Guidelines to Follow When Working with Small Numbers*

Vijaya Krishnan, Ph.D.

ECMap
Early Child Development
Mapping Project Alberta

19 April 2011

Why are Small Numbers a Concern in EDI Analysis?

1. Small numbers raise ***statistical issues***, and thereby the accuracy and usefulness of the data.
2. Problems with ***confidentiality*** arise when there are small denominators (N) in rates and percentages.
3. Problems with *reliability* arise when there are small numerators (n). Why the question of reliability? Rates and percentages based on (almost) full population counts are subject to random variation. The random variation may be substantial when rates and percentages are calculated using the small 'n' in the numerator. Rates based on small n's may fluctuate over time and geographic area.

1. How to Address the Statistical Issues?

- Populations with small tails or little skewness won't require a sample that is large, and populations with wide tails or highly skewed ones will require a much larger sample size.
- Although the technical statement of the *Central Limit Theorem* does not have a statement about how large 'n' should be, the convention is to consider the small sample size situation as less than 30 and large sample size situation as $n \geq 30$. A sample size greater than 30 approximates the normal distribution, that's probably the magic behind this number.
- Estimates based on fewer than 30 cases in the denominator or a relative standard error greater than 30% are considered statistically unreliable and are suppressed in analyses based on National Health and Nutrition Examination Survey (NHANES) in the US for Healthy People 2010 objectives.[1] It is very important to show the confidence interval (CI)[2], especially when the number is less than 20. Rates with fewer than 20 cases in the numerator have very wide CI. For example, if physical health has a domain score 7, and has a CI between 4 and 9, it cannot be precise and it needs to be acknowledged in writings. Suppress both numbers and CI, if fewer than 5.
- The specific criteria for data suppression used by 22 of the 23 major data systems to track Healthy People 2010 objectives noted that if the numerator is less than 20 (20 is the threshold for reliability used by public health assessments, CDC, US), an effort to increase the size of the numerator should be made by:
    a. Combining multiple years of data
    b. Collapsing data categories
    c. Expanding geographic boundaries
- Always report the numerator for the rate, percentage, and so on, and also Include a footnote stating that the rate or percentage is based on 'X' cases, and cannot be reliable and precise.

[1] See Klein, Proctor, Manon, & Turczyn (2002). Healthy people 2010 criteria for data suppression. *Healthy People 2010, 24*, Statistical Notes (CDC), July

[2] The confidence interval (also called margin of error) is the plus-or-minus figure usually reported in newspaper or opinion poll results. For example, if we use a confidence interval of 10 and 80% of our children are developing appropriately, we can be sure that if we had the same information on the entire cohort of kindergarten children, the percentages will be between 70% and 90%; the wider the confidence interval we are willing to accept, the more certain we can be that the whole cohort of kindergarten children would fall within that range.

Most populations with 30 or more cases will have an approximate normal distribution for its sample mean, as determined by empirical evidence. This warrants the use of 30 as the threshold.

2. How to Reduce the Risk of Confidentiality Breach?

- A general approach is to aggregate data values before analysis (e.g., combine sub-communities).
- Apply *cell suppression*, *noise infusion*, or both. Totals and values for cells that are not suppressed remain unchanged.[3] Noise infusion is an alternative to cell suppression that allows for the publishing of more data. By marginally adjusting each sub-community data (e.g., replacing with estimates), data for individual sub-communities can be camouflaged. Most aggregate measures are distorted by a small amount. These measures should be treated as a last resort, however.
- When the above steps do not produce satisfactory results: (a) combine multiple years, if available; and (b) omit certain fields from analysis (e, g., omit sub-community analysis if they don't meet the threshold).

What is a Good Participation Rate or Response Rate?

- There is no national or international standard for response rate. The response rate is affected by a host of factors. For example, thirty five questionnaires for a sub-community may be expected to provide accuracy if: (1) we have a very small population to begin with; (2) we have very little variance in the responses, or (3) we are willing to accept very low accuracy.
- As a *very rough rule of thumb*, 95 children will provide fairly good accuracy (with a 95% certainty), if we are willing to accept a confidence interval ±10% (the mean of the repeated samples will likely fall within the -10% and +10% interval) from a population of children, say 8000, based on statistical computations.[4] Some calculations take into account the harmonic mean because it takes into account situations where fewer cases have high values on a variable, while most cases have low values on the same variable:

[3] Data withheld are replaced with 'D's in appropriate cells in tables. Ranges are sometimes used in place of 'D's to suppress sensitive data.
The Healthy People 2010 baseline data use the following abbreviations:
DNC-Data are not collected by the data system to monitor the objective.
DNA-Data have been collected but have not yet been analyzed.
DSU-Do not meet the criteria for statistical reliability, data quality, or confidentiality (data are suppressed). Major reasons outlined for suppression are: (a) Number of cases too small to produce reliable estimates or may violate confidentiality requirements; (b) Sample design does not produce representative estimates for a particular group; and (c) High item non-response or large number of unknown entries (Klein et al., 2002, pp.1-2).

[4] Sample size calculations can be done using different methods. All methods take into account the amount of certainty we like to have (95%, 99%, etc.).

Adequacy of the number of cases for a sub-community is important to detect differences within and across (sub)communities. The reporting of the CI is important, but it is also important to estimate the error utilizing the harmonic mean of unequal n's.

The formula for means ***within*** (sub)community is:

$$Harmonic\ mean \pm (1.96) S.E. \left(1 - \frac{n}{N}\right)$$

Where S.E. is the standard error within sub-community, n is the sample size and N, the population

If we have, say, 10 sub-communities, the formula for means across sub-communities:

$$Harmonic\ mean \pm (1.96) \frac{2\ MS\ within\ groups}{\sum_1^{10} \frac{10}{1/n}}$$

Where 10 is the number of sub-communities.

What is a Reasonable Threshold for Participation or Coverage Levels?

- The participation rates in Censuses and response rates in surveys vary across nations and regions within nations. The participation rate in the 2000 census in the US was 70% and in the 2010 Census, it was 71%.[5] The rates varied from 65% to 73% in 2000 and 66% to 75% in 2010, across states. "I would consider 71% a very serious accomplishment," said Kenneth Prewitt, Census Bureau director in 2000, now professor at Columbia University. "If it does hit 72%, it's even better."
- In his presidential address to the Public Opinion Quarterly, Bradburn (1992) suggested that the response rates of 70% and 77% can be regarded as rather "good." [6]

> A percentage greater than or equal to 74 is a reasonable response (participation) rate that may serve as a meaningful threshold level in community analyses, based on evidence from national and international censuses and surveys.

[5] http://www.usatoday.com/news/nation/census/2010-04-20-census-participation-rate (Retrieved on April, 8, 2011, pp.1-4).

[6] See Bradburn, N. (1992). Presidential address: a response to the nonresponse problem. *Public Opinion Quarterly, 56*, 391-397.