THE RANDOM VARIATION IN RATES BASED ON TOTAL ENUMERATION OF EVENTS

J. Richard Udry
Charles Teddlie
C.M. Suchindran
University of North Carolina, Chapel Hill

ABSTRACT

The instability of rates based on total enumeration of events, although not sampling error, may be thought of as being generated by random processes operating in the population. It is therefore necessary to use probability statistics to estimate a "true" rate to determine whether two rates based on total enumeration of events are different from one another. The binomial model has customarily been used to generate predicted variances on the basis of which such determinations are made. Using birth rates for five years from population units of various sizes from Taiwan, North Carolina, and Costa Rica, we obtained empirical estimates of variance in rates which are much larger than those predicted by the binomial model, even when corrections are made for time trends and unit effects. Some of the possible sources of the discrepancy in estimates are discussed. If the binomial model is used to test null hypotheses about the differences in such rates, non-conservative assertions will result.

INTRODUCTION

Social scientists and policy makers are often interested in observing changes in the rate of occurrence of events in discrete population aggregations. For example, one may be interested in knowing whether homicide rates in Pocono County are different for whites and non-whites, or whether the birth rate in a census tract in Manhattan has fallen during the last five years, or whether the motor vehicle accident rates in two counties are significantly different. Such rates are usually derived through complete enumeration of the events rather than sampling, and hence are not subject to sampling variations (errors). Thus, one may think that observed rates pretty much tell the "true" situation. It is well known, however, that the smaller the population base on which such a rate is compiled, the more unstable is the rate over time.

The purpose of this paper is to show how the instability over time of such rates, although not sampling error, may be thought of as being generated by random processes operating in the population. Statistically speaking these events are the outcome of a random experiment. These outcomes (such as birth, death, accident) are subject to chance. Thus, the observed rate may deviate from the "true" rate. Such deviation is called random error. If the experiments are repeated, a measure of this random error can be obtained by obtaining the average deviation of the observed rates around a "true value". Since in this case the experiments cannot be repeated, other procedures have to be developed to obtain measured of random variation. This paper will describe some procedures based on empirical data of birth rates, but the same logic is applicable to the rate of occurrence of many events in a population.

The considerations proposed here are important in many scientific and practical decisions concerning changes in rates. In the development of statistics for small areas, we need to consider the minimum sized population which will provide useful information. In the conduct of field experiments, the investigator often selects small population aggregates as units of "treatment." It is helpful to have a logic for the selection of the size of such units which takes into consideration the random errors in the rates of interest. Any city, considering whether some policy has been effective in changing rates of traffic accidents, crimes, fires, etc., is confronted with the same problems of determining whether the change in rates is "real".

Consider the following hypothetical data from a rural district in Costa Rica containing 5,000 people, and having the following number of births in five consecutive years: 200, 175, 215, 180, 160. For simplicity let us assume that these figures represent the true number of births which occurred. Question: is the birth rate in year five lower than in year one? On the face of it, it seems obvious that the answer is yes. But let us consider that this population contains perhaps 1,000 women of child-bearing age. In any one year about one in five gets pregnant. Which ones? Imagine that the process determing which women get pregnant is stochastic. Some will go through periods of non-exposure to risk, through illness, spouse-absence, etc. Among those exposed during the year, we can imagine pregnancy as a random gift. Whether the birth rate is actually lower in year five than in year one depends not only on the absolute difference between the two rates, but also on the random variation within the rates.

Chiang and Linder[1] seem to have been the first to examine the statistical variations in such vital rates. They have examined random errors and sampling errors of death rates in a variety of situations. They state, "The random error is associated with experimentation, whereas sampling error is due to sampling. These two kinds of error have a subtle but important difference. An understanding of these errors and their difference is essential for the understanding of the standard error of a rate." Keyfitz[2] has discussed the idea of statistical variations of some life table functions. Keyfitz[3] also discussed some measures of random deviations in crude death rates, and direct and indirect adjusted death rates under binomial and poisson

conditions. Walsh[4] and Wilson[5] have also considered some measures of life table death rate and expectation of life at birth. Kupper and Kleinbaum[6] and Kupper[7] have discussed testing equality of k indirect age-adjusted death rates from p($\geq$k) populations in which they derive a measure of random variations of some functions of indirect age specific death rate under the binomial condition. Most of the estimates of measures of random deviation obtained in the above papers make use of binomial or poisson conditions and make several simplifying assumptions to derive them.

An empirical approach to the problem of random variation in vital rates was made by Spencer[8]. She considered the problem of the effect of size of population on variability of demographic data in a historical population. Suchindran et al.[9] approached the problem using Monte Carlo simulation techniques to obtain estimates of random variations in several fertility measures.

These approaches to determining the random variation in demographic rates all assume that probability statistics may be used to estimate the "true" demographic rates. Quite often the binomial model is used to generate these "true" rates. This paper will compare estimates of variability generated by the binomial model with variations found in actual birth rate data to determine how accurate a predictor the binomial model is.

DATA SOURCES AND RESULTS

Selection of Data

The birth rate data come from three separate sources. Annual data for small units were available from North Carolina, Taiwan, and Costa Rica for the 1968-1972 period.[10] These three countries were selected because they had accurate birth rate data for areas as small as 5,000 in population. Seven population size categories were selected: (1) 0-5,000; (2) 5-10,000; (3) 10-15,000; (4) 15-20,000 (5) 20-30,000; (6) 30-40,000; (7) 40-50,000.

Data from fifty-seven cantons in Costa Rica which fell in the 0-50,000 population range in 1968 were used. Originally sixty-two cantons were in this range, but five had to be omitted due to geographic subdivisions during the 1968-1972 period. Data from the four precincts in Taiwan which had at least one township in the 0-5,000 population range were used. This resulted in data from ninety-one townships in Taiwan.

In North Carolina, birth rates were available for whites only, non-whites only, and total combining whites and non-whites. Data for the 0-5,000 population category were based only on non-whites, since white populations exceeded 5,000 in almost all of the counties. Data for the other six categories were based on total rates combining whites and nonwhites. Thus, twenty-two counties using nonwhite birth rates only constituted the 0-5,000 category, while sixty-six counties combining white and nonwhite rates were used in the other six population categories.

Estimates Under the Binomial Model

Table 1 presents a summary of the estimated standard errors based on the binomial model. For each population size category, a mean population (N) was calculated by averaging the 1970 p population of all units within that category. The estimated value of the binomial parameter ($\hat{p}$) for each category was calculated by averaging the crude birth rates over the five year time period of all units within that category. The standard error of the birth rate for each category was then estimated using the formula, $\sigma = \frac{\hat{p}(1-\hat{p})}{N}$, where $\hat{p}$ is the estimated value of the binomial parameter.

As expected, the estimated standard errors decrease as the population size increases. Within each specific population size category, North Carolina shows the smallest estimated error (except for the 0-5,000 category).

Coefficients of variation were calculated by dividing the estimated standard error of each category by the average crude birth rate for that category. These coefficients also decrease as the population size increases. The coefficients of variation for North Carolina are higher than those for Taiwan or Costa Rica, because the crude birth rate is lower in North Carolina than in the other two countries. The estimated coefficients of variation for Taiwan and Costa Rica are very similar.

Observed Standard Errors of Birth Rates

Standard errors based on observed birth rate data from these three countries were calculated next. Assuming time homogeneity, the variance in crude birth rate for each individual unit over the five year time period was calculated (Appendix A, Formula A-1). Next the average variance for each population size category was calculated by averaging the variances of the units within that category under the assumption that rates are unit homogeneous (Appendix A, Formula A-3). Finally, the standard error for each category was calculated by taking the square root of the average variance for that category. These standard errors and their corresponding coefficients of variation are presented in Table 2.

Coefficients of variation based on this analysis generally decrease as the population size increases, but there are a number of exceptions (for example, in Costa Rica at the 20-30,000 level and the 40-50,000 level). The standard errors and coefficients of variation generated from these observed rates are much higher than those that were generated using the binomial distribution.

Observed Standard Errors of Estimate Eliminating Linear Trends

Since there appeared to be a generally decreasing trend for the birth rates over the years observed, a re-analysis was performed. In this analysis, linear trends were eliminated by fitting straight line regressions (Appendix A, Assumption 3). Separate regression lines were fitted for each unit within a population size category, and the mean square error for the deviation from the regression line was calculated

for each unit. The mean square errors of the units within each population size category were then averaged. The standard error of the estimate for each category was then determined by taking the square root of the average mean square error for that category. The results of this analysis are presented in Table 3.

Elimination of linear trends brought about significant reductions in the size of the standard errors of the estimate as compared to the observed standard errors. These reductions are greater in Taiwan and Costa Rica than in North Carolina. Similarly, the coefficients of variation based on the standard errors of the estimate eliminating linear trends are reduced compared to those based on observed standard errors. Despite this reduction, these coefficients of variation and standard errors of the estimate are still consistently higher than those predicted from the binomial model.

Estimates Using Two-way Analysis of Variance

The estimates derived so far have assumed homogeneity of the units within a population category. Under this assumption we have averaged the within unit variation to get a single index of variation. However, when the assumption of homogeneity of units is not satisfied, the true variance of rates will be over-estimated. On the other hand, the process of obtaining separate variances for each unit and then averaging the variances usually results in a reduced estimate of the variance compared to the one obtained by taking a single estimate of variance ignoring the unit classification. These two biases have conflicting effects which may not balance one another.

In order to eliminate both biases, it was decided to re-analyze the data eliminating the assumption of homogeneity of units and the averaging procedure. In this new procedure, a two way analysis of variance was performed with time and units as the two factors. (Appendix A, Assumption 4). This analysis of variance gives an estimate of the random variation in the rates after eliminating the unit and time variations from the total variation. The procedure also allows for testing for trends in time effects, and the deletion of variance associated with linear, quadratic, and cubic time trends.

The analysis was carried out only for Taiwan and the results are presented in Table 4. This analysis revealed that there were significant differences among the units for all the population categories considered. The results also showed that apart from significant linear time trends in all categories, there were four categories with significant quadratic effects and four categories with significant cubic effects. The standard error for a given category presented in Table 4 was obtained by taking the square root of the mean square error for that particular category.

The coefficients of variation in Table 4 show a pattern of decline (with some exceptions) as the population size increases. The estimates in general are larger than the estimates based on the binomial model (Table 1) and smaller than the observed standard errors (Table 2). A comparison of the rates for Taiwan in Tables 3 and 4 shows that the two-way analysis of variance procedure gives smaller estimates in the smaller population size categories and larger estimates in the larger population size categories.

DISCUSSION

The variances generated by actual data consistently display larger values than those predicted by the binomial model for hypothetical populations. We have explored some possible reasons for the over-estimates, but found them unhelpful in reducing the discrepancy. There are other possible reasons which need exploration

1. The binomial model predicts the variance for an infinite number of replications, while we have relatively small numbers of replications available. However, if this were the primary explanation, then the discrepancy between the predicted and the obtained variance should be inversely proportional to the number of replications available. This is not the case, as can be seen from comparison of Tables 1 and 3.

2. We have used crude birth rates, which include in the denominator all persons in the population. However, not everyone in the population is at risk of birth. Denominators should include only the number of women at risk of birth. Birth rates per thousand women at risk would have been preferable, but we did not have these data available. However, this cannot explain the discrepancies. Assuming that perhaps the number of women at risk is one-fifth the number of persons in the population, we would use the reduced denominators in both the empirical estimates and the binomial estimates. The discrepancies would be exactly the same size, but the size of population to which they applied would be one-fifth as large.

3. The mean square error of the crude birth rate (which is the square of the standard error of the estimate) is equal to the true error plus the correlation between error and time. If there is a correlation between error and time (a circumstance which we can rarely evaluate), the standard error of the estimate would be slightly larger than the true error.

4. The simple binomial model assumes that every woman is at the same risk of birth. Surely this is an erroneous assumption. If one assumes that the risk of birth varies, then the simple binomial model will underestimate the variance in birth rates. Consider the following example.

Suppose a population of 1,000 women with the probability (p) of a birth in a year of .01. Assuming p is constant for all women, the expected number of births in a year is 10, and the variance is equal to 1,000 x .01x.99=9.9, or variance in the birth rate of 9.9/1,000. Now assume that p varies among women with a mean of .01, and a variance of only .00001. The expected number of births is still 10, but the variance is now 19.9/1,000. (See Appendix B for the equation). By adding a very small variance to p, we have more than doubled the variance in the birth rate.

If we could decompose any population into sub-populations with the same probability of experiencing the criterion event, our estimates

would probably more closely approach those predicted by the binomial model. But from a practical point of view, this observation is of little assistance, since the circumstances under which we can either decompose the population into groups with the same p, or alternately, estimate the variance of p, are extremely unusual. Assuming that we are usually dealing with populations in which p has some unknown distribution, our predicted variances based on the simple binomial model seem doomed to be over-estimates.

SUMMARY AND CONCLUSIONS

This paper has explored the estimation of random variation in rates based on total enumeration of events. It is not concerned with variations due to sampling and response errors. Assessment of random variation in rates is necessary to detect changes with time as well as differentials in rates between regions or groups. It is necessary to determine minimum sample size needed to detect change or differentials, or minimum change in rates which cannot be attributed to random factors. It is necessary in establishing the size of statistical reporting units which will provide sufficiently stable rates for various purposes.

Several measures of random variation are presented. The variance generated by the most widely used binomial model displayed smaller values than any of those generated by our empirical data. We have identified difficult-to-eliminate sources of random variance which may make any empirically derived variance estimates substantially larger than those predicted by the binomial model. The use of the binomial model to estimate predicted variances against which to test null hypotheses can therefore be expected routinely to result in the rejection of null hypotheses which should in fact have been accepted. It will therefore lead to nonconservative assertions of true differences in rates where none in fact exist. If the experience with birth rates in other populations and the experience with other types of rates is similar to that we have presented, conservative inferences will require estimates of predicted variances made from detailed data on the actual population being studied.

APPENDIX A
Measures of Random Variation

Let b_{tk} denote the rate at time t (t = 1,2, . . .s) and for unit k (k = 1,2, . . .ℓ).

The following measures of random variation can be obtained.

Assumption 1. Rates are time homogeneous

A measure of random variation for unit k is given by (A. 1) $S^2_{1k} = \frac{1}{s-1} \Sigma(b_{tk} - \bar{b}_k)^2$, when $\bar{b}_k = \frac{1}{s} \Sigma b_{tk}$

Assumption 2. Rates are time and unit homogeneous

The following measures of random variations can be constructed.

(A.2) $S^2_2 = \frac{1}{\ell s-1} \Sigma_t \Sigma_k (b_{tk} - \bar{b})^2$ when $\bar{b} = \frac{1}{tk} \Sigma\Sigma b_{tk}$

(A.3) $S^2_3 = \frac{1}{\ell} \Sigma S_{1k})^2$

(A.4) $S^2_4 = \frac{(1}{\ell} \Sigma S_{1k})^2$

Assumption 3. Assume that the rates change with time. $b_{t,k} = B_{o,k} + B_{1,k}t + B_{2K}t^2 + \ldots + B_{2k}th + E_k$, when E_k is $N(0, \sigma^2)$.

Then an estimate of the variance of observed $b_{t,k}$ is given by the mean square error for the deviation from the best fitted regression line.

If all units are assumed to be homogeneous, then an improved estimate can be obtained by taking an average of the standard error obtained for each region.

Assumption 4. Rates are not homogeneous with respect to time and region. In this case, it is better to eliminate region and time effects from the total variation of the rates. This can be done using the analysis of variance technique. Using orthogonal polynomials one can also test for the linear, quadratic, cubin etc.. time trends of the rates. (For a standard reference, see Snedecor and Cochran.[11])

An estimate of the variance of the rate is obtained from the mean squares due to error in the analysis of variance table.

APPENDIX B
Variance in Binomial Model

Assume that p is the probability of occurrence of an event in a year for a member of the population. Then, for a population of size N, the observed rate will have an expected value of p, and variance p(1-p)/N.

Now assume that p varies among women with mean value of p* and variance σ^2_p. Then, it can be shown that the observed rate has an expected value of p* and variance equal to

$$V_{\hat{p}} = \frac{1}{N^2} [Np^*(1 - p^*) + N(N - 1)\sigma^2_p]$$

Note that, when N is large, the second term of the sum does not disappear.

FOOTNOTES

*Partial support for this project was provided through grants from the Ford Foundation and the National Institute of Child Health and Human Development, (Grant #HD05798) to the Carolina Population Center.

1. C.L. Chiang and F.E. Linder, "On the Standard Errors of Death Rates", (mimeographed) Population Laboratories, University of North Carolina at Chapel Hill, 1969.

2. N. Keyfitz, "Sampling Variance of Demographic Characteristics", Human Biology, 38, 1966, pp. 22-41.

3. N. Keyfitz, Introduction to the Mathematics of Population, Addison-Wesley, 1968.

4. J.E. Walsh, "Large Sample Tests and Confidence Intervals for Mortality Rates", Journal of the American Statistical Association, 45, 1950, pp. 225-237.

5. E.B. Wilson, "The Standard Deviation of Sampling for Life Expectancy", Journal of the American Statistical Association, 33, 1938, pp. 705-708.

6. L.L. Kupper and D.G. Kleinbaum, "On Testing Hypothesis Concerning Standardized Mortality Ratios", Theoretical Population Biology, 2, 1971, pp. 290-298.

7. L.L. Kupper, "Some Further Remarks on Testing Hypothesis Concerning Standardized Mortality Ratios", Theoretical Population Biology, 2, pp. 431-436.

8. B. Spencer, "Size of Population and Variability, of Demographic Data (17th - 18th centuries)." Paper presented at the annual meetings of the Population Association of America, April, 1975.

9. C.M. Suchindran, J.W. Lingmer, A.N. Sirha, and E.J. Clark, "Sensitivity of Alternative Fertility Indices," Proceedings of the Social Statistics Section of the American Statistical Association 1976, pp. 798-805, Washington, D.C.

10. The birth rate data from North Carolina were obtained from North Carolina Vital Statistics 1968, 1969, 1970, 1971, and 1972 published by the North Carolina State Board of Health, Public Health Statistics Division. Data from Costa Rica were obtained from the Republica de Costa Rica Estadistica Vital 1968, 1969, 1970, 1971, and 1972 published by the Departamento Estadisticas Sociales, Seccion Estadistica Vital. The Taiwanese data were obtained from Taiwan Demographic Fact Book 1968, 1969, 1970, 1971, and 1972 published by the Ministry of the Interior of the Republic of China.

11. G.W. Snedecor and W.G. Cochran, Statistical Methods, The Iowa State University, 1968.

TABLE 1. Estimated standard errors of birth rates based on binomial model

Data source	Population size category	Number of units	Average crude birth rate	Estimated standard error	Coefficient of variation
Costa Rica	0 - 5,000	4	23.6	2.216	9.4%
	5 - 10,000	12	29.5	1.903	6.5%
	10 - 15,000	22	32.4	1.526	4.7%
	15 - 20,000	7	34.4	1.360	4.0%
	20 - 30,000	6	32.0	1.050	3.3%
	30 - 40,000	2	33.6	.942	2.8%
	40 - 50,000	4	34.6	.820	2.4%
North Carolina	0 - 5,000	22	22.9	2.894	12.6%
	5 - 10,000	11	16.2	1.487	9.2%
	10 - 15,000	10	16.4	1.101	6.7%
	15 - 20,000	12	17.2	.946	5.5%
	20 - 30,000	17	17.9	.818	4.6%
	30 - 40,000	7	18.0	.703	3.9%
	40 - 50,000	9	17.7	.610	3.4%
Taiwan	0 - 5,000	14	33.2	3.010	9.1%
	5 - 10,000	10	29.7	1.934	6.5%
	10 - 15,000	12	29.7	1.455	4.9%
	15 - 20,000	19	29.3	1.263	4.3%
	20 - 30,000	20	28.0	1.045	3.7%
	30 - 40,000	9	28.9	.856	3.0%
	40 - 50,000	7	29.2	.785	2.7%

TABLE 2. Observed standard errors of birth rates

Data source	Population size category	Number of units	Average crude birth rate	Average standard error	Coefficient of variation
Costa Rica	0 - 5,000	4	23.6	4.375	18.5%
	5 - 10,000	12	29.5	4.782	16.2%
	10 - 15,000	22	32.4	3.593	11.1%
	15 - 20,000	7	34.4	3.577	10.4%
	20 - 30,000	6	32.0	3.485	10.9%
	30 - 40,000	2	33.6	2.910	8.7%
	40 - 50,000	4	34.6	5.551	16.0%
North Carolina	0 - 5,000	22	22.9	4.799	21.0%
	5 - 10,000	11	16.2	1.807	11.2%
	10 - 15,000	10	16.4	1.832	11.2%
	15 - 20,000	12	17.2	1.963	11.4%
	20 - 30,000	17	17.9	1.131	6.4%
	30 - 40,000	7	18.0	1.133	6.3%
	40 - 50,000	9	17.7	0.976	5.5%
Taiwan	0 - 5,000	14	33.2	4.082	12.3%
	5 - 10,000	10	29.7	3.171	10.7%
	10 - 15,000	12	29.7	2.338	7.9%
	15 - 20,000	19	29.3	2.951	10.1%
	20 - 30,000	20	28.0	2.649	9.5%
	30 - 40,000	9	28.9	2.217	7.7%
	40 - 50,000	7	29.2	2.345	8.0%

TABLE 3. Observed standard errors of estimate of birth rates eliminating linear trends

Data source	Population size category	Number of units	Average crude birth rate	Average standard error of the estimate	Coefficient of variation
Costa Rica	0 - 5,000	4	23.6	2.496	10.6%
	5 - 10,000	12	29.5	2.438	8.3%
	10 - 15,000	22	32.4	2.344	7.2%
	15 - 20,000	7	34.4	2.397	7.0%
	20 - 30,000	6	32.0	1.611	5.0%
	30 - 40,000	2	33.6	1.322	3.9%
	40 - 50,000	4	34.6	1.758	5.1%
North Carolina	0 - 5,000	22	22.9	3.910	17.1%
	5 - 10,000	11	16.2	1.564	9.7%
	10 - 15,000	10	16.4	1.444	8.8%
	15 - 20,000	12	17.2	1.552	9.0%
	20 - 30,000	17	17.9	0.987	5.5%
	30 - 40,000	7	18.0	1.062	5.9%
	40 - 50,000	9	17.7	0.806	4.6%
Taiwan	0 - 5,000	14	33.2	3.667	11.0%
	5 - 10,000	10	29.7	2.167	7.3%
	10 - 15,000	12	29.7	1.463	4.9%
	15 - 20,000	19	29.3	1.731	5.9%
	20 - 30,000	20	28.0	1.258	4.5%
	30 - 40,000	9	28.9	1.230	4.3%
	40 - 50,000	7	29.2	1.140	3.9%

TABLE 4. Observed standard errors of estimate for Taiwan birth rates eliminating time and unit effects

Population size category	Number of units	Average crude birth rate	Standard error	Coefficient of variation
0 - 5,000	14	33.2	3.558	10.7%
5 - 10,000	10	29.7	2.147	7.2%
10 - 15,000	12	29.7	1.390	4.7%
15 - 20,000	19	29.3	1.718	5.9%
20 - 30,000	20	28.0	1.533	5.5%
30 - 40,000	9	28.9	1.290	4.5%
40 - 50,000	7	29.2	1.365	4.7%