

Running the Numbers

*A Periodic Feature to Inform North Carolina Healthcare Professionals
About Current Topics in Health Statistics*

Paul A. Buescher, PhD

Errors in Rates and Percentages Based on Small Numbers

It is widely recognized that measures based on a random sample from a population are subject to sampling error. It is perhaps less well known that measures based on a complete count of events in a population are also subject to some degree of statistical 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 health programs should seek to address this underlying rate, rather than annual rates which may fluctuate dramatically.

Take a rather extreme case of the infant mortality rate (infant deaths per 1,000 live births) in a rural North Carolina county. Many small counties have only one or two infant deaths per year. If the number of infant deaths in this county increased from 1 in 2001 to 2 in 2002 and the number of births remained about the same, the doubling of the infant mortality **rate** would erroneously suggest that the problem had become twice as great. Examining the numbers behind the rates is always a good idea, and in some cases just looking at the numbers makes more sense.

It should be noted that a percentage is simply a rate per 100. In 2001, Wake County had a resident population of approximately 658,000. During that year, 3,341 residents of Wake County died. The proportion who died is $3,341 / 658,000$ or .005078. For the percentage who died, multiply by 100; the result is .5078%. For a rate per 1,000, multiply the proportion by 1,000; the result is 5.078 deaths per 1,000 population. The number of deaths per 100,000 population is 507.8. 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.

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. A useful rule of thumb is that a rate with 20 events in the numerator will have a 95% confidence interval width approximately the size of the rate itself. For example, if there were 20 infant deaths in a county out of a population of 1,900 live births, the infant mortality rate would be 10.5 infant deaths per 1,000 live births. The 95% confidence interval is 10.5 plus or minus 4.6 (lower and upper limits of 5.9 and 15.1, or a width of 9.2). For details on the calculation of simple confidence intervals, see *Statistical Primer* No. 12 of the State Center for Health Statistics at <http://www.schs.state.nc.us/SCHS/pubs/title.cfm?year=1997>

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 general, you have to quadruple the population to cut the error in half.

From the State Center for Health Statistics
www.schs.state.nc.us/SCHS
North Carolina Department of Health and Human Services