# Stat 854 Project

**Stat 454/854: Sampling Theory & Practice (Winter 2024)**

Professor: Christian Boudreau
Last updated: Apr 5, 2024

Due at 11:59 pm on Tuesday April 23, 2024

---

# 1 Introduction

Up to now, all of the variance estimates we have discussed in STAT 454/854 are the so called Taylor expansions. This should be particularly easy to notice, when looking at how we derive the variance (and thus indirectly the variance estimate) of $\hat{R}$ in section 5.3 of the slides. Since the Hájek estimator is a special case of $\hat{R}$, all the variations of $\widehat{V}(\bar{y}_{\text{Ha}})$ discussed so far (this includes $\widehat{V}(\widehat{p}_{\text{Ha}})$, as $\widehat{p}_{\text{Ha}}$ is a special case of $\bar{y}_{\text{Ha}}$) are Taylor expansions. This is also well documented in the online help for `PROC SURVEYMEANS` and `PROC SURVEYFREQ`; see Details ⇒ Statistical Computations. Note that Taylor expansion variance estimates are obtained in SAS by specifying the option `VARMETHOD=TAYLOR`. However, since this is the default option, we never had to specify it.

Although Taylor expansion variance estimates are the ones most commonly used, this is not the only method to obtain variance estimates for survey data. Other methods include: balanced repeated replication (`VARMETHOD=BRR` in SAS), Jackknife (`VARMETHOD=JK` in SAS) and the bootstrap (`VARMETHOD=BOOTSTRAP` in SAS). This project is concerned with last of those 3 methods. The bootstrap was introduced by Efron (1979), but this was not in the context of survey data. Efron's bootstrap was first "adapted" to survey data by McCarthy and Snowden (1985). Slightly modifying their original bootstrap algorithm to allow for stratification (in addition to sampling weights), we get the following (very simple) algorithm:

Let $y_{hi}$ be the observation from the $i^{\text{th}}$ subject/respondent ($i = 1, \ldots, n_h$) of the $h^{\text{th}}$ stratum ($h = 1, \ldots, H$) and $w_{hi}$ be the sampling weight corresponding to that observation.

Step 1: Let $b = 1$. Independently in each of the $H$ strata, draw a simple random sample with replacement (SRSWR) of size $n_h$ is drawn from the original sample $\{y_{h1}, \ldots, y_{hn_h}\}$ to obtain the bootstrap sample $\{y^{*b}_{h1}, \ldots, y^{*b}_{hn_h}\}$ with corresponding weights $\{w^{*b}_{h1}, \ldots, w^{*b}_{hn_h}\}$.

Step 2: Repeat step 1 for $b = 2, \ldots, B$, where $B$ is a large number (e.g., 1000), thus obtaining the following two matrices for each of the $H$ strata

$$Y^*_h = \begin{bmatrix} y^{*1}_{h1} & y^{*2}_{h1} & \ldots & y^{*b}_{h1} & \ldots & y^{*B}_{h1} \\ y^{*1}_{h2} & y^{*2}_{h2} & \ldots & y^{*b}_{h2} & \ldots & y^{*B}_{h2} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ y^{*1}_{hn_h} & y^{*2}_{hn_h} & \ldots & y^{*b}_{hn_h} & \ldots & y^{*B}_{hn_h} \end{bmatrix}$$

$$W^*_h = \begin{bmatrix} w^{*1}_{h1} & w^{*2}_{h1} & \ldots & w^{*b}_{h1} & \ldots & w^{*B}_{h1} \\ w^{*1}_{h2} & w^{*2}_{h2} & \ldots & w^{*b}_{h2} & \ldots & w^{*B}_{h2} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ w^{*1}_{hn_h} & w^{*2}_{hn_h} & \ldots & w^{*b}_{hn_h} & \ldots & w^{*B}_{hn_h} \end{bmatrix}$$

Step 3: Compute

$$\bar{y}_{\text{Ha}}^{*b} = \frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h} y_{hi}^{*b}\, w_{hi}^{*b}}{\sum_{h=1}^{H}\sum_{i=1}^{n_h} w_{hi}^{*}}$$

the bootstrap analog of $\bar{y}_{\text{Ha}}$, for $b = 1, \ldots, B$.

Step 4: Compute the bootstrap variance estimator of $\bar{y}_{\text{Ha}}$ using the following formula

$$\widehat{V}^{*}(\bar{y}_{\text{Ha}}) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\bar{y}_{\text{Ha}}^{*b} - \bar{y}_{\text{Ha}}\right)^2 \tag{1}$$

Because sampling in step 1 is done with replacement, this is called the with-replacement bootstrap. It is also often referred to the naive bootstrap, as $\widehat{V}^{*}(\bar{y}_{\text{Ha}})$ is biased.

Notes:

1. Steps 3 and 4 are performed by `PROC SURVEYMEANS` (when estimating a population mean) or `PROC SURVEYFREQ` (when estimating a population proportion).
2. There is an alternate version of (1) given by

$$\widehat{V}^{*\bullet}(\bar{y}_{\text{Ha}}) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\bar{y}_{\text{Ha}}^{*b} - \bar{y}_{\text{Ha}}^{*\bullet}\right)^2 \tag{2}$$

where

$$\bar{y}_{\text{Ha}}^{*\bullet} = \frac{1}{B-1}\sum_{b=1}^{B}\bar{y}_{\text{Ha}}^{*b}$$

However, like $\widehat{V}^{*}(\bar{y}_{\text{Ha}})$, $\widehat{V}^{*\bullet}(\bar{y}_{\text{Ha}})$ is also biased.

To fix the bias of $\widehat{V}^{*}(\bar{y}_{\text{Ha}})$ (and of $\widehat{V}^{*\bullet}(\bar{y}_{\text{Ha}})$) various modifications/improvements to the with-replacement bootstrap have been proposed. For this project, we will focus on the following three:

i) the rescaling bootstrap (BRS) proposed by Rao & Wu (1988) and Rao et al. (1992)

ii) the mirror-match bootstrap (BMM) proposed by Sitter (1992*b*); but also look at Sitter (1992*a*)

iii) the without-replacement bootstrap or pseudo-population bootstrap proposed by Gross (1980) and Chen et al. (2019); but also look at section 10.2 of Wu & Thompson (2020)

In addition to the above references, chapter 10 of Wu & Thompson (2020) contains a nice discussion of bootstrap methods for survey data. Note that the basic algorithm of those 3 methods

## 1.1 What you need to do

Step 1: Find 1 or 2 teammate(s) to work on the project with you, and email me the names of all team members. Please make sure to cc your teammate(s) so there are no confusion about who is part of your team. Pretty much all of you have already done this, but I'm missing a couple individuals.

Step 2: When I've received the team information, I will assign your team to one of the 3 bootstrap methods listed above, and email that to you and your teammate(s).

Step 3: Read about the bootstrap method you've been assigned; paying special attention on making sure that you correctly understand how to implement it.

Step 4: Write an `R` function (see below) that will compute bootstrap weights for the method that was assigned to your team.

Step 5: Use your newly created `R` function to compute bootstrap weights for the Survey of Youth in Custody (SYC). To keep things relatively simply, use $B = 100$ (i.e., compute 100 sets of bootstrap weights).

Step 6: Create a short video presentation (see below) to present your work.

## 1.2 Things to submit

Things to submit (i.e., upload on UW Learn*) on Tuesday April 23 (by 11:59 pm):

☐ Your `R` code (or SAS code if you opted to do everything in SAS)

☐ The bootstrap weights you computed (as an Excel or csv file); see notes below

☐ Your SAS code and corresponding output (as a pdf file obtained using SAS ODS)

☐ Video file of your presentation; see notes below

☐ Slides of your presentation (either as a pdf or PowerPoint file)

Notes:

- There are many ways to export an `R` data frame to Excel, but a simple way is to use the `write_xlsx` function from the `writexl` package. Exporting to csv is even easier as the `write.csv` function does not required the installation of any additional `R` package.
- What to talk about in your video/slide (in no particular order):
  - Explain the bootstrap method you are using; making sure to clearly describe the bootstrap algorithm you are using

*Crowdmark does not allow to upload videos, we will thus use the Dropbox feature of Learn.

- Explain the key elements/lines of your `R` code and, equally importantly, how to use your `R` function
- Explain how to use bootstrap weights with `PROC SURVEYMEANS` and `PROC SURVEYFREQ`
- Describe how you applied the bootstrap weights to the Survey of Youth in Custody (SYC), and compare your results from bootstrap weights to those from Taylor expansion
- Your target audience are your fellow STAT 454 classmates; i.e., someone who has taken a sampling course, but likely does not know about bootstrapping and definitively does not know about bootstrapping with survey data
- Though I am not planning on using a stopwatch, I fully expect that team members will speak an equal amount of time; e.g., in a 10 minute video made by a team of 2, I fully expect that each team member will speak about 5 minutes each (I realize that splitting time exactly at 5 minutes might not be practical; I am okay with that, but a 60%–40% is not acceptable)
- [Apr 5] Though you can record using any device or software of your choosing, I need to be able to clearly see both the slides and the person who is speaking. In accordance with the above bullet point, I need to see who is talking, as it's unlikely that I will be able to recognize your voice.
- Your video should be about 10 minutes long (± 1 minute; i.e., between 9 and 11 minutes) for a team consisting of 2 members, and about 13 minutes long (± 1.5 minutes; i.e., between 11.5 and 14.5 minutes) for a team consisting of 3 members
- You will be evaluated on the content of your video and slides (as well as your `R` and SAS code behind it) and how well you explain your work, not on your video production and editing skills; this is not an Hollywood production. That being said, CapCut is a nice, simple and free video editing software. It is available for Windows, MacOS, iOS, iPadOS and Android. Alternatively, you can also edit your video online at capcut.com. They are many YouTube videos on CapCut; here are a few nice ones.
- [Apr 5] As mentioned in one of the bullet points above, you are free to use the device and software of your choosing to record your video. However, using Zoom on your laptop is likely a good option. It will allow you to easily record both the slides (using the Screen Sharing feature) and the person who is talking. Furthermore, you do not even need to all be at the same physically location. Lastly. Zoom is free to use to UWaterloo graduate students (click here for more information).

- `R` code
  - An important aspect of writing computer code is to properly document what you did and what the code does using comments. I am thus fully expecting that your code will content sufficient comments so that somenone knows about the bootstrap method you are using (e.g., myself) will easily understand what your code does
  - Your `R` code must be a function, where the key arguments are the weights, strata information, and $B$ (the number of bootstrap replicates) and the key output is a matrix/data frame containing the bootstrap weights

# References

Chen, S., Haziza, D., Léger, C. & Mashreghi, Z. (2019), 'Pseudo-population bootstrap methods for imputed survey data', *Biometrika* **106**, 369–384.

Efron, B. (1979), 'Bootstrap methods: another look at the jackknife', *The Annals of Statistics* **7**, 1–26.

Gross, S. (1980), Median estimation in sample surveys, *in* 'Proceedings of the Section on Survey Research Methods', American Statistical Association, pp. 181—-184.

Rao, J. N. K., Wu, C. F. J. & Yue, K. (1992), 'Some recent work on resampling methods for complex surveys', *Survey Methodology* **18**, 209–217.

Rao, J. & Wu, C. (1988), 'Resampling inference with complex survey data', *JASA* **83**, 231–241.

Sitter, R. R. (1992*a*), 'Comparing three bootstrap methods for survey data', *Canadian Journal of Statistics* **20**, 135–154.

Sitter, R. R. (1992*b*), 'A resampling procedure for complex survey data', *JASA* **87**, 755–765.

Wu, C. & Thompson, M. E. (2020), *Sampling Theory and Practice*, Springer – Verlag.