Jackknife and Other Replication Methods with a Reduced Number of Replicates

Stephen Ash
stephen.eliot.ash@census.gov
U.S. Census Bureau

2018 Joint Statistical Meetings
July 30, 2018

Any views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

United States Census Bureau
U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

Goals of the Research

G1. Identify replication variance estimators that use a reduced set of replicates.

G2. Simple expression for the estimator of a variance.

G3. Appropriate for systematic random sampling from an ordered list – which we will refer to as *sys*.

Two parts

Part 1 – Single-stage sample design.

- Estimate the variance from a *sys* sample design.
- *sys* – defined as *systematic random sample from an ordered list*.

Part 2 – Two-stage sample design.

- Estimate the second-stage variance from a general first-stage sample design and a *sys* sample design in the second stage.

Part One: Single Stage sample Designs

$$\hat{Y} = \sum_{h} \sum_{k \in s_h} w_k \, y_k$$

$$\hat{v}\left(\hat{Y}\right)$$

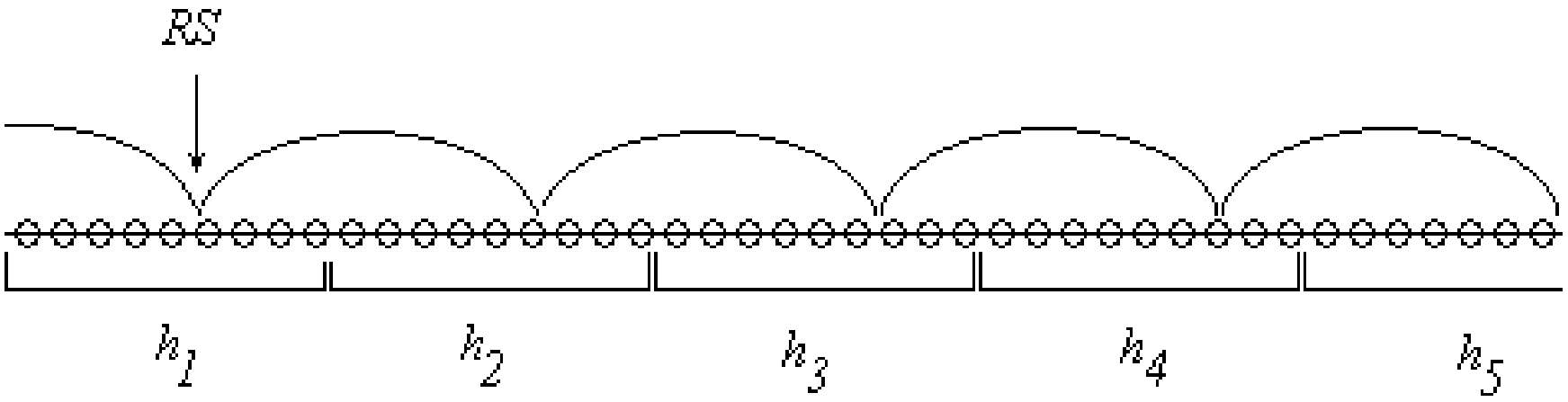
Part Two: Two Stage sample Designs

$$\hat{Y} = \sum_{h} \sum_{i \in s_h} \sum_{k \in s_i} w_i w_k y_k$$

$$\hat{v}\left(\hat{Y}\right) = \hat{v}_{\mathrm{PSU}}\left(\hat{Y}\right) + \hat{v}_{\mathrm{SSU}}\left(\hat{Y}\right)$$

$$\hat{v}_{\mathrm{SSU}}\left(\hat{Y}\right) = \sum_{h} \sum_{i \in s_h} \frac{\hat{v}\left(\hat{Y}_i\right)}{\pi_i^2}$$

Systematic Random Sampling from an Ordered List (*sys*)



-Can be treated as a cluster sample (Cochran 1977)
-Can be treated as a $n_h = 1$ – implicit stratification (Megill *et al*. 1987)

Part 1: Replication methods considered

- Jackknife replication -- Delete 1 unit (JK-1)
- Delete-a-Group Jackknife (DAGJK)
  - Kott (2001)
- Balanced Repeated Replication (BRR)
- Successive Difference Replication (SDR)
  - Fay and Train (1995) and Ash (2014)

Delete-a-Group Jackknife (DAGJK)

- Randomly assigns each sample unit to the R groups called D_r with d_r units in each group D_r.
- Used sys to form the groups.

$$\hat{v}_{JKDAG}(\hat{\theta}) = \frac{R-1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_r - \bar{\hat{\theta}}_r\right)^2$$

$$F_r = \begin{cases} n/(n-d_r) & k \notin D_r \\ 0 & k \in D_r \end{cases}$$

Successive Difference Replication

- Mimics the Successive Difference estimator.
- The successive difference estimator is often good with the *sys* sample design. See Wolter (1984).
- Collapse strata estimator. (1,2), (2,3), (3,4), (4,5),…

$$\hat{v}_{\text{SDR–SSU}}(\hat{Y}) = \frac{4}{R}\sum_{r=1}^{R}(\hat{Y}_r - \hat{Y})^2$$

$$F_{k,r} = 1 + 2^{-\frac{3}{2}}a_{a_k,r} - 2^{-\frac{3}{2}}a_{b_k,r}$$

Part 1: Balanced Repeated Replication (BRR)

- Two sample units per stratum.
- Use BRR as SDR – a collapsed strata estimator
- (1,2), (3,4), (5,6),…

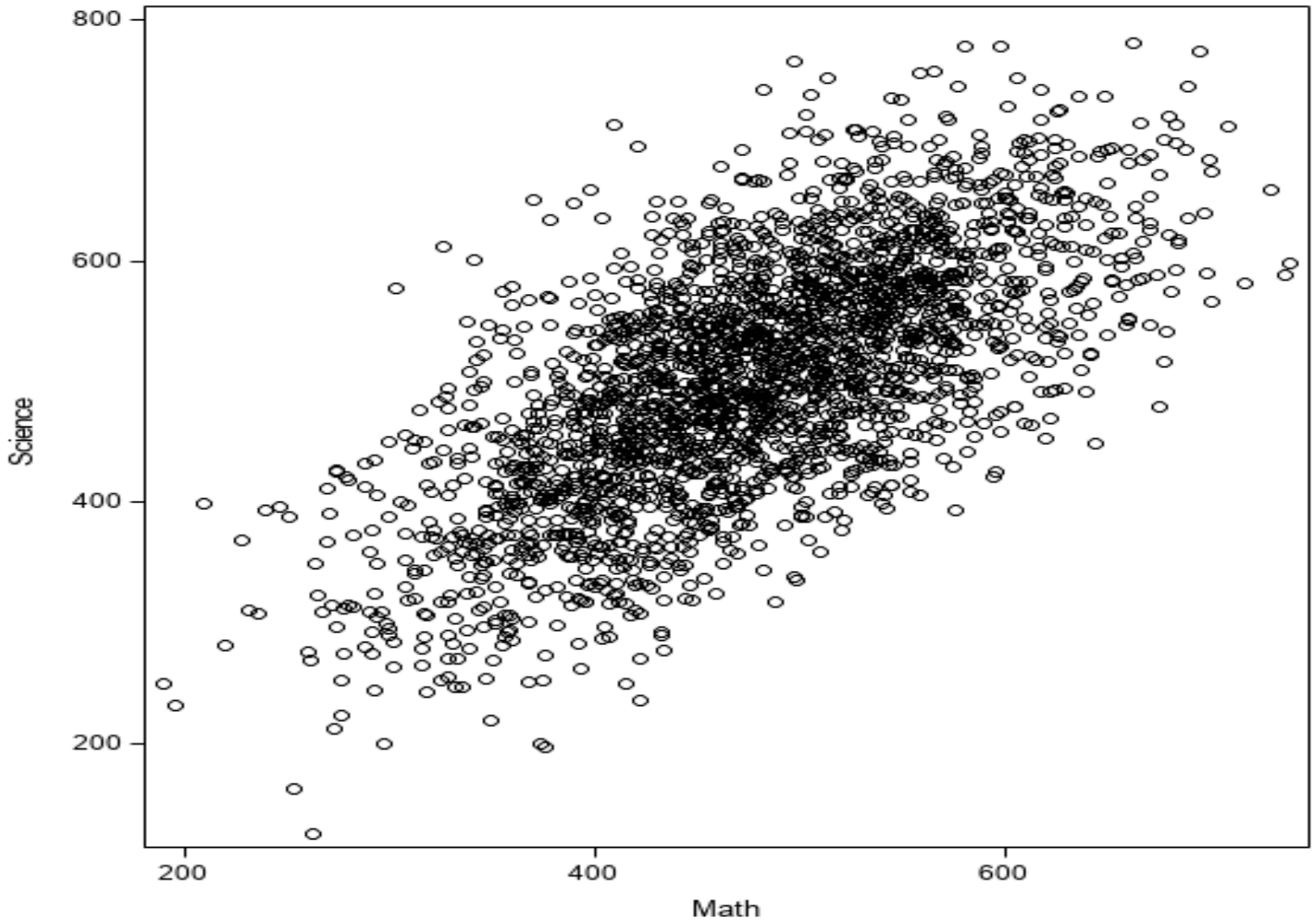
$$F_{r,h} = \begin{cases} 1 + (1-k)a_{r,h} & i = 1 \\ 1 - (1-k)a_{r,h} & i = 2 \end{cases}$$

$$\hat{v}_{\mathrm{BBR-FAY}}(\hat{Y}) = \frac{1}{R(1-k)^2} \sum_{r=1}^{R} \left(\hat{Y}_r - \hat{Y}\right)^2$$

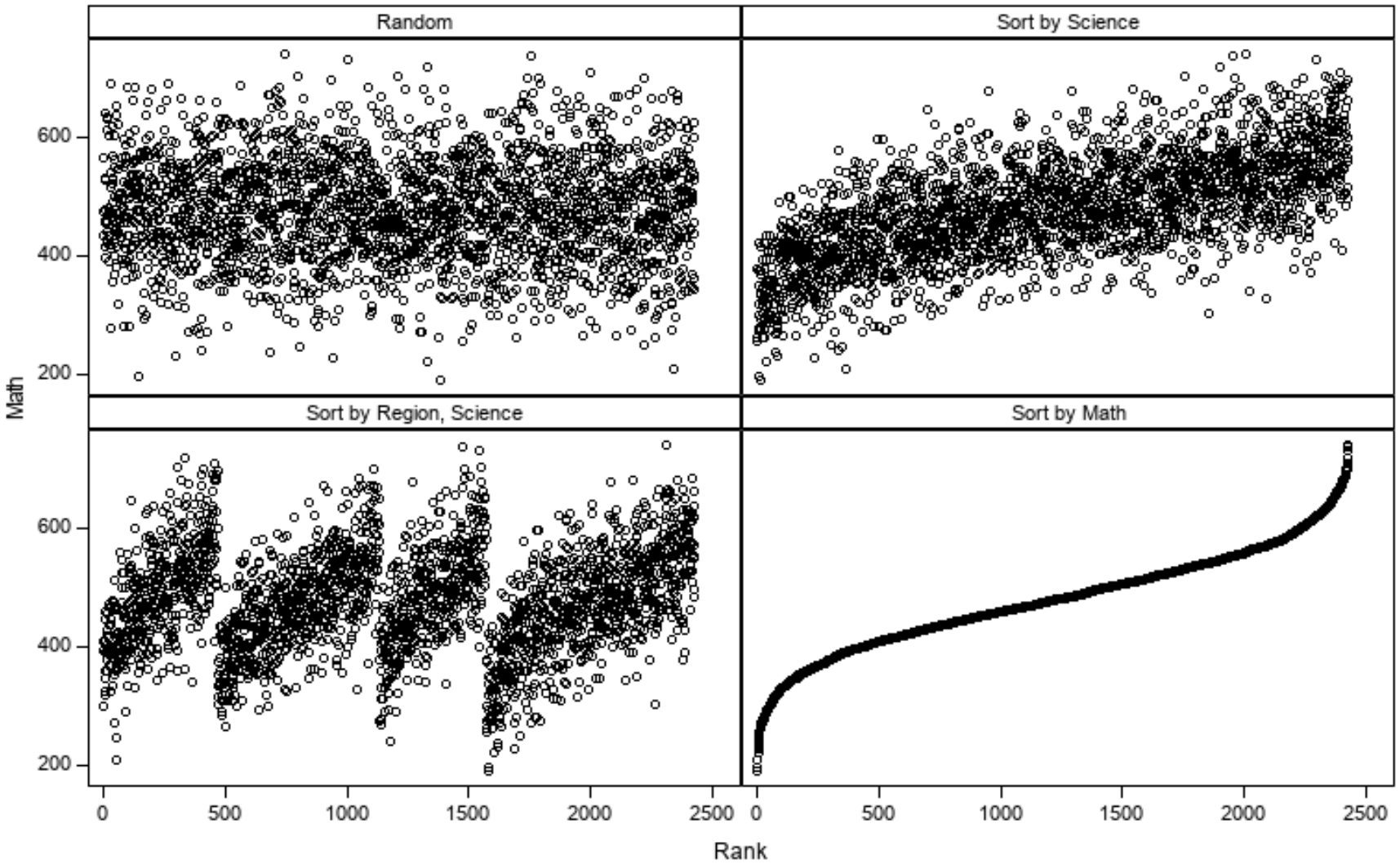
Empirical Example

- Used the complete population of 3rd Graders from Valliant, Dorfman, and Royall (2000).
- N = 2,427.
- Variable of interest total of math scores – y_k = math score
- Sort variables used
  - Science Scores
  - Region (4)

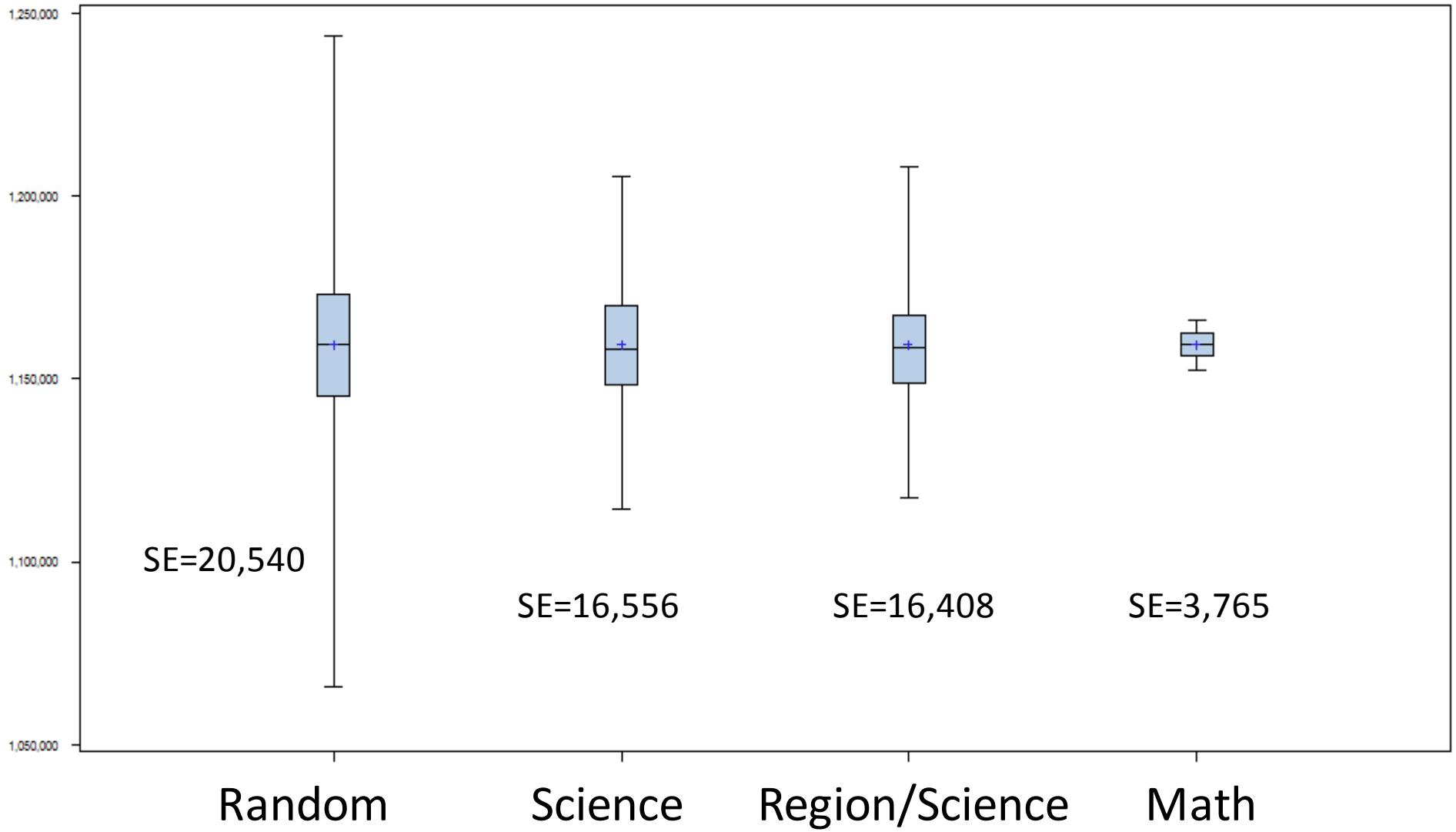
Math and Science Test Scorces



Sort Orders for *sys*

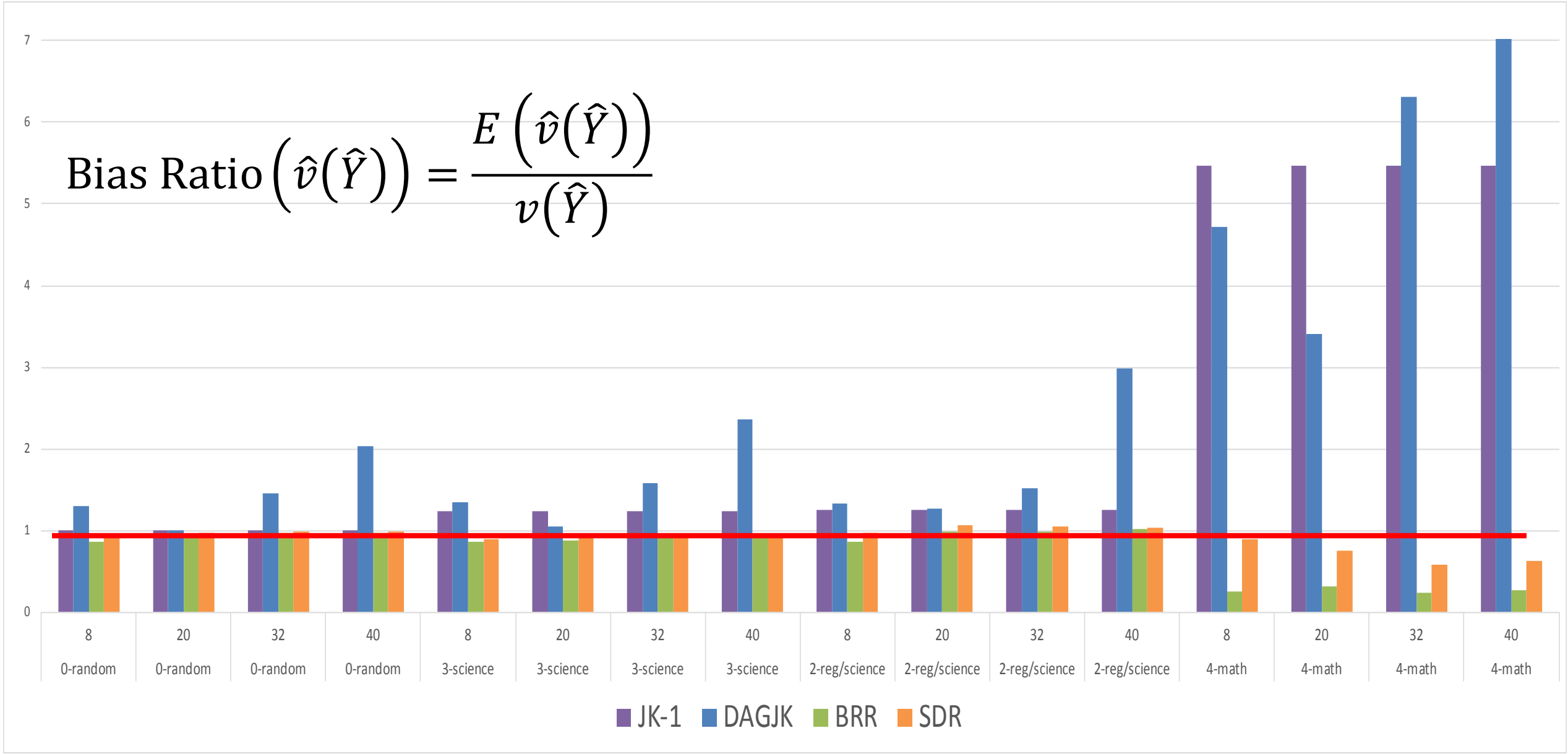


Actual Variances for Empirical Example



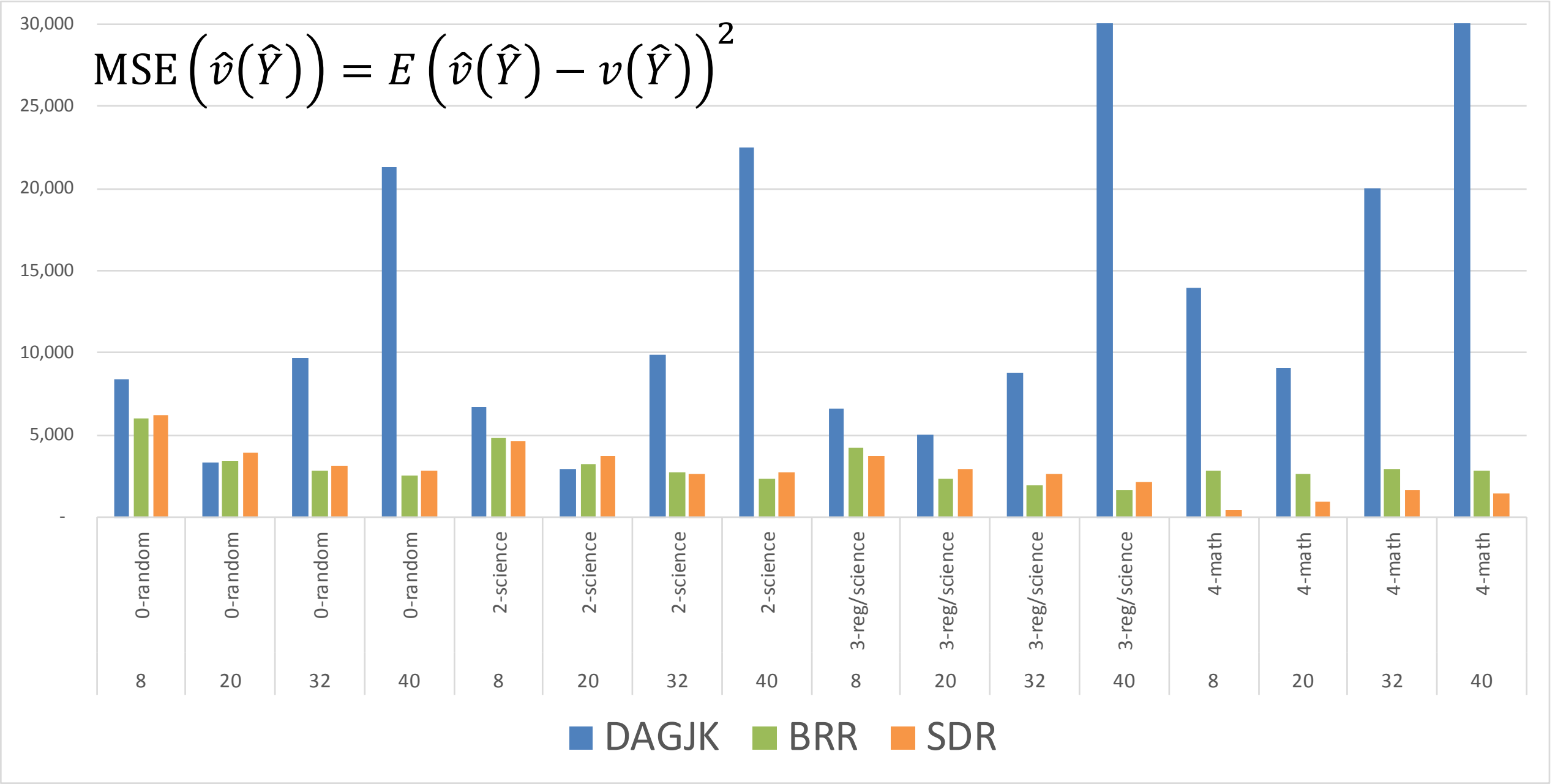
Bias Ratios

$$\text{Bias Ratio}\left(\hat{v}(\hat{Y})\right) = \frac{E\left(\hat{v}(\hat{Y})\right)}{v(\hat{Y})}$$

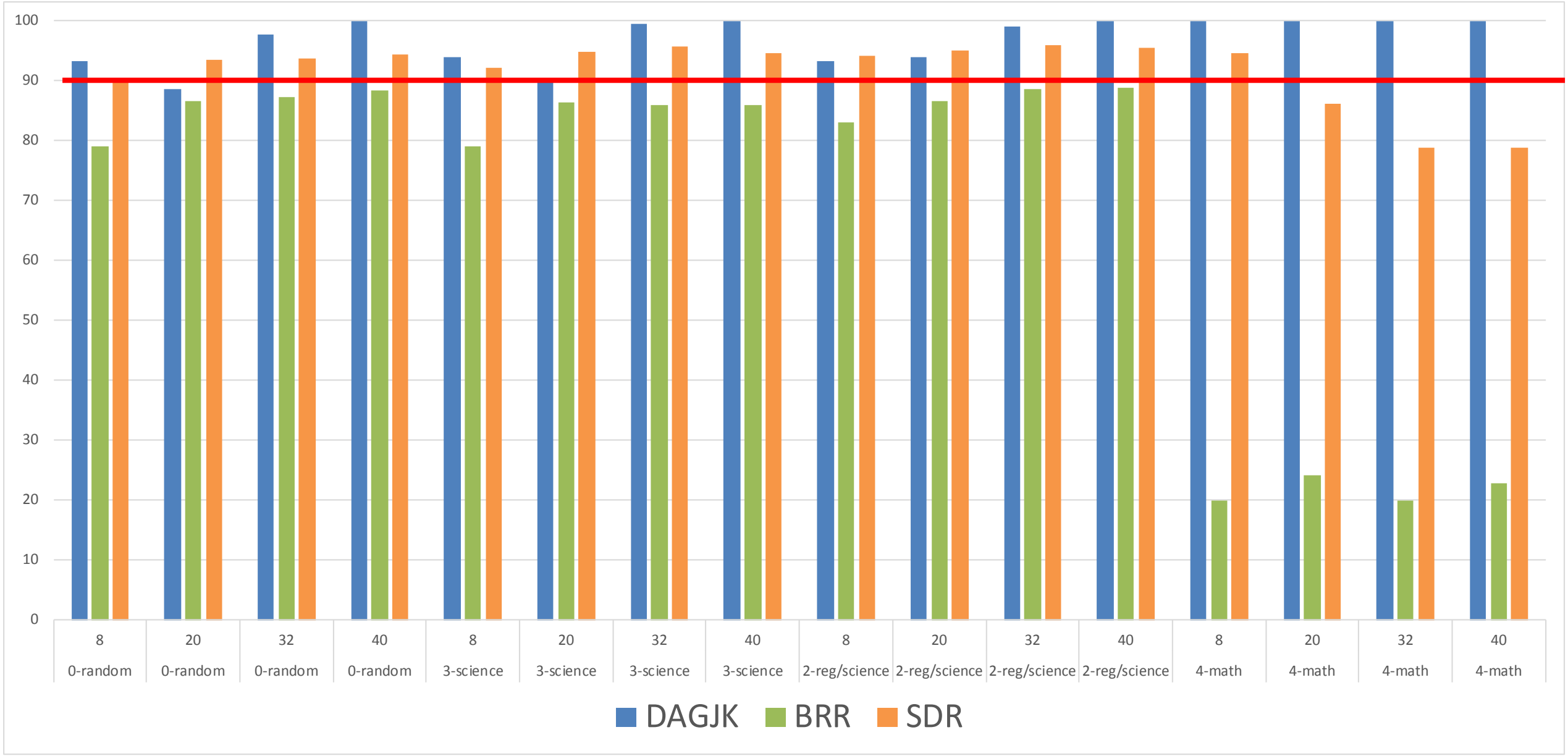


MSEs (as CVs)

$$\mathrm{MSE}\left(\hat{v}(\hat{Y})\right) = E\left(\hat{v}(\hat{Y}) - v(\hat{Y})\right)^2$$



90% Coverage Ratios



Part 1: Conclusions

C1. SDR is best for estimating the variance from a *sys* sample design.

C2. BRR was 2nd best.

C3. Leaving out a *sys* sample with the DAGJK does not do well at estimating the variance from a *sys* sample design.

Part 2: Replication methods for Second-Stage Variance

- Modified Successive Difference Replication (SDR2)
- Modified Balanced Repeated Replication (BRR2)
- Rizzo and Rust [2011] (RR)

$$v(\hat{Y}) = v_{\text{PSU}}(\hat{Y}) + v_{\text{SSU}}(\hat{Y})$$

We want an estimator for:

$$\hat{v}_{\text{SSU}}(\hat{Y}) = \sum_{h} \sum_{i \in s_h} \frac{\hat{v}(\hat{Y}_i)}{\pi_i^2}$$

Second-Stage Replicate Variance Estimation

- Modified Successive Difference Replication (SDR2)

$$F_{k,r} = 1 + \sqrt{1 - f_i}\left(2^{-\frac{3}{2}} h_{a_k,r} - 2^{-\frac{3}{2}} h_{b_k,r}\right)$$

- Modified Balanced Repeated Replication (BRR2)

$$F_{k,r} = \begin{cases} 1 + (1-k)a_{r,k} & k = 1 \\ 1 - (1-k)a_{r,k} & k = 2 \end{cases}$$

- Rizzo and Rust [2011] (RR)

$$F_{k,r} = \begin{cases} 1 + (1-k)a_{r,h} & j \in A_i \\ 1 - (1-k)a_{r,h} & j \in D_i \\ 1 & \text{otherwise} \end{cases}$$

Thanks

stephen.eliot.ash@census.gov