# Using raking to reweight the New Hampshire BRFSS for substate estimates

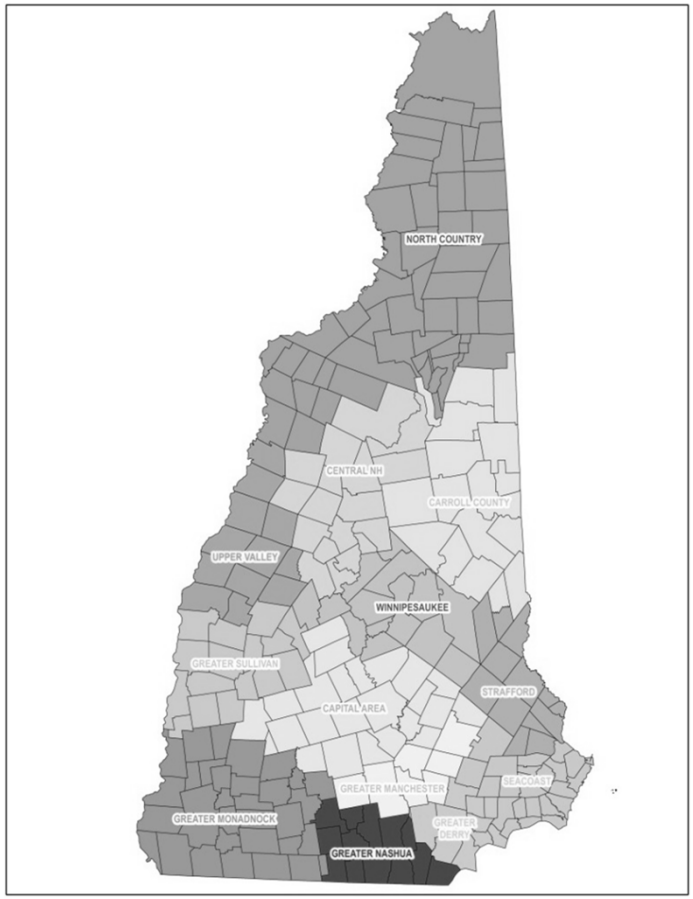
Michael Laviolette, PhD, MPH
Public Health Statistician
New Hampshire Department of Health and Human Services
Bureau of Public Health Statistics and Informatics

Northeast Epidemiology Conference
October 19, 2017
Northampton, MA

# New Hampshire regions for surveillance

**Counties and Major Cities**


**Public Health Regions**




Objective: weight New Hampshire BRFSS for direct substate estimates

# Raking

- Adjusts weights of sample survey so that sample demographics match those of target population
- Introduced with 2011 BRFSS, along with surveying mobile phones
- Used to adjust estimates for undercoverage and response bias
- Methodology dates to 1940's, but only feasible on large scale with modern computing
- Also known as "iterative proportional fitting" or "sample balancing"

# Simple example of raking

- Employee satisfaction survey of 200 employees out of 10,000
- Sample was simple random sample, with each respondent assigned weight of 50
- Population of employees consists of 3800 men, 6200 women
- Of all employees, 1600 are under age 30, 6800 are age 30-44, and 1600 are age 45 and older
- Use raking to make total sample weights for age and sex, respectively, match those of population

# Simple example of raking (2)

| Sex/Age | < 30 | 30-44 | 45+ | Row Total | Target |
|---|---|---|---|---|---|
| Male | 800 | 3800 | 1000 | 5600 | 3800 |
| Female | 800 | 3000 | 600 | 4400 | 6200 |
| Col Total | 1600 | 6800 | 1600 | Initial | |
| Target | 2000 | 5000 | 3000 | | |

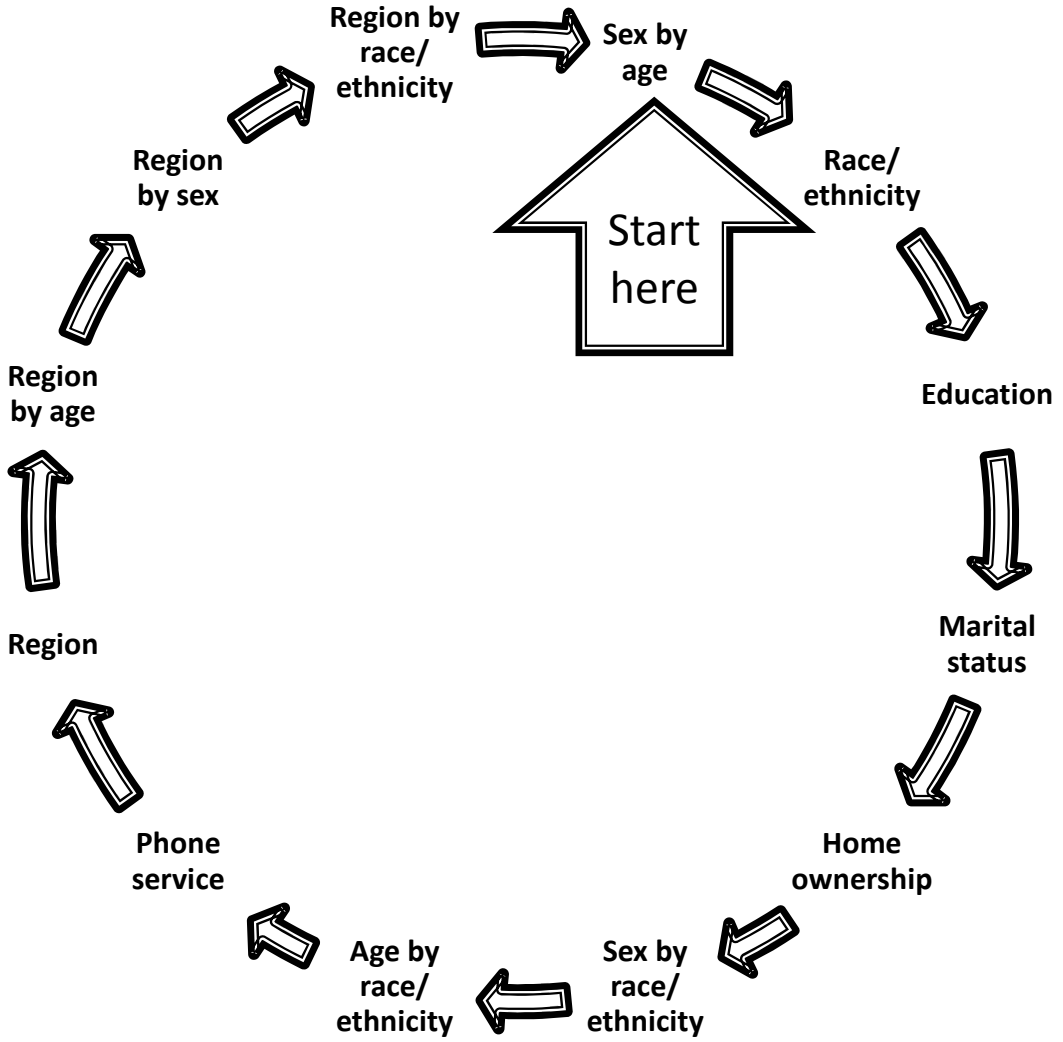
| Sex/Age | < 30 | 30-44 | 45+ | Row Total | Target |
|---|---|---|---|---|---|
| Male | 650.08 | 1894.38 | 1335.75 | 3880.21 | 3800 |
| Female | 1349.92 | 3105.62 | 1664.25 | 6119.79 | 6200 |
| Col Total | 2000 | 5000 | 3000 | Iteration 1 | |
| Target | 2000 | 5000 | 3000 | | |

| Sex/Age | < 30 | 30-44 | 45+ | Row Total | Target |
|---|---|---|---|---|---|
| Male | 635.29 | 1854.65 | 1310.67 | 3800.61 | 3800 |
| Female | 1364.71 | 3145.35 | 1689.33 | 6199.39 | 6200 |
| Col Total | 2000 | 5000 | 3000 | Iteration 2 | |
| Target | 2000 | 5000 | 3000 | | |

Adjust row and column totals so that all totals are within specified tolerance of targets

# Raking on multiple dimensions

Basic principle is same as with two dimensions (or "margins"), but computationally intensive.
Dimensions are adjusted one at a time.
Process continues up to 125 times, or until total weights of each dimension converge to population estimates.



# Raking dimensions

CDC

1 Sex by age
2 Race/Hispanic origin
3 Education
4 Marital status
5 Home tenure
6 Sex
7 Age
8 Telephone service
9 Region
10 Region by age
11 Region by sex
12 Region by race/Hispanic
13 County
14 County by race/Hispanic
15 County by age
16 County by sex

NH

1 Sex by age
2 Race/Hispanic origin
3 Education
4 Marital status
5 Home tenure
6 Sex
7 Age
8 Telephone service
9 Public Health Region (PHR)
10 PHR by age
11 PHR by sex
12 PHR by race/Hispanic
13 County*
14 County* by race/Hispanic
15 County* by age
16 County* by sex
17 County* by education
18 County* by marital status

*with Manchester and Nashua broken out

# Needed for raking

- State raking report from CDC
  - Control totals for all raking dimensions
  - Convergence information, including number of iterations
- Software for preprocessing and raking
  - R used for preprocessing (creating and collapsing variables, aggregating control totals)
  - SAS raking macro
    - Modified by CDC from original developed by Abt Associates
    - R “survey” package will also rake
- Population estimates for control totals

# Approach

1. Reproduce CDC raking
2. Impute towns where needed
3. Choose additional or alternate dimensions for raking
4. Create variables in survey data corresponding to chosen dimensions, collapsing as needed
5. Construct target (control) totals for each dimension using population estimates
6. Modify SAS macro as needed and run
7. Once raking converges, fine-tune to reduce variation in new weights

# Reproduce CDC raking

- Adjust initial weights for dual landline-cell frame
- Truncate weights so that overly large weights don’t enter the raking
  - Split by region and telephone service type
  - CDC documentation not sufficiently detailed
- Run SAS raking macro
- Check resulting weights against final weights from survey data

# Impute towns

- State-added question “What town do you live in?”
  - Town = County subdivision in New England states
- Respondent’s town is identified as stated
  - If not available, primary zip code used
  - If neither is available, town imputed by hotdeck method from county imputed by CDC

# Collapsing categories

- Rule of thumb: cells should have minimum count of 25
  - Sparsity in table can cause lack of convergence
- PHR by race/Hispanic origin: 26 categories collapsed to 15

| HSDM Raking Margin 6: PHR by race/Hispanic origin | |
|---|---|
| HSDM06 | Frequency |
| North Country, All races | 422 |
| Upper Valley, All races | 237 |
| Central NH, All races | 161 |
| Carroll County, All races | 344 |
| Greater Sullivan, All races | 318 |
| Winnipesaukee, All races | 401 |
| Strafford County, All races | 710 |
| Greater Monadnock, All races | 574 |
| Capital Area, All races | 739 |
| Greater Nashua, WH NH | 873 |
| Greater Nashua, not WH NH | 72 |
| Greater Manchester, WH NH | 913 |
| Greater Manchester, not WH NH | 66 |
| South Central, All races | 544 |
| Seacoast, All races | 648 |

WH NH = “White, non-Hispanic”

# Control totals

| third_margin | Output Weight Sum of Weights | Target Total |
|---|---|---|
| Less than HS | 87,625 | 87,619 |
| HS Grad | 314,744 | 314,670 |
| Some College | 330,012 | 329,989 |
| College Grad | 344,258 | 344,361 |

| sixth_margin | Output Weight Sum of Weights | Target Total |
|---|---|---|
| Male, ALL RACES | 527,605 | 527,605 |
| Female, ALL RACES | 549,033 | 549,033 |

All margin totals = 1,076,638

# Constructing control totals

- Need population estimates
  - Use same control total as CDC whenever raking on same dimension
- Population estimate sources
  - Census publishes annual population estimates at county level by age group, race, Hispanic origin
  - NH purchases town-level data from Claritas
  - CDC also purchases data from Claritas
  - American Community Survey also used
    - For town level, need five-year summary tables

# Control totals consistency

| ninth_margin | Target Total |
|---|---|
| Region 01 | 89,286 |
| Region 02 | 97,187 |
| Region 03 | 101,002 |
| Region 04 | 322,097 |
| Region 05 | 120,174 |
| Region 06 | 243,829 |
| Region 07 | 103,063 |

| eleventh_margin | Target Total |
|---|---|
| Region 01,Male | 43,663 |
| Region 01,Female | 45,623 |
| Region 02,Male | 47,215 |
| Region 02,Female | 49,972 |
| Region 03,Male | 50,003 |
| Region 03,Female | 50,999 |
| Region 04,Male | 158,444 |
| Region 04,Female | 163,653 |
| Region 05,Male | 58,809 |
| Region 05,Female | 61,365 |
| Region 06,Male | 119,805 |
| Region 06,Female | 124,024 |
| Region 07,Male | 49,666 |
| Region 07,Female | 53,397 |

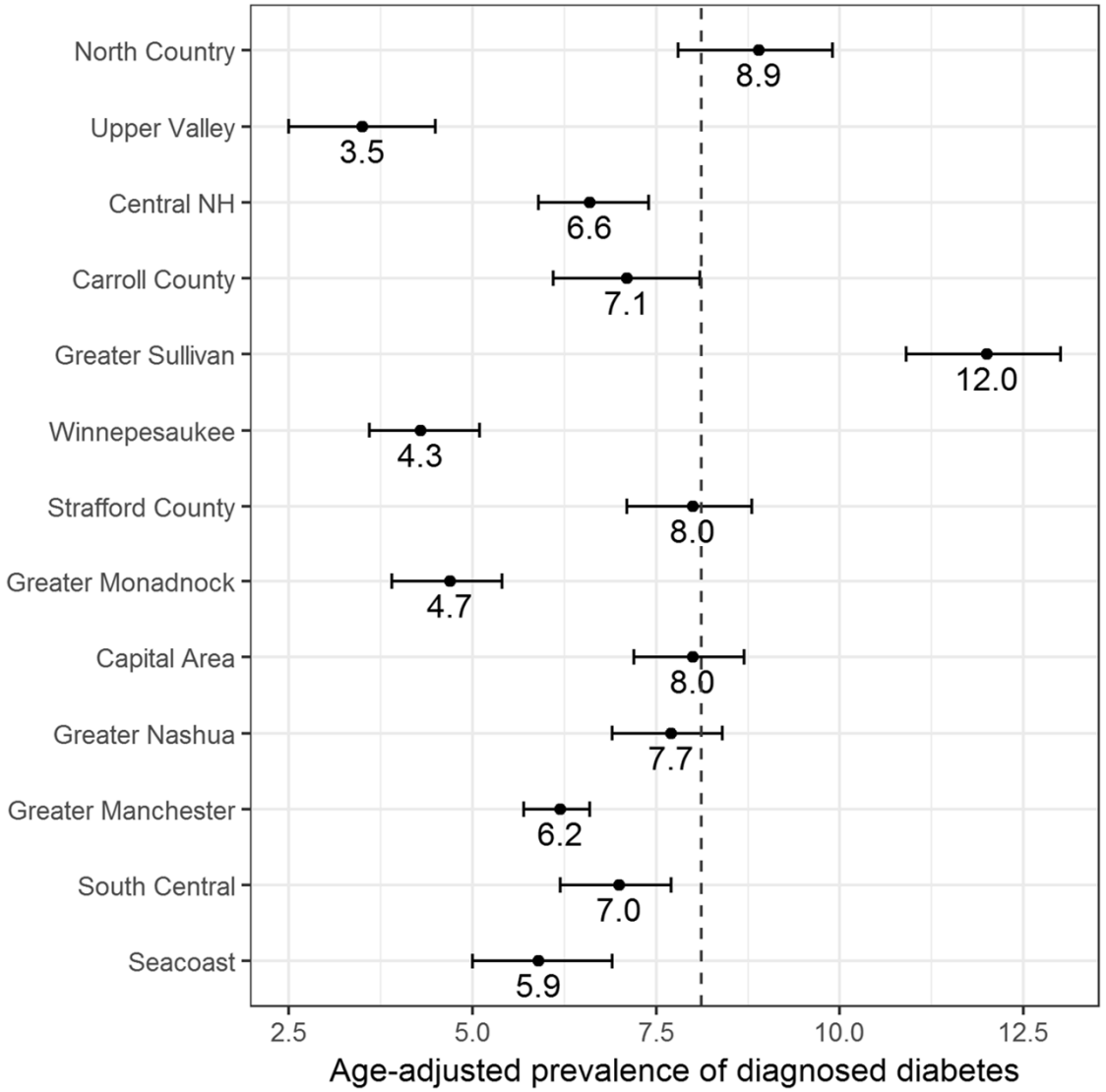
| sixth_margin | Target Total |
|---|---|
| Male, ALL RACES | 527,605 |
| Female, ALL RACES | 549,033 |

- All margins must add to population total
- All corresponding population totals must match, or raking will not converge
- Sometimes necessary to adjust totals to match (raking within raking)

# Final processing

- R used for data pre-processing
  - Add-on package “VIM” for hotdeck imputation
  - Function “interaction” for composite dimensions
  - Functions from “tidyverse” collection of packages for collapsing categories and aggregating population data
  - Custom function for auxiliary adjustments
- Export data for SAS raking macro
- Run macro and troubleshoot in case of nonconvergence
- Convergence took 62 iterations for 2015 data and ran about 15 minutes

# Application



# Conclusions

- Raking produces a single working data set to produce direct prevalence estimates for NH major cities, counties, and public health regions
- Raking a state's BRFSS data for substate estimates is feasible, though extremely detailed
- Interest in small-area methods for model-based estimates by smaller geographies like town and census tract

# References


Addinsoft corp. (2013). Raking a survey sample using XLSTAT. http://www.xlstat.com/en/learning-center/tutorials/raking-a-survey-sample-using-xlstat.html

Battaglia MP et al. (2009). Practical considerations in raking survey data. http://www.abtassoc.us/presentations/raking_survey_data_2_JOS.pdf

Lumley T (2017). "survey: analysis of complex survey samples," R package version 3.32. http://r-survey.r-forge.r-project.org/survey/index.html

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

# Discussion

Michael Laviolette PhD MPH
State of New Hampshire
Department of Health and Human Services
Bureau of Public Health Statistics and Informatics
michael.laviolette@dhhs.nh.gov