Population Review

Volume 58, Number 2, 2019

Type: Article, pp. 1-22

Estimating the Underlying Infant Mortality Rates for Small Populations, Including those Reporting Zero Infant Deaths: A Case Study of Counties in California

Authors: David A. Swanson, Augustine Kposowa and Jack Baker

Affiliations: Department of Sociology, University of California Riverside, Riverside, California, and Center for Studies in Demography and Ecology, University of Washington, Seattle, Washington (Swanson); Department of Sociology, University of California Riverside, Riverside, California (Kposowa); Health Fitness Corporation, Minneapolis, Minnesota (Baker)

Corresponding author/address: David A. Swanson, Department of Sociology, University of California Riverside, Riverside, California, USA 92521; email: dswanson@ucr.edu

Abstract

Infant mortality is an important population health statistic that is often used to make health policy decisions. For a small population, an infant mortality rate is subject to high levels of uncertainty and may not indicate the "underlying" mortality regime affecting the population. This situation leads some agencies to either not report infant mortality for these populations or report infant mortality aggregated over space, time or both. A method is presented for estimating "underlying" infant mortality rates that reflect the intrinsic mortality regimes of small populations. The method is described and illustrated in a case study by estimating IMRs for the 15 counties in California where zero infant deaths are reported at the county level for the period 2009-2011. We know that among these 15 counties there are 50 infant deaths reported at the state level but not for the counties in which they occurred. The method's validity is tested using a synthetic population in the form of a simulated data set generated from a model life table infant mortality rate, representing Level 23 of the West Family Model Life Table for both sexes. The test indicates that the method is capable of producing estimates that represent underlying rates. In this regard, the method described here may assist in the generation of information about the health status of small populations.

Keywords

Beta-binomial model, data confidentiality, health policy, underlying mortality regime, stochastic, superpopulation, uncertainty

Introduction

The infant mortality rate (IMR) is widely used. It is an indicator that can measure the risk of infant death and the availability/quality of health care services, poverty levels, and socio-economic status differentials (Hummer, 2005; Kitagawa and Hauser, 1973; Link and Phelan, 1995; Stockwell, Goza and Balisteri, 2005; Stockwell et al., 1987).[1]

Because statistical data are often used to guide health policy decisions, it is not surprising that the IMR also is used in this regard (Chen, Oster, and Williams, 2016; Kleinman, 1996; Misra et al., 2004; Stockwell et al., 1987). Moreover, as observed by VanEenwyk and Macdonald (2012), questions concerning health outcomes and related health behaviors and environmental factors often are studied within small subgroups of a population because many activities to improve health affect relatively small populations. Fortunately, the advent of geographic information systems and high volume, fast computer-based information systems – often involving the matching of records from different sources – means that this type of information is technically feasible. However, the demand for this information along with the technical feasibility of obtaining it is not always compatible with the need to preserve data confidentiality. This means that even when it is possible to provide data for a small population, it is not always the case that they are, in fact, provided, something not often encountered when dealing with large populations (Office for National Statistics, 2015). The Centers for Disease Control, for example, does not present or publish death or birth counts of nine or fewer or rates based on counts of nine or fewer (in figures, graphs, maps, table, etc.) at the subnational level (US National Center for Health Statistics, no date).

Data representing small populations are not only subject to limitations posed by confidentiality concerns, they are also subject to higher levels of uncertainty than those typically found in larger populations (Reeske and Razum, 2011; Swanson and Tayman, 2012: 216).[2] Along with confidentiality concerns, awareness of the effect of uncertainty on infant mortality and other rates associated with small populations leads to strategies aimed at limiting this source of variability. A typical strategy for dealing with the combination of these two issues is to aggregate data for small populations and generate what amounts to an arithmetic average from them.[3] Another strategy is to gain permission to access individual level records, match them and then construct statistics (Kinge and Kornstad, 2014). However, unlike the strategy of aggregation, this approach inevitably requires administrative approval and requires both a substantial amount of time and personnel costs to implement.

As a means of directly examining the issue of uncertainty (and indirectly the issue of confidentiality limitations) associated with data representing small populations, two empirical examples are used, one for an area with a large population and the other for an area with a small population. The large population example is based on the reported 2015 IMR for the State of Washington and the small population example is based on the 2015 IMR for Asotin County, which is one of the smallest counties in the State of Washington that reported at least one infant death in 2015. The estimated 2015 population for the State as a whole is 7,061,410 (Washington State Office of Financial Management, 2017: Table 3), with 89,000 live resident births (Washington Department of Health, 2016a) and 431 resident infant deaths reported, respectively, for 2015 (Washington Department of Health, 2016b). For Asotin County, the estimated 2015 population is 22,010 (Washington State Office of Financial Management, 2017: Table 3), with 228 live resident births (Washington Department of Health, 2016a) and three resident infant deaths reported,

respectively, for 2015 (Washington Department of Health, 2016b). Given these data, the 2015 IMR for the State as a whole is 4.8427 per 1,000 live resident births, while for Asotin County, it is 13.1579 per 1,000 live resident births. Note the striking difference, which leads to the question: Is the IMR for Asotin County really almost three times higher than that for the State as a whole or is it the result of uncertainty acting on a small population?

To answer this question, we start with work by Voss et al. (1995) and Swanson and Tayman (2012: 189-190) who viewed the crude death rate of a given area i at a given time t as the marginal probability of death for the area's inhabitants. Using this framework with infant deaths, we find the distribution of infant deaths in a given area i at a given time t is (approximately) binomial, with parameter d, where

$$d_{i,t} = D_{i,t}/B_{i,t} \quad [1]$$

where
i = area (i = 1 to n)
t = time
D = infant deaths
B = births

As this example progresses it will show the effect of the uncertainty in d (IMR) by using it in conjunction with the reported number of infant deaths in order to (hypothetically) estimate the number of births for the population in question. It is not the actual intent that we are proposing that IMR be used for this purpose; rather, the example is used to show the effect of uncertainty on small populations. To do this, we start by using statistical concepts discussed by Swanson and Tayman (2012: 29-42) to show that Equation [1] can be re-written so that the expected number of births at a given time t in area i is:

$$E[B_{i,t}] = D_{i,t} / d_{i,t} \quad [2]$$

The preceding equation leads to defining the variance of $B_{i,t}$:

$$V[B_{i,t}] = [D_{i,t} (1 - d_{i,t}) / d_{i,t}]^2 \quad [3]$$

Finally, the coefficient of variation (CV) for Bit is defined as

$$CV[B_{i,t}] = [(1 - d_{i,t}) / D_{i,t}]^{0.5} \quad [4]$$

Keep in mind that: (1) we view B as subject to uncertainty while D is not; and (2) because d = D/B it also is subject to the uncertainty associated with B.

As can be seen in Equation [4], the CV is defined as the ratio of the standard deviation to the mean. It is most useful for variables that are always positive, which is the case in the discussion here. In the case of IMR, as the number of births decreases, the size of the CV increases and a large CV indicates that uncertainty is large.

Using equation [4], we find for Washington as a whole that the coefficient of variation for the estimated number of births (using infant deaths and the infant death rate as estimators of the number of births, as shown in the following equation) is $0.04806 = [(1-0.00484)/(431)]^{0.5}$; for Asotin County, the coefficient of variation using its infant deaths and infant death rate as estimators is $0.57354 = [(1-0.01316)/(3)]^{0.5}$. The CV for Asotin County is nearly 12 times the size of the CV for the State as a whole, 11.93 = 0.57354/0.04806. Thus, we expect that over time, the relationship between the number of infant deaths and the number of births for the state of Washington as a whole is far more stable than the case for Asotin County. Put another way, the IMR for the State of Washington will be much more stable over time than the IMR for Asotin County, providing a much clearer view of the "underlying" IMR for the State than can be expected for Asotin County.

The difference in the CVs for the State of Washington and Asotin County illustrates the potential 'instability' inherent in small populations, which implies that reported IMRs for small populations can vary dramatically over time even though there is no substantive change in their respective "underlying" infant mortality rates. Awareness of this situation has led to two general approaches for dealing with the effect of uncertainty on IMRs for small populations. One approach is "non-reporting," which is to simply not report IMRs for small populations, as is the case with the Centers for Disease Control (US National Center for Health Statistics, no date). Unfortunately, this approach discards related information (e.g., reported births) that may be of use in estimating IMRs for small populations – a point to which we return later. Another general approach is to provide an estimate by embedding small population information within a "larger context," which takes us back to the "aggregation strategy" discussed earlier. This approach is used by, among other agencies, the US National Center for Health Statistics (2018), for which the "larger context" is defined both in terms of time and space. In terms of time, the NCHS data on infant mortality rates by county are aggregated for the period 2007-2015 and in terms of space, counties with small populations are aggregated.

One drawback to both approaches is that they typically yield simple arithmetic averages and neither is specific to the time and county of interest. Related to this issue is the fact that these averages are biased unless appropriate weights or other procedures are used to reduce bias (Voss et al., 1995), steps that may not be feasible in a given situation. Another "contextual" approach that we refer to as the "representational approach," is taken in this paper. It uses a beta-binomial model, which, unlike the "non-reporting" approach, has the potential to provide estimates of the IMRs underlying small populations, while also avoiding the drawbacks found in the aggregated approach.[4] Another benefit of this approach is that it is a statistical estimator and, as such, is not in conflict with confidentiality issues. To this end, a publication by Link and Hahn (1996) was used as a guide in generating the approach described, tested, and applied here.

California is used as a case study because infant mortality rates reported as zero were found in 15 of its 58 counties for 2009-11, a period over which the California Health & Human Services Agency (no date) aggregated infant deaths and births in order to develop IMRs.[5]

Exhibit 1. Map of California by County

Source: Geology.com (no date): https://geology.com/county-map/california.shtml

Method and data

Method

The method employed here is aimed at generating an estimate of the underlying IMR specific to a given small population for which the reported number of infant deaths is zero (or where the number is not reported due to confidentiality issues). As noted earlier, the method employs a distribution of IMRs taken from a "representational context" from which the two parameters of a "beta model" fitted to a binomial distribution made up IMRs (where the IMR is divided by 1,000) are used to develop an IMR estimate for a small population in conjunction with its reported births.[6] Even where it is the case that the population is so small that neither infant deaths nor births are reported, the two parameters may be used to develop an estimate of the underlying IMR. However, in such a case, the underlying estimated IMR is an "average" based on the parameters generated by fitting a beta model to a distribution of IMRs selected as representative of the small population(s) of interest. We now turn to a more formal definition.

Infant mortality rates measure the proportion of births that result in deaths during the first year of life. As such, they measure the relationship between events (deaths) and trials (births) with the distribution of infant deaths in a given area i at a given time t is (approximately) binomial, with parameter d, where, as shown in Equation [1] and is typically described as a beta-binomial random process with a probability mass function defined by two parameters: α and β. The first parameter, α, can be interpreted as the count of the event of interest, which in our case is the number of infant deaths, the number of births in which the infant dies before achieving the first year of life. The second parameter, β, can be interpreted as the count of "non-events," which in this paper is the number of children born who survive to reach one year of age. Note that "rate" = $\alpha/(\alpha + \beta)$, which in this paper is equivalent to" infant mortality rate" = infant deaths/(infant deaths + survivors to age 1), which reduces to infant deaths/births. Thus, parameter α is the numerator in the expression defining a rate, and when added together, the parameters α and β represent the denominator. Together, the IMR may be re-expressed the IMR as the compound distribution of α and β captured in the beta-binomial probability model:

$$\text{IMR} = \alpha/(\alpha + \beta) = \text{infant deaths}/(\text{infant deaths} + \text{infant survivors}) \quad [5]$$

Since the IMR may be conceptualized directly using the beta-binomial model, IMRs may be thought of as stochastic processes that occur within each county while also contributing to higher-level meta-populations within which they are nested (Karlin and Taylor, 1993; Graham and Talay, 2013).

A potential number of strategies exist for dealing with small sample size dynamics or confidentiality suppression in making estimates of infant mortality rates. First, one might simply use the national IMR in place of highly uncertain localized estimates of IMR. This would stabilize estimates for IMR on the local level, but at the expense of potentially masking heterogeneity in IMRs across geographic units. For purposes of capturing spatial patterns in IMR, a main priority in smaller-level analyses, this solution is less acceptable. A second alternative might be to make local adjustments based on judgment. While this may improve estimates overall, especially when judgments are made by applied demographers with significant experience, this approach is subject to the criticism that non-standard methods are applied across different geographies and/or population groupings. With resource allocation decisions often tied to demographic estimates, this solution may not be satisfactory either. An ideal approach would be to utilize a principled

method for adjusting local estimates of IMR. Simple model averaging, based on the beta-binomial model represents a viable approach for achieving this goal.

Because it has been established that the IMR constitutes a beta-binomial probability process, think of two estimates of this process as constituting samples of the mean and variance of the underlying process. Therefore, these can be considered as samples obtained from the same underlying mortality process and in averaging them it can be anticipated that a superior estimate of the mean proportion is obtained (Gardiner, 1983; Graham and Talay, 2013; Karlin and Taylor, 1993). As such, the averages of two estimates based on the model may also be averaged as:

$$IMR_{averaged} = (\alpha_1 + \alpha_2)/ ((\alpha_1 + \beta_1) + (\alpha_2 + \beta_2)) \quad [6]$$

where the subscripts (1,2) now represent estimates of death and survivorship counts for two groups. This method can, of course, be extended to k groups as desired. Such model averaging yields an estimate where a larger-scale and representationally-appropriate model IMR is leveraged to make smaller-scale estimates more precise in a manner similar to that observed in the literature on indirect estimation in demography (Brass, 1968; Moultrie et al., 2013; Siegel and Swanson 2004; UN, 1967). Recent attempts to extend indirect estimation based on stochastic process theory have been introduced (Baker et al., 2011), and here this idea is leveraged further in developing indirect estimates of IMR based on model averaging.

Before turning to a discussion of the data, it is appropriate here to discuss in some detail the averaging process just described. Because an IMR is typically expressed per 1,000 births, it can be turned into a binomial variable by dividing it by 1,000 (or more generally if IMR is expressed as infant deaths per *k* births, it would be divided by *k*). In this form, IMR is strictly bound in that it cannot be less than zero nor greater than ($0 \leq IMR \leq 1$). In practice, it is substantially less than one. Once in this form, a beta model (binomial) can be fitted to a distribution of IMRs, which when fitted, produces two estimated parameters, α and β. The first parameter, α, can be interpreted as the number of births in which the infant dies before achieving the first year of life. The second parameter, β, can be interpreted as the number of children born who survive to reach one year of age. Note that "rate" = $\alpha/(\alpha + \beta)$, which in our case is equivalent to" infant mortality rate" = infant deaths/(infant deaths + survivors to age 1), which reduces to infant deaths/births. Thus, parameter α is the numerator in the expression defining a rate, and when added together, the parameters α and β represent the denominator.

The two parameters estimated by fitting the beta model to a distribution of IMRs are then used to adjust the reported infant deaths (a) and survivors (b) for the population in question, even when either one or both is equal to zero. The adjustment is straightforward: adjusted IMR = $(a + \alpha)/ ((a + b) + (\alpha + \beta))$. Note, as stated earlier that if a = zero then the adjusted IMR = $\alpha/(b + \alpha + \beta)$ and that if both a and b are zero, then the adjusted IMR = $\alpha/(\alpha + \beta)$.

Once in this form, we can fit a beta-binomial model to a distribution of IMRs, which when fitted, produces two estimated parameters, α and β, as discussed earlier. Note that "rate" = $\alpha/(\alpha + \beta)$, which in our case is equivalent to" infant mortality rate" = infant deaths/(infant deaths + survivors to age 1), which reduces to infant deaths/births. Thus, parameter α is the numerator in the expression defining a rate, and when added together, the parameters α and β represent the denominator.

As an example of this process, consider again our earlier examples of the State of Washington and Asotin County, where IMR = 0.00484 and 0.01316, respectively. A beta-binomial model was fitted to IMRs representing the 31 of 39 counties of the State of Washington for which an IMR > 0.00 was reported in 2015. Although we discuss the data and the beta-binomial process for the California counties in the Results section, we discuss neither in terms of the estimates for our Washington example. Here, we simply note that for Washington, the beta-binomial process yielded the following values for the model's two parameters: $\alpha = 3.5897$, and $\beta = 523.105$. Applying these parameters to the infant deaths and births reported in 2015 for Washington as a whole yields an adjusted IMR = .00483 = (431 + 3.5897)/((431 + 89,000) + (3.5897 + 523.105)); applying them to Asotin County yields an adjusted IMR = 0.00870 = (3 + 3.5897)/((3 + 228) + (3.5897 + 523.105)). Note that the adjusted IMR (0.00483) for the State as a whole is virtually the same as the reported IMR (0.00484). This is consistent with the argument that this large population is not substantially affected by uncertainty and, as such, its reported IMR represents the state's underlying IMR. However, the adjusted IMR (0.00870) for Asotin County is substantially less than the reported IMR (0.01316), which also is consistent with the argument that this small population is affected by uncertainty. As such, the adjusted IMR for Asotin County is likely to be closer to the county's "underlying" IMR than is the reported IRM because it is the result of a process that smoothed out some of the uncertainty.

Data

As noted earlier, the IMR data for California's counties representing the period 2009-2011 are taken from the open data portal provided by the California Health and Human Services Agency (no date, see endnote 3 for details). For the "representative" set of IMRs, the IMRs for the 43 counties reporting IMRs are used. Table 1 shows the IMRs, births, and deaths, for all of the 58 counties, including the 15 where neither infant deaths nor IMRs are reported. Note that the mean IMR across all 58 counties is 0.00376, the standard deviation, 0.00247, and the CV (coefficient of variation), 0.66. The highest IMR is 0.008023 (Tehama County) and the lowest non-zero IMR is 0.00285 (San Mateo County). This range and the CV strongly suggest that the IMR is not constant across counties. As such, we use the beta-binomial model, which, as discussed earlier, is designed to deal with binomial variables that exhibit variation across cases.

Here, it also is important to note that the sum of number of infant deaths is 7,448 across the 58 counties shown in Table 1 is 50 less than that reported for the State as a whole. These represent the cases where the state reported zero deaths instead of the recorded number in order to preserve confidentiality at the county level. When these 50 infant deaths are included, the state's IMR increases from 0.00484 to 0.00487.

Table 1. Infant Deaths, Births & IMRs by County, California, 2009-11

COUNTY	INFANT DEATHS	BIRTHS	IMR
Alameda	263	58,637	0.00449
Alpine	0	14	0.00000
Amador	0	837	0.00000
Butte	36	7,286	0.00494
Calaveras	0	1,010	0.00000
Colusa	0	1,001	0.00000
Contra Costa	185	37,095	0.00499
Del Norte	0	1,042	0.00000
El Dorado	17	4,969	0.00342
Fresno	339	48,714	0.00696
Glenn	0	1,250	0.00000
Humboldt	26	4,552	0.00571
Imperial	42	9,294	0.00452
Inyo	0	644	0.00000
Kern	268	43,537	0.00616
Kings	46	7,716	0.00596
Lake	16	2,166	0.00739
Lassen	0	947	0.00000
Los Angeles	1,969	403,343	0.00488
Madera	44	7,227	0.00609
Marin	22	7,253	0.00303
Mariposa	0	432	0.00000
Mendocino	22	3,227	0.00682
Merced	68	12,937	0.00526
Modoc	0	291	0.00000
Mono	0	446	0.00000
Monterey	95	20,649	0.00460
Napa	24	4,750	0.00505
Nevada	13	2,314	0.00562
Orange	469	116,786	0.00402
Placer	49	11,463	0.00427
Plumas	0	489	0.00000
Riverside	469	92,899	0.00505
Sacramento	329	60,493	0.00544
San Benito	13	2,259	0.00575
San Bernardino	609	93,937	0.00648
San Diego	586	133,462	0.00439
San Francisco	82	26,422	0.00310
San Joaquin	192	31,793	0.00604
San Luis Obispo	42	7,983	0.00526
San Mateo	79	27,696	0.00285
Santa Barbara	80	17,666	0.00453
Santa Clara	242	72,795	0.00332
Santa Cruz	29	9,723	0.00298
Shasta	39	6,229	0.00626
Sierra	0	67	0.00000
Siskiyou	10	1,383	0.00723
Solano	86	15,598	0.00551
Sonoma	77	16,231	0.00474
Stanislaus	125	23,486	0.00532
Sutter	19	4,119	0.00461
Tehama	19	2,310	0.00823
Trinity	0	348	0.00000
Tulare	123	24,488	0.00502
Tuolumne	0	1,342	0.00000
Ventura	148	33,160	0.00446
Yolo	22	7,252	0.00303
Yuba	15	3,751	0.00400

Source: California Department of Public Health (see endnote 5)

Results

The beta binomial model procedure found within the "survival/reliability" module of the NCSS statistical analysis package (release 8) was used to obtain the two beta model parameters using the infant mortality rates for the 43 counties for which IMRs were reported (see Table 1). The major results of interest found

in running this procedure with the data are found as Exhibit 2. Note that there two different estimates of the parameters, α and β, presented in the exhibit, one accomplished by the method of moments and the other by Maximum Likelihood Estimation. The parameters of the latter are used here, namely: α = 15.48574, and β = 3042.145.

Exhibit 2. Report of the Fit of the Beta Model to 43 California Counties Reporting 1 or More Infant Deaths

Dataset ...\CA IMR MORE THAN ZERO BY COUNTY 200911.NCSS
Variable IMR

Parameter Estimation Section

Parameter	Method of Moments Estimate	Maximum Likelihood Estimate	MLE Standard Error	MLE 95% Lower Conf. Limit	MLE 95% Upper Conf. Limit
Minimum (A)	0	0			
Maximum (B)	1	1			
α	15.70821	15.48574	3.304395	9.009242	21.96223
β	3085.83	3042.145	659.7058	1749.145	4335.145

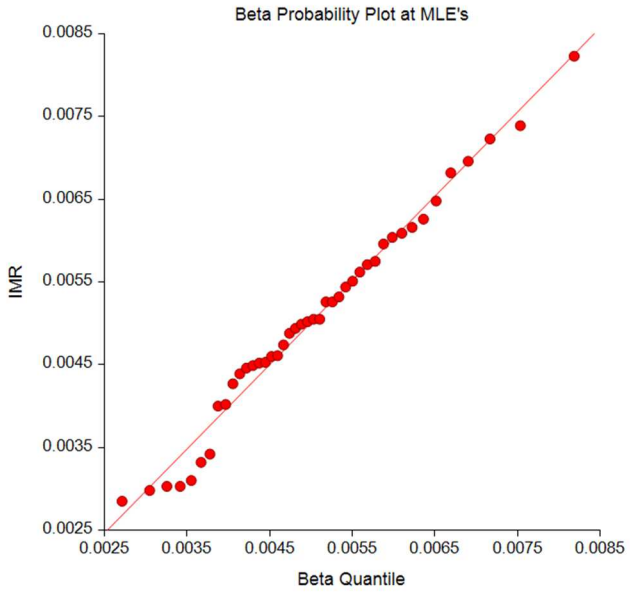


Table 2 shows both the IMRs shown in Table 1 for the 43 counties reporting them and the estimated 2015 underlying IMRs for the 15 counties (identified by the yellow background). Note that if estimates are calculated by hand they may not exactly match those shown in Table 2 because of precision levels and rounding algorithms. The estimated underlying IMRs are found by using the two beta parameters in conjunction with reported 2015 infant deaths (zero in each case) and reported births by county using the formulas described earlier in the examples for the State of Washington and Asotin County. We note that

Table 2. Estimates of Underlying IMR by County

COUNTY	2009-11 IMR
Alameda	0.00449
Alpine	0.00500
Amador	0.00400
Butte	0.00494
Calaveras	0.00380
Colusa	0.00380
Contra Costa	0.00499
Del Norte	0.00380
El Dorado	0.00342
Fresno	0.00696
Glenn	0.00360
Humboldt	0.00571
Imperial	0.00452
Inyo	0.00420
Kern	0.00616
Kings	0.00596
Lake	0.00739
Lassen	0.00390
Los Angeles	0.00488
Madera	0.00609
Marin	0.00303
Mariposa	0.00440
Mendocino	0.00682
Merced	0.00526
Modoc	0.00460
Mono	0.00440
Monterey	0.00460
Napa	0.00505
Nevada	0.00562
Orange	0.00402
Placer	0.00427
Plumas	0.00460
Riverside	0.00505
Sacramento	0.00544
San Benito	0.00575
San Bernardino	0.00648
San Diego	0.00439
San Francisco	0.00310
San Joaquin	0.00604
San Luis Obispo	0.00526
San Mateo	0.00285
Santa Barbara	0.00453
Santa Clara	0.00332
Santa Cruz	0.00298
Shasta	0.00626
Sierra	0.00500
Siskiyou	0.00723
Solano	0.00551
Sonoma	0.00474
Stanislaus	0.00532
Sutter	0.00461
Tehama	0.00823
Trinity	0.00450
Tulare	0.00502
Tuolumne	0.00350
Ventura	0.00446
Yolo	0.00303
Yuba	0.00400

by multiplying the IMR by the number of births, an estimate of the number of infant deaths by county is found. These estimated infant deaths sum to 7,489, which is one less than that reported for the state as a whole. This shows that the beta-binomial process is not only dealing with uncertainty but also with the 50 infant deaths that are not reported by county (Table 1). As such, the IMR for the State as a whole as estimated by the beta-binomial process is 0.00487, which is virtually identical to that reported for the State as whole (California Health and Human Services Agency, no date).

Discussion of results

The estimated IMRs for the 15 counties reporting zero infant deaths range from a low of 0.00350 (Tuolomne) to a high of 0.00500 (Sierra) which is less than the range found for the 43 counties reporting one or more infant deaths, which is from a low IMR of 0.00285 (San Mateo) to a high of 0.00823 (Tehama County). This suggests that the process used to create the estimated IMRs may, in fact, represent the IMRs underlying these counties in that the estimates do not display a high level of variation, which would reflect a high level of uncertainty. The results also suggest that the process also deals with the 50 infant deaths that are suppressed at the county level. These results represent what the method is intended to do.[7]

As discussed in Endnote 2, we use the "superpopulation" concept as a means of interpreting what the method does. Thus, we view the 15 "small population" counties for which no infant deaths are reported as being subject to the high levels of uncertainty associated with small samples. The actual "sample" IMRs for these counties are suppressed by the fact that the State of California reports zero infant deaths for each of them because of confidentiality and related concerns. However, we know from the state total that there are 50 infant deaths spread among these 15 counties. In using the beta-binomial process to estimate IMRs for these 15 counties, we are attempting to estimate the "underlying" IMRs associated with the 15 county "superpopulations" from which, respectively, 15 "samples" were taken (and for which the sample IMRs are known by the State of California), but for which zero infant deaths were reported, even for those for which the sample IMRs were not zero. Our estimation attempt may not provide the exact IMR underlying each of these 15 counties, but we do know that it generates 50 infant deaths among them, reproducing the number that was suppressed by the State of California.

Validity test

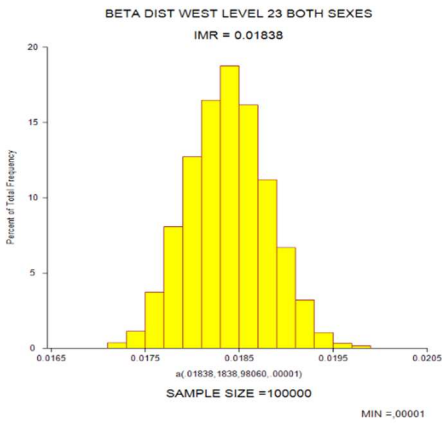
Because we argue that the method is producing a revised IMR that is likely to be close to the underlying IMR for a small population and therefore reflective of its intrinsic mortality regime, one would expect the method to do this where one could observe the intrinsic mortality regime. Model stable populations afford this opportunity because they have known intrinsic mortality regimes, the model life tables associated with a given set of model stable populations. To examine how the method works in this environment, we employed the IMR associated with a model stable population found in *Manual IV, Methods of Estimating Basic Demographic Measures from Incomplete Data* (United Nations, 1967). For this purpose, we selected the infant mortality rate associated with West Level 23 for both sexes (United Nations, 1967: Table 1.2, p.93), which shows that of 100,000 births, 98,166 are expected to reach the first birthday. This yields an IMR of 0.0184 = 1 -.98166.

Using the IMR of 0.0184 and a seed population of 100,000, we generated a random sample of 5,000 IMRs using the beta model simulation provided by the NCSS statistical system (release 8). The sample is sufficiently large to allow the simulation program the opportunity to generate outliers, which it did. As can be seen in Exhibit 3, the mean is 0.01838 with a standard deviation of 0.000423 and a coefficient of variation equal to 0.02305. The minimum IMR is 0.016849 and the maximum is 0.020147.

Exhibit 3. Descriptive Statistics for the 5,000 Simulated IMR observations

Descriptive Statistics of Simulated Data

Statistic	Value	Statistic	Value
Mean	0.01838248	Minimum	0.01684878
Standard Deviation	0.0004237251	1st Percentile	0.01744547
Skewness	0.07195781	5th Percentile	0.01769227
Kurtosis	2.934381	10th Percentile	0.0178332
Coefficient of Variation	0.02305049	25th Percentile	0.01808544
Count	5000	Median	0.01838394
		75th Percentile	0.01866444
		90th Percentile	0.0189272
		95th Percentile	0.01908542
		99th Percentile	0.01937772
		Maximum	0.02014658



From the 5,000 randomly generated observations, we extracted two sets of data. For the first set, we extracted the initial 43 IMR randomly generated observations from the simulation. For the second, we rank-ordered the 5,000 observations: from high to low and then from low to high, and extracted the eight highest IMR and seven lowest IMRs, respectively from them. The idea is that the entire set represents a synthetic population with 58 observations while the set of 43 simulated IMRs represents the subset of "large populations" in the synthetic population for which IMRs are reported, and the set of 15 simulated IMRs represents a subset of "small populations" in the synthetic population, which is subject to a high level of uncertainty. The 43 observations are expected to be closer, on average, to the "underlying" IMR of 0.01838 and have variation, respectively, than that found in the 15 observations.

For the set of 43 observations, the mean IMR is 0.01834 and the coefficient of variation is .02305. For the set of 15 observations, the mean IMR is .01855 and the coefficient of variation is .07692. Thus, the set of

43 “large population” observations has a mean and a coefficient of variation closer to the mean and coefficient of variation found in the full set of 5,000 observations than does the set of 15 observations.

To the set of 43 observations, a beta model was fit and its parameters were used to revise the IMRs in the set of 15 observations. The expectation is that the revised IMRs will yield a mean IMR closer to that found for the full 5,000 set of simulated observations and that the variation among these revised means will decline, yielding a smaller coefficient of observation.

The results show that the beta model moved the initial IMR estimates for the 15 observations closer to the underlying IMR. As such, they are more reflective of the West Level 23 mortality regime that is intrinsic to them: the mean of the original IMRs for the 88 observations is 0.01855 while the mean for the revised IMRs is 0.01839, which is closer to the underlying IMR of 0.01838. In terms of variation, the coefficient of variation for the initial set of 14 IMRs is 0.07692, while that for the revised set is 0.00338. These results support our argument that the method we describe in this paper is capable of moving IMRs subject to uncertainty closer to the underlying IMRs.[8]

Concluding remarks

Because of the “representational context” selection, the estimates are subject to judgment. However, even still the entire process is transparent, which means that the results are not subject to arbitrary and capricious judgments that render them difficult to replication. Keep in mind that with a different “representational context,” one will have a different model and different IMR estimates. However, as the validity test indicates, a different model, can be expected to move, on average, the IMRs for the California counties closer to their underlying IMRs, better reflecting their “underlying IMRs.” This argument can be generalized to other potential data sets that could be used to build different beta-binomial models. This feature of the beta-binomial approach suggests that while a model built from a given “representational” data set may move the estimated IMRs closer, on average, to their underlying values, than a model built from a different “representational” data set, even a less than optimal model should provide reasonable estimates. This and the evidential support provided by the validity test that, in fact, our method is capable of producing estimates of underlying IMRs, suggests that the method is not only capable of generating reasonable IMR estimates in the absence of reported infant deaths, but that these are valid in terms of the “underlying” mortality regimes affecting small populations and also as a means of compensating for the situation of suppression, namely that infant deaths may be reported as zero to preserve confidentiality for a given county when in fact infant deaths have occurred. Because these estimates can be efficiently generated by the process described here also suggests that they have the potential to support policy decisions while keeping time and resource requirements low, characteristics that Swanson and Tayman (2012: 304) suggest are important components in deciding what methods to use in developing estimates.

While the beta-binomial model has been used in medical research (Kim and Lee, 2013; Arostegui, Nuñez-Antón, and Quintana, 2007; and Young-Xu and Chan, 2006), consumer studies (Chatfield and Goodhardt, 1970), bioinfomatics (Pham et al., 2010) and public health research (Alanko and Lemmons, 1996; Gakidou and King, 2002), it has not found much traction in demographic research. This is surprising on two counts: (1) the components of demographic change, births, deaths, and migration, can all be constructed as rates that are inherently binomial variables; and (2) the method is simple to use, explain, and understand.[9] This

paper illustrates one such use with a subset of the mortality component, the infant mortality rate. Although the paper focuses on a specific application, namely infant mortality rates for counties with small populations in California, the method can be applied to many other situations where small numbers are present and affected by uncertainty. As such, it could be used in conjunction not only with other mortality measures such as neo-natality rates, crude death rates, age-specific death rates and cause-specific death rates, but with fertility measures such as crude birth rates and age-specific birth rates. Even more broadly, it could be used with any binomial variable of interest affecting small populations, such as a housing occupancy (or vacancy) rate, employment (or unemployment) rate, and cigarette smoking (or non-smoking) rate.

Acknowledgments

We are grateful to comments made by three anonymous reviewers on an earlier version of this paper and to Tom Pullum for advice he kindly provided regarding the beta-binomial method and its use.

Endnotes

1. Murray (1996) has argued that the infant mortality rate is flawed when it is used as an index of overall mortality (i.e., the mortality regime affecting a given population) and that Disability Adjusted life Expectancy (DALE) should be used in its place. However, it has been pointed out by Reidpath and Allotey (2003) that the infant mortality rate and the DALE are so highly correlated that it merely goes to reinforce the intuition that the causes of infant mortality are strongly related to those structural factors like economic development, general living conditions, social well-being, and environmental factors, and, and such, the infant mortality rate remains a useful and comparatively inexpensive indicator of population health. Guillot et al. (2013) also note that infant mortality is very useful because it involves a short lag between the timing of mortality exposures and the timing of corresponding births.

2. Uncertainty can be described in different ways. In our paper, it can be viewed similarly to that found in sampling theory. That is, one has a variable of interest (the infant mortality rate) and its value in the population of interest becomes more uncertain as sample size decreases. When one is looking a finite population (such as the population of either the State of Washington as a whole or Asotin county, Washington), it can itself be viewed as the realization of a "sample" from an infinite set of possibilities. Because Washington has a very large population, there is less uncertainty in its reported infant mortality rate (IMR) than that reported for Asotin County, which has a very small population. This view is based on the idea of a "superpopulation" (Swanson and Tayman, 2012, 172; Williams, Frederick, and Nichols, 2011), which dates back to a paper by Deming and Stephan (1941). In this view, there is a "Washington State superpopulation" underlying the Washington State IMR "sample" and an "Asotin County superpopulation" underlying the Asotin County IMR sample. The method we employ is aimed at estimating the IMRs in these superpopulations. The superpopulation view is not dissimilar to the stochastic perspective, which sees a variable as being subject to a random process that can result in changes in the variable over time (Doob, 1952; Graham and Talay, 2013; Karlin and Taylor, 1993). We also make use of this perspective in the paper. For an interpretation using the stochastic perspective consider the following hypothetical case. Suppose there was a 2017 car accident in our small population example of Asotin County in which six teens aged 15-19 all died, but there was no such accident in either 2016 or 2018. With only about 1,100 teens

aged 15-19, there would be a huge spike in the 2017 death rate for those 15-19 in Asotin County compared to 2016 and 2018. It is due to the random occurrence of an accident – the fickle finger of fate. However, for the State of Washington, our example of a large population, the six deaths in Asotin County would not create a huge spike in the 2017 death rate for those aged 15-19 compared to 2016 and 2018 because there were approximately 457,000 teens aged 15-19 in the state as a whole, about 415 times more than found in Asotin County.

3. Another strategy is to gain permission to access individual level records, match them and then construct statistics (Kinge and Kornstad, 2014). This strategy is used by the California Department of Public Health (see endnote 4). However, unlike the strategy of aggregation, this approach inevitably requires both a substantial amount of time and personnel costs to implement and it also may be subject to limitations due to confidentiality and the uncertainty associated with small populations.

4. In addition to ideas taken from the approach described by Link and Hahn (1996), we use ideas from both the "superpopulation" and the "stochastic" traditions found in demographic analysis in this paper (See endnote 2). For examples of these traditions in demography, see, on the one hand, Deming and Stephan (1941) and Swanson and Tayman (2011), and on the other, Baker, Alcantara and Ruan (2011).

5. The data for California are taken from the Open Portal service provided by the California Health and Human Services Agency (no date) via a download of a CVS data set assembled by the California Department of Public Health. This data set can be accessed by going to
https://data.chhs.ca.gov/dataset/infant-mortality-deaths-per-1000-live-births-lghc-indicator-01/resource/ae78da8f-1661-45f6-b2d0-1014857d16e3
and then clicking on the "download" tab, which downloads the file, "Infant Mortality, Deaths Per 1,000 Live Births (LGHC Indicator 01) (CSV)" in CVS form. Once downloaded, it can be saved as an excel file. The data in this file include the infant mortality rates (identified as "rate" in the file) and the infant deaths (identified as "numerator" in the file) and live births (identified as "denominator" in the file) used to calculate the IMRs for all counties and other administrative areas, including the State as a whole. The data represent the period 2009-2011. A description of the methods, caveats, and so forth associated with this data set found on the ULR shown above is reproduced below.

This is a source dataset for a "Let's Get Healthy California" indicator, which can be found online (https://letsgethealthy.ca.gov/). Infant Mortality is defined as the number of deaths in infants under one year of age per 1,000 live births. Infant mortality is often used as an indicator to measure the health and well-being of a community, because factors affecting the health of entire populations can also impact the mortality rate of infants. Although California's infant mortality rate is better than the national average, there are significant disparities, with African American babies dying at more than twice the rate of other groups. Data are from the Birth Cohort Files. The infant mortality indicator computed from the birth cohort file comprises birth certificate information on all births that occur in a calendar year (denominator) plus death certificate information linked to the birth certificate for those infants who were born in that year but subsequently died within 12 months of birth (numerator). Studies of infant mortality that are based on information from death certificates alone have been found to underestimate infant death rates for infants of all race/ethnic groups and especially for certain race/ethnic groups, due to problems such as confusion about event registration requirements, incomplete data, and transfers of newborns from one facility to another for

medical care. Note there is a separate data table "Infant Mortality by Race/Ethnicity" which is based on death records only, which is more timely but less accurate than the Birth Cohort File. Single year shown to provide state-level data and county totals for the most recent year. Numerator: Infants deaths (under age 1 year). Denominator: Live births occurring to State of California residents. Multiple years aggregated to allow for stratification at the county level. For this indicator, race/ethnicity is based on the birth certificate information, which records the race/ethnicity of the mother. The mother can "decline to state"; this is considered to be a valid response. These responses are not displayed on the indicator visualization.

6. If IMR variation across the counties was either constant or very small, a binomial model could be applied (see, e. g., the discussion in Chatfield and Goodhardt 1970: 400). However, the variation in county IMR levels found and discussed in Table 1 (even discounting the 15 counties for which zero is reported) is noticeable and as such we used the beta-binominal model, which is designed to handle binomial variables with variation across cases.

7. Keep in mind that small populations, however defined, with approximately the same total populations may have different age compositions. For example, one may have a relatively large aged population and another a relatively large young population. This simple example is meant to illustrate the effect of demographic heterogeneity, which can affect measures of mortality (Vaupel and Missov, 2014). In situations where substantial heterogeneity may be present, a model with additional covariates may prove useful because the latter can potentially take into account the effects of demographic heterogeneity. In a related vein, we note that if one applied the IMR for California as a whole (0.00487) to the reported births of the 15 counties reporting zero infant deaths to estimate infant deaths for each of them, the sum would come to 49, nearly capturing the 50 infant deaths not reported at the county level for data confidentiality reasons. However, the estimated IMRs for each of these counties using this ultra-simple approach vary from the IMRs estimated by the beta-binomial model because the underlying assumption is that the IMR underlying IMR is constant across all of them; the beta-binomial model allows for the possibility that the underlying IMRs themselves vary.

8. In the validity test, different populations are simulated from a common beta distribution, and the result is that the two sets of populations, large and small, are normally distributed around the underlying mean IMR of the "superpopulation." The simulation shows that the adjusted IMRs of the small populations move closer to the superpopulation's underlying IMR, which indicates that the method works when both the small and large populations represent samples taken from the same underlying superpopulation. If the small populations represent a sample from a different superpopulation than the sample of large population, then the adjustment may yield a "biased" estimate of the former's underlying IMR. This shows the importance of having a reference set that conceptually represents a sample from the same underlying population as the small population sample. One way to visualize the unbiased and biased outcomes is to picture the case where the method yields: (1) an "unbiased" estimate, which is when the mean IMR of the large populations is between the underlying superpopulation IMR and the mean IMR of the small populations; and (2) a "biased" estimate when the method does not move the mean IMR for the small population closer to its underlying IMR, which occurs where the mean IMR of the small population is between the underlying IMR and the mean IMR of the large populations.

9. The beta-binomial process shares similarities with "Empirical Bayesian" methods (Link and Hahn 1996), which also are known as "shrinkage estimates" (Assunção et al., 2005). These methods have been used by demographers to estimate a range of population characteristics, including mortality (Atkin Liu and Chadwick 2009; Assunção et al., 2005; and Marshall 1991). However, we believe the beta-binomial approach is likely to be more understandable to a wider range of demographers than the Empirical Bayesian approach. The starting point for our argument is found in Green and Armstrong (2015). Although they discuss simple vs. complex methods in terms of forecasting, their discussion applies here in that the beta-binomial approach falls into the simple methodological category rather than the complex category, where Empirical Bayesian methods fall. Adapting their discussion to methods in general, the work of Green and Armstrong (2015) suggests that while there is no evidence that shows complexity improves accuracy, complexity remains popular among: (1) researchers, because they are rewarded for publishing in highly ranked journals, which favor complexity; (2) methodologists, because complex methods can be used to provide information that support decision makers' plans; and (3) clients, who may be reassured by incomprehensibility. We believe that the argument by Green and Armstrong (2015) can be applied to Bayesian methods, which represents the "complex" alternative to the "simple" beta-binomial approach. We prefer the beta-binomial approach, however, not only because of the argument presented by Green and Armstrong, but also because the application of a Bayesian approach can be difficult, effortful, opaque and even counter-intuitive (Goodwin, 2015).

References

Alanko, T., and P. Lemmens (1996). Response effects in consumption surveys: An application of the beta-binomial model to self-reported drinking frequencies. *Journal of Official Statistics* 12 (3): 253-273.

Arostegui I., V. Nuñez-Antón, and J. Quintana (2007). Analysis of the Short Form-36 (SF-36): The beta binomial distribution approach. *Statistics in Medicine* 26: 1318-1342

Atkin, M. C. Liu and T. Chadwick (2009). Bayesian Model Comparison and Model Averaging for Small-Area Estimation. *The Annals of Applied Statistics* 3 (1): 199-221

Assunção, R., C. Schmertmann, J. Potter and S. Cavenaghi (2005). Empirical Bayes Estimation of Demographic Schedules for Small Areas. *Demography* 42 (3): 537-558

Baker J, A. Alcantara, and X. Ruan. (2011), A stochastic version of the Brass PF Ratio adjustment of age-specific fertility schedules. *PLoS ONE* 6(8): e23222. https://doi.org/10.1371/journal.pone.0023222.

Brass W., A. Coale, P. Demeny, D. Heisel, F. Lorimer, A. Romaniuk, and E. Van de Walle (1968). *The Demography of Tropical Africa*. Princeton: Princeton University Press.

California Health and Human Services Agency (no date) Births and infant deaths by county. https://data.chhs.ca.gov/dataset/infant-mortality-deaths-per-1000-live-births-lghc-indicator-01/resource/ae78da8f-1661-45f6-b2d0-1014857d16e3

Chatfield, C., and G. Goodhardt (1970). The Beta-Binomial Model for Consumer Purchasing Behavior. *Applied Statistics* 19:240–250.

Chen, A., E. Oster, and H. Williams (2016). Why is infant mortality higher in the United States than in Europe? *American Journal of Economic Policy* 8(2): 89–124.

Chung, K., D. Yang, and R. Bell (2004). Health and GIS: Toward spatial statistical analyses. *Journal of Medical Systems* 28 (4): 349 – 360.

Deming, W. and F. Stephan (1941). On the interpretation of censuses as samples. *Journal of the American Statistical Association* 36 (21): 45–49.

Doob, J. (1953). *Stochastic Processes*. New York. John Wiley & Sons.

Gakidou, E., and G. King (2002). Measuring total health inequality: adding individual variation to group-level differences. *International Journal of Equity in Health* 1:3 doi: 10.1186/1475-9276-1-3

Gardiner, C. (1983*). Handbook of Stochastic Methods for Physics, Chemistry, and the Natural Sciences*. New York: Springer.

Goodwin, P. (2015). When simple alternatives to Bayes formula work well: Reducing the cognitive load when updating probability forecasts. *Journal of Business Research* 68: 1686-1691.

Graham, C., and D. Talay (2013). *Stochastic Simulation and Monte Carlo Methods*. New York: Springer.
Green, K. and J.S. Armstrong. (2015). Simple versus complex forecasting: The evidence. *Journal of Business Research* 68: 1678-1685.

Guillot, M., S. Lim, L. Torgasheva, and M. Denisenko (2013). Infant mortality in Kyrgyzstan before and after the break-up of the Soviet Union. *Population Studies* 67(3): 335-352.

Hummer, R. (2005). Income, Race, and Infant Mortality: Comment on Stockwell et al. *Population Research and Policy Review* 24: 405–409.

Karlin, S, and H. Taylor (1993). An Introduction to Stochastic Modeling. Elsevier.

Kim J., and J. Lee (2013). Simultaneous confidence intervals for a success probability and intraclass correlation, with an application to screening mammography. *Biometrical Journal* 55 (6):944–954

Kinge, J. and T. Kornstad (2014). Assimilation effects on infant mortality among immigrants to Norway: Does maternal source country matter? A *Demographic Research* 31 (available online at https://www.demographic-research.org/volumes/vol31/26/default.htm)

Kitagawa, E. and P. Hauser (1973). *Differential Mortality in the United States: A Study in Socioeconomic Epidemiology*. Cambridge: Harvard University Press.

Kleinman, J. (1996). Underreporting of infant deaths: Then and now. *American Journal of Public Health* 76 (4): 365-366.

Link, B., and J. Phelan (1995). Social Conditions as Fundamental Causes of Disease. *Journal of Health and Social Behavior* (extra issue): 80–94.

Link, W. and D. Hahn (1996). Empirical Bayes Estimation of proportions with application to cowbird parasitism rates. *Ecology* 77 (8): 2528-2537.

Misra, D., H. Grason, M. Liao, D. Strobino, K. McDonnell, and A. Allston (2004). The nationwide evaluation of fetal and infant mortality reviewed (FIMR) programs: development and implementation of recommendations and conduct of essential maternal and child health services by FIMR programs. *Maternal and Child Health Journal* 8(4); 217-229.

Murray C. (1996). Rethinking DALYs. pp. 1-98 in C. Murray and A. Lopez (eds.) *The global burden of disease: a comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020*. Cambridge, MA: Harvard School of Public Health.

Office for National Statistics (2015). *Small populations tables from the 2011 Census – user guide*. Newport, South Wales, United Kingdom.

Pham, T., S. Piersma, M. Warmoes, and C. Jimenez (2010). On the beta-binomial model for analysis of spectral count data in label-free tandem mass spectrometry-based proteomics. *Bioinformatics* 26 (3): 363–369.

Reeske, A., and O. Razum (2011). Maternal and child health – from conception to first birthday. pp. 139 – 153 in B. Recehl, P. Mladovsky, W. Devillé, B. Rijks, R. Petrova-Benedict, and M. Mckee (Eds.). *Migration and health in the European Union. Berkshire, England*. Open University Press.

Reidpath, D. and P. Allotey (2003). Infant mortality rate as an indicator of population health. *Journal of Epidemiology and Community Health* 57: 344-346.

Siegel, J. and D. a. Swanson (2004). *The methods and materials of demography, 2nd Edition*. Los Angeles: Academic/Elsevier Press.

Stockwell, E., F. Goza, and K. Balisteri (2005). Infant Mortality and Socioeconomic Status: New bottle, Same Old Wine." *Population Research and Policy Review* 24: 387–39

Stockwell, E., M. Bedard, D. A. Swanson, and J. Wicks (1987). Public Policy and the Socioeconomic Mortality Differential in Infancy. *Population Research and Policy Review* 6 (Fall):105-121.

Swanson, D. A. and J. Tayman (2012). *Subnational Population Estimates*. Dordrecht, The Netherlands: Springer.

United Nations. (1967). *Manual IV, Methods of Estimating Basic Demographic Measures from Incomplete Data.* New York, NY: United Nations.

US National Center for Health Statistics (no date). Data Use Restrictions (https://wonder.cdc.gov/datause.html)

US National Center for Health Statistics (2018). *Linked Birth / Infant Death Records 2007-2015, as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program, on CDC WONDER On-line Database*, http://wonder.cdc.gov/lbd-current.html.

VanEenwyk, J. and S. Macdonald (2012). *Guidelines for Working with Small Numbers.* Olympia, WA: Washington Department of Health, Environmental Public Health Division. (available at https://www.doh.wa.gov/Portals/1/Documents/1500/SmallNumbers.pdf

Vaupel, J, and T. Missov (2014). Unobserved population heterogeneity: A review of formal relationships. *Demographic Research* 31 (22): 659-686.

Voss, P. R., C. Palit, B. Kale, and H. Krebs (1995). Censal ratio methods. pp.70–89, in N. W. Rives, W. J. Serow, A. S. Lee, H. F. Goldsmith, and P. R. Voss (Eds.) *Basic methods for preparing small-area estimates.* Madison: Applied Population Laboratory, University of Wisconsin.

Washington State Department of Health (2016a). Excel file, "Births 2015." Olympia, WA. Center for Health Statistics, Department of Health. (available by year, online at https://www.doh.wa.gov/DataandStatisticalReports/HealthStatistics/Birth/BirthTablesbyYear),
Washington State Department of Health (2016b). Excel file, "Infant Deaths 2015." Olympia, WA. Center for Health Statistics, Department of Health. (available by year, online at

Washington State Office of Financial Management (2017). *State of Washington 2017 Population Trends*. Olympia, WA. Forecasting and Research Division, Office of Financial Management. (available online at https://www.ofm.wa.gov/sites/default/files/public/dataresearch/pop/april1/ofm_april1_poptrends.pdf

Williams, K., P. Frederick, and J. Nichols. (2011). Use of the superpopulation approach to estimate breeding population size: an example in asynchronously breeding birds. *Ecology* 92 (4): 821-828

Young-Xu, Y. and K. Chan (2008). Pooling overdispersed binomial data to estimate event rate BMC *Medical Research Methodology* 8:58