Update on CPS Research, Discussion for the Session

Edwin L Robison
Bureau of Labor Statistics

1. Importance of the Current Population Survey

The Current Population Survey (CPS) of households is a source of national and subnational labor force data for the United States. It is the source of the monthly official unemployment rate which is a Primary Federal Economic Indicator (PFEI). It is a major input for the Local Area Unemployment Statistics program (LAUS). The distribution of some federal funds to states and localities is tied to estimates generated from CPS data.

Given its importance, the sponsoring agencies (Bureau of Labor Statistics and the Census Bureau) only implement changes that are supported by sufficient analysis. Extreme care is taken when making methodological changes that affect "topside" national estimates of unemployment and employment. But there are many other estimates that the agencies examine for possible disruption to the time series.

2. The Session

Topics for ongoing Current Population Survey research topics include: subnational sample allocation, variance estimation, Generalized Variance Functions, nonresponse adjustment, benchmarking to population controls, composite estimation, small area estimation, time series models, and seasonal adjustment. The four presentations represent some of this research:

- Current Population Survey Sample Size Study (Daniel Sommers, Stephanie Chan Yang, and Yang Cheng)An Iterative Composite Estimator in the Current Population Survey (Yang Cheng, Zhou Yu, and Jun Shao)
- An Overview of Current Population Survey Variance Methodology (Aaron Gilary, Yang Cheng, and Eric Slud)

Calculating Generalized Variance Functions with a Single-Series Model in the Current Population Survey (Justin McIllece)

The following discussion is restricted to the presented topics. The intent is not to summarize or critique, but to add to the overall context of the research. Three sections follow: Sampling, Variance Estimation and Generalized Variance Functions, and Composite Estimation.

3. Sampling

The Current Population Survey is typically described as having a two-stage sample design with rotating panels. In the first stage, Primary Sampling Units (PSUs) are selected. In the second stage, clusters of housing units (HUs) are selected. An independent sample of housing units is selected in each state (state-based design).

1st-stage Primary Sampling Units

- Each PSU is a county or cluster of counties.

- Self-representing (SR) PSUs in a state are the PSUs with the largest populations and are taken with certainty. For some states all PSUs are SR PSUs.
- The remaining non-self-representing (NSR) PSUs in a state are stratified and sampled. The contribution to variance is called between-PSU variance.
- Probability Proportional to Size sampling in each stratum
- PSU overlap with prior design maximized (for less administrative disruption, lower cost)

2nd-stage sample of housing unit clusters

- Ultimate Sampling Unit (USU) clusters of about 4 housing units
- Sort HUs in a PSU; form USUs and systematically sample them.

The Current Population Survey sample has been redesigned after each decennial census. Historically (prior to the 2010 redesign):

1st-stage PSUs

- SR PSUs and sampled NSR PSUs to be kept for a decade
- Census long-form data used to guide PSU sampling (available for all counties)
- For states, between-PSU variances computed using long form data on labor force (Adjusted through the decade only for overall growth)

2nd-stage HU clusters

- Decennial census used to create a list of housing units for sampling
- HU sample for a decade selected all at once
- Supplemented through the decade with new housing
- Occasional "sample maintenance reductions" needed to control costs (Since new housing made the sample grow)

-

3.1 2010 Redesign Flexibility

New Census products led to incorporating some potential flexibility in Current Population Survey sampling and in estimating the between-PSU portion of state variance.

- American Community Survey (ACS) – Released each year, its 5-year estimates cover every county. The ACS made a long form unnecessary for the 2010 decennial census.
- Master Address File (MAF) – Census has a continuously updated MAF. New housing units are added; out-of-scope and destroyed units are identified and dropped.

1st-stage PSUs

- 5-year ACS replaces long form data; updated each year
- Can reselect PSUs in states with large shifts, using more recent data
- Possibility of annually recomputing state between-PSU variances
- Possibility of model-based modification of NSR PSU selection probabilities

2nd-stage HU clusters

- Annual sampling – sample for a year selected from the Master Address File
- Incremental yearly adjustments can eliminate sample maintenance reduction
    - (Design issues covered by Sommers, Yang, and Cheng address this)
- Eases sample increases/decreases in response to budgets and data needs

Continuously modifying the CPS sample in response to recessions or other economic changes is not envisioned. Too much flexibility, that is modifying the sample too frequently, can make survey administration a nightmare. Sudden large increases in the sample are not tenable under CPS personal/telephone computer data collection methods; administering the relatively long questionnaire requires more than superficial training.

Note that current CPS design criteria aim at stability. An assumed 6 percent unemployment rate (UER) is referenced, not a real unemployment rate. An assumed 6 percent UER is fixed and does not react to changes in national UER and does not take into account state differences in UER. Neither of the coefficient of variation (CV) criteria given here called for more/less sample when UER roughly doubled after the recession hit in 2008. The criteria provide for stable state allocations that are appropriate for a wide variety of labor force data. The assumed 6 percent UER is just a convenient tool.

- State Design Criterion – Assuming 6 percent UER, obtain an 8 percent CV (or better) on the annual average unemployment level.
- National Design Criterion – Assuming 6 percent UER, obtain a 1.9 percent CV on the monthly unemployment level. This is set so that if the national UER really is 6 percent, and using a 90 percent confidence interval, then a 0.2 percent monthly change will be significant.

4. Variance Estimation and Generalized Variance Functions (GVFs)

Replicate weighting using 160 replicates has been in Current Population Survey monthly processing since 1996. The paper by Gilary, Cheng, and Slud covers the current Successive Difference Replication methodology. Replicate estimates of variance tend to be noisy, hence the need for national Generalized Variance Functions covered by McIllece.

Assignment to replicates is based on the first noncertainty stage of sampling. For self-representing Primary Sampling Units, the sampled clusters of housing units are used. For non-self-representing PSUs, it is the sampled NSR PSUs that are used.

The choice of 160 replicates is appropriate for the number of NSR PSUs for the nation. If a term is needed, the replicate weights used for the nation are “total” replicate weights. For the nation there is no need to separately compute the between-PSU and within-PSU contributions to variance. The 160 “total” replicates are also appropriate for states where all PSUs are self-representing.

There is a problem for states where some PSUs are non-self-representing. There are too few NSR PSUs to support 160 replicates for computing “total” state variances. The solution:

- Express the state variance as $V_{total} = V_{between\text{-}PSU} + V_{within\text{-}PSU}$
- Compute $V_{within\text{-}PSU}$ (noisy) using within-PSU replicate weights
- Add on $V_{between\text{-}PSU}$ computed using American Community Survey data

4.1 State GVFs

For states, the replicate estimates of the within-PSU portion of variance are noisy. In recently completed Bureau of Labor Statistics work, a simple SRS-type formula is modified by a design effect to create a GVF. (For estimates of "total" state variance there is an added adjustment for the between-PSU contribution to variance. State between-PSU variances are estimated using American Community Survey data.)

- SRS-type formula:
  $V \approx N^2p(1\text{-}p)/n$, where
  N= adult population
  p is a ratio such as unemployment/N or employment /N
- Can be converted to the form:
  $ax^2 + bx$, where
  b = SI = N/n the sampling interval
  a = -b/N
  x = Np
  n=sampled adults
  or: b=N/r where r is the number of sampled adults

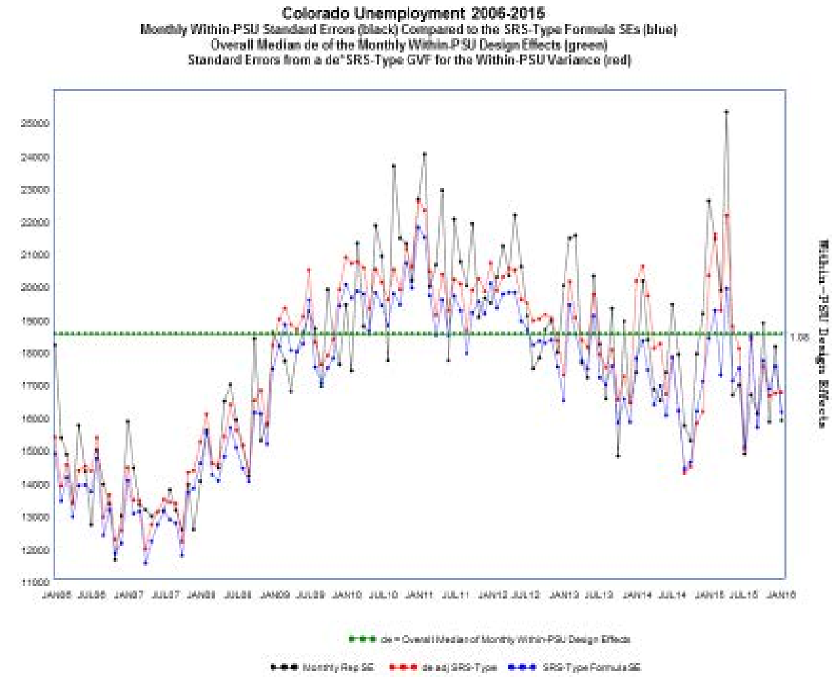
This is called an "SRS-type" formula since the Current Population Survey has a sample of households, not of adult persons. It turns out an effective b parameter is the average of final weights in a state. It accounts for just about everything that is expected to affect variances except changes in clustering over time:

- Accounts for growth in N and sample changes
- Accounts for monthly nonresponse differences
- Accounts for monthly benchmarking differences in complex CPS weighting

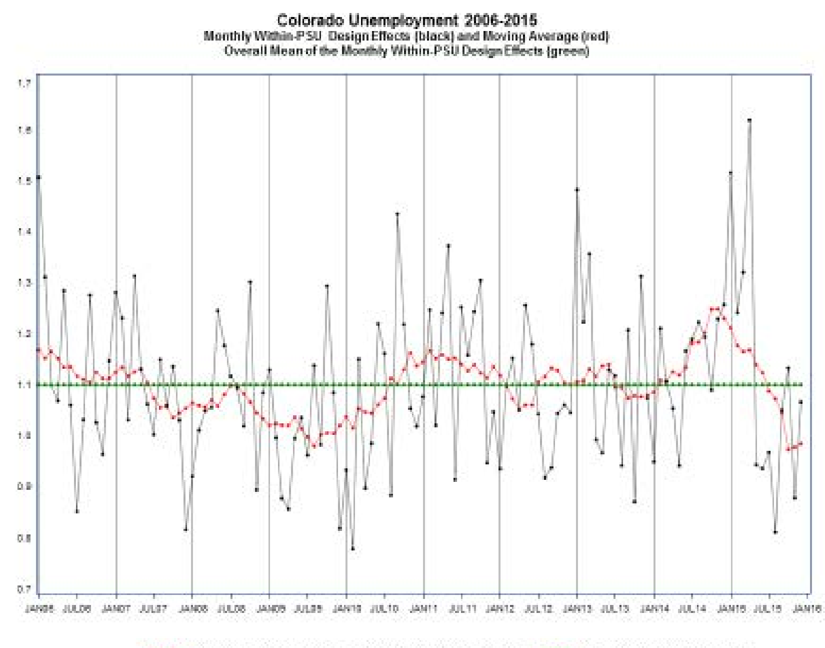
The SRS-type tracks $V_{within\text{-}PSU}$ well for over ten years in every state, for both unemployment and employment. The two GVFs in each state are modified by 10-year design effects to account for small long-term differences between simple SRS-type GVFs and the within-PSU replicate variances $V_{within\text{-}PSU}$ that they model.

- Monthly: $de_{mo} = V_{within\text{-}PSU}$ /SRS-type for unemployment, employment
- There are no noticeable design effect changes over ten years, even for the recession period (starting in 2008) when unemployment doubled.
- de is a 10-year average of the monthly de_{mo} for unemployment, employment.
- $de(ax^2 + bx)$ is the GVF for within-PSU variance for unemployment, employment.
- Final adjustments for between-PSU variance are not covered here.

In this plot for Colorado unemployment, blue is the SRS-type line and for ten years (2006-January 2015) it follows changes in the black line for the replicate variances $V_{within\text{-}PSU}$. Since the blue SRS-type line is for a GVF, it is less noisy than the replicate variances, but the noise observed from one month to the next are due to changes in parameters.



The following plot shows monthly design effects for Colorado unemployment that average about 1.1 for the ten years. The red line is a 12-month moving average. There is an apparent modest dip in design effect roughly coincident with the recession that started in 2008; there is also an apparent up-down swing in design effect roughly coincident with the phase-in of 2010 redesign sample for the CPS. Although some other states have apparent dips/increases/swings in design effects there is little connection in the timing to either the recession or the design phase-in. More research is needed to understand the modest design effect changes.



4.2 Relationship of State Design Effects and National Design Effects

There are relatively small differences in the state within-PSU design effects.

Unemployment design effects
- Average of states: 1.15
- Range of states: 1.05-1.26

Employment design effects
- Average of states: 0.96
- Range of states: 0.87-1.09 (excluding 1.34 outlier)

State employment design effects are lower than those for unemployment. This is due to the CPS weighting procedure that includes benchmarking to eight population controls for each state (gender by four age groups). The effect is to somewhat lower variances on employment estimates (relatively strong correlation with the benchmarks). But there is little reduction in variance for unemployment estimates (only weakly correlated with the benchmarks).

Within-PSU design effects were computed for the nation.
- National $V_{within\text{-}PSU}$ computed using within-PSU replicates (not total replicates)
- National "Stratified SRS-type" -- the sum of state SRS-type
- Monthly: $de_{mo} = V_{within\text{-}PSU}$ /Stratified SRS-type for unemployment, employment
- There are no noticeable national design effect changes over ten years, even for the recession period (starting in 2008) when unemployment roughly doubled.
- Unemployment 10-year average de 1.16 (almost the same as state average 1.15)
- Employment 10-year average de 0.75 (lower than state average 0.96)

For employment, the national design effect (0.75) is lower than the average of states (0.96). The CPS weighting procedure includes complex benchmarking to many population controls. The correlation of employment estimates with the benchmarks is relatively strong. The effect is to lower variances on national employment estimates, and consequently to lower the design effect. (There is little reduction in variances of national unemployment estimates due to the complex national benchmarking.)

5. Composite Estimation

Several weighting steps precede CPS composite estimation.
- Basic weighting for housing unit selection probabilities
- Adjustment for housing unit nonresponse
- First-Stage weighting restricted to Non-Self-Representing Primary Sampling Units
- National coverage adjustment
- State coverage adjustment
- Second-Stage weighting with complex benchmarking to population controls
    - Three steps iterated ten times: state step, race step, ethnicity step
    - Step population controls defined by geography, gender, race, ethnicity, and age

Composite estimation (refer to Technical Paper 66) is engineered to reduce variances on the most important CPS estimates: month-to-month change in unemployment and month-to-month change in employment. Composite estimation can also lower variances on monthly estimates of unemployment and employment levels. Before composite estimation even comes into the process, two aspects of CPS sampling and weighting are designed to lower variances.

- 4-8-4 panel rotation – The CPS sample is divided into panels. Each month eight panels are included. From one month to the next two panels drop out and two panels are added. Consecutive months have six panels in common. The panel overlap helps lower variances on estimates of change (compared to independent monthly samples).
- Benchmarking to population controls – Particularly for "large" variables such as unemployment that have relatively strong correlation with populations, benchmarking substantially lowers variances on estimates of levels (compared to no benchmarking). Variances on month-to-month change can also be lowered.

In 4-8-4 panel rotation a panel is included 4 consecutive months, is dropped for 8 months, then returns for another 4 consecutive months. For example, a panel may be included February 2014, March 2014, April 2014, May 2014, February 2015, March 2015, April 2015, and May 2015. The eight months that a panel is included are called months-in-sample (MIS) and are frequently labeled MIS1-MIS8. The somewhat simplified formulas that follow have a first subscript for the preceding month t-1 or current month t, and a second subscript of 1-8 for the MIS. The symbol $\Downarrow$ is used to indicate a panel that continues from one month to the next.

$$\begin{array}{llllllllll} YSS_{t-1} = & YSS_{t-1,8} + & YSS_{t-1,7} + & YSS_{t-1,6} + & YSS_{t-1,5} & & + YSS_{t-1,4} + & YSS_{t-1,3} + & YSS_{t-1,2} + & YSS_{t-1,1} \\ & & \Downarrow & \Downarrow & \Downarrow & & & \Downarrow & \Downarrow & \Downarrow \\ YSS_t = & & YSS_{t,8} & + YSS_{t,7} & + YSS_{t,6} & + YSS_{t,5} & & + YSS_{t,4} & + YSS_{t,3} & + YSS_{t,2} \quad + YSS_{t,1} \end{array}$$

$Var(YSS_{t-1}) = Var(YSS_t) \approx 8\sigma$,assuming that $\sigma = Var(YSS_{t-1,i}) = Var(YSS_{t,i})$ for all MIS i

$Var(YSS_t - YSS_{t-1}) = Var(YSS_t) + Var(YSS_{t-1}) - 2Cov(YSS_t, YSS_{t-1}) \approx 8\sigma + 8\sigma - 2(6c)$,
assuming the six continuing panels covariance terms are all c (zero otherwise)

What drove composite estimation development was the desire to do better. It is possible to develop a month-to-month change estimate $\Delta_{t-1,t}$ that uses only continuing panels. Its variance can be smaller when c is strongly positive.

$$\Delta_{t-1,t} = (8/6)\ [\ + YSS_{t,8} + YSS_{t,7} + YSS_{t,6} + YSS_{t,4} + YSS_{t,3} + YSS_{t,2} - YSS_{t-1,7} - YSS_{t-1,6} - YSS_{t-1,5} - YSS_{t-1,3} - YSS_{t-1,2} - YSS_{t-1,1}\ ]$$

$Var[\Delta_{t-1,t}] \approx (8/6)^2[6\sigma + 6\sigma - 2(6c)] = (8/6)[8\sigma + 8\sigma - (8/6)2(6c)]$

There are problems associated with using $\Delta_{t-1,t}$ as an estimate of variance. If the official estimates for two consecutive months are YSS_{t-1} and YSS_t, then it can be a source of confusion that $\Delta_{t-1,t}$ won't equal their difference. You could force agreement by using a modified set of official estimates. The following example starts with a December second-stage estimate and successively adds Δ terms. By the time June is reached, Y_{June} can stray quite a bit from the "unbiased" YSS_{June}.

December	YSS_{Dec}
January	$Y_{Jan} = YSS_{Dec} + \Delta_{Dec,Jan}$
February	$Y_{Feb} = YSS_{Dec} + \Delta_{Dec,Jan} + \Delta_{Jan,Feb}$
March	$Y_{Mar} = YSS_{Dec} + \Delta_{Dec,Jan} + \Delta_{Jan,Feb} + \Delta_{Feb,Mar}$
April	$Y_{Apr} = YSS_{Dec} + \Delta_{Dec,Jan} + \Delta_{Jan,Feb} + \Delta_{Feb,Mar} + \Delta_{Mar,Apr}$
May	$Y_{May} = YSS_{Dec} + \Delta_{Dec,Jan} + \Delta_{Jan,Feb} + \Delta_{Feb,Mar} + \Delta_{Mar,Apr} + \Delta_{Apr,May}$
June	$Y_{June} = YSS_{Dec} + \Delta_{Dec,Jan} + \Delta_{Jan,Feb} + \Delta_{Feb,Mar} + \Delta_{Mar,Apr} + \Delta_{Apr,May} + \Delta_{May,June}$

A simple composite estimatior makes a clever use of the Δ terms. Assuming unbiasedness, two estimators can be computed for a given month (March in the example). And a weighted average of the two is also unbiased ($0 < K < 1$).

- YSS_{Mar} "unbiased" for March
- $YSS_{Feb} + \Delta_{Feb,Mar}$ "unbiased" for March
- Weighted average of the two also "unbiased"
    - $YCOMP_{Mar} = (1\text{-}K)\ YSS_{Mar} + K(YSS_{Feb} + \Delta_{Feb,Mar})$
    - $YCOMP_t = (1\text{-}K)\ YSS_t + K(YSS_{t-1} + \Delta_{t-1,t})$
- Change to an iterative process by replacing YSS_{t-1} with previous $YCOMP_{t-1}$
    - $YCOMP_{Mar} = (1\text{-}K)\ YSS_{Mar} + K(YCOMP_{Feb} + \Delta_{Feb,Mar})$
    - $YCOMP_t = (1\text{-}K)\ YSS_t + K(YCOMP_{t-1} + \Delta_{t-1,t})$

To complete the process, the "unbiased" estimates $YCOMP_{t-1}$ and $YCOMP_t$ replace the second stage ones as the official estimates of monthly levels. In general for CPS, a lower variance on month-to-month change can be obtained.

- New month-to-month changes/differences $YCOMP_t - YCOMP_{t-1}$ "unbiased"
- In general, it is possible to have $Var(YCOMP_t - YCOMP_{t-1}) < Var(YSS_t - YSS_{t-1})$
- $V(YCOMP_t) < V(YSS_t)$ is also possible (reduction in variance on levels, not necessarily mean squared error)

The variance reduction is free, in the sense that no more sample is needed. But in practice there are two problems:

- Reduction in variance on month-to-month change is not as great as anticipated
- $YCOMP_t$ can stray too much from YSS_t

The chief underlying problem is Month-in-Sample bias; the previous estimates labeled as "unbiased" do have troublesome biases. For MIS1-MIS8 there are observed systematic differences in $YSS_{t,i}$ for the eight panels active in a given month. Alternatively, single panels can be tracked and systematic differences found for MIS1-MIS8. It is not clear which MIS is least biased.

A term $A\beta_t$ can be added to the composite estimation. The A is a constant and the computation of β_t using only month t panels is shown.

- β_t is a difference between incoming MIS5 and MIS1 panels

and overlap panels MIS8, MIS7, MIS6, MIS4, MIS3, and MIS2
- Recovers some of the gains on the variance of composited month-to-month change
- Also reduces how much composite $YCOMP_t$ strays from YSS_t
- Sometimes called a bias correction term

$$\begin{array}{lllllllll}
YSS_{t-1} = & YSS_{t-1,8} + & YSS_{t-1,7} + & YSS_{t-1,6} + & YSS_{t-1,5} & + YSS_{t-1,4} + & YSS_{t-1,3} + & YSS_{t-1,2} + & YSS_{t-1,1} \\
 & & \Downarrow & \Downarrow & \Downarrow & & \Downarrow & \Downarrow & \Downarrow \\
YSS_t = & & YSS_{t,8} & + YSS_{t,7} & + YSS_{t,6} + YSS_{t,5} & & + YSS_{t,4} & + YSS_{t,3} & + YSS_{t,2} + YSS_{t,1}
\end{array}$$

$$\begin{array}{lll}
\beta_t = & [& + YSS_{t,5} \qquad\qquad\qquad + YSS_{t,1} \,] \\
-(1/3) & [& + YSS_{t,8} + YSS_{t,7} + YSS_{t,6} \qquad + YSS_{t,4} + YSS_{t,3} + YSS_{t,2} \,]
\end{array}$$

$$YCOMP_t = (1\text{-}K)\, YSS_t + K(YCOMP_{t-1} + \Delta_{t-1,t}) + A\beta_t$$

Certainly MIS bias makes things tricky for composite estimation. There is a tradeoff between reducing the variance of month-to-month change and limiting the degree to which monthly composite estimates stray from second-stage estimates. The Cheng, Yu and Shao paper considers a family of composite estimators and concentrates on the mean square estimates of levels – and in one sense examines how much composite estimators stray from unbiasedness. Erkens (2012) used a more general composite and multiple objective functions (variance of month-to-month change and bias/stray).

Here are recommendations for future research on composite estimation.
- Examine both the variance of month-to-month change and the difference between composite estimates and second-stage estimates (the "stray").
- MIS bias is the driver, and keep in mind that it is known to change over time.
- Analyze many years of data and check how optimum parameters change over time. (2003 forward is a good choice since methods and data files contents are stable.)
- Currently there are separate parameters for employment and unemployment. Also examine not-in-labor-force.
- Current parameters were developed based primarily on topside employment and unemployment. Give more attention to application by gender/race/ethnicity/age.

References

U.S. Census Bureau. 2006. *Current Population Survey Design and Methodology, Technical Paper 66*.
(www.census.gov/cps/methodology/techdocs.html)

Erkens, Greg. 2012. Changes in Panel Bias in the U.S, Current Population Survey and its Effects on Labor Force Estimates. *JSM Proceedings*, Section on Survey Research Methods. Alexandria, VA; American Statistical Association.