Estimation of Population Denominators for Public Health Studies at the Tract, Gender, and Age-Specific Level

Mikel Aickin, PhD, Clara N. Dunn, MPA, and Timothy J. Flood, MD

ABSTRACT

In epidemiologic and public health studies of disease incidence in geographic subpopulations, attention is properly directed toward the ascertainment of accurate numerators. Population or person-years denominators are generally given less consideration, under the assumption that estimates produced by sources other than the state health department are sufficiently accurate. Here, we report our experience in estimating person-years denominators in the highly urbanized, rapidly expanding population of Maricopa County, Arizona. The usual sources of population estimates were found to be of little use for public health purposes, and so we report on a method for obtaining smoothed person-years figures that can accurately reflect population acceleration which varies from one time period to another. Our method is to regress the logarithm of census enumerations on quadratic or tertic polynomials in time. We describe how differential reliability of census figures can be incorporated into our procedure, and how the problem of missing census data can be handled by an iterated regression method. Our evidence suggests that the logarithmic regression model works well, even in the face of rapid and erratic population growth or decline. (*Am J Public Health*. 1991;81:918–920)

Introduction

The accuracy of person-years denominators is of considerable importance for epidemiologic analysis. In forming ratios of incidence rates, or carrying out a statistical analysis using a method such as Poisson regression,[1] one is either explicitly or implicitly using ratios of denominator terms. Errors in denominator terms can have a non-trivial impact on the results. Our experience in a very rapidly growing, suburbanized region suggests that in many cases 20 percent is a lower bound on the error resulting from using straight-line population models. When discrepancies of this magnitude occur in opposite directions in numerators and denominators, an estimated incidence rate ratio can be 1.5 times the true ratio.

In the absence of sufficient data (birth, death, migration rates) for a strong, demographic analysis, we have employed a practical approach that appears to work well even at the small unit level.

Methods and Materials

Official and semi-official population estimates are available from several sources. The United States Bureau of the Census can produce subcounty estimates for incorporated places and minor civil divisions. The state Department of Economic Security uses a mixture of techniques to produce population estimates and projections at a variety of subcounty levels.[2] Moreover, consultants are often engaged to produce estimates for traffic analysis zones, districts, and municipal planning areas.

A review of these approaches reveals that they produce errors that range from below 1 percent to above 35 percent for units much larger than census tracts.[3] They also involve complex formulas, and mixtures of data sources. In many cases the geographical units for which estimates are made not only fail to coincide with census tract boundaries, but actually change their definitions from year to year. Data that are available on incident cases are frequently insufficient to place them in these alternative geographic units.

Moreover, these estimates are made for economic, political, or marketing reasons, and make no attempt to avoid confounding. Denominators are generated on the basis of assumptions about relationships between numerators and denominators, while these latter relationships are precisely the focus of public health studies.

For these reasons, it appears to us to be desirable from the public health perspective to generate historical and current estimates of person-years of experience that are based solely on official census figures.

Log-Linear Population Model

If p_t represents the size of a population at time t, then the velocity is defined by

$$\text{Population Velocity} = \frac{d}{dt}\ln(p_t)$$

where ln is the natural logarithm. We have used the tertic polynomial model

$$\ln(p_t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3$$

in order to allow for smooth changes in population, and to include the special case of a constant percent change, when β_2 and β_3 are set to zero. In our situation, we used the decennial censuses of 1970 and 1980, and special census in 1965, 1975 and 1985.

In order to allow for the fact that approximately one-half of all tracts were not covered or only partially covered in the mid-decade censuses, and that the latter three censuses might not be as accurate as

From the Arizona Department of Health Services, Division of Disease Prevention, Office of Chronic Disease, Phoenix (all authors) and the University of Arizona, Department of Family and Community Medicine, Tucson (Aickin).

Address reprint requests to Mikel Aickin, PhD, Department of Family and Community Medicine, Arizona Health Sciences Center, Tucson, AZ 85724. This paper, submitted to the Journal June 25, 1990, was revised and accepted for publication January 23, 1991.

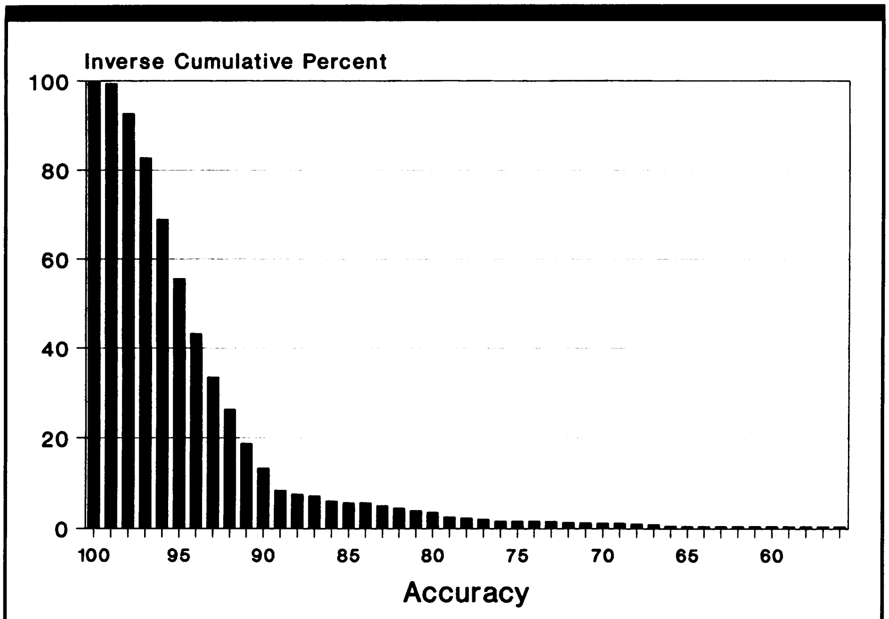

FIGURE 1—Percent of Total Person-Years Experience in Tracts having Specified Accuracy, Determined by Comparing the Exponential Polynomial Population Function with the 1970 and 1980 Census

the two decennial censuses, we weighted the decennial censuses three times as heavily as the others. We further computer weighted orthogonal tertic polynomials by the Gram-Schmidt procedure in order to simplify the many regressions that were required.

At the first step, we regressed $\ln(p_t)$ on the tertic polynomials, using only tracts that had complete data for all five censuses. The accuracy of the fitted population functions were assessed by defining

$$\text{Accuracy} = 100 -$$

$$(\%\ \text{error in 1970} + \%\ \text{error in 1980})/2$$

Thus, an accuracy of 95 means that the average percent error for the two decennial censuses was 5 percent.

In order to fill in the missing data, we carried out the next step separately for each gender and age-group. Using only the tracts with full data that had accuracy at least 90, we computed b_0 = the average of the ln of the 1970 and 1980 figures, and b_1 = one-half the difference between the latter and former numbers. These are the constant and slope estimates of a constant rate model, and their advantage is that they can be computed using only the 1970 and 1980 census counts, which are available for all tracts. Finally, using only the tracts indicated above, we regressed each of the β_i-estimates on b_0 and b_1. We refer to this as an iterated regression, because we are regressing one set of regression coefficient estimates on another set. For each tract having missing data, we then estimate the β_i-coefficients from the iterated regression, which we can do because we can compute b_0 and b_1 for any tract. We then fill in the missing counts from the fitted population function.

With complete census data (either original or filled-in) we passed through the tracts one last time, computing the coefficients for the third-degree model. As a practical matter, we were willing to set the estimate of β_3 equal to zero in those cases where the quadratic model results in an accuracy above 90.

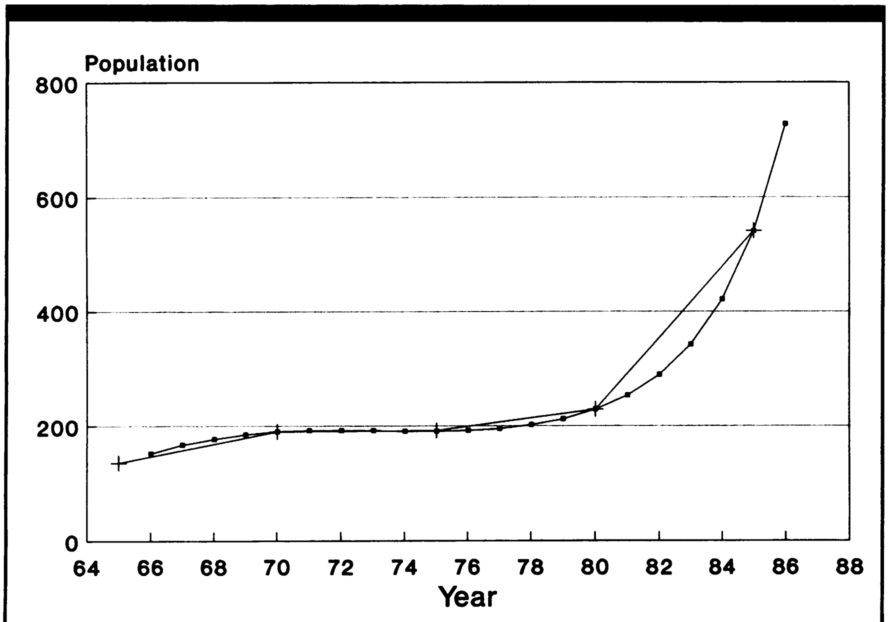

FIGURE 2—Census Enumerations (Crosses) and Smoothed Estimates (Boxes) for Arizona Tract 1092, for Males Ages 15–19, Exhibiting 100 Percent Accuracy

Results

Figure 1 summarizes the accuracy results for combinations of gender, age-group, and tract over the period 1966–86. Each bar represents the percent of the total person-years of experience that occurred in a tract having accuracy at or below the value indicated on the horizontal axis. Thus, for example, we see that 8.4 percent of the total person-years occurred in combinations with accuracy of 89 percent or worse. In the light of the small size of the units for which we produced estimates, the results shown in Figure 1 are very encouraging.

In applying our estimates to a study of childhood leukemia incidence (1966–86), we found that the median accuracies of units comprising our 12 study areas ranged from 93.9 to 96.4, indicating an absence of bias from population miss-estimation. We also reviewed a number of specific tracts, particularly those with low accuracies. Figure 2 is an example of 100 percent accuracy for a tract with a recent growth spurt, while Figure 3 shows a tract with fairly steady rapid growth since 1975, and with an accuracy of only 77 percent. These cases are typical of the range of fits that we obtained.

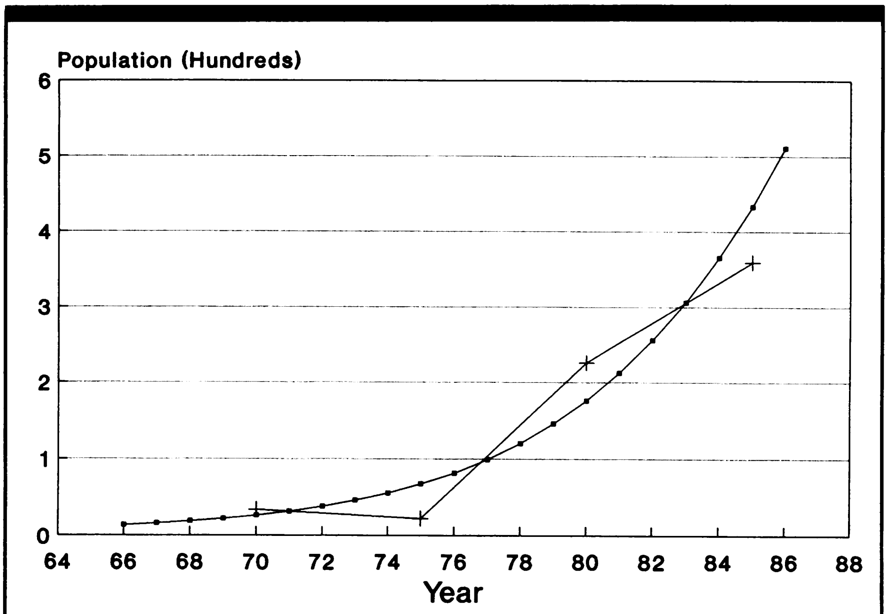

FIGURE 3—Census Enumerations (Crosses) and Smoothed Estimates (Boxes) for Arizona Tract 303, for Males Ages 15–19, Exhibiting 77 Percent Accuracy

Discussion

The straight-line interpolation (and sometimes, extrapolation) method is often used, but has some drawbacks. Because it uses the model $p_t = a + bt$, it imposes a necessary relation between population size and population velocity, which may be unrealistic. Our approach allows size and velocity to vary independently of each other. Further, our approach produces smooth estimates over an entire range, whereas the linear method produces unrealistic and counterintuitive abrupt changes in velocity at census years. Our method is susceptible to automation, so that the population functions can be computed over a large number of tracts and groupings in a reasonable amount of computer time. Moreover, we have indicated how differential census reliability and the associated missing data issues can be addressed. □

References

1. Frome EL, Checkoway H: Use of Poisson regression models in estimating incidence rates and ratios. Am J Epidemiol 1985; 121:309–323.
2. Swanson DA, Tedrow LM: Improving the measurement of temporal change in regression models used for county population estimates. Demography 1984; 21:373–381.
3. Galdi D: Evaluation of 1980 subcounty population estimates. Current Population Reports, US Department of Commerce, Bureau of the Census, Series P-25, No. 963.

APHA's Epidemiology Section Invites Abstracts for "Late Breaker" Poster Session

The Epidemiology Section of the American Public Health Association will again be sponsoring a "Late Breaker" poster session at the annual meeting. The session will take place on Wednesday, November 12, 1991, from 12:30 to 2 PM. Work completed in 1991 is eligible for consideration. Submit an abstract of fewer than 200 words (in any format) and a return envelope to Cathey Falvo, MD, MPH, Graduate School of Health Science, Munger Pavilion, New York Medical College, Valhalla, NY 10595 (tel: 914/993-4323). Abstracts must be received by October 4, 1991. Decisions will be made and mailed on October 7. Students and recent graduates are particularly encouraged to submit.