Constructing Better "Confidence" Intervals for Proportions and Tests of Differences Among Proportions Using a Stratified Multistage Sample

Phillip S. Kott
RTI International, 7413 Tupelo Drive, Rockville, MD 20855

Abstract
Coverage intervals for a parameter estimate computed using complex survey data are often constructed by assuming that the parameter estimate has an asymptotically normal distribution. The size of the sample and the nature of the parameter being estimated render the conventional "Wald" methodology and the term "confidence interval" dubious in many applications. A revised method of coverage-interval construction has been developed that "speeds up the asymptotics" of a complex-survey estimator by incorporating a measure of its third-central moment. Public-use data from the 2019 National Survey on Drug Use and Health (NSDUH) are employed to demonstrate the construction of third-moment-adjusted coverage intervals for population proportions of cocaine and crack use within seven race/ethnicities. A third-moment-adjusted Bonferroni methodology is proposed for making comparison among these seven subdomains. The NSDUH public-use data set contains two variance primary sampling units within each variance stratum and little additional information about the sample design. This complicates measuring third-central moments.
Key Words: Edgeworth Expansion, Third-Central Moment, Holm-Bonferroni, Variance Stratum

1. Introduction

Government surveys often estimate population proportions, such as the fraction of the US population that had used cocaine in a previous year, based on a stratified multistage sample. A two-sided $(1 - \alpha) \times 100\%$ coverage interval for a population proportion purports to contain all values P such that there is no more than $(1 - \alpha) \times 100\%$ probability that absolute difference between the nearly unbiased estimate p and P is as large as it is. It is common, but misguided, to call this interval a "confidence" interval for the population proportion. Were the estimator normally distributed and based on a simple random sample, one could be *confident* that such an interval could be constructed . That is not the case for a proportion estimated from complex survey data. In such a case, there is no certainty involved, one can only expect that the interval will "cover" the true population proportion, regardless of what that true proportion is, around $(1 - \alpha) \times 100\%$ of the time in repeated sampling.

Heedless of the above caveat, it is common practice the rely on the asymptotic normality of the estimator p, and a nearly unbiased estimate of its variance v, and to assert that all P such that $(p - P)^2 < v(z_{1-\alpha/2})^2$ are in a $(1 - \alpha) \times 100\%$ confidence interval for the true proportion, where $z_{1-\alpha/2}$ is the value normal score at $1-\alpha/2$ (an estimator like p is "nearly unbiased" when its bias is of a smaller asymptotic order than the value it is estimating). In fact, it is well known that this may not even constitute a good coverage interval. Brown

et al. (2001), among others, demostrates this for simple random samples, and Franco et al. (2019) for complex samples.

Given a stratified multistage sample and treating the sampling of primary sampling units (PSUs) within strata as if they were drawn with replacement, Kott and Liu (2010) argue that instead of the usual two-sided $(1 - \alpha) \times 100\%$ coverage interval for the population proportion, one should construct the following third-moment-adjusted interval:

$$p + \delta - \sqrt{(z_{1-\alpha/2})^2 v + \delta^2} \le P \le p + \delta + \sqrt{(z_{1-\alpha/2})^2 v + \delta^2}\ , \qquad (1)$$

where $\delta = \left(1/6 + (z_{1-\alpha/2})^2/3\right)b$, $b = m_3/v$, and m_3 is a nearly unbiased estimator for the third-central moment of p, that is, the expected value of the cube of the difference between p and its expected value.

Kott and Liu discuss one-sided intervals, but the extension to two-sided intervals is obvious. Andersson and Nerman (2000) had previously proposed two-sided intervals that attempt to adjust for the impact of the v and p being correlated. In place of δ in equation (1), their intervals effectively feature $\delta^* = \frac{1}{2}(z_{1-\alpha/2})^2 b$ (observe that $\delta^* > \delta$ when $z_{1-\alpha/2} > 1$). Andersson-Nerman intervals are very similar to Wilson and logistic-transformation intervals. Kott (2017) provides a discussion of their similarities and differences.

It is important to realize that b in equation (1) is itself an estimate. Let $B = M_3/V$ be the value it is estimating, where V and M_3 are respectively the second- and third-central moments of p.

When the number of strata grows arbitrary large while the numbers of PSUs within strata are bounded, δ converges to 0, and the interval in equation (1) to the conventional "Wald" interval centered at p. Otherwise, the center of the interval moves in the same direction as the sign of m_3, and the range of the interval increases, usually not by very much.

When the sample design is with-replacement simple random sampling, B can be estimated in a nearly unbiased fashion by

$$b_{srs\_wr} = \frac{m_3}{v} = \frac{\frac{p(1-p)(1-2p)}{n^2}}{\frac{p(1-p)}{n}} = \frac{1-2p}{n}\ .$$

This suggests that a more general *ad-hoc estimator* for B is $b_{alt} = (1–2p)/n^*$, where $n^* = p(1–p)/v$ is the so-called effective (element) sample size given the actual sample design. By using b_{alt} as an estimate for B, one assumes that unequal weighting, clustering, and stratification have the same impact on the third-central moment as they do on the second.

Kott and Liu derive an equation like (1) using an Edgeworth expansion and replacing the variance of p with $v - b(p - P)$, which is a more efficient – if idealized – measure of the variance than v when v and p are correlated (as was noted by Andersson and Nerman). In their formulation the estimand need not be a population proportion so long as p is a nearly unbiased estimator for it. The intervals in Kott and Liu are sensitive not only to the

possible correlation between v and p, but also to p being asymetric when P is not equal to ½.

Our primary focus here will be on estimating proportions and differences among proportions computed from the 2019 National Survey on Drug Use and Health (NSDUH) public-use file (PUF) available at https://www.datafiles.samhsa.gov/dataset/national-survey-drug-use-and-health-2020-nsduh-2020-ds0001. What is novel here is the focus on estimating the differences among proportions and on problem of third-order-moment estimation when there are only two PSUs in a stratum, as is often the case in multistage government surveys.

The NSDUH PUF, like the NSDUH itself, features two *variance* PSUs within each of 50 *variance* strata (the variance strata in the PUF are distinct unions of the 750 variance strata in the NSDUH itself). One of the two design PSUs in a design stratum has been assigned to a variance PSU within the variance stratum containing its design stratum. The other design PSU has been assigned to the other variance PSU in the variance stratum. Design PSU and stratum identifiers are not available on the PUF, which makes it very difficult for an intruder to connect survey values on the PUF with particular respondents.

By treating the design PSUs as if they were selected with replacement, an inverse-probability-weighted (i.e., design-weighted) sum computed from a variance PSU is independent of the corresponding weighted sums of the other variance PSUs, both within a variance stratum and across variance strata. Moreover, the variance PSU's weighted sum estimates half of the total for its variance-stratum. This is very useful for variance estimation but not for estimating a third-central moment.

Section 2 discusses the estimation of proportions, differences of proportions, and their second and third-central moments calculated from a stratified multistage sample assuming, as is common, with-replacement selection of design PSUs within each design stratum. It offers two imperfect solution to the problem of estimating third-central moments with two PSUs per stratum. Section 3 applies equation (1) and the estimators for the previous section to the estimation of the proportions of cocaine and crack use within seven race ethnicities using the 2019 NSDH PUF. Section 4 proposes and implements a Bonferroni methodology for assessing differences among the seven subdomains. Section 5 offers some concluding remarks and proposes a new suppression rule based on the range where the conventional coverage interval can be used.

2. Some Estimators (a Bit of Theory)

Suppose we have a stratified multistage sample of elements from a population. For simplicity, assume there is no nonresponse or measurement error in the sample or coverage error in the sampling frame. Let H, the number of (variance) strata, be large, and n_h, the number of with-replacement sampled (variance) PSUs in stratum h be at least 3 in every stratum but small compared to H. Let Y and X be population totals for two variables of interest, and Y_h and X_h the respective subtotals in stratum h. Now suppose y_{hi} and x_{hi} are inverse-probability-weighted estimates of Y_h/n_h and X_h/n_h respectively based on the sample in PSU i of stratum h. Under mild conditions, we assume to hold, a nearly unbiased estimator for $R = Y/X$ is

$$r = \frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h} y_{hi}}{\sum_{h=1}^{H}\sum_{i=1}^{n_h} x_{hi}} = \frac{y}{x}. \tag{2}$$

We further assume mild conditions under which $r - R$ is $O_P(1/H^{1/2})$, and $E(r - R) = O(1/H)$, so that the second and third central-moments of r are nearly equal to $E[(r - R)^2]$ and $E[(r - R)^3]$, respectively. Moreover, conditions are such that the former, the variance/mean-squared error of r, is $O(1/H)$, and the latter is $O(1/H^2)$. Finally, the same can be said about the central moments of the unbiased estimators y and x, the numerator and denominator of r.

As a result of these assumptions, nearly unbiased estimators for the second and third-central moments of r (when H is large) are

$$v = \frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h}(n_h z_{hi} - z_h)^2}{n_h(n_h - 1)}, \text{ and } m_3 = \frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h}(n_h z_{hi} - z_h)^3}{n_h(n_h - 1)(n_h - 2)}, \tag{3}$$

where $z_{hi} = \frac{y_{hi} - r x_{hi}}{\sum_{g=1}^{H}\sum_{k=1}^{n_g} x_{gk}}$, and $z_h = \sum_{i=1}^{n_h} z_{hi}$.

Each z_{hi} is a nearly unbiased estimator of Z_h/n_h, where

$$Z_h = \frac{Y_h - R X_h}{X}. \tag{4}$$

Note that $\sum_{h=1}^{H} Z_h = 0$.

For our purposes, x, the denominator of r, estimates a total such as the number of non-Hispanic blacks in the US in 2019, while y, the numerator, correspondingly estimates the number of non-Hispanic blacks who had ever used cocaine. Consequently, r estimates the fraction of non-Hispanic blacks in the US in 2019 who had ever used cocaine. It is not hard to see that when the h denote variance strata, and $n_h = 2$ the number of variance PSUs in variance stratum h, the variance estimator v in equation (3) remains nearly unbiased assuming H (now the number of variance strata) remains large, which is needed to justify using asymptotic arguments. This is because the $(y_{hi} - R x_{hi})/X$ are independent across the variance PSUs (the hi), and the sum, $(y_{h1} - R x_{h1})/X + (y_{h2} - R x_{h2})/X$ estimates Z_h in equation (4). That is one reason why variance strata and variance PSUs are used in the NSDUH PUF.

Unfortunately, the third-central-moment estimator m_3 in equation (3) cannot be used with the NSDUH PUF because there are only two variance PSUs in a variance stratum. Instead, the appendix shows that a reasonable measure of the third central moment, one we will use in the sequel, is the following:

$$m_3' = \frac{4}{3}\left(\sum_{h=1}^{H} z_{h1}^3 + z_{h2}^3\right) - \frac{1}{3}\sum_{h=1}^{H}\left(z_{h1} + z_{h2}\right)^3. \tag{5}$$

It is nearly unbiased when $\sum_{h=1}^{H} Z_h[(z_{h1} - \frac{1}{2}Z_h)^2 + (z_{h2} - \frac{1}{2}Z_h)^2] \approx 0$.

The appendix also investigates two other potential estimates for the M_3, the third-central moment of r. One estimates M_3 with m_3 in equation (3) as if all $2H$ PSUs came from a single stratum. The other requires there to be an even number of strata. It randomly divides the strata into groups of two and treats each group of four PSUs as if they came from a single stratum.

To estimate the difference between two proportion using the same sample of PSUs, one need estimate r in equation (2) and the z_{hi} in equation (3) and (5) separately for each proportion. Labels these $r^{(k)}$ and $z_{hi}^{(k)}$, where $k = 1$ or 2 denote the two proportions being differenced (note that the n_h and H are the same for both estimated proportions). The difference between the two estimated r is simply $p^{(1-2)} = r^{(1)} - r^{(2)}$. For the estimates of the second and third-central moments of $p^{(1-2)}$ one can replace the z_{hi} in equations (3) and (5) with $z_{hi}^{(1)} - z_{hi}^{(2)}$.

3. Some Coverage Intervals Computed from the 2019 NSDUH PUF

The 2019 NSDUH PUF comes with analysis weights that are adjusted for nonresponse and undercoverage. In this section, we will treat those weights as perfect inverse-probability weights and ignore the impact of the implicit model fitting involved in their determination. We use those weights to estimate the proportions (in percentages) of lifetime and past year cocaine (in any form) and crack use ($4 = 2 \times 2$ variables in all) within the following seven race/ethnicities for US adults 12 or above residing in noninstitutional dwelling places:
Non-Hispanic **White**
Non-Hispanic **Black**
Non-Hispanic American Indian/Alaska Native (**AIAN**)
Non-Hispanic Native Hawaiian/Other Pacific Islander (**NHOPI**)
Non-Hispanic **Asian**
non-Hispanic more than one race (**Multi**)
Hispanic.

In addition to the estimated proportions, 95% coverage intervals were computed using equation (1) estimating B in three different ways: 1, using m_3' from equation (5), yielding the *proposed adjusted interval*) 2, with $(1-2p)v/n^*$, the *ad-hoc adjusted interval*; and 3, with 0, the conventional or *Wald interval*. The results are displayed in Table 3.1.

Table 3.2 digs into the results. It displays sample sizes and the percent increases in the center (from p to $p + \delta$,) and the range (from $\sqrt{(1.96)^2 v}$ to $\sqrt{(1.96)^2 v + \delta^2}$) of the interval caused by using the proposed adjusted interval in place of the conventional interval. It also displays the estimated skewness (ideally $\tau = m_3/v^{3/2} = b/v^{1/2}$) computed using m_3' in place of m_3 (the proposed approach) and then $(1-2p)/n^*$ in place of b (the *ad-hoc* approach).

All the skewness measures, when they exist, are positive (they do not exist when $p = 0$). The proposed adjusted skewness measure is more often larger than the *ad-hoc* adjusted skewness measure. The two appear to be correlated, but not perfectly. The larger the skewness measure the larger the percent increases in the center and the range. The latter increase is always smaller than the former.

Table 3.1. Some Estimates and the Associated Two-Sided 95% Coverage Interval

Race/ Ethnicity	Estimated Proportion (in percent)	Lower Bound: Proposed	Lower Bound: *Ad-hoc*	Lower Bound: Wald	Upper Bound: Proposed	Upper Bound: *Ad-hoc*	Upper Bound: Wald
Cocaine Lifetime Use							
White	18.25	17.57	17.57	17.56	18.95	18.95	18.94
Black	8.71	7.38	7.39	7.32	10.17	10.18	10.10
AIAN	14.90	11.08	11.30	10.99	18.90	19.15	18.82
NHOPI	10.35	7.54	7.55	7.25	13.77	13.78	13.45
Asian	3.75	2.72	2.76	2.65	4.92	4.96	4.84
Multi	20.34	17.78	17.66	17.56	23.37	23.24	23.13
Hispanic	11.20	10.10	10.11	10.07	12.37	12.38	12.34
Cocaine: Past Year Use							
White	2.01	1.82	1.82	1.81	2.23	2.23	2.22
Black	1.66	1.20	1.16	1.10	2.36	2.30	2.23
AIAN	1.46	0.89	0.83	0.69	2.47	2.38	2.22
NHOPI	2.96	2.00	1.87	1.69	4.63	4.44	4.23
Asian	0.95	0.58	0.57	0.49	1.52	1.50	1.41
Multi	2.94	2.08	2.07	1.95	4.08	4.05	3.92
Hispanic	2.17	1.84	1.84	1.82	2.55	2.54	2.52
Crack: Lifetime Use							
White	3.84	3.52	3.52	3.51	4.18	4.17	4.16
Black	4.00	3.19	3.18	3.11	4.97	4.95	4.88
AIAN	4.35	2.23	2.15	1.60	7.93	7.80	7.10
NHOPI	3.56	1.63	1.39	0.66	7.94	7.45	6.47
Asian	0.69	0.32	0.30	0.18	1.37	1.35	1.20
Multi	8.49	6.55	6.44	6.25	11.08	10.94	10.73
Hispanic	1.85	1.50	1.48	1.45	2.29	2.28	2.24
Crack: Past Year Use							
White	0.25	0.19	0.19	0.19	0.32	0.32	0.32
Black	0.60	0.30	0.27	0.17	1.21	1.15	1.03
AIAN	0.00	.	.	.	.	.	.
NHOPI	0.91	0.12	0.10	-0.50	3.45	3.37	2.33
Asian	0.01	0.00	0.00	-0.01	0.04	0.04	0.03
Multi	0.36	0.13	0.10	-0.03	1.01	0.93	0.74
Hispanic	0.10	0.05	0.04	0.02	0.24	0.21	0.19

AIAN- American Indian/Alaska Native: NHOPI – Native Hawaiian/Other Pacific Islander

Table 3.2. Some Statistics About the Coverage Intervals

Race/Ethnicity	*n*	*n**	$\tau_{proposed}$	$\tau_{ad\text{-}hoc}$	Percent Center Increase	Percent Range Increase
Cocaine: Lifetime Use						
White	32089	12017	0.02	0.01	0.05	0.01
Black	7256	1580	0.06	0.07	0.73	0.11
AIAN	752	319	0.03	0.11	0.58	0.02
NHOPI	292	371	0.13	0.14	2.88	0.46
Asian	2697	1153	0.09	0.14	1.96	0.22
Multi	2202	803	0.11	0.05	1.12	0.34
Hispanic	10848	2966	0.04	0.05	0.28	0.04
Cocaine: Past-Year Use						
White	32089	18106	0.08	0.05	0.63	0.19
Black	7256	1968	0.28	0.17	7.05	2.13
AIAN	752	948	0.39	0.26	15.25	4.16
NHOPI	292	684	0.38	0.21	12.08	3.88
Asian	2697	1717	0.28	0.24	9.98	2.12
Multi	2202	1128	0.20	0.17	4.85	1.04
Hispanic	10848	6633	0.10	0.08	1.16	0.26
Crack: Lifetime Use						
White	32089	13286	0.06	0.04	0.39	0.11
Black	7256	1882	0.13	0.11	2.13	0.46
AIAN	752	211	0.36	0.31	16.72	3.43
NHOPI	292	156	0.57	0.40	34.32	8.50
Asian	2697	1029	0.42	0.37	22.60	4.64
Multi	2202	594	0.20	0.12	3.85	1.06
Hispanic	10848	4436	0.16	0.11	2.50	0.68
Crack : Past-Year Use						
White	32089	22722	0.17	0.13	3.16	0.74
Black	7256	1262	0.49	0.36	25.68	6.34
AIAN	752	.	.	.	.	.
NHOPI	292	174	0.83	0.78	95.15	17.38
Asian	2697	14492	0.72	0.72	74.70	13.19
Multi	2202	910	0.74	0.55	59.39	13.99
Hispanic	10848	5922	0.74	0.40	42.99	13.83

4. Comparing Proportions Across Race/Ethnicities Using the 2019 NSDUH PUF

The usual way of testing whether the proportions are the same across subdomains is with some version of a multivariate *F* test. Korn and Graubard (1999) recommend using either what SUDAAN calls the *Adjusted Wald F* or the *Satterthwaite Adjusted F* (Research Triangle Institute, 2012, p. 315). Both rely on the asymptotic normality of the estimated proportions, and both find variance estimation difficult when some of the estimated proportions are very small.

A conservative test for whether or not 21 (i.e., $\binom{7}{2}$) pairs of race/ethnicity proportions are significantly different from each other at the .05 level is with a Bonferroni

adjustment (Holm, 1979 describes the Bonferroni adjustment). With this procedure, if any of the 21 pairs are significant different at the two-sided .05/21 level, then the null hypothesis of the equality of all seven proportions is rejected at the .05 level.

Using this procedure our four variables, lifetime and past-year use of cocaine and crack, all show significant differences across the seven race/ethnicities. This result obtains whether we compare estimated proportions for subdomains f and g by computing $b_{proposed}^{(f-g)} = m_3'/v$ with a combined m_3' as described in Section 2, pretending the third-central moment of the difference is 0 (the conventional or Wald approach), or pretending that the estimated proportions p_f and p_g are close to independent andcomputing the following *ad-hoc* estimator for $B^{(f-g)}$ when p_f and p_g are both between 0 and 1:

$$b_{ad\text{-}hoc}^{(f-g)} = \frac{\dfrac{p_f(1-p_f)(1-2p_f)}{(n_f^*)^2} - \dfrac{p_g(1-p_g)(1-2p_g)}{(n_g^*)^2}}{\dfrac{p_f(1-p_f)}{n_f^*} + \dfrac{p_g(1-p_g)}{n_g^*}}. \qquad (6)$$

When $p_f = 0$ or 1, n_f^* in this equation is replaced by 1. Observe that $b_{alt}^{(f-g)}$ can be positive, negative, or zero depending on the relative sizes of the estimated proportion and the relative sizes effective sample sizes. It exists when either p_f or p_g is 0, but not when both are 0.

The multivariate *F*-test results are the same for three of the four variables. SUDAAN warns against their use for the fourth, past-year crack use, because the estimated proportion of users among Non-Hispanic American Indian/Alaskan Natives (AIAN) is 0.

One advantage of the Bonferroni procedure over a multivariate *F* test in general is that the former can be used to assess *which* pairs of subdomains have significantly different proportions. For a pair of subdomains f and g, let

$$Test^{(f-g)} = \frac{|\, p^{(f-g)} + \delta^{(f-g)} \,|}{\sqrt{(z_{1-\alpha/(21\times 2)})^2 v^{(f-g)} + \delta^{(f-g)2}}}, \qquad (7)$$

where $p^{(f-g)}$ is the difference in their estimated proportions, $v^{(f-g)}$ the estimated variance of that difference, and $\delta^{(f-g)}$ the appropriate redefinition of δ in equation (1). Surely, when the test statistic in equation (7) is greater than 1 for a pair f and g, the difference $P^{(f-g)}$ is significantly different from zero, and P^f and P^g are significantly different from each other. Setting aside the c pairs (if any) where the right-hand side of equation (7) is greater than 1, we can assess if any of the remaining pairs are also significantly different from each other by replacing 21 in the above formulation by $21 - c$, then continuing the process and stopping when no additional pairs are deemed significantly different.

The top half of Table 4.1 displays the nine pairs of subdomains that have proportions of past-year crack use that are significantly different at the .05 level using the conservative Holm-Bonferroni procedure described above with $\delta^{(f-g)}$ determined using either the proposed adjusted or *ad-hoc* adjusted coverage interval (note that Pair 5-6 was in the original batch of significant differences when the proposed interval was used, but not the *ad-hoc* interval). Using the Wald interval only two pairs are so assessed.

A common practice after the original conservative Bonferroni procedure determines that at least one pair have significantly different proportions (conservative because there may be *less than* 5% probability of finding a significant difference when there is none), an unadjusted *t* test should be used to assess which pairs are significantly different. That more liberal criteria (one more likely to find a significant difference when there isn't any) is employed in the bottom half of the table. Using the proposed adjusted coverage intervals, 16 pairs have significantly different proportions; using the *ad-hoc* adjusted intervals, 14 pairs; and using the Wald, only eight.

Table 4.1 Significantly Different Pairs at the .05/Q Level for Past-Year Crack Use

Q	Pair	Difference	$\tau_{proposed}$	$\tau_{ad\text{-}hoc}$	$Test_{proposed}$	$Test_{Wald}$	$Test_{ad\text{-}hoc}$
Significantly Different Pairs According to Holm/Bonferroni Procedure							
21	1-3	0.25097	0.165	0.128	2.62	2.49	2.64
21	1-5	0.23762	0.159	0.100	2.48	2.35	2.39
21	2-3	0.60033	0.490	0.378	1.27	Not	1.19
21	2-5	0.58698	0.489	0.377	1.25	Not	1.17
21	2-7	0.49668	0.474	0.364	1.12	Not	1.02
21	3-6	-0.35698	-0.741	-0.553	1.09	Not	1.02
21	3-7	-0.10365	-0.736	-0.390	1.26	Not	1.15
21	5-6	-0.34363	-0.731	-0.549	1.06	Not	Not (yet)
21	5-7	-0.09030	-0.731	-0.362	1.18	Not	1.01
13	5-6	-0.34363	-0.731	-0.549	1.08	Not	1.02
Significantly Different Pairs without a Bonferroni Adjustment							
1	1-2	-0.34936	-0.481	-0.367	1.10	Not	1.02
1	1-3	0.25097	0.165	0.128	3.95	3.86	4.00
1	1-5	0.23762	0.159	0.100	3.73	3.64	3.63
1	1-7	0.14732	-0.345	-0.159	1.14	1.43	1.30
1	2-3	0.60033	0.490	0.378	1.66	1.41	1.59
1	2-5	0.58698	0.489	0.377	1.63	1.37	1.56
1	2-7	0.49668	0.474	0.364	1.43	1.17	1.33
1	3-4	-0.91484	-0.833	-0.796	1.07	Not	1.06
1	3-5	-0.01335	-0.718	-0.720	1.10	Not	1.10
1	3-6	-0.35698	-0.741	-0.553	1.29	Not	1.23
1	3-7	-0.10365	-0.736	-0.390	1.59	1.26	1.51
1	4-5	0.90148	0.837	0.798	1.07	Not	1.05
1	4-7	0.81118	0.833	0.792	1.01	Not	Not
1	5-6	-0.34363	-0.731	-0.549	1.25	Not	1.20
1	5-7	-0.09030	-0.731	-0.362	1.45	1.10	1.31
1	6-7	0.25333	0.697	0.523	1.03	Not	Not

Pair Numbers:

1. Non-Hispanic White
2. Non-Hispanic Black
3. Non-Hispanic American Indian/Alaska Native
4. Non-Hispanic Native Hawaiian/Other Pacific Islander
5. Non-Hispanic Asian
6. Non-Hispanic more than one race
7. Hispanic

Not – Not significant at the .05/Q level.

For convenience, let us say that the pairs of proportions significantly different using the Holm-Bonferroni procedure are *certain* to be significantly different, while those pairs significant different only under the more liberal criteria *may* be significantly different.

For the other three variables studied here the following number of pairs were deemed certain to be significantly different at the .05 level (not shown): 13 pair of proportions of lifetime cocaine users employing each of the three intervals, 4 pairs of past-year cocaine users again employing each of the three intervals, and 9 pairs of lifetime crack users employing Wald intervals and 10 employing either one of the other two intervals. 16, 9, and 14 pairs may have been significantly different for lifetime cocaine, past-year cocaine, and lifetime crack use employing either the proposed adjusted or the *ad-hoc* adjusted intervals. Employing Wald intervals, those numbers decreased W to 15, 6, and 12.

In all cases, all the pairs deemed significant employing Wald intervals were also deemed significant using either one of the other two intervals. The pairs deemed significant employing the proposed adjusted intervals and the *ad-hoc* adjusted interval were the same. There is no guarantee of these relationships for all variables.

5. Some Concluding Remarks

5.1 Summarizing the Findings

An initial motivation for this endeavor was to apply two-sided versions of the coverage intervals described and empirically justified in Liu and Kott (2009) and Kott and Liu (2009, 2010) to subdomain proportions estimated from a complex US government survey like the 2019 NSDUH, and in so doing, display their practical relevance. It turned out, however, that the two-sided interval in equation (1) could not be computed directly because the 2019 NSDUH PUF, like many multistage government surveys, features only two variance PSUs in each variance stratum.

Two competing methods were used here for measuring the third-central moment of an estimator for a proportion. One, called the *proposed method*, makes a sensible assumption (contained below equation (5)) but requires computer code not yet readily available. The other, the *ad-hoc method*, pretends the impact on the third-central moment of an estimated proportion of unequal weighting, clustering, and stratification, jointly measured in the effective sample size, is the same as the impact on the second. The two methods yielded somewhat, but not exactly, similar results and neither was investigated empirically here.

The two methods were used to construct two-sided coverage intervals for lifetime and past-year use of cocaine and crack for seven race/ethnicity subdomains. The results are displayed in Tables 3.1 and 3.2. We can see in the tables that the smaller the element sample size of the subdomain, the more sensitive its coverage interval is to the skewness of the investigated estimator. The skewness tends to increase as the estimated proportion gets smaller (that is for estimates less than .5, which was always the case here).

Of growing concern in the US is testing differences in proportions among tightly defined race/ethnicity subdomains of various sizes. That is something not frequently addressed in the literature but faced directly here. For testing differences among proportions, the Bonferroni procedure, described in Section 4, is recommended over more conventional multivariate *F* tests. The Bonferroni procedure can capture the nonnormality of the estimated differences (see Table 4.1), and it can be used even when one of the estimated proportions is 0, which may be due to the associated subdomain's small sample size rather

than the value of the estimand. Here, again the two method of measuring the impact of third-central moments yield similar but not identical results.

5.2 Cochran's Conjecture (and a new suppression rule)

Tables 3.1 and 3.2 appear to confirm what Cochran (1963, p. 41) advises for simple random samples: When the estimated skewness is less the .2, it is relatively safe to use the conventional coverage interval. Note that this was always the case in 2019 for the Non-Hispanic White proportions and for all lifetime cocaine-use proportions. Cochran's advice suggests one should estimate skewness first before constructing a conventional interval. Using the proposed method is preferred, but the *ad-hoc* method is relatively trivial to compute, especially when it can be determined that the conventional interval is adequate for one's needs.

Cochran's focus on the estimated skewness τ rather than on $b = m_3/v$ makes some sense because the range of the coverage interval in equation (1) is dominating by a multiple of $v^{1/2}$, while δ, the displacement of the center of the interval, is linearly related to the τ times $v^{1/2}$, which is of a smaller asymptotic order.

The *ad-hoc* estimated skewness of an estimated proportion p is

$$\tau_{AH} = \frac{1-2p}{[n^* p(1-p)]^{1/2}} .$$

Some algebra reveals that this values is safely less than .2 when

$$\frac{25}{n^*} \leq p \leq 1 - \frac{25}{n^*} .$$

This suggests when p is within this range, the conventional [confidence] interval can be used (and p published if that is the issue). Otherwise, prudent practice when p is less than $25/n^*$, is to replace p as *the* estimate with the assertion that the true P is estimated to be less than $25/n^*$. Analogously, when p is greater than $1 - 25/n^*$, P is estimated to be greater than $1 - 25/n^*$. In both cases, a conventional upper (or lower) bound can be computed, treating the estimated standard error as if it was $[(25/n^*)(1 - 25/n^*)/n^*]^{1/2}$.

5.3 Calibration Weights

One subject not yet addressed here is how to use the interval in equation (1) when the estimated p is computed with calibrated weights rather than inverse-probability weights. If the calibration is to population totals of the calibration variables, the z_{hi} in the equations (3) and (5) can be replaced by regression residuals when estimating a total and by regression residuals divided by the (estimated) population size when estimating a mean. When calibrating to pre-nonresponse sample totals or estimated totals from a previous phase of sampling, how to estimate third-central moments is a question for future research. A subject given much undo attention in the literature (starting with Korn and Graubard (1998)) is the possible need to adjust for the variance of the variance estimator v in equation (1). Standard practice is to treat $(p - P)/v^{1/2}$ as if it had close to a Student's t distribution with degrees of freedom equal to the number of variance PSUs minus the number of variance strata. That won't do here because $(p - P)/v^{1/2}$ is unlikely to have anything close to a t distribution. Moreover, it is not the variance of v that is relevant here, but the variance of $v - b(p - P) \approx v - B(p - P)$, which is smaller by design. Realizing that equation (1) is

only an approximation (with most terms of smaller asymptotic order than $O_P(1/H)$ removed), we make no "degrees of freedom" adjustment to the $z_{1-\alpha/2}$ in the equation.

Appendix

The equality $(a + b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$ is used repeatedly in what follows.

Estimating M_3 (when H is large and all $n_h = 2$) with m_3' in equation (5)

Suppose z_1 and z_2 are independent and unbiased estimators of ½ Z, so that $z_1 + z_2$ is an unbiased estimator of Z. The third-central moment of $z = z_1 + z_2$ is

$$E[(z_1 + z_2 - Z)^3] = E\{([z_1 - \tfrac{1}{2} Z] + [z_2 - \tfrac{1}{2} Z])^3\} = E(e_1^3) + E(e_2^3),$$

where $e_j = z_j - Z/2$, $j = 1$ or 2, which are nearly independent with mean 0.

Observe that

$$\begin{aligned} z_1^3 + z_2^3 &= (\tfrac{1}{2} Z + e_1)^3 + (\tfrac{1}{2} Z + e_2)^3 \\ &= (\tfrac{1}{2} Z)^3 + 3(\tfrac{1}{2} Z)^2 e_1 + 3(\tfrac{1}{2} Z) e_1^2 + e_1^3 + (\tfrac{1}{2} Z)^3 + 3(\tfrac{1}{2} Z\ 2)^2 e_2 + 3(\tfrac{1}{2} Z) e_2^2 + e_2^3 \end{aligned}$$

implies $E[z_1^3 + z_2^3] = (1/4)E(Z^3) + 3(\tfrac{1}{2} Z)E[(e_1 + e_2)^2] + E(e_1^3) + E(e_2^3)$,
while

$$\begin{aligned} (z_1 + z_2)^3 &= [(\tfrac{1}{2} Z + e_1) + (\tfrac{1}{2} Z + e_2)]^3 = [Z + (e_1 + e_2)]^3 \\ &= Z^3 + 3Z^2(e_1 + e_2) + 3Z(e_1 + e_2)^2 + e_1^3 + e_2^3 \end{aligned}$$

implies $E[(z_1 + z_2)^3] = E(Z^3) + 3Z\,E[(e_1 + e_2)^2] + E(e_1^3) + E(e_2^3)$.

As a result, $E[(4/3)(z_1^3 + z_2^3) - (1/3)(z_1 + z_2)^3] = E(e_1^3) + E(e_2^3) + Z\,E[(e_1 + e_2)^2]$.

Replacing z_j with z_{hj} from Section 2, e_j with e_{hj}, and Z with Z_h, where h and j respectively denote variance strata and PSUs, we have the near equality ("near" because the z_{hj} are nearly unbiased for ½ Z_h when H is large),

$$E\left[\frac{4}{3}\sum_{h=1}^{H}(z_{h1}^3 + z_{h2}^3) - \frac{1}{3}\sum_{h=1}^{H}\left(z_{h1} + z_{h2}\right)^3\right] \approx \sum_{h=1}^{H}[E(e_{h1}^3) + E(e_{h2}^3)] + E\left[\sum_{h=1}^{H} Z_h\left(e_{h1} + e_{h2}\right)^2\right].$$

The right-hand size of the above near equality is approximately

$$\sum_{h=1}^{H}[E(e_{h1}^3) + E(e_{h2}^3)] \approx \sum_{h=1}^{H} E[(z_{h1} + z_{h2} - Z_h)^3] \text{ when } \sum_{h=1}^{H} Z_h\left(e_{h1} + e_{h2}\right)^2 \approx 0. \text{ (Recall } \sum_{h=1}^{H} Z_h = 0.)$$

The $z_h = z_{h1} + z_{h2}$ are nearly independent, so $\sum_{h=1}^{H} E[(z_{h1} + z_{h2} - Z_h)^3]$ is a good measure of the third-central moment of r in equation (2) when $\sum_{h=1}^{H} Z_h\left(e_{h1} + e_{h2}\right)^2 \approx 0$. Obseve that the

expection of the last expression on the right is $\sum_{h=1}^{H} Z_h \left[\mathrm{Var}(e_{h1}) + \mathrm{Var}(e_{h2}) \right]$ or $\sum_{h=1}^{H} Z_h \left[\sigma_{h1}^2 + \sigma_{h2}^2 \right] = \sum_{h=1}^{H} Z_h \sigma_h^2$, where $\sigma_{hj}^2 = \mathrm{Var}(e_{hj})$, and $\sigma_h^2 = \sigma_{h1}^2 + \sigma_{h2}^2$.

Heuristically, if Z_h was positively correlated with σ_h^2 (neither the Z_h nor the σ_h^2 are random variables), then we would expect the bias of this method to be positive. We would expect the reverse if Z_h was negatively correlated with σ_h^2, and no bias if they were uncorrelated.

Estimating M_3 (when H is large and all $n_h = 2$) with m_3 in equation (2) treating all the PSUs as if they came from the same stratum; that is,

$$m_3'' = \frac{(2H)^2}{(2H-1)(2H-2)} \sum_{h=1}^{H} \sum_{i=1}^{2} (z_{hi} - \tfrac{1}{2H} z)^3, \text{ where } z = \sum_{h=1}^{H} z_h$$

Assuming z_{h1} and z_{h2} are nearly independent and nearly estimators of ½ Z_h, so that $z_h = z_{h1} + z_{h2}$ is a nearly unbiased estimator of Z_h. The third-central moment of r (from equation (2)) is nearly

$$\mathrm{E}\left[\left(\frac{\sum_{h=1}^{H} \sum_{i=1}^{2} y_{hi}}{\sum_{h=1}^{H} \sum_{i=1}^{2} x_{hi}} - R \right)^3 \right] = \mathrm{E}\left[\left(\frac{\sum_{h=1}^{H} \sum_{i=1}^{2} y_{hi} - R \sum_{h=1}^{H} \sum_{i=1}^{2} x_{hi}}{\sum_{h=1}^{H} \sum_{i=1}^{2} x_{hi}} \right)^3 \right]$$

$$\approx \mathrm{E}\left[\left(\frac{\sum_{h=1}^{H} \sum_{i=1}^{2} y_{hi} - r \sum_{h=1}^{H} \sum_{i=1}^{2} x_{hi}}{\sum_{h=1}^{H} \sum_{i=1}^{2} x_{hi}} \right)^3 \right]$$

(when H is large and $r - R = O_P(1/H^{1/2})$)

$$= \mathrm{E}\left\{ \left(\sum_{h=1}^{H} \sum_{i=1}^{2} z_{hi} \right)^3 \right\}$$

$$= \mathrm{E}\left\{ \left(\sum_{h=1}^{H} [Z_h + e_{h1} + e_{h2}] \right)^3 \right\}$$

$$\approx \sum_{h=1}^{H} [\mathrm{E}(e_{h1}^3) + \mathrm{E}(e_{h2}^3)],$$

where the $e_{hj} = z_{hj} - Z_h/2$, j = 1 or 2, are nearly independent with mean 0.

Now

$$z_{hi} - \frac{1}{2H} \sum_{h'=1}^{H} z_{h'} = \tfrac{1}{2} Z_h + \left(e_{hi} - \frac{\sum_{h'=1}^{H} (e_{h'1} + e_{h'2})}{2H} \right),$$

so that

$$\left(z_{hi}-\frac{1}{2H}\sum_{h'=1}^{H}z_{h'}\right)^3 =$$

$$\tfrac{1}{8}Z_h{}^3+\tfrac{3}{4}Z_h{}^2\left[e_{hi}-\frac{\sum_{h'=1}^{H}(e_{h'1}+e_{h'2})}{2H}\right]+\tfrac{3}{2}Z_h\left[e_{hi}-\frac{\sum_{h'=1}^{H}(e_{h'1}+e_{h'2})}{2H}\right]^2+\left[e_{hi}-\frac{\sum_{h'=1}^{H}(e_{h'1}+e_{h'2})}{2H}\right]^3,$$

and

$$\mathrm{E}(m_3'') = \mathrm{E}\left[\frac{(2H)^2}{(2H-1)(2H-2)}\sum_{h=1}^{H}\sum_{i=1}^{2}(z_{hi}-\tfrac{1}{2H}z)^3\right]\approx$$

$$\frac{(2H)^2}{4(2H-1)(2H-2)}\sum_{h=1}^{H}Z_h{}^3+\frac{(2H)^2}{4(2H-1)(2H-2)}\left(\frac{3}{2}\right)\sum_{h=1}^{H}Z_h\left[\sigma_h^2\left(1-\tfrac{1}{H}\right)+\frac{\sum_{h'=1}^{H}\sigma_{h'}^2}{4H^2}\right]$$

$$+\sum_{h=1}^{H}[\mathrm{E}(e_{h1}{}^3)+\mathrm{E}(e_{h2}{}^3)]..$$

Consequently, if all the Z_h were 0, m_3'' would be nearly unbiased. Otherwise, it has two bias terms. When H is large, the first is nearly $\sum_{h=1}^{H}Z_h{}^3$, which can be positive, negative, or zero, while the second is nearly 3/2 times the bias of m_3', which also can be positive, negative, or zero. There is no guarantee that m_3'' has the greater absolute bias, but it seems likely.

Estimating M_3 (when H is large and all $n_h = 2$) when H is even by randomly pairing the strata and then treating each pair as a single stratum when computing m_3 in equation (2).

Let $L = H/2$ be the number of stratum pairs, and h and h' denote the two strata in a particular pair. The contribution to M_3 coming from the pair (our estimation target is)

$$C \approx \mathrm{E}(e_{h1}{}^3)+\mathrm{E}(e_{h2}{}^3)+\mathrm{E}(e_{h'1}{}^3)+\mathrm{E}(e_{h'2}{}^3).$$

The question we need to address is how good an estimator for C is

$$c=\frac{\sum_{i=1}^{2}\left\{\left[4z_{hi}-(z_h+z_{h'})\right]^3+[4z_{h'i}-(z_h+z_{h'})]^3\right\}}{24}\ .$$

Obseve that $4z_{hi}-(z_h+z_{h'})=2(Z_h-Z_{h'})+\left[4\varepsilon_{hi}-(\varepsilon_{h1}+\varepsilon_{h2}+\varepsilon_{h'1}+\varepsilon_{h'2})\right]$, so, for example,

$$\mathrm{E}\{[4z_{h1}-(z_h+z_{h'})]^3\}=$$
$$8(Z_h-Z_{h'})^3+$$
$$6(Z_h-Z_{h'})\left(9\sigma_{h1}^2+\sigma_{h2}^2+\sigma_{h'1}^2+\sigma_{h'2}^2\right)+\left[27\mathrm{E}(e_{h1}{}^3)-\mathrm{E}(e_{h2}{}^3)-\mathrm{E}(e_{h'1}{}^3)-\mathrm{E}(e_{h'2}{}^3)\right].$$

When the PSUs *h*2, *h*'1, or *h*'2 replace *h*1, the results are analogous, so that a little work reveals

$$E(c) = 2(Z_h - Z_{h'})\left(\sigma_h^2 - \sigma_{h'}^2\right) + C.$$

There is a potential bias coming from this pair of strata only when $Z_h \neq Z_{h'}$. The bias could be in either direction. It need not be in the same direct for every pair.

Heuristically, suppose Z_h was positively or negatively correlated with σ_h^2 (again, neither the Z_h nor the σ_h^2 are random variables), while Z_h' is uncorrelated with σ_h^2 because of the randomness of the pairing. It is then easy to see that the bias of this method is twice that of the proposed method.

References

Anderson, P. and Nerman, O. (2000). A balanced adjusted confidence interval procedure applied to finite population sampling, Presented at the Second International Conference of Establishment Surveys, Buffalo, NY.

Brown, L., Cai, T. and Dasgupta,A. (2001). Interval estimation for a binomial proportion. *Statistical Science,* 16, 101–133.

Cochran, W. (1977). *Sampling Techniques*, third edition. New York: John Wiley and Sons.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6, 65–70.

Franco, C., Little, R., Louis, T., and Slud, E. (2019). Comparative study of confidence intervals for porportions in complex surveys. *Journal of Survey Statistics and Methodology*, 7, 334-364.

Korn, E. and Graubard, B. (1998). Confidence interval for proportions with small expected number of positive counts estimated from survey data. *Survey Methodology*, 1030-1039.

Korn, E. and Graubard, B. (1999). *Analysis of Health Surveys.* New York: John Wiley and Sons.

Kott, P. (2017). A note on Wilson coverage intervals for proportions estimated from complex
Samples. *Survey Methodology*, 43, 235-240.

Kott, P. and Liu, Y. (2009). One-sided coverage intervals for a proportion estimated from a stratified simple random sample. *International Statistical Review,* 77, 251–265.

Kott, P. and Liu, Y. (2010). Speeding up the asymptotics when constructing one-sided coverage intervals with survey data. *Metron* 63, 137–151.

Liu, Y. and Kott, P. (2009). Evaluating alternative one-sided coverage intervals for a proportion. *Journal of Official Statistics,* 25, 569–588.

Research Triangle Institute (2012). *SUDAAN Language Manual*, Volumes 1 and 2, Release 11. Research Triangle Park, NC: Research Triangle Institute.