Statistical Approaches for Small Numbers: Addressing Reliability and Disclosure Risk

NAHDO-CDC Cooperative Agreement Project
CDC Assessment Initiative

December 2004

THE NATIONAL ASSOCIATION OF
HEALTH DATA ORGANIZATIONS

Acknowledgements

The National Association of Health Data Organizations (NAHDO) developed this document with funding and support from the Centers for Disease Control and Prevention (CDC) Assessment Initiative. We are grateful for the leadership and guidance provided by the CDC Project Team, Patricia Schumacher and Alex Charleston.

The lead author of this document is Barbara Rudolph, Ph.D., NAHDO's Data and Research Senior Scientist Consultant. Dr. Rudolph integrated documents and research conducted by Michael Stoto, Ph.D., Harvard University; and Luis Paita, Ph.D., NAHDO's former Deputy Director. Guidance for this paper was provided by the NAHDO-NAPHSIS Expert Panel consisting of Garland Land, Missouri Department of Health and Human Services; Lois Haggard, Ph.D., Utah Department of Health; David Solet, Ph.D., King County Health Department; Gregory Crawford, Kansas Department of Health and Environment; and Daniel Goldman, M.D., M.P.H., Vital Net. We thank the Assessment Initiative states and the private and public state health data organization staff who selflessly provided input and shared their data dissemination practices with us.

The NAHDO-CDC Cooperative Agreement is a joint project promoting collaboration and partnership with the National Association of Public Health Statistics and Information Systems (NAPHSIS). As national associations representing statewide population-based health data bases and the agencies that maintain them, NAHDO and NAPHSIS, with support from CDC, are working to build data and technical capacity at the state and local levels.

Table of Contents

Statistical Approaches for Small Numbers: Addressing Reliability and Disclosure Risk

A. Introduction

Public health data when queried or displayed in web-based data tables can often have cells with a small number of individuals or events especially when the query is focused on small geographic areas (Zip codes) rare events, population subgroups, provider groups, payers, or other small samples. The primary statistical concern is reliability of results from queries in which the results contain small cell sizes or a small underlying population. Without some intervention to increase cell size or population, or the interpretation of the results, there may be misinterpretation by the user. It should be noted however, that the definition of "small" varies across political boundaries; the database, the state, and the application often influence how the term "small" is defined.

Most public health professionals are aware of reliability problems resulting from small numerators; fewer consider a small denominator as a contributor to poor reliability of the data. Web-based data systems' developers should be aware that the reliability of rates based on case reports where the denominator is from a smaller population will be affected negatively. "For example, if *x* forms the numerator of a rate *p,* population = *n, w*hen p is small Var (*p*) = Var (*x*/*n*) = *p*/*n,* the resulting standard deviation for the rate is significantly larger in smaller populations. In the table below we see that a denominator with 100 cases results in a less reliable rate than one with 10,000 cases where both have the same case numerator."[1]

x = 4,	n = 100 or 10,000	
x = 4	n = 100	SD(p) = √0.04/100 = 0.02
x = 4	n = 10,000	SD(p) = √0.0004/10,000 = 0.0002
p = 0.04,	n = 100 or 10,000	
p = 0.04	n = 100	SD(p) = √0.04/100 = 0.02
p = 0.04	n = 10,000	SD(p) = √0.04/10,000 = 0.002

Table by Michael Stoto[2]

While small cell size is a concern for most public health statistical publications, it is more acutely so in web-based data dissemination systems for several reasons. First, because web-based data dissemination systems are particularly desirable for immediate answers to questions about the public's health, and generally, the users of the systems are interested in data for small geographical areas and other small groups of individuals. Second, the

[1] Statistical Issues in Interactive Web-based Public Health Data Dissemination Systems." Michael Stoto, RAND, paper developed for NAPHSIS, Sept 19, 2002, p. 18.
2 Ibid.

information reaches a much broader audience than a paper publication, and often this includes individuals without statistical or epidemiologic training. Third, web-based systems generally provide less documentation on how to interpret the results than do paper publications which usually provide extensive bibliographies, appendices, footnotes, caveats, etc. The web-based data dissemination systems (WDDS) that attempt to provide documentation still have the issue of varying types of queries—it may be difficult to direct the user to the appropriate documentation.

Small numbers are of also of great concern when reporting sensitive information that might lead to violation of individuals' right to anonymity and privacy with respect to attributes that are typically stigmatized. While this guideline is primarily focused on data reliability, it also provides statistical approaches that can better assure anonymity and privacy.

This document is part of a set of guidelines supported by the CDC Assessment Initiative and designed to assist data managers, epidemiologists, and analysts in public health when releasing public health data on the web. These guideline sets include: Statistical Approaches for Small Numbers**:** Addressing Reliability and Disclosure Risk, Security of Data for Web-based Data Dissemination Tools, and Management and Institutional Controls for Reducing Disclosure Risk in Web-based Data Dissemination of Public Health Data; as a package, the guideline set will assist in assuring reduction of risk of inappropriate disclosure of sensitive information, meaningful statistics, and security of data. We have attempted in the guideline set to artificially isolate methods to reduce redundancy across the guideline set, however, in practice one would use methods from each of the guidelines.

This specific guideline will address statistical approaches for releasing public health data on web-based dissemination systems; approaches that impact on the reliability of statistical tests and/or, at the same time protect individuals from disclosure of sensitive information.[3] The first part of the document covers methods operating as modifiers of the data in the underlying database—the section is titled "Summary of Data Modification Statistical Approaches for Addressing Small Cell Size. First, there is a summary of the approaches, followed by individual sections discussing in greater detail each approach. The second major section is titled "Summary of Formal Statistical Approaches to Improve Interpretation of Results from Small Cells." This section includes statistics for modifying the interpretation of the results of statistical tests. It too, has a brief summary, following by descriptions of the individual approaches. Within each description we offer references, web-sites where the approach has been used and contact information.

[3] Non-statistical approaches for reducing disclosure risk are found in the "Management and Institutional Controls for Reducing Disclosure Risk in Web-based Data Dissemination of Public Health Data" guideline.

B. Summary of Data Modification Statistical Approaches for Addressing Small Cell Size

There are a variety of approaches for increasing the reliability of statistical tests in situations where cell sizes are small; these statistical techniques address the issue with both modifications to existing data and the use of synthetic information to achieve larger cell or population sizes.

The "data modification" statistical approaches include the following:

1. Aggregation or combining results over geographic areas, or multiple years, or subgroups (e.g., age groups) is done in order to achieve a larger denominator that will produce a larger result in the table.
2. Statistical Noise/Data Perturbation. Introduction of uncertainty to all cell values in a table less than a pre-determined threshold (e.g., < 10 observations). To implement one can add or multiply the values of a continuous data element by a randomly-determined factor.
3. Smoothing Techniques--including maximum likelihood, simple weighted averages, and the moments methods (multivariate signal extraction) are all classified as approaches for smoothing or signal extraction. These are designed to improve the reliability of the estimates by removing noise from the data.
4. Other Bayesian Methods--small area model-based estimation and bootstrapping. Essentially these techniques impute information from either direct or indirect sources to create a new estimate for the geographic area, or for a specific demographic characteristic. Bayesian methods are primarily used for improving data reliability, but also serve to reduce disclosure risk.

Aggregation of results is the most commonly used statistical approaches to address small cell sizes; the other approaches are more complex and require more statistical sophistication to implement and potentially more time to compute. This latter point will impact a dynamic web-based query system—users will wait only seconds, rather than minutes or hours for computations to occur.

We will describe in greater detail each of the data modification and interpretation approaches—indicating their strengths and weaknesses. Also provided are public health agencies/systems that have utilized the method described. In addition, we provide the user with useful references for further investigation.

C. Data Modification Approaches

1. Aggregation This approach combines results over geographic areas, or multiple years, or subgroups (e.g., age groups) in order to achieve a minimum number in the combined cell. For example, if the results for a 5 year age group (ages 1-5 years of age) do not yield an adequate number of cases for statistical testing, then the age group is extended to cover more ages (1-10 years of age). This however, precludes an analyst from drawing any conclusions about the children age 5 and

under. In rural areas, it is very difficult to construct grants or planning documents for specific health care conditions, due to small cells. One cannot make reliable statements for 1 or 2 cases of cancer, or conditions potentially resulting from environmental factors. So, to achieve that reliability there is a loss of detail. If aggregation isn't used, then results are often blocked from display via cell suppression techniques and the utility of the data is reduced significantly. These cell suppression methods are described in greater detail in the guideline "Management and Institutional Controls for Reducing Disclosure Risk in Web-based Data Dissemination of Public Health Data".

The actual number of cases required for statistical reliability may vary by type of statistical test used, by the alpha level, and whether confidence intervals or coefficients of variation are produced. The other issue is how the results are going to be used, that is, if the results are for exploratory investigation versus determining how millions of community dollars will be spent. Clearly, the more important the results are—the greater the reliability that is required.

Agencies Using This Approach:

It is likely that all web-based systems have been designed to include some aggregation of results.

Strengths and Limitations of Approach:

This approach results in loss of information, e.g., aggregating data for two or more entities may hide differences between or among those entities. Yet, aggregating data for multiple years provides more stable estimates—estimates that are less likely to be caused by random variation. But, the data is consequently older when aggregating across years.

While aggregation is relatively easy to do, often in dynamic web-based query systems it still may take the user a number of tries to select age groups or geographic units to achieve the cell size that is necessary. The return of information is generally limited by suppression algorithms to avoid any disclosure.

Some systems are designed to avoid this by pre-aggregations within the database. This may at times unnecessarily reduce information for more common events or conditions. It does, however, assure that you will actually receive information in cells for statistical testing. For example, a web-based data dissemination system could be designed to pre-aggregate age groups to reach a minimum of 30 individuals in a cell. This would require advance programming; and perhaps, loss of information for certain types of questions in the upper and lower ends of age categories.

2. Standard Data Perturbation Methods – Statistical Noise, Data Swapping and Controlled Rounding to Reduce Disclosure Risk

These Bayesian methods include the addition of statistical noise, data swapping and controlled rounding. The addition of statistical noise to a data file, data swapping or controlled rounding results in "pseudo" information that reduces disclosure risk. Other Bayesian methods described in the next section use synthetic data to increase the size of small cells, thereby increasing data reliability.

There are two types of "noise" that can be introduced into data, natural noise and statistical noise (Zarate, 1999). Natural noise consists of errors in the data, such as: coding errors, keying errors, or missing data. Statistical noise is introduced into the data to add uncertainty to all cell values in a table that are less than a prescribed threshold, such as < 10. When the results return 4 cases and the threshold is 10, an addition is made to some case values in the table. This results in blurring of the data, assuring that identifying an individual within the data is protected to a degree of certainty. While the blurring does change the data, the simple statistics and distributions remain the same. The add-on factor is not made available to data users.

"Controlled tabular adjustment" (CTA) is another form of statistical noise—which can be used in two-dimensional tables to reduce disclosure risk, as an alternative to cell suppression. Unfortunately, it cannot be extended to tables with three or more dimensions (Ernst 1989). It relies on a probability measure for rounding "down" or "up" for each of the table cells and uses a mathematical programming approach called a *stepping stones algorithm*. Using an "unbiased" controlled rounding approach will preserve original values with respect to the statistical criterion expectation. The results deliver the same statistical distribution, assuring reduction of disclosure risk (Cox, 1987). The National Center for Health Statistics has funded the development of software for tabular data protection using controlled rounding and a method to preserve additivity of the sub-totals along one of the dimensions (rows or columns.) The software uses a synthetic substitution for replacing a cell value; it substitutes the current value of the cell with its "closest safe value" and uses linear programming to adjust other cells to preserve additivity (Gonzalez and Cox 2004).

"Data swapping" is a method that swaps information from one individual within the same sample to another individual with similar characteristics in the sample. This results in "pseudo-cases." The individual records (after swapping) do not represent any one individual, yet these pseudo cases still produce the same simple statistics and distributions as those produced by the original data. This allows for display of small cell sizes without risk of individual identification.

An example of data swapping can be found in the Census Bureau's "confidentiality edit." The Census Bureau developed the "confidentiality edit" (CE) to prevent the disclosure of personal data in tabular presentations. The CE selects a small sample of cases and interchanges their data with other cases which have same characteristics on a pre-selected set of variables but who live in different geographic locations (Jabine, 1993).

One survey currently uses multiple imputation methods; the Survey of Consumer Finances, which is conducted by the Federal Reserve Board and holds sensitive

financial information from a high-wealth population, has been a test bed for this type of method. This method is computationally intense, and is not likely to be applied in any web-based data dissemination system that allows for dynamic queries.

Any of these standard data perturbation methods add expense to the preparation of a file for web-based data dissemination, but for web-based micro-data files, this method can be useful in preventing users from matching the database with other databases explicitly identifying the individuals in the second database. Thus, the perturbation methods are useful for protecting confidentiality of the data, as well as increasing data reliability.

References:

Cox, L.H. (1987) "A constructive procedure for unbiased controlled rounding." Journal of the American Statistical Association, 82: 520-524. (lcox@cdc.gov)

Cox, L.H. (2003) Balancing Data Quality and Confidentiality for Tablular Data. An Invited Paper for the United Nations Statistical Commission and Economic Commission for Europe Conference of European Statisticians. Luxembourg, April 2003. (http://www.unece.org/stats/documents/2003.04.confidentiality.htm)

Ernst, L.R. (1989) "Further applications of linear programming to sampling problems." Technical Report—Census/SRD/RR-89-05. Washington, D.C., US Census Bureau.

Gonzalez, J.F., and L.H. Cox. "Software for Tabular Data Protection." Slides dated September 29, 2004. (lcox@cdc.gov)

Jabine, T.B., (1993) "Statistical Disclosure Limitation Practices of United States Statistical Agencies." *Journal of Official Statistics*, Vol 9, No.2, pp 427-454.

National Research Council. (2000) "Improving Access to and Confidentiality of Research Data: Report of a Workshop." Committee on National Statistics, Christopher Mackie and Norman Bradburn, Eds., Commission on Behavioral and Social Sciences and Education, Washington, D.C.: National Academy Press.

Zarate, A.O. "The ICDAG Checklist on Disclosure Potential of Proposed Data Releases—A Tool for Disclosure Review." A presentation at the NCHS National Conference on Health Statistics, August 2, 1999.

Agencies/Systems using approach:

U.S. Census Bureau
National Center for Health Statistics

Strengths and Weaknesses of the Approach

Perturbation methods such as the addition of statistical noise, data swapping, and controlled rounding limit disclosure risk while maximizing information available to the user. Although it may distort the actual information it maintains the statistical distribution. The distortion however, may result in misinterpretation by users and produce unnecessary concern about specific health conditions or environmental risks—when examining cell sizes that are "rounded up."

Should perturbation methods (if less computationally intense) be applied to data in a web-based data dissemination system? It appears that they may work for some simple statistics, but could be problematic depending on the use of the results. Given that many social scientists are skeptical of analyses that are not based on the original data, the perturbation methods should be applied last, when other approaches cannot prevent disclosure (National Research Council, 2000). And if such methods are used, there must be greater effort to education and convince the data users that the key properties of the data are preserved, even with the addition of statistical noise (imputation).

3. Data Smoothing for Improving Reliability

Data smoothing is a technique that adjusts for differences in the reliability of data resulting from small cell sizes. The information taken from tabular cells that hold small numbers are less reliable than information taken from those cells that hold larger numbers. There are a variety of approaches to producing smoother data, including maximum likelihood, simple weighted averages, and the moments methods (multivariate signal extraction). They can all be classified as approaches for "smoothing" or signal extraction.

Geographic Smoothing Methods

Geographic smoothing techniques have often been used in conjunction with the creation of disease incidence rate maps, where the raw rates for rare events (such as cancers) are unstable for regions with small populations at risk. Rather than report small numbers for a specific geographic region (Zip code or census block), typically disease incidence has been reported only as a summary count or rates for a larger defined region, such as county. Geographic smoothing techniques can be used to produce counts for smaller geographic regions with small numbers at risk. The statistical models draw upon the "strength offered by adjacent geographic areas" to create more stable estimates for the small area. The smoothing methods may also rely upon Bayesian or empirical Bayesian modeling approaches described below.

In terms of actual use of this type of information, while it has not yet become normative for public health systems to utilize smoothing techniques in their web-

based data dissemination systems, there are several systems that do currently rely on geographic smoothing methods.

Recently, HCUPnet has utilized geographic smoothing in their risk-adjustment methodology for reporting hospital indicators. While not the same usage, this methodology was again used with the idea that small cell sizes could be made more reliable during the development of the risk-adjustment methodology. There is a report available on this technique on the AHRQ website (www.ahrq.gov) under HCUP.

The State of Washington's EPI QMS system uses both smoothing and Bayesian methods for addressing small cell sizes. These included adding additional data from other years, or drawing on data from surrounding "neighborhoods." This system is designed primarily for epidemiologists and not the lay public, although some data is available to the public.

Agencies/Systems Using Geographic Smoothing Approach

Utah Department of Health, Indicator Based Information System (IBIS)
Contact: Lois Haggard, Ph.D., Utah Department of Health
loishaggard@utah.gov

State of Washington--Epidemiologic Query and Mapping System (EPI QMS)
https://fortress.wa.gov/doh/epiqms

State of Washington—VISTA system
GIS and Spatial Epidemiology
www.doh.wa.gov/OS/Vista/HOMEPAGE.HTM
Contacts: Dick Hoskins
reh0303@hub.doh.wa.gov
360-236-4270

AHRQ, HCUPnet http://www.ahrq.gov/data/hcup/
Uses smoothing techniques in the QIs.

4. Bayesian Methods[4] for Improving Reliability

Statistical procedures have been developed to address small numbers in sample data. These procedures draw upon Bayesian methods and include small area estimation (see for example, Shen and Louis, 1999 & 2000). Essentially, these techniques impute information from either direct or indirect sources to create a

[4] Bayesian statistics rely heavily on the formulation that "posterior is proportional to prior times likelihood." This translates to—the basis of various alternative hypotheses is knowledge at a particular point in time, modifying those hypotheses is based on collecting new information from relevant data to arrive at "posterior probabilities," essentially being able to predict both sensitivity and specificity of the estimates.

new "estimate" for the geographic area, or for a specific demographic characteristic. The techniques use information from other sources or from population and cell averages, replacing the sample data by using only the new information, or by averaging the new information with the sample information. The latter, is called a composite estimate (Ghosh and Rao, 1994)[5], it creates new estimates that are based on the mean of the sample data and the other external source's mean. Sometimes the estimates are also weighted. Using these methods assumes that the other source of information is an equal or better representation of the population than the sample. While sampling statisticians have used these techniques for some time, they have not been widely used in public health web-based data dissemination systems, because calculating the variance for these techniques is quite complex. For model specifications, see the Ghosh and Rao, 1994, article.

Bayesian modeling approaches are used to address the reliability of data (given small cell sizes) and to predict a better estimate using information from prior quarters or years of data or other sources, thereby reducing the variability or error from estimations of the small cell value based on using only the population mean. Census Bureau researchers (Fay and Herriot, 1979)[6] used Bayesian methods with census data; they proposed that an estimate of per capita income (PCI) for a small place in the census could be a weighted average of the census sample estimate and a "synthetic" estimate obtained by fitting a linear regression equation to the sample estimates of PCI, using other data sources for the independent variables, such as county averages and tax-return data. The Census Bureau adopted this approach for estimates of PCI in small areas in 1974. The National Center for Health Statistics also adopted this synthetic approach for creating state estimates of disability for the National Health Interview Survey data.

A Bayesian modeling approach could be used for small cell size estimates in hospital discharge data reporting by taking information from previous years for the variable of interest, and using this synthetic estimate along with a weighted average of the current year. Estimation can occur using a fairly general Bayesian regression model. However, the Bayesian methods may pose a challenge for dynamic web-based dissemination systems, given the computational time for these estimations. And, in some cases where normality is violated, the models may not assign the appropriate weights. In order to assess whether the Bayesian method used is appropriate, various regression diagnostics may be necessary. Software, such as: LISREL 8.0[7] allows an assessment of the reliability of the prediction and thus could provide a test of whether reliability was increased using

[5] Ghosh, M. and Rao, J.N.K. (1994). "Small area estimation: an appraisal" (with discussion). Statistical Science, 9, pp 65-93.
[6] R. E. Fay and R. A. Herriott. (1979) "Estimates of income for small places: an application of James-Stein procedures to census data." Journal of the American Statistical Association, 74, pp. 269-277.
[7] K. Joreskog, D. Sorbom, S. duToit, and M. duToit. (2000) LISREL 8: New Statistical Features, Scientific Software International, Lincolnwood, Illinois.

Bayesian methods.[8] But, integrating software like this would be difficult in a dynamic web-based system.

a. Bootstrapping Approaches

Bootstrapping approaches for estimating various parameters from the sample for the purpose of studying the mean and variance of these parameter estimates can be used for ascertaining a reliable estimate of a small cell in a table. Monte Carlo techniques (a form of bootstrapping) are essentially computer-generated data based upon the available sample. There are a variety of software packages that provide for this approach. Bootstrapping allows you to produce estimates of standard errors by repeated random sampling (with replacement) from the available sample.[9] Generally, users of this technique will draw anywhere from 100 to 1000 sub-samples from the existing data to generate the estimates. While the procedure itself is not as complicated as some of the other Bayesian methods, it does require substantial computing resources. For example, Waller et al., (1997)[10] stated that a model using bootstrapping (with 500 iterations) to estimate disease incidence in a geographic area took 20 minutes on a Sparc10 workstation, thus this approach may be untenable for dynamic web-based data dissemination systems, where responses to queries should not take longer than 1-2 minutes. This could be resolved, however, with faster computers, and a static response (to the dynamic query) based on pre-aggregated tables.

References:

"Statistical Issues in Interactive Web-based Public Health Data Dissemination Systems." Mike Stoto, RAND, presentation at NAPHSIS Meeting, New York, NY, June 2003.

Shen W, Louis TA (1999). Empirical Bayes Estimation via the Smoothing by Roughening Approach. J. Computational and Graphical Statistics, 8: 800-823.

Shen W, Louis TA (2000). Triple-Goal estimates for Disease Mapping. Statistics in Medicine, 19: 2295-2308.

[8] If you want to predict the contents of a small cell, predictions can be estimated using general Bayesian regression models. Alternative Bayesian approaches use covariance matrices and the likelihood function in multilevel models, where the actual value is the fixed part of the model, the random component is the estimated population parameter, or predicted cell count. With this approach variances from the estimated or predicted cell count can be estimated using software such as: LISREL 8.0, this software allows an assessment of the reliability of the estimated prediction.

[9] W. Paul Vogt (1993) Dictionary of Statistics and Methodology: A Nontechnical Guide for the Social Sciences. SAGE Publications, Newbury Park, NJ.

[10] L.A. Waller, B.P. Carlin, H. Xia and A.E. Gelfand (1997). "Hierarchical Spatio-Temporal Mapping of Disease Rates." Journal of the American Statistical Association, 92 (438:607-617.

Strengths and Weaknesses of Bayesian Approaches:

Critics of these methods suggest that the end user's assumption that data created from iterative statistical sampling processes is the true data may result in misinterpretations. Critics are also concerned that the new estimate may distort the true relationships in more complex modeling efforts.

In addition to some criticism of data integrity, there is also an implementation challenge in terms of web-based systems. The computer processing time is significant when creating the new cell estimates. It is unlikely that a dynamic or "query-on-the-fly" system could utilize this method, given the processing time demands.

These more advanced methods require highly trained public health data managers, IT staff, and analysts. Alternatively, financing is needed for contracts to build the systems that incorporate these more complex methods. On the side of the data users, there are needs for training in interpretation of the query results and a need for sophisticated, but easy to understand documentation.

D. Summary of Formal Statistical Approaches to Improve Interpretation of Results from Small Cells

Formal statistical approaches (confidence intervals, hypothesis tests χ^2, and coefficients of variation) allow maximum information to be disseminated while honestly communicating statistical reliability to the user. The evaluation of reliability of any measurement procedure consists in determining how much of the variation in scores among individuals is due to inconsistencies in measurement.[11] If measurement is free from random or variable errors it is considered reliable. Some measures of reliability focus on measuring different sources of variation. We describe three formal statistical approaches to improving the quality of interpretation of web-based query system results.

1. Calculation of confidence intervals is a strategy to provide the end user with a more accurate interpretation of the results. The width of the confidence interval provides a good picture of the potential variability in the results. It is used frequently in public data reporting to indicate that results should not be compared given the range of randomness (e.g., rural vs urban results for a specific condition or event, such as mortality rates). It is a measure of whether or not the results seen are within the range one could expect given the cell size. The smaller the number of cases in the numerator and denominator the greater the width of the confidence interval. If the result falls outside of the confidence interval one cannot be confident that it is not the result of randomness.

[11] C. Seltiz, M. Yahoda, M. Deutsch and S.W. Cook. Research Methods in Social Relations. Holt, Rinehart and Winston, New York, NY, 1967.

2. χ^2 tests are used when the distribution of the data is not normal, and the data are not at least at a level of interval scaling. These are considered non-parametric tests because they fail to meet the basic parameters as described necessary for parametric tests of normal distributions. Certain types of chi square tests are designed specifically to adjust for small cell size.
3. Coefficient of variation (C.V.) is a measure of the stability of the estimate, compared with the magnitude of the estimate.

We further describe each of the above three statistical approaches that can be used with small cell sizes for increasing the understanding of the results. These methods are available to designers of web-based data dissemination tools. It should be remembered however, that many individuals will not understand how they work or what they mean.

1. Confidence Intervals

The variability, within a cell across time, in health data systems is increased when cell sizes are small. Wide fluctuations in the data can be seen from quarter-to-quarter, or year-to-year, when the cell size is small. For example, if examining hospital discharge data, it is likely that some hospitals will experience very small numbers of deaths for certain types of diagnoses or procedures, and the number of procedures may also vary across institutions. Analysis of changes from one year to the next could result in significant findings, yet these findings could be due solely to chance. Using a confidence interval (C.I.) assists the data user in determining the reliability of the information being provided. The wider the confidence interval and the smaller the sample, the less precision there is in the estimate. Narrow confidence intervals suggest that the estimate is nearly precise, especially with large sample sizes, and that chance plays a smaller role in the outcome of interest.

Confidence intervals avoid many of the problems inherent in simply reporting the statistical significance of a test statistic, and provide considerably more information. A significant statistic only gives us the information that the true population effect is probably not zero. Confidence intervals, focus on the magnitude of the effect, and provide an estimate on the precision with which the effect is estimated. For example, if the confidence interval includes both negative and positive values, it generally indicates a lack of precision.

When to use confidence intervals? Oakes (1986) suggests that confidence intervals be used in lieu of significance testing. Confidence intervals should be used whenever there is a need to understand the uncertainty in a point estimate. That need often arises due to small cell sizes. In hospital reporting, if there is significant variation across time in the hospital's mortality or length of stay, or average charges—a confidence interval will help the user to understand the contribution of uncertainty to that fluctuation. The analysis of fluctuations across time (for institutions whose population is smaller) using confidence intervals will allow for greater variability for the event of interest and provide a better look at the range of possible values. It can also help to reduce the misinterpretation of random variation when cells are small.

Institutions that have large numbers of cases for cells will have narrow confidence intervals, suggesting that there is greater precision and a smaller range of possible values, i.e., less margin for error in interpreting the outcomes.

In addition, confidence intervals can be fit around an odds ratio and be determined for different levels of error (alpha), an important contribution to epidemiological research. For example, reporting on the odds of death in a particular hospital for a particular diagnosis or procedure is of interest to consumers. It is important that users not compare mortality rates across institutions by examining overlapping confidence intervals.

To create a confidence interval requires only an approximate normal distribution and knowledge of the standard error of the sample.

The State of Washington Department of Health has provided excellent documentation of the methods by which they produce confidence intervals for their web-based data systems. They provide the methods used for producing confidence intervals for the following: age-adjusted rates, crude and age specific rates, standardized mortality rates (for cells with <100 and >100 cases), and for non-independence of events (such as those including multiple re-admissions), binomial proportions and for complex survey design. See the web address below to download a PDF version of their guidelines.

References:

Oakes, M. Statistical inference: A commentary for the social and behavioral sciences. New York: Wiley, 1986.

Chiang, CL. Standard Error of the Age-Adjusted Death Rate. Vital Statistics Special Reports, 47(9):275-285, 1961.

Chiang, CL. Introduction to Stochastic Processes in Biostatistics. New York: John Wiley & Sons, 1968.

Newton, R.R., and Rudestem, K.E. Your Statistical Consultant: Answers to Your Data Analysis Questions. Thousand Oaks, CA: SAGE Publications, 1999.

Agencies/Systems Using Approach:

State of Washington, Department of Health
Documentation available on website (generally designed for public health professional use) www.doh.wa.gov/data/guidelines/
Vista/PHw - Washington State Center for Health Statistics
Contact: David Solet, Washington Center for Health Statistics.
(David.Solet@doh.wa.gov)

Utah Department of Health, Indicator Based Information System (IBIS)
Contact: Lois Haggard, Ph.D., Utah Department of Health

loishaggard@utah.gov

State of Wisconsin
Bureau of Health Information and Policy
Wisconsin Interactive Statistics on Health (WISH)
www.dhfs.state.wi.us/wish
Contact: Richard Miller (millere1@dhfs.state.wi.us)

Missouri Department of Health and Senior Services
Missouri Information for Community Assessment (MICA), (C.I. for Rates)
Contact: Garland Land, Director
Center for Health Information Management and Evaluation
Missouri Department of Health and Senior Services
P.O. Box 570
Jefferson City , Mo. 65102
573-751-6272
landg@dhss.mo.us

Strengths and Weaknesses of Approach

In conclusion, use of the confidence interval may be necessary to report in a web-based data dissemination system when the results in some cells are significantly smaller than in other cells. For example, in a study of mortality related to a specific surgery, some hospitals will have very few surgeries to consider and deaths may be subject to wide variation across time periods. As results are shown some users might think that the institution is doing better or worse than what one would expect given the true underlying population—when the differences in fact may be due to random variation. The confidence interval allows the user to assess whether the rate of mortality for one hospital is a good estimator of the true rate of mortality in the population, allowing greater confidence in the results. For those with scientific backgrounds, confidence intervals (and coefficients of variation) are quite useful in assessing reliability of the outcomes under study.

While confidence intervals offer a good indicator of statistical power, they should generally not be used to draw comparisons across cells because you cannot necessarily interpret the certainty of the statistical significance (Newton and Rudestam, 1999).

The level of understanding of the user of the data must also be a consideration; it may be difficult for the lay public to understand the information provided by confidence intervals.

2. Use of χ^2 tests

Most statistical tests of significance are based upon normal distributions and measurements that are in the form of at least interval scales. These conditions are however, not present when there are very small cell sizes or very small populations.

There are statistical tests designed to address these conditions—they are called non-parametric or distribution-free statistics. The chi-square test is one of the more frequently used non-parametric tests; it is relatively easy to meet the assumptions for this test. But there are some nuances to be aware of in relation to the number of variables and cell sizes. For example, a simple contingency table which is called a 2 x 2 table would require use of the Continuity Correction rather than a simple chi square test. Also, if the smallest expected frequency[12] for any cell in a 2 x 2 table is less than 5 then one should use the Fisher Exact Test. In a larger table, there is a requirement that no more than 20% of the expected frequencies in the table can be less than 5. Other requirements include: 1) no cells may be less than 1, and 2) no respondent (individual) may be in more than one cell in the table (independence). If the test is invalid due to cell size, then simply aggregate the data to increase the size of the cell. If we affirm that a difference is present in the two samples with a chi square test, then we reject the null hypothesis that the two samples are the same. Thus, if we are examining rates of breast cancer for women in one county versus rates in another county, if the test is significant, we are rejecting the null hypothesis that the two groups (counties) have equal rates of breast cancer in women. Chi square tests are used because rates of breast cancer may be very low in one or both counties, or because one or both counties have a very small population base.

The chi square test allows us to examine differences where the distribution is not normal—a significant result suggests that the difference is not due to error. This means that we can reliably state that the differences between the two populations are not due to chance.

In public health, statistical tests are frequently used for determining significant trends over time (see State of Washington's Vista/PH system as an example), and for assessing significant differences in rates between groups. The chi square test is a useful tool to assess these differences, even under conditions of small populations or small cell sizes.

There are a variety of different chi square distributions—one that is frequently used in public health is the Mantel-Haenszel Chi Square Statistic. This is used for a stratified analysis of health risks when statistical control of a few variables is required. It can also be useful in exploring more complex relationships and can sometimes be used to effectively quantify risk when there are numerous variables to control. This test can only be used if both variables lie on an ordinal scale. For more detailed information see Kleinbaum, Kupper and Morganstern (1982). An example of how this method is used in public health practice can be found in the following article, "Firearm ownership and health care workers," Public Health Reports, May-June, 1996 by Bruce W. Goldberg, Evelyn Whitlock, and Merwyn Greenlick.

References:

Armitage, P. Statistical Methods in Medical Research, 2e. Boston: Blackwell Scientific Publications, 1987.

[12] The expected frequencies are calculated for each cell in the table by multiplying the appropriate row and column totals and dividing by N (Foster, 2001).

Brogan, D.J. Pitfalls of Using Standard Statistical Software Packages for Sample Survey Data. Rollins School of Public Health, Emory University, Atlanta, April 15, 1997

Kleinbaum, Kupper and Morganstern. *Epidemiologic Research* by (Lifetime Learning Publications), Wadsworth, Inc; Belmont California; 1982

Mantel, N. and Haenszel, W. (1959), "Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease," *Journal of the National Cancer Institute*, 22, 719 -748.

Agencies/Systems Using Approach:

CDC EPI INFO™ in the analysis section – allows for chi square statistics to be reported in the output. http://www.cdc.gov/epiinfo

State of Washington, VISTA/PH system
http://www.doh.wa.gov/OS/Vista/Statistical_calculations.htm

Pennsylvania Department of Health
Health Statistics - Technical Assistance
(717)783-2548

Strengths and Weaknesses of Approach:

While the chi square test is useful in determining whether a significant difference can be found in a contingency table, it still requires at least 5 cases in each cell in a 2 x 2 table, or no more than 20% of cells with < 5 in a larger contingency table. Also, because the chi square distribution is really a family of distributions based on degrees of freedom, it may require further research to assess which distribution (chi square test) is most appropriate for the data under consideration.

Programming to build chi square tests into a system is relatively easy given that it is available in most statistical programming packages, such as SAS and SPSS and other public health oriented statistical packages. However, if you are working with survey sample data you may need to use SUDAAN or another package to account for stratified sampling. For more information on this see the article by Donna Brogan listed in the references section.

3. Coefficient of Variation

While confidence intervals assist us in understanding the margin of error, it is sometimes not sufficient to assess quality of the estimate. The coefficient of variation (C.V.) is a measure of the stability of the estimate, compared with the magnitude of the estimate. The coefficient of variation provides a relative measure of data dispersion compared to the mean: $Cv = s/\chi$ for the normal (bell shaped) distribution. The coefficient of variation has no units. It may be reported as a simple decimal value or it may be reported as a percentage $100 \times Cv = s/\chi$. Thus, it is defined as the ratio of the standard deviation to its mean. A smaller C.V. suggests less variability due to magnitude.

For example, if you were looking at performance of two hospitals in terms of number of deaths over a ten year period, you could take the average number of deaths in Hospital A and the average in Hospital B, but because hospital A has a Type I trauma center (more deaths) and Hospital B is a general community hospital (serious cases are transferred out), the confidence interval is not sufficient to determine whether your estimate reflecting performance differences is accurate. Instead, the coefficient of variation provides a relative measure of the data dispersion compared to the mean in both hospitals. Because Hospital B has fewer deaths, they have more instability in an estimate in any given year in terms of the number of deaths (if you look at the data over the 10-year period). Thus, the C.V. provides additional information for an assessment of the reliability of the information.

The Ontario Ministry of Health has set specific respondent numbers for using the Coefficient of Variation in their Ontario Health Survey. The guidelines for the release of their data state that "if the number of sampled respondents is less than 30, the weighted estimate should not be released regardless of the value of the coefficient of variation for this estimate. For weighted estimates based on sample sizes of 30 or more, the coefficient of variation will conclude whether the estimate is unqualified, qualified, confidential or not releasable. Generally, larger sample sizes provide more reliable estimates of health risks and related health behavior." This suggests that the underlying population must at least contain 30 individuals before using this stability measure. Cell sizes can be less as long as there are at least 30 individuals (cases) in the population.

In conclusion, use of the confidence interval and the coefficient of variation may be necessary to report in a web-based data dissemination system when the system provides data across communities and facilities, and where small cell sizes exist. In addition to a good understanding of how the data will be used, the level of understanding of the user of the data might also be a consideration; it may be difficult for consumers to understand the information provided by confidence intervals.

References:

Lois Haggard. 'Reporting Small Numbers: How Small is Too Small?' June 15, 2000. (Utah Department of Health, Salt Lake City, UT 84114-2101, Phone: 801-538-9191 or contact: loishaggard@utah.gov)

Rae Newton and Kjell Rudestam, Your Statistical Consultant: Answers to Your Data Analysis Questions, SAGE Publications, 1999.

Agencies/Systems Using Approach:

State of Washington, Department of Health
Contact: David Solet (David.Solet@doh.wa.gov)

Ministry of Health, Ontario Health Survey

http://www.cehip.org/DataInfo/

Strengths and Weaknesses of Approach

The use of the coefficient of variation is rarely seen in web-based data dissemination systems, it is less commonly known and used than the confidence interval. It is also more difficult to explain to lay persons using the data.

That being said, it is a useful addition to the confidence interval, and should be considered for inclusion in the web-based system for use by professionals.

E. Software Tools

State and private organizations are developing open source or proprietary software products that apply multiple approaches for improving statistical reliability and reducing disclosure risk of public health data. In addition to the approaches listed above, there are now several new software packages that provide technical support for protecting public health data. The National Center for Health Statistics (NCHS) sponsored the development of disclosure limitation software for two-way tables by OptTek Systems, Inc.. The OptTek software includes the following functionality:

- cell suppression
- controlled rounding (minimum-distance controlled rounding)
- unbiased controlled rounding
- controlled rounding subject to subtotal constraints
- synthetic substitution (controlled tabular adjustment)

For more information on the NCHS tool contact Larry Cox at NCHS. The second tool was created by RTI International and it is called MASSC^{SM} and it focused on reducing disclosure risk for surveys where sampling methods have been used. For additional information on this tool, contact Dr. Michael Samuhel at samuhel@rti.org .

F. Recommendations

This review of the statistical approaches for both protecting data from disclosure of sensitive information, and increasing the reliability of the data, should be used in tandem with the other two sets of guidelines. It would generally be appropriate to use both approaches (technical/statistical) for reducing disclosure risk. It would also be useful to improve the reliability of the data offered in web-based data dissemination systems.

We suggest that public health agencies review the current methods that are in use in their web-based data dissemination systems and determine whether addition of other approaches would provide that extra protection and result in more reliable information for the user. Yet, we also support the notion of keeping it as simple as possible—while providing the necessary protection.

We also encourage statisticians to expand their efforts on new or enhanced statistical methods to assure that individuals will not be identified via public health web-based data dissemination systems.

In summary, we suggest that a substantial investment will be necessary if public health agencies are going to take advantage of the more advanced statistical methods. Investments could be targeted at upgrades to systems currently in place, additional research on new statistical approaches, or for training state data system developers and data users.

Guideline Use

We hope users of this document will notify the National Association of Health Data Organizations with additions or corrections. Please send an email to: Barbara Rudolph, Ph.D., Senior Scientist for Research and Data. brudolph@nahdo.org.