Effect of Nearest Neighbor Imputation on Variances Calculated by Fay's Balanced Repeated Replication

Brad Rhein, Chester Ponikowski, Leland Righter

Bureau of Labor Statistics

FCSM 2018

March 8, 2018

Outline

- Introduce the Occupational Requirements Survey (ORS)
- Variance Issue
- Empirical Evaluation of the *current* and *proposed* methods
- Conclusion

Introduction to ORS

Disability Programs

Introduction to ORS

Job Preparation	Cognitive Elements	Physical Demands	Environmental Conditions
	Decisions?		
 	Supervised?		
	Adaptability?		
	Contact - With whom? Frequency? Type?		

ORS Products

- Published estimates for 338 occupations (2017), including percentage of workers, means, and percentiles (82% of US employment represented, www.bls.gov/ors)
  - Percentage of statisticians that occasionally lift more than 10 lbs
  - Mean hours of standing or walking for a college professor
- News releases, The Economics Daily (TED), infographics
- Job Profiles
- Data Finder

ORS Sample Design

- Establishment survey; collected nationally, annually
- Stratified by ownership/industry groups and census region; allocated by establishment employment size
- Two stage selection: probability proportional to employment size (PPS) sample of establishments, then jobs (quotes)
- Panel survey with 3-year rotation

Non-response Types

Weight Adjustment

Job
(Quote)

Establishment
(Unit)

Degree, for example
(Item)

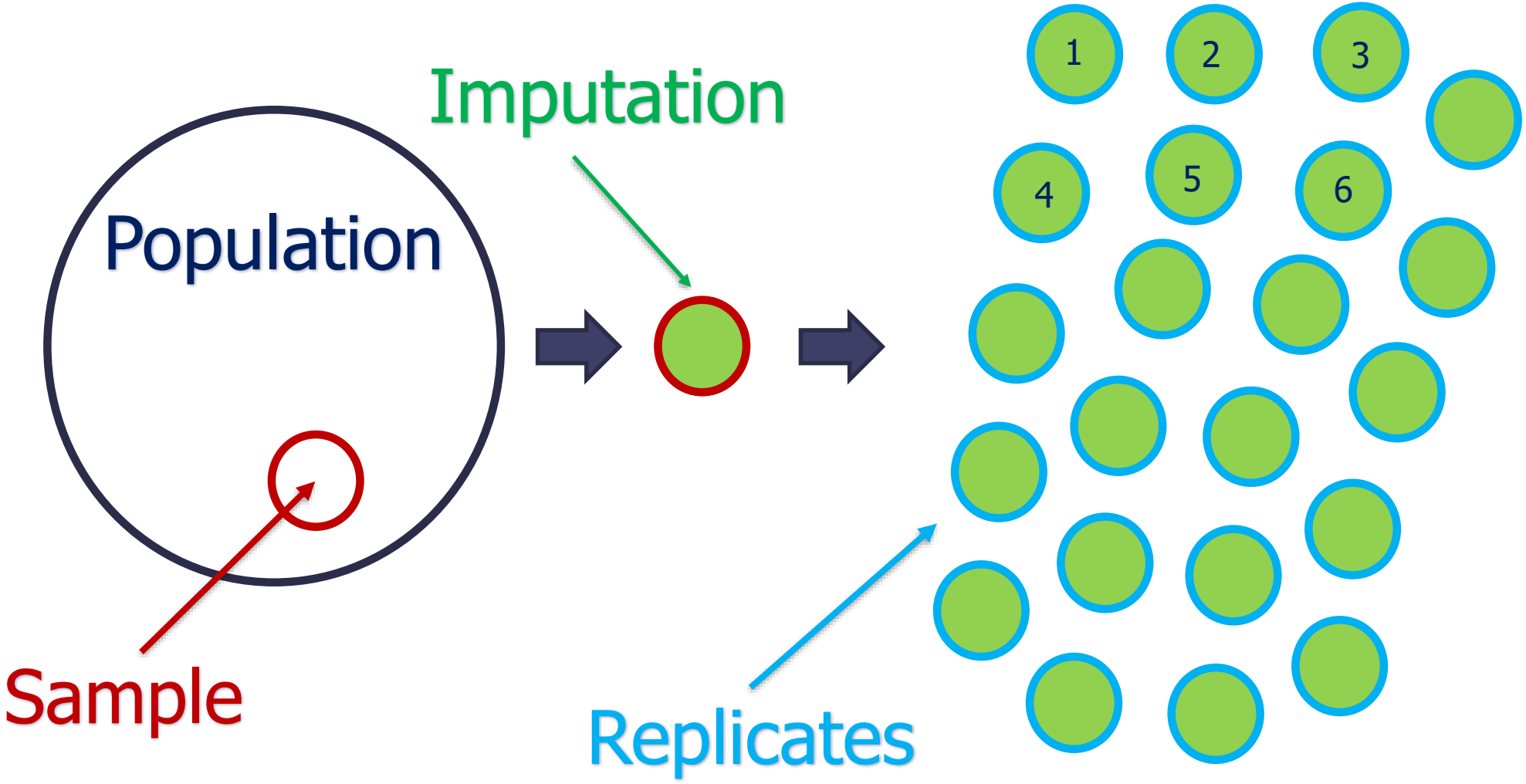
Imputation Method

- Random hot deck nearest neighbor imputation
  - Imputation cell defined by ownership, SOC, industry, size, union, full time status
  - Nearest neighbor within the cell is chosen by the smallest difference in establishment employment size
  - **Ties decided by a random process**
  - Donor reuse capped at 3, initially
  - Collected values donated to recipients

Variance Method

- ORS estimated variances are computed by Fay's Balanced Repeated Replication (BRR) method
  - Variance strata are defined by the sampling strata
  - Form replicate "half-samples" facilitated by a Hadamard matrix, using the variance strata and PSUs defined in sampling
  - All quotes used in all replicates

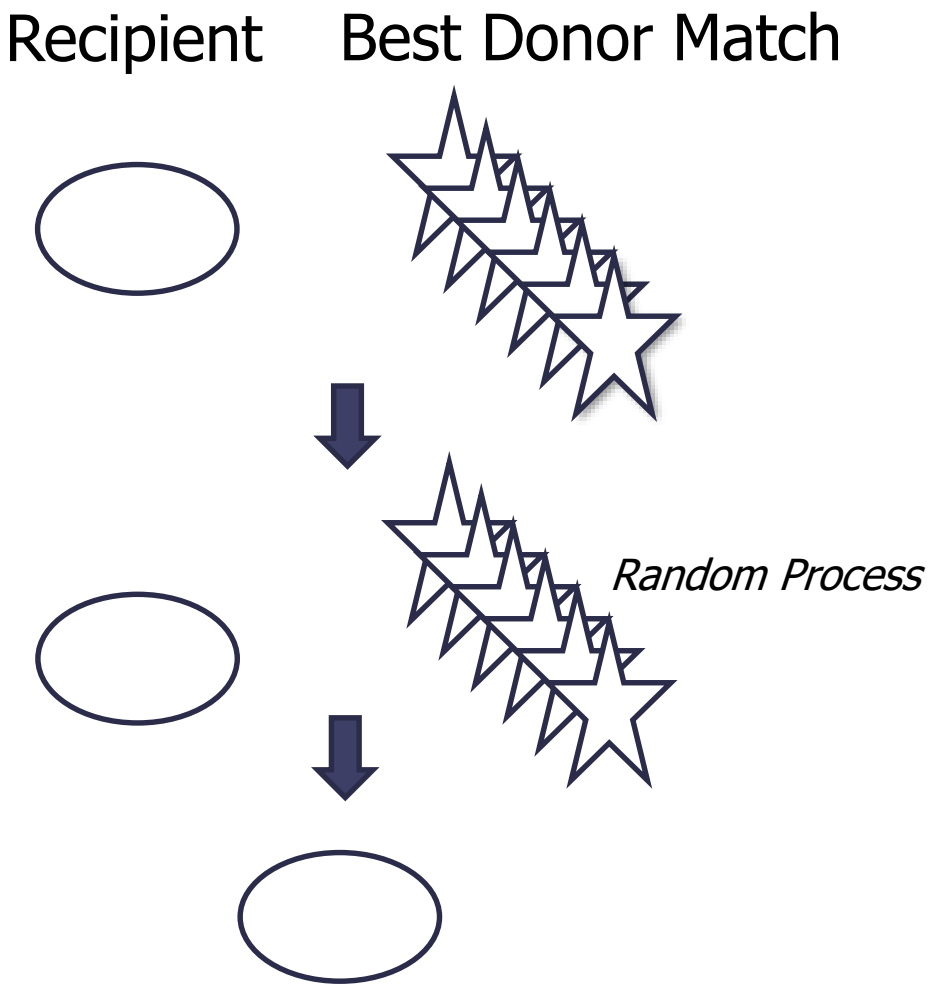
Current Method



Variance Issue?

- Ties are a source of randomness… the current variance estimation does not attempt to account for this variability
  - Is it necessary to do the imputation by replicate, or is it sufficient to do the imputation only on the full sample

Ties! What? How many?



Group Label	Percentage of matches determined by ties
Driving	11.42%
Vision	14.34%
Hearing/speaking	13.62%
Heat/cold/humid	14.23%
Hazardous	13.86%
Climbing	14.94%
Postural	13.84%
Keyboarding	15.00%
Manipulation	10.05%
Pushing Legs	13.84%
Pushing Arms	13.77%
Sitting/Standing	11.43%
Lift/carry	13.86%
Education	11.39%

Ties! Why?

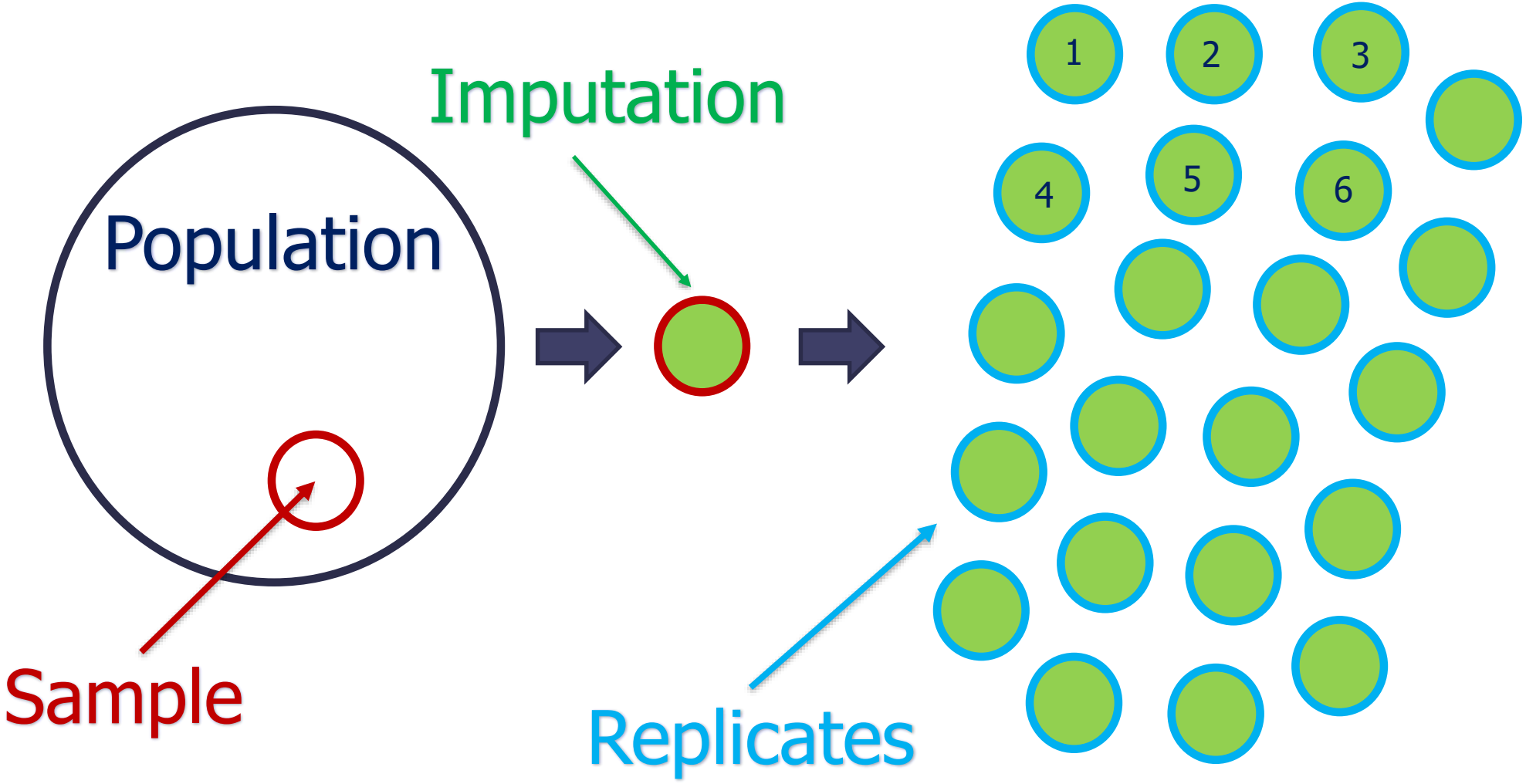
Job #	SOC	Title	Full/Part Time	Salary/Incentive
1	119000	Manager	Full	Salary
2	412000	Sales Rep	Part	Incentive
3		2		
4	499000	Maintenance	Full	Salary
5	412000	Senior Sales Rep	Full	Salary
6	412000	Sals Rep	Part	Incentive
7		4		
8		5		

"Unique quotes" cause 90% of the ties

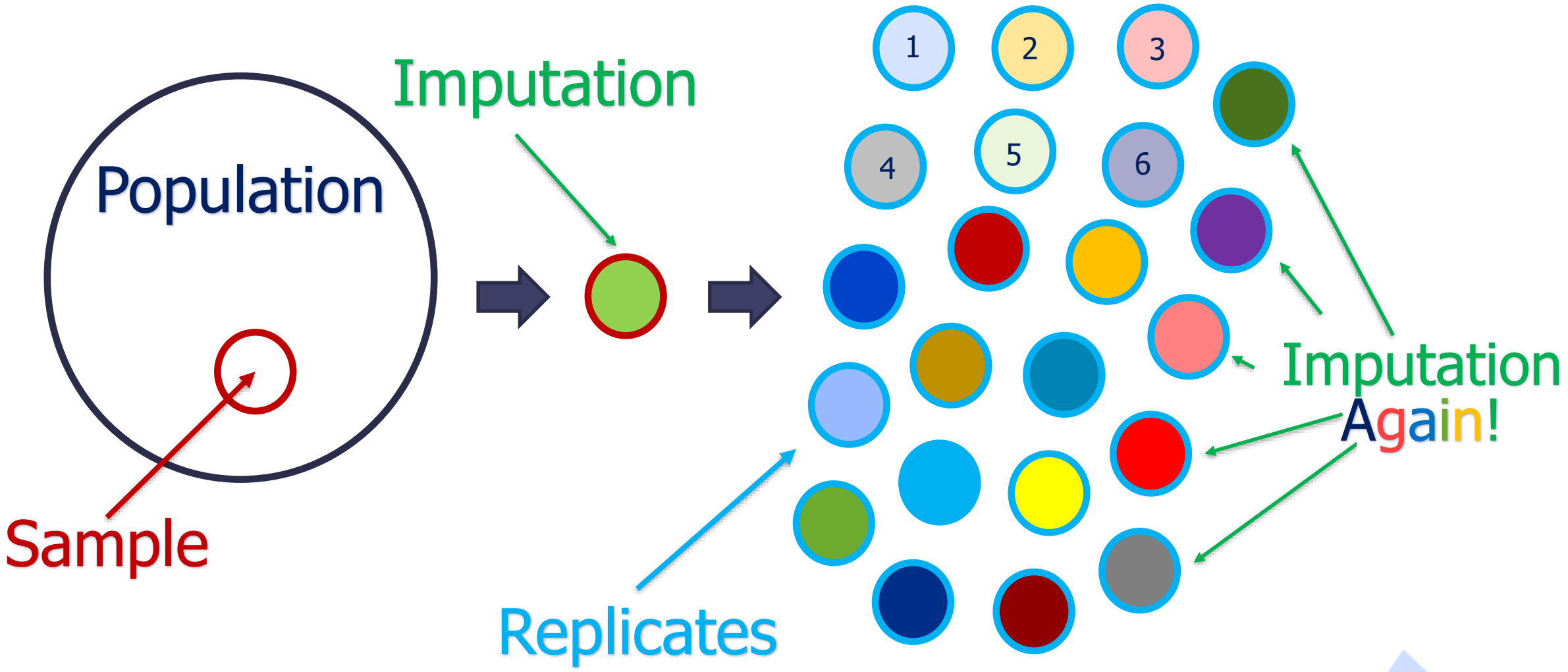
- Same SOC code
- Same job title
- Same worker characteristics
- Same ORS data

Or the same everything!

Current Method



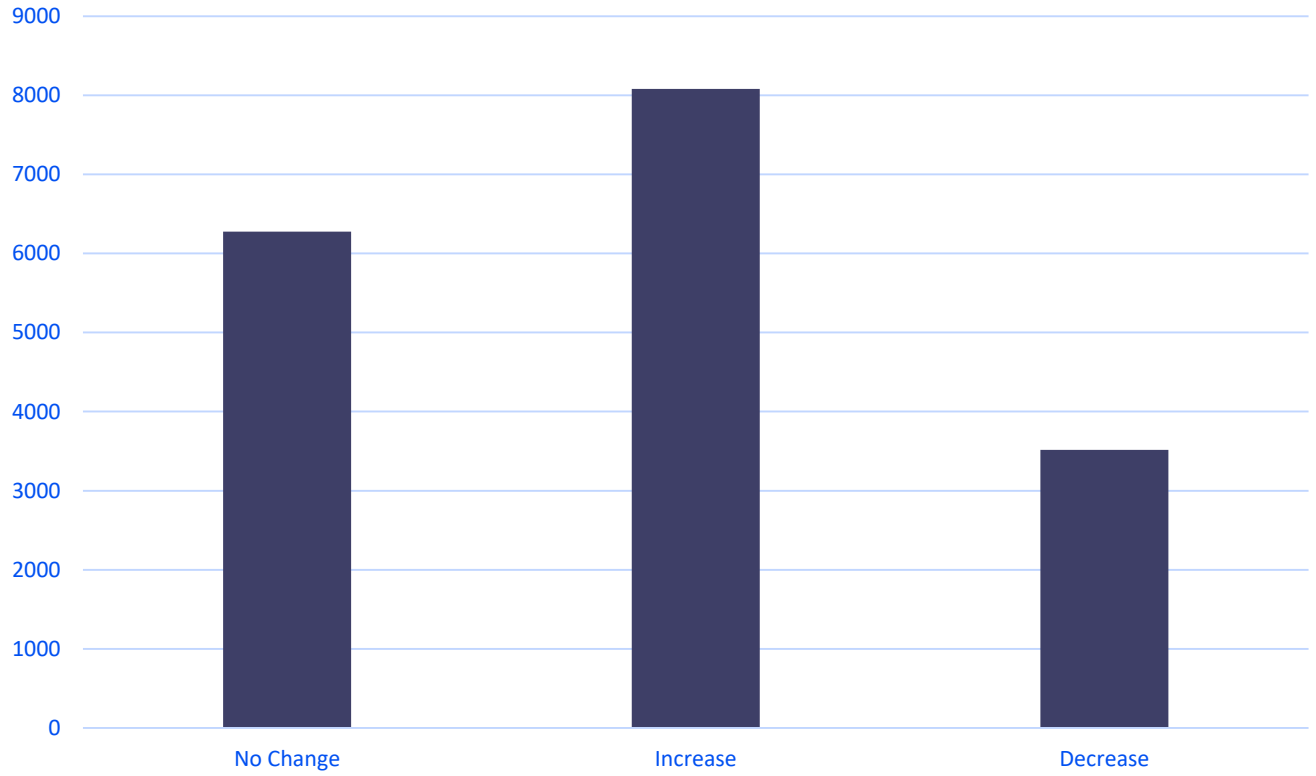
Proposed Method



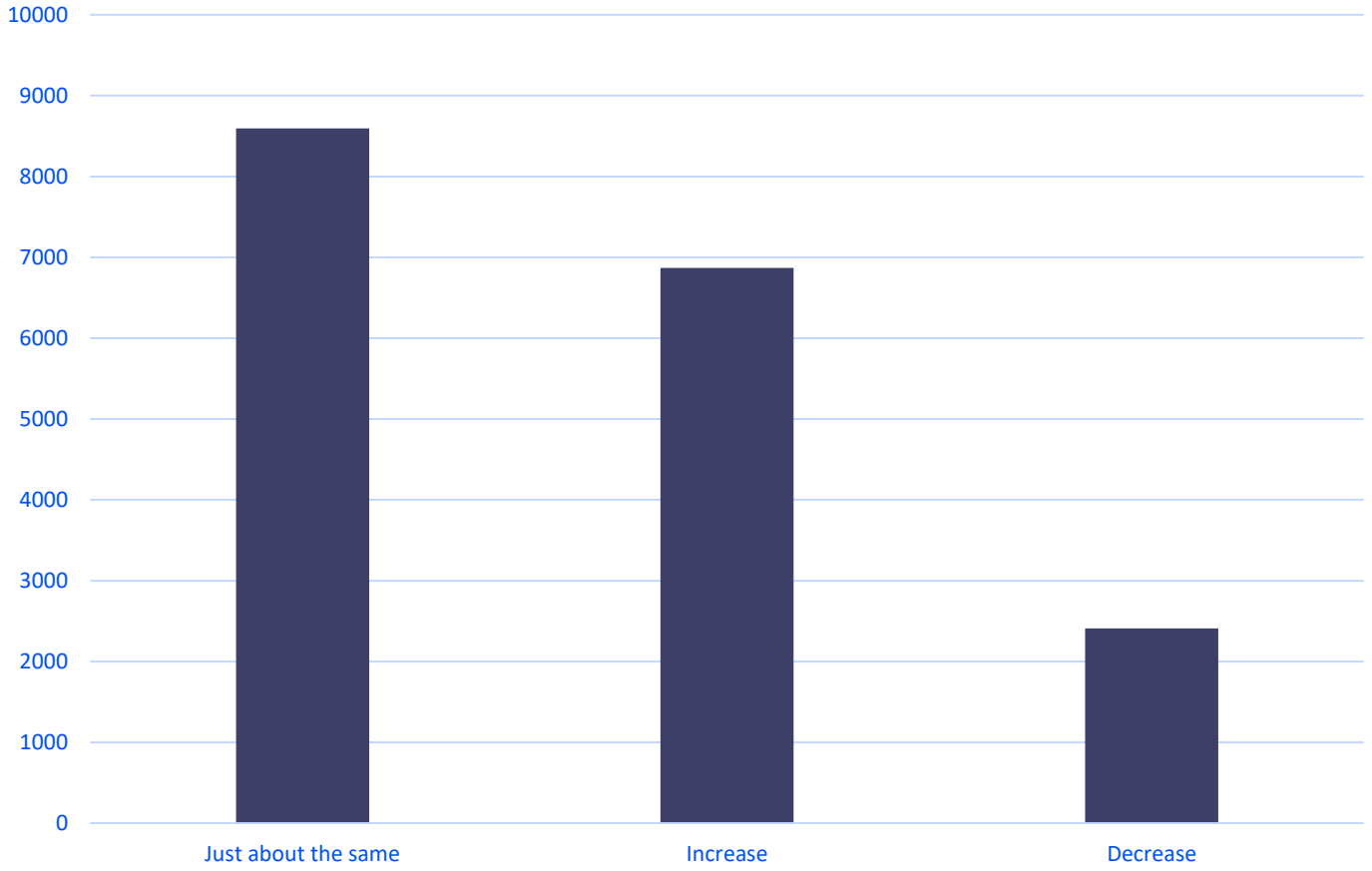
Empirical Evaluation

- Chose 6 data elements that are “representative” of all elements
  - On the job training (11% matches decided randomly)
  - Standing/walking (11%)
  - Pushing with the upper body (14%)
  - Peripheral vision (14%)
  - Exposure to hazardous contaminants (14%)
- Ran variance estimates for both methods
- Compared around 17,000 publishable variances

How do the Proposed Standard Errors Compare to the Current Method?



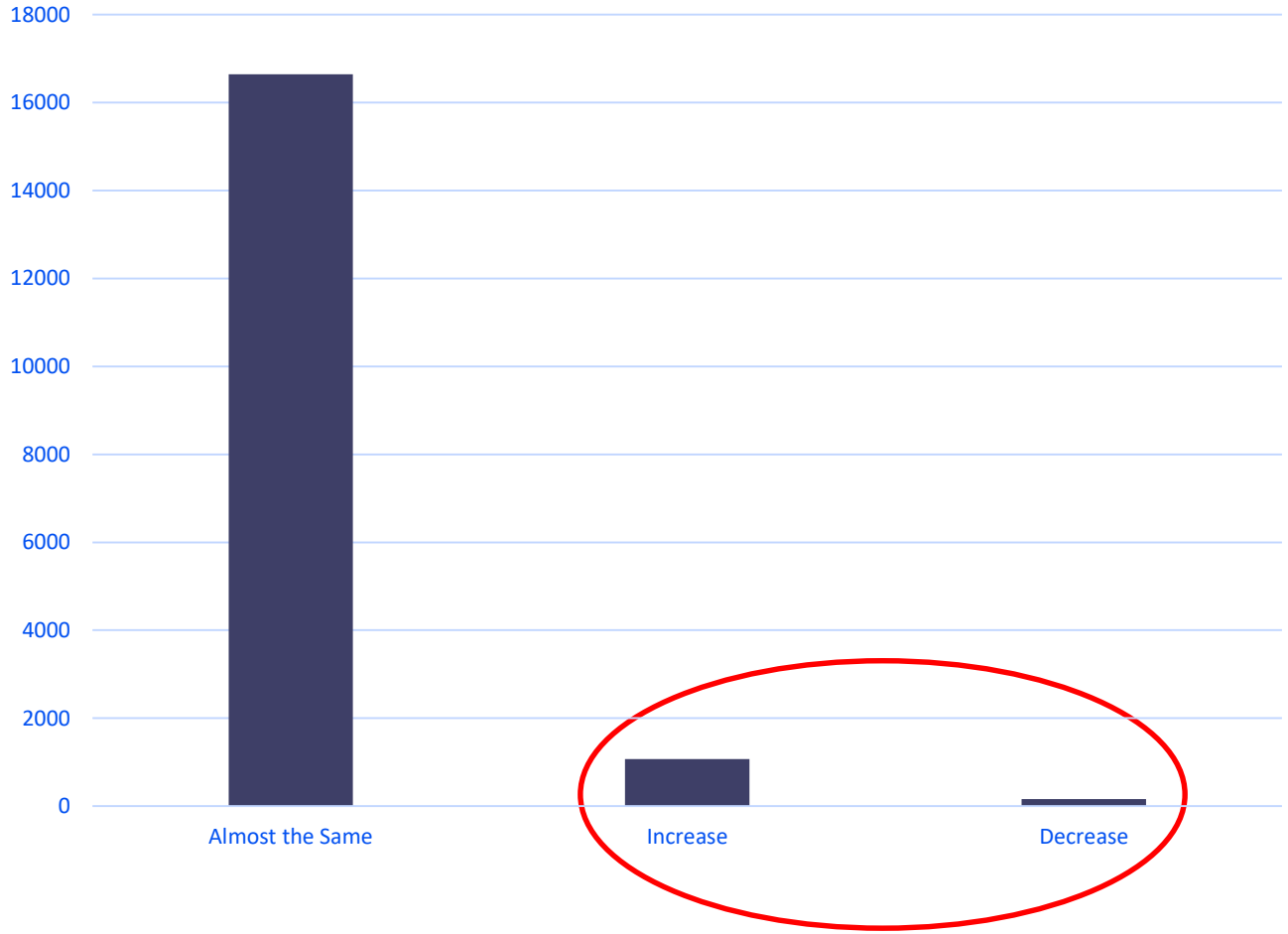
What About Comparing at the 0.0001 Level?



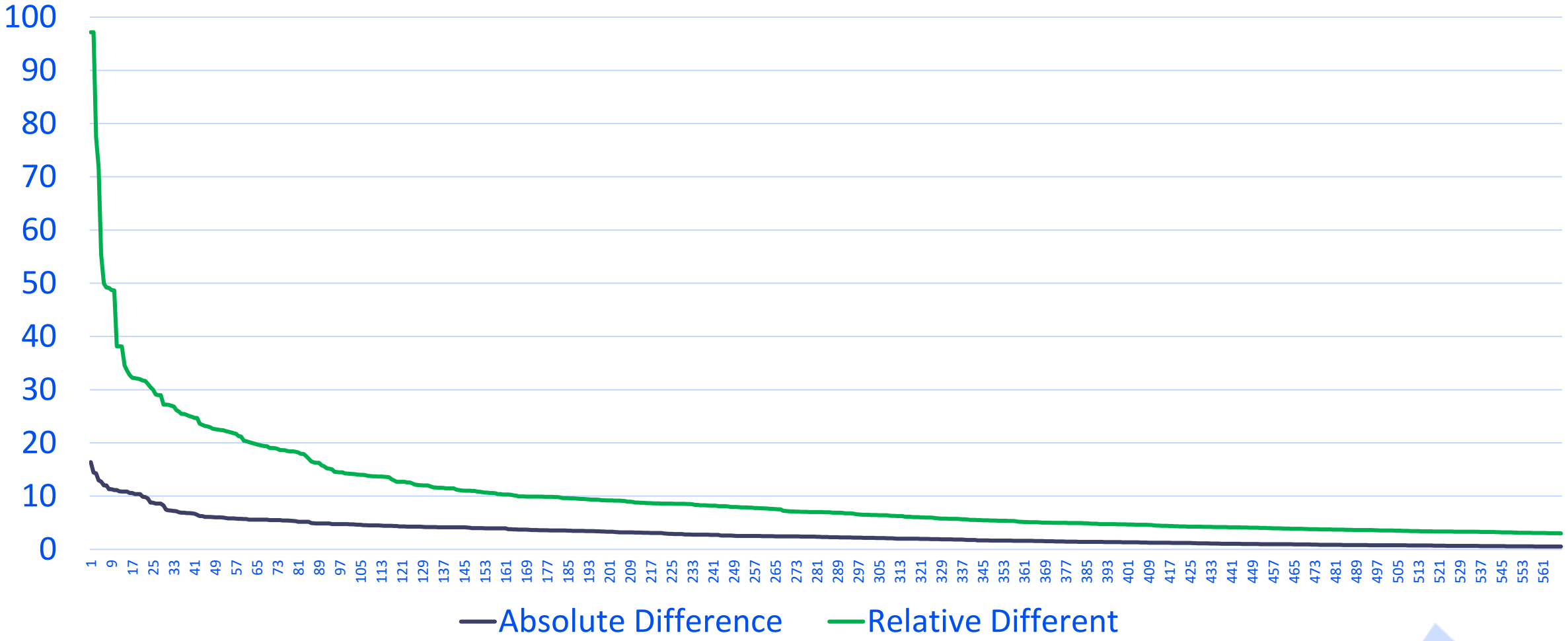
Empirical Evaluation

- Comparing standard errors, the more complex way
  - Find the absolute difference in standard errors
  - Find the relative difference in standard errors
    - *(absolute standard error difference/estimate value)*
  - Combine these metrics to form a new criteria
- "Almost the same" is now:
  - Absolute standard error difference < 0.5 and
  - Relative standard error difference < 3%

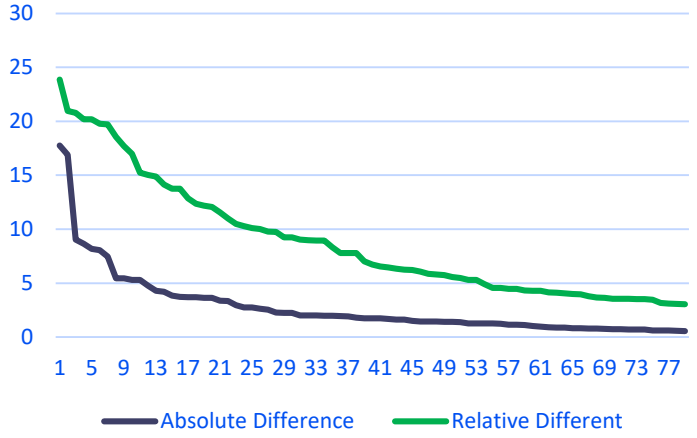
How do the Proposed Standard Errors Compare to the Current Method?



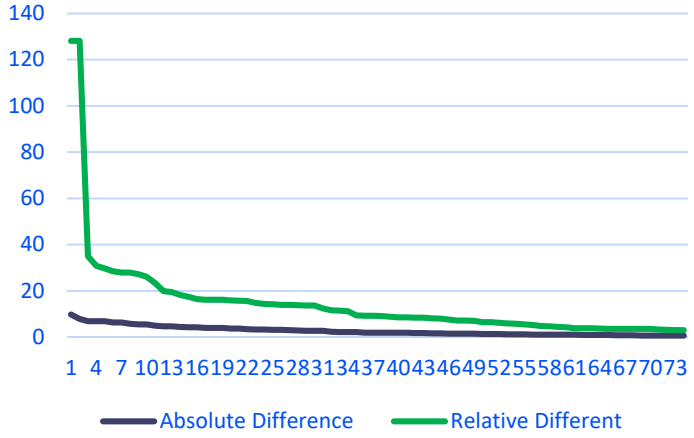
Absolute vs Relative Differences for Percentage Estimates



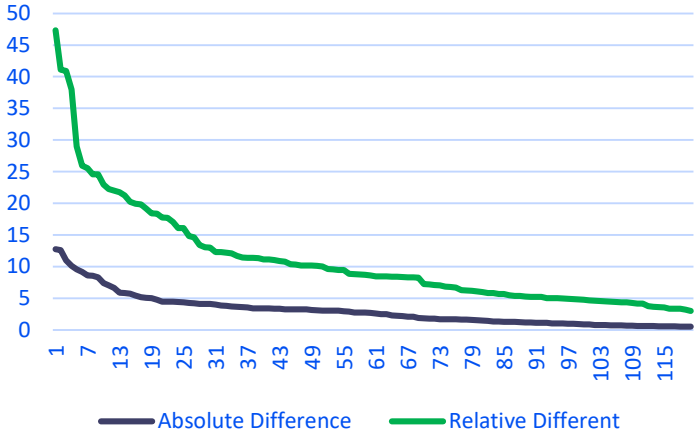
Means



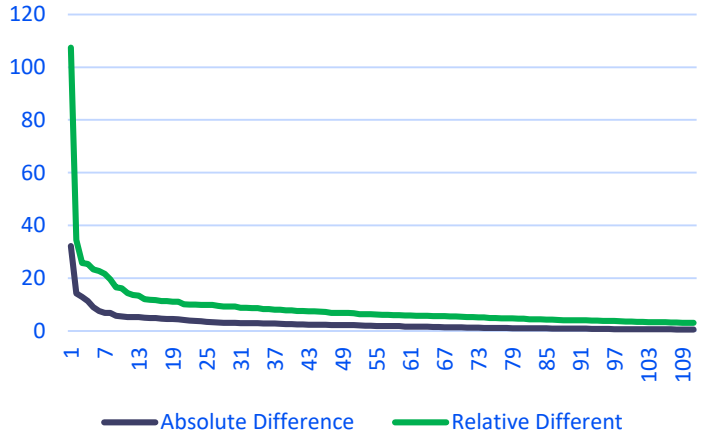
10th Percentile



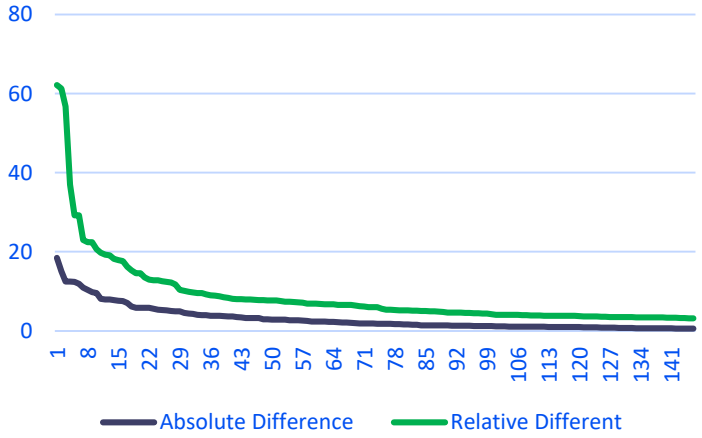
25th Percentile



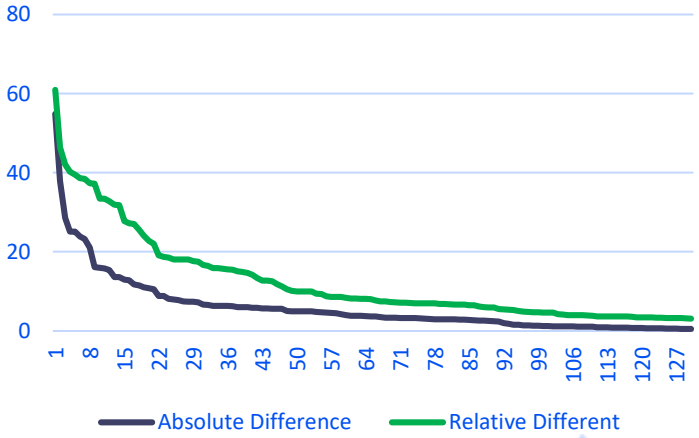
50th Percentile



75th Percentile



90th Percentile



Conclusions

- Differences in the estimated variances show no substantive pattern – some increase, some decrease, most are the same
  - When there is a difference, the proposed variance is usually larger than the current variance
- The effect of ties on estimated variances is minimal
  - 93% of the published variance data is “almost the same” when comparing the current and proposed methods

Contact Information

Bradley Rhein
Mathematical Statistician
Statistical Methods Group
www.bls.gov/ors
202-691-6116
rhein.bradley@bls.gov