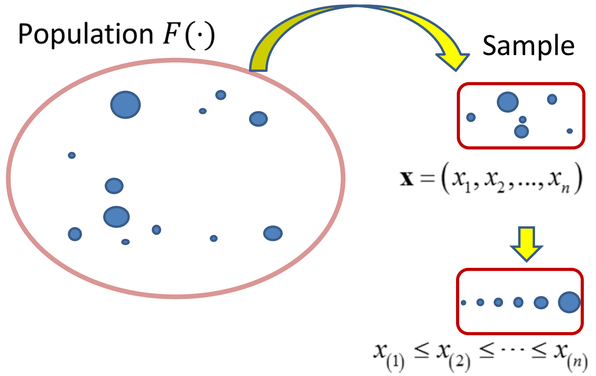
Assessing the Relationship between Balanced Sample and Sample Representativity*

Yonil Park (Yonil.Park@census.gov), John Chesnut (Thomas.J.Chesnut@census.gov)
Demographic Statistical Methods Division, U.S. Census Bureau, Washington, DC 20233

* Disclaimer: The views in this poster are those of the authors, and do not necessarily represent those of the Census Bureau. Disclosure Avoidance Officer Approval number : CBDRB-FY19-ROSS-B0165

1. Introduction



Representative Index (RI)†

$$R(\boldsymbol{x}, F) = 1 - \frac{12n}{4n^2 - 1} \sum_{r=1}^{n} \left(F(x_{(r)}) - \frac{2r-1}{2n} \right)^2$$

$$0 \le R(\boldsymbol{x}, F) \le 1$$

- measure of how much the observed sample represents its population
- min-max normalized index using Smirnov-Cramér-von Mises test statistic
- distribution-free measure

† Bertino, S. (2006), A Measure of Representativeness of a Sample for Inferential Purposes. *International Statistical Review*, 74: 149-159.

2. Methodology

Representative Sample

$RI, R(\mathbf{x}, F)$

Simulation

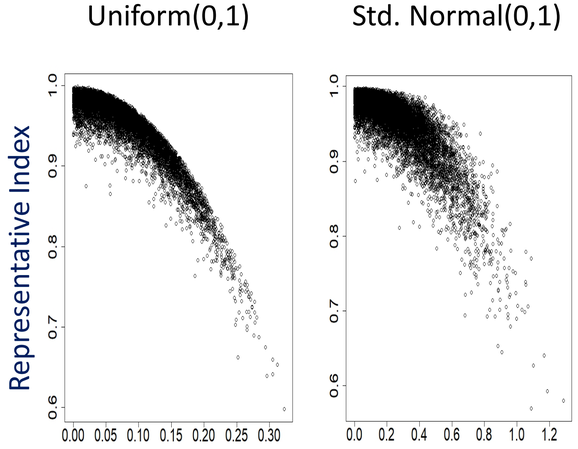
$$\sum_{i \in s} \frac{x_i}{\pi_i} = \sum_{i \in U} x_i$$

Balanced Sample

3. Simulation Study

Samples from N(0,1)	RI
balanced	0.99
left Skewed	0.60
decentralized	0.88

RI VS. Sampling Error



10,000 runs with each sample size = 10

Length-biased Correction

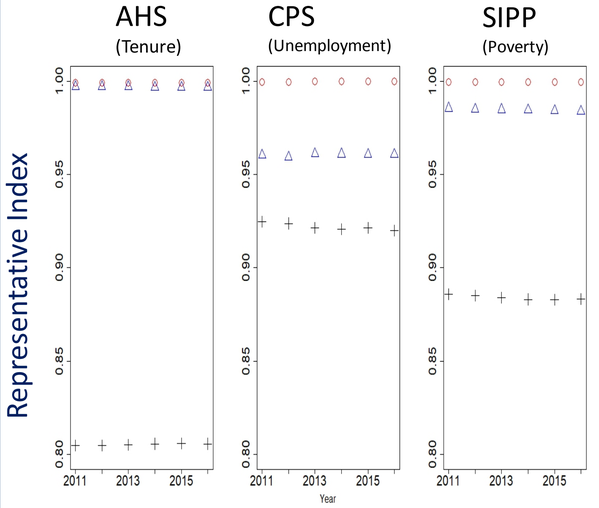
$$f(x) \leftarrow \frac{x f(x)}{E(X)}$$

- Inspection paradox in renewal process
- Randomly selected sample tends to be larger length than regular sample.

2010 Sample Redesign PSU Selection Probability

$$f(x_i) = \frac{\pi_i}{\sum_{i=1}^{N} \pi_i} = \frac{\frac{m_i}{M_i}}{\sum_{i=1}^{N} \frac{m_i}{M_i}} = \frac{\frac{m_i}{M_i}}{n} = \frac{m_i}{nM_i}$$

Evaluating the 2010 Sample Redesign PSUs



o : probability proportional to size (Sample Redesign PSUs)
Δ : length-biased correction
+ : simple random sampling probability

4. Conclusion

- There is a strong positive relationship between balanced sample and representative sample.
- 2010 Sample Redesign PSUs showed very high RI.
- Length-biased correction to RI has a potential to improve the utility of the RI for a simple random sampling.