Optimal Cutoff Sampling for the Annual Survey of Public Employment and Payroll

Brian Dumbacher
Carma Hogue

Governments Division
U.S. Census Bureau

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

Outline

- Survey Overview
- Sampling Design
- Small Area Estimation
- Objective
- Notation
- Methodology
- Results
- Future Work

Survey Overview

- The Annual Survey of Public Employment and Payroll (ASPEP) is conducted by the Governments Division of the U.S. Census Bureau to collect data on government civilian employees and their gross payrolls

- ASPEP consists of three parts:
  - Census of select federal agencies
  - Census of the 50 state governments
  - Sample of about 10,500 local governments

- Small area methods are used to estimate local government totals for each combination of state and government function

Local Government Functions

- Air transportation
- Corrections
- Education
- Police protection
- Fire protection
- Financial admin.
- Social insurance admin. (DC)
- Other government admin.
- Judicial and legal
- Health
- Hospitals
- Highways
- Libraries
- Housing and community development
- Natural resources
- Parks and recreation
- Public welfare
- Sewerage
- Solid waste management
- Water transport and terminals
- Water supply
- Electric power
- Gas supply
- Transit
- Other and unallocable

Sampling Design

- Sampling frame for ASPEP is a list of the approximately 90,000 governments identified in the 2007 Census of Governments

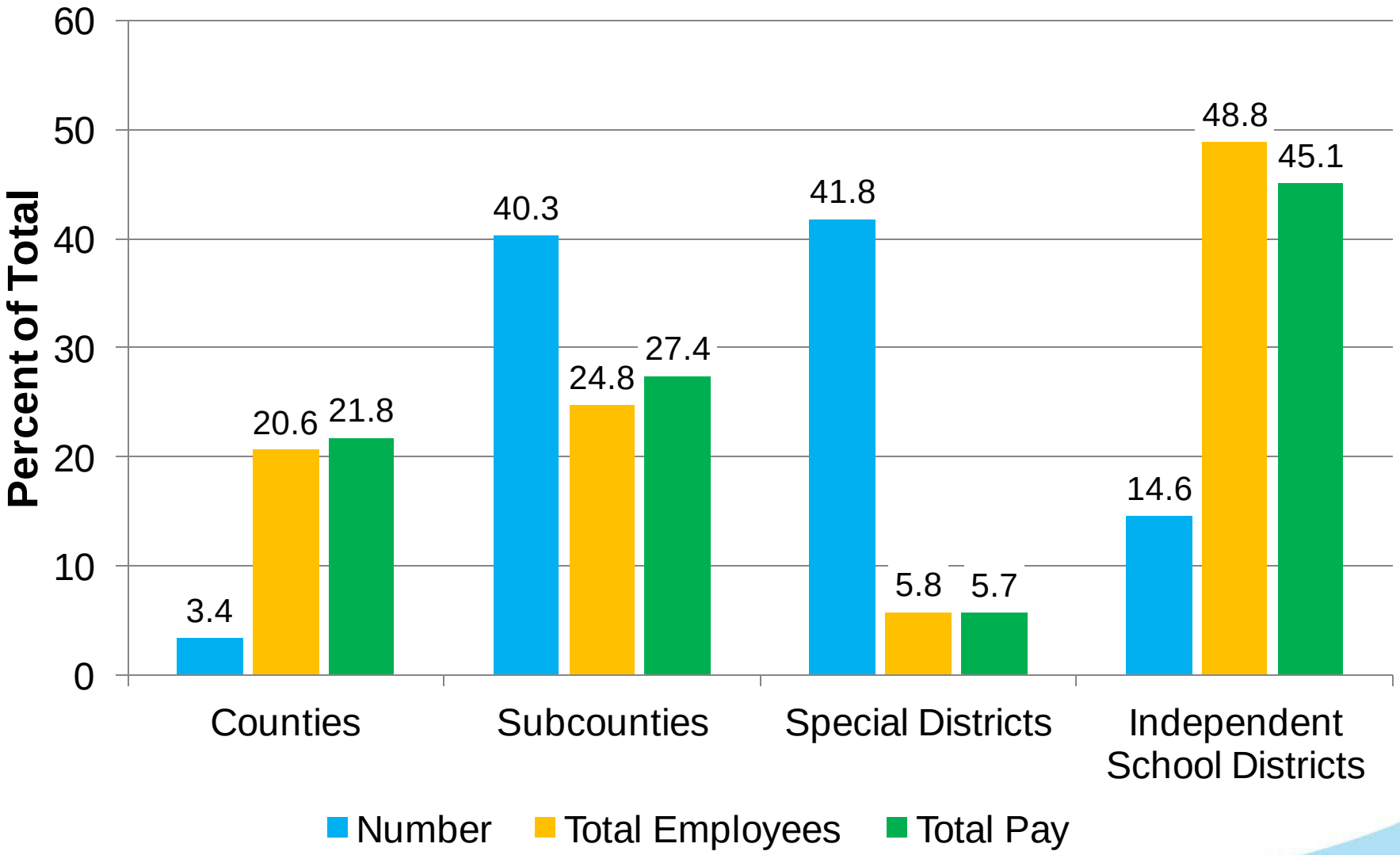
- Updated annually with births, deaths, and mergers

- Two-stage, stratified, πps design
  - Certainty criteria
  - Measure of size is total pay in 2007
  - Strata formed by state × government type

Types of Local Government

- Counties
- Subcounties
  - Cities
  - Townships
- Special districts
- Independent school districts

Local Governments in 2007



Source: 2007 Census of Governments: Employment

Modified Cutoff Sampling

- Modified cutoff sampling in the subcounty and special district strata
- Cumulative Square Root of the Frequency (CSRF) method used to split the first-stage samples into a small and large cutoff stratum
- Subsampling performed in the small cutoff strata
  - Subsampling rate = 0.56
  - Target reduction of 800 units

Small Area Estimation

f = state g = function

- Composite estimator

$$\hat{t}_{fg} = \hat{\varphi}_f \, \hat{t}_{fg}^{HT} + \left(1 - \hat{\varphi}_f\right) \hat{t}_{fg}^{SYN}$$

- Synthetic estimator

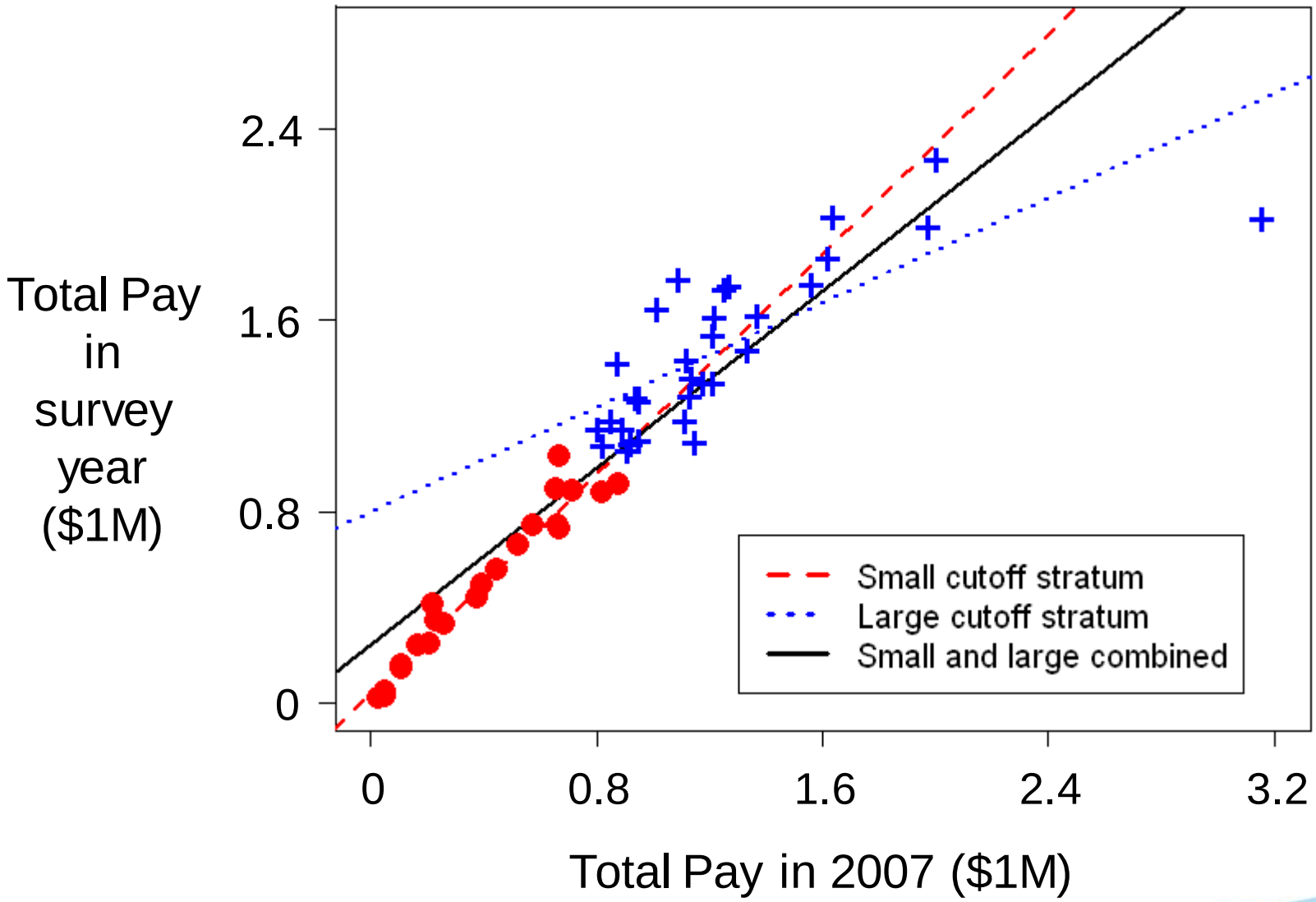
$$\hat{t}_{fg}^{SYN} = \hat{t}_f^{DB} \, K_{fg}^{07}$$

K_{fg}^{07} = 2007 proportion for function within state

- Decision-based estimator

$$\hat{t}_f^{DB}$$

Illustration of Decision-Based Methodology



Source: 2002 and 2007 Censuses of Governments: Employment

Objective

- Find an optimal combination of cutoff and subsampling rate that
  - Complements the decision-based methodology
  - Attempts to control cost in some sense
- Build on work of Corcoran and Cheng (2010) who investigated numerical methods
- Minimize the average of mean squared errors ($MSEs$) from weighted linear regressions fitted separately to small and large cutoff strata

Corcoran, C. and Cheng, Y. (2010). Alternate Sample Approach for the Annual Survey of Public Employment and Payroll. Governments Division Report Series, Research Report #2010-5.

Notation

- Given a first-stage sample and specified cost C_0, find cutoff c and subsampling rate p that minimize $MSE(c,p)$ subject to $COST(c,p) \leq C_0$

n_1 Size of small cutoff stratum

n_1^* Size of subsample from small cutoff stratum

n_2 Size of large cutoff stratum

Notation (cont.)

y_i Value of total pay 2007 for unit i

$\hat{y}_i$ Predicted value of total pay 2007 for unit i from a weighted regression

w_i Sampling weight for unit i

Weighted MSE for the small cutoff stratum

$$MSE^1 = \frac{1}{n_1^* - 2} \sum_{i=1}^{n_1^*} w_i (y_i - \hat{y}_i)^2$$

Weighted MSE for the large cutoff stratum

$$MSE^2 = \frac{1}{n_2 - 2} \sum_{i=1}^{n_2} w_i (y_i - \hat{y}_i)^2$$

Methodology

- Two measures of MSE

$$MSE^{Simple} = \frac{1}{2}[MSE^1 + MSE^2]$$

$$MSE^{Pooled} = \frac{1}{n_1^* + n_2 - 4}[(n_1^* - 2)MSE^1 + (n_2 - 2)MSE^2]$$

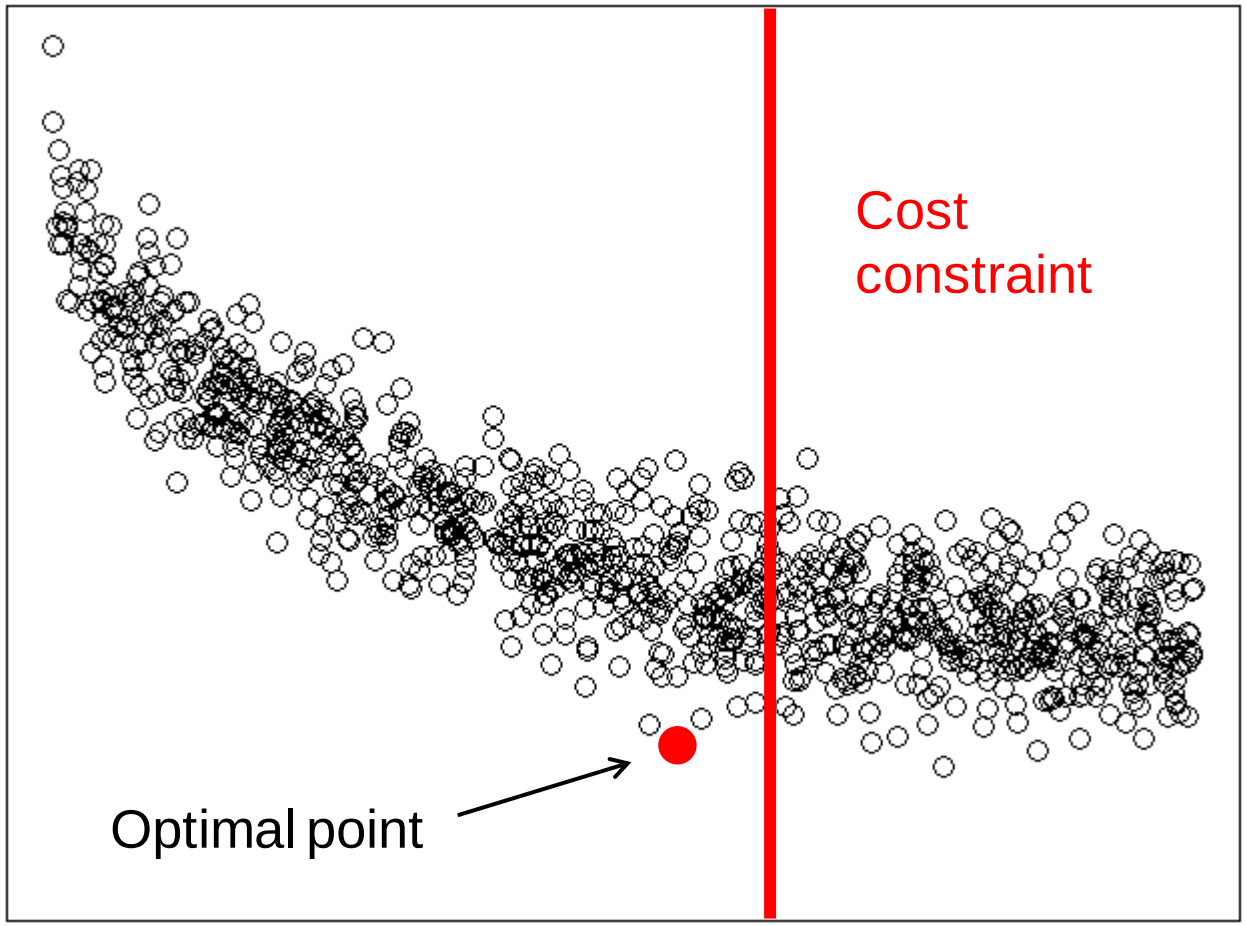
- One measure of cost

$COST_i = D$, some constant

Algorithm

- Iterate through combinations of c and p
- Select 50 subsamples for each combination
- Calculate $\overline{MSE}(c,p) = \frac{1}{50}\sum_{b=1}^{50} MSE_b(c,p)$
- Calculate $\overline{COST}(c,p) = \frac{1}{50}\sum_{b=1}^{50} COST_b(c,p)$
- Determine optimal c and p subject to cost constraint

Hypothetical Scatterplot of $\overline{MSE}(c,p)$ vs. $\overline{COST}(c,p)$



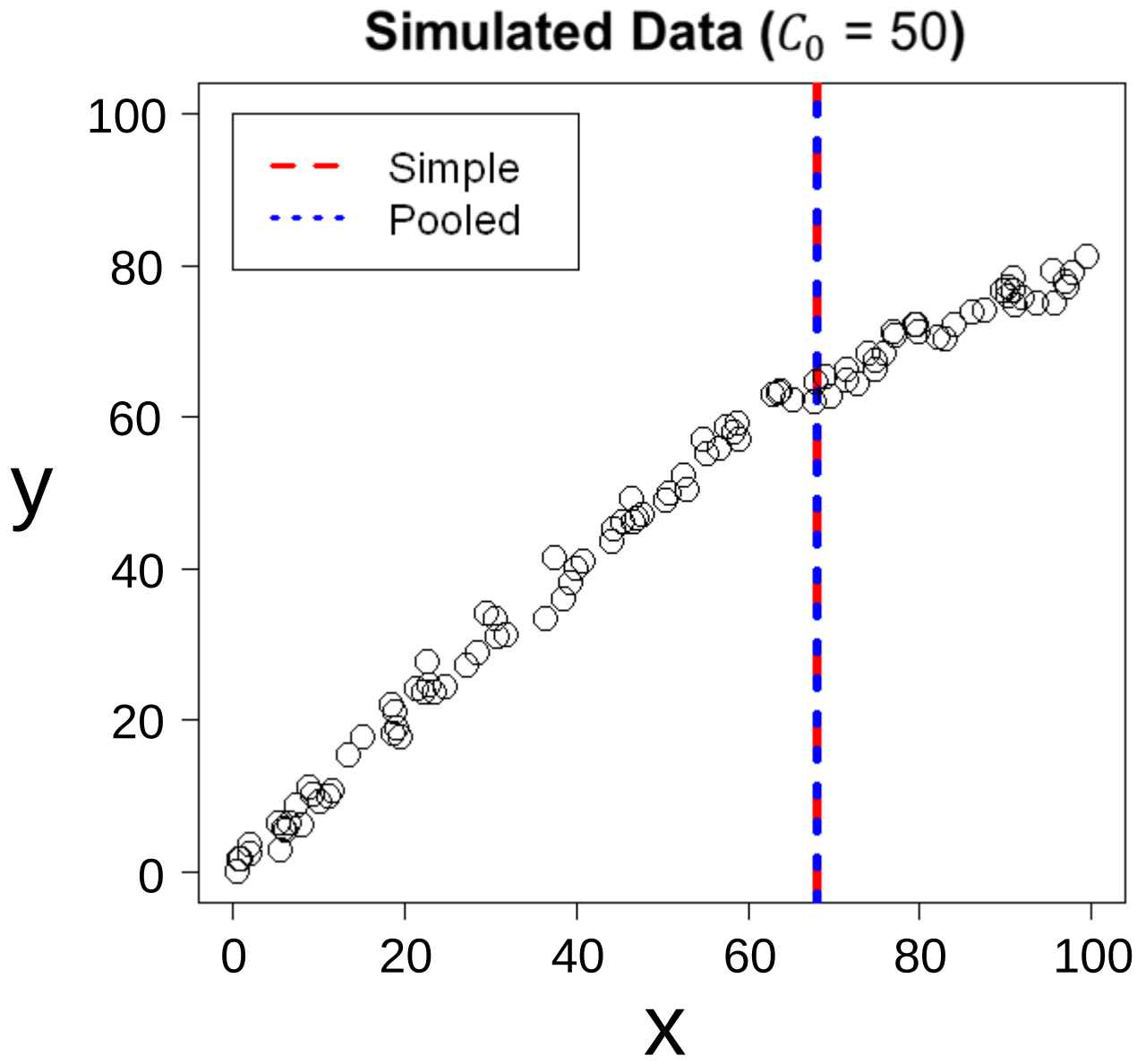
$\overline{MSE}(c,p)$

$\overline{COST}(c,p)$

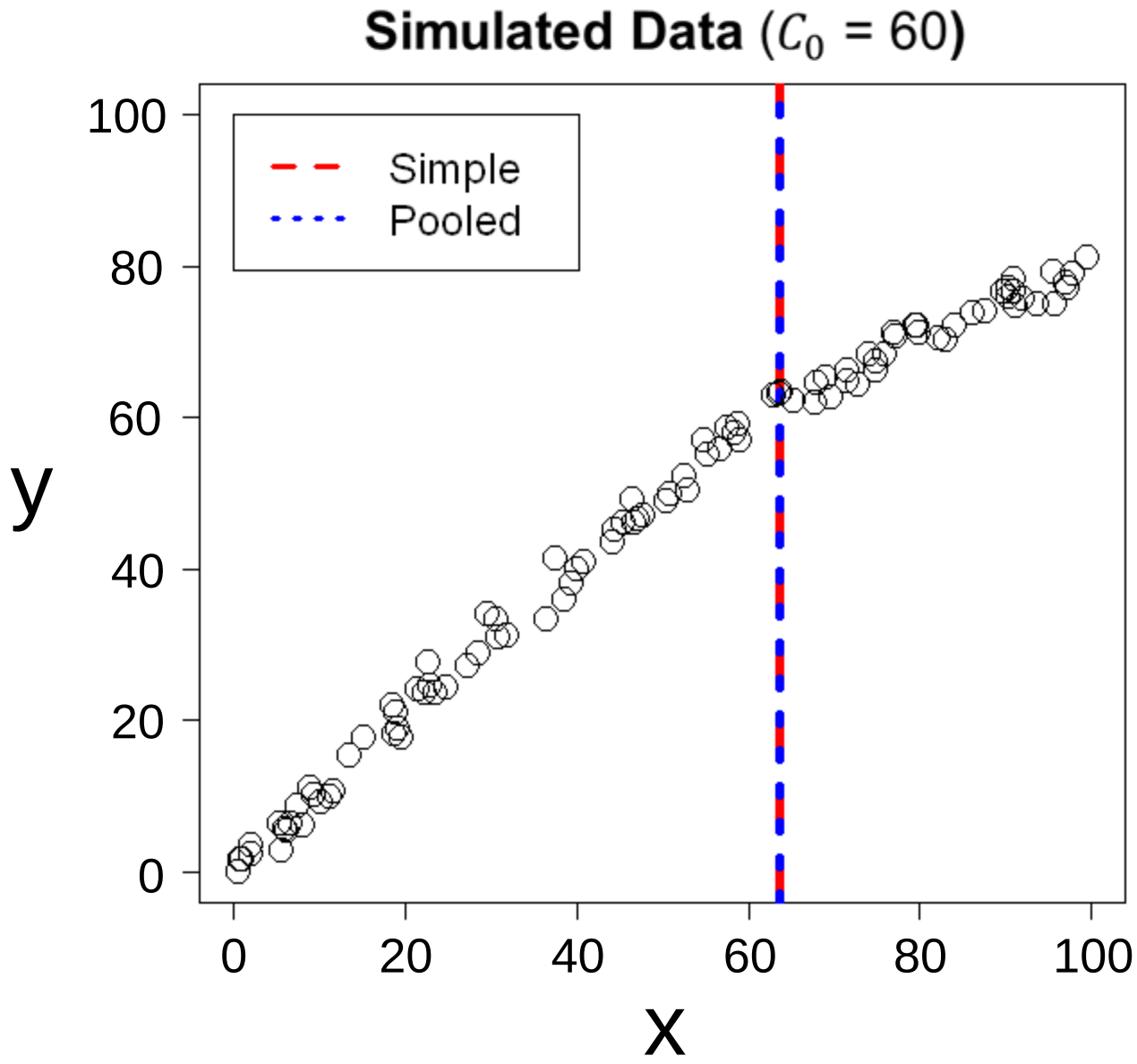
Source: Simulated data for illustrative purposes only

Results: Simulated Data

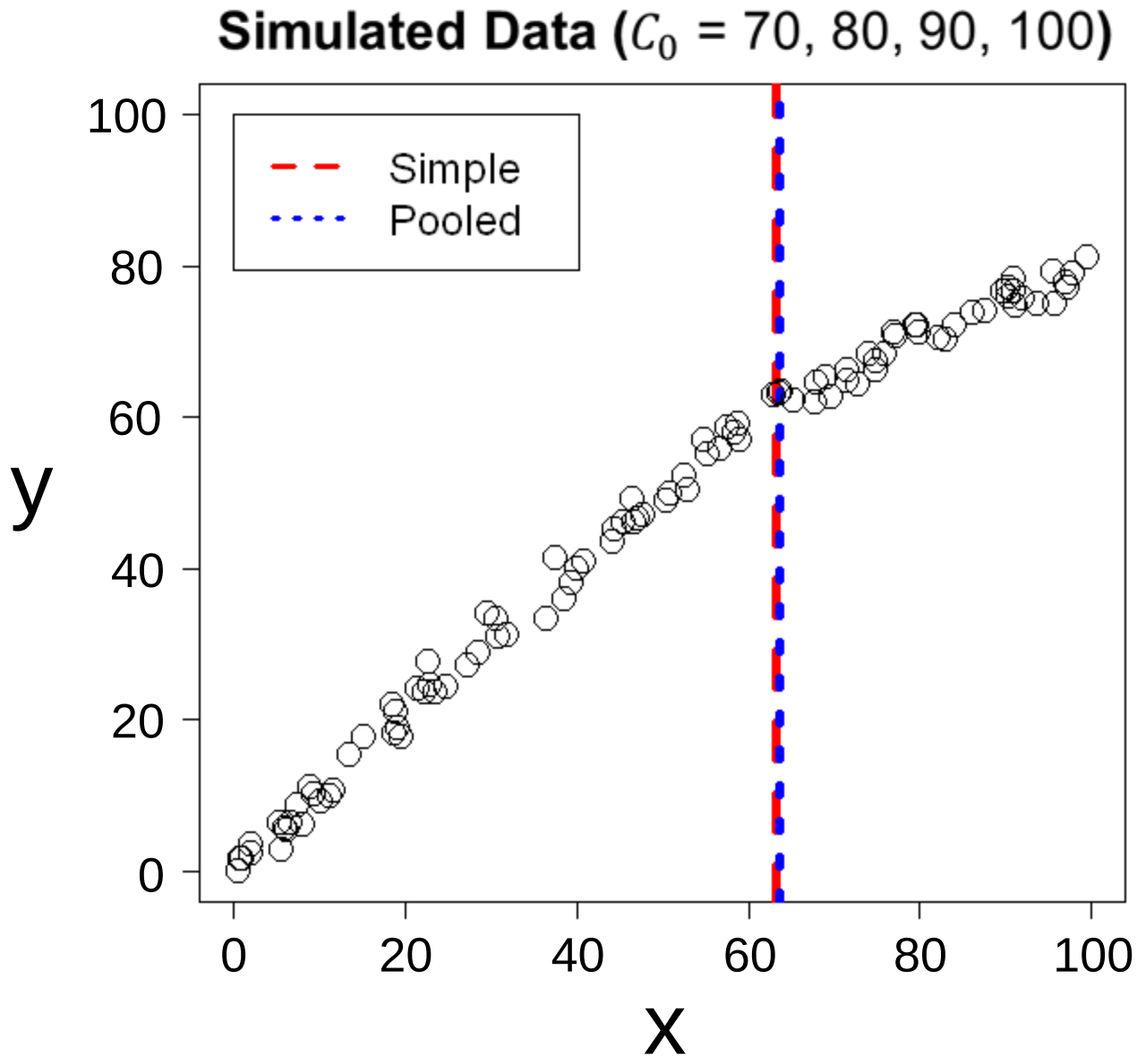
- 100 data points with equal weights
- X values
  - $x_1, \ldots, x_{60} \sim \text{Uniform}(0,60)$
  - $x_{61}, \ldots, x_{100} \sim \text{Uniform}(60,100)$
- Y values
  - $y_i = x_i + \varepsilon_i, \quad i = 1, \ldots, 60$
  - $y_i = 0.5x_i + 30 + \varepsilon_i, \quad i = 61, \ldots, 100$
- Random noise
  - $\varepsilon_1, \ldots, \varepsilon_{100} \sim \text{N}(0,1.5)$



Source: Simulated data for illustrative purposes only

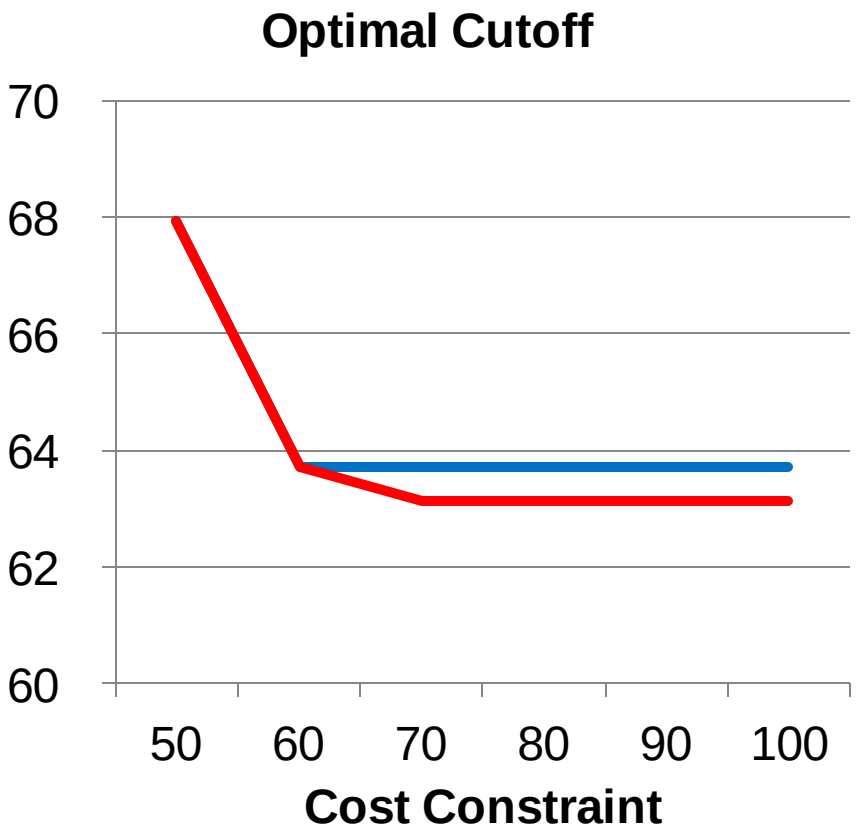


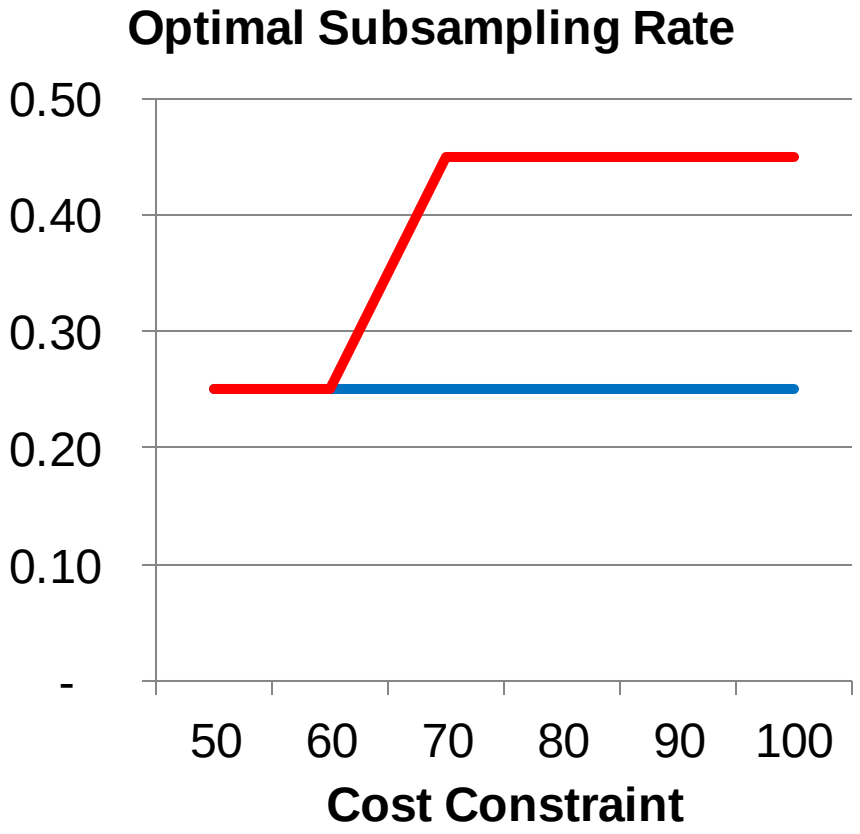
Source: Simulated data for illustrative purposes only



Source: Simulated data for illustrative purposes only

Results: Simulated Data (cont.)



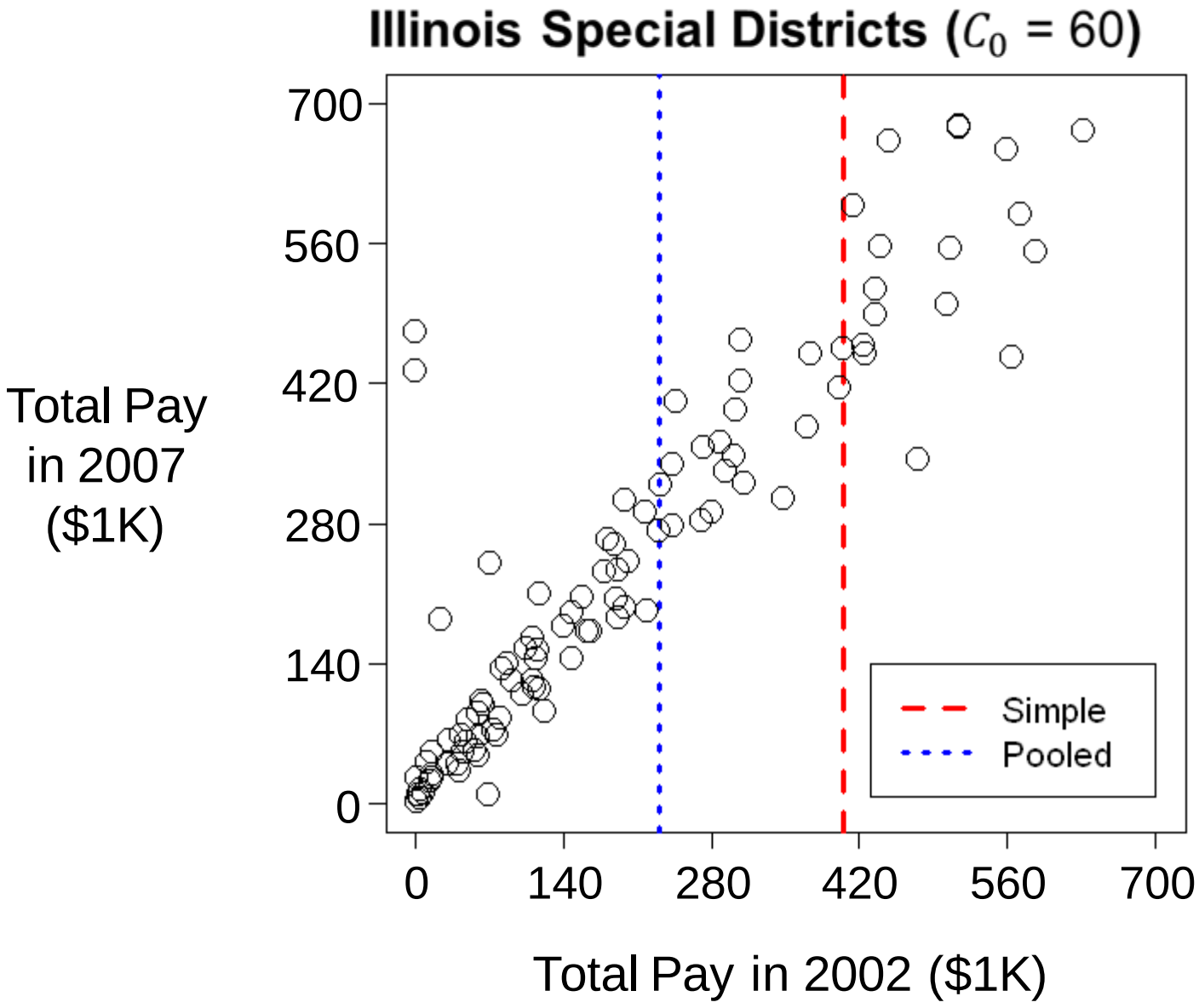


Simple Pooled

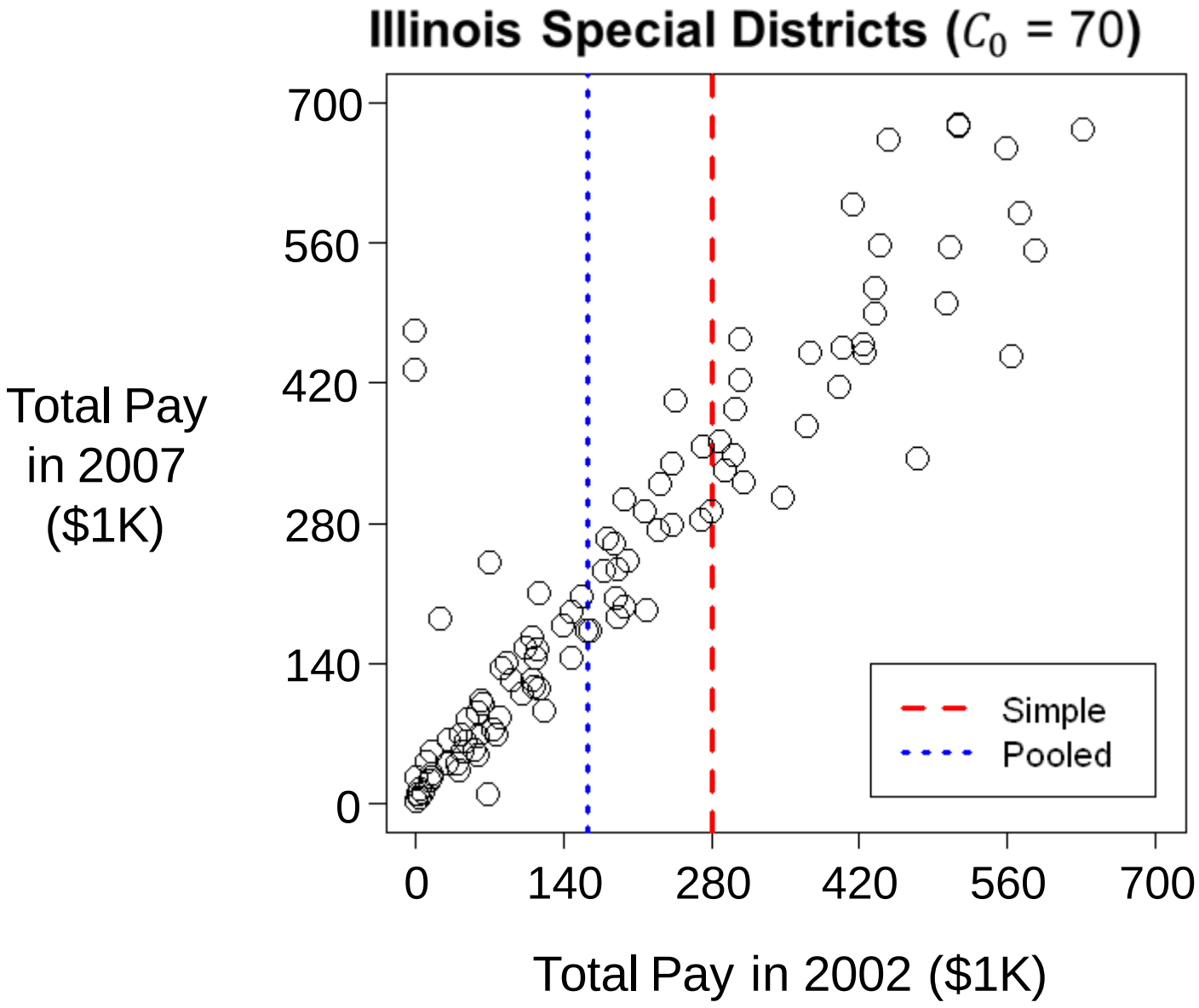
Source: Simulated data for illustrative purposes only

Results: Census Data

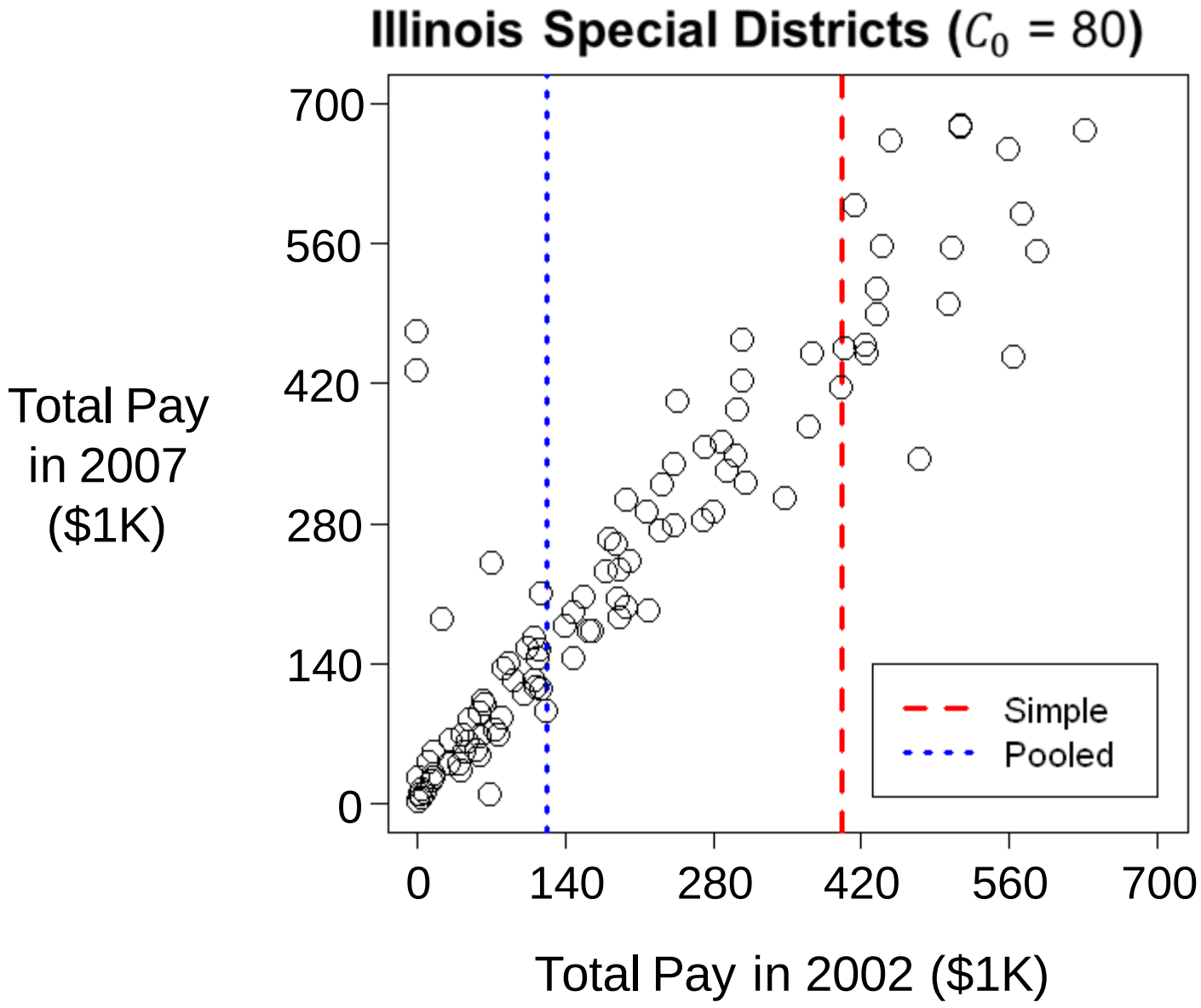
- Data from the 2002 and 2007 Censuses of Governments
- Six strata
  - California special districts
  - Illinois special districts
  - Kentucky special districts
  - New York subcounties
  - Pennsylvania subcounties
  - Wisconsin subcounties



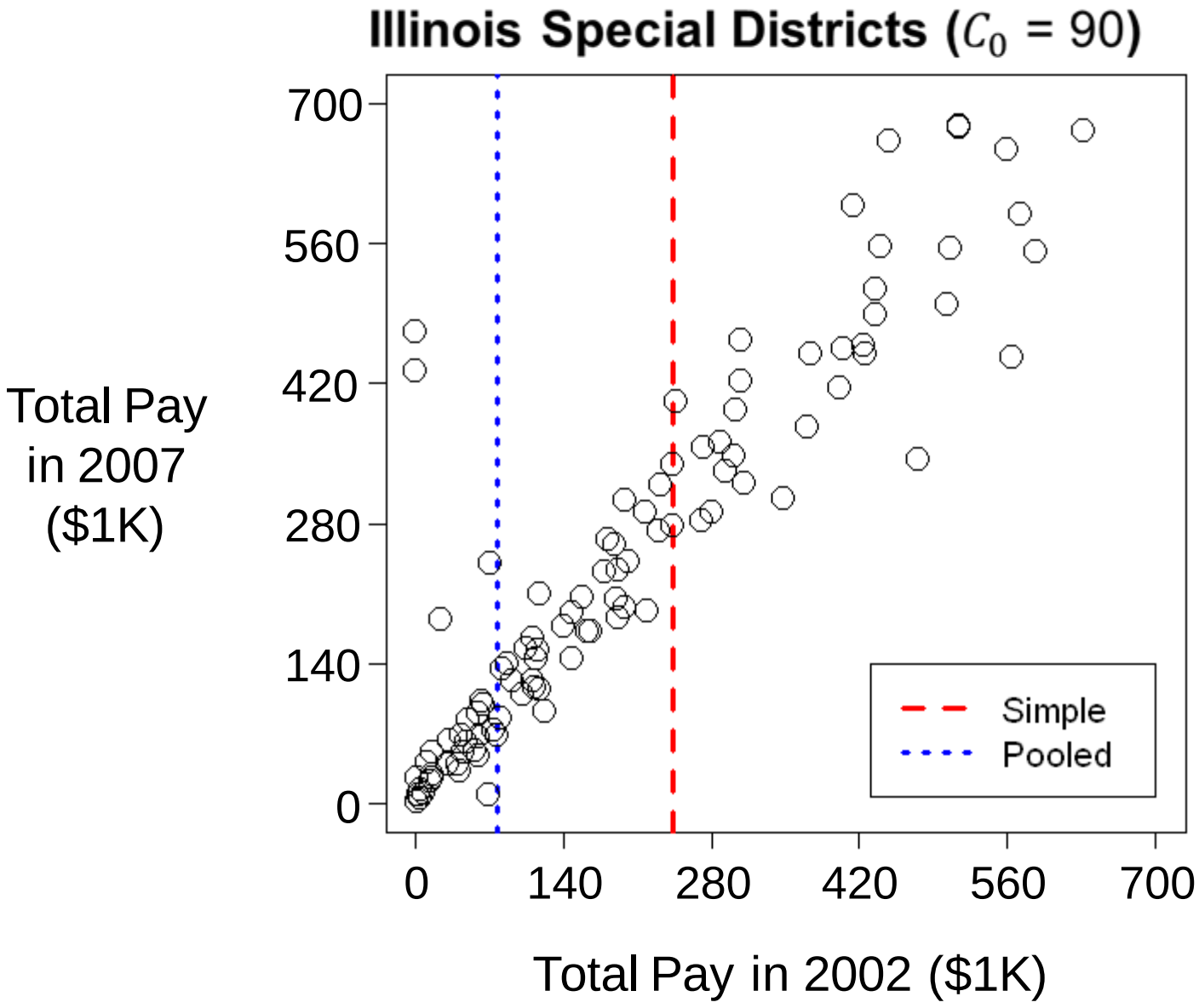
Source: 2002 and 2007 Censuses of Governments: Employment



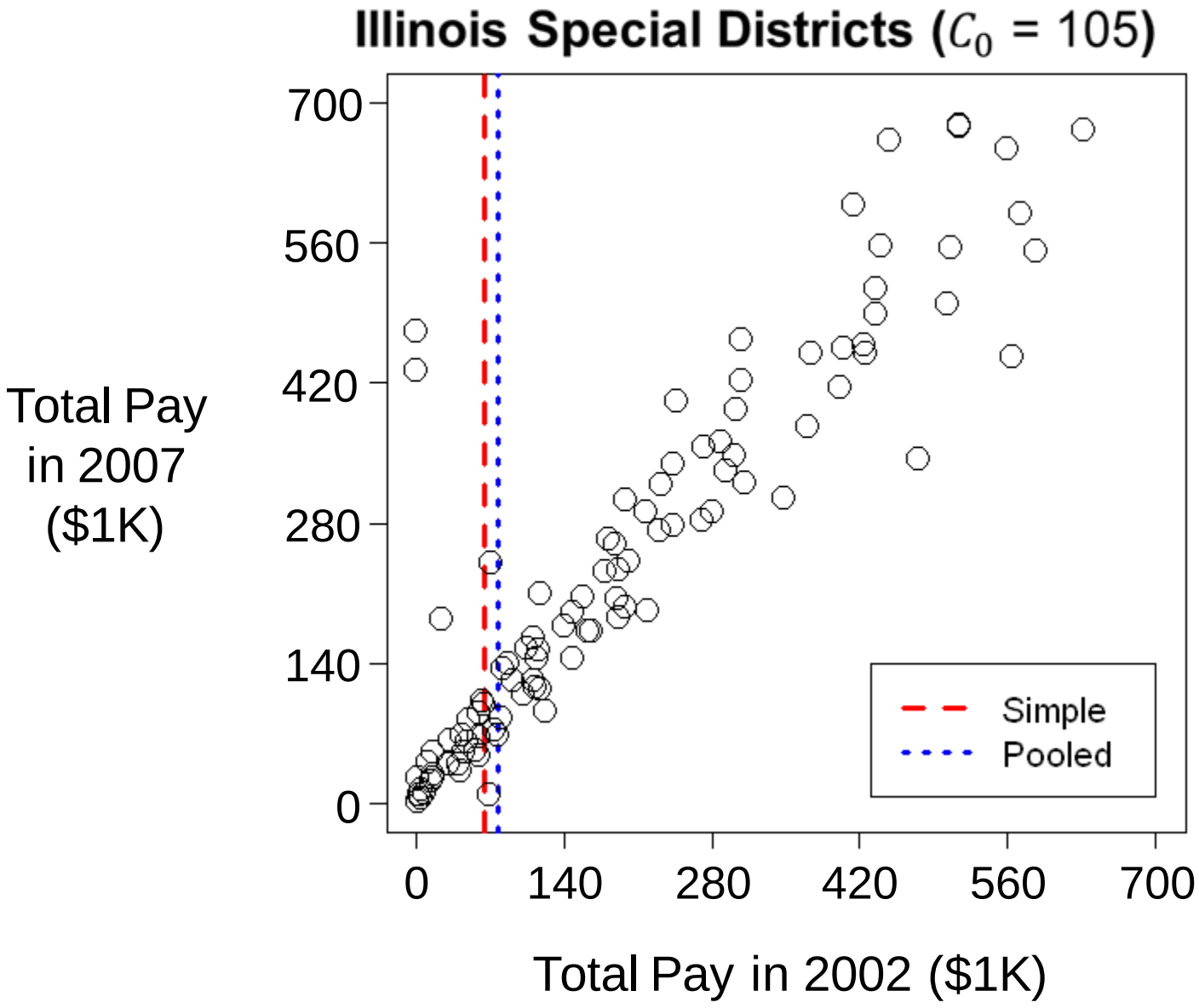
Source: 2002 and 2007 Censuses of Governments: Employment



Source: 2002 and 2007 Censuses of Governments: Employment

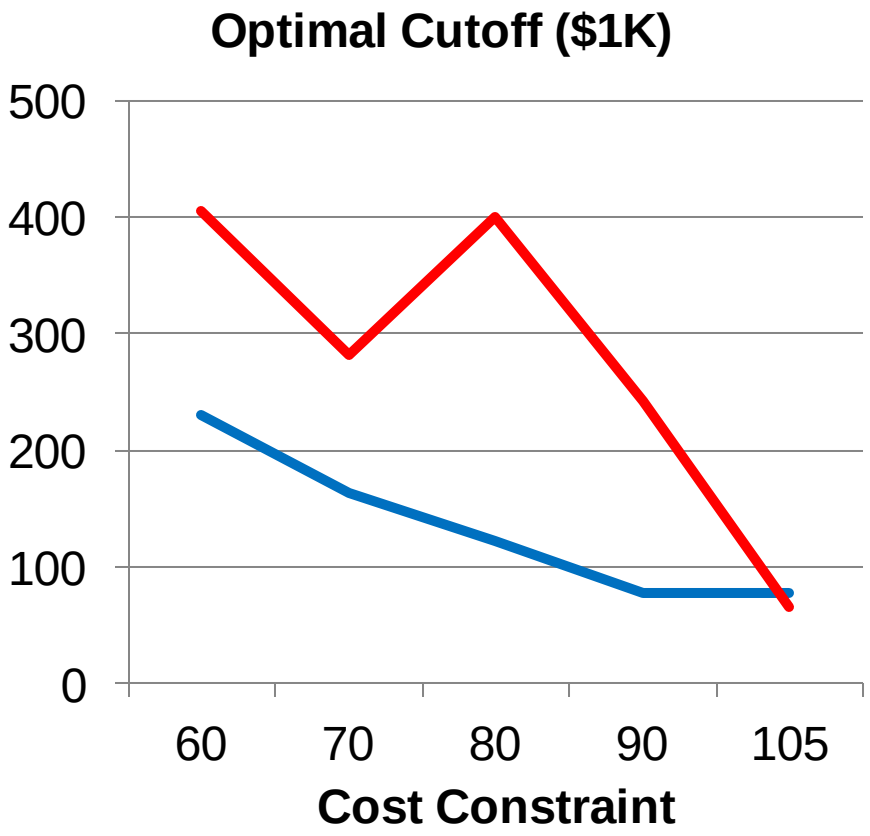


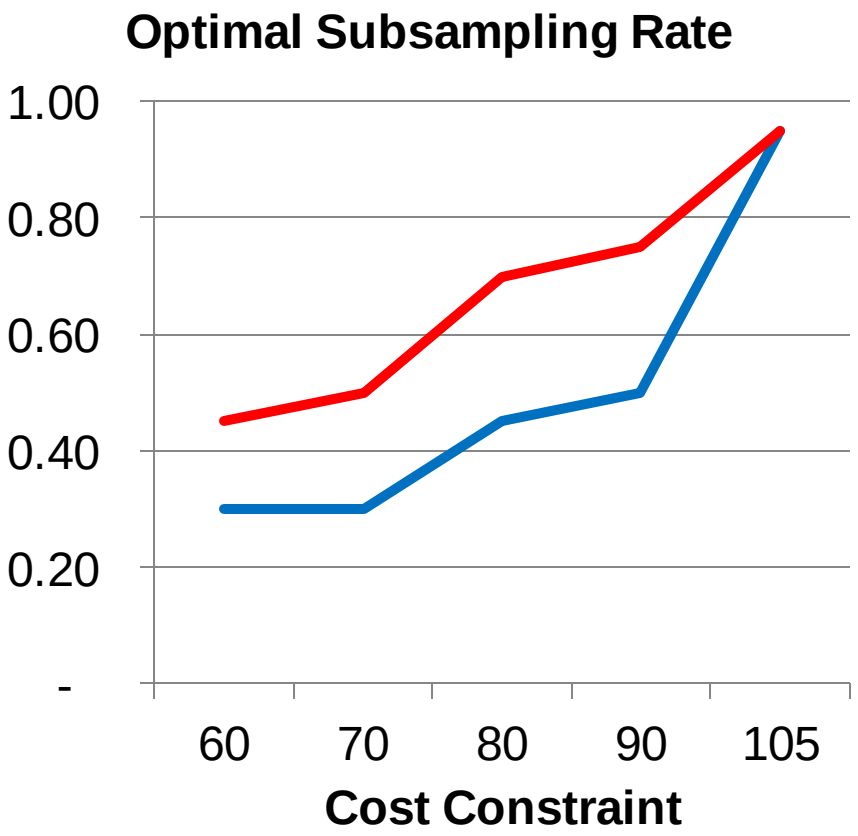
Source: 2002 and 2007 Censuses of Governments: Employment



Source: 2002 and 2007 Censuses of Governments: Employment

Results: Illinois Special Districts

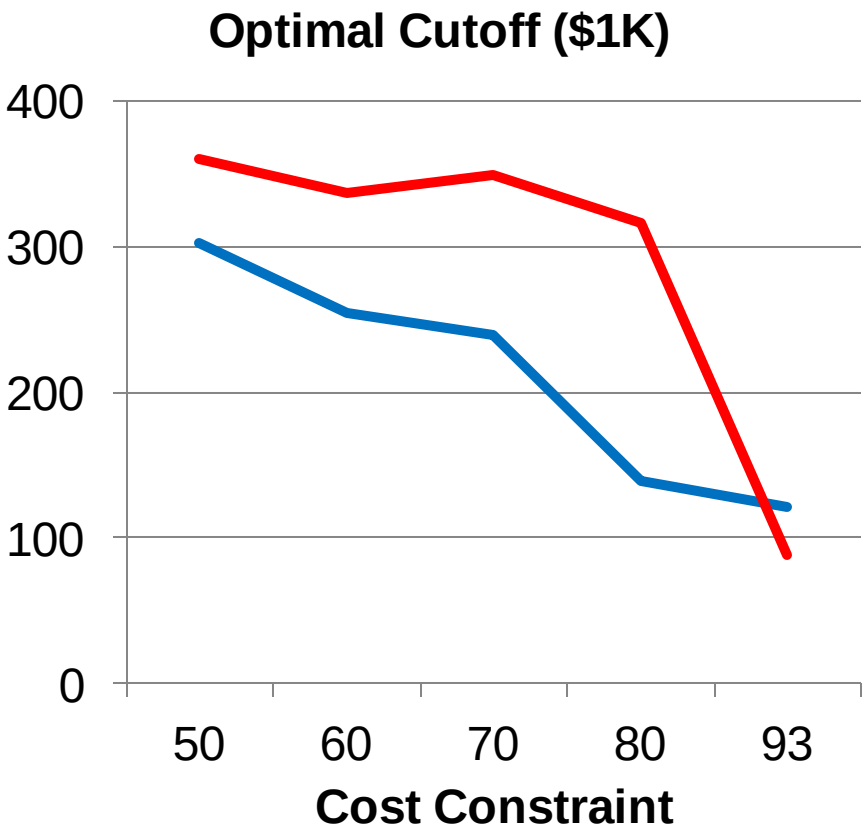


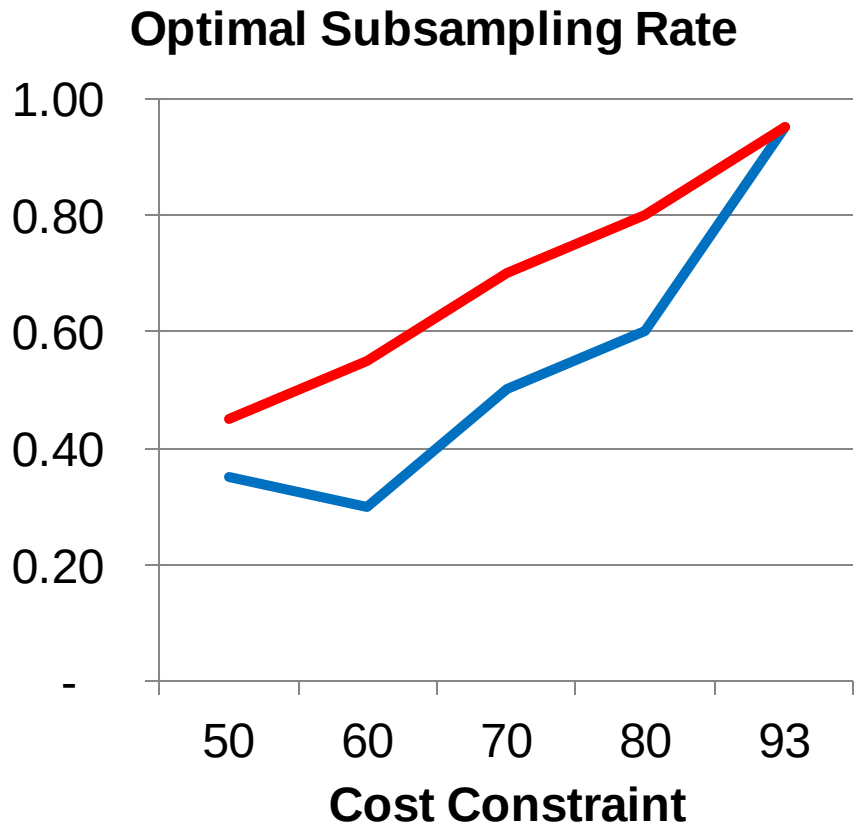


Simple Pooled

Source: 2002 and 2007 Censuses of Governments: Employment

Results: Wisconsin Subcounties

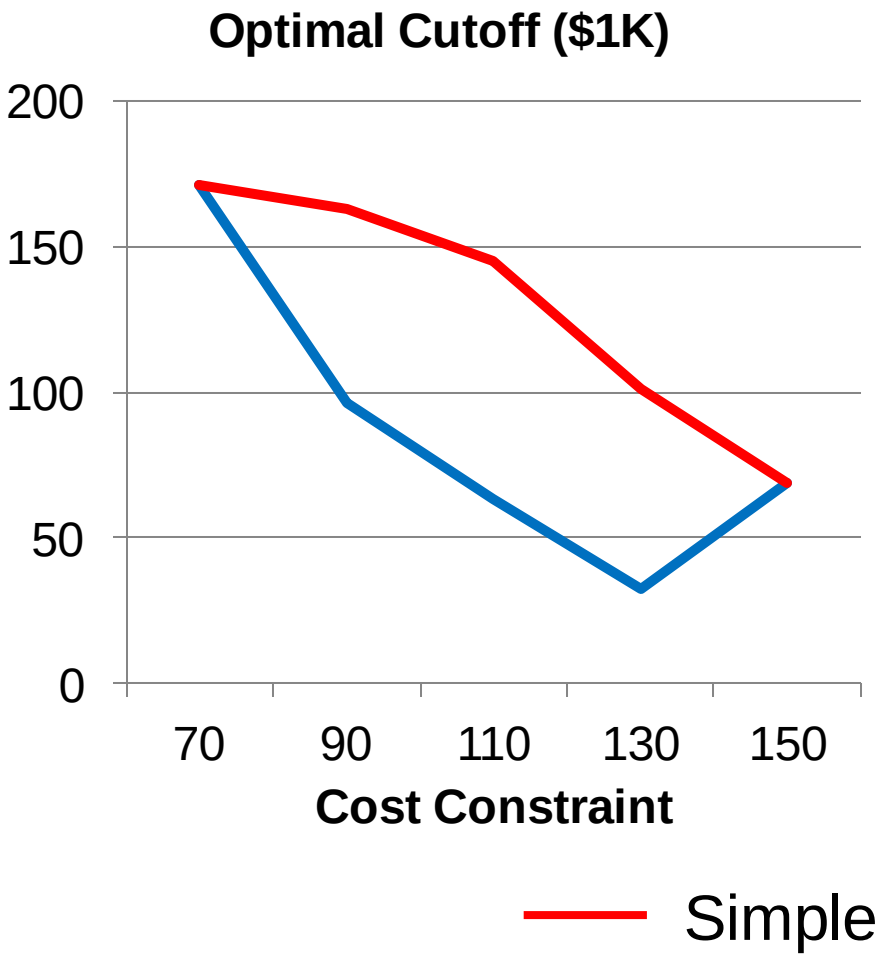


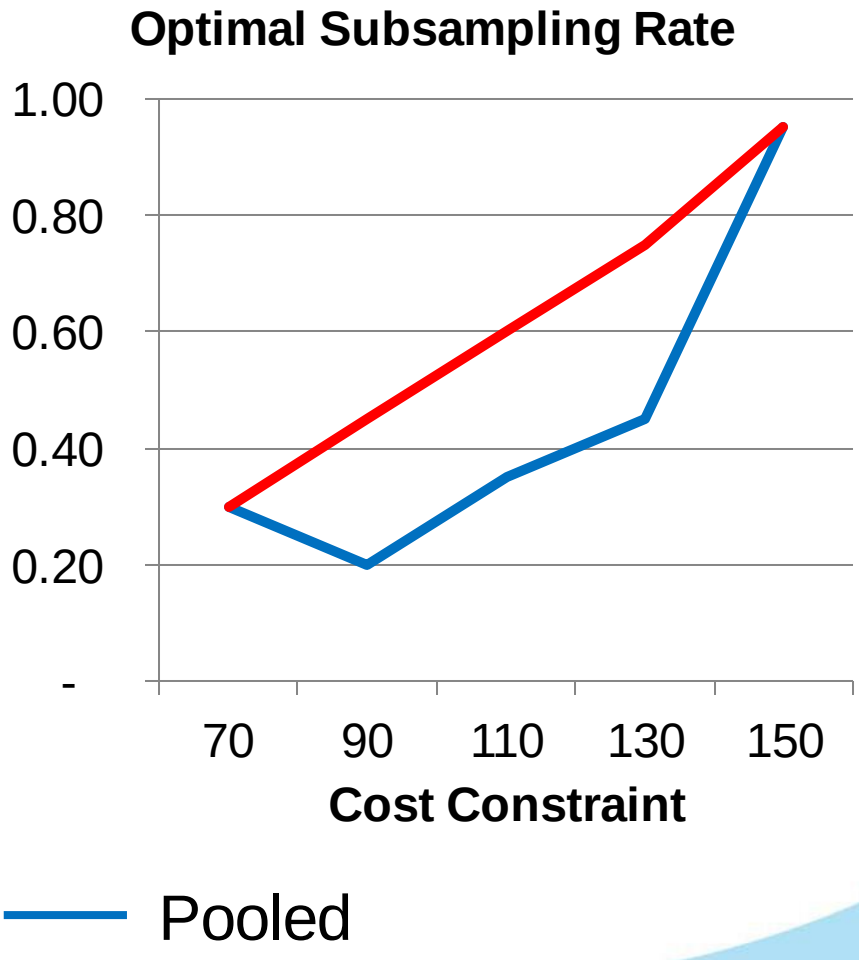


Simple Pooled

Source: 2002 and 2007 Censuses of Governments: Employment

Results: Pennsylvania Subcounties





Simple Pooled

Source: 2002 and 2007 Censuses of Governments: Employment

Results: Census Data (cont.)

- MSE^{Pooled} gives less extreme cutoffs and more stable cutoffs as C_0 varies
- MSE^{Pooled} and MSE^{Simple} give very similar optimal values when there is no cost constraint
- Using sampling weights could be accounting for heteroskedasticity

Future Work

- Perform an empirical evaluation to see which measure of MSE yields more accurate decision-based state totals

- Compare with
  - Current CSRF cutoffs
  - Geometric cutoffs
  - Lavallée and Hidiroglou cutoffs

- Construct a more adequate measure of cost

Contact Information

- Brian.Dumbacher@census.gov
- Carma.Ray.Hogue@census.gov