On Generalized Variance Functions for Sample Means and Medians November 2018

Justin J. McIllece[1]
[1]U.S. Bureau of Labor Statistics, 2 Massachusetts Ave NE #4985/02, Washington, DC 20212

Abstract
Generalized variance functions (GVFs) and their associated parameters are most often applied to binomial characteristics of complex surveys, usually within the Federal Statistical System. Non-binomial estimates have less GVF support in the general body of literature. This paper presents a framework for modeling the variances of sample means and quantiles and calculating generalized parameters. The models are applied to the mean and median duration unemployed characteristics from the Current Population Survey (CPS), the only two estimates in the primary CPS estimation tables – from *The Employment Situation*, a Principal Federal Economic Indicator published monthly by the Bureau of Labor Statistics – for which no GVF parameters are produced. The performance of these GVFs is evaluated, and plans for implementation are discussed.

Key Words: CPS, GVF, generalized variance functions, replication, quantiles

1. Introduction

Generalized variance functions (GVFs) provide a convenient solution to the problem of variance estimation in complex surveys, in which sample designs involve multiple stages. A cursory review of the literature regarding GVFs suggests that their applications tend to be concentrated in the Federal Statistical System, primarily for estimating the variances of proportions and totals from binomially-distributed data. Valliant (1987) extends GVF relative variance models of the form $a + b/X$ beyond the binomial case and provides conditions for theoretical support. Most GVFs adhere closely to this relative variance model, which was used by the Current Population Survey (CPS) from 1947 through 2015 and is described in *Technical Paper 66* (2006); specifically, these functions relate relative variance to the inverse of the associated point estimate for a cluster of estimation series. McIllece (2016) presented an alternative construction of GVFs for CPS variances, forming single-series models that rely on each series' longitudinal histories instead of cross-sectional grouping, a clustering variation also briefly suggested (but not exposited upon) by Wolter (2007). This model, though of similar formulaic construction, differs in that does not use the relative variance as the dependent quantity (Section 2). The CPS has employed single-series GVFs for published standard errors since August 2015.

[1] Views expressed are those of the author and do not necessarily reflect the views or policies of the U.S. Bureau of Labor Statistics.

Adaptations to the variances of sample means and medians have been applied in a limited number of surveys. The American Community Survey (ACS 2010) generalized variances for its 2005 public-use data, utilizing "design factors" multiplied by base variance formulas that vary depending upon estimate type—counts, means, or medians. To estimate standard errors for means in the 1990-1991 Schools and Staffing Survey, the National Center for Education Statistics (NCES 1995) modified the GVF of a total by dividing by the corresponding sum of the weights. The variance estimates in the cited survey applications are of inconsistent, and arguably insufficient, quality, as noted by the authors themselves:

- In ACS 2010, it is stated that less than one percent of the standard errors of means at the state level fall in the acceptable range, while about 60 percent of the standard errors of medians are considered acceptable[2]. The authors conclude: "future research may explore new methods for producing DFs [design factors] for the mean and medians."

- In NCES 1995, a comprehensive evaluation of the quality of the standard errors of means is not given, but in one of the two examples it is noted that "the result from using GVF seems not to give satisfactory accuracy." Notably, this conclusion was drawn from a relative difference comparison to the replicate variance, but given the volatility of replication-based variance estimators, this quality benchmark can be a moving target.

The extension of the GVF framework to order statistics, in particular, seems almost nonexistent, excepting ACS 2010. The variance properties of order statistics do not naturally lend themselves neatly to design-based estimates nor to a modeling process, but considering the abundance of published medians derived from complex surveys in the Federal Statistical System, a reliable GVF solution would seem to be of widespread utility to both survey agencies and public data users.

In this paper, GVF models are developed for the mean and median duration unemployed statistics (reported in weeks) published by the CPS. Of 665 CPS estimates series in *The Employment Situation*[3] news release tables, a Principal Federal Economic Indicator produced monthly by the Bureau of Labor Statistics, mean and median duration unemployed are the only two for which no GVF parameters are currently computed[4]. The precision of the GVF standard errors is evaluated for these two series over both the modeling period and a two-year projection. A discussion of the internal and external usefulness of these models concludes the paper.

2. GVF Models for Sample Rates and Totals

The CPS has utilized GVFs for binomial estimates, such as rates and totals, since 1947 (*Technical Paper 66*). Until 2015, relative variance for a group of estimation series was modeled by iterative weighted least squares estimation, and the parameters from this model

[2] It may be more accurate to report this number as 36 percent, since the 60 percent of acceptable standard errors were only based on approximately 60 percent of the original set of estimates. About 40 percent of the replicate standard errors of the medians were either zero or undefined (due to falling in the maximum category).

[3] https://www.bls.gov/news.release/empsit.toc.htm

[4] https://www.bls.gov/cps/documentation.htm#reliability

were used for all rates and totals in *The Employment Situation* news release tables derived from CPS data:

$$\frac{V(\hat{X})}{\hat{X}^2} = a + b\hat{X}^{-1}$$

The inverse of the expected variance was used as the vector of series weights for this regression model. Once a and b parameters were estimated, standard errors could be calculated from this "classical" GVF:

$$SE(\hat{X}) = \sqrt{V(\hat{X})} = \sqrt{a\hat{X}^2 + b\hat{X}} \quad (1)$$

In the ideal scenario, estimation series with similar variance properties were clustered prior to model fitting, resulting in a GVF model that produced reliable, accurate standard error estimates for all component series.

In August 2015, several methodological changes were implemented, detailed in McIllece (2016). Instead of grouping across estimation, GVFs were constructed separately for each individual series, using lengthier longitudinal histories (usually ten years of monthly estimates) as the grouping mechanism. This change all but guaranteed that the similarity condition of the cluster was met, since it is unlikely for a series to experience consequential shifts of its own variance properties even during turbulent economic periods, absent the presence of some identifiable interruption, such as a redefinition or changing bounds in a censored response item. The consistency of CPS design effects through the Great Recession of the late 2000s is demonstrated at both the national level (McIllece 2016) and state level (Zimmerman and Robison 2018) in recent research.

Additionally, the product of the sampling interval (which changes monthly) and the design effect (computed via replication) replaced the relative variance as the dependent quantity of interest. In the classical model, this is the b parameter itself. Given a period without substantial changes to the CPS sample size, an ordinary least squares (OLS) regression between the centered b and the centered civilian noinstitutional population (CNP) forms the simple foundation for a GVF that accounts for changing population size while smoothing through the volatility inherent in the replicate weights. Model (1) had not changed, but its parameter estimation had, resulting in more reliable, accurate standard error estimates.

The reconfigured models have been extended to publication tables beyond the *Employment Situation*. Recent adjustments allow for the a and b parameters to be replaced by α and β (*Technical Paper 77*) in a variation of model (1). The former parameters are based on a specific CNP, and therefore work well when applied to months with a similar population total. The latter parameters are estimated by the same fundamental model but allow the CNP to vary, making them effective over a much longer timeframe.

Despite recent advancements, still no standard errors for mean and median weeks unemployed have been published. Sections 3 and 4 present GVF models intended to fill these gaps.

3. GVF Model for Sample Means

For a survey with weights w_i, the variance of the weighted sample mean $\bar{x}$ is given by the function

$$V(\bar{x}) = V\left(\frac{\sum_i w_i x_i}{\sum_i w_i}\right)$$

and assuming the broad condition that x_i are independently distributed with $V(x_i) = \sigma^2$:

$$V(\bar{x}) = \frac{V(w_1 x_1 + \cdots + w_n x_n)}{(\sum_i w_i)^2} = \frac{\sum_i w_i^2 \sigma^2}{(\sum_i w_i)^2}$$

Using the sample variance s^2 to estimate the population variance σ^2:

$$s^2 = \frac{\sum_i w_i (x_i - \bar{x})^2}{\sum_i w_i}$$

Then the variance of the weighted sample mean is estimated by

$$V(\bar{x}) \cong \frac{\sum_i w_i^2 \sum_i w_i (x_i - \bar{x})^2}{(\sum_i w_i)^3} \tag{2}$$

where $\sum_i w_i = \hat{Y}$, the estimate of total unemployed persons in the CNP, since the mean estimate is only derived from the unemployed subset. Since (2) is an approximation that does not fully account for the complex sample design of the CPS, a modified design effect d is included multiplicatively, yielding the approximate standard error formula

$$SE(\bar{x}) = \sqrt{V(\bar{x})} \cong \sqrt{\frac{\sum_i w_i^2 \sum_i w_i (x_i - \bar{x})^2 * d}{\hat{Y}^3}} \tag{2.1}$$

Letting $\sigma_0 = \sqrt{\sum_i w_i^2 \sum_i w_i (x_i - \bar{x})^2 * d}$ to simplify notation, the standard error can be approximated as

$$SE(\bar{x}) \cong \frac{\sigma_0}{\hat{Y}\sqrt{\hat{Y}}} \tag{2.2}$$

As in other CPS variance models, such as McIllece (2016) or the historical models detailed in *Technical Paper 66*, replicate variances are computed and form the basis for modeling. As noted in those references, replicate variances are typically not published directly because of their volatility. According to Wolter (2007), "GVFs simultaneously estimate variances for groups of statistics rather than individually...it may be that some additional stability is imparted to the variance estimates when they are so estimated. He also remarks that "at present...there is no theoretical basis for this claim." In the CPS, the instability of monthly replicate variances has been observational, and the need for smoothing models apparent. Charts in later sections demonstrate this visually.

The framework for building GVF models (for both the sample mean in this section and the sample median in the next section) is to use replication to measure the hard-to-obtain quantities in the formulaic approximation, then model those replicated quantities by a model that delivers both quality fits and efficient application. For the mean, that results in substituting the replicate standard error SE_r for $SE(\bar{x})$ and rearranging (2.2) to isolate σ_0:

$$\sigma_0 \cong SE_r * \hat{Y} * \sqrt{\hat{Y}}$$

which implicitly accounts for the modified design effect d. An OLS regression model, using as predictors the estimated totals of unemployed persons ($\hat{Y}$) and weeks unemployed ($\hat{X}$), is then fit to σ_0:

$$\hat{\sigma}_0 = \hat{\beta}_0\hat{Y} + \hat{\beta}_1\hat{X} \qquad (2.3)$$

Then, substituting (2.3) into (2.2), and noting that the sample mean weeks unemployed $\bar{x} = \hat{X}/\hat{Y}$:

$$\widehat{SE}(\bar{x}) = \frac{\hat{\beta}_0\hat{Y} + \hat{\beta}_1\hat{X}}{\hat{Y}\sqrt{\hat{Y}}} = \frac{\hat{\beta}_0 + \hat{\beta}_1\bar{x}}{\sqrt{\hat{Y}}} \qquad (2.4)$$

The final GVF model (2.4) for the mean weeks unemployed ($\bar{x}$) requires two published parameters ($\hat{\beta}_0, \hat{\beta}_1$) and two published estimates ($\bar{x}, \hat{Y}$) as inputs. The results of fitting this model to 2011 – 2017 data are displayed in Figure 1.

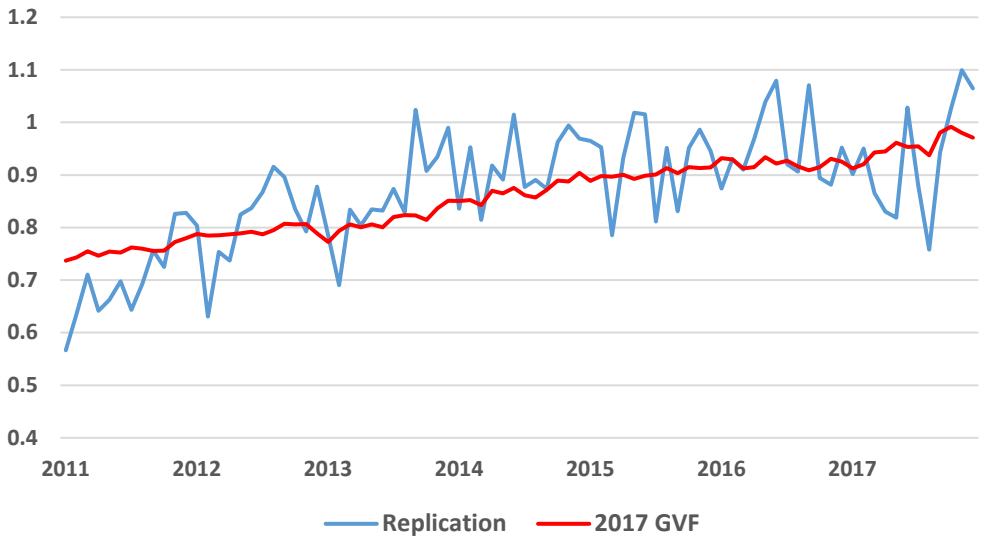

Figure 1: Estimated standard errors of seasonally adjusted mean weeks unemployed

The blue line, labeled "Replication," is not the directly replicated standard error estimate, which is based on not seasonally adjusted data. Instead, the relative standard error is replicated, and that result is multiplied by the seasonally adjusted mean weeks unemployed estimates. This series is compared to the standard error produced from (2.4), using seasonally adjusted estimates of $\bar{x}$ and $\hat{Y}$. Mitigating the effects of seasonality allows focus

on the gains in stability, relative to the replication method, and the overall quality of fit. Further, seasonally adjusted estimates are often of greater economic interest.

The GVF standard errors exhibit much less volatility than the replication-based estimates. The model fit generally hews toward the center of the replicates, although there are extended periods in which the GVF result tends to be consistently higher (2011) or lower (2013, 2014[5]). Such periods are common for GVF models in the CPS, likely due to the panel design, which induces correlations into the response data over time. Robustness against the peaks and valleys of the unstable replicates is a desirable attribute of the modeled standard error, facilitating more effective analyses of change over time.

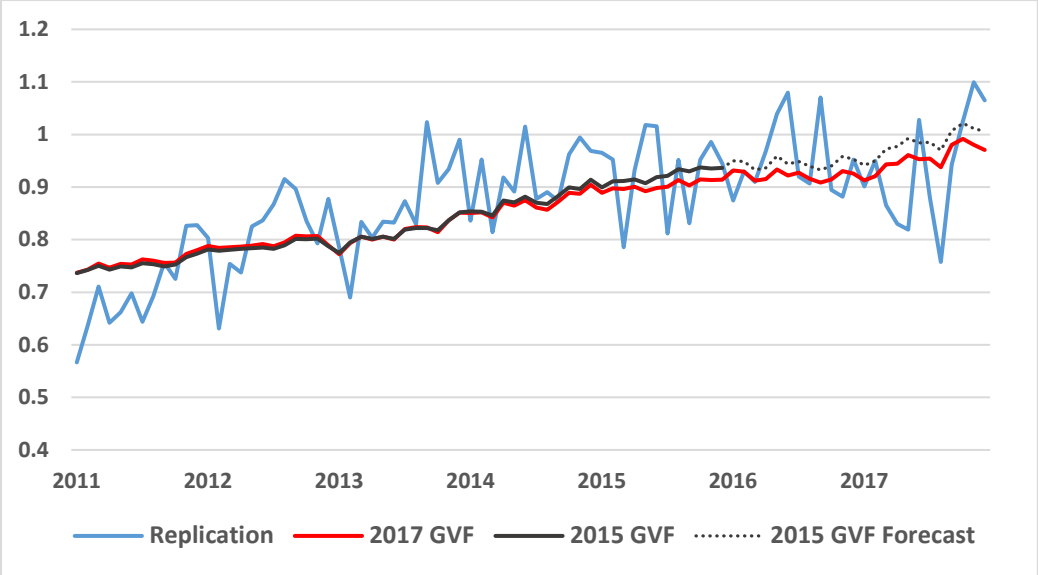

Figure 2: Comparison of 2015 GVF forecast to 2017 "full model" GVF standard errors

Some forecast error is expected when projecting a model forward. To test the quality of the GVF model applied to future dates, the model parameters were refitted using only 2011-2015 data, and those parameters were applied to 2016 and 2017 published, seasonally adjusted estimates. That forecast line appears in Figure 2. The average forecast error, relative to the full 2011-2017 GVF model, is 2.9 percent—2.5 percent in 2016; 3.2 percent in 2017. Intuitively, removing further in time from the modeling period should generally lead to increasing forecast errors, stressing the need for routinely updated model fits. However, in practical terms, the magnitude of the difference in estimated standard errors between the 2015 GVF and the 2017 GVF tends to be small, which suggests at least a reasonable robustness of the model to proximate time periods.

[5] In April 2014, the 2010 CPS sample began its "phase in," while the 2000 decennial CPS sample started being "phased out." This PIPO (phase in/phase out) is a 16-month process that concluded in July 2015. Since longitudinal correlations attenuate during the PIPO process, variances tend to be larger due to the noneconomic effect of changing samples. The GVFs are not explicitly attuned to this variance inflation, although they are implicitly affected during parameter estimation, if the reference period includes the PIPO months, as is true for (2.4).

4. GVF Model for Sample Medians

Under Central Limit Theorem conditions, the asymptotic variance of a sample quantile $q \in (0,1)$ is given by the function

$$V(q) \cong \frac{q(1-q)}{n[f(q)^2]} \tag{3}$$

where $f(q)$ is the density function of the sample quantile q, and n is the sample size.

For a sample median, (3) is simplified by setting $q = 0.5$. Given the unequal weighting of the CPS, the sample size term (n) is replaced by sum of the weights ($\sum_i w_i$) for the conditional subset of unemployed respondents, since the estimated median under consideration is derived only from those respondents. Therefore, $\sum_i w_i$ is the estimate of total unemployed persons ($\hat{Y}$), which is published monthly in *The Employment Situation* news release. Inserting these quantities into (3):

$$V(0.5) \cong \frac{0.5^2}{\hat{Y}[f(0.5)^2]} = \frac{1}{4\hat{Y}} f^{-1}(0.5)^2 \tag{3.1}$$

Let V_r equal the replicate variance of the median estimate, and $SE_r = \sqrt{V_r}$ the replicate standard error. Equating V_r and the approximated formulaic variance, modified by an adjustment ratio d, for the same estimate yields

$$V_r = \frac{1}{4\hat{Y}} f^{-1}(0.5)^2 d$$

$$SE_r = \sqrt{V_r} = \sqrt{\frac{1}{4\hat{Y}} f^{-1}(0.5)^2 d} \tag{3.2}$$

where SE_r is the replicate standard error of the median.

The quantity $\hat{Y}$ in (3.2) is readily available in the published tables. The squared inverse density function and adjustment ratio, however, are difficult to obtain; thus, rearranging (3.2), the square root of their product becomes the dependent quantity (f^*) in the GVF model:

$$f^* = f^{-1}(0.5)\sqrt{d} = 2SE_r\sqrt{\hat{Y}}$$

This formulation allows computation of the d-adjusted density function by replication. The ratio d effectively comprises a complex variance design effect and a scaling parameter (to account for the replacement of n by $\sum_i w_i$) and cannot be disambiguated from the density function in this construction. Given a parsimonious model for f^*, calculating a GVF-based standard error estimate of median weeks unemployed is simply an algebraic extension.

While GVFs are customarily defined as models that approximate variances by evaluation at the survey estimate of interest, in the case of the CPS estimate of median weeks unemployed, it was observed that a variation provided a better fit, particularly when

projecting the GVF beyond the model reference period. Specifically, rather than building a model off the median estimate, the model instead utilizes the estimated total weeks unemployed. This indirect GVF implies, at least among models considered in this research, that the density function is better evaluated by the total rather than by a specific quantile. While total weeks unemployed is not published as a standalone series, it can be obtained as the product of two published series defined previously: average weeks unemployed ($\bar{x}$) and total unemployed persons ($\hat{Y}$).

Applying an OLS model to f^* and substituting into (3.2) results in a standard error estimate possessing the desirable properties of a GVF (as shown in Figures 3 and 4):

$$\hat{f}^* = \hat{\beta}_0 + \hat{\beta}_1(\bar{x}\hat{Y})$$

$$\widehat{SE}(q = 0.5) \cong \sqrt{\frac{1}{4\hat{Y}} f^{-1}(0.5)^2 d} \cong \frac{\hat{f}^*}{2\sqrt{\hat{Y}}} = \frac{0.5\left[\hat{\beta}_0 + \hat{\beta}_1(\bar{x}\hat{Y})\right]}{\sqrt{\hat{Y}}} = \frac{\hat{\beta}_0^* + \hat{\beta}_1^*(\bar{x}\hat{Y})}{\sqrt{\hat{Y}}} \quad (3.3)$$

The final GVF for the median (3.3) is similar in form to the GVF for the mean (2.4). Both models require as inputs the estimated mean weeks unemployed ($\bar{x}$) and total unemployed persons ($\hat{Y}$), which are published monthly. While the beta parameter notation was reused for simplicity, the values of those parameters are not equal, as should be evident from their respective formulations.

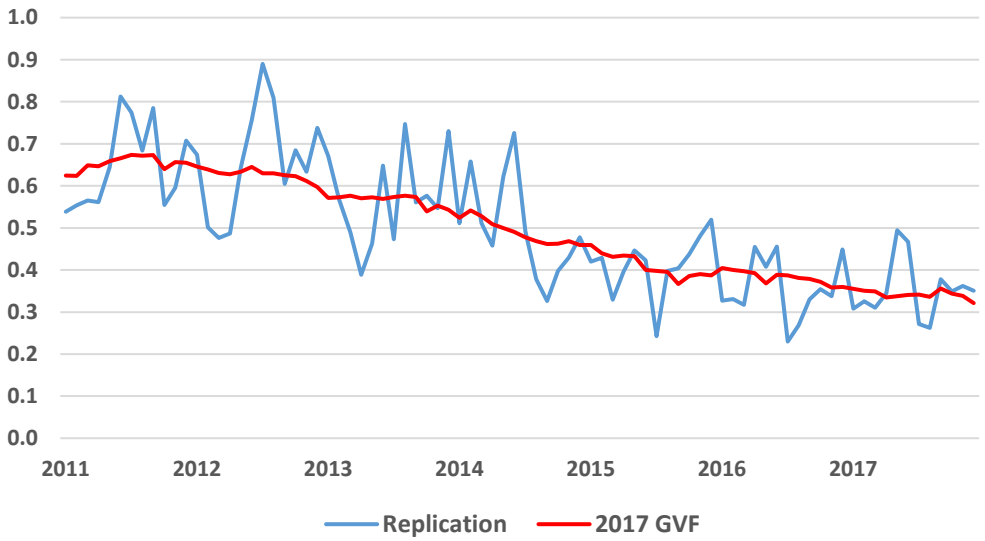

Figure 3: Estimated standard errors of seasonally adjusted median weeks unemployed

Figure 3 displays the results of fitting model (3.3) to the median weeks unemployed series. Analogous to Figures 1 and 2, the blue line is the replicated relative standard error multiplied by the seasonally adjusted median weeks unemployed. This series is compared to the standard error produced from (3.3), using seasonally adjusted estimates of $\bar{x}$ and $\hat{Y}$.

As demonstrated in Figure 3, the GVF model produces standard error estimates that track the replicate-based estimates well over time and are clearly of greater stability. It is not uncommon for the replication-based estimates to vary by a relative magnitude of 25 to 50 percent over the course of several months, which attenuates their utility for evaluating the significance of short-term change.

Following the same forecasting approach as for the mean, the model parameters were refitted using only 2011-2015 data, and those parameters were applied to 2016 and 2017 published, seasonally adjusted estimates. That forecast line appears in Figure 4. The average forecast error, relative to the full 2011-2017 GVF model, is 3.6 percent—3.1 percent in 2016; 4.1 percent in 2017. The conclusions are the same as for the mean: the projections do not vary significantly from the full model estimates, suggesting that model (3.3) parameters remain effective for proximate time periods.

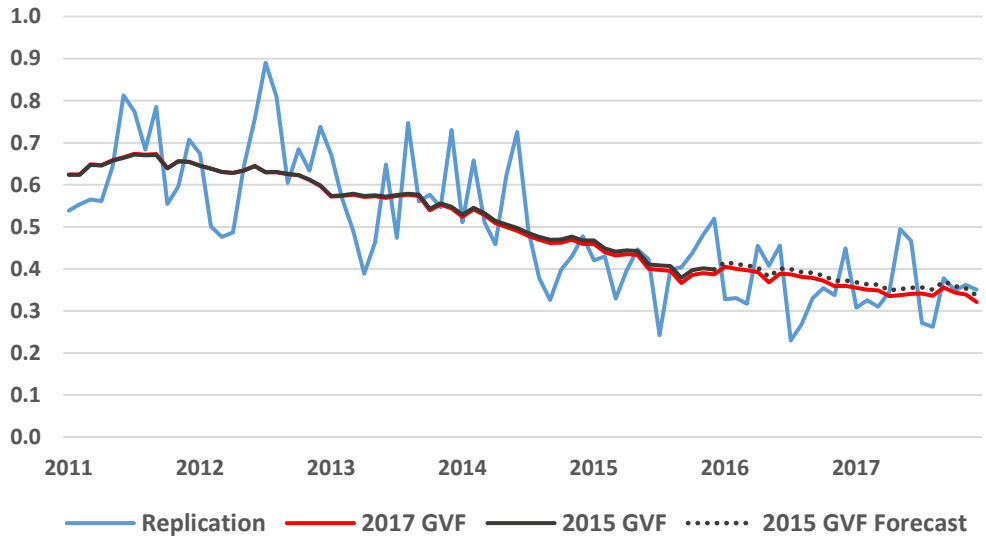

Figure 4: Comparison of 2015 GVF forecast to 2017 "full model" GVF standard errors

As is generally true for variance models, parameters should be updated regularly, based on the periodicity and stability of the observed series. Figures 2 and 4 suggest that annual or biennial updates for models (2.4) and (3.3) are sufficiently frequent.

5. Discussion

The GVF models presented in this paper are, in one sense, specific to the needs of the CPS, as they were derived to fill the standard error gap in *The Employment Situation* news release tables produced from CPS data. However, since GVFs apparently receive greatest attention and application in the Federal Statistical System, the current lacunae in the research related to GVFs for estimating the standard errors of means and medians lends potential usefulness to other official statistics.

It should foremost be repeated that the framework of these GVF models assumes the availability of replicate weights. Various methods to create such weights are well known and presented by Wolter (2007) and others.

In practice, the few GVFs attempted for mean estimates tend to either rely on the classical $ax^2 + bx$ binomial GVF model with an adjustment factor to bridge the gap between a weighted mean variance and a weighted binomial variance; or, more recently, a weighted sample variance s^2 multiplied by a design factor. Improvements seem possible utilizing a formulation that more closely approximates the theoretical variance that the GVF is intended to resemble. Incorporating (2) into the construction of a model for the variance of the mean, given a sufficient time series upon which to fit parameters, could support better estimation of standard errors of complex survey means.

For complex survey medians, there have been few attempts to build GVF models[6]. This is likely due to the inherent challenge in estimating the density function. However, as shown in (3.2), the ratio-adjusted density function can be approximated via replication, utilizing the asymptotic variance of the median. More broadly, though as yet untested in this research, there is nothing obvious to prevent the extension of this model to other quantiles besides q equal to 0.5. While the only quantile estimates the CPS publishes are medians, there may be external need for modeled standard errors of other quartiles, deciles, et cetera.

Returning to the objectives of this research:

As of September 2018, these GVFs are internal research series only, having not yet been fully evaluated for potential publication. However, as methodology for producing standard errors for mean and median weeks unemployed in *The Employment Situation*, the GVF models constructed in this paper demonstrate the necessary reliability across the reference period, and robustness in near-future time periods, to be considered reliable and accurate. Further, they possess the public usability—requiring only published estimates and parameters—to fit within the existing framework of GVFs produced by the Current Population Survey.

References

Valliant, R.L. (1987). "Generalized Variance Functions in Stratified Two-Stage Sampling," in *Journal of the American Statistical Association* Vol. 82, No. 398, pp.499-508.

U.S. Census Bureau (2006). *Design and Methodology, Current Population Survey, Technical Paper 66*. Washington, DC: Author.

McIllece, J.J. (2016). "Calculating Generalized Variance Functions with a Single-Series Model in the Current Population Survey," in *Proceedings of the 2016 Joint Statistical Meetings*, Survey Research Methods Section.

Wolter, K.M. (2007). *Introduction to Variance Estimation* (2nd ed.), New York, NY, Springer.

American Community Survey (2010). *DSSD 2010 American Community Survey Memorandum Series #ACS10-RE 02*. U.S. Washington, DC, U.S. Census Bureau.

[6] The Current Population Survey has long used a different method for approximating standard errors of median earnings estimates. The technique involves estimating a one percent interval around the median and multiplying the range of this interval by a binomial standard error estimate, where p is 50 percent.

National Center for Education Statistics (1995). *Design Effects and Generalized Variance Functions for the 1990-1991 Schools and Staffing Survey (SASS)*. Washington, DC: Author.

The Employment Situation. *https://www.bls.gov/news.release/empsit.toc.htm*. Bureau of Labor Statistics. Regularly revised.

The Current Population Survey technical documentation and reliability statement (2018). *https://www.bls.gov/cps/documentation.htm#reliability*. Bureau of Labor Statistics.

Zimmerman, T.S. and Robison, E.L. (2018). "Current Population Survey State GVFs and Design Effects," in *Proceedings of the 2018 Joint Statistical Meetings*, Survey Research Methods Section.

U.S. Census Bureau (expected 2019). *Design and Methodology, Current Population Survey, Technical Paper 77*. Washington, DC: Author.