



State Center for Health Statistics
P.O. Box 29538 • Raleigh, NC 27626-0538
919/733-4728

STATISTICAL PRIMER

No. 12

April 1997

PROBLEMS WITH RATES BASED ON SMALL NUMBERS

by

Paul A. Buescher

INTRODUCTION

Most health professionals are aware that estimates based on a random sample of a population are subject to error due to sampling variability. Fewer people are aware that rates and percentages based on a full population count are also estimates subject to error. Random error may be substantial when the measure, such as a rate or percentage, has a small number of events in the numerator (e.g. less than 20). A rate observed in a single year can be considered as a sample or estimate of the true or underlying rate. This idea of an “underlying” rate is an abstract concept, since the rate observed in one year did actually occur, but it is this underlying rate that health policies should seek to address rather than annual rates which may fluctuate dramatically. The larger the numerator of the observed rate, the better the observed rate will estimate the underlying rate.

Many publications of the State Center for Health Statistics contain rates or percentages with a small numerator. This is a problem with a measure such as the infant mortality rate. In a single year many counties may have only one or two infant deaths and such rates in a small population may fluctuate dramatically from year to year. One means of addressing this problem is to look at five-year rates where the numerator will be larger. Even with five-year rates, however, many counties will have few events and therefore unstable rates.

Many cause-specific death rates for individual counties will have small numerators. This statistical problem is exacerbated when adjusted rates are produced since, in the process of calculating an adjusted rate by the direct method, the deaths and population are broken out and rates are calculated for a number of specific age or age-race-sex groups.



Some customers of the State Center for Health Statistics may treat our published rates and percentages as completely accurate. There is the danger of making unwarranted comparisons between geographic areas or comparisons over time when the rates or percentages have small numerators. We do not consider it feasible to completely suppress all rates based on small numbers. In one sense, they do describe what actually happened in a year, but there is the potential for misinterpretation when uncritical comparisons are made. The following section provides some methods for quantifying random errors in rates as a basis for making decisions about when changes or differences in rates are meaningful.

CALCULATION OF ERRORS IN RATES

The formulas presented here provide a means of estimating the confidence interval around a single rate or for determining whether the difference between two rates is statistically significant. A confidence interval is a range above and below an observed rate within which we would expect the “true” rate to lie a certain percentage of the time (often 95% is used). Calculation of a confidence interval recognizes that an observed rate is not a precise estimate of the underlying rate. These formulas are exactly the same as the ones used for a random sub-sample from a larger population. The rate measured for a population in a given year based on a complete count can be considered as a sample of one of a large number of possible measurements, all of which cluster in a normal distribution (bell curve) around the “true” (unknown) rate of the population. The larger the numerator of the measured rate, the better the rate will estimate the true or underlying rate of the population. These formulas assess only random measurement error. Systematic errors or biases in measurement may still be present and cannot be assessed by these formulas.

These formulas apply to any proportion or simple (“crude”) rate. Random errors may also be estimated for adjusted rates and other more complex measures, but a description of this is beyond the scope of the present Primer.

Proportions vs. Percentages vs. Rates

The formulas below are expressed in terms of p , or the **proportion** or fraction of a population that has a certain characteristic (e.g., death, low birthweight, early prenatal care). In this context, the terms proportion, percentage, and rate are interchangeable. For example, in 1995 Wake County had a resident population of approximately 518,000 out of which approximately 2,900 died during the year. The proportion who died is $2,900 / 518,000$ or .005598. For the percentage who died, multiply by 100; the result is .5598%. A percentage is simply a rate per 100. For a rate per 1,000, multiply the proportion by 1,000; the result is 5.598 deaths per 1,000 population. The number of deaths per 100,000 is 559.8. So the multiplier is completely arbitrary, though for rare events we usually use 1,000 or higher so that the rate is not a decimal fraction. The formulas presented below use p , or the proportion, so a percentage or rate has been converted back to the proportion (by dividing by the multiplier) in these examples. The formulas may also be applied with a percentage or rate.

Infant Death Rates

The infant death rates usually reported in publications produced by the State Center for Health Statistics, expressed per 1,000 live births, are not strictly proportions since the deaths and births are those occurring during a particular calendar year. Though approximately one-half of infant deaths occur on the first day of life, some of the infant deaths that occur in a given year are to babies born in the previous calendar year. The more technically correct way to compute the proportion of babies who die as infants would be to use the linked birth/infant death file to track a population or “cohort” of births through the first year of life. But in practicality this difference is small. We suggest that the formulas below may reasonably be used for infant deaths rates reported in the usual manner based on year of occurrence, expressed as the proportion of babies who die.

Confidence Intervals

We can compute a **confidence interval** around a proportion or rate, which is the interval within which we would expect the “true” rate to fall a certain percentage of the time. A 95% confidence interval is frequently used, which means using a multiplier (“Z” value) of 1.96. For a 99% confidence interval, one would use the multiplier 2.57. Let’s take the example of Nematode County with 20 infant deaths (d) out of a population of 1,900 live births (n) in a single year. The proportion dying (p) is $20 / 1,900 = .0105$. This is the same as 1.05 percent, or 10.5 per 1,000. The formula for the 95% confidence interval is:

$$p \pm 1.96 \sqrt{\frac{p q}{n}}$$

where $q = 1-p$, or in this case .9895. This formula works for any value of p, though for small values of p (.01 or less), the value of q is very close to 1 and may therefore be ignored. In the current example this calculates out to:

$$.0105 \pm 1.96 \sqrt{.0105(.9895) / 1900}, \text{ or } .0105 \pm .0046.$$

Expressed in the traditional way in terms of infant deaths per 1,000 live births, we can say that we are 95% sure that the true infant death rate for this population is between 5.9 and 15.1. These limits are quite large. A useful rule of thumb is that any rate with fewer than 20 events in the numerator will have a confidence interval that is wider than the rate itself. In the current example of a rate of 10.5 per 1,000 with a numerator of 20, the width of the confidence interval is 9.2.

Combining Data for Greater Precision

One way to reduce the error of a rate is to combine several years of numerator and denominator data. Another way is to combine geographic areas; for example, look at regional rather than county-level rates. In the example above, let’s assume that over a five year period in Nematode County we observed five times as many infant deaths and live births (100 and 9,500 respectively).

The five-year infant death rate would still be 10.5, but with the larger numerator, the range of the 95% confidence interval would be 8.5 to 12.5. Try the calculations so you can verify this result. In general, you have to quadruple the sample size (n) to cut the error in half.

Differences Between Rates

In many cases it is desirable to assess the statistical significance of a change in a rate over time, or of the difference between two rates in one period of time (for example between two geographic areas or population groups). The **standard error of the difference between two rates** is computed as:

$$\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

where p_1 and p_2 are the two rates to be compared, expressed as proportions. The difference between the two proportions may be regarded as statistically significant at the 95% confidence level if the difference exceeds 1.96 standard errors, as defined above.

As an example, take a county where the percentage of women who smoked during pregnancy (from the birth certificates) declined from 21.4% in 1990 to 16.7% in 1995. We want to know if this change is statistically significant at the 95% confidence level. In 1990, for 150 births (d_1) out of 700 total births (n_1) the mother smoked. In 1995, for 125 births (d_2) out of 750 total births (n_2) the mother smoked. The proportions are $p_1 = d_1 / n_1 = .214$ and $p_2 = d_2 / n_2 = .167$ (or 21.4% and 16.7%). The calculation of 1.96 standard errors of the difference for this example is as follows:

$$1.96 \sqrt{\frac{.214 (.786)}{700} + \frac{.167 (.833)}{750}} = .0404$$

Since the difference between the two proportions of .047 (i.e. $.214 - .167$) exceeds 1.96 standard errors of the difference (i.e. .0404), we can say that the decline in the smoking percentage in this county is statistically significant at the 95% confidence level. Or stated another way, the probability that the observed decline was due just to random variation in the percentages is less than .05 (or 5%).

The formula for the standard error of the difference can be used to solve for any unknown in the equation. For example, if one wanted to know the exact level of statistical significance of an observed difference between two proportions, solve for the multiplier (“Z”) by dividing the observed difference by the standard error of the difference and look up the probability value for Z in a table of areas under the normal curve. In the smoking example presented above, the probability that the observed decline would occur just due to random variation in the percentages is .02. For assistance with this or for other questions, contact the State Center for Health Statistics.

Other Issues

These formulas are based on parameters of the normal curve and in some cases will be only an approximation. If n (the denominator of the proportion or rate) is less than 30, or if the numerator of the proportion is less than 5, these formulas become less reliable and readers should contact the State Center for Health Statistics for more appropriate alternatives.

Another important consideration is the issue of practical versus statistical significance. If the n 's are large enough, almost any difference will be **statistically** significant. However, the same difference may be of very little practical or clinical significance. It is the responsibility of the user of statistics to evaluate whether observed differences, which may be statistically significant, are of real public health importance.

Finally, the issue of using rates versus numbers should be mentioned. Rates or proportions allow more standardized comparisons between populations of different size, but there may be substantial random measurement error involved. In many cases just looking at the number of events is appropriate; do not always rush to calculate a proportion or rate. If the number of infant deaths in a county increased from 1 in 1994 to 2 in 1995 and the number of births remained about the same, looking at the infant mortality **rate** would erroneously suggest that the problem had become twice as great. In this case, each infant death could be investigated as unique "sentinel health event." Examining the numbers behind the rates is always a good idea, and in some cases just looking at the numbers makes more sense.

This section on calculation of errors in rates demonstrates that an observed rate or proportion should not be taken as an exact measure of the true value in a population. Even measures based on complete reporting from a population may have a substantial random error component.

STATISTICAL POLICY

To address the problems of rates based on small numbers, the State Center for Health Statistics has adopted the following statistical policy guidelines:

- ▶ All publications of the State Center for Health Statistics that contain rates or percentages should contain a caution about interpreting rates or percentages based on small numbers. This caution should be featured prominently in the introductory material, and then discussed in more detail in the methods or technical notes section. See the *1995 NC Vital Statistics, Volume 1 and Volume 2*, for examples of this.
- ▶ Such a caution should accompany any information that is sent out to a customer as a special data request, if the information contains rates or percentages based on small numbers.
- ▶ When rates or percentages are published or distributed, the numerators should also be shown if possible.

- ▶ When adjusting rates by the direct method, only adjustment for age should be done. This divides the data into fewer cells and thus will increase the stability of the cell-specific rates, and thus the stability of the adjusted rate. Many western North Carolina counties have a very small minority population and sometimes one or two deaths in a specific minority age group causes a severely inflated age-race-sex adjusted death rate. There are also other reasons for not adjusting rates for race. When age adjusting death rates, ten age groups should be used, following the convention of the National Center for Health Statistics: 0-4, 5-14, 15-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75-84, 85+.

The State Center will continue to produce five-year age-adjusted death rates by county and cause of death for white males, white females, minority males, minority females, and total population. We will add the numbers of deaths to the reports we have produced in the past. Caution should be taken in using this information: look at the numbers of deaths as well as the rates, especially in counties with a small minority population.

- ▶ When maps of rates are produced, where possible there should be a legend of “interpret with caution” for rates or percentages based on a very small numerator, e.g. less than 10 events.
- ▶ Tables with rates or percentages should contain a footnote cautioning the user about problems of interpreting measures based on a small number of events. See the *1995 N.C. Vital Statistics, Volume 2* for an example of this.
- ▶ At every opportunity, customers of the State Center for Health Statistics should be educated about statistical issues, and especially about the potential for misinterpretation when comparisons are made using rates or percentages based on small numbers.

Readers with questions or comments about this Statistical Primer may contact Paul Buescher at (919) 715-4478 or through e-mail at paul_buescher@mail.ehnr.state.nc.us.

Department of Environment, Health, and Natural Resources
State Center for Health Statistics
P. O. Box 29538
Raleigh, N.C. 27626-0538
919/733-4728

BULK RATE
U.S. Postage
PAID
Raleigh, N.C. 27626-0538
Permit No. 1862