Ratio of Counts of Undiagnosed+Seropositive Over Diagnosed

Michael Fay

March 30, 2021

## Update from December 16, 2020 version

In response to reviewers, we added two new sections at the end. Plots of the proportion of the final data set wihtin categories (e.g., regions), and estimates of the rates of previously undiagnosed infections as a fraction of the diagnosed cases, by region. Also, the document was edited for clarity.

## Update from December 13 Version

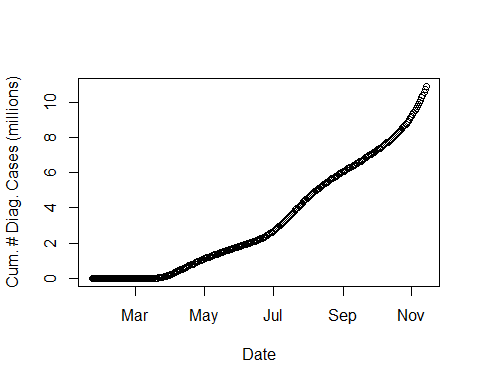
I added a section at the end with weighted population and diagnosed counts. I also added a plot of the weights by collection time.

## Overview Ratio of Counts

We are interested in the ratio of the number of "undiagnosed infections" (US adults previously undiagnosed with SARS-CoV-2 infection by PCR, that are seropositive for SARS-CoV-2 antibodies) over "diagnosed cases" (US adults previously diagnosed by PCR) during the time period of the study. Our main estimate of the paper is the rate of undiagnosed infections per total undiagnosed US adults over the study collection time. Loosely speaking, our estimate of the ratio for US adults uses

What makes this estimator difficult is that the data are collected over a few months during which the numbers diagnosed as changing dramatically. We describe the data used to estimate the number previously diagnosed and undiagnosed at any specific date in two separate sections. Then we describe how the time element is used to get our final estimate.

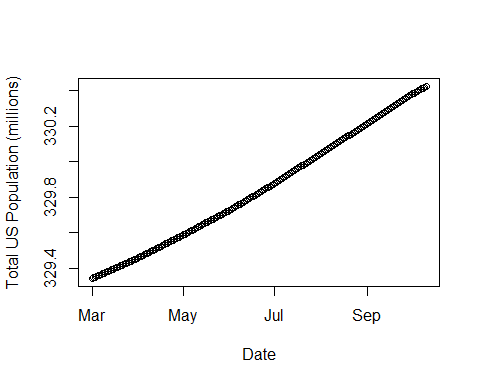
## Number Previously Diagnosed

To get the total number of cumulative cases for any date during the study period, we use The COVID Tracking project (covidtracking.com), a nd code derived from Chow et al (2020), to get, for any date of the study, the total cumulative number of virologically-confirmed cases in the US. Here are the total number of cumulative cases over time (in millions). 

For our analysis, we only want the adult cases. By September 19, 2020 there were 277,285 laboratory-confirmed cases of COVID-19 in children ages 5-17 (Leeb, et al, 2020). On the same date the total number of cases was 6755233. We estimate the number of adult cases on September 19 by subtracting off the number of ages 5-17 (ignoring cases in children under 5). This gives 6477948, which gives the percentage of cases that are adults as 95.9 %. So we approximate the cumulative adult cases at any date t, as 0.959 times the total cumulative cases at t.

## Number Previously Undiagnosed

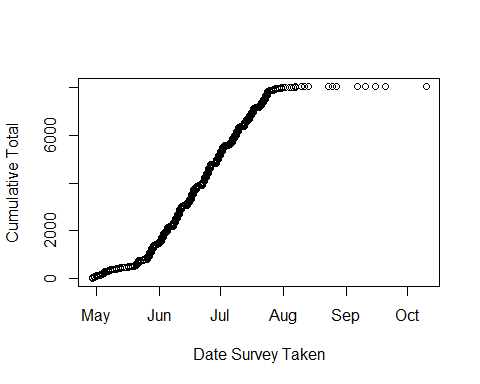
To get the numer of adults previously undiagnosed, we can just subtract the number previously diagnosed by time t from the total number of adults in the US at time t. To get the number of adults in the US at time t, we start with the total number of people in the US at time t based on the US census population clock (US Census, 2020).

Here is the population over time: 

We can use the 2018 data to adjust that estimate (that is the data we are using for the quota sampling). In 2018, the adult population was

253623658 / 327167434 = 77.52 % of the US population. So assume that the proportion of the population that is adult at each date of the study is 77.52 % of the estimated total population at that time.

## Rates During Study

For the final analysis data set we used 8058 individuals. We plot the date that the survey was taken for these individuals: 

Using the methods described previously, the main statistic is the overall rate of undiagnosed US adults that were seropositive per undiagnosed US adults. This rate is an average over the time of the study. The rate is weighted for many variables, but there is no weighting for time. The time of sample collection (and survey taken) is approximately independent of the quota variables, because the quota sampling was done periodically throughout the study so that for each of those variables the marginal distributions are approximately constant over the course of the main part of the study.

Despite the approximate independence of the quota variables and time, there is a large effect of time in terms of seropositivity -- later samples are more likely to be positive. There are two strategies for adusting for this. First, we can pick a timepoint when about half of the sample is collected and calculate the ratio of counts at that time. In that first approach, the population of previously diagnosed and previously undiagnosed adults in the US are estimated at that middle time point. A second approach is to estimate average the population of previously diagnosed and undiagnosed indivudals over the sample survey times of the individuals. We describe each of those approaches in turn.

## Estimating Counts at the Midpoint of the Study

To be precise, we use notation to define the values we are trying to estimate. Let

* the number of US adults that are were previously undiagnosed by PCR by time t.
* the number of US adults that were previously diagnosed by PCR by time t.
* the number of US adults that are seropositive at time t and were previously undiagnosed by PCR at time t.
* the overal estimate of the rate of seropostive individuals out of the population of US adults that were previously undiagnosed. This is a weighted average of binary seroprevalence assay values for the.

Since is a weighted average of the seropostive rate per undiagnosed US adult, one method for estimating the ratio of the counts is to estimate populations at the midpoint of the study. For a midpoint of the study, we have two choices. First, we could use t= July 7, 2020, which is close to the time when half of the study population had been included. This is the time when about half of the non-region 3 states had their data collected. (Recall, region 3 started early in a pilot phase of the study, but the non-region 3 states only participated in the main part of the study when the quota sampling is taking place. Region 3 states are in both the pilot and the main parts of the study.) Alternatively, we can use 2020-06-22 which is the median time of all individuals in the final analysis data set (pilot and main part of the study combined).

Let be the chosen midpoint. Then we want to estimate

where is the total US adult population at . As described earlier, we estimate using the Census values at (which is 329910412 on July 7, 2020, or 329832142 on 2020-06-22) multiplied by 0.7752, the approximate proportion that are adults. Also described eariler, we estimate by taking the total counts of cumulative confirmed cases by PCR at time (which is 2987076 on July 7, 2020, or 2304358 on 2020-06-22 ), and multiplying it by the approximate proportion that are adults, 0.959.

In other words, we estimate when July 7, 2020 by

and when 2020-06-22 by

Here are the estimates and 95% confidence intervals for for the two different threshold defintions:

## estimate LCL UCL   
## 0.04555983 0.02631447 0.06490880

## estimate LCL UCL   
## 0.06568599 0.03321048 0.09404430

Here is a table of the estimates of :

## threshold t\_m estimate lowerCL  
## 1 One.Spike.One.RBD.3SD 07/07/20 4.022192 2.323140  
## 2 One.Spike.One.RBD.3SD 2020-06-22 5.226104 3.018496  
## 3 Any.Single.Positive.4SD 07/07/20 5.799004 2.931945  
## 4 Any.Single.Positive.4SD 2020-06-22 7.534747 3.809527  
## upperCL  
## 1 5.730391  
## 2 7.445597  
## 3 8.302581  
## 4 10.787689

## Estimating Counts by Averaging Over Collection Times

Using the midpoint of the study to estimate the population will give a reasonable estimate if the population is changing approximately linearly during the study time, the cumulative counts of diagnosed adults is changing approximately linearly during the study time, and the data are collected uniformly over the study time. Because we want a method that relies less on those assumptions (i.e., one that would give reasonable estimates if the diagnosed cases increased closer to exponentially than linearly, or one that would give reasonable estimates if a small prortion of the same was collected during an extended pilot period and an extended wrap up period [see above Figure]). If those assumptions are not met, then we may get a biased estimate. An alternative estimator uses the average of the cumulative case counts, averaged over the dates when the individual samples in the final analysis set were collected.

To be precise, we introduce notation to describe the estimator of the count ratio. First, let be the overal estimate of the rate of seropostive individuals out of the population of US adults that were previously undiagnosed. For the 8058 individuals in the final analysis data set, let be the binary seroprevalence assay results (postive=1/negative=0), let be the associated dates that the survey (and sample) was taken, and let be the list of vectors of covariates used in the weighting model. Let be the standardized weight for the th individual, so that Then

By basing the weights only on the covariates, , and responses, , and not on the time values, , we are ignoring the changes in the cumulative diagnosis rates over the course of the study. One way to account for those changes over time, is to take an average of the cumulative diagnosed population over the collection times.

In other words, we use the same estimators of and as described above, but take averages over the study collection times, rather than use a middle time point of the study. Notationally, instead of using

to estimate the ratio of counts, we use

where

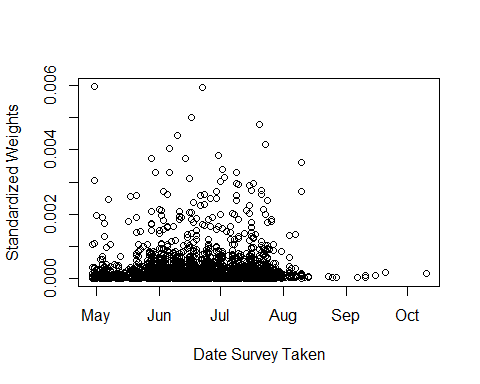
and

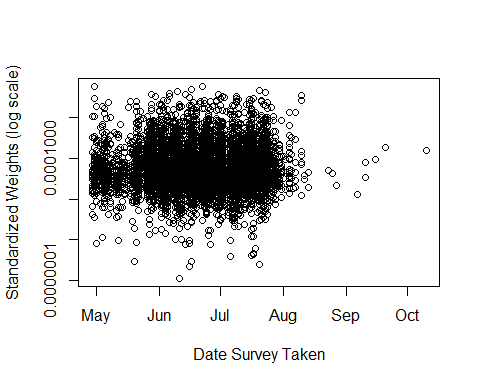
This gives a factor of

which gives estimates and 95% confidence intervals for the rate of undiagnosed infections to diagnosed cases:

## threshold estimate lowerCL upperCL  
## 1 One.Spike.One.RBD.3SD 4.786564 2.764626 6.819388  
## 2 Any.Single.Positive.4SD 6.901040 3.489129 9.880395

## Weighted Average of Population Counts

Here is a plot of the standardized weights by date survey taken: 

Plotting again on the log scale for the standardized weights, we see that there is little correlation:  There is no significant correlation:

##   
## Pearson's product-moment correlation  
##   
## data: as.numeric(D$survey\_taken\_date) and D$stdwgts  
## t = 0.73482, df = 8056, p-value = 0.4625  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.01365044 0.03001594  
## sample estimates:  
## cor   
## 0.008186648

##   
## Pearson's product-moment correlation  
##   
## data: as.numeric(D$survey\_taken\_date) and log10(D$stdwgts)  
## t = 1.1731, df = 8056, p-value = 0.2408  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.008768478 0.034893373  
## sample estimates:  
## cor   
## 0.01306868

So we try the weighted estimates,

where

and

This gives a factor of

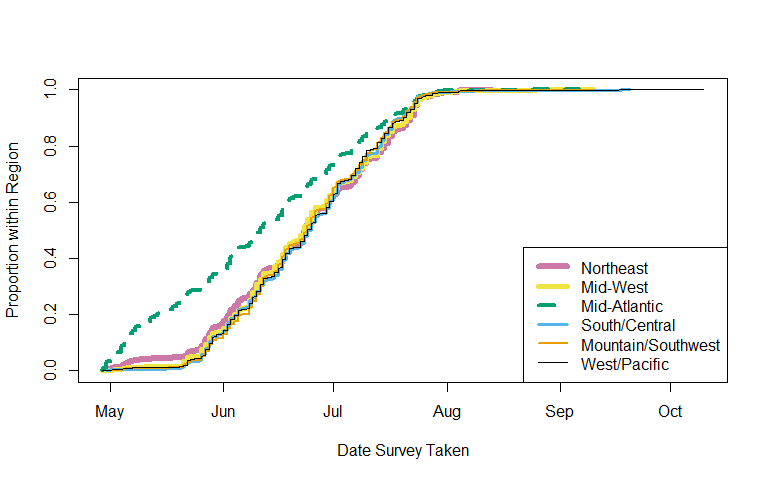
which gives estimates and 95% confidence intervals for the rate of undiagnosed infections to diagnosed cases:

## threshold estimate lowerCL upperCL  
## 1 One.Spike.One.RBD.3SD 4.741785 2.738763 6.755591  
## 2 Any.Single.Positive.4SD 6.836480 3.456487 9.787962

## Data Collection of Time by Covariates

To doublecheck that the data are spread out evenly over the time of the study, we plot the date that the survey was taken for these individuals by different covariates.

### By Region

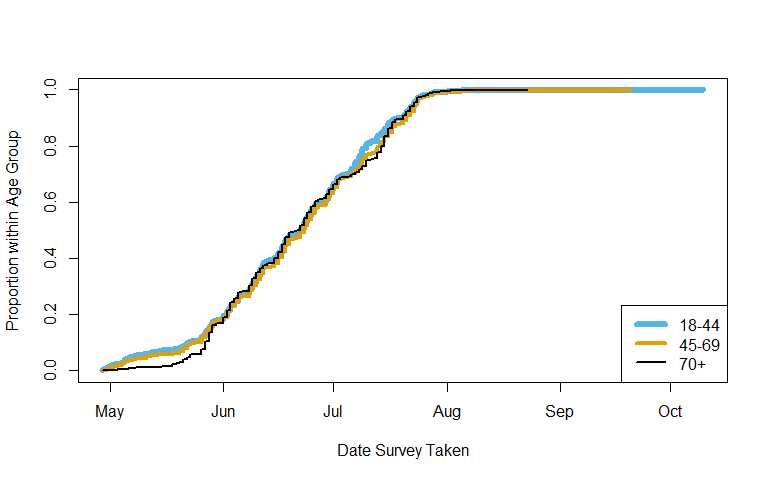


Dates Survey Taken by Region for Final Analysis Data Set

We see that the date when 20% of the participants from the region are included in the study is different for the Mid-Atlantic region.

## Northeast Mid-West Mid-Atlantic   
## "2020-06-03" "2020-06-04" "2020-05-14"   
## South/Central Mountain/Southwest West/Pacific   
## "2020-06-05" "2020-06-06" "2020-06-05"

### By Age Group

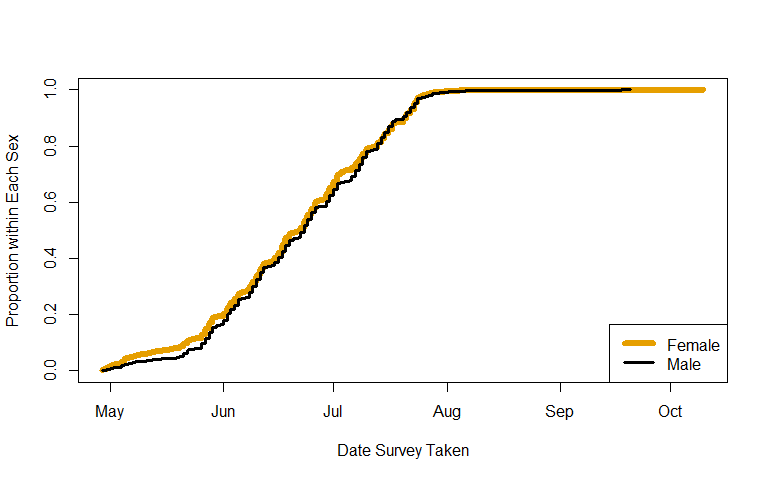


Dates Survey Taken by Age Group for Final Analysis Data Set

Here are the dates when 20% of the participants from each age group are included in the study.

## 18-44 45-69 70+   
## "2020-06-02" "2020-06-02" "2020-06-02"

### By Sex



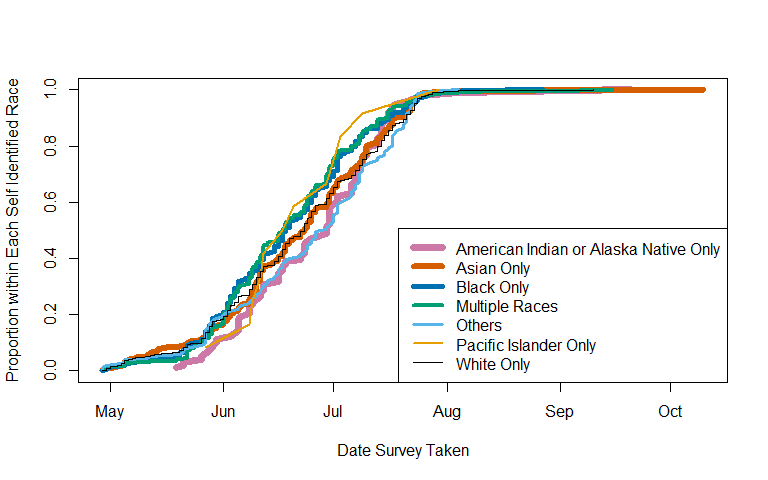
Dates Survey Taken by Sex for Final Analysis Data Set

Here are the dates when 20% of the participants from each sex are included in the study.

## Female Male <NA>   
## "2020-06-01" "2020-06-02" NA

### By Race

Note: "Others" are those that wrote in a race, instead of checking one or more of the specified categories.

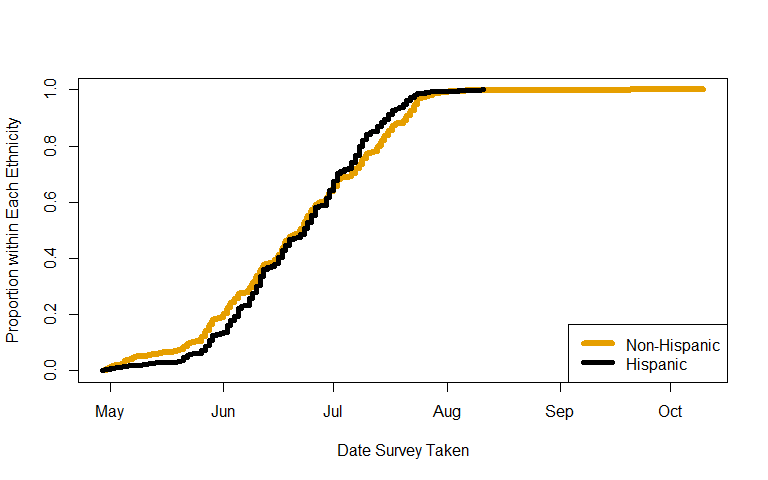


Dates Survey Taken by Race for Final Analysis Data Set

Here are the dates when 20% of the participants from each self-identified race are included in the study.

## American Indian or Alaska Native Only   
## "2020-06-08"   
## Asian Only   
## "2020-06-03"   
## Black Only   
## "2020-06-01"   
## Multiple Races   
## "2020-06-01"   
## Others   
## "2020-06-01"   
## Pacific Islander Only   
## "2020-06-08"   
## White Only   
## "2020-06-02"

### By Ethnicity (Hispanic/Non-Hispanic)

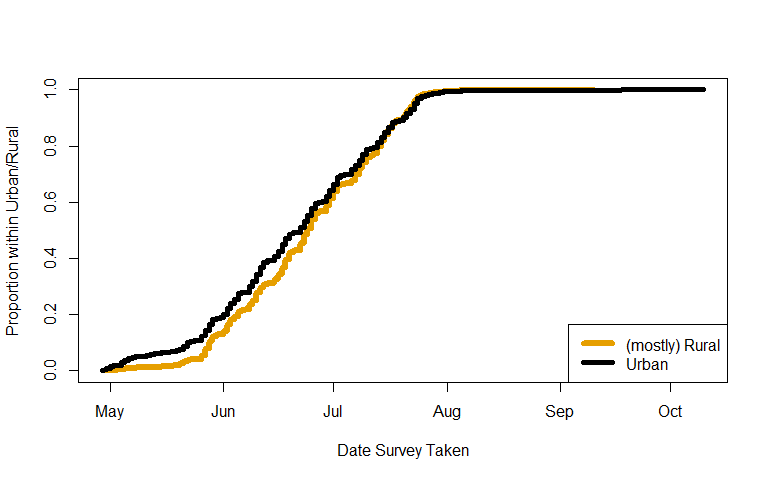


Dates Survey Taken by Ethnicity for Final Analysis Data Set

Here are the dates when 20% of the participants from each ethnicity are included in the study.

## Non-Hispanic Hispanic   
## "2020-06-01" "2020-06-05"

### By Urban/Rural



Dates Survey Taken by Urban/Rural for Final Analysis Data Set

Here are the dates when 20% of the participants from urban or rural are included in the study.

## (mostly) Rural Urban   
## "2020-06-05" "2020-06-01"

## Calculating the Ratio Separately by Region

We repeat the section "Estimating Counts by Averaging Over Collection Times", except we get the average counts for each region.

Expand the notation, adding a superscript to denote the th region. For example, is the population size in region at time , and similarly define as the cumulative number of diagnosed cases in region at time . Also let be the number in the sample from the th region, and be the set of indices from the th region. We use

where

and

We get the US Census estimates of the adult population (18 years old or older) by state on July 1, 2020 from a recent data release on December 22, 2020 (after the original analysis was already completed, data accessed on March 25, 2021):

<https://www2.census.gov/programs-surveys/popest/tables/2010-2020/state/totals/sc-est2020-18+pop-res.xlsx>

We use the previously used total US population by date from the Census population clock to get a multiplier for each date. Let be July 1, 2020, and let

be the total US population at as the proportion of the July 1, 2020 total population (by the US census population clock, see above). Let be the estimated adult US population in region (by the Dec 22, 2020 estimate released by the US Census), and we estimate the adult US population in region at by .

The average adult population in region is . For all regions, in millions is:

## Northeast Mid-West Mid-Atlantic   
## 44.42434 43.32598 43.68070   
## South/Central Mountain/Southwest West/Pacific   
## 39.50593 41.60494 44.08524

The average positively confirmed diagnosed cases in region is . For all regions, per 100,000 is:

## Northeast Mid-West Mid-Atlantic   
## 7.817734 3.826757 3.049454   
## South/Central Mountain/Southwest West/Pacific   
## 3.389338 3.202991 2.727000

The factor for the th region is then

, and for all regions are:

## Northeast Mid-West Mid-Atlantic   
## 55.82508 112.21855 142.24106   
## South/Central Mountain/Southwest West/Pacific   
## 115.55943 128.89402 160.66205

which gives estimates and 95% confidence intervals for the rate of undiagnosed infections to diagnosed cases:

## estimate lowerCL upperCL  
## Northeast 4.204 2.162 6.919  
## Mid-West 1.803 0.290 2.652  
## Mid-Atlantic 12.240 3.648 26.937  
## South/Central 3.470 1.331 5.754  
## Mountain/Southwest 5.807 1.688 12.222  
## West/Pacific 3.013 0.291 6.083

## References

Leeb RT, Price S, Sliwa S, et al. COVID-19 Trends Among School-Aged Children — United States, March 1–September 19, 2020. MMWR Morb Mortal Wkly Rep 2020;69:1410–1415. DOI: <http://dx.doi.org/10.15585/mmwr.mm6939e2external>

US Census (2020). U.S. and World Population Clock at <https://www.census.gov/popclock/> (accessed on December 12, 2020).

US Census (2020b). "Estimates of the Total Resident Population and Resident Population Age 18 Years and Older for the United States, Regions, States, and the District of Columbia: July 1, 2020 (SC-EST2020-18+POP-RES)" <https://www2.census.gov/programs-surveys/popest/tables/2010-2020/state/totals/sc-est2020-18+pop-res.xlsx> (accessed on March 25, 2021).