CE’s Sample Design

David Swanson
Statistical Methods Division
Bureau of Labor Statistics
Overview

- Selection of PSUs
- Sample size (national and PSU-level)
- Selection of addresses within PSUs
- 2006 sample cut
- Quality measures
- Improvements
New Sample of Geographic Areas and Addresses Selected Every Decade

- 2010 Census-Based Sample Design (in development)
Randomness Ensures Representativeness

- CE is a nationwide household survey.
- Geographic areas randomly selected to represent the total U.S.
- Households randomly selected to represent the geographic areas.
- Systematic sampling of households ensures every segment of the population is well-represented.
28 “A” PSUs – metropolitan CBSAs over 2 million people
42 “X” PSUs – metropolitan CBSAs under 2 million people
16 “Y” PSUs – micropolitan CBSAs
16 “Z” PSUs – non-CBSA (“rural”) areas

“A” PSUs are self-representing
“X,” “Y,” “Z” PSUs are non-self-representing
2009 Revision
CPI and CE PSUs

Notes:
Adjacent same-size dass PSUs have different shades of same color.
Labels for X PSUs dropped from CPI and former APSUs that have been reclassified to X are in parentheses.
Non-Self-Representing PSUs

X, Y, Z PSUs are grouped into “strata” according to a 5-variable geographic model:

1. Latitude
2. Longitude
3. Latitude squared
4. Longitude squared
5. Percent urban

One PSU is randomly selected from each stratum (pps)
## PSUs in Stratum X344

<table>
<thead>
<tr>
<th>PSU</th>
<th>Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Charlotte, NC-SC</td>
<td>1,114,808</td>
</tr>
<tr>
<td>Charleston-North Charleston, SC</td>
<td>549,033</td>
</tr>
<tr>
<td><strong>Greenville, SC</strong></td>
<td><strong>379,616</strong></td>
</tr>
<tr>
<td>Fayetteville-Fort Bragg, NC</td>
<td>302,963</td>
</tr>
<tr>
<td>Columbus, GA-AL</td>
<td>274,624</td>
</tr>
<tr>
<td>Gastonia, NC</td>
<td>190,365</td>
</tr>
<tr>
<td>Wheeling, WV-OH</td>
<td>153,172</td>
</tr>
<tr>
<td><strong>Warner Robbins, GA</strong></td>
<td><strong>134,433</strong></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>3,099,014</strong></td>
</tr>
</tbody>
</table>
Household Sample - National

Target Sample Size

- 7,700 interviewed households per year (Diary)
- 7,700 interviewed households per quarter (Interview, interviews 2-5 only)

Target Sample Yield

- 15,400 weekly diaries per year (=7,700×2)
- 30,800 quarterly interviews per year (=7,700×4)
Household Sample - Local

Local Target Sample Size

- Allocate 7,700 interviewed households to individual PSUs proportional to each stratum’s population (pps)
  - Sample allocated to 40 CPI_AREAs (28 “A” PSUs + 12 region-size classes), then sub-allocated to individual PSUs
  - Minimum of 80 interviewed households per CPI_AREA
  - Urban areas over-sampled (and rural areas under-sampled) to help CPI

- Minimizes CE’s nationwide variance
Translate Interviewed Households into Addresses

- 80% “eligibility” rate
- 75% response rate
- 60% “participation” rate \((0.60 = 0.80 \times 0.75)\)
Translate Interviewed Households into Addresses (continued)

<table>
<thead>
<tr>
<th>PSU</th>
<th>Interviewed households</th>
<th>Addresses</th>
</tr>
</thead>
<tbody>
<tr>
<td>A102 Philadelphia</td>
<td>169</td>
<td>322</td>
</tr>
<tr>
<td>A103 Boston</td>
<td>195</td>
<td>286</td>
</tr>
<tr>
<td>A104 Pittsburgh</td>
<td>80</td>
<td>123</td>
</tr>
<tr>
<td>A109 New York City</td>
<td>220</td>
<td>420</td>
</tr>
<tr>
<td>A110 NY-Conn suburbs</td>
<td>212</td>
<td>335</td>
</tr>
<tr>
<td>A111 New Jersey suburbs</td>
<td>182</td>
<td>291</td>
</tr>
<tr>
<td>etc.</td>
<td>etc.</td>
<td>etc.</td>
</tr>
<tr>
<td>Total</td>
<td>7,700</td>
<td>12,800</td>
</tr>
</tbody>
</table>
Select a Random Sample of Addresses

Census Bureau’s “100% Detail File”

Sort from poor to rich (information from 2000 Census)

- Urban/rural
- Number of people in household
- Tenure (owner, renter)
- Market value of home (owners)
- Monthly rent (renters)
Select a Random Sample of Addresses (continued)

Compute the sampling interval for each PSU

Sampling interval = (# addresses in sampling frame) ÷ (# addresses in CE sample)

Typical sampling intervals:
• every 5,000th address (A PSUs)
• every 1,000th address (X,Y,Z PSUs)
Select a Systematic Sample of Addresses (continued)

-- D --- I --- D --- I --- D --- I --- D --- I --- D --- I ---
- D --- I --- etc.

D=Diary, I=Interview

- 12 years of addresses selected
- Reserve sample
Sampling Frames

- Unit frame (85%)
- Area frame (10%)
- Permit frame (5%)
- Group Quarters (<1%)
2006 Sample Cut

7 PSUs changed from “A” to “X”
11 “X” PSUs cut from sample
8% of addresses cut from sample

New Geographic Sample (2006-present)
- 21 “A” PSUs – metropolitan CBSAs over 2.7 million people
- 38 “X” PSUs – metropolitan CBSAs under 2.7 million people
- 16 “Y” PSUs – micropolitan CBSAs
- 16 “Z” PSUs – non-CBSA (“rural”) areas
New Sample Size (2006 – present)

Target Sample Size
- 7,050 interviewed households per year (Diary)
- 7,050 interviewed households per quarter (Interview, interviews 2-5 only)

Target Sample Yield
- 14,100 weekly diaries per year (=7,050×2)
- 28,200 quarterly interviews per year (=7,050×4)
Sample Design Improvements

Recent
- Cluster sampling research
- PSU stratification
- “Optimal” allocation of nationwide sample to individual PSUs
- Nonresponse bias

Future
- Annual Sampling
- Master Address File (MAF)
Estimation
Overview

- Basic estimator
- Weights
- Data adjustments (allocation, imputation)
- Accuracy of estimates (Variances, PCE)
## Average Annual Expenditure per Consumer Unit

<table>
<thead>
<tr>
<th>Item</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Items</td>
<td>$49,638</td>
<td>$50,486</td>
<td>$49,067</td>
</tr>
<tr>
<td>Food</td>
<td>6,133</td>
<td>6,443</td>
<td>6,372</td>
</tr>
<tr>
<td>Housing</td>
<td>16,920</td>
<td>17,109</td>
<td>16,895</td>
</tr>
<tr>
<td>Apparel and Services</td>
<td>1,881</td>
<td>1,801</td>
<td>1,725</td>
</tr>
<tr>
<td>Transportation etc.</td>
<td>8,758</td>
<td>8,604</td>
<td>7,658</td>
</tr>
</tbody>
</table>
CE’s Basic Estimator is a Weighted Average

\[ \overline{x}_i = \frac{\sum_{c} w_c x_{ic}}{\sum_{c} w_c} \]

where

\( w_c \) = weight of consumer unit “c”

\( x_{ic} \) = expenditure of CU=c for item=i
Weights

Final weight
= (Base weight) \times (Nonresponse adjustment factor) \times (Calibration adjustment factor)

Typical values:
15,000 = 10,000 \times 1.33 \times 1.13
Weights: Nonresponse Adjustment

Nonresponse adjustment factor
= (# occupied housing units) ÷ (# responders)

64 demographic groups (64=4×4×2×2)
- 4 regions (Northeast, Midwest, South, West)
- 4 CU sizes (1,2,3-4,5+)
- 2 tenures (owner, renter)
- 2 races (Black, Non-Black)

Typical factor = 1.33
Weights: Calibration Adjustment

Calibration adjustment factor
= (CPS population estimate) ÷ (CE population estimate)

CE population estimate
= (# responders) × (# people per CU) × (nonresponse adjustment factor)

24 population controls from CPS (24=14+8+2)
- 7 Age × 2 Race categories
- 4 Regions × 2 Urban/Rural
- 2 Tenure (owner/renter)

Typical factor = 1.13
Data Adjustments

Allocation – Split combined expenditures into individual item categories

Example:
Reported: $100 clothing
Allocated: $60 shoes, $20 skirts, $20 pants

Imputation – Fill-in unreported values
1. Specific items mentioned, but not specific expenditures (e.g., $100 for shoes, skirts, pants):

Allocate the expenditure to the items reported based on the nationwide distribution of expenditures (e.g., 60% shoes, 20% skirts, 20% pants)
2. Specific items not mentioned (e.g., $100 for clothing)

2a. Identify all item categories whose median expenditure is below the reported expenditure

2b. Randomly select one of those item categories and assign the median expenditure to the item category

2c. Subtract the assigned amount from the reported expenditure

2c. Repeat 2b and 2c until the reported expenditure is allocated completely
Imputation Methods

- Percentage distributions
- Weighting class
- Traditional regression estimator
  \[ y = \beta_0 + \beta_1 x \]
- Regression with multiple imputation
  \[ y_i = \beta_0 + \beta_1 x + \varepsilon_i \quad \text{for } i=1,2,3,4,5 \]
Percentage distribution -- match on 2 or 3 characteristics (e.g., region and income class), then randomly select a value.

IF RANUNI(0) < 0.6750 THEN TENURE='OWNER';
ELSE TENURE='RENTER';

<table>
<thead>
<tr>
<th>Tenure</th>
<th>Percent</th>
<th>Cumulative percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Owner</td>
<td>0.6750</td>
<td>0.6750</td>
</tr>
<tr>
<td>Renter</td>
<td>0.3250</td>
<td>1.0000</td>
</tr>
</tbody>
</table>
Imputation - Expenditures

Weighting class – match on 2 or 3 variables (e.g., region and income class), then use the mean expenditure from all reported expenditures in the cell.

<table>
<thead>
<tr>
<th>Income quartile</th>
<th>Mean reported expenditure for pre-paid phone cards</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$78.51</td>
</tr>
<tr>
<td>2</td>
<td>67.80</td>
</tr>
<tr>
<td>3</td>
<td>61.51</td>
</tr>
<tr>
<td>4</td>
<td>53.53</td>
</tr>
</tbody>
</table>
Rental Equivalence

“If someone were to rent this home today, how much do you think it would rent for monthly, unfurnished and without utilities?”

20% of CPI’s weight.
Old: Use the reported value from the most similar housing unit (hotdeck)

New: Traditional regression estimator
(y = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n)

y = \beta_0 + \beta_1 \cdot \text{Bedrooms} + \beta_2 \cdot \text{Bathrooms} + \beta_3 \cdot \text{Property Value} + ...
Imputation - Income

- Old: None (analysis limited to “complete reporters”)
- New: Multiple imputation (5 estimates)
- Cold-deck (5 years of data)
- Backward model selection
- Variable selection & parameter estimation done quarterly

\[ \text{Salary} = \beta_0 + \beta_1 \cdot \text{Age} + \beta_2 \cdot \text{Education} + \ldots + \varepsilon_i \]
\[ \text{Pension} = \beta_0 + \beta_1 \cdot \text{Age} + \beta_2 \cdot \text{Retired} + \ldots + \varepsilon_i \]

for \( i=1,2,3,4,5 \)
### Accuracy of Estimates: Standard Errors

**CE Standard Errors for 2009**

<table>
<thead>
<tr>
<th>Item</th>
<th>Annual Expenditure</th>
<th>Standard Error</th>
<th>Coefficient of Variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Items</td>
<td>$49,067</td>
<td>$595</td>
<td>1.21%</td>
</tr>
<tr>
<td>Housing</td>
<td>16,895</td>
<td>194</td>
<td>1.15</td>
</tr>
<tr>
<td>Transportation</td>
<td>7,658</td>
<td>165</td>
<td>2.16</td>
</tr>
<tr>
<td>Food</td>
<td>6,372</td>
<td>84</td>
<td>1.31</td>
</tr>
<tr>
<td>etc.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

SEs are estimated with Balanced Repeated Replication (BRR)
# Quality Measures: Coverage Rate

<table>
<thead>
<tr>
<th></th>
<th>2004</th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>88%</td>
<td>87</td>
<td>88</td>
<td>87</td>
<td>86</td>
<td>86</td>
</tr>
<tr>
<td>Female</td>
<td>87</td>
<td>86</td>
<td>87</td>
<td>86</td>
<td>86</td>
<td>85</td>
</tr>
<tr>
<td>Non-Black</td>
<td>89</td>
<td>89</td>
<td>89</td>
<td>89</td>
<td>88</td>
<td>88</td>
</tr>
<tr>
<td>Black</td>
<td>71</td>
<td>72</td>
<td>72</td>
<td>69</td>
<td>72</td>
<td>68</td>
</tr>
</tbody>
</table>

**Coverage Rate** = \( \frac{\text{CE population estimate}}{\text{CPS population estimate}} \times 100\% \)

**CE population estimate** = \( \text{(# responders)} \times (\text{# people per CU}) \times (\text{nonresponse adjustment factor}) \)
Recent
- CAPI (fewer outliers & missing values)
- User-Friendly Diary
- Rental Equivalence imputation
- Income imputation

Future
- Coverage of the Black population
- Income tax imputation