# Understanding Variance Estimator Bias in Stratified Two-Stage Sampling

Khoa Dong[1], Tim Trudell[1], Yang Cheng[1], Eric Slud[1,2]

[1]U.S. Census Bureau, [2]University of Maryland

**Joint Statistical Meetings**

**Vancouver, CA**

**July 29, 2018**

United States Census Bureau

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

# Disclaimer

This presentation is intended to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

# Outline

1. Motivation
2. Overview of Current Population Survey (CPS)
3. Problem description
4. CPS variance estimation
5. Simulation results

# Motivation

- When estimating response rate $p$ and $var(p)$ for households in CPS **non-self-representing (NSR)** primary sampling units (PSU), we observed unusually high value of $var(p)$.
- We wanted to better understand the cause of this result.

# Current Population Survey

- One of the oldest surveys in the U.S. (in operation since 1942)
- Measuring national **unemployment rate**
- Monthly sample of ~72,000 households

# Primary Sampling Unit

- PSU - either a county or group of contiguous counties
- Two types of PSU:
  - Self-representing SR
  - Non-self-representing NSR

# CPS Sample Design

- Two-stage stratified sampling design for NSR PSUs:
  - **First stage**: select one PSU per stratum with probability proportional to size (civilian noninstitutionalized population 16+ = CNP 16+)
  - **Second stage**: do systematic sampling within selected PSUs
- Systematic sampling for SR PSUs

# CPS Sample Design

- Select PSUs once every 10 years
- 852 PSUs selected in first-stage (2010 design): 506 SR and 346 NSR
- Approximately 80% of CNP 16+ population in SR PSUs

| | Est. CNP 16+ (Sep 2017) |
|---|---|
| **SR** | 200,298,742 |
| **NSR** | 55,263,673 |

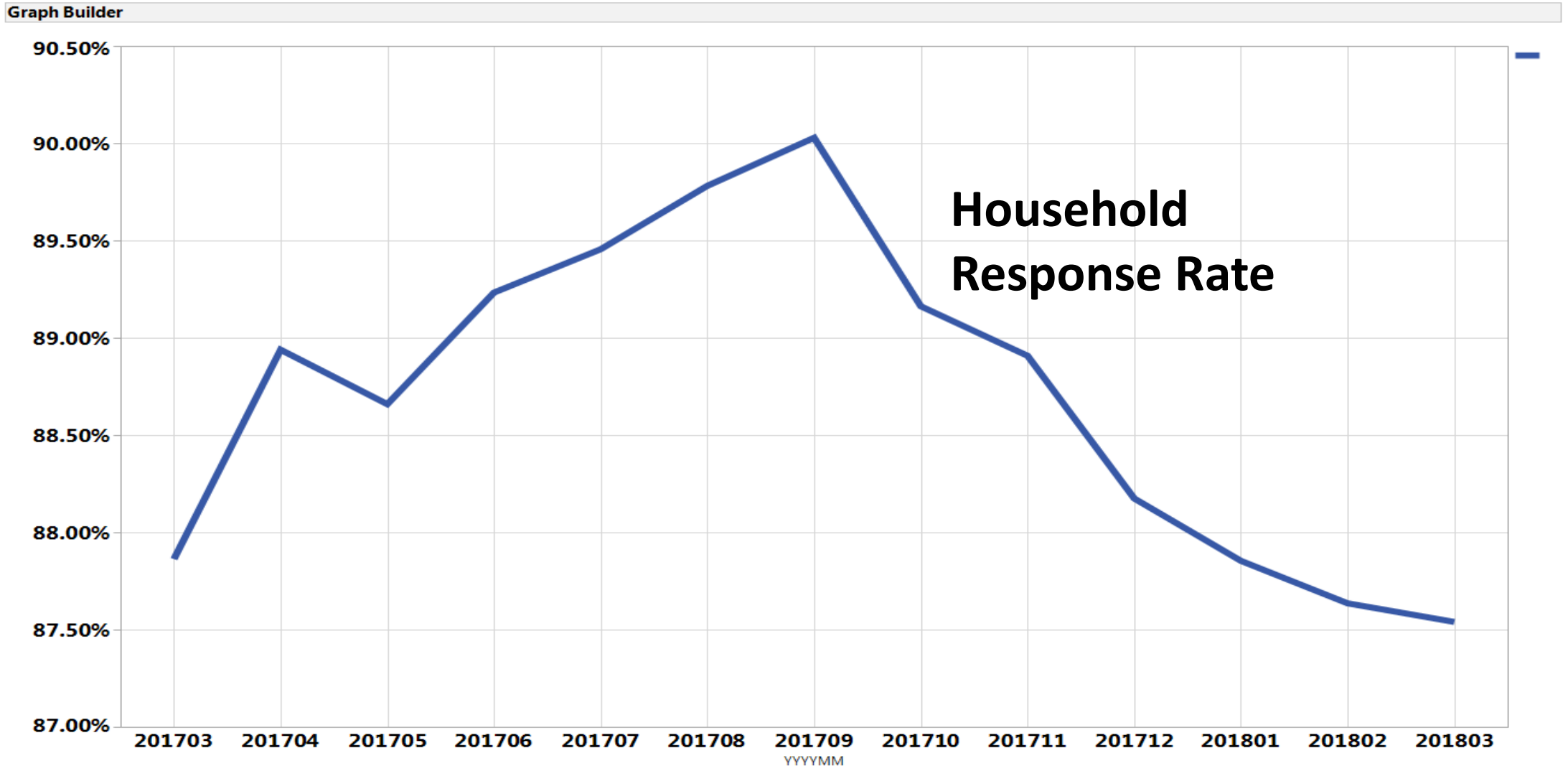
# Key Labor Force Estimates

- Noninstitutionalized civilian labor force statistics:
  - Unemployment/employment levels
  - Unemployment rate
  - Labor force participation rate

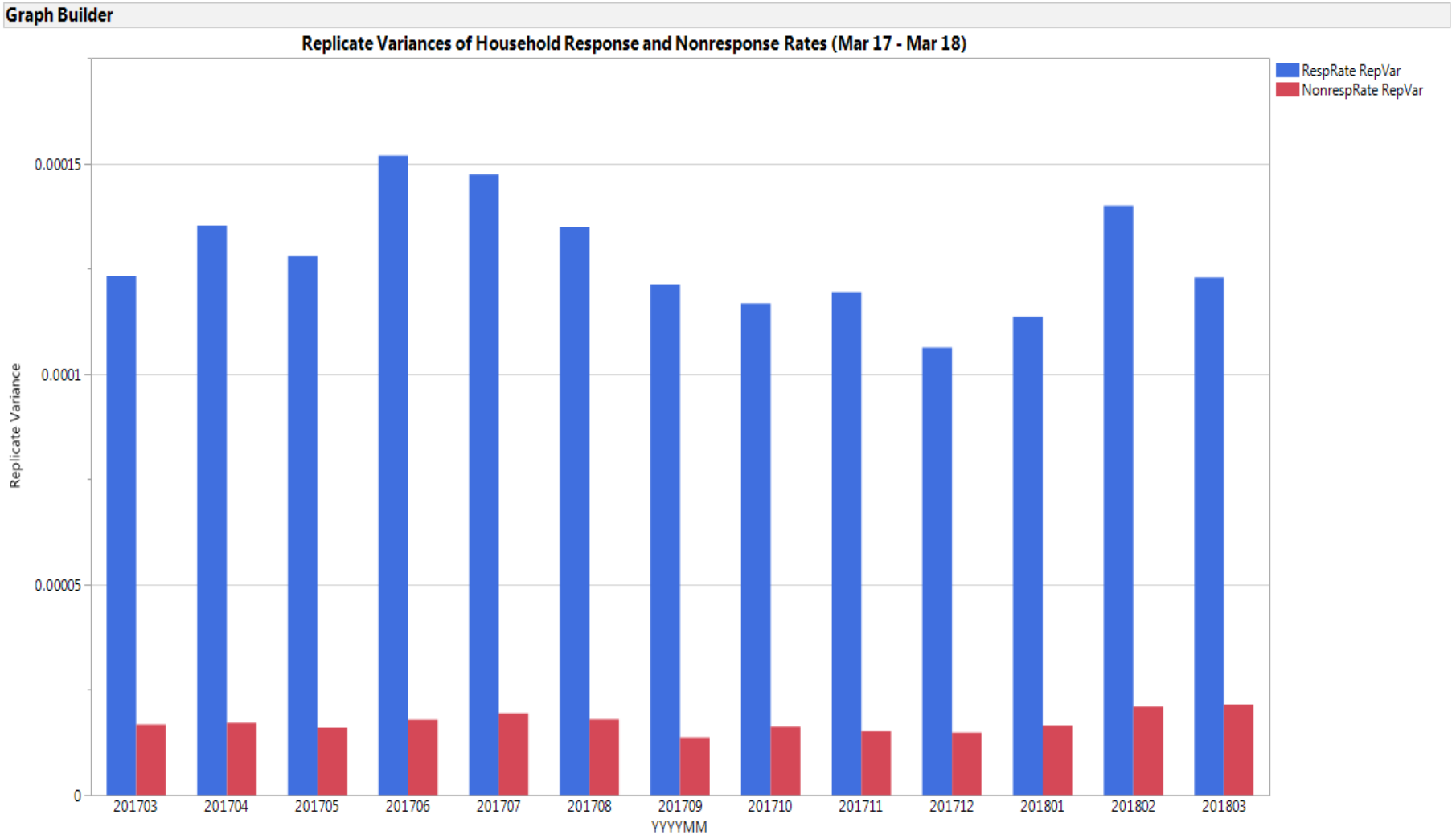
# Problem

- Estimate monthly response rate $p$, variance $var(p)$ for CPS households **in NSR PSUs**.
- The sample is at household level: 1 record for each sampled household in each month.
- Response $y_i$ has binary outcome: 1--response and 0--nonresponse.
- Time period: March 17 – March 18

# NSR Household Response Rates March 17 – March 18



# Estimated Variance for Response and Nonresponse Rates Mar 17 – Mar 18




- **Expect to see $var(p) = var(1-p)$, but they are NOT.**
- **Our chosen variance estimator introduces bias in some way.**

# CPS Variance Estimation

- Due to CPS sample design, there is no direct variance estimator formula:
  - Select only **one** PSU per NSR stratum
  - **Systematic sampling** within PSU
- Currently use **balanced-repeated replication (BRR)** method for **NSR PSUs**.

# CPS Variance Estimation

- BRR variance estimator:

$$var(\hat{Y}) = \frac{1}{R(1-K)^2} \sum_{r=1}^{R} (\hat{Y}_r - \hat{Y})^2$$

where

$\hat{Y}_r$ = the $r$-th replicate estimate of $Y$

$\hat{Y}$ = the full sample estimate of $Y$

$R$ = number of replicates

$K$ = perturbation factor; $0 \leq K < 1$

- BRR requires selecting **two** PSUs per stratum, but CPS selects only one PSU per stratum → collapse strata to make pseudo-strata.
- These pseudo-strata should **ideally contain exactly 2 perfectly matched strata**.

# BRR with Pseudo-Strata

- Suppose we want to estimate a population total $Y$ using

$$\hat{Y} = \sum_{h=1}^{L} \hat{Y}_h$$

where $L$ denotes the number of strata.

- Consider the simple case when $L$ is even, we estimate the variance of $\hat{Y}$ by combining the $L$ strata into $G$ groups of two strata each ($L = 2G$).

# BRR with Pseudo-Strata

- Hence,

$$\hat{Y} = \sum_{g=1}^{G} \hat{Y}_g = \sum_{g=1}^{G} (\hat{Y}_{g1} + \hat{Y}_{g2})$$

$$\boldsymbol{Var(\hat{Y}) = \sum_{g=1}^{G} Var(\hat{Y}_{g1}) + Var(\hat{Y}_{g2}) = \sum_{g=1}^{G} (\sigma_{g1}^2 + \sigma_{g2}^2)}$$

# BRR with Pseudo-Strata

- The $r$-th replicate estimate of $Y$:

$$\hat{Y}_r = \sum_{g=1}^{G} (1 + (1 - K)\delta_{gr})\,\hat{Y}_{g1} + (1 - (1 - K)\delta_{gr})\hat{Y}_{g2}$$

where $\delta_{gr} = 1$ if the **first** stratum in $g$-th group is selected and $\delta_{gr} = -1$ if the **second** stratum in $g$-th group is selected.

- $\delta_{gr}$ are chosen from entries of a Hadamard matrix.
- Rows of a Hadamard matrix are mutually orthogonal:

$$\sum_{r=1}^{R} \delta_{gr}\delta_{kr} = 0\ (\forall\ g \neq k)$$

# BRR with Pseudo-Strata

$$\hat{Y}_r - \hat{Y} = \sum_{g=1}^{G} (1 - K)\, \delta_{gr} \left( \hat{Y}_{g1} - \hat{Y}_{g2} \right)$$

$$(\hat{Y}_r - \hat{Y})^2 = \sum_{g=1}^{G} (1 - K)^2\, \delta_{gr}^2 \left( \hat{Y}_{g1} - \hat{Y}_{g2} \right)^2 + \sum_{g=1}^{G} \sum_{k \neq g}^{G} (1 - K)^2 \delta_{gr} \delta_{kr} \left( \hat{Y}_{g1} - \hat{Y}_{g2} \right) \left( \hat{Y}_{k1} - \hat{Y}_{k2} \right)$$

# BRR with Pseudo-Strata

$$\frac{1}{R(1-K)^2}\sum_{r=1}^{R}(\hat{Y}_r-\hat{Y})^2=\frac{1}{R(1-K)^2}\sum_{r=1}^{R}\sum_{g=1}^{G}(1-K)^2\left(\hat{Y}_{g1}-\hat{Y}_{g2}\right)^2$$

$$+\frac{1}{R(1-K)^2}\sum_{g=1}^{G}\sum_{k\neq g}^{G}(1-K)^2\left(\hat{Y}_{g1}-\hat{Y}_{g2}\right)\left(\hat{Y}_{k1}-\hat{Y}_{k2}\right)\sum_{r=1}^{R}\delta_{gr}\delta_{kr}$$

- Therefore,

$$var\left(\hat{Y}\right)=\sum_{g=1}^{G}(\hat{Y}_{g1}-\hat{Y}_{g2})^2=\sum_{g=1}^{G}(\hat{Y}_{g1}^2+\hat{Y}_{g2}^2-2\hat{Y}_{g1}\hat{Y}_{g2})$$

# Bias in BRR with Pseudo-Strata

- Taking expectation:

$$E\left\{\sum_{g=1}^{G}(\hat{Y}_{g1}^2+\hat{Y}_{g2}^2-2\hat{Y}_{g1}\hat{Y}_{g2})\right\} = Var(\hat{Y}_{g1}) + \mu_{g1}^2 + Var(\hat{Y}_{g2}) + \mu_{g2}^2 - 2\mu_{g1}\mu_{g2}$$

$$= \sum_{g=1}^{G}(\sigma_{g1}^2 + \sigma_{g2}^2) + \sum_{g=1}^{G}(\mu_{g1} - \mu_{g2})^2$$

$$= Var(\hat{Y}) + Bias^2$$

where $\sigma_{gh}^2 = \text{Var}\{\hat{Y}_{gh}\}$ and $\mu_{gh} = E\{\hat{Y}_{gh}\}$.

- Bias squared term is positive and would ADD to variance estimate.
- Bias squared term would be zero if the pair of PSUs in each group were perfectly matched.

# How are Strata Collapsed ?

- In CPS, the objective function is a function of:
  - Unemployment
  - Civilian labor force
  - Children 0-17 at or below $200\%$ poverty level

# Simulation Overview

- Use one month CPS data (Mar 18) which has pseudo-strata information.
- For each household, generate $y_i$ responses iid from Bernoulli distribution with various $p = 0.03, 0.06, \ldots, 0.99$.
- For each $p$:
  - Run 5,000 sims.
  - Compute true variance and BRR variance.
  - Compute bias squared term $\sum_{g=1}^{G}(\mu_{g1} - \mu_{g2})^2$
  - Compare true variance with BRR variance after adjusting for bias.

# Simulation Computation

- Total number of households: $\widehat{N} = \sum_{i=1}^{n} w_i$
- Full sample estimated response count: $\hat{Y} = \sum_{i=1}^{n} w_i y_i$
- Replicate $r$ estimated response count: $\hat{Y}_r = \sum_{i=1}^{n} w_i y_i f_{ir}$ where $f_{ir}$ is either 1.5 or 0.5.

# Simulation Computation

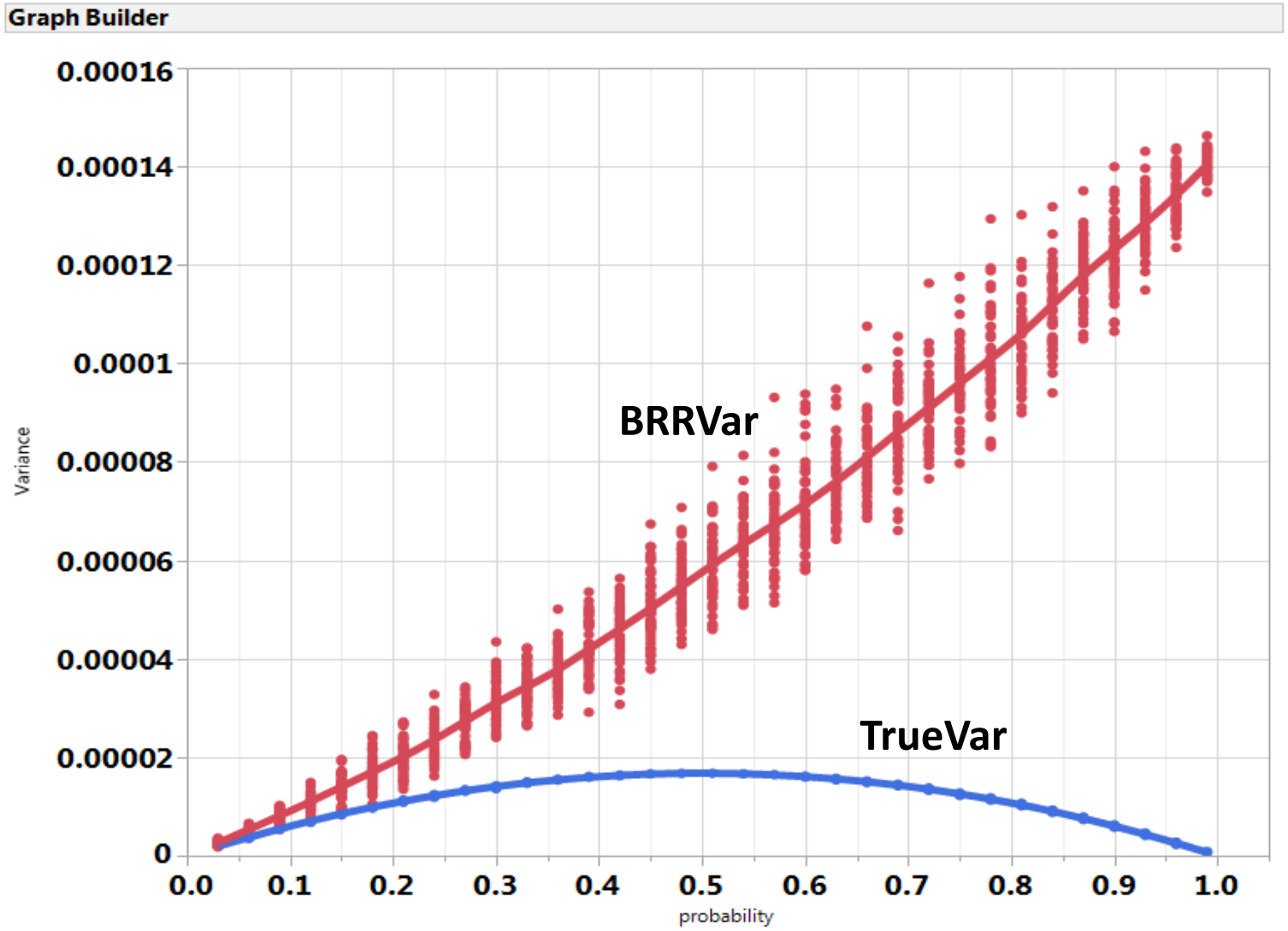
- BRR variance of $\hat{Y}$: $BRRVar(\hat{Y}) = \frac{4}{160}\sum_{r=1}^{160}(\hat{Y}_r - \hat{Y})^2$
- BRR variance of response rate $p$:

$$BRRVar(p) = \left(\frac{1}{\widehat{N}}\right)^2 BRRVar(\hat{Y})$$

assuming $N = \widehat{N}$ is fixed from outside knowledge.

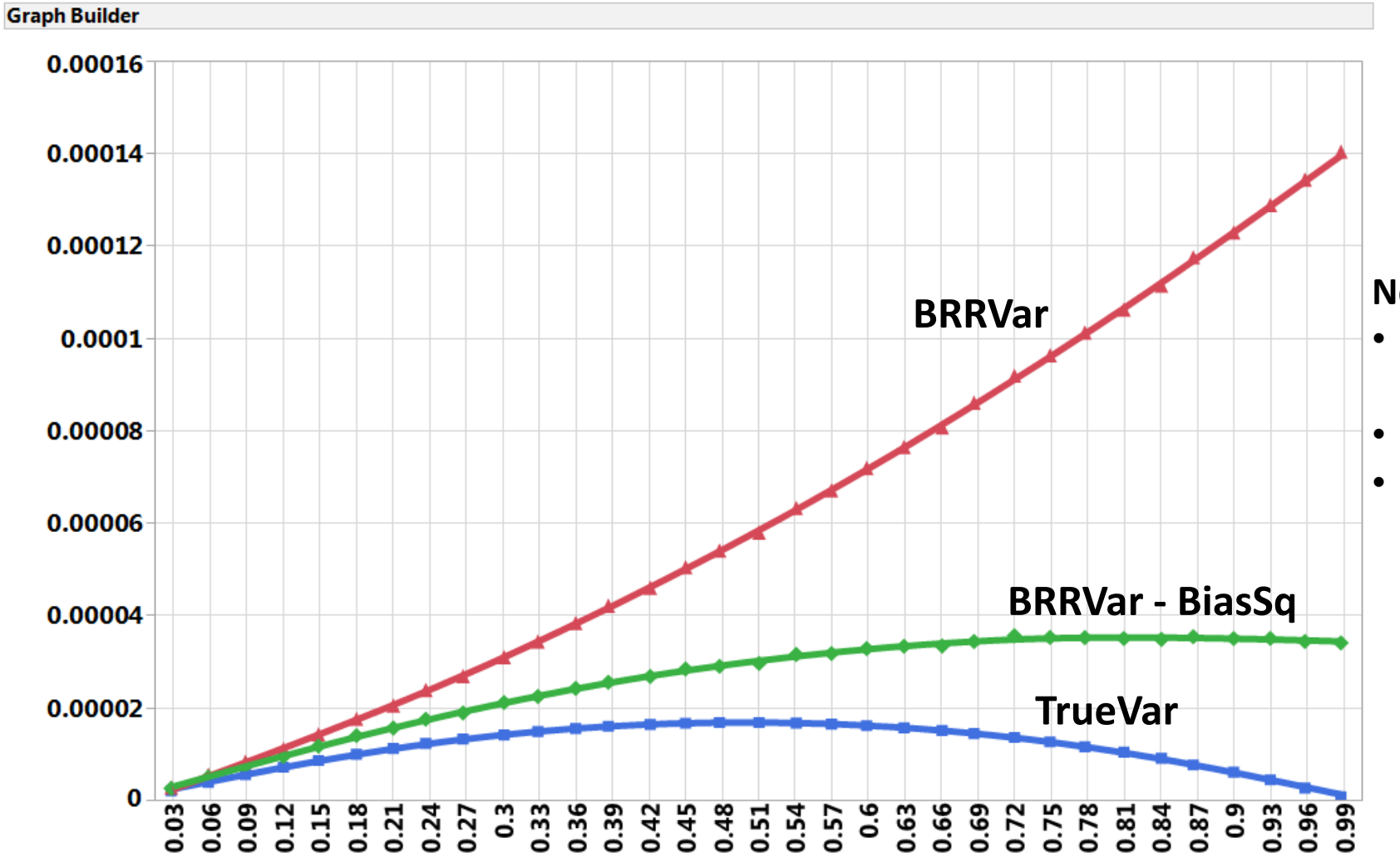
- $\mu_{gh} = p \times N_{gh}$

# BRR Variance vs. True Variance




As $p$ gets close to 1, BRR variance estimate is far different from true variance.

# BRR Variance vs. True Variance




**Not all of bias can be explained due to:**

- **Strata collapsed based on different set of covariates**
- **Use AHS MOS instead of CPS**
- **AHS MOS not currently updated (from 2010 design)**

# Summary

- For NSR component:
  - CPS collapses strata to make pseudo-strata.
  - There is no perfect matching of strata → bias in variance estimator.
  - Bias gets significantly large when $p$ gets close to 1.
  - Quick fix is to use $var(1-p)$ for large $p$.
- CPS is designed for civilian labor force statistics. Expect more bias when estimating variance of other statistics.

# Questions?

# Thank You!

khoa.dong@census.gov

# References


1. David Judkins (1990). “Fay’s method for variance estimation.” Journal of Official Statistics, Vol 6, No. 3, 1990
2. Philip J. McCarthy (1966). “Replication: An Approach to the Analysis of Data from Complex Surveys.” Vital and Health Statistics Series 2 No. 14
3. Robert E. Fay (1984). “Some Properties of Estimates of Variances Based on Replication Methods.”
4. Philip J. McCarthy (1969). “Pseudo-Replication: Half Samples.” Review of the International Statistical Institute, Vol. 37, No. 3, pp. 239-264
5. Yang Cheng (2012). “Overview of Current Population Survey Methodology.” Internal Report.
6. Wolter, K.M. (2008). Introduction to Variance Estimation, New York: Spring-Verlag.