Government Statistics Research Problems and Challenge

Yang Cheng
Carma Hogue

Governments Division
U.S. Census Bureau

USCENSUSBUREAU
Helping You Make Informed Decisions

Governments Division
Statistical Research & Methodology

Program Research Branch

- Sample design
- Estimation
- Small area estimation

Sampling Frame Research and Development Branch

- Governments Master Address File
- Government Units Survey
- Coverage evaluations

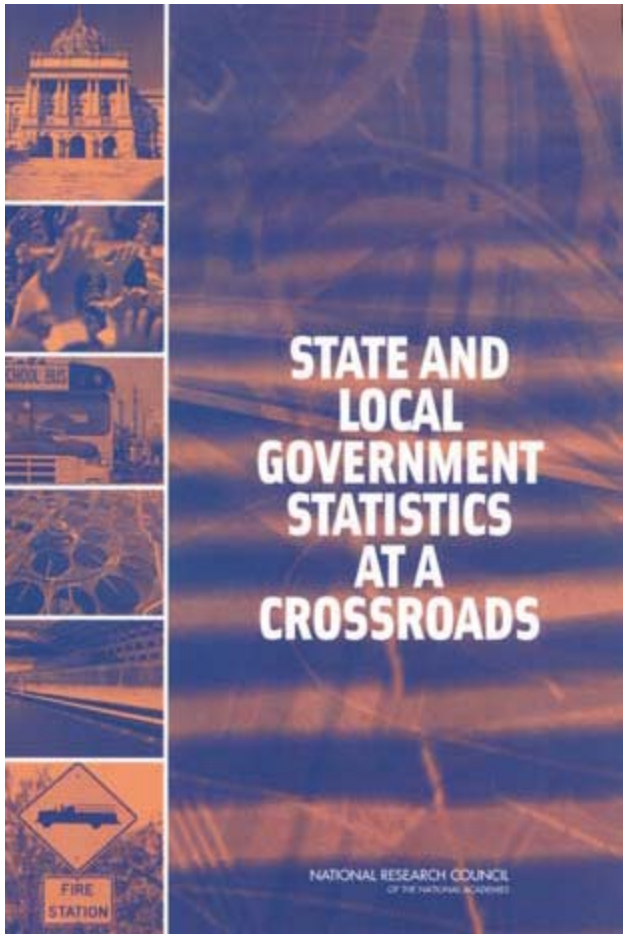
Statistical Methods Branch

- Nonresponse bias studies
- Evaluations
- Selective editing
- Imputation

Committee on National Statistics Recommendations on Government Statistics

- Issued 21 recommendations in 2007

- Contained 13 recommendations that dealt with issues affecting sample design and processing of survey data



The 3-Pronged Approach

- Data User Exchanges
- Research Program
- Modernization and Re-engineering

Dashboards

- Monitor nonresponse follow-up
  - Measures check-in rates
  - Measures Total Quantity Response Rates
  - Measures number of responses and response rate per imputation cell
- Monitor editing
- Monitor macro review

Governments Master Address File (GMAF) and Government Units Survey (GUS)

- GMAF is the database housing the information for all of our sampling frames
- GUS is a directory survey of all governments in the United States

Nonresponse Bias Studies

- Imputation methodology assumes the data are missing at random.
- We check this assumption by studying the nonresponse missingness patterns.
- We have done a few nonresponse bias studies:
  - 2006 and 2008 Employment
  - 2007 Finance
  - 2009 Academic Libraries Survey

Quality Improvement Program

- Team approach
- Trips to targeted areas that are known to have quality issues:
  - Coverage improvement
  - Records-keeping practices
  - Cognitive interviewing
  - Nonresponse follow-up
- Team discussion at end of the day

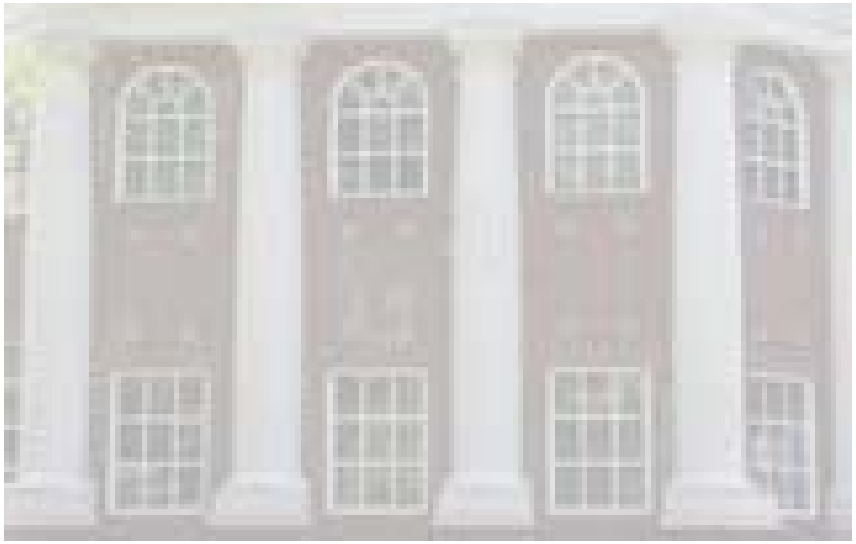
Outline

- Background
- Modified cut-off sampling
- Decision-based estimation
- Small-area estimation
- Variance estimator for the decision-based approach

Background

Types of Local Governments

- Counties
- Municipalities
- Townships
- Special Districts
- Schools



Survey Background

Annual Survey of Public Employment and Payroll

- Variables of interest: Full-time Employment, Full-time Payroll, Part-time Employment, Part-time Payroll, and Part-time Hours

Stratified PPS Sample

- 50 States and Washington, DC

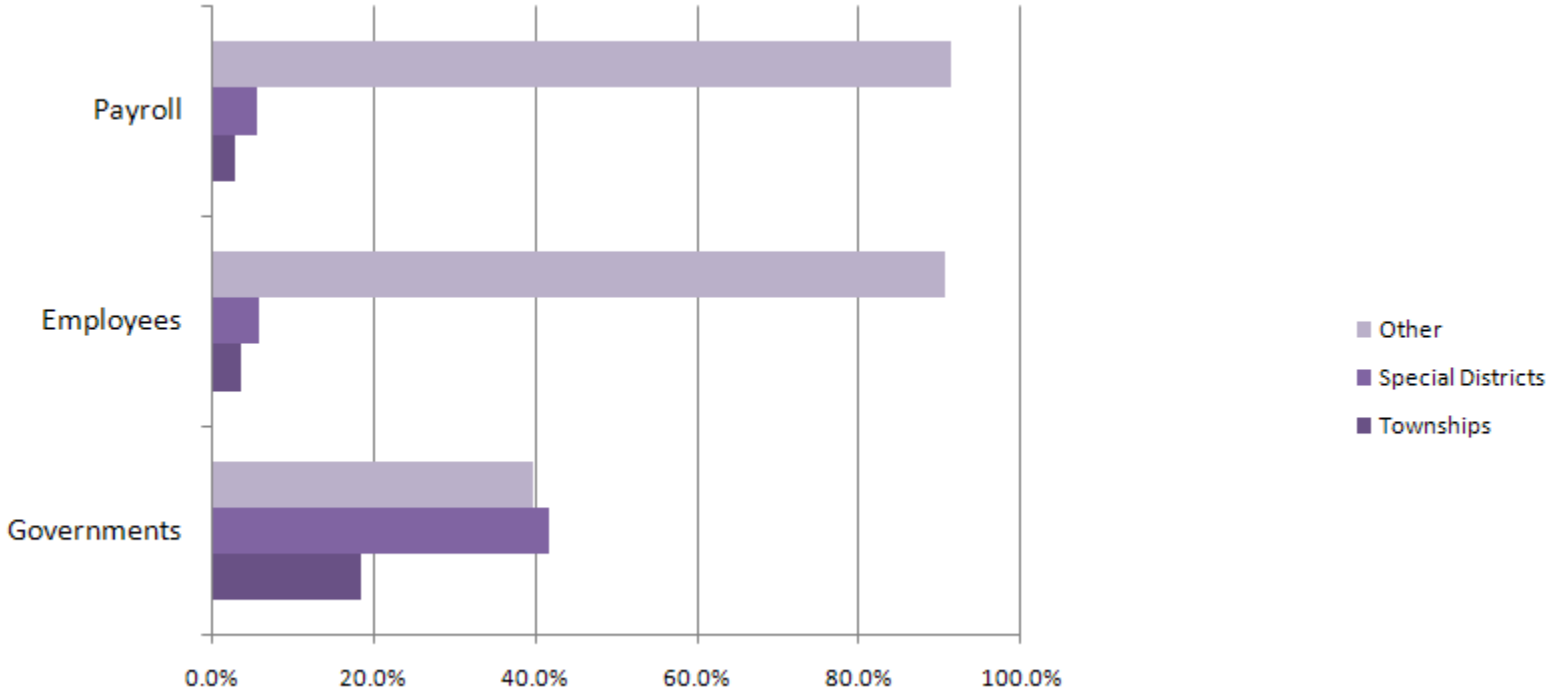
- 4-6 groups: Counties, Sub-Counties (small, large cities and townships), Special Districts (small, large), and School Districts

Distribution of Frequencies for the 2007 Census of Governments: Employment

Government Type	N	Total Employees	Total Payroll	2008 n	2009 n
State	50	5,200,347	$17,788,744,790	50	50
County	3,033	2,928,244	$10,093,125,772	1,436	1,456
Cities	19,492	3,001,417	$11,319,797,633	2,609	3,022
Townships	16,519	509,578	$1,398,148,831	1,534	624
Special Districts	37,381	821,369	$2,651,730,327	3,772	3,204
School Districts	13,051	6,925,014	$20,904,942,336	2,054	2,108
Total	**89,526**	**19,385,969**	**$64,156,489,693**	**11,455**	**10,464**

Source: U.S. Census Bureau, 2007 Census of Governments: Employment

Characteristics of Special Districts and Townships



Source: 2007 Census of Governments

What is Cut-off Sampling?

- Deliberate exclusion of part of the target population from sample selection (Sarndal, 2003)
- Technique is used for highly skewed establishment surveys
- Technique is often used by federal statistical agencies when contribution of the excluded units to the total is small or if the inclusion of these units in the sample involves high costs

Why do we use Cut-off Sampling?

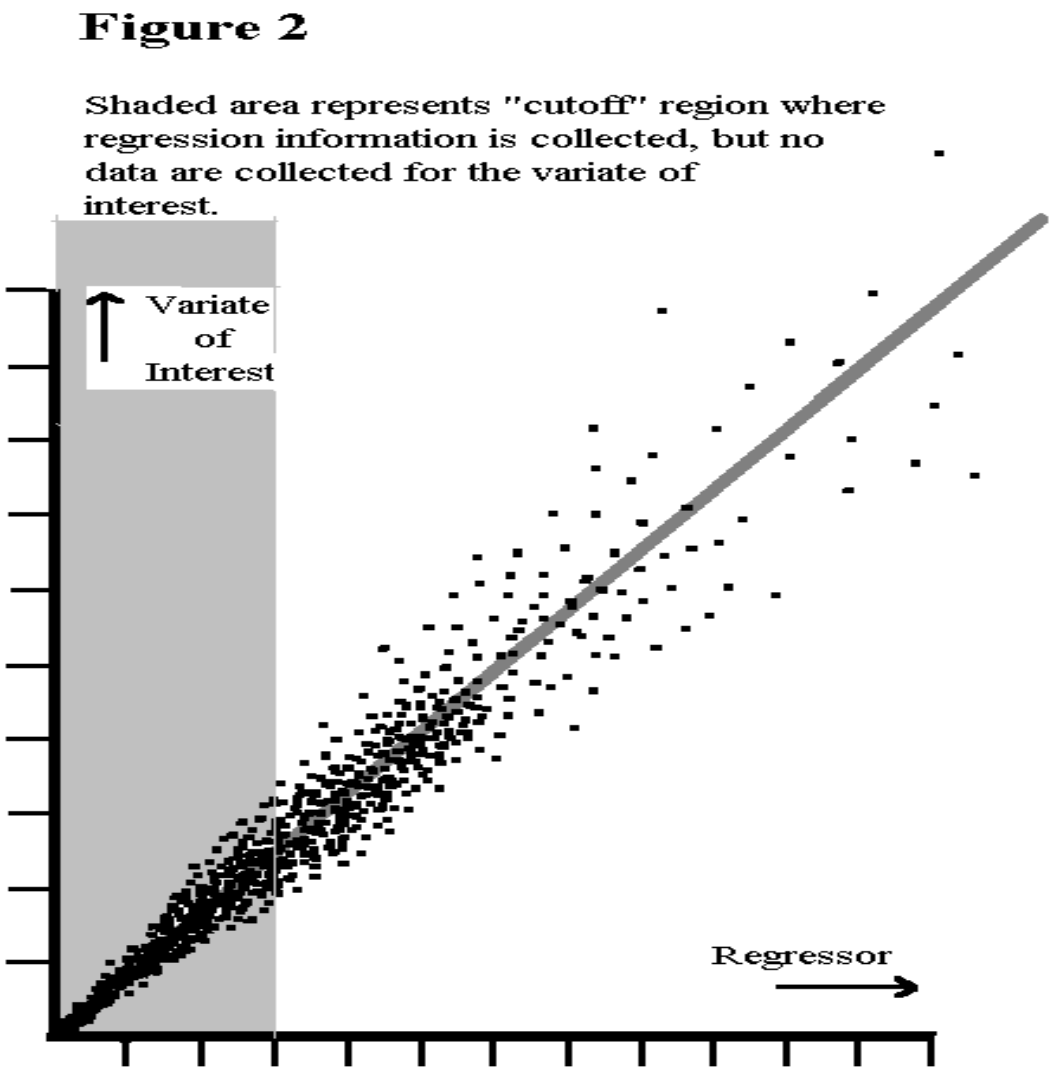
- Save resources
- Reduce respondent burden
- Improve data quality
- Increase efficiency

When do we use Cut-off Sampling?

- Data are collected frequently with limited resources
- Resources prevent the sampler from taking a large sample
- Good regressor data are available

Estimation for Cut-off Sampling

- Model-based approach – modeling the excluded elements (Knaub, 2007)



How do we Select the Cut-off Point?

- 90 percent coverage of attributes
- Cumulative Square Root of Frequency (CSRF) method (Dalenius and Hodges, 1957)
- Modified Geometric method (Gunning and Horgan, 2004)
- Turning points determined by means of a genetic algorithm (Barth and Cheng, 2010)

Modified Cut-off Sampling

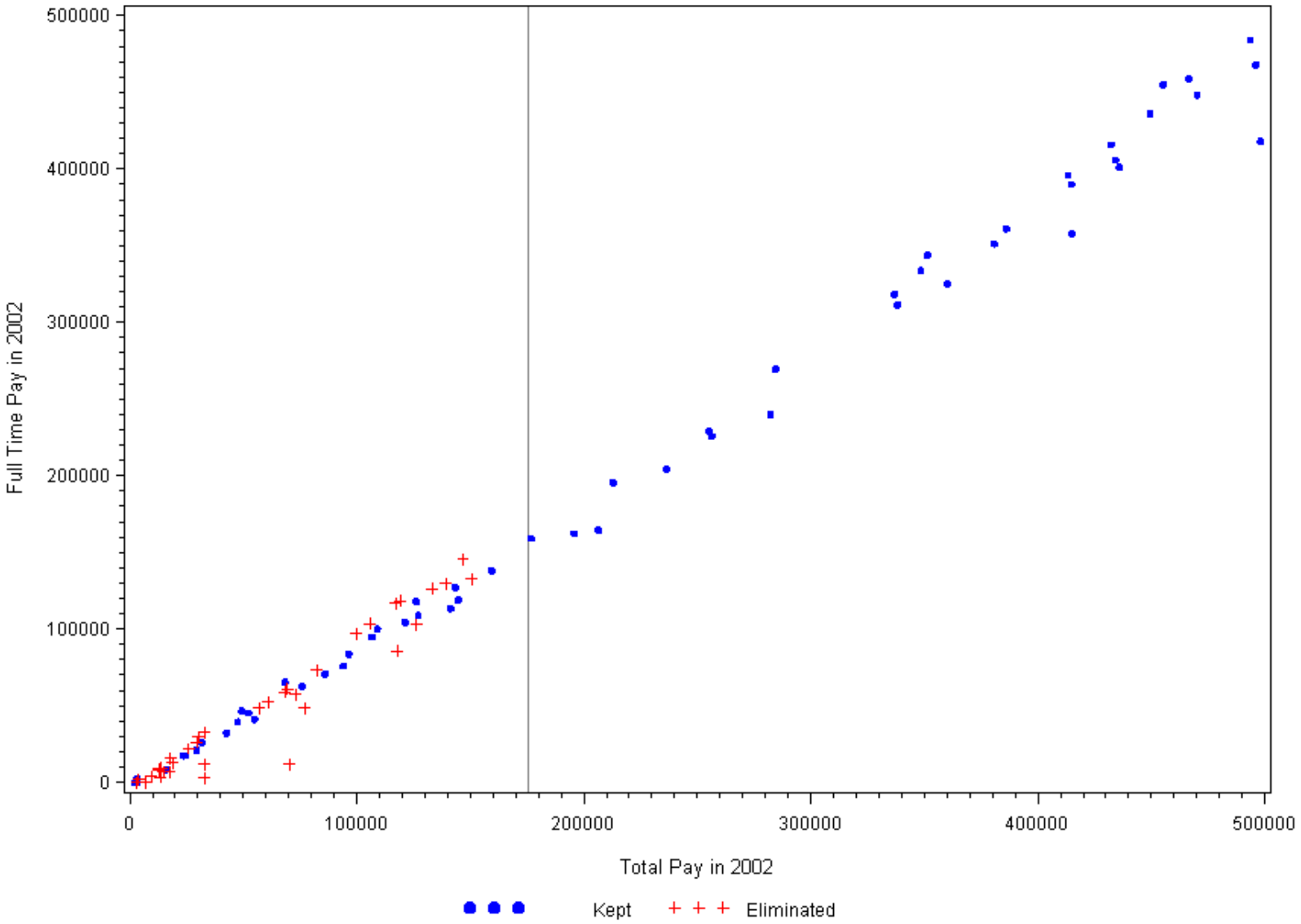
Major Concern:
Model may not fit well for the unobserved data

Proposal:

- Second sample taken from among those excluded by the cutoff
- Alternative sample method based on current stratified probability proportional to size sample design

Wisconsin Cities and Townships

After CSRF and Subsampling



Key Variables for Employment Survey

- The size variable used in PPS sampling is Z=TOTAL PAY from the 2007 Census
- The survey response attributes Y:
  - Full-time Employment
  - Full-time Pay
  - Part-Time Employment
  - Part-Time Pay
- The regression predictor X is the same variable as Y from the 2007 Census

Modified Cut-off Sample Design

Two-stage approach:

- First stage: Select a stratified PPS based on Total Pay

- Second stage: Construct the cut-off point to distinguish small and large size units for special districts and for cities and townships (sub-counties) with some conditions

Notation

- S = Overall sample
- S_1= Small stratum sample
- n_1 = Sample size of S_1
- S_2 = Large stratum sample
- n_2 = Sample size of S_2
- c = Cut-off point between S_1 and S_2
- p = Percent of reduction in S_1
- S_1* = Sub-sample of S_1
- n_1* = pn_1

Modified Cutoff Sample Method

Lemma 1:

Let S be a probability proportional to size (PPS) sample with sample size n drawn from universe U with known size N. Suppose $S_m \subset S$ is selected by simple random sampling, choosing m out of n. Then, S_m is a PPS sample.

How do we Select the Parameters of Modified Cut-off Sampling?

- Cumulative Square Root Frequency for reducing samples (Barth, Cheng, and Hogue, 2009)

- Optimum on the mean square error with a penalty cost function (Corcoran and Cheng, 2010)

Model Assisted Approach

- Modified cut-off sample is stratified PPS sample
  - 50 States and Washington, DC
  - 4-6 modified governmental types: Counties, Sub-Counties (small, large), Special Districts (small, large), and School Districts

- A simple linear regression model:

$$y_{ghi} = a_{gh} + b_{gh} x_{ghi} + \varepsilon_{ghi}$$

Where $g = 1,\ldots,G; h = 1,\ldots,H; i = 1,\ldots,N_{gh}$

Model Assisted Approach (continued)

- For fixed g and h, the least square estimate of the linear regression coefficient is:

$$\hat{b}_{gh} = \frac{S_{gh,xy}}{S^2_{gh,x}}$$

where $S_{gh,xy} = \sum_{i \in U} (x_i - \bar{X})(y_i - \bar{Y}) / (N_{gh} - 1)$ and $S^2_{gh,x} = \sum_{i \in U} (x_i - \bar{X})^2 / (N_{gh} - 1)$

- Assisted by the sample design, we replaced $\hat{b}_{gh}$ by

$$\hat{b} = \frac{\sum_{i \in S} (x_i - \bar{x})(y_i - \bar{y}) / \pi_i}{\sum_{i \in S} (x_i - \bar{x})^2 / \pi_i}$$

Model Assisted Approach (continued)

- Model assisted estimator or weighted regression (GREG) estimator is

$$\hat{Y}_{REG} = \hat{Y}_{\pi} + \hat{b}\left(X - \hat{X}_{\pi}\right)$$

where $X = \sum_{i \in U} x_i$, $\hat{X}_{\pi} = \sum_{i \in S} \frac{x_i}{\pi_i}$, and $\hat{Y}_{\pi} = \sum_{i \in S} \frac{y_i}{\pi_i}$

Decision-based Approach

Idea: Test the equality of the model parameters to determine whether we combine data in different strata in order to improve the precision of estimates.

Analyze data using resulting stratified design with a linear regression estimator (using the previous Census value as a predictor) within each stratum (Cheng, Corcoran, Barth, and Hogue, 2009)

Decision-based Approach

Lemma 2:

When we fit 2 linear models for 2 separate data sets, if $a_1 = a_2$ and $b_1 = b_2$, then the variance of the coefficient estimates is smaller for the combined model fit than for two separate stratum models when the combined model is correct.

Test the equality of regression lines

- Slopes
- Elevation (y-intercepts)

Test of Equal Slopes (Zar, 1999)

$$H_0 : b_1 = b_2$$

$$H_A : b_1 \neq b_2$$

$$t_{gh} = \frac{\hat{b}_{gh,1} - \hat{b}_{gh,2}}{s_{b_{gh,1} - b_{gh,2}}} \sim t_{n_{gh,1} + n_{gh,2} - 4}$$

where

$$s_{b_{gh,1} - b_{gh,2}} = \sqrt{\frac{\left(s^2_{gh,xy}\right)_p}{\left(x^2_{gh}\right)_1} + \frac{\left(s^2_{gh,xy}\right)_p}{\left(x^2_{gh}\right)_2}}$$

and

$$\left(s^2_{gh,xy}\right)_p = \frac{\sum_{i \in S_{gh,1}} \left(y_{gh,i} - \hat{y}_{gh,i}\right)^2 + \sum_{i \in S_{gh,2}} \left(y_{gh,i} - \hat{y}_{gh,i}\right)^2}{n_1 + n_2 - 4}$$

Test of Equal Elevation

$$t_{gh} = \frac{\left(\bar{y}_{gh,1} - \bar{y}_{gh,2}\right) - \hat{b}_{gh,c}\left(\bar{x}_{gh,1} - \bar{x}_{gh,2}\right)}{\sqrt{\left(s^2_{gh,xy}\right)_c \left[1/n_{gh,1} + 1/n_{gh,2} + \left(\bar{x}_{gh,1} - \bar{x}_{gh,2}\right)^2 \Big/ \left(\sum_{i \in S_{gh}} x^2_{gh,i}\right)\right]}}$$

$$\sim t_{n_{gh,1}+n_{gh,2}-4}$$

where $$s^2_{gh,xy} = \frac{\sum_{i \in S_{gh}} y_{gh,i}{}^2 - \left(\sum_{i \in S_{gh}} x_{gh,i}\, y_{gh,i}\right) \Big/ \left(\sum_{i \in S_{gh}} x_{gh,i}{}^2\right)}{n_{gh} - 3}$$

More than Two Regression Lines

$$H_0 : b_1 = b_2 = \ldots = b_k$$

$$F = \frac{\left(\frac{SS_c - SS_p}{k-1} \right)}{\frac{SS_p}{\sum_{i=1}^{k} n_i - 2k}} \sim F_{k-1, \sum_{i=1}^{k} n_i - 2k}$$

- If rejected, k-1 multiple comparisons are possible.

Test of Null Hypothesis

Data analysis: Null hypothesis of equality of intercepts cannot be rejected if null hypothesis of equality of slopes cannot be rejected.

The model-assisted slope estimator, $\hat{b}$, can be expressed within each stratum using the PPS design weights as

$$\hat{b} = \sum_{i \in S} \frac{1}{\pi_i} y_i \left(x_i - \hat{X}_\pi / \hat{N} \right) \Bigg/ \sum_{i \in S} \frac{1}{\pi_i} \left(x_i - \hat{X}_\pi / \hat{N} \right)^2$$

where $\hat{N} = \sum_{i \in S} \frac{1}{\pi_i}$

Test of Null Hypothesis (continued)

- In large samples, $\hat{b}$ is approximately normally distributed with mean b and a theoretical variance denoted $\sum$.

- The test statistic becomes

$$\left(\hat{b}_1 - \hat{b}_2\right)\sum_{1,2}^{-1}\left(\hat{b}_1 - \hat{b}_2\right) \sim \chi_1^2 \quad \text{where } \Sigma_{1,2} = \Sigma_1 + \Sigma_2$$

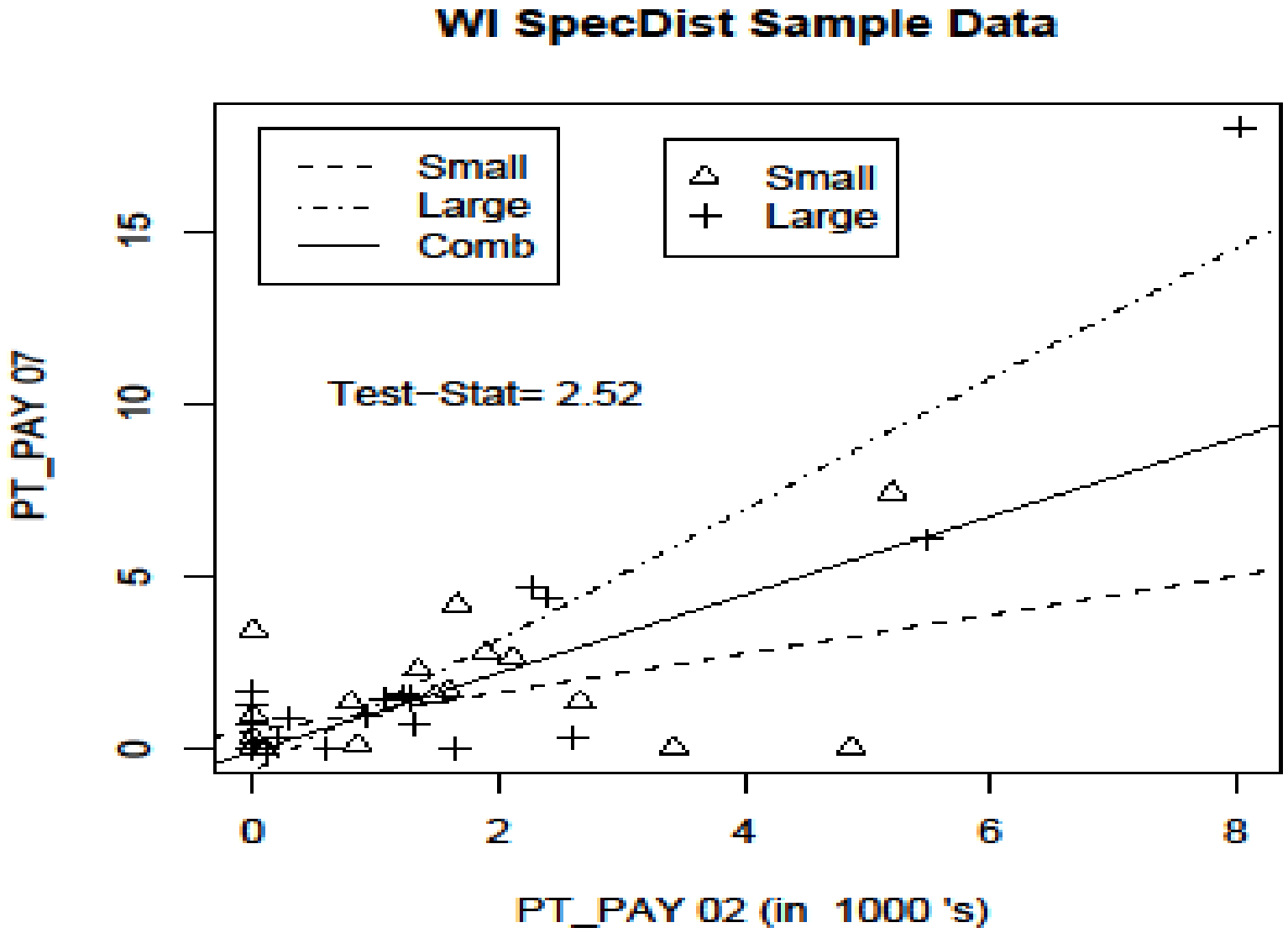
- If the P value is less than 0.05, we reject the null hypothesis and conclude that the regression slopes are significantly different.

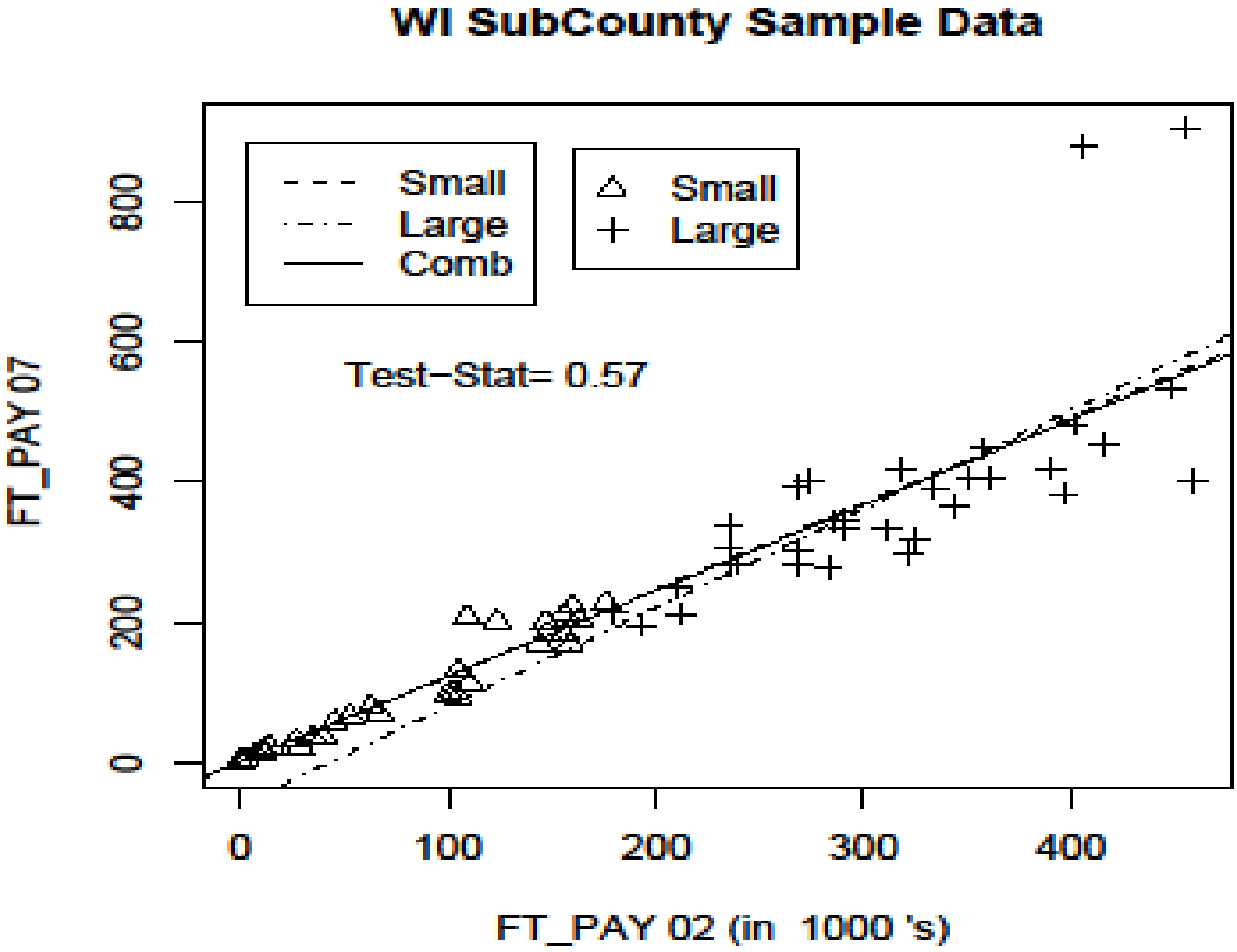
Decision-based Estimation

- Null hypothesis: $H_0: \beta_S = \beta_L$

- The decision-based estimator:

$$\hat{t}_{y,dec} = \begin{cases} \hat{t}_{y,S} + \hat{t}_{y,L} & \text{If reject } H_0 \\ \hat{t}_{y,S\&L} & \text{If cannot reject } H_0 \end{cases}$$


WI SpecDist Sample Data
Small
Large
Comb
Small
Large
Test-Stat= 2.52
PT_PAY 07
0
5
10
15
0
2
4
6
8
PT_PAY 02 (in 1000 's)


WI SubCounty Sample Data
Small
Large
Comb
Small
Large
Test−Stat= 0.57
FT_PAY 07
0
200
400
600
800
0
100
200
300
400
FT_PAY 02 (in 1000 's)

Test results for decision-based method

	FT_Pay		FT_Emp		PT_Pay	
(State,Type)	Test-Stat	Decision	Test-Stat	Decision	Test-Stat	Decision
(AL, SubCounty)	2.06	Reject	2.04	Reject	3.62	Reject
(CA, SpecDist)	0.98	Accept	1.02	Accept	0.29	Accept
(PA, SubCounty)	0.54	Accept	0.62	Accept	0.08	Accept
(PA, SpecDist)	0.24	Accept	0.65	Accept	1.09	Accept
(WI, SubCounty)	0.57	Accept	0.85	Accept	2.11	Reject
(WI, SpecDist)	1.33	Accept	0.85	Accept	2.52	Reject

Small Area Challenge

Our sample design is at the government unit level

- Estimating the total employees and payroll in the annual survey of public employment and payroll
- Estimating the employment information at the functional level.
  - There are 25-30 functions for each government unit
  - Domain for functional level is subset of universe U
  - Sample size for function f, $n_f \leq n$ and $S_f = S \cap U_f$
- Estimate the total of employees and payroll at state by function level:

$$Y_{gf} = \sum_{i \in U_{gf}} Y_{gf,i}$$

Functional Codes

001, Airports
002, Space Research & Technology (Federal)
005, Correction
006, National Defense and International Relations (Federal)
012, Elementary and Secondary - Instruction
112, Elementary and Secondary - Other Total
014, Postal Service (Federal)
016, Higher Education - Other
018, Higher Education - Instructional
021, Other Education (State)
022, Social Insurance Administration (State)
023, Financial Administration
024, Firefighters
124, Fire - Other
025, Judical & Legal
029, Other Government Administration
032, Health

040, Hospitals
044, Streets & Highways
050, Housing & Community Development (Local)
052, Local Libraries
059, Natural Resources
061, Parks & Recreation
062, Police Protection - Officers
162, Police-Other
079, Welfare
080, Sewerage
081, Solid Waste Management
087, Water Transport & Terminals
089, Other & Unallocable
090, Liquor Stores (State)
091, Water Supply
092, Electric Power
093, Gas Supply
094, Transit

Direct Domain Estimates

Structural zeros are cells in which observations are impossible

Function/ID	1	2	3	4	5	…	N-1	N
001	✓	N/A	N/A	N/A	N/A	…	✓	N/A
005	✓	✓	N/A	✓	✓	…	✓	✓
012	✓	✓	✓	✓	N/A	…	N/A	✓
023	N/A	✓	✓	✓	✓	…	✓	✓
024	✓	✓	✓	✓	✓	…	✓	✓
…	…	…	…	…	…	…	…	…
124	✓	✓	✓	✓	✓	…	✓	✓
162	✓	N/A	✓	✓	✓	…	✓	✓
Total	✓	✓	✓	✓	✓	…	✓	✓

Direct Domain Estimates (continued)

- Horvitz-Thompson Estimation

$$\hat{Y}_{gf} = \sum_{i \in S_{gf}} w_{g,i} y_{gf,i}$$

- Modified Direct Estimation

$$\hat{Y}_{gf} = \hat{Y}_{gf,\pi} + \hat{b}_f (X_{gf} - \hat{X}_{gf,\pi})$$

Synthetic Estimation

- Synthetic assumption: small areas have the same characteristics as large areas and there is a valid unbiased estimate for large areas
- Advantages:
  - Accurate aggregated estimates
  - Simple and intuitive
  - Applied to all sample design
  - Borrow strength from similar small areas
  - Provide estimates for areas with no sample from the sample survey

Synthetic Estimation (continued)

General idea:

- Suppose we have a reliable estimate for a large area and this large area covers many small areas. We use this estimate to produce an estimator for small area.
- Estimate the proportions of interest among small areas of all states.

Synthetic Estimation (continued)

- Synthetic estimation is an indirect estimate, which borrows strength from sample units outside the domain.
- Create a table with government function level as rows and states as columns. The estimator for function f and state g is:

$$\hat{y}_{gf} = \frac{\sum_{g \in G} x_{gf}}{\sum_{f \in F} \sum_{g \in G} x_{gf}} \hat{y}_{g.}$$

Synthetic Estimation (continued)

Function Code	State					Total
	1	2	3	…	50	
1	$X_{1,1}$	$X_{1,2}$	$X_{1,3}$	…	$X_{1,50}$	$X_{1,.}$
5	$X_{2,1}$	$X_{2,2}$	$X_{2,3}$	…	$X_{2,50}$	$X_{2,.}$
12	$X_{3,1}$	$X_{3,2}$	$X_{3,3}$	…	$X_{3,50}$	$X_{3,.}$
…	…	…	…	…	…	
124	$X_{29,1}$	$X_{29,2}$	$X_{29,3}$	…	$X_{29,50}$	$X_{29,.}$
162	$X_{30,1}$	$X_{30,2}$	$X_{30,3}$	…	$X_{30,50}$	$X_{30,.}$
Total	$Y_{.,1}$	$Y_{.,2}$	$Y_{.,3}$	…	$Y_{.,50}$	$X_{.,.}$

Synthetic Estimation (continued)

Bias of synthetic estimators:

- Departure from the assumption can lead to large bias.
- Empirical studies have mixed results on the accuracy of synthetic estimators.
- The bias cannot be estimated from data.

Composite Estimation

- To balance the potential bias of the synthetic estimator against the instability of the design-based direct estimate, we take a weighted average of two estimators.
- The composite estimator is:

$$\hat{y}_{gf}^{C} = w_{gf}\,\hat{y}_{gf}^{D} + \left(1 - w_{gf}\right)\hat{y}_{gf}^{S}$$

Composite Estimation (continued)

Three methods of choosing w_{gf}

- Sample size dependent estimate:

$$w_{gf} = \begin{cases} 1 & \text{if } \hat{N}_{gf} \geq \delta N_{gf} \\ \hat{N}_{gf} \Big/ \left(\delta N_{gf}\right) & \text{otherwise} \end{cases}$$

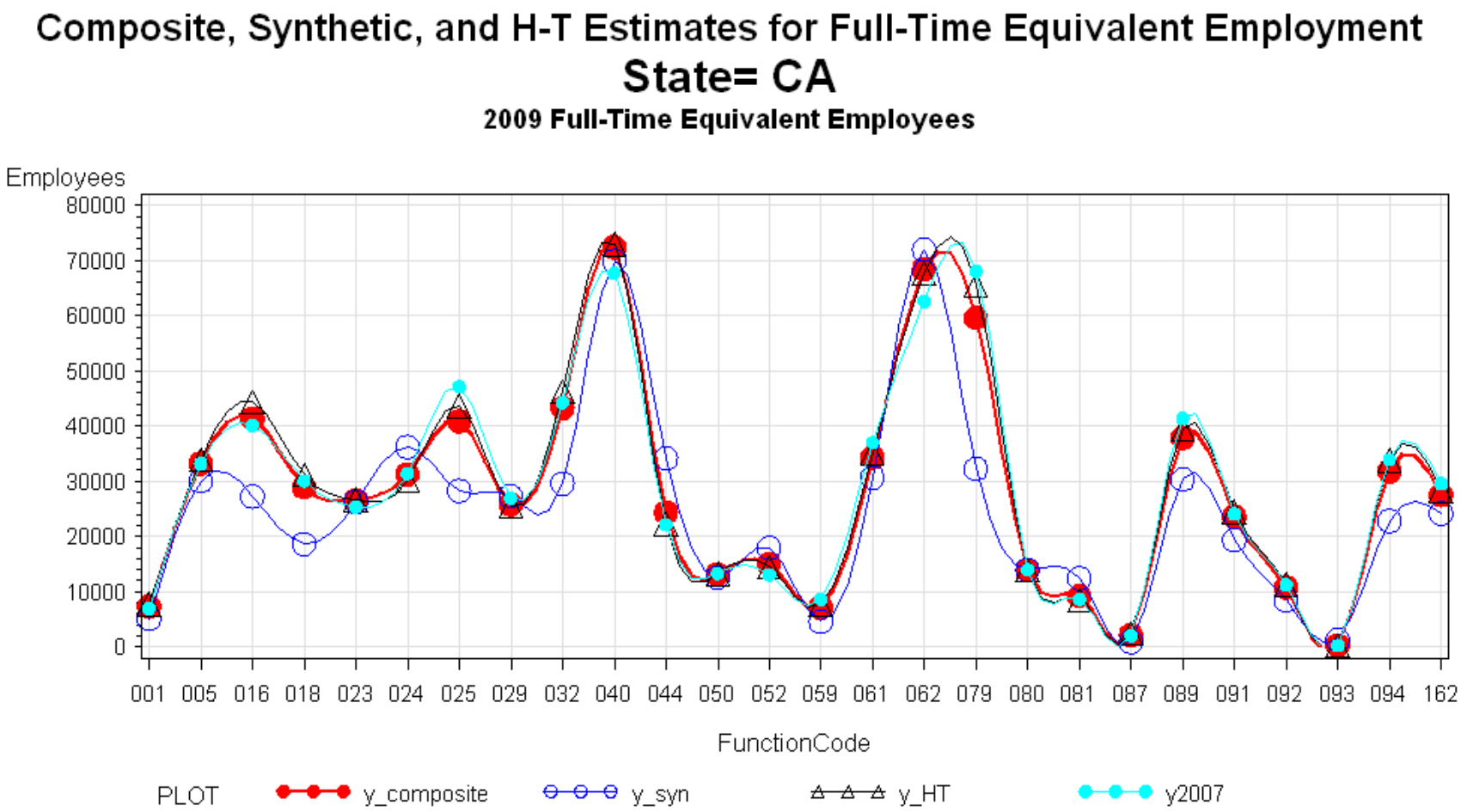
  where delta is subjectively chosen. In practice, we choose delta from 2/3 to 3/2.

- Optimal w_{gf} :

$$w_{gf}^{opt} = \frac{MSE\left(\hat{y}_{gf}^{S}\right)}{MSE\left(\hat{y}_{gf}^{S}\right) + Var\left(\hat{y}_{gf}^{D}\right)}$$

- James-Stein common weight

Composite Estimation (Cont'd)
Example



Variance Estimator

- To estimate the variance for unequal weights, first apply the Yates-Grundy estimator:

$$\hat{V}_1(\hat{y}) = -\frac{1}{2}\sum_{i,k\in S}\frac{\pi_i\pi_k - \pi_{ik}}{\pi_{ik}}\left(\frac{y_i}{\pi_i} - \frac{y_k}{\pi_k}\right)^2$$

- To compensate the variance and avoid the 2nd order joint inclusion probability, we apply the PPSWR variance estimator formula:

$$\hat{V}_2(\hat{y}) = \frac{n}{(n-1)}\sum_{i\in S}(z_i - \bar{z})^2$$

where:

$$z_i = \frac{y_i}{\pi_i} \quad \text{and} \quad \bar{z} = \frac{1}{n}\sum_{i\in S} z_i$$

Variance Estimator for Weighted Regression Estimator

- The weighted regression estimator: $\hat{t}_{y,pps}$
- The naive variance obtained by combining variances for stratum-wise regression estimators and using PPSWR variance formula within each stratum:

$$V(\hat{t}_{y,pps}) = \sum_{i=1}^{N} \frac{e_i^2}{p_i}$$

  where p_i is the single-draw probability of selecting a sample unit i

- The variance is estimated by the quantity

$$\hat{V}(\hat{t}_{y,pps}) = \frac{n}{n-1} \sum_{i \in S} \left(\frac{y_i - \hat{y}_i}{\pi_i} \right)^2$$

Data Simulation (Cheng, Slud, Hogue 2010)

- Regression predictor: $x_i \sim Gamma(\alpha, \beta)$
- Sample weights: $w_i = \dfrac{\sum_{i=1}^{N} x_i}{n x_i}$
- Response attribute:

$$y_i = \begin{cases} a x_i^2 + b x_i + c + \varepsilon_{1i} & i \in U_S, \quad \varepsilon_{1i} \sim N(0, \sigma_1^2) \\ a x_i^2 + b x_i + c + d x_i + \varepsilon_{2i} & i \in U_L, \quad \varepsilon_{2i} \sim N(0, \sigma_2^2) \end{cases}$$

Data Simulation Parameters Table

Examples	a	b	c	D	σ1	σ2	n1	n2	N1	N2
1	0	2	0.2	0	3	3	40	60	1,500	1,200
2	0	2	0	0.2	3	3	40	60	1,500	1,200
3	0	2	0	0.4	3	3	40	60	1,500	1,200
4	0	2	0	0.6	3	3	40	60	1,500	1,200
5	0	2	0	0.6	4	4	40	60	1,500	1,200
6	0	2	0	0.8	4	4	40	60	1,500	1,200
7	0	2	-0.1	0.8	4	4	40	60	1,500	1,200
8	0	2	0.2	0	3	3	20	30	1,500	1,200

Bootstrap Approach

1. Population frame: N_1 and N_2
2. Substratum values: (X_{ij}, Y_{ij}), $i = 1,2; j \in U_i$
3. Sample selection: PPSWOR with n_1 , n_2 elements
4. Bootstrap replications: b=1,...,B
5. Bootstrap sample: SRSWR with size n_1 and n_2
6. Estimation: Decision-based method was applied to each bootstrap sample
7. Results: $t_{y,dec}^{b}$ and V_{naiv}^{b}

Monte Carlo Approach

- The simulated frame populations are the same ones used in the bootstrap simulations.
- Monte Carlo replications: r = 1,2...,R
- Following bootstrap steps 3, 5, 6, and 7, we have results: $t_{y,dec}^{r}$ and V_{naiv}^{r}

Null hypothesis reject rates for decision-based methods

- Prej_MC: proportion of rejections in the hypothesis test for equality of slopes in MC method
- Prej_Boot: proportion of rejections in the hypothesis test for equality of slopes in Bootstrap method

Different Variance Estimators

- MC.Naiv: $$V_{MC.naiv} = \frac{1}{R}\sum_{r=1}^{R} V_{naiv}^{r}$$
- MC.Emp $$V_{MC.Emp} = \frac{1}{R-1}\sum_{r=1}^{R}\left(t_{y,dec}^{r} - \bar{t}_{y,dec}^{r}\right)^2$$
- Boot.Naiv: $$V_{Boot.naiv} = \frac{1}{RB}\sum_{r}\sum_{b} V_{naiv}^{rb}$$
- Boot.Emp $$V_{Boot.Emp} = \frac{1}{R}\sum_{r=1}^{R} S_r^2$$

where S_r^2 is the sample variance of $\left\{t_{y,dec}^{rb}, i = 1,2,\ldots,B\right\}$

Data Simulation with R=500 and B=60

Examples	Prej. MC	Prej. Boot	MC. Emp	MC. Naiv	Boot. Emp	Boot. Naiv	DEC. MSE	2str. MSE
1	0.796	0.719	991.8	867.9	863.6	846.9	832,904	819,736
2	0.098	0.231	920.6	873.2	871.4	856.4	846,843	857,654
3	0.126	0.277	908.3	868.6	903.2	847	826,142	845,332
4	0.258	0.333	880.9	874.7	862.8	850.6	777,871	779,790
5	0.144	0.249	1,159.5	1,139	1,192.1	1111.4	1,346,545	1,351,290
6	0.258	0.339	1,173.5	1,144.1	1,179.1	1113.7	1,374,466	1,401,604
7	0.088	0.217	1,167.7	1,148.4	1,165.3	1126.7	1,361,384	1,397,779
8	0.582	0.601	1,288.2	1,209.1	1,229.4	1149.8	1,656,195	1,656,324

Monte Carlo & Bootstrap Results

The tentative conclusions from simulation study:

- Bootstrap estimate of the probability of rejecting the null hypothesis of equal substratum slopes can be quite different from the true probability
- Naïve estimator of standard error of the decision-based estimator is generally slightly less than the actual standard error
- Bootstrap estimator of standard error is not reliably close to the true standard error (the MC.Emp column)
- Mean-squared error for the decision-based estimator is generally only slightly less than that for the two-substratum estimator, but does seem to be a few percent better for a broad range of parameter combinations.

References

Barth, J., Cheng, Y. (2010). Stratification of a Sampling Frame with Auxiliary Data into Piecewise Linear Segments by Means of a Genetic Algorithm, *JSM Proceedings.*

Barth, J., Cheng, Y., Hogue, C. (2009). Reducing the Public Employment Survey Sample Size, *JSM Proceedings.*

Cheng, Y., Corcoran, C., Barth, J., Hogue, C. (2009). An Estimation Procedure for the New Public Employment Survey, *JSM Proceedings.*

Cheng, Y., Slud, E., Hogue, C. (2010). Variance Estimation for Decision-Based Estimators with Application to the Annual Survey of Public Employment and, *JSM Proceedings.*

Clark, K., Kinyon, D. (2007). *Can We Continue to Exclude Small Single-establishment Businesses from Data Collection in the Annual Retail Trade Survey and the Service Annual Survey?* [PowerPoint slides]. Retrieved from http://www.amstat.org/meetings/ices/2007/presentations/Session8/Clark_Kinyon.ppt

References

Corcoran, C., Cheng, Y. (2010). Alternative Sample Approach for the Annual Survey of Public Employment and Payroll, *JSM Proceedings.*

Dalenius, T., Hodges, J. (1957). *The Choice of Stratification Points*. Skandinavisk Aktuarietidskrift.

Gunning, P., Horgan, J. (2004). A New Algorithm for the Construction of Stratum Boundaries in Skewed Populations, *Survey Methodology, 30*(2), 159-166.

Knaub, J. R. (2007). Cutoff Sampling and Inference, *InterStat.*

Sarndal, C., Swensson, B., Wretman, J. (2003). *Model Assisted Survey Sampling*. Springer.

Zar, J. H. (1999). Biostatistical Analysis. Third Edition. New Jersey, Prentice-Hal