Optimal and Coherent Data Visualization in R for the Empirical Study of CPI-U Standard Errors October 2015

Harold Gomes
U.S. Bureau of Labor Statistics
2 Massachusetts Ave, NE, Room 3655 Washington, D.C. 20212

Abstract

This study looks at the density distribution of standard errors (SE) of item stratum-index area level percent change using Consumer Price Index for All Urban Consumers (CPI-U) historical data in order to understand anomalous behavior. Many attributes can be determined about the SE, such as: the shape of the SE density distribution; whether the overall central tendency and overall variability of the SE distribution are smaller, larger, or similar from one year to the next, or from one month to the same month of the following year; whether the distributions tend to shift over time or stay stationary. The SE of basic level price changes are used to produce the underlying distributions for further examination. SE for May 2006 in particular were investigated as it was the month when the SE of 12-month CPI-U percent change reached its largest value of 0.19. Non-parametric methods and categorical analysis techniques were employed to assess the May 2004-May 2008 datasets. A compelling approach to data visualization is produced by combining multiple pieces of information in a single graph in order to demonstrate an optimal and coherent visual of comparisons.

Key Words: Consumer Price Index (CPI), Consumer Price Index for All Urban Consumers (CPI-U), variance, standard error, distribution, data visualization

Any opinions expressed in this paper are those of the author and do not constitute policy of the Bureau of Labor Statistics.

1. Introduction

Consumer Price Index for All Urban Consumers (CPI-U) is a measure of average price change over time in the prices of consumer items—goods and services that people buy for day-to-day living. Here "people" implies about 88% of the U.S. population, referred to as urban population. CPI-U provides an estimate of the price change between any two periods, such as, 1 month, 2 month, 6 month, or 12 month intervals. This price change is then presented in percent (%) format to indicate the degree of change between periods.

This percent change also has a corresponding standard error (SE) which can be used to determine whether this change is statistically significant or not. Standard errors (SE) are produced for all the *basic indexes* and the *aggregate indexes*, although *basic index* SE are not published for the public. There are 38 geographic areas called *index areas*, and the set of all goods and services purchased by consumers is divided into 211 categories called *item strata.* This results in 8,018 (38 x 211) item-area combinations that are *basic indexes*. The CPI is calculated in two stages. The first stage is the calculation of *basic indexes*, which shows the average price change of the items *within* each of the 8,018 CPI item-area combinations. In the second stage*, aggregate indexes* are produced by the *weighted* average across subsets of the 8,018 CPI item-area combinations. The aggregate indexes are the higher-level indexes and different types of aggregates can be produced based on usefulness, such as, 38 geographic areas, 4 census regions, 211 items, 8 major groups, etc. The *weights* for the second stage do not derive from the CPI survey but from the reported expenditures of the Consumer Expenditure Survey (CE) (Chapter 17, *The Consumer Price Index*).

This study investigates the probability density distribution (PDF) of standard errors (SE) of all 8,018 basic indexes in order to understand the anomaly in the behavior of the percent change standard errors for a particular area or item. This study also aims to examine the historical behavior of these basic index SE. Many properties can be examined about the standard errors from this underlying density distribution, such as: the shape of the SE density distribution; whether the overall central tendency and the overall variability of the SE distribution are smaller, larger, or similar from one year to the next, or from one month to the same month of the following year, and whether the distributions tend to shift over time or stay stationary.

In support of the Bureau of Labor Statistics (BLS) mission, the CPI program produces accurate and precise estimates within given constraints. One CPI program performance goal is to produce “All U.S. – All Items” 12-month percent change standard error estimates less than 0.25. Thus, this performance goal provides another motivation for this research study, which is to look at the properties of the basic level SE density distribution, anomalous behavior of SE in particular item-area level, and possible probabilistic quality control.

2. Research Frame-Work and Datasets

Twelve-month CPI-U percent change “All U.S. – All Items” aggregate SE reached its largest value of 0.19 in May 2006. It was due to a single item-area index that attributed a high proportion of variance in aggregation compare to other 8,017. In this study, historic datasets of May 2004, May 2005, May 2006, May 2007, and May 2008 are used as comparison groups that contain *basic index* level information. Choosing the same month of a different year as comparison groups reduces the source of variability due to seasonality (blocking). Additionally, these pre-post comparison groups may provide a perspective on the behavior before and after that May 2006 event.

3. Probability Density Distribution of Basic Index Variances

CPI is a complex construct that uses the following Stratified Random Group (SRG) methodology to calculate the variance for its 12-month percent change (PC) for “All U.S. – All Items”:

$$\boldsymbol{V(A,I,t,t-12)} = \sum_{\boldsymbol{a\varepsilon A}} \frac{\boldsymbol{1}}{\boldsymbol{N_a(N_a-1)}} \sum_{\boldsymbol{r\varepsilon R_a}} \boldsymbol{(PC\,[(A,I,00)-(a,I,00)+(a,I,r),t,t-12]-PC(A,I,t,t-12))^2} \quad \text{..... (1)}$$

The standard error of the 12-month price change is then equal to:

$$\boldsymbol{SE(A,I,t,t-12)} = \sqrt{\boldsymbol{V(A,I,t,t-12)}} \quad \text{or} \quad \boldsymbol{\sigma_{All\ US,All\ Item}} = \sqrt{\boldsymbol{\sigma^2_{All\ US,All\ Item}}} \quad \text{.....(2)}$$

Where:

- $r \in R_a$ refers to the set of replicates in AREA = a,
- $a \in A$ refers to the 32 self-representing and 6 non-self-representing index areas in A,

- I = All Items, and
- N_a is the number of variance replicates in AREA = a.

As an approximation for computational purpose, this "All U.S. – All Items" variance can be decomposed into its 8,018 basic index components.

$$\begin{aligned}\boldsymbol{\sigma^2_{All\ US,All\ Item}} &\approx \boldsymbol{\sigma^2_{wgt.1}} + \boldsymbol{\sigma^2_{wgt.2}} + \cdots + \boldsymbol{\sigma^2_{wgt.8,018}} \\ &= (\boldsymbol{ri_1} * \boldsymbol{\sigma_1})^2 + (\boldsymbol{ri_2} * \boldsymbol{\sigma_2})^2 + \cdots + (\boldsymbol{ri_{8,018}} * \boldsymbol{\sigma_{8,018}})^2\end{aligned}$$

Where:

- σ^2_1 is the variance for 1 of the 8,018 item-area indexes
- $\sigma^2_{wgt.1}$ is the *weighted* variance component for 1 of the 8,018 item-area indexes
- ri_1 is the corresponding expenditure *weight* for 1 of the 8,018 item-area indexes known as *relative importance*. *Weights* are derived from the reported expenditures of the Consumer Expenditure Survey (CE).

There are four selection choices for probability density distribution, data visualization and further assessment. They are:

1) decomposed variances (σ^2),
2) decomposed standard errors (σ),
3) [1]decomposed *weighted* variances (σ^2_{wgt}), or
4) decomposed *weighted* standard errors (σ_{wgt}).

For this study, decomposed standard error (σ) and decomposed *weighted* variances (σ^2_{wgt}) are used for further examination.

Since the first stage of CPI design computes standard error (σ) for all the basic indexes (area-item) before incorporating the *weights* from Consumer Expenditure survey, standard error (σ) distributions are examined. On the other hand, *weighted* variance (σ^2_{wgt}) distribution is also examined because "All U.S. – All Items" variance ($\sigma^2_{\text{All US, All item}}$) is computed by summing across all the *weighted* variances (σ^2_{wgt}) in the second stage.

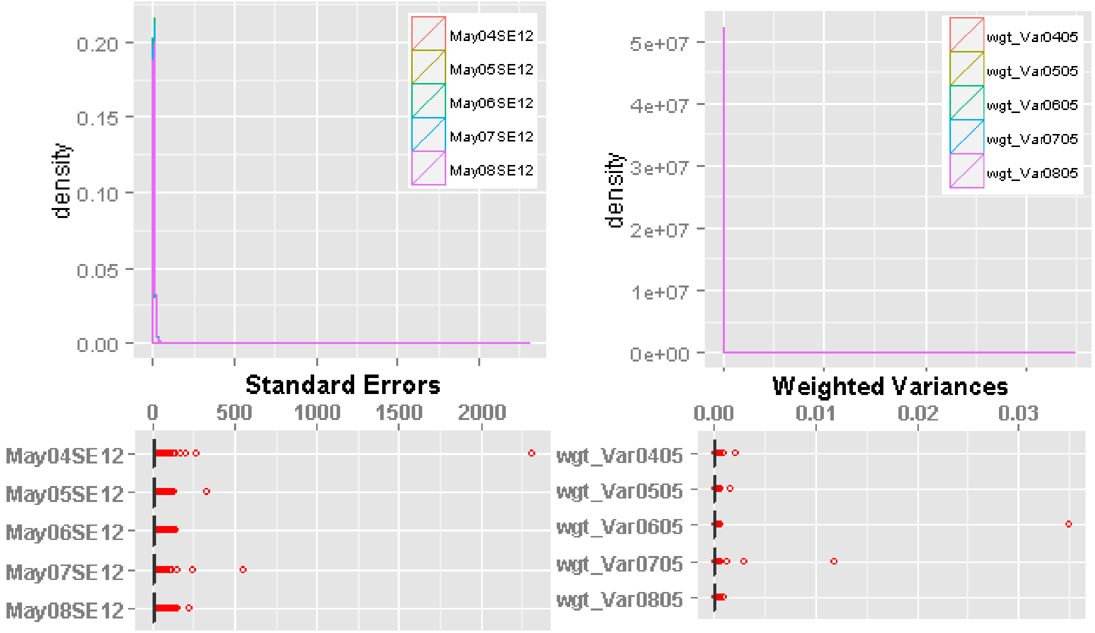
The following Table 1 shows the 12 month "All U.S. – All Items" variance ($\sigma^2_{\text{All US, All item}}$) from re-calculation based on decomposed weighted variances, originally published using formula (1), and corresponding rankings among groups.

Table 1	**Variance**	**SE**		**SE**		
	$\sigma^2_{\text{All US, All item}}$	**$\sigma_{\text{All US, All item}}$**	**Ranking**	**$\sigma_{\text{All US, All item}}$**	**Ranking**	**difference**
	calculated from	**calculated from**	**high to low**		**high to low**	**original SE -**
	approximation	**approximation**	**based on**	**as originally**	**as**	**approximated**
Month-Year	**formula**	**formula**	**approximation**	**published**	**published**	**SE**
May 2004	**0.01256**	**0.11208**	**3**	**0.13137**	**3**	**0.01929**
May 2005	**0.00856**	**0.09252**	**4**	**0.09601**	**5**	**0.00348**
May 2006	**0.04214**	**0.20529**	**1**	**0.19046**	**1**	**-0.01483**
May 2007	**0.02299**	**0.15162**	**2**	**0.13394**	**2**	**-0.01768**
May 2008	**0.00732**	**0.08557**	**5**	**0.10304**	**4**	**0.01747**
Performance Goal	**< 0.0625**	**< 0.25**	**--**	**< 0.25**	**--**	**--**

The following Figures (1, 2 and 3) display the probability density distribution (using kernel density estimation method), boxplots, and cumulative probability distribution of decomposed standard errors (σ) and decomposed *weighted* variances (σ^2_{wgt}) of 8,018 basic indexes.

[1] decomposed *weighted* variances (σ^2_{wgt}) indicates decomposed *weighted* variance **components** throughout this document.

Figure 1: **Probability density distribution (kernel density) and boxplots of SE and of decomposed weighted variances**

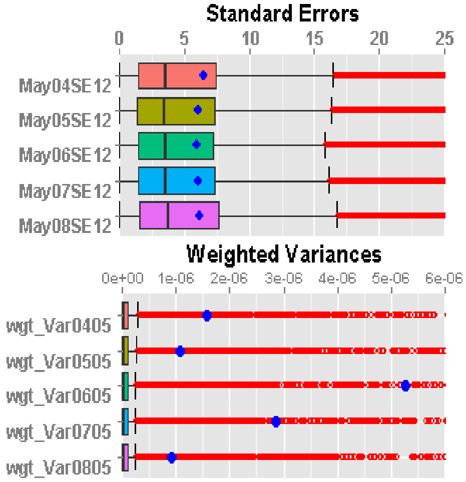


The *probability density distribution (kernel density)* and *boxplots* of SE and of *weighted* variances are superimposed for all 5 comparison groups (May 2004-May 2008), and are displayed in pairs for intuitive comparison. "May04SE12" indicates May 2004 SE of 12-month, and "wgt_Var0405" indicates weighted variances of 2004 of May. The red dots display a potential outlier.

3.1 Exploratory Analysis Summary

The superimposed probability density distributions (non-parametric, kernel density estimation) of the decomposed SE (σ) and the decomposed *weighted* variances (σ^2_{wgt}) display a positively skewed distribution (Figure 1 & 3), i.e., a large proportion of data are condensed on the lower side of the distribution across all groups (May 2004-May 2008). Potential outliers are also observed (red dots) in both distributions by assessing boxplots (Figure 1). An important feature is: once the standard errors are weighted based on their relative importance (expenditure weight from CE survey), different outliers emerge in the *weighted* variance distribution (σ^2_{wgt}) than SE distribution (σ). For example, an extreme outlier is detected in the May 2006 *weighted* variances distribution but not in the SE distribution. Similarly, the extreme outlier in May 2004 SE distribution is not observed as extreme once *weighted*.

Figure 2: **Magnified boxplots with distribution mean (blue dots)**

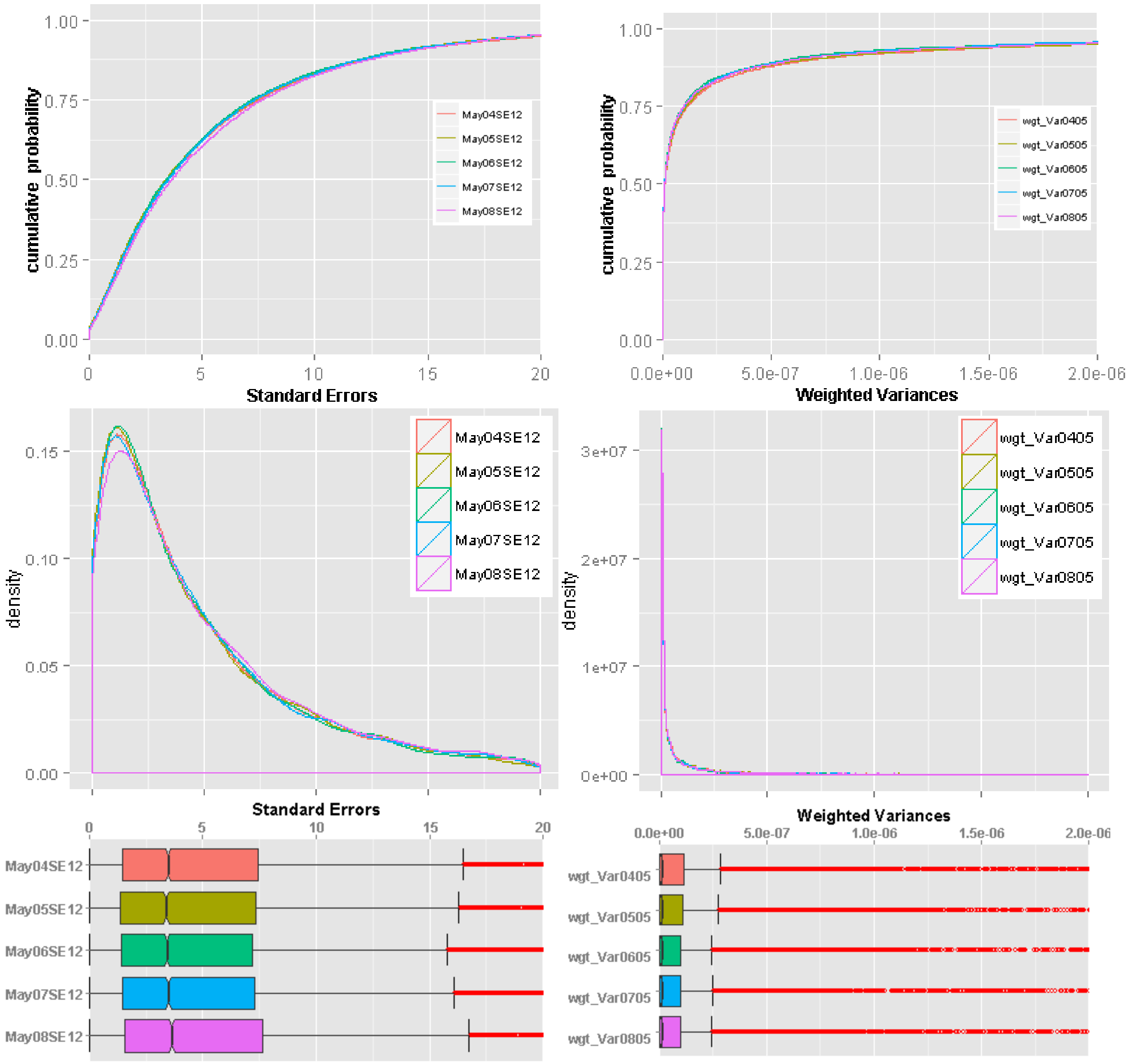


To assess the properties of median and mean (average) of each distribution, boxplots are magnified (Figure 2). Distribution means are inside the box for SE; however, they are pulled out of the box and whisker in the *weighted* variance distribution. May 2006 mean moved the highest among the groups due to an extreme outlier presence (figures 2 & 1).

To assess the data properties and distributions more intuitively, magnified plots are produced for probability density distribution (kernel density), cumulative probability distribution (see section 5 for details about *ecdf*) and boxplots for SE and *weighted* variances. It is observed that the distribution of *weighted* variances is highly skewed than the SE distribution for a similar proportion of data.

Figure 3: Magnified plots of figure 1.

Cumulative probability distribution (ecdf), probability density distribution (pdf; kernel density), and boxplots of SE and of decomposed weighted variances for May 2004-May 2008



These magnified figures display about ~95% data for SE (0-20 range) and *weighted* variances (0-2x10^{-6} range). Identical proportion of data for all these graphs provide a frame of reference for intuitive comparisons. It enables us to observe the shape of the density and shape of the cumulative distribution curve. They all graphically confirm that the SE and *weighted* variance display a positively skewed distribution, and that the *weighted* variance distribution is much skewer and highly dense in lower end than the SE distribution. Boxplots and ecdf aid us to assess the quantiles and the skewness properties. The boxplot with notches is produced to display a normally approximated 95% confidence interval around the median (Chambers, J. et al, 1983).

4. Non-Parametric Inference for Density Distributions

To compare the superimposed distributions of standard errors (σ) and of decomposed *weighted* variances (σ^2_{wgt}), a more robust and distribution free, non-parametric statistical method is employed. The rank based Kruskal-Wallis test (KW) is implemented to compare more than two groups as it can be applied to a dataset without ties or adjustment for ties. Since the distributions are skewed (not normal) as section 3.1 confirms, KW test generally provides a greater power to detect differences among groups than one way ANOVA *F*-test (Higgins, J., 2004). KW test is essentially a one-way ANOVA on ranks and does not require normality to be a prerequisite. However, it does assume that the groups have the same distributional form or shape (Higgins, J., 2004; Aho, K., 2014). Section 3.1 also confirms that the groups have similar shape. A reference distribution for KW test statistic can be generated using either a permutation-testing procedure from randomization of observed data or chi-square approximation for a large sample size. If significant difference is detected by rejecting the null, multiple pairwise comparisons can be conducted while accounting for the familywise error rate to find contrast between pairs. A conservative multiple comparison method, Bonferroni procedure, is used here (Aho, K., 2014; Higgins, J., 2004, Hothorn, T., 2006).

Initially, the skewed data was attempted to be transformed in order to produce normality, so that the parametric methods could be applied to transformed data. Log, square root, and cube root transformations were produced. However, the transformed data did not meet the criteria of a normal distribution as confirmed by an example Q-Q (Figure 4). So this was another motivation for implementing the non-parametric procedure.

The Kruskal-Wallis test of *k* populations with the following *cumulative distribution function* (cdf) is:

$$H_0: \quad F_{2004}(x) = F_{2005}(x) = F_{2006}(x) = F_{2007}(x) = F_{2008}(x)$$
$$H_a: \quad F_i(x) < F_j(x) \text{ or } F_i(x) > F_j(x)$$

for at least one pair of (i, j), where $\{(i, j)$: (2004, 2005, 2006, 2007, 2008)$\}$

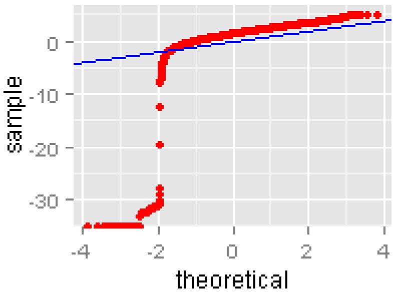
Table 3: Kruskal-Wallis Test Results P-values

	standard errors (σ)	*weighted* variances (σ^2_{wgt})
Kruskal-Wallis Test *(asymptotic Chi-Square Appr)*	**0.003037**	**0.1918**
Kruskal-Wallis Test *(permutation test 1,000,000)*	**0.00311**	**0.1917**

Multiple Pairwise Comparison with Bonferroni adj maintaining overall α < 0.05

Significant contrast between SE 2005 & SE 2008
Significant contrast between SE 2006 & SE 2008

Figure 4: QQ Plot of Log (SE $_{\text{May 2006}}$), A data transformation example that displays non-normality of transformed data



4.1 Non-Parametric Result Summary

The Kruskal-Wallis Tests results indicate that there is no significant differences (*p*-value = 0.1918) in *weighted* variance (σ^2_{wgt}) distribution among groups. However, there is a significant difference in SE distribution among groups. Two out of ten pairwise comparisons show significant contrast—between SE 2005 & SE 2008 and between SE 2006 & SE 2008 distributions (table 3). The Bonferroni procedure maintaining overall $\alpha < 0.05$ is implemented in the multiple pairwise comparison to account for familywise error rate. It is important to note that the large sample size (*n*=8,018) contributed to a significant result even for a small shift (small effect size) in a distribution (as observe from figure 3).

Intuitively it makes sense. Based on the exploratory analysis (section 3.1, figure 3), we observe that the *weighted* variance (σ^2_{wgt}) distribution is more skewed than SE distribution. More skewed implies, higher proportion of data points are condensed within smaller region. As a result, data points from this densely small region do not show any differences across groups in significance testing. On the other hand, SE are spread across larger region compare to *weighted* variance distribution (figure 3). Thus, a shift in distribution is easily detectible even if it is small.

In summary, it basically informs us that the standard error distributions of 8,018 indexes may be different initially; however, once standard errors are *weighted* based on the consumer expenditure weight (buying habits of American consumers), the *weighted* variance distributions look very similar across time. It also indicates, *weighted* variance distributions show stability (no shift) across time, while SE distributions may or may not. Please note that this comparison is observed only among five time points of the same month (May 2004 - May 2008) and not across all time points; hence, extrapolation should not be drawn without further investigation across more time points.

Based on this result, if *weighted* variance distributions do not show a shift across the comparison groups (no significant difference), then what drives the increase in the final aggregate “All U.S. – All Items” variance?

Exploratory analysis (section 3.1) provides a preliminary assessment, i.e, it’s the extreme outlier that may contribute a high proportion of variance in the final aggregation. To further investigate this idea, a categorical analysis is deployed in Section 6.

5. Detection Mechanism for Anomalous Behavior using Probabilistic Quality Control

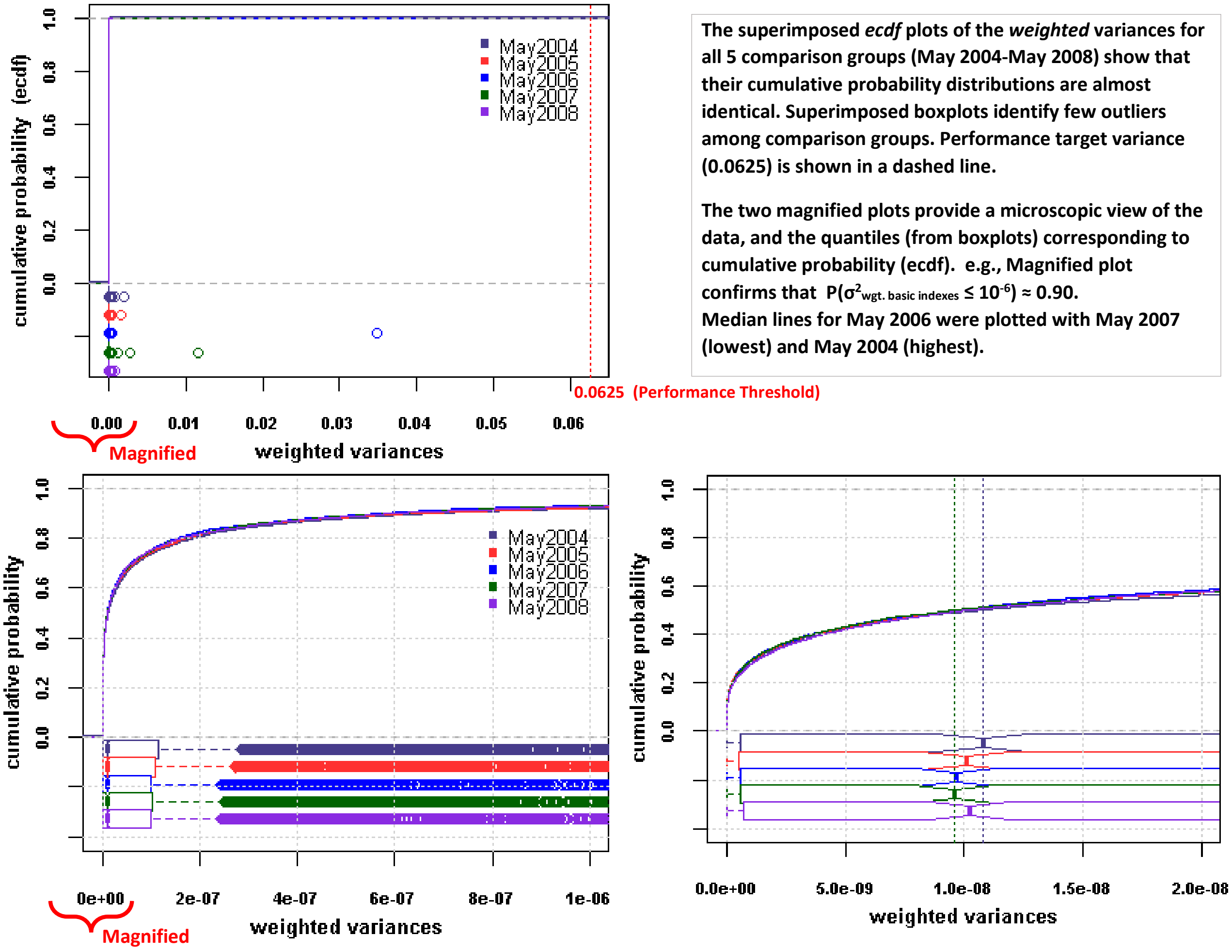
Based on the above assessment on the *weighted*-variance distributions, it is observed that only a handful of basic indexes out of 8,018 may contribute most of the "All U.S. – All Items" aggregate variance. Understanding these area/items allow us to see if there are underlying data issues that may be addressable through the sample design. As other federal agencies with their mission and goals, CPI program is also dedicated to meet its performance goal. It is to ensure that the "All U.S. – All Items" aggregate SE ($\boldsymbol{\sigma}_{\textbf{All US, All item}}$) is less than 0.25 i.e., the Variance ($\boldsymbol{\sigma}^2_{\textbf{All US, All item}}$) is less than 0.0625.

The *ecdf*, often known as the cumulative probability density curve, is a very robust tool for practical application in the field of probabilistic quality control in a production setting. For a given sample of $(x_1, x_2, \dots x_n)$, $\hat{F}_n(Var)$ is the fraction of observations less than or equal to the Var value, i.e., the ecdf is defined by:

$$\hat{F}_n(Var) = \frac{number\ of\ elements\ in\ the\ sample \le Var\ value}{n} = \frac{1}{n}\sum_{i=1}^{n}\{x_i \le Var\}$$

This *ecdf* uses a *step function* method that jumps up by *1/n* at each of the *n* data points to estimate the curve for the cumulative probability of *1*. The following plot (Figure 5) displays the cumulative probability distribution of *weighted*-Variances.

Figure 5: **Empirical Cumulative Probability Distribution (ecdf) of Weighted Variances with Boxplots**



The superimposed *ecdf* plots of the *weighted* variances for all 5 comparison groups (May 2004-May 2008) show that their cumulative probability distributions are almost identical. Superimposed boxplots identify few outliers among comparison groups. Performance target variance (0.0625) is shown in a dashed line.

The two magnified plots provide a microscopic view of the data, and the quantiles (from boxplots) corresponding to cumulative probability (ecdf). e.g., Magnified plot confirms that P($\sigma^2_{\text{wgt. basic indexes}} \le 10^{-6}$) ≈ 0.90. Median lines for May 2006 were plotted with May 2007 (lowest) and May 2004 (highest).

5.1 Result Summary and Conclusions

The superimposed *ecdf* of *weighted*-Variances for all 5 years indicate that their cumulative probability distributions are identical and cover similar density region for item-area indexes. Graphically, it confirms that about 90% of basic level indexes have less than or equal to 10^{-6} *weighted*-Variance for any of the 5 years. In other words, there is a 90% probability that a randomly picked *weighted*-Variances among any 5 years may have a less than or equal to 10^{-6} *weighted*-Variances,
i.e., $P\ (\sigma^2_{\textbf{wgt. basic indexes}} \leq 10^{-6}) \approx 0.90$.

Based on the historic data, if there is indeed such a high probability for a basic level *weighted*-Variance to be so small, what drives the increase in the aggregate variance? Superimposed boxplots on the same data visualization with the superimposed *ecdf* plots may provide guidance for interpretation. Outliers are observed looking at the boxplots, such as, May 2006 shows about 0.035 *weighted*-Variances for a data point (basic index); May 2007 shows about 0.0115 *weighted*-Variances for a basic index. "All U.S. – All Items" aggregate Variance ($\boldsymbol{\sigma}^{\mathbf{2}}_{\textbf{All US, All item}}$) for May 2006 is 0.04214 and for May 2007 is 0.02299 (Table 1). Thus, this single index in May 2006 is attributable to ~ 83% of aggregate Variance (~ 0.035 / 0.04214), and the single index in May 2007 is attributable to ~ 50% of aggregate variance (~ 0.0115 / 0. 02299).

Based on the analysis and data visualization, we learn that although the probability is very low for a basic index to display anomalous behavior (i.e., handful indexes out of 8,018), the impact may be very high if this takes place due to its overwhelming contribution to the "All U.S. – All Items" aggregate variance.

Superimposed data visualizations of *ecdf* and boxplots provide a *coherent* and meaningful representation of large datasets for an intuitive understanding and interpretation. Additionally, boxplots guide us to assess the quantiles with the corresponding cumulative probability on the *ecdf* plot. Magnification of the axis provides a microscopic view of data points for assessment. A useful application of the superimposed boxplots and *ecdf* plot is to monitor the basic level indexes for anomalous behavior by comparing contrasting with previous or relevant months, or from a hypothetical target.

6. Odds, Relative Risk, and Odds Ratio: Different Likelihoods and Probabilities of Events

Based on the above analysis, we observe that few *weighted*-variances out of 8,018 indexes act as the top players to increase the "All U.S. – All Items" aggregate variance. This observation creates another premise to look at this data from a count-data perspective. It enables one to discover additional properties, such as, how many or what proportion of indexes contribute the most variation in percent change of prices in U.S.; or a range of common likelihoods in general.

To investigate these questions, categorical data analysis was used. First, contingency tables are produced as follows:

Table 3a

Number of Indexes Contributed in Decomposed Variance (All U.S. -All Items)

Mo-YY	# of Index in Higher 80%	# of Index in Lower 20%	Total
May 2004	88	7930	8,018
May 2005	162	7856	8,018
May 2006	1	8017	8,018
May 2007	15	8003	8,018
May 2008	181	7837	8,018

Table 3b

Proportion of Indexes Contributed in Decomposed Variance (All U.S. -All Items)

Mo-YY	Prop of Index in Higher 80%	Prop of Index in Lower 20%	Total	Odds of Index in Higher 80%
May 2004	0.01098	0.98902	1	0.01110
May 2005	0.02020	0.97980	1	0.02062
May 2006	0.00012	0.99988	1	0.00012
May 2007	0.00187	0.99813	1	0.00187
May 2008	0.02257	0.97743	1	0.02310

Table 4a

Number of Indexes Contributed in Decomposed Variance (All U.S. -All Items)

Mo-YY	# of Index in Higher 99%	# of Index in Lower 1%	Total
May 2004	1785	6233	8,018
May 2005	2039	5979	8,018
May 2006	734	7284	8,018
May 2007	1188	6830	8,018
May 2008	2129	5889	8,018

Table 4b

Proportion of Indexes Contributed in Decomposed Variance (All U.S. -All Items)

Mo-YY	Prop of Index in Higher 99%	Prop of Index in Lower 1%	Total	Odds of Index in Higher 99%
May 2004	0.22262	0.77738	1	0.28638
May 2005	0.25430	0.74570	1	0.34103
May 2006	0.09154	0.90846	1	0.10077
May 2007	0.14817	0.85183	1	0.17394
May 2008	0.26553	0.73447	1	0.36152

To produce the contingency tables, the approximate decomposition equation of "All U.S. – All Items" variance into its 8,018 basic index components is used.

$$\sigma^2_{All\ US, All\ Item} \approx \sigma^2_{wgt.1} + \sigma^2_{wgt.2} + \cdots + \sigma^2_{wgt.8,018}$$
$$= (ri_1 * \sigma_1)^2 + (ri_2 * \sigma_2)^2 + \cdots + (ri_{8,018} * \sigma_{8,018})^2$$

Once the aggregate variance is decomposed into its mutually exclusive components followed by ranking of the *weighted*-variances, binomial count data is produced using a few meaningful benchmarks. In this context, the benchmarks are chosen as 80-20 and 99-1 to meet the goal of the question of interest.

Table 3 shows the number of ranked indexes that are attributable to 80% of the "All U.S. – All Items" variance, while remaining indexes are only attributable to 20% of the variance. Similarly, Table 4 shows the number of ranked indexes that are attributable to 99% of the variance, while remaining indexes are only attributable to 1% of the variance. These tables also show the associated proportion tables (probabilities) and corresponding odds.

Odds ratio, relative risk and difference in proportion can be employed for statistical inference on these contingency tables (Aho, K. 2014). In practice, when proportions are small, it is useful to compare groups with odds ratio or relative risk.
In probability, the odds of an event A is defined as:

$$\boldsymbol{\Omega(A) = \frac{P(A)}{1 - P(A)} = \frac{\pi}{1 - \pi}}$$

If P(A) < 0.5, then Ω(A) < 1, and as P(A) increase from 0.5 to 1, then Ω(A) increase from 1 to ∞. A suitable interpretation is as follows: the event A is Ω(A) times more likely to occur than to not occur.

The ratio of odds for two outcomes is their odds ratio and is defined as:

$$\boldsymbol{\theta_{1,2} = \frac{\Omega_1}{\Omega_2}}$$

If the odds ratio is greater than 1, this indicates that the odds of success for one event (Ω_1) are greater than the odds of success for the other event (Ω_2). For empirical estimation from data, the estimator for $\theta_{A,B}$ is

$$\hat{\theta}_{1,2} = \frac{\hat{\pi}_1/(1-\hat{\pi}_1)}{\hat{\pi}_2/(1-\hat{\pi}_2)} = \frac{(y_1/n_1)/(1-\,y_1/n_1)}{(y_2/n_2)/(1-\,y_2/n_2)}$$

Similar to odds ratio, relative risk is another measure and is defined as the ratio of two probabilities. Relative risk is used to contrast the relative probabilities for “success” between different groups or treatments.

$$\widehat{RR}_{1,2} = \frac{\hat{\pi}_1}{\hat{\pi}_2} = \frac{(y_1/n_1)}{(y_2/n_2)}$$

Table 5 Month-Year	Rank	Variance $\sigma^2_{\text{All U.S., All item}}$ calculated from approximation formula	SE $\sigma_{\text{All U.S., All item}}$ calculated from approximation formula	Odds 80-20 benchmark	Odds 99-1 benchmark
May 2006	1	0.04214	0.20529	0.00012	0.10077
May 2007	2	0.02299	0.15162	0.00187	0.17394
May 2004	3	0.01256	0.11208	0.01110	0.28638
May 2005	4	0.00856	0.09252	0.02062	0.34103
May 2008	5	0.00732	0.08557	0.02310	0.36152
Performance Goal		< 0.0625	< 0.250	..	..

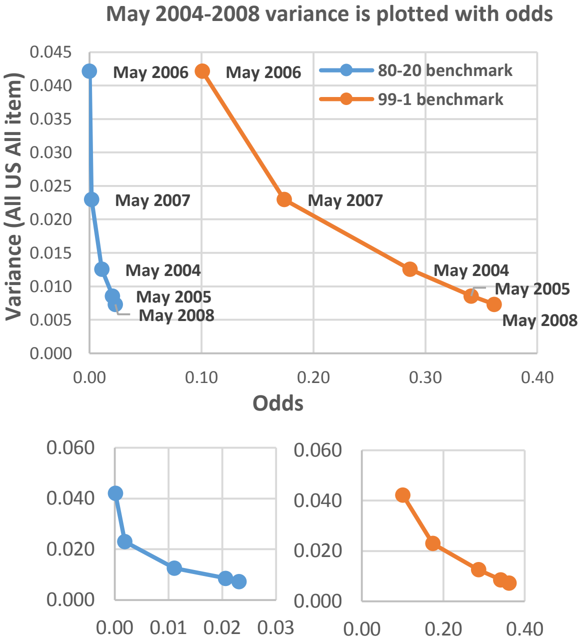
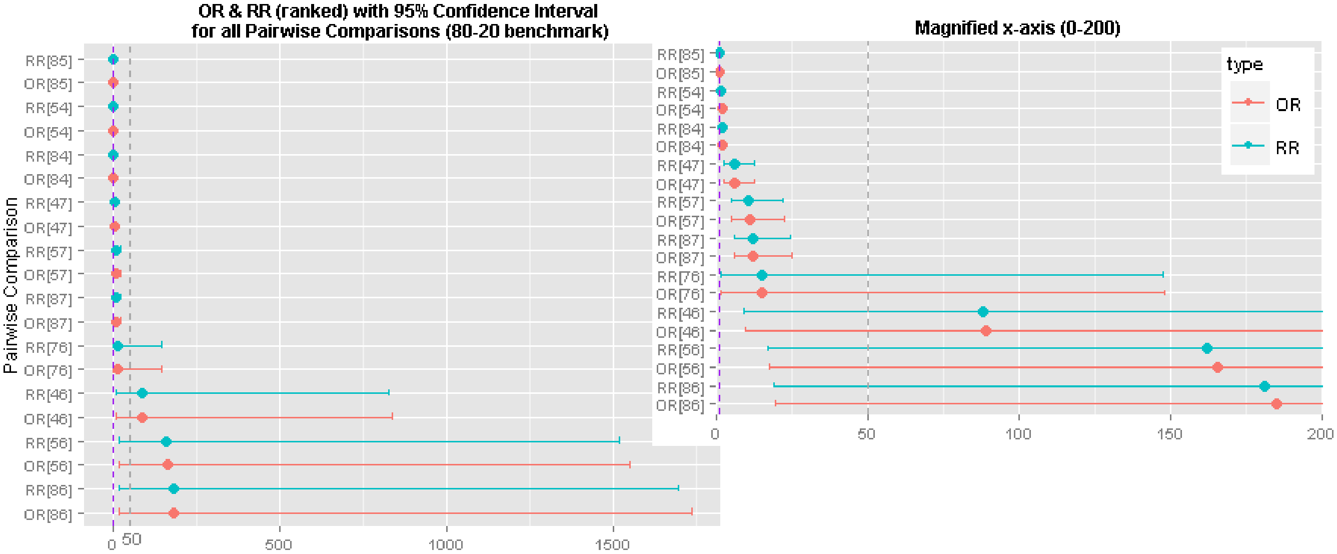


Figure 6: The odds display an inverse functional form with the “All US-All Items” variance for May 2004-2008 for both benchmarks (80-20 and 99-1). Each benchmark was separately plotted due to different scales in odds, and also superimposed in a single scale plot.

Multiple pairwise comparison between five groups (May 2004 - May 2008) can be conducted using a recently published methodology (Agresti, Alan et all. 2008) for odds ratio, relative risk and difference in proportion. A useful feature of this method is to construct a simultaneous confidence interval for statistical significance instead of ad hoc approaches, such as Bonferroni correction, to account for familywise error rate (FWER). Once the confidence bounds of odds ratios are obtained (Agresti, Alan et all. 2008), corresponding conversions are computed for relative risk confidence bound using Schmidt and Kohlmann (2008); Zhang and Yu (1998); and Holland (1989). The following tables (6 and 7) and the graph summarize the pairwise odds ratio, relative risk, difference in proportion and the corresponding 95% simultaneous confidence intervals for both benchmarks (80-20 and 99-1).

Table 6: Pairwise Comparisons Between All Groups with corresponding 95% simultaneous confidence intervals for 80%-20% Benchmark on ranked order (based on OR)

Comparison Groups	Lower Bound (OR)	Odd Ratio (OR)	Upper Bound (OR)	Lower Bound (RR)	Relative Risk (RR)	Upper Bound (RR)	Lower Bound (DiffP)	Difference in Proportion (DiffP)	Upper Bound (DiffP)
2008-2005	0.832	1.120	1.508	0.830	1.117	1.504	-0.00350	0.00237	0.00850
2005-2004	1.294	1.858	2.669	1.282	1.841	2.644	0.00450	0.00923	0.01449
2008-2004	1.458	2.081	2.972	1.440	2.057	2.937	0.00650	0.01160	0.01649
2004-2007	2.808	5.921	12.484	2.782	5.867	12.370	0.00650	0.00910	0.01249
2005-2007	5.345	11.002	22.648	5.246	10.800	22.232	0.01449	0.01833	0.02249
2008-2007	6.004	12.322	25.290	5.879	12.067	24.765	0.01649	0.02070	0.02549
2007-2006	1.526	15.026	147.823	1.523	15.000	147.564	0.00150	0.00175	0.00350
2004-2006	9.441	88.965	838.391	9.338	88.000	829.293	0.00850	0.01085	0.01449
2005-2006	17.613	165.320	1551.723	17.259	162.000	1520.561	0.01649	0.02008	0.02449
2008-2006	19.727	185.157	1737.919	19.284	181.000	1698.899	0.01849	0.02245	0.02649



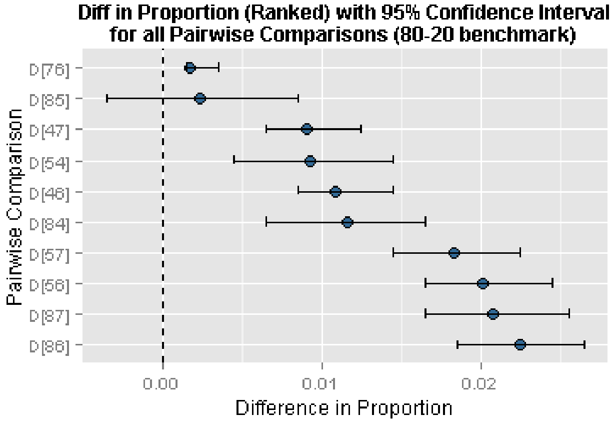
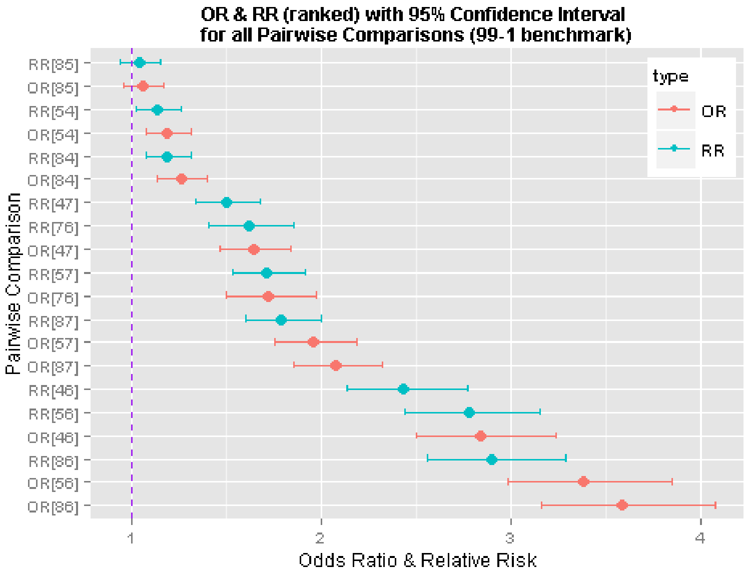


Figure 7: 80-20 benchmark. The OR and RR for pairwise comparison are plotted with their corresponding 95% CI based on ranked order (small to large). OR[85] implies odds(2008)/odds(2005). Ranking provides additional feature for interpretation, i.e., which pair is smaller or larger compare to other. Magnified plot shows a microscopic view of OR and distance between pairs. The dashed line at x=1 is to assess the statistical significance for ratio.

The *Difference in Proportion* plot shows pairwise differences with CI in ranked order. D[76] implies odds(2007) — odds(2006). The dashed line at *x*=0 is to assess the statistical significance for difference.

Table 7: Pairwise Comparisons Between All Groups with corresponding 95% simultaneous confidence intervals for 99%-1% Benchmark on ranked order (based on OR)

Comparison Groups	Lower Bound (OR)	Odd Ratio (OR)	Upper Bound (OR)	Lower Bound (RR)	Relative Risk (RR)	Upper Bound (RR)	Lower Bound (DiffP)	Difference in Proportion (DiffP)	Upper Bound (DiffP)
2008-2005	0.961	1.060	1.170	0.946	1.044	1.152	-0.00750	0.01122	0.02949
2005-2004	1.076	1.191	1.318	1.032	1.142	1.264	0.01349	0.03168	0.04948
2008-2004	1.141	1.262	1.397	1.078	1.193	1.320	0.02449	0.04290	0.06047
2004-2007	1.470	1.646	1.844	1.342	1.503	1.683	0.05847	0.07446	0.09045
2007-2006	1.506	1.726	1.978	1.412	1.619	1.855	0.04348	0.05662	0.07046
2005-2007	1.754	1.961	2.191	1.536	1.716	1.918	0.08946	0.10614	0.12244
2008-2007	1.861	2.078	2.321	1.605	1.792	2.001	0.10045	0.11736	0.13443
2004-2006	2.497	2.842	3.234	2.137	2.432	2.768	0.11644	0.13108	0.14643
2005-2006	2.980	3.384	3.844	2.446	2.778	3.155	0.14743	0.16276	0.17841
2008-2006	3.162	3.588	4.071	2.556	2.901	3.291	0.15842	0.17398	0.18941



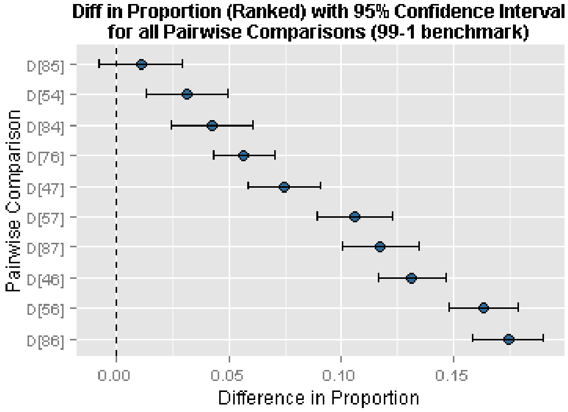


Figure 8: 99-1 benchmark. **The OR and RR for pairwise comparison are plotted with their corresponding CI based in ranked order (small to large). OR[85] implies odds(2008)/odds(2005). Ranking provides additional feature for interpretation, i.e., which pair is smaller or larger compare to other. The dashed line at x=1 is to assess the statistical significance for ratio.**

The *Difference in Proportion* plot shows pairwise differences with CI on ranked order. D[85] implies odds(2008) — odds(2005). The dashed line at x=0 is to assess the statistical significance for difference.

6.1 Result Summary and Conclusions

- **Contingency Tables:** Based on May 2004, 2005, and 2008 datasets (Table 3b), about 1-2% indexes are attributable to 80% of the variance while 98-99% indexes are attributable to 20% of the variance. May 2006 shows a different behavior—only 0.01% index (1 index out 8,018) is attributable to 80% of the variance. This anomalous behavior is reflected on the "All U.S.— All Items" variance of May 2006 that has the largest variance (0.04214) compare to other years.

Similarly, based on May 2004, 2005, and 2008 datasets (Table 4b), about 22-27% indexes are attributable to 99% of the variance while 73-78% indexes are attributable to only 1% of the variance. For May 2006, only 9% indexes are attributable to 99% of the variance. Similar explanations also hold for May 2007 aggregate variance which is the second largest among the group.

In other words, these conditional contingency tables (80-20 and 99-1 benchmarks) empirically show that across all the groups, a small proportion of indexes are attributable to a large proportion of variance in general. Additionally, in the event that a very few indexes contribute to high proportion of aggregate variance, an increase in aggregate variance may also be observed compared to other years. Thus, monitoring these outlying behavior of basic index variances is once again confirmed from the categorical analysis perspective. The result answers the question of interest and provides a technique to construct a baseline information for future reference, i.e., what proportion of the basic indexes contribute the most variance.

- **Odds:** May 2006 has the lowest odds relative to other years (2004-2008) for both benchmarks (Tables 3b & 4b). Among other properties, we observe that the ranks of "All U.S. – All Items" aggregate variance has an inverse functional form with odds (Table 5 and corresponding graph), i.e., lower the odds, higher the aggregate variance. One likely explanation of the phenomena could be: lower odds implies only a few indexes contributed to a large proportion of variance in this context; thus, generating a higher risk that an anomaly might have taken place in an index level, resulting in a larger disparity in attributable variances than expected. Additionally, it is also observed as odds are very similar, aggregate variances are also similar, such as May 2005 and May 2008 (Table 5).

- **Pairwise Comparison of Odds:** Since odds are systematically lower for the indexes that contribute the most aggregate variance, it is useful to conduct a pairwise comparison between all the groups, i.e., odds of one event with another event (odds ratio) for further examination (Tables 6 & 7; associated graphs). Odds of 2008 and 2005 are similar (1 is included within the CI), while all the pairwise comparisons of odds ratio show significant difference (1 is not within the CI) for both benchmarks. The large sample size (n=8,018) contributed for significant results in pairwise comparison.

Based on the results of 80-20 benchmark, the odds of the number of indexes attributable to 80% variance ranges about 1-12 times the other years across all years except May 2006. On the other hand, pairwise comparisons with May 2006 reveals a different pattern—other years are 15-185 times the odds of May 2006. Additionally all the pairwise comparisons with 2006 show a much wider 95% confidence interval. Similarly, for 99-1 benchmark, odds of the number of indexes attributable to 99% variance for 2004, 2005, and 2008 are 1-1.3 times likely the odds among them, 1.6-2 times the odds of May 2007, and about 2.8-3.6 times the odds of May 2006. In other words, a distinct range of OR pattern across as well as a larger odds ratios in the behavior May 2006 pairwise comparisons are observed. It also provides a baseline information for future reference for a range of possible OR.

- **Pairwise Comparison of Difference in Proportion:** Pairwise difference between all groups except 2008-2005 show significant difference (0 is not within the CI) for both benchmarks (Tables 6 & 7; associated graphs). The large sample size (n=8,018) contributed for significant results in pairwise comparisons. Because odds are systematically lower in proportion, difference in proportion is also small. Thus, OR may have provided a better interpretation than DiffP. DiffP shows a smaller range (0.00175-0.02245) for 80-20 benchmark, but a larger range (0.01122-0.17398) for 99-1 benchmark. Pairwise comparisons between May 2006 and 2004, 2005, 2008 show the largest range of difference (0.13108-0.17398) among all other groups in 99-1 benchmark. The difference between 2008-2006 is the highest in both benchmark.

7. Recommendations

All the results converge to a single idea: anomalous behavior in a handful of basic indexes may potentially contribute a large proportion of variance, that in turn may pose a risk on "All U.S. – All Items" aggregate variance to reach its performance threshold ($\sigma < 0.25$). Thus, monitoring the decomposed *weighted*-variances using exploratory and data visualization techniques may be beneficial for CPI program in the long run. Month-to-month superimposed data visualization of decomposed *weighted*-variances may be an effective monitoring tool as they provide a reference group(s) for comparison. It may also provide additional insights about data issues that may be addressable with future improvements to the sample design.

R Software and SAS

- **R** packages used:
  - ggplot2 (Wickham, Hadley. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009).
  - coin (Hothorn, T. et al. *The American Statistician*, 2006. see below).
  - asbio (Aho, K. *Foundational and applied statistics for biologists using R*, 2014. see below)
- **SAS** procedure:
  - proc npar1way

References

Agresti, A., Bini, M., Bertaccini, B., and Ryu, E. (2008). Simultaneous confidence intervals for comparing binomial parameters. *Biometrics* 64: 1270–1275. Retrieved from http://www.stat.ufl.edu/~aa/articles/agresti_bini_bertaccini_ryu_2008.pdf.

Agresti, A., Bini, M., Bertaccini, B., and Ryu, E. (2008). R code to find simultaneous confidence intervals for binomial proportions. Supplemental material for the method proposed in *Biometrics* 64: 1270–1275. Retrieved from http://www.stat.ufl.edu/~aa/cda/R/multcomp/ryu-simultaneous.pdf.

Aho, K. (2014). *Foundational and applied statistics for biologists using R*. New York: CRC Press, Taylor and Francis Group.

Brown, Morton B.; Forsythe, Alan B. (1974). "Robust tests for equality of variances". *Journal of the American Statistical Association*, **69**: 364–367. (Retrieved from: http://drsmorey.org/bibtex/upload/Brown:Forsythe:1974.pdf).

Chambers, J., Cleveland, W., Kleiner, B., & Tukey, P. (1983). *Graphical methods for data analysis*. Monterey, CA: Wadsworth.

Holland, P. W. (1989). A note on the covariance of the Mantel–Haenszel log-odds-ratio estimator and the sample marginal rates. *Biometrics*, 45: 1009–1016.

Hothorn, T., Hornik, K., van de Wiel, M. A. and Zeileis, A. (2006). Lego System for Conditional Inference. *The American Statistician*, 60(3): 257-263.

Higgins, J. J. (2004). *An introduction to modern nonparametric statistics*. Pacific Grove, CA: Thomson, Brooks/Cole.

McGill, R., Tukey, J. W., & Larsen, W. A. (1978), "Variations of box plots". *The American Statistician*, *32*, 12–16.

Schmidt, C.O., and Kohlmann, T. (2008). When to use the odds ratio or the relative risk? *International Journal of Public Health*, 53: 165–167. Retrieved from http://www.researchgate.net/publication/23762851_When_to_use_the_odds_ratio_or_the_relative_risk.

Zhang, J., and Yu, K. (1998). What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. *The Journal of the American Medical Association*, 280: 1690–1. Retrieved from http://www.research.labiomed.org/biostat/Education/Case%20Studies%202005/Session4/ZhangYu.pdf.

U.S. Bureau of Labor Statistics. (2015). Chapter 17: The Consumer Price Index. *BLS Handbook of Methods*. Retrieved from http://www.bls.gov/opub/hom/homch17.htm.