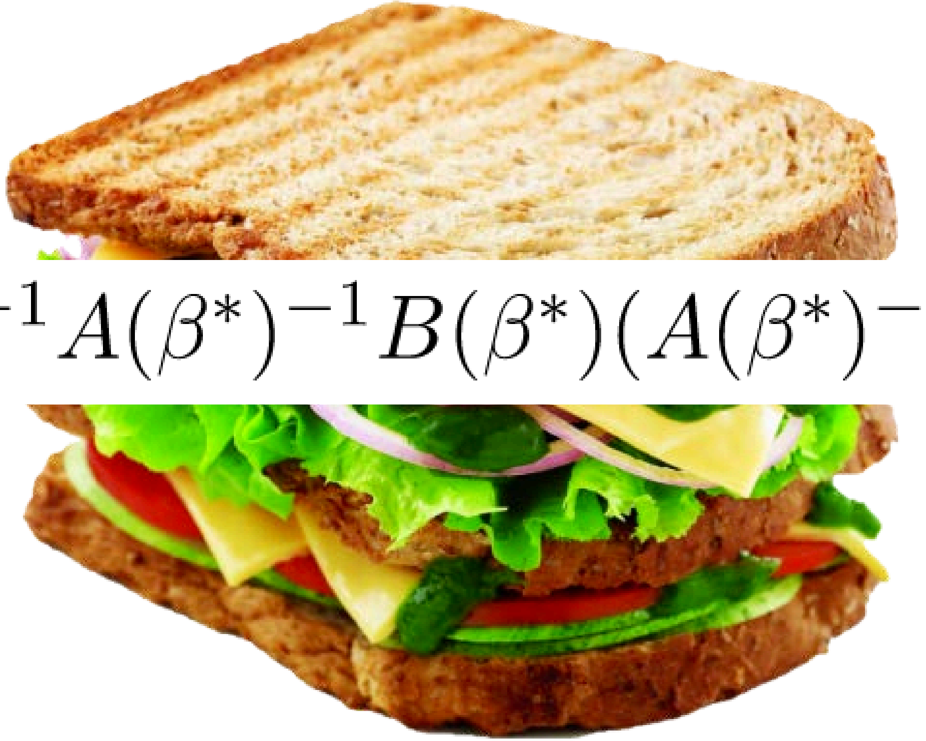
```r
model_conditional <- lmer(
  SBP_1 ~ (1 | TOWN),
  data = s5)
summary(model_conditional)$coefficients
```

```
            Estimate Std. Error  t value
(Intercept) 110.7312  0.8071857 137.1818
```

Marginal Models

$$sbp_i = \beta_0 + \beta_1 \cdot age_i + \epsilon_i$$
$$Var(sbp) = \Sigma$$

'Robust' estimators



$$n^{-1}A(\beta^*)^{-1}B(\beta^*)(A(\beta^*)^{-1})^T$$

The robust sandwich variance estimator for linear regression (theory)

MacKinnon, Nielsen, and Webb (2023)

```r
model_marginal <- gee(SBP_1 ~ 1, id = TOWN, data = s5, corstr = "independence")
```

```
(Intercept)
   110.7501
```

```r
summary(model_marginal)$coefficients
```

```
            Estimate Naive S.E.  Naive z Robust S.E. Robust z
(Intercept) 110.7501  0.5772193 191.8684   0.7857947 140.9403
```

--

```r
model_marginal <- gee(SBP_1 ~ 1, id = TOWN, data = s5, corstr = "exchangeable")
```

```
(Intercept)
   110.7501
```

```r
summary(model_marginal)$coefficients
```

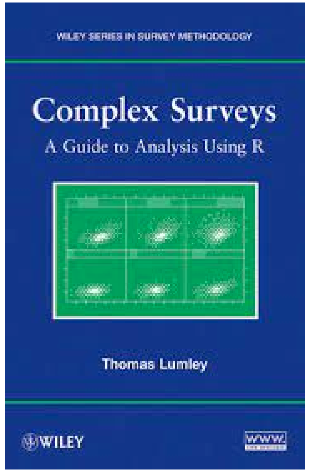
```
            Estimate Naive S.E.  Naive z Robust S.E. Robust z
(Intercept) 110.7323  0.7856546 140.9428   0.7905238 140.0747
```

Robust estimation: software

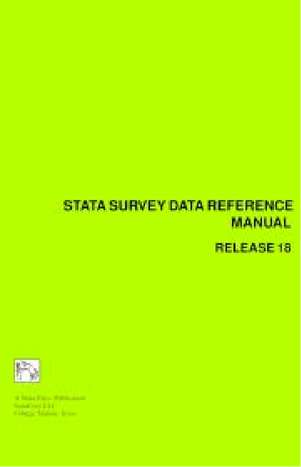
cran/survey

Package "Survey"

"SVY" prefix

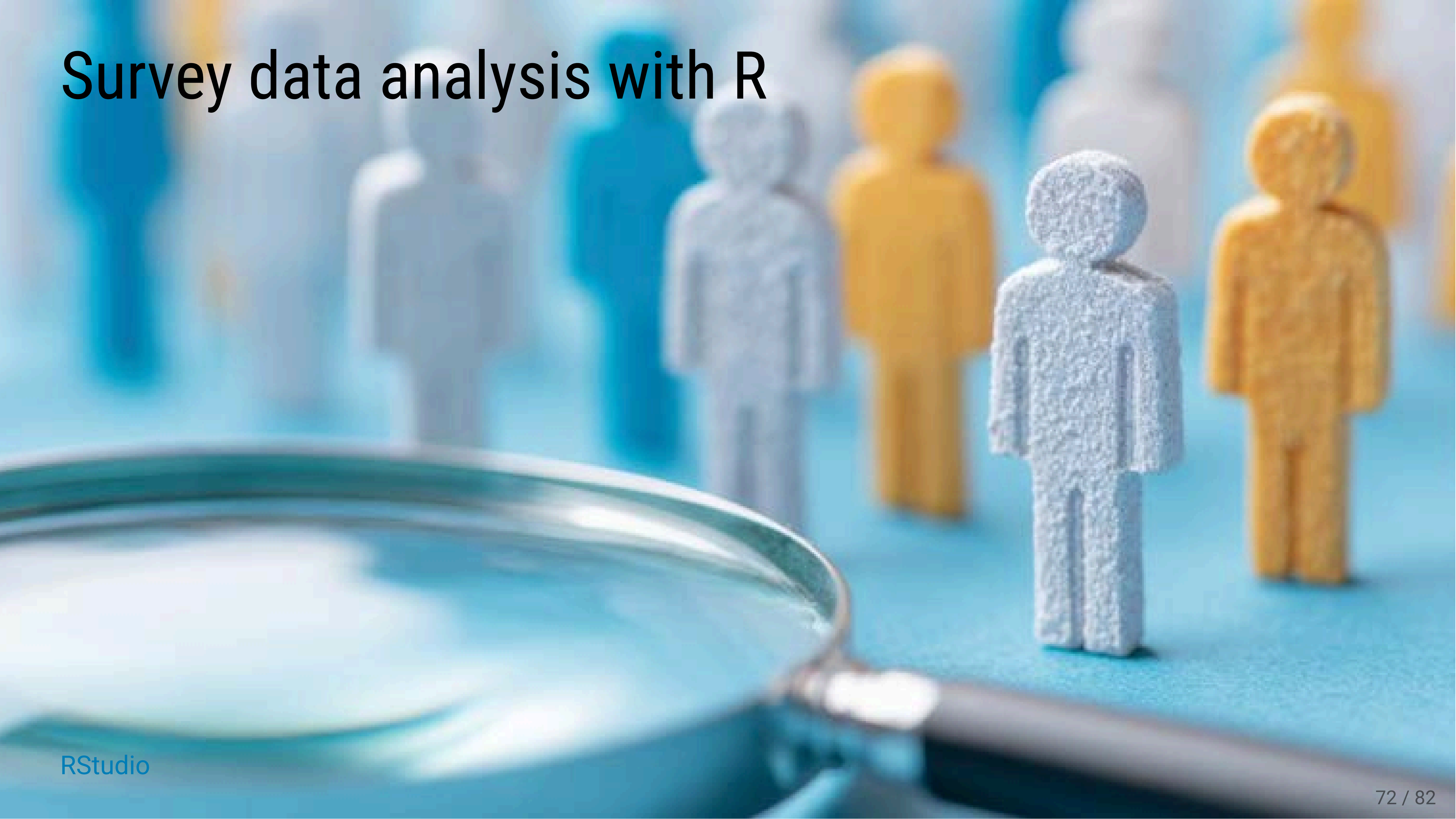
Book: 

Video:

SVY: 

Video:

Survey data analysis with R



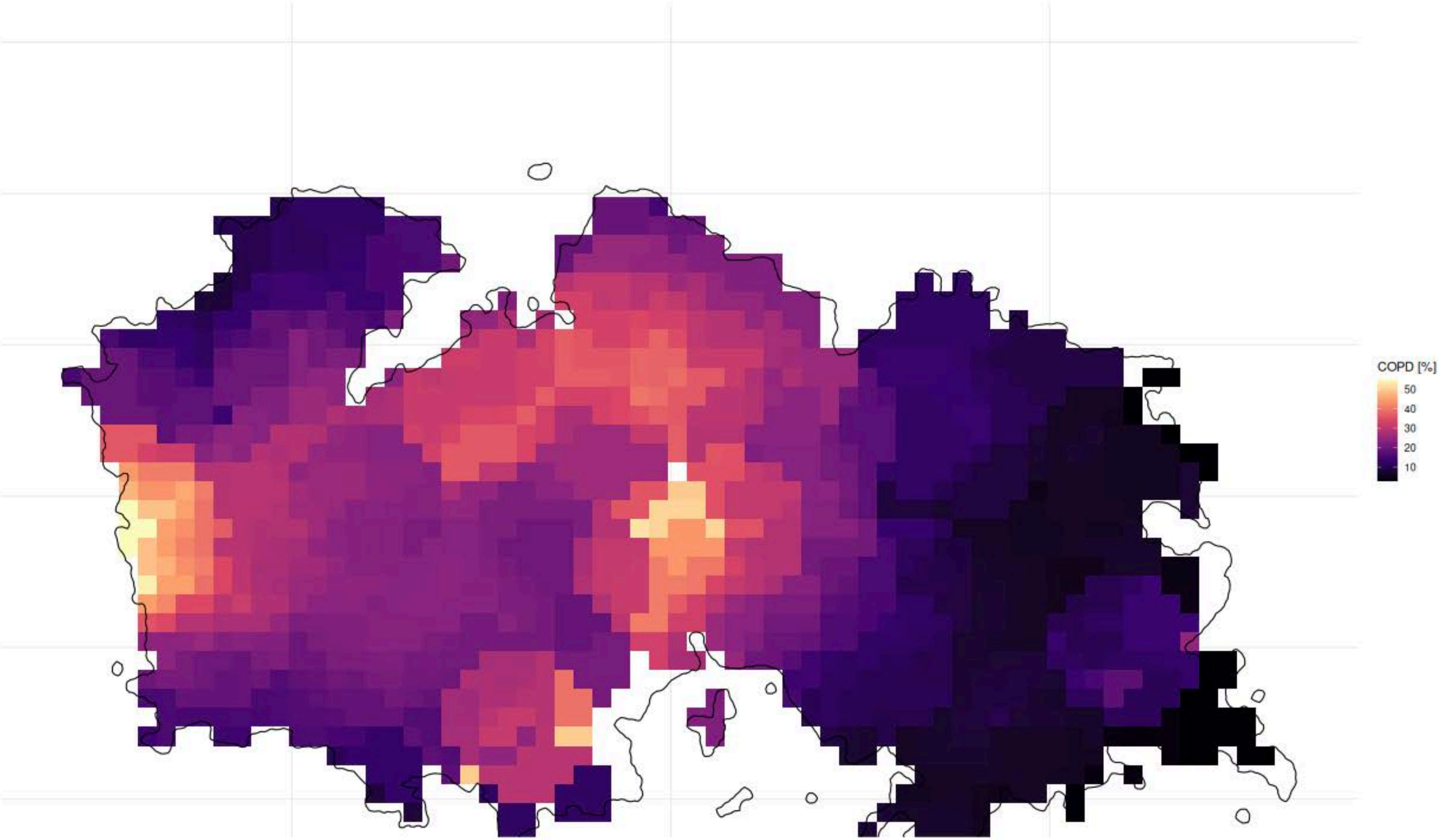
RStudio

Answering our research questions

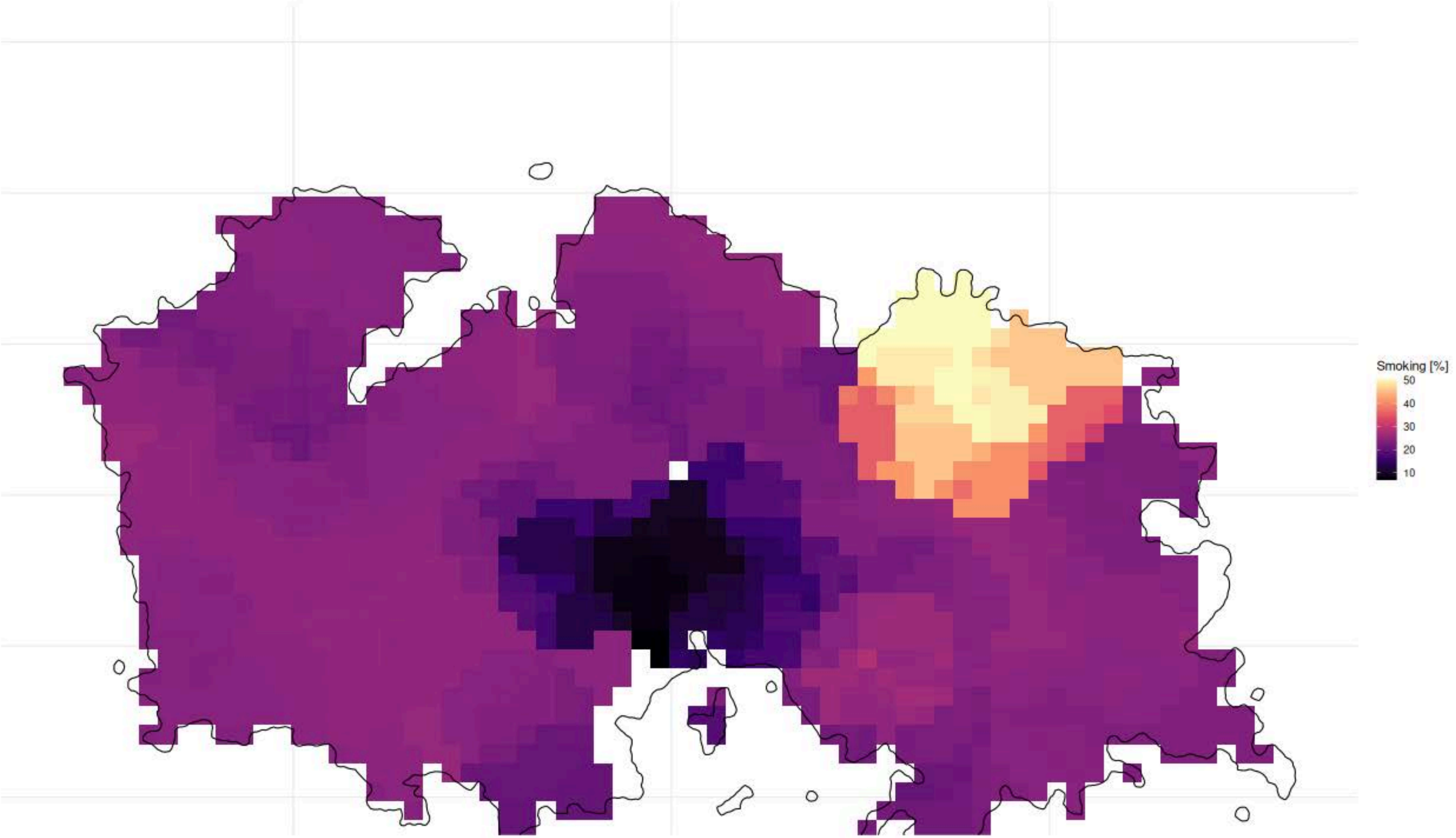
```r
copd <- function(sex, curr_smoking, age, popgroup, global_wquint, energy_cook, powerstation_coal, pm2
                 lbound = LBOUND_COPD, ubound = UBOUND_COPD, random = RANDOMVAR) {
  if (!is.na(sex)) {   # Excludes households with no members
    if (sex == "Female") {
      r2 <- case_when(age<15~1, age<45~1.94, age<55~2.66, age<65~7.92, age<75~12.96, TRUE~16.02)
      r3 <- case_when(curr_smoking=="Yes"~4.50, TRUE~1)
      r4 <- case_when(popgroup=="Kewhea"~1.17, TRUE~1)
      r5 <- 1.17^(global_wquint - 1)
    } else {
      r2 <- case_when(age<15~1, age<45~1.65, age<55~2.45, age<65~6.91, age<75~10.40, TRUE~12.05)
      r3 <- case_when(curr_smoking=="Yes"~3.81, TRUE~1)
      r4 <- case_when(popgroup=="Kewhea"~1.20, TRUE~1)
      r5 <- 1.22^(global_wquint - 1)
    }
    r6 <- case_when((energy_cook=="Wood" | energy_cook=="Coal")~4, TRUE~1)
    r7 <- case_when(powerstation_coal==1~2, powerstation_coal>1~2, TRUE~1)
    r8 <- 1.12^(pm25 - 5)     # ORIGINA ARTIGLE: COEFFICINT WAS 1.07 rather than 1.12

    risk <- r2 * r3 * r4 * r5 * r6 * r7 * r8
    return(risk * rtruncnorm(1, a = lbound, b = ubound, mean = 1, sd = random))
  } else {
    return(NA)
  }
}
```

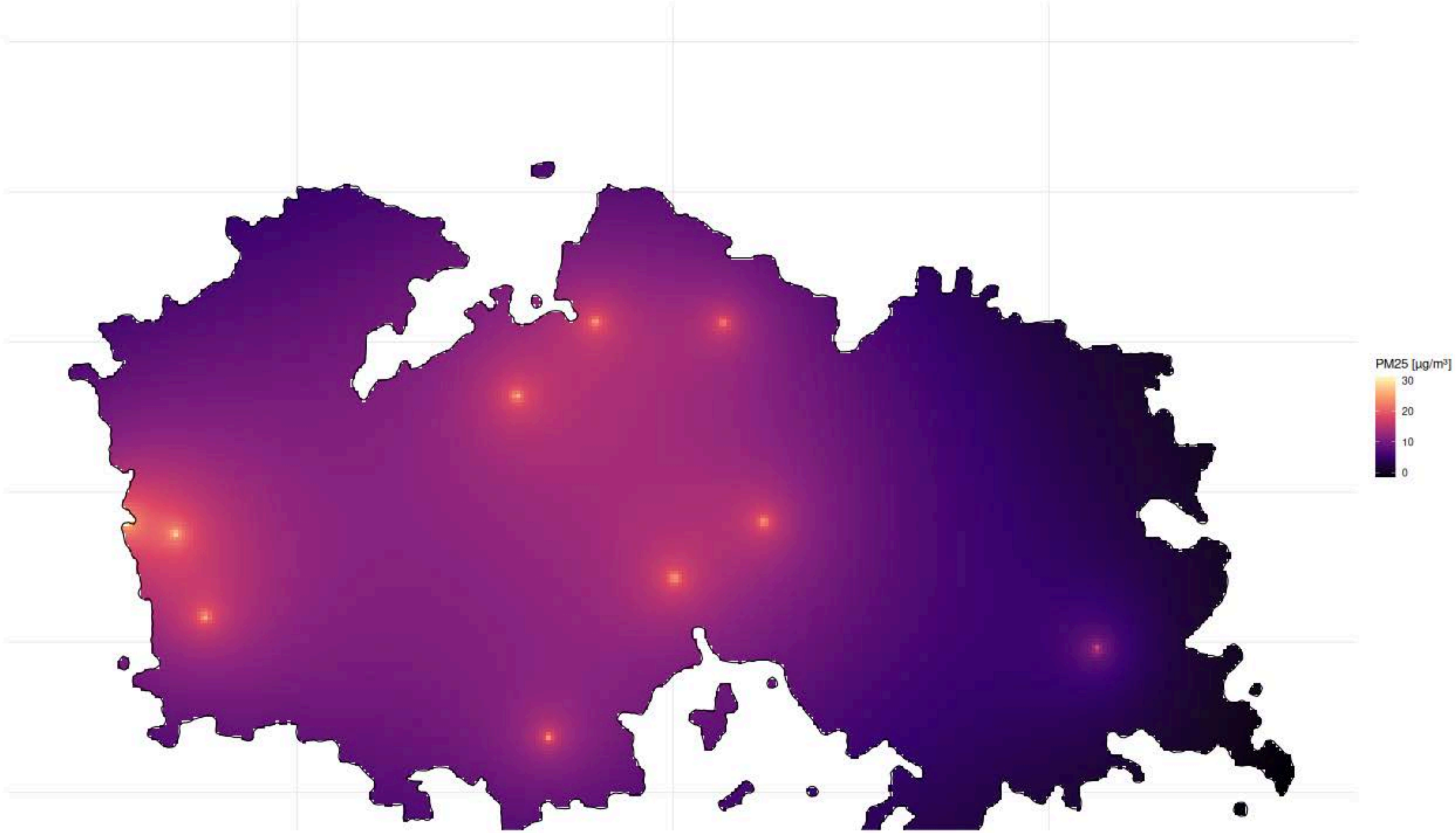
Nacul, Soljak, and Meade (2007)



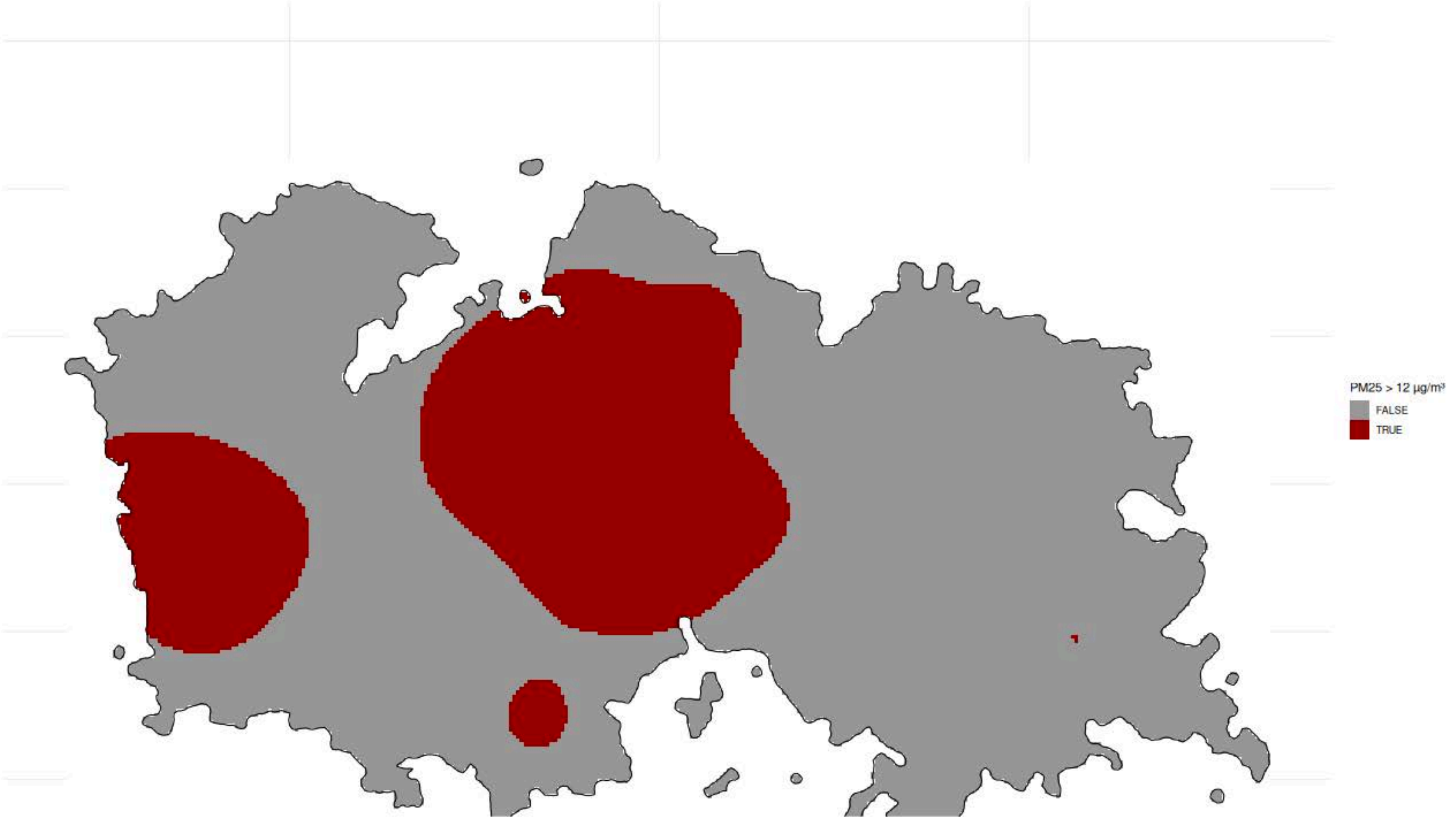
Prevalence of COPD (population 15+)



Prevalence of current smoking (population 15+)



Air concentration of pm2.5 (Annual average)



Areas with annual average concentration of pm2.5 > 12 ug/m3

From questions, to sampling to estimates: a summary

References

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).

Nacul, L. C., M. Soljak, and T. Meade (2007). "Model for Estimating the Population Prevalence of Chronic Obstructive Pulmonary Disease: Cross Sectional Data from the Health Survey for England". In: *Population Health Metrics* 5.1, p. 8. ISSN: 1478-7954. DOI: 10.1186/1478-7954-5-8.

Wittemberg, M. (2009). *Weights: Report on NIDS Wave 1. Technical Paper No. 2.* Southern Africa Labour and Development Research Unit.

Wittenberg, M. and N. Branson (2021). "Creating Household Weights for NIDS-CRAM". In: *SALDRU Working Papers*. (Visited on Sept. 26, 2025).

https://forms.gle/VUhsHHgBmgnR6ZFC9

Thank You!