Mar 9

1. Stratified sampling and sample weights
2. Estimation in R with brms

Stratified sampling and sample weights

Oversampling

The problem

A truly uniform sample from a population may not include enough cases from smaller groups for meaningful analysis. This is especially true for intersecting categories (e.g. Asian students with Black teachers).

Full sample

White	4440
Black	2191
Asian	20
Hispanic	9
Native American	9
Other	11

~5% subsample

White	225
Black	101
Asian	1
Hispanic	1
Native American	0
Other	0

Oversampling

The solution

Deliberately sample populations you know to be small with higher probability. In this case, we could sample 3% of white students, 6% of Black students, and 100% of remaining students.

Full sample

White	4440
Black	2191
Asian	20
Hispanic	9
Native American	9
Other	11

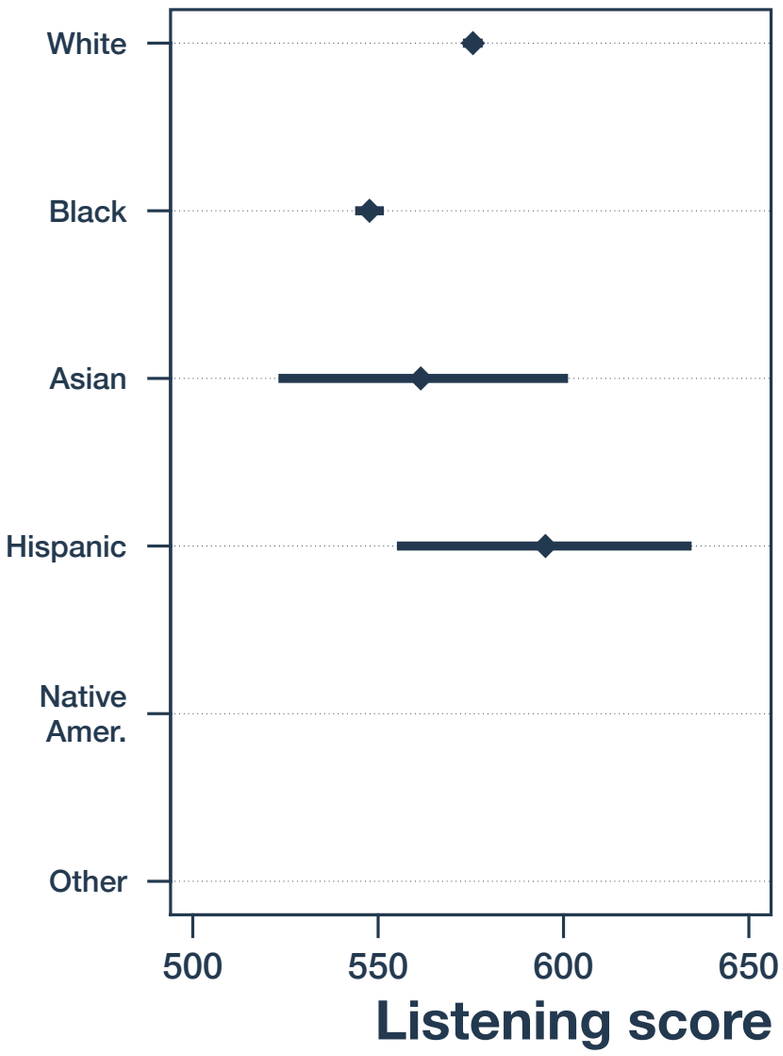
~5% subsample
(with oversampling)

White	139
Black	140
Asian	20
Hispanic	9
Native American	9
Other	11

Oversampling

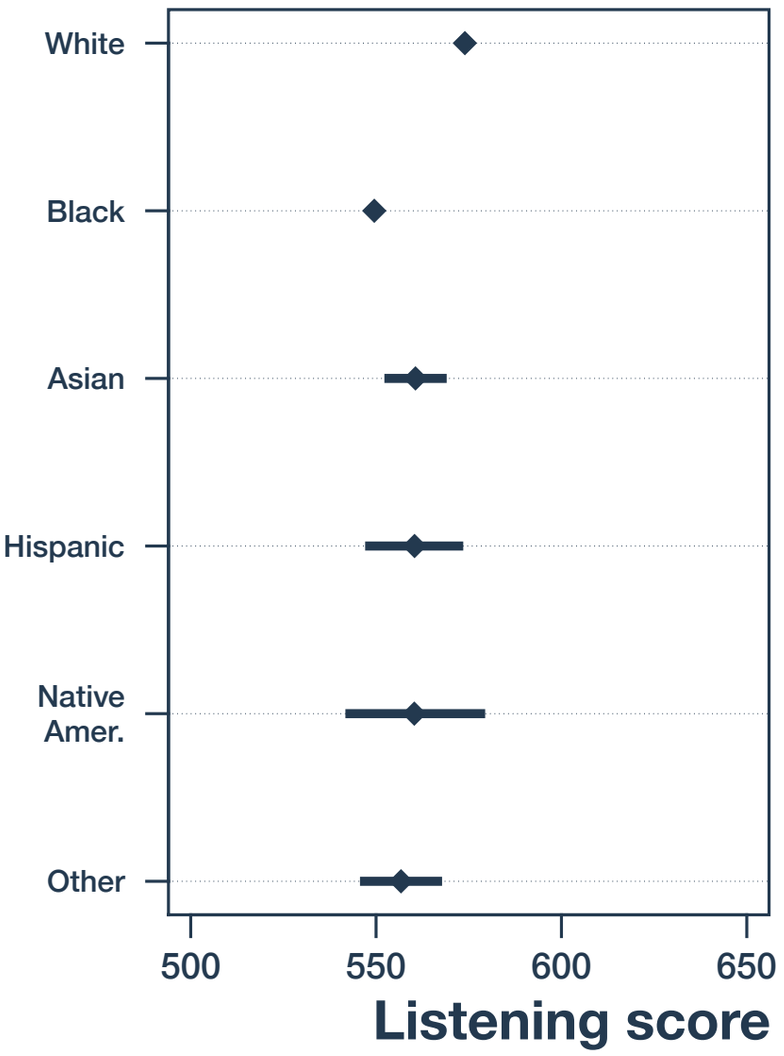
~5% subsample

White	225
Black	101
Asian	1
Hispanic	1
Native American	0
Other	0



~5% subsample
(with oversampling)

White	139
Black	140
Asian	20
Hispanic	9
Native American	9
Other	11



Using oversampled data

Sampling weights tell us how many cases this data point represents in the population.

ID	listening_score	race_ethnicity	s_w
4	556	Black	16.66667
20	—	Hispanic	1.00000
43	568	Other	1.00000
60	531	White	33.33333
86	592	White	33.33333
122	611	Asian	1.00000
⋮	⋮	⋮	⋮

Using oversampled data

More complicated scenarios

There are many *reasons* that data is non-uniformly sampled

- Stratified sampling
- Multiple rounds
- Non-response

There are many *ways* that data is non-uniformly sampled

- Multiple waves
- Levels of analysis (individual, household, region, etc.)

Data sets can have several different ‘weights’

- It is important to use the right one.

Using oversampled data

```
listening_score | weights(s_w)~
    re_black + re_asian + re_hispanic +
    re_native_american + re_other
```

Sampling weights are indicated in brms with a pipe (‘|’) after your outcome variable and the special “weights” function that indicates the variable containing case weights (in our case, ‘s_w’).

This tells brms to multiply the likelihood for each case by that case’s value of s_w.