May 19th report

Mireia Triguero

5/14/2020

Race matching algorithm

Here's the race distribution for everyone in the data

```
# A tibble: 4 x 3
  race     race_total race_pct
  <chr>         <int>    <dbl>
1 asian          4566     2.98
2 hispanic       5981     3.90
3 nh_black       5952     3.88
4 nh_white     136909    89.2
```

House 2010 race staffers

Here's the race distribution for staffers in the house in 2010.

```
# A tibble: 4 x 3
  race_staffer race_total race_pct
  <chr>             <int>    <dbl>
1 asian               521     1.74
2 hispanic           1819     6.08
3 nh_black           1216     4.06
4 nh_white          26384    88.1
```

Here's the distribution in Black house members' offices:

```
# A tibble: 4 x 3
  race_staffer race_total race_pct
  <chr>             <int>    <dbl>
1 asian                75     2.63
2 hispanic            116     4.06
3 nh_black            498    17.4
4 nh_white           2168    75.9
```

Here's the distribution in Hispanic house members' offices:

```
# A tibble: 4 x 3
  race_staffer race_total race_pct
  <chr>             <int>    <dbl>
1 asian                57     3.19
2 hispanic            766    42.9
3 nh_black             26     1.46
4 nh_white            936    52.4
```

Here's the distribution in Asian/PI house members' offices:

```
# A tibble: 4 x 3
  race_staffer race_total race_pct
  <chr>             <int>    <dbl>
1 asian                77    11.9
```

```
2 hispanic             69    10.7
3 nh_black             27     4.18
4 nh_white            473    73.2
```

Here's the distribution in White house members' offices:

```
# A tibble: 4 x 3
  race_staffer race_total race_pct
  <chr>             <int>    <dbl>
1 asian               312     1.27
2 hispanic            868     3.52
3 nh_black            665     2.70
4 nh_white          22807    92.5
```

Senate 2010 race staffers

Here's the race distribution for staffers in the house in 2010.

```
# A tibble: 4 x 3
  race_staffer race_total race_pct
  <chr>             <int>    <dbl>
1 asian               121     1.80
2 hispanic            249     3.71
3 nh_black            262     3.90
4 nh_white           6087    90.6
```

Here's the distribution in Black senators' offices:

```
# A tibble: 4 x 3
  race_staffer race_total race_pct
  <chr>             <int>    <dbl>
1 asian                 5     4.17
2 hispanic              3     2.5
3 nh_black             17    14.2
4 nh_white             95    79.2
```

Here's the distribution in Hispanic senators' offices:

```
# A tibble: 4 x 3
  race_staffer race_total race_pct
  <chr>             <int>    <dbl>
1 asian                 1     1.56
2 hispanic              9    14.1
3 nh_black              4     6.25
4 nh_white             50    78.1
```

Here's the distribution in Asian/PI senators' offices:

```
# A tibble: 4 x 3
  race_staffer race_total race_pct
  <chr>             <int>    <dbl>
1 asian                18    12.5
2 hispanic              8     5.56
3 nh_black              5     3.47
4 nh_white            113    78.5
```

Here's the distribution in White senators' offices:

```
# A tibble: 4 x 3
  race_staffer race_total race_pct
  <chr>             <int>    <dbl>
1 asian                97     1.52
2 hispanic            229     3.58
3 nh_black            236     3.69
4 nh_white           5829    91.2
```

Distribution of race by salary groups in 2010

group by quintile

```
# A tibble: 12 x 4
# Groups:   salary_rank [3]
   salary_rank race         n   pct
         <int> <chr>    <int> <dbl>
 1           1 asian      153  2.05
 2           2 asian      133  1.78
 3           3 asian      156  2.09
 4           1 hispanic   390  5.23
 5           2 hispanic   465  6.23
 6           3 hispanic   245  3.28
 7           1 nh_black   317  4.25
 8           2 nh_black   445  5.96
 9           3 nh_black   320  4.29
10           1 nh_white  6603 88.5
11           2 nh_white  6419 86.0
12           3 nh_white  6741 90.3
```

Comparing to senate report on salaries

The senate report you sent me includes aggregate salaries by racial group (page 89), which is useful to compare it with my data. However, I get pretty different numbers. For one, my overall salary numbers don't seem to match the salary numbers in that report (in the next section I look at the gender salaries as a sanity check, since gender is something in our data, and those also don't seem to match). This might be related to how I'm aggregating the data, but I'm not sure about what I could be going wrong.

Currently, I'm aggregating all the salary inputs a person got in a year (usually this means the salary for the 180 day term, in the case of the senate, but this would also include all sorts of "corrections", such as negative salaries and salaries that only span for 1 or 2 days), then I divide that number by the total number of days they worked in a year. I did this after realizing that many staffers might have only worked for part of the year and thus, it wasn't comparable to an annual salary.

The table below shows the annual salaries (computed as explained above) by racial group for Senate staffers (includes exclusively those who work directly for a Senate member, which I'm not 100% sure is the same as in the report).

```
# A tibble: 4 x 2
  race     salary_race
  <chr>          <dbl>
1 asian         36590.
2 hispanic      35722.
3 nh_black      36236.
4 nh_white      40677.
```

Here's a screenshot from the report (page 89). I think the comparable group would be staffers in Washington? But I'm also not sure. Do you know if our dataset includes staffers that work in the state offices?

Average Salary for all Positions by Race/Ethnicity

Race/Ethnicity	Total	Washington	State
Asian	$35,044	$40,477	$26,894
Black	$37,690	$38,685	$36,260
Hispanic	$35,829	$40,876	$32,780
White	$47,271	$50,462	$40,976
Other	$39,184	$42,085	$37,008

On average, Black Senate staff earn 80 cents for every dollar earned by white staff. Hispanic earn 76 cents, and for Asian staff the figure is 74 cents.

This could mean that the algorithm is biased towards white (ie., it's categorizing as "white" people who would self-identify as non-white), which would explain why the salary gap seems smaller in our data. However, given that gender and overall average salary also doesn't check out (below), it's hard to draw any conclusions yet.

Sanity check with overall salary

According to the report, the average salary in 2001 for Senate staffers in Washington was 49,202. In the data it is 40,503.

Sanity check with salaries in 2001 in the Senate by gender

```
# A tibble: 2 x 2
  gender salary_senate
  <chr>          <dbl>
1 F              39476.
2 M              43725.
```

Here's a screenshot from the report (page 87)

Average Salary for all Positions by Gender

Gender	Total	Washington	State
Female	$42,236	$45,845	$36,923
Male	$50,501	$52,876	$44,845
Differential	$8,265	$7,031	$7,922

On average, female Senate staff earn 84 cents for every dollar earned by male staff. Among Washington staff, the figure is 87 cents; among state staff, it is 82 cents.

In short, the numbers in our data don't really match those in the report (even ignoring race). However, the report says they got their numbers from only 24 offices of senators. So one possibility is that they just got a sample that is uniquely different from the total population (which is what we have). To check if that's the case, I created a function to "sample" 24 offices from our total population of senators, and I repated that procedure multiple times to get a distibution and see where the report's average salary falls within that distribution.

In the graph below you can see the distribution of these samples (I repeated the procedure 2000 times). The dashed line is the average in the report. It's a rare event, but not an impossible one, which means it is possible that the difference we are seeing in averages is due to their sample being particularly unique. The bad news about that is that this leaves us a bit on square one on having a "gold standard" to compare to our imputed race distribution to.