Inference about Age-standardized Rates with Sampling Errors in the Denominators

JIMING JIANG[1], ERIC J. FEUER[2], YUANYUAN LI[1], THUAN NGUYEN[3], AND MANDI YU[2]
University of California, Davis, USA[1], *National Cancer Institute*, USA[2], *and Oregon Health & Science University*, USA[3]

Cancer incidence and mortality are typically presented as age-standardized rates. Inference about these rates become complicated when denominators involve sampling errors. We propose a bias-corrected rate estimator as well as its corresponding variance estimator that take into account sampling errors in the denominators. Confidence intervals are derived based on the proposed estimators as well. Performance of the proposed methods is evaluated empirically based on simulation studies. More importantly, advantage of the proposed method is demonstrated and verified in a real-life study of cancer mortality disparity. A web-based, user-friendly computational tool is also in development at the National Cancer Institute to implement the new bias-corrected estimators for calculating ASRs of cancer mortality by immigration status. Finally, promise of proposed estimators to account for errors introduced by differential privacy procedures to the 2020 decennial census products is discussed.

Key Words. approximation, bias correction, cancer rates, mortality rates, Poisson, variance estimation, sampling error.

1 Introduction: Background and motivation

Despite the importance roles that cancer incidence and mortality rates play in monitoring progresses against cancer (e.g.,[1,2]), inferential method of age-standardized rate (ASR) by risk factors, for which population denominators can only be estimated from

sample surveys, such as immigration status [3], cancer screening [4], and smoking status [5], is lacking. The existing method, first formulated by Brillinger [6] and later extended by Fay [7], is not applicable because sampling errors in denominators were not considered. Failure to incorporate sampling errors leads to underestimation of the variabilities of ASRs and, as a result, possible falsely positive differences. An intuitive approach to mitigate the impact is to aggregate cancer cases and populations by space, time, and/or demographics, so that sampling errors are negligible. However, it inevitably prohibits small population studies or studies of temporal/spatial variabilities.

An existing method has been used to produce official reports of cancer incidence and mortality rates for the past twenty years in the United States ([8]). It was developed under the assumptions that both numerators and denominators are collected from legally required registrations or censuses. Although both reflect exact values, various philosophical and conceptual reasons have led to treating them as random [6] because they can be regarded as a sample drawn in time from all times in which substantially the same conditions prevail [9]. The natural variability in the census population is negligible in size, thus is often ignored in the inference. However, as discussed in Kish [10], a distinction between natural variability (from a "superpopulation" or "inferential population", e.g.,[11]) and survey sampling variability (from a "finite population") needs to be made, especially when denominators are estimated from sample surveys. The inference then involves two nested steps. The first step infers census population totals from the sample based on survey sampling mechanisms, and the second step infers the inferential population from the census population totals based on natural variability theory (e.g.,[12]). This paper is the first to incorporate survey sampling errors in making inference about ASRs using such a two-step approach.

Notably, the estimation challenge presented here is different from the classic ratio estimation (e.g.,[13]). In the classic ratio estimation, numerator and denominator are usually collected jointly in the same process, for example, by asking two questions in the same survey questionnaire. The data are also often assumed to be available at individual level. In contrast, numerators and denominators of the ASRs are collected for different purposes through distinct mechanisms, and their data are often available only in aggregated form due to confidentiality constraints. Therefore, findings from the classic ratio estimation literature do not apply.

Our current research was primarily motivated by a recent empirical study by Pinheiro et al. [14]. This study evaluated immigration disparities in cancer mortality rates by comparing foreign-born Hispanics with US-born Hispanics in California and Texas. Five-year mortality rates were computed by pooling deaths from 2008 to 2012. Population characteristics by immigration status for the same time period were estimated from the American Community Survey (ACS). We later use this as a case study to demonstrate the advantage of our new method.

In this paper, we propose a bias-corrected point estimator of ASR under the two-step inferential framework. A variance estimator is also developed. Although we demonstrate the development using mortality as an example, the new estimators can be applied to calculate cancer incidence rates and corresponding variances. The remaining part of the paper is organized as follows. After some preliminary discussion of the rate estimator's variance without bias-correction (i.e., the existing method), the bias-corrected rate estimator and the associated variance estimator are presented in Section 2. In Section 3 we demonstrate performance of the proposed estimators via simulation studies, and compare the new method with the existing method. Section 4 is

a highlight of this paper, in which we demonstrate the advantage of our method using the empirical study on immigration disparities in cancer mortality [14] as a case study. Note that this is a case where we know the ground truth, and therefore can verify the results; such an empirical evaluation is considered more important than showing the advantage of the proposed method in simulation studies. Some concluding remarks are offered in Section 5. Additional technical derivations are provided in the Appendix.

2 Bias-corrected ASR with measure of uncertainty

2.1 Preliminary

Let X_j and N_j denote the census-based death count (numerator) and person-year total (denominator), respectively, for age group j, $j = 1, \ldots, J$. The ASR is an age-weighted sum of death count divided by person-year total at-risk, that is,

$$\hat{R} \;=\; \sum_{j=1}^{J} w_j \frac{X_j}{N_j}, \tag{1}$$

where w_j is the age adjustment fraction for age group j so that and $\sum_{j=1}^{J} w_j = 1$. The age-standardization is a feature that permits comparisons of populations with different age distributions. The value X_j is collected from the National Vital Statistics System (NVSS); thus, it is regarded as the census count of all deaths that have occurred in a given area during a given time period. To reflect the stochastic process of lifetime and disease, X_j is subject to Poisson random errors. It is further assumed that N_j is a census count of lived population and thus a fixed quantity. However, in reality, N_j may also be subject to natural variability (albeit small in magnitude). Assuming that N_j is fixed (i.e., ignoring the natural variability), $1 \leq j \leq J$, a variance estimator of

the ASR is given by

$$V = \sum_{j=1}^{J} w_j^2 \frac{X_j}{N_j^2}. \tag{2}$$

Confidence interval based on modified gamma distribution was developed in [15].

However, when the denominator N_j is not available from the census, but instead estimated from a sample survey and thus subject to sampling error, formula (2) is no longer valid. To reflect this difference, we use $\hat{N}_j$ to denote the sample estimator of N_j. Using standard asymptotic techniques (e.g.,[15]), a variance of the ASR (1) can be obtained. Some regularity conditions are required for the approximation. See Appendix for detail. To state the result, first consider the rate estimator for a single age group, $\hat{R} = X/\hat{N}$. Let $\mathcal{P}$ denote the finite population with size N_{P}, and N_A the subpopulation size for age group A. Then, an estimator of $\mathrm{var}(\hat{R})$ is given by

$$\hat{V} = \frac{\hat{R}}{\hat{N}} + \left(\frac{\hat{R}}{\hat{N}}\right)^2 \widehat{\mathrm{var}}(\hat{N}|\mathcal{P}), \tag{3}$$

where $\widehat{\mathrm{var}}(\hat{N}|\mathcal{P})$ is an estimator of $\mathrm{var}(\hat{N}|\mathcal{P})$. Examples are given in Appendix.

The variance estimator for a single age group can be easily extended to the variance estimator of $\hat{R} = \sum_{j=1}^{J} w_j \hat{R}_j$, where $\hat{R}_j = X_j/\hat{N}_j$ is the rate estimator for the jth age group. Assuming that rate estimators for different age groups are independent (e.g.,[6]), the variance estimator for $\hat{R}$ is given by

$$\hat{V} = \widehat{\mathrm{var}}(\hat{R}) = \sum_{j=1}^{J} w_j^2 \hat{V}_j, \tag{4}$$

where $\hat{V}_j$ is given by the right side of (3) with $\hat{N}, \hat{R}$ replaced by $\hat{N}_j, \hat{R}_j$, respectively.

2.2 Bias-corrected ASR

The variability in the denominators not only complicates the variance estimation, it also increases the bias of the ASR. Specifically, when the denominators are subject to variability, the ASR (1) is not unbiased, even approximately. The bias can be reduced, using the following bias-corrected estimator. From now on, let R_j denote the true rate, N_j the census count of the finite population size for age group j, and $\hat{N}_j$ the estimated $N_j, 1 \leq j \leq J$. The ASR, $\hat{R}$, will now be understood as (1) with N_j replaced by $\hat{N}_j, 1 \leq j \leq J$. We assume that data from different age groups are independent. Note that this assumption means that the samplings from different age groups are independent, not that the population totals from different age groups, if considered as random variables, are independent conditional on the total population (the population totals from different age groups are, of course, negatively correlated given the population total, because the age group totals add up to the population total). Furthermore, we assume that the following hold for any $1 \leq j \leq J$:

(i) $X_j|N_j \sim \text{Poisson}(R_jN_j)$;

(ii) $\text{E}(\hat{N}_j|N_j) = N_j$, $\text{var}(\hat{N}_j|N_j) = V_j$, which can be consistently estimated by $\hat{V}_j$;

(iii) conditional on N_j, X_j and $\hat{N}_j$ are independent.

A main goal is to derive a bias-corrected rate estimator. We do this for the jth age group separately, then combine the results. Note that, by assumptions (i)–(iii), we have

$$\text{E}(\hat{R}_j) = \text{E}\left\{\text{E}(X_j|N_j)\text{E}\left(\frac{1}{\hat{N}_j}\middle| N_j\right)\right\} = \text{E}\left\{R_jN_j\text{E}\left(\frac{1}{\hat{N}_j}\middle| N_j\right)\right\}. \tag{5}$$

Next, by an elementary expansion of [16], we have

$$\frac{1}{\hat{N}_j} \approx \frac{1}{\text{E}(\hat{N}_j|N_j)} - \frac{\hat{N}_j - \text{E}(\hat{N}_j|N_j)}{\{\text{E}(\hat{N}_j|N_j)\}^2} + \frac{\{\hat{N}_j - \text{E}(\hat{N}_j|N_j)\}^2}{\text{E}(\hat{N}_j|N_j)\}^3}, \tag{6}$$

where $\approx$ is in the sense that the remaining term is of lower order than the last term. It follows, from (6), that

$$\begin{aligned} \mathrm{E}\left(\frac{1}{\hat{N}_j}\middle| N_j\right) &\approx \frac{1}{\mathrm{E}(\hat{N}_j|N_j)} + \frac{\mathrm{var}(\hat{N}_j|N_j)}{\mathrm{E}(\hat{N}_j|N_j)\}^3} + \text{lower order term} \\ &= \frac{1}{N_j} + \frac{V_j}{N_j^3} + \text{lower order term.} \end{aligned} \tag{7}$$

Combining (5) with (7), we get

$$\begin{aligned} \mathrm{E}(\hat{R}_j) &= \mathrm{E}\left\{R_j\left(1+\frac{V_j}{N_j^2}\right)\right\} + \text{lower order term} \\ &= R_j + \mathrm{E}\left(\frac{R_j V_j}{N_j^2}\right) + \text{lower order term.} \end{aligned} \tag{8}$$

If we replace the R_j, V_j, and N_j inside the expectation on the right side of (8) by their consistent estimators, $\hat{R}_j = X_j/\hat{N}_j$, $\hat{V}_j$, and $\hat{N}_j$, respectively, the difference is of lower order than the second term on the right side of (8), that is,

$$\mathrm{E}\left(\frac{R_j V_j}{N_j^2}\right) = \mathrm{E}\left(\frac{\hat{R}_j \hat{V}_j}{\hat{N}_j^2}\right) + \text{lower order term.} \tag{9}$$

Combining (8) and (9), we get

$$\mathrm{E}(\hat{R}_j) = R_j + \mathrm{E}\left(\frac{\hat{R}_j \hat{V}_j}{\hat{N}_j^2}\right) + \text{lower order term,}$$

or, writing in another way,

$$\mathrm{E}\left(\hat{R}_j - \frac{\hat{R}_j \hat{V}_j}{\hat{N}_j^2}\right) = R_j + \text{lower order term.} \tag{10}$$

The bias correction can now be seen by comparing (8) and (10). Namely, define

$$\hat{R}_{\mathrm{bc},j} = \hat{R}_j - \frac{\hat{R}_j \hat{V}_j}{\hat{N}_j^2} = \hat{R}_j\left(1 - \frac{\hat{V}_j}{\hat{N}_j^2}\right). \tag{11}$$

Then, it is seen from (8) and (10) that

$$\text{bias}(\hat{R}_j) = \text{E}(\hat{R}_j) - R_j = \text{E}\left(\frac{R_j V_j}{N_j^2}\right) + \text{lower order term}, \tag{12}$$

$$\text{bias}(\hat{R}_{\text{bc},j}) = \text{E}(\hat{R}_{\text{bc},j}) - R_j = \text{lower order term}, \tag{13}$$

where the lower-order terms are of lower order than the first term on the right side of (12). Therefore, $\hat{R}_{\text{bc},j}$ has a lower-order bias than $\hat{R}_j$. It follows that

$$\hat{R}_{\text{bc}} \;\; = \;\; \sum_{j=1}^{J} w_j \hat{R}_{\text{bc},j} \tag{14}$$

has a lower-order bias than $\hat{R}$.

The bias-correction performance of $\hat{R}_{\text{bc}}$, in comparison with that of $\hat{R}$, will be evaluated in Sections 3 and 4.

2.3 Variance estimator

We now consider variance estimation for $\hat{R}_{\text{bc}}$. Again, first consider a single age group. Define $b_j = V_j/N_j^2$, and $\hat{b}_j = \hat{V}_j/\hat{N}_j^2$. Note that b_j is the square of the coefficient of variation (c.v.) of $\hat{N}_j$. Then, we have

$$\text{var}(\hat{R}_{\text{bc},j}) = \text{E}\{\text{var}(\hat{R}_{\text{bc},j}|N_j)\} + \text{var}\{\text{E}(\hat{R}_{\text{bc},j}|N_j)\}. \tag{15}$$

We first argue that the second term on the right side of (15) is, typically, of lower order than the first term. This is because, under regularity conditions, we have

$$\text{E}(\hat{R}_{\text{bc},j}|N_j) = R_j + O(n^{-1}),\ \text{var}(\hat{R}_{\text{bc},j}|N_j) = O(n^{-1}),$$

where n is the sample size. Thus, under regularity conditions, the first term on the right side of (15) is $\text{E}\{O(n^{-1})\} = O(n^{-1})$; the second term on the right side of (15) is

$\mathrm{var}\{R_j + O(n^{-1})\} = \mathrm{var}\{O(n^{-1})\} = O(n^{-2})$. It follows that

$$\mathrm{var}(\hat{R}_{\mathrm{bc},j}) = \mathrm{E}\{\mathrm{var}(\hat{R}_{\mathrm{bc},j}|N_j)\} + \mathrm{lot}, \tag{16}$$

where, hereafter, lot stands for "lower-order term".

Next, we have, by (11) and the definition of $b_j, \hat{b}_j$ [see above (15)],

$$\hat{R}_{\mathrm{bc},j} = \hat{R}_j(1 - \hat{b}_j) = \hat{R}_j(1 - b_j) - \hat{R}(\hat{b}_j - b_j) = \hat{R}_j(1 - b_j) + \hat{R}_j o(b_j).$$

Note that $o(b_j)$ is of lower order than b_j. It follows that

$$\mathrm{var}(\hat{R}_{\mathrm{bc},j}|N_j) = \mathrm{var}\{\hat{R}_j(1 - b_j)|N_j\} + \mathrm{lot} = (1 - b_j)^2\mathrm{var}(\hat{R}_j|N_j) + \mathrm{lot}. \tag{17}$$

Combining (16), (17), we have

$$\mathrm{var}(\hat{R}_{\mathrm{bc},j}) = \mathrm{E}\{(1 - b_j)^2\mathrm{var}(\hat{R}_j|N_j)\} + \mathrm{lot}. \tag{18}$$

Furthermore, by (7), we have

$$\begin{aligned} \mathrm{E}(\hat{R}_j|N_j) &= \mathrm{E}(X_j|N_j)\mathrm{E}\left(\frac{1}{\hat{N}_j}\middle| N_j\right) \\ &= R_j N_j \left(\frac{1 + b_j}{N_j} + \mathrm{lot}\right) = R_j(1 + b_j) + \mathrm{lot}. \end{aligned} \tag{19}$$

Also, we have

$$\mathrm{E}(\hat{R}_j^2|N_j) = \mathrm{E}(X_j^2|N_j)\mathrm{E}\left(\frac{1}{\hat{N}_j^2}\middle| N_j\right) = R_j N_j(R_j N_j + 1)\mathrm{E}\left(\frac{1}{\hat{N}_j^2}\middle| N_j\right). \tag{20}$$

Finally, by (6), it can be derived that

$$\frac{1}{\hat{N}_j^2} \approx \frac{1}{N_j^2} - 2\frac{\hat{N}_j - N_j}{N_j^3} + 3\frac{(\hat{N}_j - N_j)^2}{N_j^4} - 2\frac{(\hat{N}_j - N_j)^3}{N_j^5} + \frac{(\hat{N}_j - N_j)^4}{N_j^6}.$$

It follows that

$$\mathrm{E}\left(\frac{1}{\hat{N}_j^2}\middle| N_j\right) = \frac{1}{N_j^2} + 3\frac{V_j}{N_j^4} + \mathrm{lot} = \frac{1 + 3b_j}{N_j^2} + \mathrm{lot}. \tag{21}$$

Combining (20), (21), we have

$$\mathrm{E}(\hat{R}_j^2|N_j) = R_j(R_j + N_j^{-1})(1 + 3b_j) + \mathrm{lot}. \tag{22}$$

Combing (19), (22), we have

$$\begin{aligned}
\mathrm{var}(\hat{R}_j|N_j) &= \mathrm{E}(\hat{R}_j^2|N_j) - \{\mathrm{E}(\hat{R}_j|N_j)\}^2 \\
&= R_j(R_j + N_j^{-1})(1 + 3b_j) + \mathrm{Lot} - \{R_j(1 + b_j) + \mathrm{lot}\}^2 \\
&= R_j(R_j + N_j^{-1})(1 + 3b_j) - R_j^2(1 + b_j)^2 + \mathrm{lot} \\
&= R_j(R_j b_j + N_j^{-1} + 3N_j^{-1}b_j - R_j b_j^2) + \mathrm{lot}. \qquad (23)
\end{aligned}$$

Combining (18) and (23), we obtain

$$\begin{aligned}
&\mathrm{var}(\hat{R}_{\mathrm{bc},j}) \\
= \;& \mathrm{E}\{R_j(1 - b_j)^2(R_j b_j + N_j^{-1} + 3N_j^{-1}b_j - R_j b_j^2)\} + \mathrm{lot} \\
= \;& \mathrm{E}\{\hat{R}_{\mathrm{bc},j}(1 - \hat{b}_j)^2(\hat{R}_{\mathrm{bc},j}\hat{b}_j + \hat{N}_j^{-1} + 3\hat{N}_j^{-1}\hat{b}_j - \hat{R}_{\mathrm{bc},j}\hat{b}_j^2)\} + \mathrm{lot}. \qquad (24)
\end{aligned}$$

(24) shows that an approximately unbiased estimator of $\mathrm{var}(\hat{R}_{\mathrm{bc},j})$ is

$$\widehat{\mathrm{var}}(\hat{R}_{\mathrm{bc},j}) = \hat{R}_{\mathrm{bc},j}(1 - \hat{b}_j)^2(\hat{R}_{\mathrm{bc},j}\hat{b}_j + \hat{N}_j^{-1} + 3\hat{N}_j^{-1}\hat{b}_j - \hat{R}_{\mathrm{bc},j}\hat{b}_j^2). \tag{25}$$

Now combining different age groups, it follows that an approximately unbiased estimator of $\mathrm{var}(\hat{R}_{\mathrm{bc}})$, where $\hat{R}_{\mathrm{bc}}$ is given by (14), is

$$\hat{V}_{\mathrm{bc}} = \widehat{\mathrm{var}}(\hat{R}_{\mathrm{bc}}) = \sum_{j=1}^{J} w_j^2 \widehat{\mathrm{var}}(\hat{R}_{\mathrm{bc},j}), \tag{26}$$

where $\widehat{\mathrm{var}}(\hat{R}_{\mathrm{bc},j})$ is given by (25).

A large-sample confidence interval for $R = \sum_{j=1}^{J} w_j R_j$ (e.g.,[17]), based on the bias-corrected estimator, is given by

$$\left[\hat{R}_{\mathrm{bc}} - z_{\alpha/2}\sqrt{\hat{V}_{\mathrm{bc}}},\;\; \hat{R}_{\mathrm{bc}} + z_{\alpha/2}\sqrt{\hat{V}_{\mathrm{bc}}}\right].$$

In the next section, we evaluate performance of the bias-corrected rate estimator and its variance estimator via simulation studies.

3 Simulation studies

We carry out a series of real-data motivated simulation study based on information collected from ACS with $J = 19$ age groups, with the weights $w_j, 1 \leq j \leq J$ given in ([18]). The purpose of the simulation is to observe the effect of the bias correction as well as the accuracy of the variance estimation when the denominator, $\hat{N}_j$, is subject to various error sizes, different sizes of populations, and different range of cancer rates from more common to rare types of cancer deaths.

Let $N_{\rm p}$ denote the population size. Three different population sizes are considered: $N_{\rm p} = 10,000$, $50,000$ and $100,000$, covering a variety of practical situations ranging from smaller geographic regions such as county or district, to larger populations.

In addition to the varying population sizes, we also consider different ranges of true cancer rate, ranging from $r = 0.00005$ to $r = 0.001$. The range of r covers practical situations from rare cancer death with the ASR of 5 per 100,000 person years, such as Non-Hodgkin Lymphoma or liver cancer, to all cancer death with the ASR of 100 per 100,000 person years.

Let $N_j = [N_{\rm p} w_j], 1 \leq j \leq J-1$, where $[x]$ denotes the integer part of x, and $N_J = N_{\rm p} - \sum_{j=1}^{J-1} N_j$. We then fixed these $N_1, \ldots, N_J$. Let $\mathcal{P}_{\rm age} = \{a_1, \ldots, a_{N_{\rm p}}\}$, where $a_i = 1, 1 \leq i \leq N_1$, $a_i = 2, N_1 + 1 \leq i \leq N_1 + N_2$, $\ldots$, $a_i = J, N_1 \cdots + N_{J-1} + 1 \leq i \leq N_1 + \cdots + N_J = N_{\rm p}$ (in other words, the first N_1 elements of $\mathcal{P}_{\rm age}$ are 1, the next N_2 are 2, and so on, and the last N_J elements are J). For each $j = 1, \ldots, J$,

generate X_j from Poisson(rN_j) distribution (so that $R_j = r$).

To control the size of errors in $\hat{N}_j, 1 \le j \le J$, we simulate $\hat{N}_j$ from a distribution centered at N_j but with increasing variation in terms of coefficient of variation (CV). Specifically, let $\sigma_j = \rho N_j$, where $\rho = 0.05, 0.1, 0.15, 0.2, 0.25, 0.3$. Then, generate $\hat{N}_j$ as a random variable whose values are integers between $L_j = N_j - 3\sigma_j$ and $U_j = N_j + 3\sigma_j$ such that $\mathrm{P}(\hat{N}_j = L_j) = \mathrm{P}(\xi_j \le L_j + 0.5)$, $\mathrm{P}(\hat{N}_j = k) = \mathrm{P}(k - 0.5 \le \xi_j \le k + 0.5)$, $L_j + 1 \le k \le U_j - 1$, and $\mathrm{P}(\hat{N}_j = U_j) = \mathrm{P}(\xi_j > U_j - 0.5)$, where $\xi_j \sim N(N_j, \sigma_j^2)$. It follows that $\mathrm{P}(\hat{N}_j = L_j) = \Phi\left(\frac{0.5 - 3\sigma_j}{\sigma_j}\right)$, and, for $L_j + 1 \le k \le U_j - 1$,

$$\mathrm{P}(\hat{N}_j = k) = \Phi\left(\frac{k + 0.5 - N_j}{\sigma_j}\right) - \Phi\left(\frac{k - 0.5 - N_j}{\sigma_j}\right),$$

and $\mathrm{P}(\hat{N}_j = U_j) = 1 - \Phi\left(\frac{3\sigma_j - 0.5}{\sigma_j}\right)$, where $\Phi(\cdot)$ is the cdf of $N(0,1)$. As for the $\hat{V}_j$ involved in (11), it can be shown that

$$V_j = \mathrm{var}(\hat{N}_j | N_j) = \sum_{k=L_j}^{U_j} k^2 \mathrm{P}(\hat{N}_j = k) - \left\{ \sum_{k=L_j}^{U_j} k \mathrm{P}(\hat{N}_k = k) \right\}^2, \tag{27}$$

where $\mathrm{P}(\hat{N}_j = k)$ is given above for $L_j \le k \le U_j$. Thus, $\hat{V}_j$ is given by (27) with N_j replaced by $\hat{N}_j$ and σ_j by $\rho \hat{N}_j$.

We repeat the simulation $K = 10,000$ times, and compute $\mathrm{E}(\hat{R}) = K^{-1} \sum_{k=1}^{K} \hat{R}_{[k]}$ and $\mathrm{E}(\hat{R}_{\mathrm{bc}}) = K^{-1} \sum_{k=1}^{K} \hat{R}_{\mathrm{bc},[k]}$, where $\hat{R}_{[k]}$ and $\hat{R}_{\mathrm{bc},[k]}$ are $\hat{R}$ and $\hat{R}_{\mathrm{bc}}$ from the kth replication, respectively, $1 \le k \le K$. The performance measure is percentage relative bias (%RB), where $\%\mathrm{RB}$ of $\hat{R} = 100 \times [\{\mathrm{E}(\hat{R}) - r\}/r]$, and that of $\hat{R}_{\mathrm{bc}}$ is defined similarly. Note that, because $R_j = r, 1 \le j \le J$, the true $R = \sum_{j=1}^{J} w_j r = r$.

The results are presented in Tables 1–3. Specifically, the left halves of Table 1 and Table 2 report the %RB of $\hat{R}$ and $\hat{R}_{\mathrm{bc}}$ for the cases of $r = 0.001$ and $r = 0.0002$, respectively, and varying population sizes; the left half of Table 3 reports the correspond-

ing results for $r = 0.0001$ and $r = 0.00005$ under the population size $N_{\rm p} = 100,000$. It is seen that the %RB of $\hat{R}_{\rm bc}$ is almost always smaller, and in most cases much smaller (in absolute value) than that of $\hat{R}$, indicating significant effect of bias reduction by $\hat{R}_{\rm bc}$ over $\hat{R}$. The performance of both $\hat{R}$ and $\hat{R}_{\rm bc}$ gets worse as ρ increases; on the other hand, the performance does not seem to be affected by the change of population size. Overall, the %RB of $\hat{R}_{\rm bc}$ stays in low single-digit in all cases considered.

Next we consider variance estimation for the bias-corrected ASR. Continuing with the above simulation setting, we study performance of the variance estimator given by (25), (26). The results are presented in the right halves of Tables 1–3, where %RB and CV for the variance estimation are defined, respectively, as

$$\%\mathrm{RB} = 100 \times \left[\frac{\mathrm{E}\{\widehat{\mathrm{var}}(\hat{R}_{\rm bc})\} - \mathrm{var}(\hat{R}_{\rm bc})}{\mathrm{var}(\hat{R}_{\rm bc})} \right], \ \mathrm{CV} = \frac{\sqrt{\mathrm{var}\{\widehat{\mathrm{var}}(\hat{R}_{\rm bc})\}}}{\mathrm{E}\{\widehat{\mathrm{var}}(\hat{R}_{\rm bc})\}}$$

with $\mathrm{E}\{\widehat{\mathrm{var}}(\hat{R}_{\rm bc})\}$, $\mathrm{var}\{\widehat{\mathrm{var}}(\hat{R}_{\rm bc})\}$, and $\mathrm{var}(\hat{R}_{\rm bc})$ evaluated based on the simulation replicates. It is seen that the %RB of the variance estimator stays in single-digit or low double-digit, which is generally considered satisfactory in terms of the bias. On the other hand, the CV of the variance estimator seems to be mixed, ranging from 0.11 to 2.25. Overall, the performance of the variance estimator seems to get worse as ρ increases, but it does not seem to be affected by the change in population size.

In the variance estimation, it is observed that the CV increases with ρ. This is reasonable because ρ is a parameter that controls the variation in the design-based estimators involved in our estimates, with larger ρ corresponding to larger variance. On the the hand, the relationship between %RB and ρ appears to be more complicated. Note that %RB is a measure of bias; an estimator can have a larger variance, and yet smaller (relative) bias at the same time. Other factors, such as the population size and

true rate, can play bigger roles in %RB than in CV in the variance estimation.

4 Case study: Immigration disparities in cancer mortality rates

As noted, the present study was inspired by a recent empirical study by [14], described in Section 1. Our purpose, however, is not to replicate those results, but rather to demonstrate use of the proposed bias-corrected rate estimators to improve inferential validities and increase the granularity of important cancer research.

In the present study, we compute annual ASRs of all-cancer-cause mortality for foreign-born Hispanics and US-born Hispanics from 2006 to 2013 for the states of California, Texas, and New Mexico. We add New Mexico to demonstrate the impact in areas with smaller populations. All rates are per 100,000 person-years and are age-standardized to the 2000 US standard population by 5-year age group with the last group being 85 and older. Cancer sites were coded according to the International Statistical Classification of Diseases (10th revision). Annual populations of Hispanics by 5-year age group, gender, immigration status (US-born vs. foreign-born) for California, Texas, and New Mexico are estimated using one-year ACS samples from 2006 to 2013. Sampling errors of population estimates are estimated using replicates weights.

Table 4 shows the distributions of estimated populations of Hispanics and CVs of these estimates across all 18 age groups by immigration status for California, Texas, and New Mexico. Note that CV is a precision measure and is computed as the standard error divided by the population estimate. California has the largest Hispanic populations among all three states, estimated to range from approximately 25,000 to

Table 1: **Bias of $\hat{R}$ and $\hat{R}_{\rm bc}$ and Estimation of $\mathrm{var}(\hat{R}_{\rm bc})$:** $r = 0.001$

		Bias		Variance Estimation			
$N_{\rm p}$	ρ	%RB of $\hat{R}$	%RB of $\hat{R}_{\rm bc}$	$\mathrm{var}(\hat{R}_{\rm bc})$	$\mathrm{E}\{\widehat{\mathrm{var}}(\hat{R}_{\rm bc})\}$	%RB	CV
100K	0.05	0.29	0.04	1.02×10^{-8}	1.03×10^{-8}	0.74	0.11
100K	0.10	1.03	0.02	1.08×10^{-8}	1.10×10^{-8}	2.19	0.12
100K	0.15	2.40	0.10	1.19×10^{-8}	1.23×10^{-8}	3.62	0.15
100K	0.20	4.51	0.35	1.41×10^{-8}	1.41×10^{-8}	0.28	0.19
100K	0.25	8.05	1.33	1.80×10^{-8}	1.69×10^{-8}	-6.11	0.28
100K	0.30	13.87	3.67	3.51×10^{-8}	2.31×10^{-8}	-34.29	0.82
50K	0.05	0.08	-0.17	2.02×10^{-8}	2.03×10^{-8}	0.65	0.15
50K	0.10	1.04	0.03	2.09×10^{-8}	2.14×10^{-8}	2.37	0.16
50K	0.15	2.47	0.17	2.24×10^{-8}	2.32×10^{-8}	3.68	0.18
50K	0.20	4.61	0.45	2.46×10^{-8}	2.59×10^{-8}	5.09	0.22
50K	0.25	7.97	1.25	2.90×10^{-8}	3.01×10^{-8}	3.49	0.31
50K	0.30	14.37	4.13	4.91×10^{-8}	4.06×10^{-8}	-17.34	0.86
10K	0.05	0.27	0.04	1.02×10^{-7}	1.01×10^{-7}	-0.74	0.32
10K	0.10	1.07	0.07	1.03×10^{-7}	1.05×10^{-7}	1.89	0.33
10K	0.15	2.42	0.13	1.05×10^{-7}	1.11×10^{-7}	5.85	0.35
10K	0.20	4.91	0.74	1.11×10^{-7}	1.21×10^{-7}	8.52	0.39
10K	0.25	7.84	1.13	1.17×10^{-7}	1.36×10^{-7}	15.87	0.48
10K	0.30	14.14	3.92	1.59×10^{-7}	1.77×10^{-7}	11.59	1.16

Table 2: **Bias of $\hat{R}$ and $\hat{R}_{\rm bc}$ and Estimation of $\mathrm{var}(\hat{R}_{\rm bc})$:** $r = 0.0002$

		Bias		Variance Estimation			
$N_{\rm p}$	ρ	%RB of $\hat{R}$	%RB of $\hat{R}_{\rm bc}$	$\mathrm{var}(\hat{R}_{\rm bc})$	$\mathrm{E}\{\widehat{\mathrm{var}}(\hat{R}_{\rm bc})\}$	%RB	CV
100K	0.05	0.28	0.03	2.05×10^{-9}	2.03×10^{-9}	-1.13	0.23
100K	0.10	1.00	-0.00	2.02×10^{-9}	2.10×10^{-9}	4.21	0.24
100K	0.15	2.40	0.10	2.12×10^{-0}	2.24×10^{-9}	5.82	0.26
100K	0.20	4.57	0.41	2.19×10^{-9}	2.45×10^{-9}	11.90	0.29
100K	0.25	7.38	0.70	2.53×10^{-9}	2.77×10^{-9}	9.58	0.39
100K	0.30	14.07	3.86	3.64×10^{-9}	3.69×10^{-9}	1.26	1.00
50K	0.05	-0.48	-0.73	3.95×10^{-9}	4.02×10^{-9}	1.71	0.32
50K	0.10	0.58	-0.42	4.08×10^{-9}	4.17×10^{-9}	2.15	0.33
50K	0.15	2.61	0.31	4.26×10^{-9}	4.44×10^{-9}	4.23	0.35
50K	0.20	4.93	0.76	4.50×10^{-9}	4.82×10^{-9}	6.95	0.39
50K	0.25	8.21	1.48	4.75×10^{-9}	5.44×10^{-9}	14.33	0.49
50K	0.30	14.38	4.14	6.22×10^{-9}	7.05×10^{-9}	13.26	1.10
10K	0.05	-0.34	-0.57	1.96×10^{-8}	2.01×10^{-8}	2.73	0.71
10K	0.10	0.56	-0.44	2.01×10^{-8}	2.08×10^{-8}	3.23	0.73
10K	0.15	1.31	-0.96	2.09×10^{-8}	2.17×10^{-8}	3.96	0.77
10K	0.20	4.61	0.44	2.16×10^{-8}	2.37×10^{-8}	9.73	0.81
10K	0.25	7.49	0.80	2.32×10^{-8}	2.67×10^{-8}	14.98	1.00
10K	0.30	11.73	1.72	2.83×10^{-8}	3.41×10^{-8}	20.70	2.25

Table 3: **Bias and Variance Estimation:** $N_{\mathrm{p}} = 10^5$**;** $r = 0.0001$ **and** $r = 0.00005$

		Bias		Variance Estimation			
r	ρ	%RB of $\hat{R}$	%RB of $\hat{R}_{\mathrm{bc}}$	$\mathrm{var}(\hat{R}_{\mathrm{bc}})$	$\mathrm{E}\{\widehat{\mathrm{var}}(\hat{R}_{\mathrm{bc}})\}$	%RB	CV
0.0001	0.05	0.57	0.32	9.97×10^{-10}	1.02×10^{-9}	1.77	0.32
0.0001	0.10	1.46	0.45	1.01×10^{-9}	1.05×10^{-9}	4.52	0.33
0.0001	0.15	3.18	0.87	1.07×10^{-9}	1.12×10^{-9}	4.47	0.35
0.0001	0.20	4.37	0.22	1.12×10^{-9}	1.20×10^{-9}	7.50	0.40
0.0001	0.25	7.86	1.15	1.21×10^{-9}	1.36×10^{-9}	13.00	0.50
0.0001	0.30	13.35	3.20	1.58×10^{-9}	1.75×10^{-9}	0.29	1.20
0.00005	0.05	0.79	0.54	5.04×10^{-10}	5.08×10^{-10}	0.83	0.45
0.00005	0.10	0.67	-0.33	5.11×10^{-10}	5.20×10^{-10}	1.76	0.46
0.00005	0.15	1.79	-0.49	5.12×10^{-10}	5.47×10^{-10}	6.88	0.48
0.00005	0.20	5.16	0.98	5.38×10^{-10}	5.98×10^{-10}	11.14	0.52
0.00005	0.25	8.30	1.56	5.90×10^{-10}	6.69×10^{-10}	13.45	0.66
0.00005	0.30	13.43	3.27	7.41×10^{-10}	8.61×10^{-10}	16.17	1.46

1,300,000 across all age groups and years. All Californian population estimates are very precise with CVs less than 0.003. In comparison, Hispanic population in Texas are slightly smaller. The estimated populations range from 17,000 to 900,000 with CVs ranging from less than 0.02 to about 0.1. US-born Hispanics are about 2 times the sizes of foreign-born Hispanics. In New Mexico, although Hispanics make up almost 50% of the state overall population, only about 20% are foreign-born. In addition, Hispanic populations in New Mexico are much smaller compared to California and Texas. The estimated Hispanics populations in New Mexico range from about 500 to 75,000 with CVs ranging from about 0.1 to about 0.8$\sim$0.9 in certain age groups. A closer examination reveals that the high CVs mostly occur in those subpopulations corresponding to foreign-born and 0-4 years old (data not shown); the CVs for the remaining age groups are almost all below 0.3.

Table 5 compared the estimated rates and standard errors using the bias-corrected method and the simple-ratio method. As expected, in California, all estimated ASRs and variances are almost identical between the two methods. The results also suggest that mortality rates are stable for foreign-born Hispanics, but have decreased for US-born Hispanics over the study period. In Texas, bias-corrected rates are slightly lower than simple-ratio rates, and bias-adjusted variances are slightly higher than simple-ratio variances. This pattern is also as expected as the simple-ratio method overestimates the rates and underestimates the variance. However, differences are too small in magnitude to affect inferences of trends in immigration disparities. In both California and Texas, immigration disparities in mortality have increased from 2006 to 2013. In contrast, the effect of over-estimation is more pronounced among New Mexico Hispanics. Particularly in 2006, the mortality rate is artificially inflated by 25%

for foreign-born Hispanics, whereas the inflation is only 0.7% for US-born Hispanics. This striking differential impact by immigration status would have produced a spurious significant result suggesting that foreign-born Hispanic is at a higher risk dying from cancers than their US-born counterparts. Although the differential impact is not as pronounced in other years as in 2006, the tendency persists.

5 Concluding remarks

In this paper, we develop and evaluate a new inference method about ASRs for situations where population denominators involve sampling errors. This method, to the best of our knowledge, is the first in the cancer statistics literature that tackles the coexistence of errors that are unique to one of two competing theories of inference, that is, natural variability in the numerator according to the super-population model, and survey sampling variability in the denominator according to the finite population theory. When the sampling error is small (less than 10% of the denominator estimate), the simple ratio estimator, as implemented in the existing method, produces nearly unbiased results (relative bias less than 1% of the true ASR), and its variance estimator is approximately unbiased and reasonably reliable. The simple-ratio estimator is attractive because it is simple to calculate and has the same form as the standard ASR. However, the nontrivial bias limits its use in situations with moderate or large sampling errors. We developed a bias-corrected estimator of ASR and its variance estimator. The bias-corrected ASR is as accurate as the simple-ratio estimator when sampling errors are small, and it outperforms the simple-ratio estimator when sampling errors are moderate (less than 30% of the denominator estimates). The proposed variance estimator is

Table 4: **Distribution of Population Estimates of Hispanics and Corresponding Coefficient of Variations (CVs) Across 18 Age Groups by Immigration Status in California, Texas, and New Mexico, ACS 2006-2013**

		Foreign-born					US-born				
		Estimated Population			C.V.		Estimated Population			C.V.	
	Year	Total	Min	Max	Min	Max	Total	Min	Max	Min	Max
CA	2006	5,460,855	36,685	742,718	0.000	0.002	7,627,126	26,342	1,343,100	0.000	0.002
	2007	5,547,810	31,892	732,657	0.000	0.002	7,671,537	24,118	1,352,314	0.000	0.002
	2008	5,389,763	36,757	725,535	0.000	0.002	8,045,133	27,870	1,382,469	0.000	0.002
	2009	5,434,335	29,729	701,645	0.000	0.002	8,247,852	30,647	1,438,445	0.000	0.001
	2010	5,489,479	31,433	705,306	0.000	0.002	8,602,513	36,731	1,320,472	0.000	0.001
	2011	5,450,231	23,972	692,912	0.000	0.003	8,908,162	36,622	1,336,897	0.000	0.001
	2012	5,419,477	26,446	694,671	0.000	0.003	9,120,101	42,014	1,335,478	0.000	0.001
	2013	5,422,604	26,624	694,088	0.000	0.003	9,293,717	45,773	1,305,963	0.000	0.001
TX	2006	2,781,931	17,567	377,823	0.018	0.085	5,598,061	25,714	914,651	0.011	0.063
	2007	2,865,727	17,180	377,689	0.017	0.111	5,725,625	25,596	954,159	0.011	0.080
	2008	2,866,522	19,258	386,599	0.016	0.092	5,949,060	30,963	983,198	0.008	0.068
	2009	2,931,212	21,295	391,403	0.014	0.098	6,220,043	33,152	1,022,611	0.008	0.057
	2010	3,016,333	19,647	385,495	0.014	0.085	6,516,698	29,655	959,562	0.008	0.059
	2011	3,073,933	20,476	385,062	0.016	0.095	6,720,304	31,694	976,911	0.008	0.061
	2012	3,049,590	18,890	367,776	0.018	0.082	6,910,265	36,073	968,280	0.008	0.050
	2013	3,128,519	19,958	388,577	0.016	0.098	7,026,483	35,628	966,921	0.009	0.055
NM	2006	157,121	431	20,005	0.087	0.518	717,004	7,627	79,304	0.034	0.104
	2007	150,747	788	18,293	0.088	0.479	721,879	8,290	74,545	0.039	0.121
	2008	147,603	708	18,887	0.092	0.652	747,547	7,773	85,311	0.034	0.143
	2009	162,305	919	21,292	0.082	0.847	754,055	7,939	84,625	0.029	0.103
	2010	171,004	823	19,432	0.075	0.713	788,850	7,524	84,355	0.024	0.115
	2011	173,804	633	21,067	0.085	0.536	798,400	8,191	85,223	0.026	0.117
	2012	153,665	1,074	20,027	0.097	0.633	826,312	8,909	85,496	0.029	0.106
	2013	180,286	562	22,368	0.081	0.983	806,431	8,195	86,117	0.031	0.112

Table 5: **Comparisons of Age-standardized Rates of All Cancer Cause Mortality (Per 100,000 Person Years) Estimated using the Bias-Corrected Method and the Simple-Ratio Method for Hispanics by Immigration Status in California, Texas, and New Mexico, 2006-2013**

		Foreign-Born						US-Born						Difference	
		Bias-Corrected		Simple-Ratio				Bias-Corrected		Simple-Ratio				Bias-Corrrected	
	Year	ASR	SE	ASR	SE	Deaths	Population	ASR	SE	ASR	SE	Deaths	Population	Diff.	SE
CA	2006	141.4	4.3	141.4	4.3	4,133	5,460,855	105.7	4.3	105.7	4.3	3,498	7,627,126	35.7	6.1
	2007	141.4	4.2	141.4	4.2	4,320	5,547,810	117.4	4.8	117.4	4.8	3,446	7,671,537	24.0	6.4
	2008	134.1	3.9	134.1	3.9	4,396	5,389,763	106.9	4.2	106.9	4.2	3,671	8,045,133	27.2	5.7
	2009	140.9	4.0	140.9	4.0	4,617	5,434,335	99.4	3.8	99.4	3.8	3,822	8,247,852	41.5	5.5
	2010	132.4	3.6	132.4	3.6	4,806	5,489,479	91.4	3.5	91.4	3.5	3,906	8,602,513	40.9	5.0
	2011	135.1	3.5	135.1	3.5	5,234	5,450,231	84.5	3.2	84.5	3.2	3,930	8,908,162	50.6	4.7
	2012	137.8	3.4	137.8	3.4	5,287	5,419,477	89.4	3.2	89.4	3.2	4,097	9,120,101	48.5	4.7
	2013	143.9	3.4	143.9	3.4	5,574	5,422,604	91.4	3.3	91.4	3.3	4,126	9,293,717	52.4	4.7
TX	2006	124.8	7.9	125.5	6.1	2,053	2,781,931	117.5	5.9	117.8	4.6	3,566	5,598,061	7.3	9.8
	2007	128.1	8.0	128.7	6.2	2,101	2,865,727	123.8	6.6	124.3	4.8	3,620	5,725,625	4.3	10.4
	2008	124.8	7.2	125.5	5.5	2,258	2,866,522	108.9	5.2	109.2	4.1	3,722	5,949,060	16.0	8.9
	2009	126.2	7.0	126.7	5.5	2,365	2,931,212	105.5	4.7	105.7	3.9	3,759	6,220,043	20.7	8.4
	2010	127.4	6.6	127.8	5.4	2,496	3,016,333	115.6	5.4	115.9	4.3	4,095	6,516,698	11.7	8.5
	2011	131.4	7.2	132.0	5.4	2,662	3,073,933	105.2	4.9	105.4	4.0	4,103	6,720,304	26.2	8.7
	2012	134.0	6.3	134.4	5.2	2,855	3,049,590	103.5	4.3	103.7	3.7	4,223	6,910,265	30.5	7.7
	2013	145.9	7.4	146.4	5.7	2,930	3,128,519	99.5	4.2	99.7	3.6	4,363	7,026,483	46.3	8.5
NM	2006	196.5	66.1	247.1	56.8	115	157,121	91.8	7.0	92.4	6.4	823	717,004	104.7	66.5
	2007	132.6	33.9	153.5	30.6	155	150,747	101.3	8.3	102.1	7.0	825	721,879	31.3	34.9
	2008	171.4	44.3	194.7	37.1	152	147,603	99.1	8.0	100.1	6.9	828	747,547	72.3	45.0
	2009	120.2	27.2	129.3	23.1	151	162,305	111.4	8.9	112.2	7.7	821	754,055	8.7	28.6
	2010	142.1	29.7	151.8	24.4	170	171,004	105.1	9.0	105.8	7.6	834	788,850	37.1	31.0
	2011	130.3	33.8	152.4	29.0	182	173,804	93.8	7.5	94.5	6.6	876	798,400	36.6	34.6
	2012	92.4	15.5	96.0	14.2	167	153,665	97.2	7.1	97.8	6.3	966	826,312	-4.8	17.0
	2013	116.8	33.0	128.4	30.7	191	180,286	97.8	7.0	98.4	6.2	950	806,431	18.9	33.7

accurate. We have also observed that the confidence intervals constructed based on the rate and variance estimators achieve approximately nominal levels.

It is important to note that, with large sampling errors (more than 30% of the denominator estimates), the bias-corrected estimator can produce slightly biased results. This is due to omission of the third and higher-order terms in the asymptotic expansion (e.g., Jiang [16], p. 103). Fortunately, we anticipate few such applications as survey-based population estimates with low reliabilities are rarely useful in empirical analyses because of excessive random noise.

The new inferential method developed in this study do not require individual-level survey data for estimating the denominators. Thus, it can be easily applied to situations in which only aggregated population data are available. Individual-level sample data, if available, can be used to pre-calculate population denominators and corresponding sampling errors using survey statistical software to incorporate sampling features.

Although this development is motivated mostly by the need to estimate cancer incidence or mortality rates by immigration status, the setup is very generic, and the proposed methods can tremendously improve accuracies or granularities of cancer studies on a wide range of topics. For example, local plannings often requires cancer rates for small geographic areas (e.g., towns and cities), where population estimates are commonly derived from ACS. Incidence rates of cervical, uterine, or ovaries cancers corrected for hysterectomy requires population estimates by hysterectomy status from national health surveys. Furthermore, the proposed methods are not limited to cancer research; they can be applied to other types of diseases or issues related to health.

A web-based computational tool is also in development at the National Cancer Institute to implement the new bias-corrected estimators for calculating ASRs of can-

cer mortality separately for US-born and foreign-born Americans. This tool is user friendly and users can inquire annual ASRs of state-level mortality by cancer site, age, gender, race and ethnicity group, and immigration status for 2006 to 2014. The underlying annual population estimates stratified by immigration status are derived from the Integrated Public Use Microdata Series (IPUMS) developed and maintained by the he University of Minnesota ([19]). For more information about this tool and its availability timeline, interested users are encouraged to contact the authors.

Finally, our new method has a considerable promise as a solution to errors introduced to the publicly released 2020 US decennial census data products by differential privacy (DP) and post-processing (PP) procedures. Decennial census data is the primary data source of population denominators for generating official reports of cancer rates. DP is a new change made to 2020 decennial census enumeration data to protect confidentiality. Considerable concerns have been expressed about the impact of DP on data usability as noticeable discrepancies are observed between DP-modified 2010 demonstration census populations and enumerated 2010 populations. Drawn from our current study, it is not difficult to foresee the impact of DP errors on the accuracy and precision of cancer rates. Research has shed lights on similarities between sampling errors and DR errors in general settings. Our method has demonstrated the case of incorporating sampling errors by inferring decennial census populations from sample-based population estimates in the first step of our two-step inferences, and it holds great promise to be adapted to deal with DP errors instead. However, at the writing of this paper, the US Census Bureau has not provided a releasing plan for DP/PP errors for the 2020 decennial census data products. Undoubtedly, information about DP procedures and DP/PP errors is as important as DP-modified population estimates, as they

together are the building blocks to inferring the true decennial population totals from the DP-modified census population estimates. Further rigorous research is needed to advance the understanding of DP and, more importantly, to ensure continued thrive of cancer research in the era of DP protected census data.

References

[1] Haenszel, W, Loveland, DB and Sirken, MG. Lung-cancer mortality as related to residence and smoking histories. I. White Males. J Natl Cancer Inst 1962; 28: 947–1001.

[2] Extramural Committee To Assess Measures of Progress Against Cancer; Measurement of Progress Against Cancer, JNCI: J Natl Cancer Inst 1990; 82: 825–835.

[3] Banegas, MP, John, EM, Slattery, ML, Gomez, SL, Yu, M, LaCroix, AZ, Pee, D, Chlebowski, RT, Hines, LM, Thompson, CA, and Gail, MH. Projecting Individualized Absolute Invasive Breast Cancer Risk in US Hispanic Women. J Natl Cancer Inst 2017: 109: 2.

[4] DeSantis, C, Ma, J, Bryan, L and Jemal, A. Breast cancer statistics, 2013. CA: Cancer J Clin 2014; 64: 52–62.

[5] Wakelee, HA, Chang, ET, Gomez, SL, et al. Lung cancer incidence in never smokers. J Clin Oncol 2007; 25: 472–478.

[6] Brillinger, DR. A Biometrics Invited Paper with Discussion: The Natural Variability of Vital Rates and Associated Statistics. Biometrics 1986; 42: 693–712.

[7] Fay, R. E. Theory and application of replicate weighting for variance calculation. In Proc Survey Res Meth Sec. Alexandria, VA: American Statistical Association; 212–217.

[8] SEER Cancer Statistics Review, 1975-1993, National Cancer Institute. Bethesda, MD, https://seer.cancer.gov/archive/csr/1973_1993/

[9] Keyfitz, N. Sampling variance of standardized mortality rates. Hum biol 1966; 38: 309–317.

[10] Kish, L. Discussion on the Paper by David R. Brillinger: A Biometrics Invited Paper with Discussion: The Natural Variability of Vital Rates and Associated Statistics. Biometrics 1986; 42: 724–725.

[11] Korn, EL and Graubard, BI. Variance estimation for superpopulation parameters. Stat Sin 1998; 8: 1131–1151.

[12] Graubard, BI and Korn, EL. Inference for Superpopulation Parameters Using Sample Surveys. Stat Sci 2002; 17: 73–96.

[13] Cochran, WG. *Sampling techniques*. New York: Wiley, 1977.

[14] Pinheiro, PS, Callahan, KE, Gomez, SL, Marcos-Gragera, R, Cobb, TR, Roca-Barcelo, A and Ramirez, AG. High cancer mortality for US-born Latinos: evidence from California and Texas. BMC Cancer 2017; 17: 478.

[15] Singh, HP and Espejo, MR. On linear regression and ratio-product estimation of finite population mean. The Statistician 2003; 52: 59–67.

[16] Jiang, J. *Large Sample Techniques for Statistics*. New York: Springer, 2010: p. 103.

[17] Fuller, WA. *Sampling Statistics*. Hoboken, NJ: Wiley, 2009.

[18] 2000 standard populations: https://seer.cancer.gov/stdpopulations/stdpop.19ages.html

[19] Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Matthew Sobek. IPUMS USA: Version 10.0 [dataset]. Minneapolis, MN: IPUMS, 2020. https://doi.org/10.18128/D010.V10.0

[20] Lohr, SL. *Sampling: Design and Analysis*. 2nd ed. Brooks/Cole, 2010.

Appendix: Regularity conditions regarding Section 2.1, and examples

1. Regularity conditions:

A1. $\hat{N}$ is a design-unbiased estimator of N_A, that is, $\mathrm{E}(\hat{N}|\mathcal{P}) = N_A$.

A2. An estimator of $\mathrm{var}(\hat{N}|\mathcal{P})$, $\widehat{\mathrm{var}}(\hat{N}|\mathcal{P})$, is available such that $\widehat{\mathrm{var}}(\hat{N}|\mathcal{P}) \approx \mathrm{var}(\hat{N}|\mathcal{P})$ in the sense that the difference between the two sides of the $\approx$ is of lower order than the right side.

A3. $X|N_A \sim \mathrm{Poisson}(\lambda N_A)$, where λ is an unknown constant.

Assumptions *A1*, *A2* are design-based while assumption *A3* is model-based.

2. Examples.

Example 1 (SRS). Under simple random sampling (SRS), a standard estimator of the age-specific population total is $\hat{N} = N_{\mathrm{p}}\bar{y}$, where N_{p} is the population size (assumed known), $\bar{y} = n^{-1}\sum_{i\in S} y_i$, n is the sample size, S is the set of sampled indexes, and $y_i = 1_{(i\in A)}$, the indicator that index i belongs to the designated age A. Intuitively, $\bar{y}$ is the proportion of individuals in the sample that belong to age group A. It can be shown that a design based estimator of $\mathrm{var}(\hat{N}|\mathcal{P})$ is given by

$$\widehat{\mathrm{var}}(\hat{N}|\mathcal{P}) = \frac{N_{\mathrm{P}}\hat{N}(N_{\mathrm{P}} - \hat{N})}{N_{\mathrm{P}} - 1}\left(\frac{1}{n} - \frac{1}{N_{\mathrm{P}}}\right),$$

where n is the SRS sample size.

Example 2 (STR). In the case of stratified random sampling (STR), suppose that the population $\mathcal{P}$ is divided into H strata, $\mathcal{P}_h = \{y_{hj}, j = 1, \ldots, N_h\}$, $h = 1, \ldots, H$, where N_h is the population size for the hth stratum. Let $y_{hj}, j \in s_h$ be a SRS from the hth stratum, where s_h denotes the set of sampled indexes with $|s_h| = n_h$ ($|\cdot|$ denotes cardinality), so n_h is the sample size for the hth stratum. Then, a stratified estimator of the population total is given by $\hat{N} = \sum_{h=1}^{H} N_h\bar{y}_h$, where $y_h = n_h^{-1}\sum_{j\in s_h} y_{hj}$ is the sample mean for the hth stratum. For such an estimator $\hat{N}$, we have (e.g.,[20])

$$\mathrm{var}(\hat{N}|\mathcal{P}) = \sum_{h=1}^{H}\left(1 - \frac{n_h}{N_h}\right)N_h^2\frac{S_h^2}{n_h},$$

where S_h^2 is the population variance for the hth stratum. A design-unbiased estimator of $\mathrm{var}(\hat{N}|\mathcal{P})$ is thus given by

$$\widehat{\mathrm{var}}(\hat{N}|\mathcal{P}) = \sum_{h=1}^{H}\left(1 - \frac{n_h}{N_h}\right)N_h^2\frac{s_h^2}{n_h},$$

where $s_h^2 = (n_h - 1)^{-1}\sum_{j\in s_h}(y_{hj} - \bar{y}_h)^2$ is the sample variance for the hth stratum.

The above results are general. They apply, in particular, to the case where y_{hj} is the indicator of membership to age group A, that is, $y_{hj} = 1_{(a_{hj} \in A)}$, where a_{hj} denotes the age of the j individual in the hth stratum. In this case, the population total is $\sum_{h=1}^{H} \sum_{j=1}^{N_h} y_{hj} = \sum_{h=1}^{H} N_{A,h} = N_A$, where $N_{A,h}$ is the total number of individuals in stratum h that belong to age group A. It is easy to see that, in this case, $\bar{y}_h = \hat{p}_h = n_{A,h}/n_h$, where $n_{A,h}$ is the number of sampled individuals from stratum h that belong to age group A. Also, we have $s_h^2 = n_h(n_h - 1)^{-1}\hat{p}_h(1 - \hat{p}_h)$. The general formulae now become $\hat{N} = \sum_{h=1}^{H} N_h \hat{p}_h$, and

$$\widehat{\text{var}}(\hat{N}|\mathcal{P}) = \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) N_h^2 \frac{\hat{p}_h(1 - \hat{p}_h)}{n_h - 1}.$$