Tools of the Trade:

About Bayesian and Nearest Neighbor Rates

Three special rates can be used to analyze the data when the number of events is small (3 to 10) - Bayesian Smoothed, Bayesian Nearest Neighbor, and Mean Nearest Neighbor rates. These rates are especially designed to take into account the unique problems associated with highly variable rates based on small numbers of events and are particularly suited for spatial or mapping analysis. When age adjusted and age specific rates are based on less than 10 events the Bayesian/nearest neighbor rates offer the advanced health statistics user an alternative for health status assessment.

The maps produced in our interactive web site, Epidemiologic Query and Mapping System (EpiQMS), include Bayesian Smoothed, Bayesian Nearest Neighbor, and Mean Nearest Neighbor rates in the Map and County Profile modules for the Births, Deaths, and Cancer Incidence datasets. Please note that these rates are not available in the Table or Chart modules. In EpiQMS, these rates are reliable enough to be calculated and shown when the number of events is 3 or higher. With all other rates/ratios/percents, the number of events must be 10 or more to be calculated and shown. Click on the EpiQMS logo below to enter this site. However, we recommend that you review the discussions below regarding the calculation and appropriate use of these special rates to determine their value in your analysis/assessment.

Bayesian Smoothed Rate

The empirical Bayesian Smoothed Rate (BRATE) is a sophisticated tool for smoothing maps. However, you do not have to understand the details to use it. A general idea of how it works is sufficient.

Smoothing refers to using some process to level off the excesses on a map with respect to rates, especially rates that are unstable (based on small numbers of events) and may frequently appear very high but have very wide confidence intervals which often happens in areas with small populations. Rates can vary wildly, way out of the expectation of epidemiologists, other things in a county being equal. Bayesian smoothing deals with small number problems and offers the analyst some idea about whether a rate is valid in an area or not. However, it is not possible to develop a simple decision algorithm. To use Bayesian smoothed rates, one must develop some intuition about other disease rates in the area of interest and the geography of risk factors in the area.

For example, if men over the age of 85 are dying from complications of hip fracture in county A at a rate of 50,000/100,000 this would certainly be a cause of great alarm in the health

department. Of course, investigators would be quick to realize that 1 of the 2 men in that county over age 85 had died from hip fracture complications.

However, unstable rates can present themselves in more subtle ways. Suppose an SMR (standardized mortality rate) is 125, i.e. 25% higher than the overall state rate but it is in a small county for a relatively rare disease. Suppose the confidence interval does not span 100 indicating that indeed, the rate is significantly higher than the state rate. Can this indication of a significantly higher county rate be trusted?

Empirical Bayesian smoothing offers a tool to approach finding an answer. The details are discussed in (Marshall 1991; Devine, Louis et al. 1994; Lawson, Böhning et al. 1999; Louis and Shen 1999). This discussion closely follows (Bailey and Gatrell 1995)

Bayesian methods depend on a prior distribution of the disease rates that is also multiplied with a likelihood function to get a posterior distribution from which the Bayesian rates are determined. The prior distribution is determined from prior knowledge concerning how the data are distributed based on prior facts or even intuition and experience. The prior distribution in the case of empirical Bayes is the data itself. In the case of BRATE, all the counties' disease rates in the state are used to develop a prior distribution. The formulation presented here is a 2nd moment approximation that is accurate enough for our purposes.

Suppose we have an unknown rate which might be the true rate

$$\hat{T}r_j = true\ rate \quad j = 1 \quad to\ number\ of\ counties$$

Then it can be shown that an estimate of the true rate is

$$\hat{T}r_j = \hat{\gamma} + \frac{\hat{\phi}(r_j - \hat{\gamma})}{(\hat{\phi} + \hat{\gamma}/n_j)}$$

where

$$\hat{\gamma} = \frac{\sum all\ events}{\sum population}$$

is a pooled mean of the observed rates, the state rate. The weighted sample variance is

$$\hat{\phi} = \frac{\sum n_i (r_i - \hat{\gamma})^2}{\sum n_i} - \frac{\hat{\gamma}}{\bar{n}}$$

where $\bar{n}$ is the average population across all areas or counties and i = 1 to number of counties. Those counties that have larger populations will be adjusted less than those that have smaller populations.

What happens is that the observed rate r_j is shrunk towards a mean value resulting in $\hat{T}r_j$, an estimate of the "true" value. Low values for r_j can get larger and high values smaller. If an area has a large variance, then the best estimate for that area's rate is near the state rate $\hat{Y}$. If the variance is small, for example in a populous county, then $\hat{T}r_j$ will be close to r_j, the rate calculated the usual way

Note: BRATE is calculated for age and sex stratifications and then combined in the usual way to produce age adjusted empirical Bayesian rates. Confidence intervals are not presented because Bayesian confidence intervals have a completely different interpretation than the usual confidence intervals.

Nearest Neighbor Bayesian Smoothed Rate

As an extension to the Bayesian Smoothed Rate where the correction of a rate uses all the counties to develop the estimate to the "true" value, the Nearest Neighbor Bayesian Smoothed Rate presents the correction or shrinkage factor based only on the nearest neighbors, that is, those counties that share a common border, not the entire state.

This might make more sense since with respect to assessment of a disease rate because rates for county A would be adjusted by using information from counties B, C, D, and E which because they are closer, are more likely to be similar to A. Whether this makes sense or not, is left to the analyst who must make that judgment based on local geographic and demographic knowledge.

Mean Nearest Neighbor Rate

This rate is a simple rate, age adjusted or age and sex stratified using the pooled counts and pooled population for counties that share a boundary with a particular county. For example, if county A has neighboring counties B, C, D and E; then, age stratified and age-adjusted rates by sex and race are calculated from the total events in counties B, C, D and E and the total population in counties B, C, D and E.

The purpose of this rate is to give some indication of the rates around county A. It would be a smoothed rate if the events and population of county A were included as well.

If the Mean Nearest Neighbor Rate for a county is different and the confidence intervals of the nearest neighbors and county A do not overlap, then, to a first approximation, they are statistically different. Often the Mean Nearest Neighbor Rate has a narrower confidence interval than county A because the pooled number of events and population is higher. But this may not always be the case, such as when an urban county is surrounded by rural counties

Other than just indicating that nearest neighbors may have higher or lower rates, the main use of this rate is to give some idea of the variability of the rate for the single county A. Small numbers, or better, unstable rates permeate public health assessment. The purpose of presenting the rates for the adjacent counties is to help you understand to some degree how unstable the rate is. If county A and its neighbors share similar characteristics with respect to urban or rural and

socioeconomic characteristics or, in general, risk factors important in the disease being investigated, and county A's rate is very different than its neighbors, then this might be an indication that further investigation is warranted.

Interpretation of Bayesian/Nearest Neighbor Rates

When one makes a map of rates with the usual crude or age-adjusted/specific rate calculation, there may appear hotspots of high rates but with a high standard error, that is, unstable rates. For any particular county that appears to have an elevated rate, the question remains -- if the variability is high (high standard error, wide confidence intervals), is the disease rate really elevated or is the rate a statistical artifact?

One way to approach this question is to look at maps of the original rates, the nearest neighbor rates, the Bayesian rate and the nearest neighbor (often called spatial) Bayesian rate and see how they compare. If county A has a rate higher than the pooled rate of its neighbors and the populations or other factors are equal, then this may warrant concern. If the Bayesian rate for county A shrinks to the state rate and the spatial Bayesian rate does the same, then perhaps the rate in county A is more about high variability than about a truly elevated rate. However, if maps of Bayesian rates and spatial Bayesian rates still show elevated rates, then, this may be an indication that the rates are really elevated in spite of being based on a small number or for a sparsely populated area.

References

Bailey, T. C. and A. C. Gatrell (1995). *Interactive spatial data analysis*. Essex, England, Longman Scientific & Technical.

Devine, O. J., T. A. Louis, et al. (1994). "Empirical Bayes methods for stabilizing incidence rates before mapping." *Epidemiology* **5**(6): 622-630.

Lawson, A. B., D. Böhning, et al. (1999). Disease mapping and its uses. *Disease Mapping and Risk Assessment for Public Health*. R. Bertollini. Chichester, West Sussex, P019 1UD, England, John Wiley & Sons Ltd.**:** 3-13.

Louis, T. A. and W. Shen (1999). "Innovations in Bayes and empirical Bayes methods: Estimating parameters, populations and ranks." *Statistics in Medicine* **18**: 2493-2505.

Marshall, R. J. (1991). "Mapping Disease and Mortality Rates using Empirical Bayes Estimators." *Applied Statistics* **40**(2): 283-294.