Denver Bicycling Survey

Dongdong Lu

1 Introduction

Biking as a transportation is beneficial for health, local business, and environment. It is estimated that 2/3 of the population wanted to use bikes more often but did not, due to concerns of traffic safety. Many bikers and potential bikers are afraid of using bike lanes where thousands of cars whizzing past. To combat with issue, the Denver Bike Streets (a local non-profit organization) crafted a map that uses low-traffic, low-speed side streets that could take a biker everywhere in Denver. They call it “Bike Streets”. The city government also recognize the value of biking and are building extensive bike lanes.

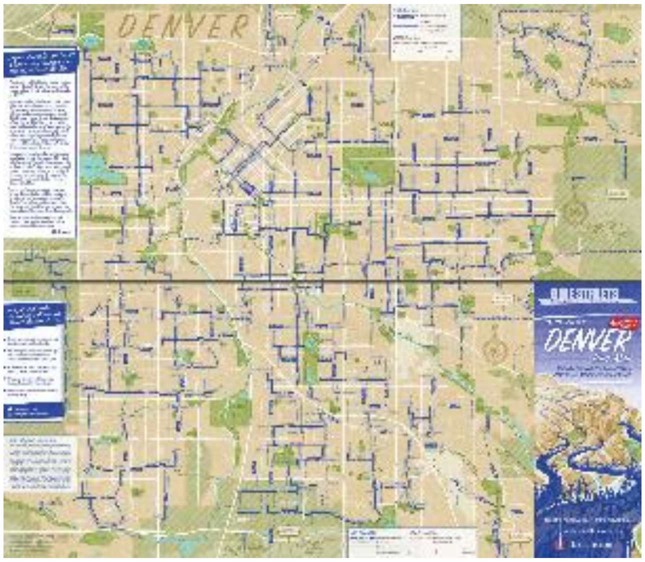

Figure 1 Denver bikes street map (cited from Bikes street website)

However, building the lanes and making the map are just the first steps, the ultimate goal is to get more people ride bikes. So, the question naturally arises: “How many people ride bikes in Denver?” Additionally, we are interested in a few other aspects of people’s biking behaviors: What is the percentage of people who ride on designated trails? What is the percentage of people who have heard Denver Bike Streets Project? What are their average riding time?

To address these questions, we explored methods such as wildlife counting at the beginning of this semester. But then we find out that the community survey is the only feasible method that could be finished within a reasonable cost and produce a relatively low-bias estimation. Ideally, we would finish the following work during this semester: create and deliver the survey, analysis the responses.

Due to the outbreak of COVID-19, it is not possible to conduct the survey during this semester. Instead we have the following goals:

- Design the survey
- Write the delivery methods of the survey
- Create a simulated response set
- Data analysis based on the simulated data set

In this report, I will include the “Denver Bicycling Survey” with the explanation of its design. Then, I will present the delivery methods of survey along the discussion of each method’s pros and cons. Thirdly, since we do not have a real response set available, we created a simulated response set which we can perform the data analysis (bias-adjusting) on. In the discussion section, we will discuss the results on the simulated dataset, provide caveats of the bias-adjusting and state the next steps of work.

2. Analysis and Results

2.1 Survey

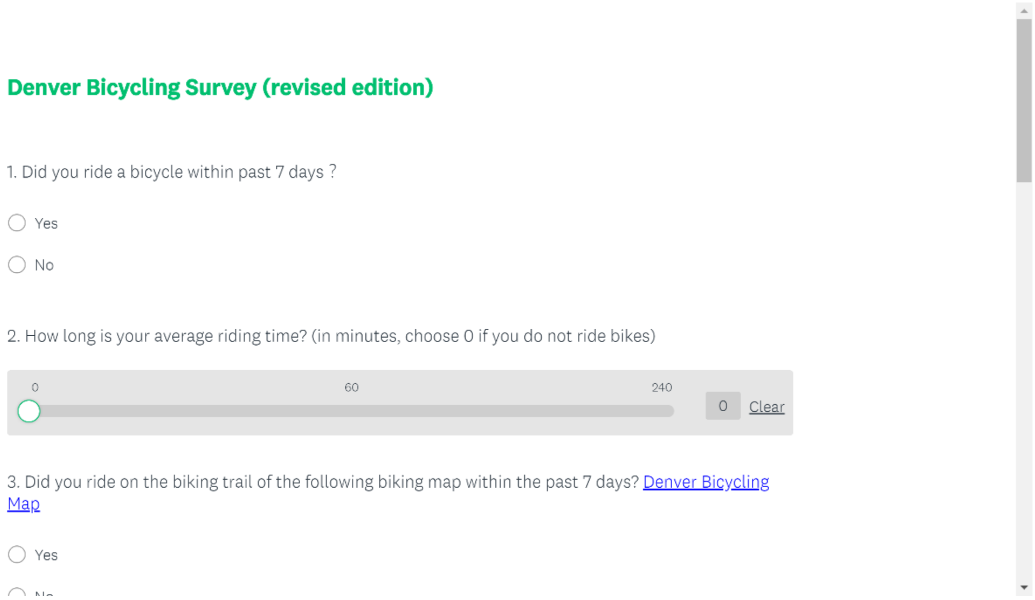

Figure 2 Denver Bicycling Survey (click on it for redirection to the webpage)

Detailed design about this survey:

1. Did you ride a bicycle within past 7 days?

   Question 1 directly address my client's key question. Originally, I considered to ask "Did you ride a bicycle **yesterday**?" But later, I realize that people's behavior more likely to fall into the cycle of weeks since what people do this week is typically not substantially (statistically) different from what they do last week, but what they do on Monday should be statistically different from what they do on Saturday. So, it is better to inquire their biking behavior on a weekly basis.

2. How long is your average riding time? (in minutes, choose 0 if you do not ride bikes)

   Question 2 addresses the riding time of people.

3. Did you ride on the biking trail of the following biking map within the past 7 days? (Denver Bicycling Map)

   Question 3 has a link to the “Denver Bike streets route Map”. This question wants to know what the percentage of people is who ride bike on the routes.

4. Have you ever heard of Denver Bike Streets Project?

   Question 4 wants to know the percentage of Denver residents who have heard of “Denver Bike Streets”.

5. What is your Gender? Which of the following best describes your age? What best describes your race group?

   Question 5 – 7 are for adjusting variables: gender, age, and race. I choose those three as adjusting variables since those data are publicly available. Additionally, I read a paper saying that the biggest drop of bike usage happened around driving age (16-18), indicating age is related with biking behaviors. There are also other researches indicating the correlation of bicycling behavior with gender and race.

6. Any suggestion to make a Denver more biking-friendly city?

   Question 8 is an open question which gathering ideas to make Denver more biking friendly.

To avoid bias, all the above questions are neutrally worded, and they are compressed into the minimized size.

2.2 Delivery methods of the questionnaire

Pros and cons of each of the delivery method.

- By Mail/Text messages

  Pro: Easy to send in a large quantity

  Con: May get a very low response rate

- By Volunteers' distribution at:

  1. Social events where people have a little bit of time while waiting in lines: such as food pantry, exhibitions

  2. Randomly select households in volunteers' neighborhood

  3. Senior living community

  Pro: Can get responses from (possibly) under-represented population (seniors)

  Con: The per response (volunteer's time) cost may be high.

- By Facebook advertisement (cause 20$ to reach 1000 people)

  Pro: Fast to implement, cost little volunteering time is distributing the questionnaire

  Con: May get a very low response rate.

2.3 Simulated response set

Table 1 Simulated response set

case_id	gender	race	age	ride	rideOT	Rtime	heard	weight
1	female	white	under 18	no	no	62	no	0.305362
2	male	black	45-64	yes	no	65	yes	1.236772
3	female	black	45-64	no	no	45	no	0.297467
4	male	white	45-64	no	no	66	no	0.270115
5	male	black	45-64	no	no	81	no	1.042227
6	female	black	25-44	no	no	64	no	1.269597
7	male	white	25-44	no	no	60	yes	1.269597
8	female	black	25-44	no	no	53	yes	1.269597
9	male	asian	25-44	no	no	66	yes	0.270115
10	male	asian	45-64	no	no	47	no	1.269597
11	female	asian	25-44	no	no	42	yes	1.042227

Imagine we have collected the responses from the survey and processed it into the above form:

The first 4 headers are named as the word's direct meaning.

RideOT: the respondent rides on the designated trail or not.

Rtime: the respondent's average riding time

Heard: whether the respondent have heard the Denver Bikes Streets project or not.

Weight: the final assigned weight for each case

I also intentionally introduced differences among different racial and age groups (which may not reflect the reality). For example, white and pacific origins are assigned with the lowest rate of riding (0.1) and highest rate (0.4) of riding. Even further, people who are above 65 and under 18 are assigned with the lowest (0.4) and highest rate of riding (0.05). We will test the introduced difference before and after raking. The table below shows the changes to of each group's weight. "Target" means the ratios got from census data which our raking is performed upon. "Unweighted N" is the number of respondents before weighting.

"Unweighted %" are the percentages of each group before weighting.

"Wtd N" is the number of each respondent group after weighting.

"Change in %" shows the how the percentages are changed after weight, positive means increases, and negatives mean decreases.

"Rsid. Disc." means the difference between the after-raking percentages with the targeted percentages.

"Orig. Disc." Shows the original difference the before-raking percentages and the targeted percentages.

The Latino group which were previously under-represented sees the largest increase (0.279). And the age group 45-64 which were previously over-represented sees the largest decrease (-0.247).

Table 2 Changes for variable race after raking

	Target	Unweighted N	Unweighted %	Wtd N	Wtd %	Change in %	Resid. Disc.	Orig. Disc.
2orMore	0.0465380	789	0.0789	465.39091	0.0465391	-0.0323609	-1.10e-06	-0.0323620
asian	0.0385925	1747	0.1747	385.93410	0.0385934	-0.1361066	-9.00e-07	-0.1361075
black	0.1157775	1786	0.1786	1157.80136	0.1157801	-0.0628199	-2.60e-06	-0.0628225
latino	0.3609535	817	0.0817	3609.38877	0.3609389	0.2792389	1.46e-05	0.2792535
native Am	0.0158910	410	0.0410	158.91383	0.0158914	-0.0251086	-3.00e-07	-0.0251090
pacific	0.0011351	97	0.0097	11.35098	0.0011351	-0.0085649	0.00e+00	-0.0085649
white	0.4211124	4354	0.4354	4211.22007	0.4211220	-0.0142780	-9.60e-06	-0.0142876
Total	1.0000000	10000	1.0000	10000.00000	1.0000000	0.5584778	2.92e-05	0.5585069

Table 3 Changes for variable age after raking

	Target	Unweighted N	Unweighted %	Wtd N	Wtd %	Change in %	Resid. Disc.	Orig. Disc.
18-24	0.1084093	284	0.0284	1084.104	0.1084104	0.0800104	-1.1e-06	0.0800093
25-44	0.3657548	2957	0.2957	3657.506	0.3657506	0.0700506	4.2e-06	0.0700548
45-64	0.2026342	4501	0.4501	2026.379	0.2026379	-0.2474621	-3.7e-06	-0.2474658
above 65	0.1053698	755	0.0755	1053.697	0.1053697	0.0298697	1.0e-07	0.0298698
under 18	0.2178318	1503	0.1503	2178.314	0.2178314	0.0675314	5.0e-07	0.0675318
Total	1.0000000	10000	1.0000	10000.000	1.0000000	0.4949241	9.5e-06	0.4949315

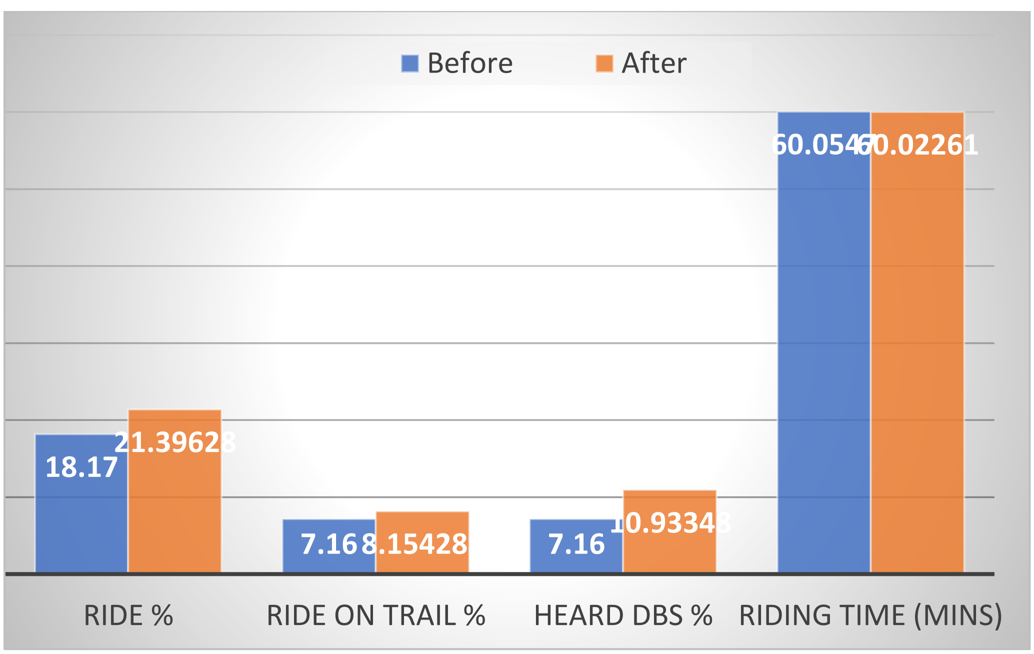

Figure 3 Changes of the 4 reponses

The above graph shows how the percentage of 4 questions change before and after raking. Blue shows the ration before raking and orange shows the ratio after raking. We can see people who ride bicycle within the past 7 days increased from 18.17% to 21.40%. Similarly, the second and third percentage all see positive changes, but the last one riding time sees no significant change since we did not introduce any bias on the response to this question.

Table 4 Range of weights

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.03323	0.26336	0.36357	1.0000	1.17139	5.00009

From the above table, we see our weight range from 0.03323 to 5.00009. The mean of those weights is always 1 since we are inflating some subgroups and deflating others while keeping the weighted population total the same. As we will discuss in the discussion section, extreme weights are an issue, so we put 5 as a cap on the weights. Finally, the general design effect (variance ratio) is 2.603833

3 Discussion

3.1 Discuss the results

As we have noticed in figure 3, 3 out the 4 questions' response showed significant change. However, as we will mention below in the weakness of raking (or any other weighting method), raking will increase the variance of the estimates derived from the survey. The "general.design.effect" function tells us that there is a 160.4% increase after raking. Undesirably, this is telling us that our conclusion is less statistically significant. But I think this is a worthwhile trade-off since the new result is unbiased in the sense that if we could repeat 1000 times, the average of any statistic after raking will approach to the one of real population, or in the words we get better representatives of the population after raking.

3.2 Caveats to my analysis

The Pew (2018) suggests the limitedness of raking as well as other bias-adjusting method: even the best would not be able to eliminate the bias. But we can combine different methods together to archive better performance.

In the paper, the researcher examined the effectiveness of raking on a simulated data which the bias is known by design. Raking itself would only eliminate 6.4% of the bias at the sample size 2000. With the combination of propensity and matching, together they are reduced 6.2% of the bias at the same sample size.

3.3 Next steps

- Update census data

Currently, we performed raking using the 2010 census data. But this year is already 2020 and Denver is a rapidly developing city where all the demographic ratios may have changed substantially after 10 years. To get a more accurate estimate, the above procedures should be performed again using the 2020 census data.

- Make a shiny app

The Denver Bikes Streets program want to repeatedly measure the percentage of Denver's biking population over time. And they need to process the data at times when statistician is not available. I think Shiny app is a great tool to help them archive such goal. *Shiny app* is a small user-friendly program generated by *R* which can perform specific tasks without the assistance of professionals. Once made, it is easy to learn and easy to use.

3.4 Recommended resources

For anyone who want to have a quick look into the raking method, Fricker (2015) provides an excellent introduction of raking.

The Collier (2018) is a webpage showing an interesting example which our case study is based upon. It is easy to understand and come with enough explanation.

Lumley (2011) provide a comprehensive guidance on complex surveys for anyone who want to broaden their knowledge in the survey analysis area. But it does not provide rigorous math.

For people who interested in the math behind raking, please refer to Lu (2003). It states the raking algorithm mathematically and derived the variances of estimates.

The statistics library of “peopleforbikes” has a great collection of research of biking behavior with demographics, ethnic diversity, incentives, and events. I decided to choose gender, age and race as adjusting variables after reading their collection of publications.

4 Mathematics of raking

Raking is a method which assign a weight to each case of the sample so that the sample’s weighted distribution of the adjusting variables matches with the ones of target dataset such a population census. In this report ,we use the *American Neighborhood Survey* as our target dataset and demographic variables such as age, race, and gender as our adjusting variables. In the raking iterations, each case started with the sample design weight and terminates with an updated weight when the pre-set convergence criterion is attained. The final weights of cases may show significant variability with certain units have much higher or lower weights than others which will increase the sampling variances of the survey estimates.

The main assumption of raking is that the responses to the survey questions are reasonably correlated to whatever population demographics that you are raking on. Before, as with any data set, we should carefully explore the data and clean it up, as necessary - look for outliers, typos, and other errors and correct them as possible and appropriate. Raking does not do anything for missing data. If there is missing data, we need to do certain imputations first and then perform the raking.

Define the values on unit i in the population as $Y_i, i = 1,2, \dots, N$ and in the sample as $y_i, i = 1,2, \dots, n$. The population mean can be estimated as $\theta = \bar{Y} = \sum_{i=1}^{N} Y_i/N$. Also divide the population with J stratification cells with population N_j and sample size n_j in each cell $j = 1,2, \dots, J$.

$$N = \sum_{j=1}^{N} N_j \text{ and } n = \sum_{j=1}^{N} n_j.$$

Population is cross-classified into $J = J_1 \times \ldots \times J_D$ cells and K= $\sum_{d=1}^{D} J_d$ is the number of marginal cells. For example, suppose that the UC Denver students' population can be classified as sex (male or female), state origin (50 states), 4 categories of racial group (White, Latino, Black or Asian/pacific), 4 categories of year, then $J = 2 \times 50 \times 4 \times 4 = 1600$. The cell population N_j can be known from university's record and define $\vec{N} = (N_1, \ \ldots, \ N_J)^t$ be the vector of populations in each cell.

We define each cell with enough specificity so that among each there all units are having the same probability of being sampled. Define π_j as the probability of including a unit in cell j in the population in the response sample. Denote the population mean and standard deviation within cell j as $\theta_j = \bar{Y}_j$ and $\sigma_j = S_j$. The sample mean within each cell is $\bar{y}_j$. The overall mean of the population is $\theta = \bar{Y} = \frac{\sum_{j=1}^{J} N_j \theta_j}{N}$ which will be referred as basic stratification identity. For the discussion below, our focus is the weighted estimates $\hat{\theta} = \sum_{j=1}^{J} W_j \, \hat{\theta}_j$ and $\sum_{j=1}^{J} W_j = 1$.

The weighting method we used here, does not depends on y_j's values, it only depends on the n_j's and N_j's. So, we have a population estimate

$$\widehat{\theta_W} = \sum_{j=1}^{J} W_j \, \bar{y}_j = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}$$

where $w_i = \frac{W_{j(i)}}{n_{j(i)}}$ is the unit weight of sample case i in cell j.

Only K margins $\vec{M} = \left(M_{1,1}, \ldots, M_{1,J_1}, \ldots, M_{D,J_D}\right)^t$ are known adjusting variables. We want to estimate N_j from sample sizes and margins.

Let A be the $K \times J$ indicator matrix that satisfies

$$A\vec{N} = \vec{M}$$

Write $A = (A_1, \ldots, A_D)^t$ here A_i is a $J \times J_i$ matrix.

Define the vector of sample margins $\vec{m} = \left(m_{1,1}, \ldots, m_{1,J_1}, \ldots, m_{D,J_D}\right)^t = A\vec{n}$, where $\vec{n} = \left(n_1, \ldots, n_J\right)^t$ is the vector of sample margins.

Raking algorithm works as the following steps:

1. For $d = 1$, calculate $\vec{r} = \left(M_{d,1}/m_{d,1}, \ldots, M_{d,J_d}/m_{d,J_d}\right)^t$ and compute $\vec{w} = A_d\vec{r}$

2. Update $\vec{n}$ by multiplying each element n_j by w_j. For $j = 1, \ldots, J$, update $\vec{m} = \left(m_{1,1}, \ldots, m_{1,J_1}, \ldots, m_{D,J_D}\right)^t = A\vec{n}$

3. Repeat steps 1, 2

4. Repeat steps 1,2,3 until $\vec{m}$ converges to $\vec{M}$

The updated n_j will be the IPF estimate of N_j , say $\widehat{N_j}$, for $j = 1, \ldots, J$

Then the weights for raking are $W_j = \widehat{N_j}/N$ and $w_j = \widehat{N_{j(i)}}/(Nn_{j(i)})$ for $i = 1, \ldots, n$

5 Reference

Fricker, Ron & Anderson, Lew. (2015). Raking: An Important Often Overlooked Survey Analysis Tool. Phalanx. 36-42.

Lu, H. and Gelman, A. (2003). A method for estimating design-based sampling variances for surveys with weighting, poststratification and raking. *J. Official Statistics* **19** 133—151

Lumley, Thomas. *Complex surveys: a guide to analysis using R*. Vol. 565. John Wiley & Sons, 2011.

Pew Research Center, January 2018, “For Weighting Online Opt-In Samples, What Matters Most?

Survey Raking: An Illustration - Andrew Collier - https://datawookie.netlify.app/blog/2018/12/survey-raking-an-illustration/

Fricker, Ronald. “Re: Ask questions about “raking”.” Message to Dongdong Lu. 26 April 2020. E-mail.

Statistics Category. Peopleforbikes

https://peopleforbikes.org/our-work/statistics/statistics-category/?cat=participation-statistics

6 Appendix

6.1 Code

```
# case id

case_id = 1:10000



# gender

gender_groups = c("male","female")

gender = sample(gender_groups,10000,replace = 1)



# race

race_groups = c("white","latino","black","asian","native Am","pacific","2orMore")

race = sample(race_groups,10000, replace = 1, prob = c(0.5,0.1,0.2,0.2,0.05,0.01,0.09))



# age

age_groups = c("under 18","18-24","25-44","45-64","above 65")

age = sample(age_groups,10000, replace = 1, prob = c(1,0.2,2,3,0.5))
```

```
#rided

ride_status = c("yes","no")

ride = rep("no",10000)



# introduce bias for different racial group

for (i in 1:10000) {

# assign white group with 0.1 chance of riding

  if (race[i] == "white"){

    ride[i] = sample(ride_status,1,prob = c(0.1,0.9))

  }

# assign latino group with 0.1 chance of riding

  if (race[i] == "latino"){

    ride[i] = sample(ride_status,1,prob = c(0.3,0.7))

  }
```

```
# assign black group with 0.2 chance of riding
  if (race[i] == "black"){
    ride[i] = sample(ride_status,1,prob = c(0.2,0.8))
  }
# assign asian group with 0.15 chance of riding
  if (race[i] == "asian"){
    ride[i] = sample(ride_status,1,prob = c(0.15,0.85))
  }
# assign pacific group with 0.4 chance of riding
  if (race[i] == "pacific"){
    ride[i] = sample(ride_status,1,prob = c(0.4,0.6))
  }
# assign native American group with 0.1 chance of riding
  if (race[i] == "native Am"){
    ride[i] = sample(ride_status,1,prob = c(0.1,0.9))
  }
```

```
 # assign 2ormore group with 0.15 chance of riding

 if (race[i] == "2orMore"){

   ride[i] = sample(ride_status,1,prob = c(0.12,0.88))

  }

# assign “under 18” group with 0.4 chance of riding

  if (age[i] == "under 18"){

   ride[i] = sample(ride_status,1,prob = c(0.4,0.6))

  }

# assign “above 65” group with 0.05 chance of riding

  if (age[i] == "above 65"){

   ride[i] = sample(ride_status,1,prob = c(0.05,0.95))

  }



}



# assign “above 65” group with 0.
```

```
05 chance of riding rideOT = rep("no",10000)

for (i in 1:10000)

  {if (ride[i] == "yes")

# set 0.4 people on trail

{rideOT[i] = sample(c("yes","no"), 1, prob = c(0.4,0.6))}

}



# average riding time

for (i in 1:10000)

  {if (ride[i] == "yes")

# set the average riding time to 60 mins

{Rtime = rpois(10000, 60)}

}



# assign the chance of heard or not

heard = rep("no",10000)
```

```
for (i in 1:10000)

# assign 0.1 people heard the project

{heard[i] = sample(c("yes","no"),1,prob = c(0.1,0.9))}



# create the data.frame for the survey

survey = data.frame(case_id,gender,race,age,ride,rideOT,Rtime,heard)

library(readr)

# load the  population (census ratios)

population <- read_csv("C:/Users/Don/Desktop/SC/population.csv")



library(readxl)

# load the data

age_fraction <- read_excel("C:/Users/Don/Desktop/SC/age_fraction.xlsx")

age_fraction
```

``

```
library(weights)

# incorporate population (gender and race) data into target which we will perform raking on.

target <- with(population, list(

  gender = wpct(gender, fraction),

  race  = wpct(race, fraction)

))

# build the list for age ratios

target2 <- with(age_fraction, list(

  age = wpct(age,fraction)

))

# combine two target info into one

target = c(target,target2)

# make a target a sting

str(target)
```

```
# calculate the percentages of ride or not, or trail or not, heard or not and riding time

wpct(survey$ride)

mean(survey$Rtime)

wpct(survey$rideOT)

wpct(survey$heard)


# load anesrake package

library(anesrake)

# apply the code raking

raking <- anesrake(target,survey,survey$case_id,cap = 5,choosemethod = "total",type =
"pctlim",pctlim = 0.05)


# create the summary for raking

raking_summary <- summary(raking)
```

```
# load the knitr package for table making

library(knitr)

library(kableExtra)



# make a table for raking summary of race

kable(raking_summary$race)

kable(raking_summary$race) %>%

kable_styling(bootstrap_options = c("striped", "hover"))



# make a table for raking summary of age

kable(raking_summary$age) %>%

  kable_styling(bootstrap_options = c("striped", "hover"))

# create a weight vector

survey$weight <- raking$weightvec



# re-calculate those weights’ percentages
```

```
wpct(survey$ride, survey$weight)

mean(survey$Rtime*survey$weight)

wpct(survey$rideOT, survey$weight)

wpct(survey$heard, survey$weight)



# get the general design effect

raking_summary$weight.summary

raking_summary$general.design.effect
```