GUIDELINES FOR STATISTICAL ANALYSIS OF PUBLIC HEALTH DATA WITH ATTENTION TO SMALL NUMBERS

A COLLABORATIVE EFFORT OF

THE FAMILY HEALTH OUTCOMES PROJECT TECHNICAL ADVISORY GROUP

Revised

July, 2003

Roy R. McCandless, MA, MPA, DrPH
Geraldine Oliva, MD, MPH

FAMILY HEALTH OUTCOMES PROJECT
Department of Family and Community Medicine
University of California, San Francisco

Funding provided by the Maternal and Child Health Branch of the California Department of Health Services

Acknowledgements

This documents was developed by the Family Health Outcomes Project at the University of California, San Francisco, with review and guidance from the Technical Advisory Group. The following individuals contributed:

Vicki Alexander, City of Berkeley Health Department

Candice L. Diamond, California Office of Statewide Health Planning and Development

Gwendolyn Doebbert, Center for Health Statistics, California Department of Health Services.

Mary Graebner, Nevada County Department of Health

Mark Gustafson, Primary Care and Family Health Division, California Department of Health Services

Barbara Hartz, Maternal and Child Health Branch, California Department of Health Services

Lisa Lindvall, Family Health Outcomes Project, University of California, San Francisco

C. Jane McKendry, Center for Health Statistics, California Department of Health Services.

Randy Reiter, San Francisco Department of Public Health

Mike Rodrian, Center for Health Statistics, California Department of Health Services.

Cynthia Schmidt, Center for Health Statistics, California Department of Health Services

Eugene Takahashi, Office of Epidemiology, Maternal and Child Health Branch, California Department of Health Services.

Roger Trent, Injury Surveillance and Epidemiology Section, California Department of Health Services

Evelyn G. Treviño, County of San Bernardino Department of Public Health

Please direct comments and suggestions to:
The Family Health Outcomes Project
3333 California Street, Suite 365
San Francisco, CA 94118
Tel 415-476-5283
fhop@itsa.ucsf.edu

TABLE OF CONTENTS

I. INTRODUCTION 1

II. BACKGROUND 3

A. The Basic Measures 3
B. Comparing Measures 5
C. Types of Studies 6
D. Validity and Reliability 7
E. Confidence Intervals 7

III. CONFIDENCE INTERVALS FOR NUMERATOR DATA 10

A. When There are 100 or More Cases or Events 10
B. When There are 20-99 Cases or Events 10
C. When There are Less Than 20 Cases or Events 10

IV. CONFIDENCE INTERVALS FOR PROPORTIONS 12

A. When the Numerator is at Least 20, and N Minus the Numerator is at Least 5..... 12
B. When the Numerator is Less Than 20, or N Minus the Numerator is Less Than 5.. 12

V. CONFIDENCE INTERVALS FOR RATES 13

A. With 100 or More Events in the Numerator 13
B. With 20-99 Events in the Numerator 13
C. With Less Than 20 Events in the Numerator 13

VI. ANALYSIS OF DIFFERENCES IN PROPORTIONS 14

A. When Comparing with a Benchmark Derived from a Very Large Population 14
B. When Comparing Two Proportions and Both Proportions are Subject to Variability . 14

VII. ANALYSIS OF DIFFERENCES IN RATES 16

A. When Comparing with a Benchmark Derived from a Very Large Population 16
B. When Comparing Two Rates, Both Rates are Subject to Variability, and Both Rates are Based on 100 or More Events in the Numerator 16
C. When Comparing Two Rates, Both Rates are Subject to Variability, and One Rate Has 20-99 Events in the Numerator 17
D. When Comparing Two Rates, Both Rates are Subject to Variability, and One Rate Has Less Then 20 Events in the Numerator 17

VIII. ANALYSIS OF RATIOS OF RATES OR PROPORTIONS 18

A. When the Benchmark Rate or Proportion is Derived from a Large Population and is Considered Fixed 18
B. When Rates or Proportions for Both Groups are Viewed as Variable 18

IX. WHEN NUMBERS ARE TOO SMALL FOR STATISTICS……….. 19

A. Aggregating Data Across Years …… 19
B. Geographic Aggregation …… 19
C. Case Studies …… 20

X. REPORTING YOUR FINDINGS …….. 21

FIGURE 1. CONFIDENCE INTERVAL WIDTH AND NUMBER OF CASES ……. 8

TABLE 1. FACTORS FOR CALCULATING 95% CONFIDENCE LIMITS WHEN THE NUMBER OF CASES OR EVENTS IS LESS THAN 100 …….. 11

GUIDELINES FOR STATISTICAL ANALYSIS OF PUBLIC HEALTH DATA WITH ATTENTION TO SMALL NUMBERS

I. INTRODUCTION

This guide was developed by the University of California, San Francisco, Family Health Outcomes Project as a collaborative effort with members from the Project's Technical Advisory Group. The guidelines are intended to serve as an informal standard of practice for data analysts and program planners. No background in epidemiology or statistics is required, although some basic algebra is helpful.

The analytic strategies offered in these guidelines were identified and developed to assist local health jurisdictions and health programs in applying clear and consistent approaches to the analysis and presentation of data on health status and outcomes, with particular attention to situations involving small numbers of cases or events. These simplified analytic techniques are intended for use by program and data managers who may not have training in statistics and who may have limited access to epidemiologists and biostatisticians. Where such expertise is available, more sophisticated and varied analytic techniques can be considered.

Many consumers of public health data have a limited understanding of statistics and scientific methods. When these consumers are policy makers, they are often pressed to make decisions with insufficient information. In such situations, the temptation to rely on "weak data" rather than "no data" is strong. One of the most common ways of pushing data beyond their limits is to draw conclusions from rates based on a small number of events. Relying on small numbers for analysis may be appropriate as long as the accompanying risks are recognized and alternative approaches are appreciated. Tests of statistical significance, such as those described in these guidelines, may be applied to ensure reliability of conclusions based on statistical data.

This updated monograph retains emphasis on the problems of analyzing data in situations with small numbers. However, such analysis takes place within the context of the broader problems of statistical analysis. Thus, we have broadened the scope of the monograph to incorporate some basic orientation to statistical methods in general. Also, we have modified the monograph to conform to practices employed by the U.S. National Center for Health Statistics for analysis and reporting of public health statistics.[1]

Public heath data, typically derived from vital records, usually represent complete counts of events and are not derived from samples. However, there remains a need to employ techniques for statistical analysis such as calculation of confidence intervals. This is because the number of events that actually occur is viewed by statisticians as one outcome among a range of possible outcomes. The number of births or deaths may vary due to chance (random variation).

[1] For greater detail, the reader is referred to the Technical Notes section of Ventura SJ, Martin JA, Curtin SC, Mathews TJ, Park MM. Births: Final data for 1998. National Vital Statistics Reports; vol 48 no. 3. Hyattsville, Md: National Center for Health Statistics. 2000.

With public health measures, we often encounter situations where relatively few events occur in a given year. The smaller the number of events and the smaller the probability of such an event, the greater the importance of random fluctuations within the specified time period. The more rare an event, the less stable the rate calculated on the basis of that event. Therefore, counties with only a few deaths or a few teen births can have unstable rates from year-to-year, and observed rates for such events may substantially differ from the true underlying rate.

Some alternative statistical techniques are available for such situations. Alternative formulas are available for small numbers. Also, through techniques such as averaging rates over a number of years or aggregating (adding) events for multiple years, these random variations can be smoothed (evened out) and more stable numbers can be used to estimate the rates. These guidelines present a variety of approaches to assessing the adequacy of an observed rate as an estimate of its true value and strategies to manage the problems stemming from year-to-year fluctuation in the occurrence of small numbers of events.

A number of basic statistical and epidemiological terms and concepts are used in these guidelines. The background provides some basic definitions and is intended for those with limited statistical background or those who just want a quick refresher.

II. BACKGROUND

A. The Basic Measures

Incidence – refers to events occurring over a period of time. In public health, we talk about disease incidence to describe the number of people getting a condition (newly diagnosed or reported cases), typically over a one year period.

Prevalence – refers to health status – the number of people having a condition. In public health, we talk about disease prevalence to describe the number of people who have a certain condition. We usually mean **point prevalence** to describe the number having the disease at a point in time, or we say **period prevalence** to describe the number having the disease at any point over a period of time, usually one year.

Prevalence equals incidence times duration. If 10 people per day in your neighborhood get a 3-day cold, then about 30 people will have the disease at any time.

Ratio – the result of dividing one quantity by another when the numerator and denominator are separate and distinct quantities, that is, neither is included in the other. Examples: ratio of males to females; ratio of fetal deaths to live births.

Proportion – a type of ratio in which the numerator is included in the denominator. Examples: males as proportion of total population; newborns having low birth weights as a proportion of all newborns; women who obtained prenatal care in the first trimester of pregnancy as a proportion of all women who gave birth. A proportion will have a range between zero and one.

Percent – one way of characterizing a proportion, obtained by multiplying 100 times the proportion. Percent means parts per 100. If the low birth weight proportion is .35, then we can say that low birth weight was 35% of births.

The reader will note that we often see errors in grammar when people write about percents – sometimes people incorrectly write “percent of low birth weight”. This is incorrect because it could be taken to refer to a proportion (subset) of low birth weight infants. The correct statement about low birth weight is to say “percent low birth weight” or “low birth weight percent”.

While a ratio or proportion can be calculated for a point in time (male/female ratio, for example), for public health measures, we typically look at a period of time, usually one year (low birth weight as a percent of births over one year). This practice of measuring over a single year can create confusion about the nature of our measures. This will be explained in more detail later.

Rate – a measure of change in one quantity per unit of another quantity, typically measured over time.

Consider the example of velocity. While driving on the highway, we glance at the speedometer that tells us we are traveling at a rate of 55 miles per hour. This is an instantaneous measure that describes the rate of speed we are traveling at a point in time. Over a long trip, we might calculate the total distance traveled and our driving time. This gives us an average rate of speed. So, we might travel 200 miles in four hours and conclude that our average rate was 50

miles/hour. However, in fact, some of the time our rate of speed was 60 miles per hour, and some of the time it was 40 miles per hour.

In public health, rates are typically used to describe mortality. Epidemiologists speak of mortality rates as a measure of the “force” of mortality. From a technical point of view, we would like to measure that force just as our speedometer measures velocity – we would like to measure the force of mortality at a point in time. In practice, however, we calculate an average rate of mortality over a period of time – typically one year.

As a practical manner, when we calculate mortality rates among human populations, we use the mid-year population for our denominator. This approach works fine for most populations, because it assumes that the deaths were evenly spaced during the year, and it averages out for migration to and from the study area. Thus, the mid-year population is used as a proxy for the number of person-years at risk for mortality over the one year period.

Here is an example of a rate calculation: In a community with 10,000 women ages 15-44 at mid-year, there were 500 births during the year. The **fertility rate** was 500/10000 = 0.05. For convenience, we multiple by 1,000 (or 10,000 or 100,000 if you wish) to say that the rate was 50 births per 1,000 women of fertile age who were at risk for the year.

Population at Risk – Public health measures work best when they focus on the population which is truly susceptible to the condition or event in question. For example, fertility rates focus on women of fertile age. Sometimes we are unable to make the appropriate adjustment. For example, the cervical cancer mortality rate among women in a community might be very low. But we would think differently if we discovered that there had been a substantial number of hysterectomies in the community.

Infant Mortality Rate – For practical purposes, we usually calculate an infant mortality rate by looking at a given year and dividing the number of babies who died before their first birthday by the number of births during the year. This approach is not technically correct, but it provides a reasonable approximation. It is not technically correct because some of those who died may actually have been born in the prior year; because babies born in the current year were not at risk for an entire year; and because we did not deduct the months during which those who died were no longer at risk of death.

The calculation provides a reasonable approximation because the number of infant deaths is fairly small, and because the number of births does not greatly change from year to year. The calculation would not work during a severe fatal epidemic. Nor would it work if there were some massive increase (or decrease) in the number of births or infant deaths from one year to the next.

B. Comparing Measures

While it is useful to know the proportions or rates for various public health measures of our communities, such measures often have little meaning until we make comparisons within the community or with other communities.

Relative Measures – The rate for one group divided by the rate for another group is a **rate ratio**.

> If the infant mortality rate for babies of women who smoked during pregnancy was 6/1000 and the rate for women who did not smoke was 4/1000, then the rate ratio was 6/1000 divided by 4/1000 = 6/4 = 1.5.

Ratios also can be applied to proportions. For example, in a community with 1000 births where 62 of the newborns had low birth weights, we would say that the low birth weight percent was 6.2%. If we knew the number of births to smokers and the number to non-smokers, and if we knew the number of low birth weight babies born to each group, then we could calculate a ratio of proportions.

> If 14/158 babies born to smokers had low birth weights, and 48/842 babies born to non-smokers had low birth weights, then the relative risk was: 14/158 divided by 48/842 = 8.86/5.70 = 1.55. This means that a baby born to a woman who smoked was about one and a half times more likely to have a low birth weight as a baby born to a woman who did not smoke.

Differences – subtracts the rate (or proportion) of one group from the rate (or proportion) of the other group. In the example of infant mortality rates given above, the **rate difference** is: 6/1000 minus 4/1000 = 2/1000. Note that these two approaches to comparing groups do not differ in any complicated way. All that has changed is a bit of algebra.

Attributable Risk Among the Exposed – the rate (or proportion) among the exposed minus the rate (or proportion) among the non-exposed.

> Using the above data for smoking and birth weight, we would calculate as follows:
>
> For babies of smokers, the low birth weight proportion is 14/158 = .0886.
> For babies of non-smokers, the low birth weight proportion is 48/842 = .0570.
>
> Thus, the attributable risk for babies of smokers = 0.0886 - 0.0570 = 0.0316. This means that, among babies of smokers, low birth weight was attributable to smoking for 3.16 per 100 births. With 158 births to smokers, we can attribute 0.0316 * 158 = 5 low birth weights to smoking.
>
> Note that 3.16% of babies born to smokers had low birth weights that were attributable to smoking. The rest of the 5.70% low birth weight babies among smokers would have had low birth weights anyway – as determined by comparison with the baseline population of babies born to non-smokers.

Population Attributable Risk (or Etiologic Fraction) – cases due to the exposure as a proportion of the total number of cases in the entire population. Given that we have already attributed 5 low birth weights to smoking, we can simply divide by the total number of low birth weight babies (62) to get a proportion of 0.08. The **population attributable risk percent** means that about 8% of the low birth weights among all births could be prevented by eliminating smoking.

Note that calculations for population attributable risk consider the prevalence of the exposure as well as the effects of the exposure. Thus, if an exposure is rare, then the population attributable risk will be small, even if the effect of the exposure is substantial.

An alternative approach is to calculate the proportion of newborns who were exposed to smoking (p = 158/1000 = 0.158) and the relative risk (RR = 1.55) and then apply the formula:

$$
\begin{aligned}
& [p * (RR - 1)] / [1 + \{p * (RR - 1)\}] \\
= \; & [0.158 * (1.55 - 1)] / [1 + \{0.158 * (1.55 - 1)\}] \\
= \; & [0.158 * 0.55] / [1 + \{0.158 * 0.55\}] \\
= \; & 0.0869 / [1 + 0.0869] \\
= \; & 0.0869 / 1.0869 \\
= \; & 0.08 \text{ or } 8\%
\end{aligned}
$$

Keep in mind that, in algebra, one solves first the part within **( )**, then within **{ }**, then within **[ ]**. Also, remember that ***** means multiply.

C. Types of Studies

There are three basic types of studies and they are each defined by the relationship of the researcher to the timing of outcomes and the predictive factors.

Prospective (Cohort) Studies – In this type of study, the researcher starts with two or more groups of subjects (exposed and not exposed) and follows them over time to see how many become cases. This is the most desirable type of study, particularly if the researcher is in the position to expose one group (and not the other) to the factor of interest. In studies of health, such experimental manipulation is usually impossible for ethical reasons.

Cross-Sectional Studies – Most of our public health data are cross-sectional. We survey both exposures and outcomes at the same time. Thus, we know that one sub-population has rates which differ from those of another sub-population for a given year, but we are unable to say why. For example, the infant mortality rate in one county may differ from that in another, and we can probably find differences between the counties in terms of race/ethnicity, income, educational levels, etc. But we can not discern which is the key factor, or even if those are the most important factors.

Retrospective (Case Control) Studies – samples cases and non-cases and measures past exposures for both groups. In this type of study, the researcher has a population with the disease and a population without the disease, but does not know the cause. The critical distinctions in this type of study, in comparison to the prospective study, are that no

experimental control is possible and that population denominators are unknown. Statistical methods for this kind of study differ from those in a prospective study.

The reader will note that, in certain rare situations, a researcher can look back in time and reconstruct a prospective study. For example, if one city has an industrial accident which exposes the population to some chemical, and the population at the time of the exposure is similar to the population of another city, then the researcher may be able to reconstruct all of the requirements of an experimental prospective study.

D. Validity and Reliability

Statistical data have two types of error and both must be addressed.

Validity – deals with systematic error. The notion of validity is closely connected to the problem of bias. For example, we might want to take a sample of our community, but, for some reason, we take a sample which is not representative (**selection bias**). Or, we might incorrectly categorize groups within the population (**misclassification bias**). For example, given that 15% of newborns in California have parents from different race/ethnic groups, we should recognize that categorization of newborns by race/ethnicity entails some problems with classification. There can also be measurement bias, recall bias and rounding bias, among others.

The problem of bias refers to issues of **internal validity**. **External validity**, on the other hand, refers to whether findings can be applied to a broader population. Researchers also are concerned with **face validity** and **construct validity** which deal with conceptual problems concerning the relationship of the measure to the phenomenon of interest.

Reliability – deals with random error and the question of whether replication of the study might produce different results simply on the basis of random chance. We can usually measure reliability and our estimates can be made more precise by taking larger samples. Reliability is described by **confidence intervals**. For example, we might say that our **point estimate** for the infant mortality rate is 5/1000, and that we are 95% certain that the true rate is between 4/1000 and 6/1000.

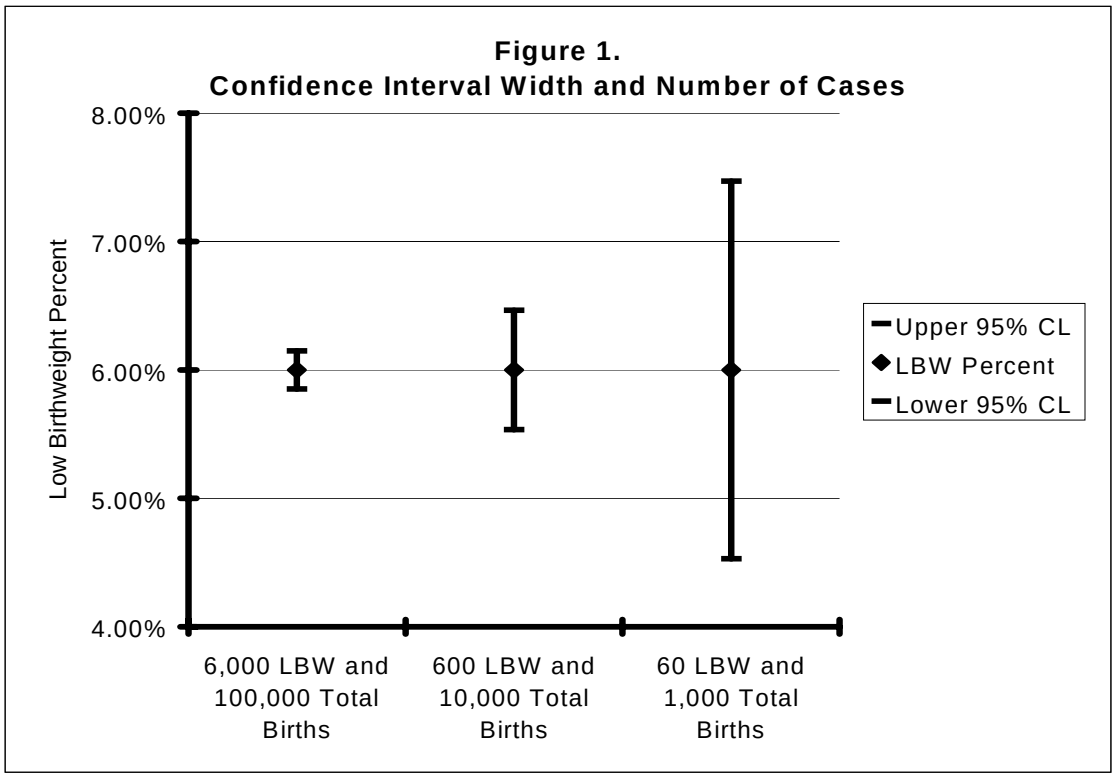
E. Confidence Intervals

At the heart of any comparison is the problem of testing whether a difference is meaningful or not. Significance testing lets us apply a standard to the question of whether differences between groups are likely to be due to chance. The first step is to determine the **level of confidence** we desire. In public health, we are generally satisfied by stating at the outset that we want to be 95% certain of our findings – meaning that, if we encounter a significant difference between two populations, then we want to be 95% certain of that finding.

One might well ask whether public health data need statistical tests, because the infant mortality rate (or any other popular measure) normally reflects actual counting of 100% of the events in question. The answer is, yes, statistical testing is needed because those events were subject to chance variation – the outcome for any given case could have been different.

We begin with a **null hypothesis** by stating that the populations do not differ. Then, we do the test and, if findings are positive, we reject the null hypothesis. For example, we might calculate a **point estimate** for the infant mortality rate in the county of 7/1000 live births, and we want to know if that figure is significantly different from a national rate of 6/1000. In this situation, the national rate is viewed as a fixed benchmark, and we use statistical analysis to establish a **95% confidence interval** around the point estimate of 7/1000 and look to see if the national rate is within or outside that interval.

As the number of cases or events (numerator) and the study population (denominator) increase, the relative width of the confidence interval becomes narrower. This means that we can be more certain that the observed rate is close to the true rate. For example, in Figure 1, we see findings from three different studies, each with a point estimate of 6% for the low birth weight percentage. However, the confidence interval for the study of 100,000 births is much narrower than is the interval for the study of only 1,000 births.



There is a convention for presenting information on the limits of confidence intervals. Consider the example of a calculated infant mortality rate of 5.3/1,000 with a 95% confidence interval that ranges from 4.5/1,000 to 6.1/1,000. The most common way this is expressed in published reports is: Infant Mortality Rate of 5.3 infant deaths per 1,000 live births, and this statement is followed by: **(95% CI = 4.5, 6.1)**.

There is a problem when we test various hypotheses at the same time. For example, if we have decided beforehand that we want 95% certainty, and we test 100 different phenomena each

with a positive finding, then we can assume that about 5 of those findings are incorrect. This problem can be addressed by adopting stricter criteria at the onset of the study.

Statistical calculations differ depending on whether we are comparing proportions or rates; and they differ for large and small populations. Also, in situations where the comparison population is very large, we can sometimes view proportions and rates for that population as statistically fixed, and not subject to statistical variation.

III. CONFIDENCE INTERVALS FOR NUMERATOR DATA

This procedure is used when you want to know the confidence interval around a number of cases or events.

A. When There are 100 or More Cases or Events: 95% CI = y +/- 1.96 * SQRT (y)

Please keep in mind that the lower case "**y**" refers to the numerator, and **SQRT** means "take the square root of". The notation "**+/-**" means that one should add the number to the point estimate to get the **upper limit** of the confidence interval, and subtract the number from the point estimate to get the **lower limit** of the confidence interval.

> With 256 births, the confidence interval would be calculated as:
>
> 95% CI = 256 +/- 1.96 * SQRT (256)
> = 256 +/- 1.96 * 16
> = 256 +/- 31.4
>
> Describe findings as 256 births (95% CI = 225, 287).

B. When There are 20-99 Cases or Events: 95% CI = y * (upper and lower factors in Table 1).

> With 60 births, refer to Table 1 on the following page and calculate the lower limit as: 60 * 0.76311 = 46.
>
> Calculate the upper limit as: 60 * 1.2872 = 77.
>
> Then report a finding of 60 births (95% CI = 46, 77).

C. When There are Less Than 20 Cases or Events:

We suggest that three successive years of data be summed and, if the combined number of cases or events is at least 10, then statistical analysis can consider the total. Findings can be reported for the middle year by reporting the average number of cases or events along with the lower and upper limits, each divided by three.

> With 10 deaths in 1997, 8 in 1998 and 15 in 1999, the sum is 33. Look up 33 in Table 1 to find the corresponding lower and upper factors. Multiply each by 33 to find a confidence interval ranging from 22.7 to 46.3. Divide each figure by 3, and report a 3-year average number of births for 1997-99 of 11 (95% CI = 7.6, 15.4). Be sure to include a statement in your narrative or table that the number of deaths and the statistical analysis cover three years of information.

Table 1. Factors for Calculating 95% Confidence Limits When the Number of Cases or Events is Less Than 100

N	Lower	Upper	N	Lower	Upper	N	Lower	Upper
1	0.02532	5.57164	34	0.69253	1.39740	67	0.77499	1.26996
2	0.12110	3.61234	35	0.69654	1.39076	68	0.77654	1.26774
3	0.20622	2.92242	36	0.70039	1.38442	69	0.77806	1.26556
4	0.27247	2.56040	37	0.70409	1.37837	70	0.77955	1.26344
5	0.32470	2.33367	38	0.70766	1.37258	71	0.78101	1.26136
6	0.36698	2.17658	39	0.71110	1.36703	72	0.78244	1.25933
7	0.40205	2.06038	40	0.71441	1.36172	73	0.78384	1.25735
8	0.42173	1.97040	41	0.71762	1.35661	74	0.78522	1.25541
9	0.45726	1.89831	42	0.72071	1.35171	75	0.78656	1.25351
10	0.47954	1.83904	43	0.72370	1.34699	76	0.78789	1.25165
11	0.49920	1.78928	44	0.72660	1.34245	77	0.78918	1.24983
12	0.51671	1.74680	45	0.72941	1.33808	78	0.79046	1.24805
13	0.53246	1.71003	46	0.73213	1.33386	79	0.79171	1.24630
14	0.54671	1.67783	47	0.73476	1.32979	80	0.79294	1.24459
15	0.55969	1.64935	48	0.73732	1.32585	81	0.79414	1.24291
16	0.57159	1.62394	49	0.73981	1.32205	82	0.79533	1.24126
17	0.58254	1.60110	50	0.74222	1.31838	83	0.79649	1.23965
18	0.59266	1.58043	51	0.74457	1.31482	84	0.79764	1.23807
19	0.60207	1.56162	52	0.74685	1.31137	85	0.79876	1.23652
20	0.61083	1.54442	53	0.74907	1.30802	86	0.79987	1.23499
21	0.61902	1.52861	54	0.75123	1.30478	87	0.80096	1.23350
22	0.62669	1.51401	55	0.75334	1.30164	88	0.80203	1.23203
23	0.63391	1.50049	56	0.75539	1.29858	89	0.80308	1.23059
24	0.64072	1.48792	57	0.75739	1.29562	90	0.80412	1.22917
25	0.64715	1.47620	58	0.75934	1.29273	91	0.80514	1.22778
26	0.65323	1.46523	59	0.76125	1.28993	92	0.80614	1.22641
27	0.65901	1.45495	60	0.76311	1.28720	93	0.80713	1.22507
28	0.66449	1.44528	61	0.76492	1.28454	94	0.80810	1.22375
29	0.66972	1.43617	62	0.76669	1.28195	95	0.80906	1.22245
30	0.67470	1.42756	63	0.76843	1.27943	96	0.81000	1.22117
31	0.67945	1.41942	64	0.77012	1.27698	97	0.81093	1.21992
32	0.68400	1.41170	65	0.77178	1.27458	98	0.81185	1.21868
33	0.68835	1.40437	66	0.77340	1.27225	99	0.81275	1.21746

Source: Ventura SJ, Martin JA, Curtin SC, Mathews TJ, Park MM. Births: Final data for 1998. National Vital Statistics Reports; vol 48(3). Hyattsville, Md: National Center for Health Statistics. 2000, page 95.

IV. CONFIDENCE INTERVALS FOR PROPORTIONS

The notation for a proportion is **p**. For example, with 6 low birth weight infants out of 100 births, p = .06 describes the proportion of newborns having low birthweights. The notation also includes **q** where **q = 1 - p**. In this example, q = .94, which respresents the proportion not having low birth weights.

A. When the Numerator is at Least 20, and N Minus the Numerator is at Least 5

Please keep in mind that the upper case "**N**" refers to the denominator.

The 95% confidence interval is calculated by: p +/- 1.96 * SQRT (pq / N)

If 40 out of 250 newborns had low birth weights, calculate a proportion of 0.16. The 95% confidence interval would be:

0.16 +/- 1.96 * SQRT (0.16 * .84 / 250)
= 0.16 +/- 1.96 * SQRT (0.1344 / 250)
= 0.16 +/- 1.96 * SQRT (.0005376)
= 0.16 +/- 1.96 * 0.0232
= 0.16 +/- 0.045

Thus, p = 0.16 (95% CI = 0.115, 0.205). Note that these figures can be expressed as percents whereby low birth weight was 16% (95% CI = 11.5%, 20.5%)

B. When the Numerator is Less Than 20, or N Minus the Numerator is Less Than 5

We suggest that numerators and denominators be combined for three successive years. Then, if the conditions are met, calculations can take place as described above. There should be notes to make it clear that all calculations considered three years of information. If the total number of cases is less then 10 after combining three years of data, it is best not to report the proportion, just the numerator.

10 out of 200 newborns had low birth weights in 1998, 8 out of 198 in 1999 and 11 out of 203 in 2000. The 95% confidence interval would be:

(10+8+11)/(200+198+203) +/- 1.96 * SQRT ((10+8+11)/(200+198+203) *
1- (10+8+11)/(200+198+203)/(200+198+203))
= 0.05 +/- 1.96 * SQRT (0.0475 / 601)
= 0.05 +/- 1.96 * SQRT (.000079)
= 0.05 +/- 1.96 * 0.0088881
= 0.05 +/- 0.017

Thus, p = 0.05 (95% CI = 0.033, 0.067). Note that these figures can be expressed as percents whereby low birth weight was 5% (95% CI = 3.3%, 6.7%)

V. CONFIDENCE INTERVALS FOR RATES

A. With 100 or More Events in the Numerator

The 95% confidence interval can be calculated as:

R +/- 1.96 * R / SQRT (y)

where "**R**" is the rate, and the lower case "**y**" refers to the numerator.

With 250 births during the year and 5,000 women of fertile age, calculate the fertility rate as follows: (250/5000) * 1000 = 0.05 * 1000 = 50 births per 1,000 women ages 15-44. The 95% confidence interval is calculated as:

= 50 +/- 1.96 * 50 / SQRT (250)
= 50 +/- 1.96 * 50 / 15.8
= 50 +/- 1.96 * 3.16
= 50 +/- 6.2

Therefore, the fertility rate is 50 births per 1,000 women (95% CI = 43.8, 56.2).

B. With 20-99 Events in the Numerator

The upper and lower limits of the 95% confidence interval are calculated by multiplying the rate times the appropriate factors in Table 1 (see page 11).

C. With Less Than 20 Events in the Numerator:

It is useful to sum numerators and denominators for three successive years and perform the calculations. In such situations, it is important to show that multiple year figures were used. If, even after summing data across 3 years, there are not at least 10 events, then it is best not to report the rate, but to show only the numerator.

VI. ANALYSIS OF DIFFERENCES IN PROPORTIONS

A. When Comparing with a Benchmark Derived from a Very Large Population:

In many situations, we compare findings from a small population (a county) to benchmark figures derived from a large population (California or national data). It is sufficient to calculate the confidence interval for your study group and to determine if the benchmark proportion lies within the confidence interval. If so, then you accept that the proportion for the study population does not differ from the proportion for the benchmark population. If not, the difference is statistically significant.

B. When Comparing Two Proportions and Both Proportions are Subject to Variability

We often compare proportions for populations where the proportions for both groups are subject to substantial variability. For example, we might compare two groups within a community or we might compare two counties.

The proportions can be compared when both meet the following conditions: N * p is at least 5; and N * q is at least 5. In this situation, it is important to determine if the difference between the two proportions is greater than zero. Thus, it is necessary to calculate a confidence interval for the difference between proportions.

To compare low birth weight (LBW) rates between smokers and non-smokers:

If the LBW proportion for infants of mothers who smoked was 14/158 = .0886
And for babies of non-smokers, the LBW proportion was 48/842 = .0570
The difference between the proportions is 0.0316

The appropriate formula is:

$$.0316 \ +/- \ 1.96 * SQRT \{ pq (1/n_1 + 1/n_2) \}$$

In this formula, $\mathbf{p = (y_1+y_2) / (n_1+n_2)}$ and $\mathbf{q = 1-p}$. Note that the subscripts $_1$ and $_2$ mean that the calculation calls for data for group 1 and group 2, respectively. And $\mathbf{y_1}$ refers to the numerator for group 1, and $\mathbf{n_1}$ refers to the denominator for group 1.

p = (14+48) / (158+842)
= 62/1000
= 0.062

q = 1 - 0.062
= 0.938

The 95% confidence interval for the difference in proportions:

= 0.0316 +/- 1.96 * SQRT { 0.062 * 0.938 * (1/158 + 1/842) }
= 0.0316 +/- 1.96 * SQRT { 0.062 * 0.938 * (0.0063 + 0.0012)}
= 0.0316 +/- 1.96 * SQRT { 0.062 * 0.938 * 0.0075 }
= 0.0316 +/- 1.96 * SQRT {0.00044}
= 0.0316 +/- 1.96 * 0.021
= 0.0316 +/- 0.041

The point estimate for the difference between the two proportions is 0.0316 (95% CI =

VII. ANALYSIS OF DIFFERENCES IN RATES

A. When Comparing with a Benchmark Derived from a Very Large Population

Calculate the confidence interval for your study group's rate and see if the benchmark rate lies within the confidence interval. If so, then accept that the rate for the study population does not differ from the rate for the benchmark population.

B. When Comparing Two Rates, Both Rates are Subject to Variability, and Both Rates are Based on 100 or More Events in the Numerator:

At times, we compare rates for populations where the rates for both groups are subject to substantial variability. For example, to compare two groups within a community or across counties, use the following formula:

$$(R_1 - R_2) \ +/- \ 1.96 * \text{SQRT} [(R_1^2 / Y_1) + (R_2^2 / Y_2)]$$

In this situation, we want to know if the confidence interval for the difference between the two rates excludes zero.

Consider a county with 140 births and 2,000 women ages 15-44, with a fertility rate of 0.07 or 70 births per 1,000 women of fertile age. A neighboring county had 400 births and 8,000 women ages 15-44 with a fertility rate of 50/1,000. The difference between the rates is: 70-50 = 20 and the test statistic:

$$= 20 +/- 1.96 * \text{SQRT} [(R_1^2 / Y_1) + (R_2^2 / Y_2)]$$

where $\mathbf{R_1^2}$ means "the square of the rate for the first group", and
$\mathbf{Y_1}$ means the number of cases in the numerator of the first group.

The problem is solved as follows:

$$= 20 +/- 1.96 * \text{SQRT} [(70^2 / 140) + (50^2 / 400)]$$
$$= 20 +/- 1.96 * \text{SQRT} [(4900 / 140) + (2500 / 400)]$$
$$= 20 +/- 1.96 * \text{SQRT} [35 + 6.25]$$
$$= 20 +/- 1.96 * \text{SQRT} [41.25]$$
$$= 20 +/- 1.96 * 6.423$$
$$= 20 +/- 12.6$$

Therefore, the point estimate for the rate difference is 20 (95% CI = 7.4, 32.6), meaning that there is a statistically significant difference between the two rates (the confidence interval does not include zero). Even when we are 95% confident of a statistically significant difference between rates, there is still a 5% probability that the difference is due to chance.

Note that the formula given here works whether the rate is given as a decimal or calculated by the 1,000 (or 10,000 or 100,000), as long as there is consistency throughout.

C. When Comparing Two Rates, Both Rates Are Subject to Variability, and One Rate Has 20-99 Events in the Numerator

The simplified procedure is to calculate the confidence intervals for each of the rates, then to see if the two confidence intervals overlap. This approach is mathematically conservative, and it is reasonable to substitute the “1.96” in the test statistic with “1.41”.

D. When Comparing Two Rates, Both Rates Are Subject to Variability, and One Rate Has Less Than 20 Events in the Numerator

Comparisons are unlikely to be worthwhile, and only the number of events can be reported. An alternative is to sum numerators and denominators for three successive years, and then perform calculations as above. In such situations, it is important to report that multiple years of data were employed for the comparison.

VIII. ANALYSIS OF RATIOS OF RATES OR PROPORTIONS

In the examples given in earlier sections, the calculations examined differences between two rates or between two proportions. An alternative approach is to present the ratio of rates or ratio of proportions.

A. When the Benchmark Rate or Proportion is Derived from a Large Population and is Considered Fixed

Calculate the point estimate and confidence interval for the study population using methods for proportions or rates (as appropriate to the measure). Divide each figure by the point estimate for the benchmark population.

> If 12% of the study population has the condition (95% CI = 9%,15%), and the percent for the benchmark population is 6%, then we can describe the risk ratio as 2.0 (95% CI = 1.5, 2.5). This means that the condition is estimated to be twice as common among the study population with a 95% confidence interval which ranges from 1.5 to 2.5 times as common. Note that, if the 95% confidence interval includes the number “1”, then you cannot conclude that the rates are significantly different.

For rates when the benchmark rate is considered as fixed, then the same comparison can be made.

B. When Rates or Proportions for Both Groups are Viewed as Variable

The point estimate is calculated in the same manner as above. The calculation for the confidence interval, however, is complicated and beyond the scope of this guide.

IX. WHEN NUMBERS ARE TOO SMALL FOR STATISTICS

A. Aggregating Data Across Years

We have already discussed calculation of statistics for multiple years when the number of cases or events is less than 20. The numerators and denominators can both be summed over three years to get a proportion or rate for the total period. The convention is to note that calculations used three years of data. It is important to point out that this approach is not useful for analysis of trends. This is because the same pieces of data would be used repeatly. Thus, aggregation of data across years should be used only to compare across populations and not across time. The exception would be if comparison is made between two time periods which do not overlap.

If aggregation of data for three years still does not produce at least 20 cases or events, then one could go further and aggregate data for five years.This approach is uncommon because it undermines one's ability to look at trends over time.

Such **aggregate rates** (or proportions), where both numerators and denominators are summed, are called **weighted averages**, because the rates for each of the years are weighted for the respective denominators. There is a short-cut when the denominators do not substantially change from year to year, or when they change very evenly. This is typical for denominators using total population. In such situations, it is acceptable to take a simple average of the rates or proportions for the respective years. While this approach helps to smooth the data and reduce the annual fluctuation, it cannot be used when any kind of statistical testing is desired or if you want to calculate a confidence interval. In such situations, the actual numerators and denominators for each of the years are needed for statistical power.

B. Geographic Aggregation

This is an epidemiological technique that can be used with relatively uncommon events or diseases in order to gain statistical power and decrease variability in the area under study. In this technique, data geographic areas are aggregated by proximal geographic location or are grouped by defined characteristics such as the extent of poverty or average level of education.

To calculate the numerator, add the number of cases or events that occur in a given time period for neighboring health jurisdictions (e.g., zip codes, census tracts or other geographic regions) or for geographic areas grouped by characteristics (e.g., poverty, race/ethnicity, hazard exposure, etc.). To calculate the denominator, add the specified populations (e.g., total population, specified age, gender, or race/ethnicity groups) for the different geographic areas of the cluster. Then the rate or proportion can be calculated in the normal manner.

The reader is advised that, when aggregating geographic areas, it is inappropriate to group areas according to the outcome variable. Doing so would force a positive statistical finding. A better approach is to group geographic areas by exposure, then to test whether the grouped areas differ in outcomes.

C. Case Studies

When numbers are too small for statistical comparisons, we suggest the use of case studies. These are often conducted by a Child Death or Fetal and Infant Mortality Review committee. Such committees generally review medical, social service, or coroner case records and other written documents that are available. Additional information can be collected through interviews with medical providers, parents, social service or other community agency providers in order to determine the factors which may have contributed to the death. These findings are reviewed by an interdisciplinary, interagency group to determine whether the case was preventable, and findings can be used to formulate intervention strategies.

This process can be extended to study of ***case series***, which entails exploring a handful of individual case records to determine whether there are common risk factors among the cases. Factors of interest may include: cause of death; diagnoses; geographic location; type or source of health care received; individual or family behavioral conditions; environmental and socio-economic factors; etc. This process can be conducted using vital statistics records, such as birth and death certificates, or by reviewing medical records. For both case studies and case series, if the data are computerized, **line listing reports** of all the variables of interest in a particular data set can be generated to facilitate consistent record review.

X. REPORTING YOUR FINDINGS

The National Center for Health Statistics does not publish a rate when the rate is based on fewer than 20 cases or events.[2] Instead, they insert an asterisk (*) at the appropriate point in the table. The reason for this is that, when the rate is based on small numbers of events, the associated confidence interval can be relatively wide. We agree with this practice in reports where numerators are not contained within the report.

On the other hand, research scientists frequently do statistical analyses with fairly small numbers. But they invariably show the confidence interval along with the rate. Thus, in an ideal situation, a good report would contain information on the numerator, the rate or proportion, and the confidence limits. We recognize that such complete reporting of data can confuse readers who typically have no statistical training.

Therefore, we recommend the following practices:

1. As much as possible, the number of cases or events should be shown in published reports. This is because, while rates and proportions are useful, we are ultimately concerned with the overall dimensions of public health problems.

2. When rates or proportions are not based on at least 20 cases or events, then they should be omitted from the report, or the report should contain clear notation about the confidence interval and a simple statement of its meaning.

3. We encourage the practice of placement of information about confidence intervals in reports which contain rates or proportions, partly because we believe it important to educate our readers about the nature of statistical issues.

That said, it is important to point out that too much can be made of statistical comparisons of rates and proportions between groups or geographic areas. A county can compare itself with another locale. But, even if, by every measure, the county appears advantaged, this would not mean that the county has no problems. There are still the problems of access to prenatal care, low birth weight, infant mortality, etc. All that would be learned from statistical comparisons is that the county's problems are not as great as in the comparison community.

Thus, while a positive finding from statistical comparison can help to focus attention on a problem, it is always important to first look at the raw numbers in order to appreciate the dimensions of the problem.

[2] For greater detail, the reader is referred to the Technical Notes section of Ventura SJ, Martin JA, Curtin SC, Mathews TJ, Park MM. Births: Final data for 1998. National Vital Statistics Reports; vol 48 no. 3. Hyattsville, Md: National Center for Health Statistics. 2000.