Sampling Frames for Seroprevalence Study

Michael Fay

4/20/2020

# Overview

# US Census Data Sources

I wanted to get data by county and race. From the best I would tell from the US Census website, the most recent publically available data with that level of detail is 2018. So I got that. Here is the link for it (accessed April 16, 2020)

<https://www2.census.gov/programs-surveys/popest/datasets/2010-2018/counties/asrh/cc-est2018-alldata.csv>

with the file description

<https://www2.census.gov/programs-surveys/popest/technical-documentation/file-layouts/2010-2018/cc-est2018-alldata.pdf>

We can use this data to make tables by State (and DC), county, age, race.

I also needed to get the FIPS codes.

<https://www2.census.gov/programs-surveys/popest/geographies/2018/state-geocodes-v2018.xlsx>

## Description and Exploration of Census Data

To make sure I understand the age codes, check 2018 data against this table (see page 7)

<https://www2.census.gov/programs-surveys/popest/tables/2010-2018/counties/asrh/PEPAGESEX.pdf>

## AGECAT AGEGRP POP POP\_MILLIONS  
## 1 All 0 327167434 327.167434  
## 2 0 - 4 1 19810275 19.810275  
## 3 5 - 9 2 20195642 20.195642  
## 4 10 - 14 3 20879527 20.879527  
## 5 15 - 19 4 21097221 21.097221  
## 6 20 - 24 5 21873579 21.873579  
## 7 25 - 29 6 23561756 23.561756  
## 8 30 - 34 7 22136018 22.136018  
## 9 35 - 39 8 21563587 21.563587  
## 10 40 - 44 9 19714301 19.714301  
## 11 45 - 49 10 20747135 20.747135  
## 12 50 - 54 11 20884564 20.884564  
## 13 55 - 59 12 21940985 21.940985  
## 14 60 - 64 13 20331651 20.331651  
## 15 65 - 69 14 17086893 17.086893  
## 16 70 - 74 15 13405423 13.405423  
## 17 75 - 79 16 9267066 9.267066  
## 18 80 - 84 17 6127308 6.127308  
## 19 85+ 18 6544503 6.544503

Check sex. It looks like the Census estimates have MALE+FEMALEf=TOTAL.

## AGEGRP POP POP\_MALE POP\_FEMALE MALE\_PLUS\_FEMALE  
## 1 0 327167434 161128679 166038755 327167434

Check that the race categories are distinct and add up to the total. The code is

* WA=White alone
* BA=Black or African American alone
* IA=American Indian and Alaska Native alone
* AA= Asian alone
* NA=Native Hawaiian and Other Pacific Islander alone
* TOM= Two or More Races

Check that when you add up all the 6 categories, you get the total.

## AGEGRP POP\_TOT POP\_ADD\_RACES POP\_WA POP\_BA POP\_IA POP\_AA POP\_NA  
## 1 0 327167434 327167434 250139096 43804319 4147521 19330600 799418  
## POP\_TOM  
## 1 8946480

### Defining Regions

I defined regions by picking groups of states that are contiguous, and such that the total population in each of the 6 groups are approximately equal. First, here are the populations of the states.

Here are the regions (state groups) I ended up with

## STNAME sgrp  
## 1 Connecticut 1  
## 2 Maine 1  
## 3 Massachusetts 1  
## 4 New Hampshire 1  
## 5 Rhode Island 1  
## 6 Vermont 1  
## 7 New Jersey 1  
## 8 New York 1  
## 9 Pennsylvania 1  
## 10 Illinois 2  
## 11 Indiana 2  
## 12 Michigan 2  
## 13 Ohio 2  
## 14 Wisconsin 2  
## 15 Iowa 2  
## 16 Minnesota 2  
## 17 Delaware 3  
## 18 District of Columbia 3  
## 19 Georgia 3  
## 20 Maryland 3  
## 21 North Carolina 3  
## 22 South Carolina 3  
## 23 Virginia 3  
## 24 West Virginia 3  
## 25 Kentucky 3  
## 26 Tennessee 3  
## 27 Kansas 4  
## 28 Missouri 4  
## 29 Florida 4  
## 30 Alabama 4  
## 31 Mississippi 4  
## 32 Arkansas 4  
## 33 Louisiana 4  
## 34 Oklahoma 4  
## 35 Nebraska 5  
## 36 North Dakota 5  
## 37 South Dakota 5  
## 38 Texas 5  
## 39 Arizona 5  
## 40 Colorado 5  
## 41 Idaho 5  
## 42 Montana 5  
## 43 New Mexico 5  
## 44 Utah 5  
## 45 Wyoming 5  
## 46 Nevada 6  
## 47 Alaska 6  
## 48 California 6  
## 49 Hawaii 6  
## 50 Oregon 6  
## 51 Washington 6

and here are the total populations in the 6 state groups (because Census uses the term REGION, I use sgrp for State Group)

## sgrp tot\_pop  
## 1 1 56111079  
## 2 2 55699207  
## 3 3 55261495  
## 4 4 49828565  
## 5 5 53791418  
## 6 6 56475670

# Proportion Tables on Adults (age 20 or more years)

Our study is based only on adults (18 and up), but Census data gives ages in 5 year age groups. So we split the AGEGRP=4 (15 to 19 year olds) population values into two parts, putting 3/5 into a group defined AGEGRP=4.0 (15,16,17) and 2/5 into a group defined as AGEGRP=4.6 (18,19). Then we create the adult data set using AGEGRP=4.6 and higher only.

We change the data so that it only has 3 age groups, 18-44, 45-69, and 70+. Then we fix the data so that each row has a unique value for: State/County, AgeGrp3 (3 categories for ages), sex (M/F), hispanic (NH/H), race (6 categories). Because there are 3142 unique state/county values, the number of rows in the data set is 3142*3*2*2*6 = 226224. For each row we have a column called “pop” for the population, and a column called “R10.000” for the proportion of the adult population times 10,000.

Here are the first few rows of the adult data file:

## STATE COUNTY STNAME CTYNAME AGEGRP sex hispanic race pop  
## 1 1 1 Alabama Autauga County 1 M NH WA 6678.0  
## 2 1 1 Alabama Autauga County 2 M NH WA 6700.0  
## 3 1 1 Alabama Autauga County 3 M NH WA 2193.0  
## 4 1 3 Alabama Baldwin County 1 M NH WA 25990.8  
## 5 1 3 Alabama Baldwin County 2 M NH WA 30682.0  
## 6 1 3 Alabama Baldwin County 3 M NH WA 13105.0  
## R10.000 AGEGRPNAME RACENAME HISPNAME  
## 1 0.2633035 18-44 White Alone Not Hispanic  
## 2 0.2641709 45-69 White Alone Not Hispanic  
## 3 0.0864667 70+ White Alone Not Hispanic  
## 4 1.0247782 18-44 White Alone Not Hispanic  
## 5 1.2097452 45-69 White Alone Not Hispanic  
## 6 0.5167105 70+ White Alone Not Hispanic

## Check DA3 Data set

We check the DA3 database. First check tables by race. We check it two ways.

## const POP\_WA POP\_BA POP\_IA POP\_AA POP\_NA POP\_TOM  
## 1 1 197029928 32690602 2979521 15385539 580589.6 4957479

## [1] 197029928

## race sum  
## 1 AA 15385538.8  
## 2 BA 32690601.8  
## 3 IA 2979520.8  
## 4 NA 580589.6  
## 5 TOM 4957478.8  
## 6 WA 197029927.6

Now check by age group.

## [1] TRUE

## AGEGRP POP R10.000  
## 1 1 117288129 4624.495  
## 2 2 100991228 3981.932  
## 3 3 35344300 1393.573

Check by state

## STNAME POP R10.000  
## 1 Alabama 3796865.0 149.70469  
## 2 Alaska 554750.2 21.87297  
## 3 Arizona 5523286.6 217.77490  
## 4 Arkansas 2310787.8 91.11089  
## 5 California 30551645.0 1204.60549  
## 6 Colorado 4427360.6 174.56418  
## 7 Connecticut 2830480.6 111.60160  
## 8 Delaware 762241.2 30.05403  
## 9 District of Columbia 568503.0 22.41522  
## 10 Florida 17068454.8 672.98354  
## 11 Georgia 8011377.0 315.87657  
## 12 Hawaii 1117656.6 44.06752  
## 13 Idaho 1309164.4 51.61839  
## 14 Illinois 9886288.0 389.80149  
## 15 Indiana 5121246.2 201.92305  
## 16 Iowa 2419774.8 95.40809  
## 17 Kansas 2203938.6 86.89799  
## 18 Kentucky 3458213.0 136.35215  
## 19 Louisiana 3566426.2 140.61883  
## 20 Maine 1086856.4 42.85312  
## 21 Maryland 4700361.6 185.32820  
## 22 Massachusetts 5509592.4 217.23496  
## 23 Michigan 7825433.8 308.54510  
## 24 Minnesota 4310116.6 169.94143  
## 25 Mississippi 2277599.2 89.80232  
## 26 Missouri 4749019.8 187.24672  
## 27 Montana 832548.2 32.82613  
## 28 Nebraska 1451530.6 57.23167  
## 29 Nevada 2351109.2 92.70071  
## 30 New Hampshire 1095885.4 43.20912  
## 31 New Jersey 6963027.0 274.54170  
## 32 New Mexico 1612987.8 63.59769  
## 33 New York 15451595.8 609.23322  
## 34 North Carolina 8070683.2 318.21492  
## 35 North Dakota 579158.8 22.83536  
## 36 Ohio 9091579.8 358.46734  
## 37 Oklahoma 2986259.0 117.74371  
## 38 Oregon 3316609.0 130.76891  
## 39 Pennsylvania 10137036.0 399.68811  
## 40 Rhode Island 846385.4 33.37171  
## 41 South Carolina 3971590.6 156.59385  
## 42 South Dakota 663854.6 26.17479  
## 43 Tennessee 5265787.0 207.62207  
## 44 Texas 21319147.2 840.58196  
## 45 Utah 2232833.0 88.03725  
## 46 Vermont 506214.0 19.95926  
## 47 Virginia 6638571.6 261.74891  
## 48 Washington 5874942.6 231.64017  
## 49 West Virginia 1440807.8 56.80889  
## 50 Wisconsin 4532988.2 178.72892  
## 51 Wyoming 443086.2 17.47022

## [1] TRUE

# Two Dummy Demographic Data Sets

I create two dummy data sets with 1000 people in each. The first data set is an actual probability in proportion to size sampling. The second data set is each cell (county, age, sex, race, ethnicity) has equal probability. This will give an extreme oversampling of minorities and rural areas.

Check Tables AgeGRP:

## AGEGRP per10.000 dummyDemo1\_per10.000 dummyDemo2\_per10.000  
## 1 1 4624.495 4530 3290  
## 2 2 3981.932 3930 3440  
## 3 3 1393.573 1540 3270

Tables by Sex:

## sex per10.000 dummyDemo1\_per10.000 dummyDemo2\_per10.000  
## 1 F 5128.345 5020 5320  
## 2 M 4871.655 4980 4680

Tables by Race:

## race per10.000 dummyDemo1\_per10.000 dummyDemo2\_per10.000  
## 1 AA 606.62869 630 1710  
## 2 BA 1288.94134 1440 1480  
## 3 IA 117.47803 80 1720  
## 4 NA 22.89178 10 1710  
## 5 TOM 195.46595 190 1840  
## 6 WA 7768.59421 7650 1540

Tables by HIspanic/Non-Hispanic:

## hispanic per10.000 dummyDemo1\_per10.000 dummyDemo2\_per10.000  
## 1 H 1623.649 1680 5000  
## 2 NH 8376.351 8320 5000

Check Tables State:

## STNAME per10.000 dummyDemo1\_per10.000 dummyDemo2\_per10.000  
## 1 Alabama 149.70469 110 210  
## 2 Alaska 21.87297 0 50  
## 3 Arizona 217.77490 100 60  
## 4 Arkansas 91.11089 60 360  
## 5 California 1204.60549 1260 180  
## 6 Colorado 174.56418 220 220  
## 7 Connecticut 111.60160 160 30  
## 8 Delaware 30.05403 0 10  
## 9 District of Columbia 22.41522 40 10  
## 10 Florida 672.98354 640 330  
## 11 Georgia 315.87657 280 530  
## 12 Hawaii 44.06752 40 20  
## 13 Idaho 51.61839 60 110  
## 14 Illinois 389.80149 440 270  
## 15 Indiana 201.92305 170 260  
## 16 Iowa 95.40809 90 300  
## 17 Kansas 86.89799 120 320  
## 18 Kentucky 136.35215 120 380  
## 19 Louisiana 140.61883 150 170  
## 20 Maine 42.85312 0 40  
## 21 Maryland 185.32820 240 70  
## 22 Massachusetts 217.23496 230 30  
## 23 Michigan 308.54510 310 240  
## 24 Minnesota 169.94143 160 330  
## 25 Mississippi 89.80232 60 180  
## 26 Missouri 187.24672 180 380  
## 27 Montana 32.82613 30 250  
## 28 Nebraska 57.23167 40 240  
## 29 Nevada 92.70071 130 40  
## 30 New Hampshire 43.20912 40 0  
## 31 New Jersey 274.54170 370 100  
## 32 New Mexico 63.59769 50 110  
## 33 New York 609.23322 620 190  
## 34 North Carolina 318.21492 320 380  
## 35 North Dakota 22.83536 20 220  
## 36 Ohio 358.46734 400 270  
## 37 Oklahoma 117.74371 90 220  
## 38 Oregon 130.76891 130 130  
## 39 Pennsylvania 399.68811 460 220  
## 40 Rhode Island 33.37171 10 0  
## 41 South Carolina 156.59385 250 220  
## 42 South Dakota 26.17479 20 160  
## 43 Tennessee 207.62207 150 360  
## 44 Texas 840.58196 830 640  
## 45 Utah 88.03725 100 120  
## 46 Vermont 19.95926 0 60  
## 47 Virginia 261.74891 230 390  
## 48 Washington 231.64017 270 150  
## 49 West Virginia 56.80889 50 140  
## 50 Wisconsin 178.72892 140 230  
## 51 Wyoming 17.47022 10 70

## Adding Urban/Rural to County Values

On the Census website, it states that every decenial census they define each county into three categories:

* mostly urban (total=1,253 colored BLUE)
* mostly rural (total=1,185 colored GREEN)
* completely rural (total=704 colored ORANGE)

So we get that data from the 2010 Census at (May 15, 2020 accessed):

<http://www2.census.gov/geo/docs/reference/ua/County_Rural_Lookup.xlsx>

Then Jing Wang took that data, and added an extra column to the master file. The code for that merging and the checks are below (not shown). It is not rerun in this R markdown file.

During the merging process, Jing noticed that two county names (shown in the following table) were not quite consistent across the two source data. She kept the names in the original Census3AgeGroupsWithNames.csv file and updated the corresponding names in the County\_Rural\_Lookup.xlsx file to complete the data merging.

From datasets, here are the two counties with different names:

* from Census3AgeGroupsWithNames.csv  
  “Petersburg Borough, Alaska”,
* from County\_Rural\_Lookup.xlsx “Petersburg Census Area, Alaska”
* from Census3AgeGroupsWithNames.csv  
  “LaSalle Parish, Louisiana”
* from County\_Rural\_Lookup.xlsx “LaSalle Parish, Louisiana”