Guidelines for State Analyses of PRAMS Data

Table of Contents

Part 1CIntroduction

These guidelines present an approach for doing descriptive analyses of PRAMS data using specialized survey software called SUDAAN. Parts 2 and 3 review general concepts that relate to analysis of survey data and define the technical terms that will be used in subsequent parts of the guidelines. Part 4 defines three types of analyses of PRAMS data that will be discussed in these guidelines. Although the general analytic approach is similar for each of these three types of analysis, the interpretation of the results varies. Parts 5 and 6 present technical aspects of SUDAAN, the pc survey software that is recommended for analysis of PRAMS data. Part 7 describes approaches for dealing with missing data. Part 8 presents a discussion of general approaches for analyzing a subpopulation of the state. And the final section, Part 9, introduces SUDAAN's modeling procedures.

Two appendixes are referred to in these guidelines. The first, found in the PRAMS Mail/Telephone Protocol Appendix J, gives a detailed technical explanation about the computation of analysis weights. The second, found in the PRAMS Mail/Telephone Protocol Section 12, is a glossary of technical terms.

These guidelines present some basic epidemiologic principles and are intended as a starting point for analysis of PRAMS data. In some instances, the analytic approaches recommended are not the only ones possible. To avoid undue complexity, for each analytic situation, we presented one approach that seemed the most straightforward. However, there may be alternative approaches in these situations. While the guidelines do not provide a full discussion of epidemiologic terms or analytic techniques, familiarity with the analytic approaches described will give the reader the groundwork needed to conduct epidemiologic analyses in the future.

Part 2CGeneral Concepts Relating to Data Analysis

Categorical and Metric Variables

Throughout this discussion, we will talk about two types of variables: *categorical* and *metric*. Each type requires different statistical techniques.

***Categorical variables* have a finite number of values, each representing a different category. Categorical variables can be divided into nominal and ordinal types. *Nominal variables* do not have an inherent order. Examples of nominal variables are gender (male/female), race, and, religion. Variables that derive from a yes/no/don't know type of question are also categorical nominal variables. An example from PRAMS is the variable for question 22, which asks the mother whether she was on WIC during her pregnancy. In contrast, *ordinal variables* can be ordered into qualitative categories with a distinct order, but with no defined numeric distance between them. An example from PRAMS is the variable for satisfaction with prenatal care. We know that a response of "satisfied" conveys greater satisfaction than a response of "dissatisfied," but we are not able to measure this difference numerically.**

For basic analysis of PRAMS data, the distinction between nominaland ordinal variables can be ignored; nominal and ordinal variables can both be analyzed as nominal variables. However, it becomes important to recognize a variable as ordinal when one wants to compute statistical tests for trends. These tests can be computed for ordinal variables, but not nominal variables.

Categorical variables have as many values as there are categories. When categorical variables have 2 values, they are sometimes called *dichotomous* variables. Categorical variables with more than 2 values are sometimes called *polychotomous* variables. An example of a polychotomous variable is the categorical variable from PRAMS Question 7 for wantedness/timing of pregnancy. The question, "Thinking back to just before you were pregnant, how did you feel about becoming pregnant?" has 6 possible values (1=earlier; 2=later; 3=then; 4=never wanted to be pregnant; .d=don't know; and .b=blank). Categorical variables are usually coded as 1 if the response is in category 1, 2 if it is in category 2, and so on. It is not meaningful to calculate the arithmetic mean or average of responses to a categorical variable. Instead, categorical variables are analyzed by computing the percentage of responses in each category. For example, consider the PRAMS variable for wantedness/timing of pregnancy. Let's restrict this example to the following 4 response categories: sooner, later, then, and never. They are coded 1, 2, 3, and 4. It would not be meaningful to compute an average of the responses. Instead, we would compute the percentage of responses that are in each category. *Metric variables* have a number of possible values along a scale that ranges from low to high values. Each point along the scale is a specific, numerical distance from the next point. As a result, the value

of a metric variable can be used to make quantitative comparisons. For example, a mother who had 20 visits for prenatal care had twice as many visits as a mother who had 10 visits; or, 200.2 lbs is twice as much as 100.1 lbs.

One type of metric variable—called a *continuous* variable—can take on an infinite number of values. For example, if one had an accurate enough scale, one could measure maternal prepregnancy weight as 123.45 lbs, 123.452 lbs, or 123.4527 lbs. Therefore, we would consider weight to be a continuous variable. In contrast, another type of metric variable, called a *discrete variable,* can take on only integer values. Examples of discrete variables in PRAMS are the number of visits for prenatal care (someone could have 8 or 9 visits, but not 8.5 visits) and the number of cigarettes smoked (someone could report 1 or 2 cigarettes, but not 1.2 cigarettes).

Discrete variables are often analyzed using the same techniques as for continuous variables and we will follow that approach here. However, when applied to discrete variables, these techniques may give results that are not possible. A well-known example is the joke about the average number of children per family being 2.3. We know that it is impossible to have 0.3 of a child. This type of result comes from treating number of children—which is a discrete variable—as a continuous variable.

Continuous and discrete variables can be converted to categorical variables by grouping values together. This is commonly done with birthweight, which is a continuous variable. For example, by putting birthweights < 2,500 grams in one category and birthweights $\geq$ 2,500 grams in another category, birthweight is converted into a categorical variable with two values, low and normal birthweight. Categorical variables created in this manner can be analyzed as ordinal variables.

An advantage of converting continuous variables into categorical variables is that, from a statistical perspective, categorical variables can be easier to work with because fewer statistical assumptions are needed. In addition, the results of categorical analyses may be easier to understand and present.

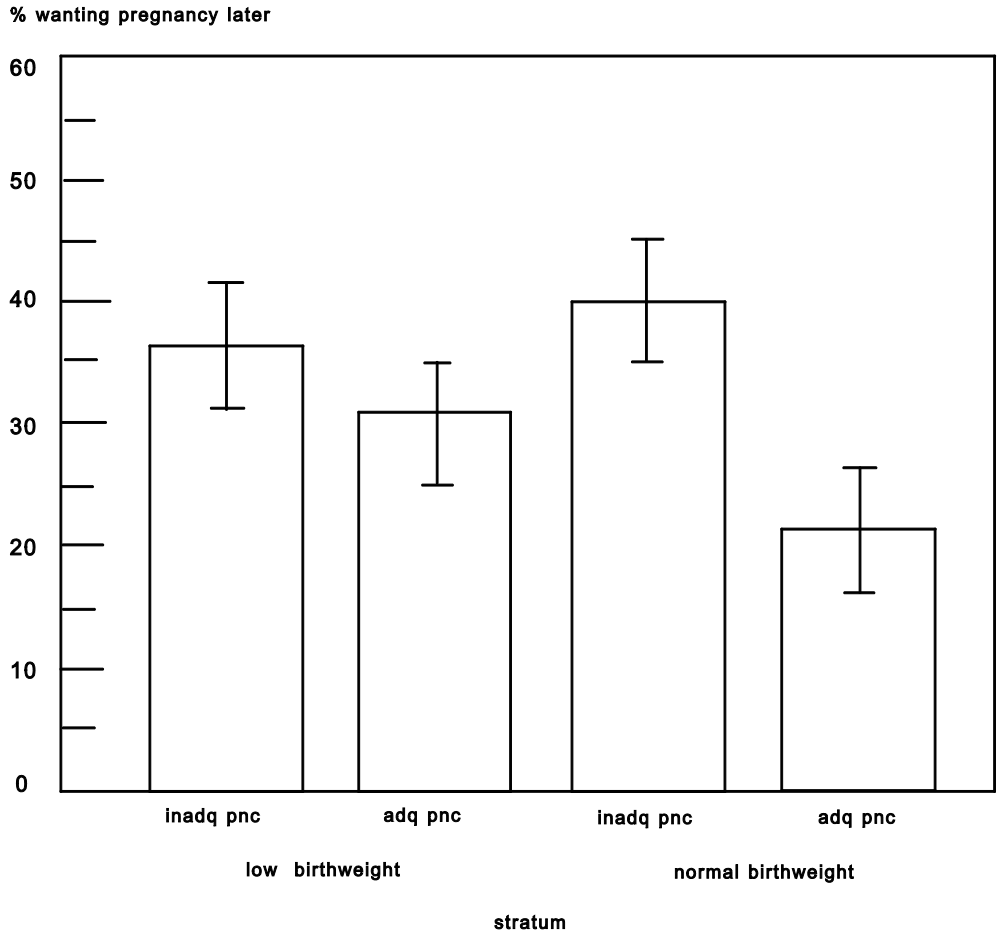
Disadvantages of converting continuous variables into categorical variables are that:

1. **This conversion may over-simplify relationships among the variables in the analysis.**

2. **By failing to use all of the available information about the continuous variable's distribution, the categorical variable can be statistically less efficient.**

When analyzing categorical variables, one usually computes the percentage distribution of respondents in each category or for each value of the variable. The sum of the percentages for each of the categories equals 100%. A graphic way of depicting these categories is to draw a bar chart such as shown in Figure 1.

Figure 1.
Percentage of mothers who wanted their pregnancy to have occurred later by their infant's birthweight and adequacy of prenatal care. State data 1988-89.
(bars show 95% confidence intervals)



The percentages computed from *sample data* are point estimates of the corresponding percentages in the *entire population* of mothers, i.e., the *population parameters.* The population parameters are what we ideally would like to measure, but, to do so we would have to contact every mother in the population. Because it is too hard to contact everyone, we contact a sample of the population instead. Then we use the information about this sample to estimate the population parameters.

A concept related to the estimate of the population parameter is that of the *confidence interval.* The confidence interval is a range of possible values for the population parameter; this range is constructed so that it has a specified probability (usually 95%) of including the population parameter. For example, we might estimate that, based on our sample data from PRAMS, the percentage of women in stratum "X" who said their pregnancy occurred sooner than they wanted was 19.8%; in other words, 19.8% is calculated from the sample data and it is the point estimate of the population parameter. The 95% confidence interval, 16.7%-22.8%, is also calculated from the sample data.

The interpretation of the 95% confidence interval is as follows. If the same sampling procedure were used to obtain many samples, and if a 95% confidence interval for a given population parameter were calculated from each sample, then 95% of the confidence intervals would actually include the value of the population parameter; 5% would not. Of course, a researcher selects only one sample and calculates one confidence interval for a given population parameter. A "practical" interpretation of the 95% confidence interval is it includes the value of the population parameter with a probability of 0.95. Using the example in the previous paragraph, we could say that there is a probability of 0.95 that the percentage of women in the population who said their pregnancy occurred sooner than they wanted lies between 16.7% and 22.8%.

The "width" of the confidence interval--in other words, the distance between its upper and lower limitsCgives an impression of the variability of the data. A wide confidence interval means that the data are highly variable and should be interpreted with caution. A narrow confidence interval indicates that the data are fairly stable and therefore, more suitable for use in decision-making. A larger sample size will yield a narrower confidence interval. As shown in Figure 1, the confidence interval can be depicted on a bar graph.

It is often of interest to examine whether the *percent distribution* for one categorical variable varies in relation to another categorical variable. For example, in PRAMS, we will often be interested to see whether the percent distribution varies among strata (the variable representing stratum is categorical, with as many values are there are strata). Does the percent distribution of wantedness/timing of pregnancy vary across strata? Usually the first step in answering this question involves constructing a *contingency table.* This is a tabular cross-classification of the two categorical variables, stratum and wantedness/timing of pregnancy as shown in Table 1. Each of the two variables in a

contingency table can have two or more categories. Table 1 presents "row percentages." For example, in row 1 (the stratum of low birthweight/inadequate prenatal care), 46% of the mothers wanted their pregnancy then/sooner and 54% wanted their pregnancy later/never. Note that the percentages for this row sum to 100%; the same is true for the other rows.

Table 1CWantedness/Timing of Pregnancy

Percent distribution of wantedness of pregnancy at the time conception occurred, by adequacy of prenatal care and infant birthweight, West Virginia, 1988-89.

Stratum	n	Later/Never	Sooner/Then	Total
Low BW, Inadeq PNC	**321**	**53.1%**	**46.9%**	**100%**
Low BW Adeq PNC	**306**	**34.7%**	**65.3%**	**100%**
Normal BW Inadeq PNC	**401**	**49.6%**	**50.4%**	**100%**
Normal BW Adeq PNC	**489**	**29.4%**	**70.6%**	**100%**

NOTE: **Low BW** = birthweight < 2,500 g;
Normal BW = birthweight ≥ 2,500 g.
Adeq PNC = adequate prenatal care, defined as starting in the first trimester and resulting in a total number of visits appropriate for length of gestation;
Inadeq PNC = inadequate PNC, defined as all other prenatal care, including unknown month of first visit, unknown number of visits, and/or unknown length of gestation.

Often contingency tables show the actual frequencies (numbers) of subjects in each "cell" of the table. (A cell is one of the cross-classified entries in the table; Table 1 has 8 cells: 53.1%, 46.9%, 34.7%, 65.3%, 49.6%, 50.4%, 29.4%, and 70.6%). As a rule of thumb, we feel that when percentages are presented, the numbers (*n*) or frequencies from which they derive should also be presented. Therefore, for each stratum, Table 1 includes the number of mothers who were included in the analysisCthe number of mothers in each stratum who answered this question.

Information about the number of mothers included in the analysis gives another clue about the likely reliability of the findings. All other factors being equal, we would infer that a larger number of mothers are likely to produce more stable results. Because the analysis weights were used to compute the row percentages, the number of respondent mothers in each cell cannot be obtained from the table.

The chi square test is used to determine whether the two variables in a contingency table are *associated* with each other. By "associated," we mean related to each other in some systematic manner. This application of the chi square test uses two variables, both of which must be categorical. For the data in Table 1, an association seems to exist since the percentage distribution of wantedness is not the same for all four strata (all four rows of the table). If the row percentage distributions differ enough, then such differences are often described as *statistically significant.* In effect, a chi square test tells us the likelihood or probability that the observed distribution occurred by chance, or whether the observed distribution occurred in the *absence* of an association between the two categorical variables. If the probability is low, for example, less than 5%, we interpret it as meaning that the results are probably not a chance occurrence and, instead, may represent a true association. (Of course the ultimate decision about whether an association exists depends on several other considerations besides the results of a statistical test. Among these factors are biologic plausibility, strength of the association, and validity of the data.)

When interpreting the results of a chi square test, you need to keep in mind that it is an "omnibus" test of a difference among the cells in the contingency table. In other words, a statistically significant result of a chi square test means that one (or more) of the cells differs from the value that would be expected if no association existed. By itself, a statistically significant chi square test does not tell you *which* cell(s) differs from its expected value; a statistically significant chi square test tells you only that at least one cell differs from its expected value. To identify the cells that differ from the values that would be expected for them if no association existed, you need to conduct additional *pairwise* analysis.

In Table 1, adequacy of prenatal care appears to be associated with wantedness of pregnancy. Mothers with adequate prenatal care (strata 2 and 4) were less likely to say that they wanted the pregnancy to occur later or not at all. Birthweight also appears to be

associated with wantedness of pregnancy. In fact, the result of the chi square test (X^2 = 82.034, 3 degrees of freedom, $p < .001$) confirms our suspicion of a statistically significant association. To specifically identify the cells that differ from the values expected in the absence of an association, we need to perform a pairwise analysis.

If you have determined that a statistically significant association is present and identified the cells that differ from their expected values, you will probably want to describe the size or magnitude of the association. For example, from a public health point of view, it may be important to know whether, compared to mothers of normal weight babies, mothers of low birthweight babies were twice as likely to report that they did not start prenatal care as early as they wanted or four times as likely to do so. This can be done by making a contingency table using weighted data (such as the data in Table 1), and from the table, computing an epidemiologic measure called an *odds ratio.* We will defer further discussion of the computation and interpretation of odds ratios until the future. However, at this point, we want to emphasize that computation of a chi square test and pairwise analysis may not be the end of the analysis. Additional steps in the analysis may involve computing a measure of the magnitude of the association and evaluating other factors for their possible influence or bias on the association.

Analysis of Metric Variables

Analysis of metric variables often begins by computing three measures of central tendency: the mean (often called the average), the median (the 50th percentile, i.e., the value that separates the observations into two equal halves), and the mode (the most frequent value). In addition, one often computes quartiles or other percentiles. The 25th percentile (equivalent to the lower quartile) is the value below which 25% of the observations lie; the 75th percentile (equivalent to the upper quartile) is the value below which 75% of the observations lie. Examined together, these measures give an impression of how evenly and closely or widely the data are distributed. A graphic depiction of the data can be developed by plotting the frequency distribution, such as shown in Figure 2, or the cumulative percentile distribution.

Along with computing the sample point estimates for the population mean and population quartiles/percentiles, one often computes the confidence intervals for the population parameters. As mentioned under the section for categorical variables, computation of the confidence interval permits assessment of the variability of the data.

After looking at the data from one stratum, you will probably want to compare the data among strata. If you want to preserve the data in its metric form, there are several ways to do this:

1. **By computing a one-way analysis of variance (ANOVA).**

2. **By comparing two strata at a time.**

3. **By estimating specific contrasts among the strata.**

All of these options are available in SUDAAN. A pairwise comparison and some contrasts are conceptually comparable to performing a *t-test.* If you are willing to convert the data into categorical form, another analytic alternative is to construct a contingency table and compute a chi square test and pairwise analysis. As noted earlier in the discussion of metric variables in Part 2, when you convert metric data into categorical data, you trade an easier analysis for possible loss of statistical efficiency and possible over-simplification of the association under study.

Interpreting Results When Many Statistical Tests Have Been Done

When doing a pairwise analysis to compare strata or to compare levels of any categorical variable, you will probably compute many statistical tests. If each one is done at the common alpha (Type I) level of 0.05, then the *overall risk* of a Type I error from *any* of the tests can be much larger than 0.05. *(Type I* error is the probability of obtaining a statistically significant result, when, in fact, no association exists). Stated simply, if you do enough tests at an alpha level of 0.05, you will eventually get statistical significance for one of the tests, even though there are no real associations in the data.

Some analysts are concerned about the problem of increased Type I error due to multiple testing. It is possible to make adjustments in the Type I error for each test so that the overall Type I error for all tests is approximately 0.05. One such adjustment is the Bonferroni method. However, this method, as well as other methods, are not perfect fixes for the problem.

In general, be aware that multiple testing using a fixed Type I error can result in an overall Type I error that is actually much larger. This problem underscores the need to take into account other relevant information when interpreting statistically significant results. Examples of other information include biologic plausibility, correct temporal sequence (did the effect happen after the cause?), and magnitude of the association.

Interpreting Results When the Data are Sparse

A concept related to Type I error is the concept of *Type II* error. Type II error is the probability of *not* obtaining a statistically significant result when, in fact, an association exists. Type II errors are most likely to occur when the number of subjects included in the analysis is small, in other words, when the data are sparse. Because of the possibility of Type II errors, when interpreting a statistically nonsignificant result, you need to consider the possibility that it is due to sparse data.

Part 3CRationale for Development of Analysis Weights: Taking into Account Sampling Design and Nonresponse

PRAMS Population, Subpopulations, and Sample

For PRAMS, we want to be able to infer behaviors among the entire group of women who delivered a live born baby within a specified period of calendar time in a particular state and who were residents of that state. This group of women forms the *population* of interest; the individual women whom we have selected to contact are a *sample* from this population.

Because of the public health importance of certain subpopulations of mothers, PRAMS uses a sampling approach that oversamples some subpopulations within the population. For example, mothers of low birthweight infants have a higher probability of being selected than mothers of normal birthweight infants. With this approach, we are assured of having an adequate sample size for analyses of subpopulations of particular interest. For the purpose of this discussion, we will define a *stratum* as a subpopulation of mothers with similar characteristics from whom a sample was selected. All of the mothers within a stratum have an equal probability of being selected. However, the probability of being selected generally differs from one stratum to another.

What about subpopulations that are not defined as strata? For example, the group of women who smoked during pregnancy forms a subpopulation that includes women from all strata. For this discussion, we will refer to a subpopulation that is *not* a stratum as a *domain*. The reason for making a distinction between a stratum and a domain is that the statistical formulas for computing variances differ for strata and domains. However, conceptually, both terms refer to subpopulations.

What is an Analysis Weight?

In the context of PRAMS, an *analysis weight* is a number that is used to adjust for the effects of the sampling design, nonresponse pattern, and omissions from the sampling frame. In the PRAMS data set, each mother who completed (or partially completed) a questionnaire has a computed weight that will be used in analyses. We will refer to this weight as the analysis weight. The analysis weight for a mother who responded to the questionnaire is the number of women like her whom she is representing in the population. In general, mothers in the same stratum with similar characteristics (such as age, education, or marital status) will have the same analysis weights.

A result of the stratified sampling in PRAMS is that the unweighted distribution of maternal characteristics in the PRAMS sample differs from the "true" distribution of

these characteristics in the population of all mothers. Therefore, the first and most important component of the analysis weight adjusts for the fact that not everyone had the same chance of being selected for the sample (i.e., PRAMS is not an *equal probability* sample of mothers).

For each stratum, this first component derives from the *sampling fraction.* This is a ratio that is computed separately for each stratum. The sampling fraction has the number of mothers sampled from a particular stratum in the numerator and the number of mothers on the sampling frame for that stratum in the denominator. If n_h = the number of mothers sampled from stratum h and N_h = the number of mothers on the sampling frame in stratum h, then the sampling fraction can be expressed as:

$$n_h / N_h$$

The sampling fraction is also the *selection probability* for any mother within the stratum. The selection probability can range from greater than 0 to 1, with larger probabilities representing higher chances of selection. A selection probability of 1 means that the mother has 100% chance of being selected. For PRAMS, the sampling fractions (and selection probabilities) vary among strata.

Response rates have also varied between strata. These differences could distort the analysis by causing over- or under-representation of a particular stratum. To avoid this, a second component of the analysis weight adjusts for differences in response rates between strata. (In the jargon of statistical sampling, failure to participate by a member of the sample is referred to as *unit nonresponse.* In practical terms, unit nonresponse means that all of the questionnaire data for an individual are missing).

The *sampling frame* is the list of mothers from whom the sample is selected. Because PRAMS uses a stratified sampling design, each stratum has a separate sampling frame. Ideally, the combination of the frames of all the strata should include all mothers in the population. However, sometimes this does not occur because one or more of the stratum-specific frames are incomplete. An omission in PRAMS that probably affects all of the stratum-specific sampling frames is birth certificates that were processed too late for inclusion.

Therefore, a third component of the analysis weight adjusts for omissions from the sampling frame. In PRAMS, this adjustment relates to mothers who were not "covered" by the frame. The need to adjust for omissions depends on their extent and the differences between omitted and included mothers. If the magnitude of the omissions is trivial, this third component of the analysis weight is not needed. Because assessment of the adequacy of the sampling frame has not been completed for all states, the analysis weights computed

at present do not adjust for omissions from the frame. They *do,* however, include adjustment for sample design and nonresponse.

Why are Weights Needed?

Analysis of the data without using an analysis weight that at least includes the first component (adjustment for the sampling design) produces incorrect results (referred to in statistical jargon as *biased estimators)* for descriptive measures (proportions, means), variance estimates, and statistical tests. Many standard statistical software packages can incorporate weights; however, they do not correctly compute variance estimates or perform statistical tests for data that have resulted from a stratified or other more complex sampling design. In particular, although PC-SAS can be run on PRAMS data using weights, the statistical tests computed by PC-SAS will not be performed correctly. For this reason, we strongly recommend that states use software specifically developed to account for PRAMS' stratified sample design. The Research Triangle Institute (RTI) has developed pc software for analysis of stratified data as well as data from other complex sampling schemes. This software CSUDAAN C computes means, proportions, and contingency tables. The software also includes linear regression, logistic regression, categorical data analysis, and survival analysis.

What are the Limitations of Using Weights?

Weights may not compensate adequately for low response rates. The component of the analysis weight that adjusts for nonresponse rests on the assumption that the average of the answers of the respondents within the particular stratum and response category under consideration is the same as the average of the answers of the nonrespondents in that stratum and response category (see Appendix J in the PRAMS Mail/Telephone Protocol for a definition of response category).

For example, consider the stratum in West Virginia of mothers who had inadequate prenatal care and delivered a low birthweight infant. Analysis of response patterns for this stratum showed that marital status was related to the likelihood of response. Therefore, separate analysis weights were developed for the respondents in this stratum who were married and those who were not married. The adjustment for nonresponse that goes into these analysis weights rests on the assumptions that, overall, the answers given by the married respondents are the same as the answers by the married *non*respondents and that the answers given by the unmarried respondents are the same as the answers by the unmarried *non*respondents.

While these assumptions seem reasonable for strata with response rates of 70% or higher, they become increasingly tenuous for strata with lower response rates. For strata with response rates below 50%, these assumptions seem unjustified. Similar reasoning leads to

the conclusion that weights cannot adequately compensate for large omissions in the sampling frame. Therefore, low response rates or large omissions in the sampling frame may preclude interpretation of the data.

How are Analysis Weights Developed?

Details of the procedure used by CDC to develop analysis weights are given in Appendix 1. New weights need to be computed every time the sampling fractions change. Therefore, it is essential that a record of changes in the sampling fractions for each stratum be maintained. New weights also need to be computed if stratum-specific response rates change markedly. CDC will compute weights for each state yearly and more frequently if the sampling fractions or stratum-specific response rates change. States are advised to use these weights in their analyses; CDC will use them in its comparisons of data among states.

Because stratum-specific response rates are needed for computation of the second component of analysis weights, they cannot be computed until data for the interval under consideration are available for analysis.

Part 4CThree Types of Analyses

Stratum-Specific Analyses

As part of developing the PRAMS sampling strategy, each state identified subpopulations of women that were of particular public health interest. For the purpose of selecting the sample, these subpopulations became strata. Implicit in the identification of strata was the assumption that health behaviors vary between strata. To develop a feeling for the variation between strata, we will begin the analysis by looking at each stratum separately. In other words, we will compute stratum-specific estimates.

Understanding the amount of variation between strata is an important step *before* computing a statewide estimate. If there is a lot of variation between strata, the overall statewide estimate may not accurately portray the variation and therefore could be misleading. In this situation, careful interpretation of the data is needed. Specifically, it must be recognized that the statewide estimate is not necessarily representative of each stratum.

Statewide Analyses

For purposes of PRAMS, we will define statewide analyses as those that combine all of the mothers into one group. This contrasts with stratum-specific analyses, which treat each stratum as a separate group or subpopulation, and domain analyses (described below), which treat each domain as a separate group or subpopulation.

Statewide analyses that describe the experience of the state as a whole provide estimates of the number, as well as the percentage, of mothers with certain behaviors/conditions. Information about the number of mothers in a state with a certain behavior can be useful for estimating the need for services related to that behavior. Information about the percentage of mothers with a certain behavior is useful for identifying the prevalence of behaviors and monitoring overall trends. Statewide analyses can also be done on the relationship between two variables. This information is useful in summarizing the experience in the state of the relationship between these variables.

When interpreting statewide data, it is important to consider that the overall results for the state may not be the same as the results for particular subpopulations in the state. For example, consider the situation where the statewide prevalence of maternal smoking during the 3 months before pregnancy began is 30%. However, the prevalence among mothers with 16 or more years of education is 15% and the prevalence among mothers with 12 or fewer years of education is 38%. A difference between the statewide prevalence and prevalences for subpopulations in the state does not mean that either set of figures is invalid; both sets are appropriate to specific uses and should not be "over-interpreted."

Analyses of Domains

For the purposes of PRAMS, we have defined a domain as a subpopulation that is not a stratum. For a particular analysis, the domains are mutually exclusive. Although the examples will compare two domains, any number of domains may be compared as long as they do not overlap.

By comparing domains, we can learn about differences that may point toward opportunities for prevention activities or, conversely, that may help in assessing utilization patterns of these activities. For an example of the former, consider a comparison between the domain of women who received most of their prenatal care at a health department clinic and the domain of women who received most of their prenatal care at a private doctor's office or HMO. We might be interested to know if the proportion of women who were asked if they smoked (Ql6) was the same or different for both providers. If detected, a difference might suggest that one of the providers needs to be encouraged to ask mothers about smoking.

We may also want to compare domains to gain greater insight into the presence or absence of relationships between variables. For example, consider a comparison of the domain of mothers who, in response to Q7, said they wanted their pregnancy to occur sooner or at that time (i.e., planned pregnancies) and the domain of mothers who wanted their pregnancy to occur later or not at all (i.e., unplanned pregnancies). We may be interested to see if these domains differ with respect to initiation of prenatal care, smoking during pregnancy, infant birthweight, breast-feeding, and use of routine baby care.

A consideration to keep in mind when comparing domains or strata is that any differences observed between domains or strata may be due to factors beyond those that define the domains, i.e., confounding factors. At this point, we are not presenting techniques that would permit you to take account of or control for these potentially confounding factors. In the absence of these control techniques, results of domain comparisons must be interpreted cautiously.

Part 5CIntroduction to SUDAAN

How SUDAAN Differs From Other Statistical Software Packages

SUDAAN is a specialized statistical software package for analysis of sample survey data. Its formulas for standard errors and statistical tests of significance take into account the survey design as described to SUDAAN by the user. The common statistical software packages, e.g., SAS, SPSS-X, BMDP, SYSTAT, and EPI-INFO calculate standard errors and variances based upon the assumption of simple random sampling.

When a sample survey departs radically from simple random sampling, as most surveys do, use of the standard statistical software packages will yield incorrect analyses of the data. In surveys with clustering, use of standard statistical software generally will yield smaller standard errors than is actually the case, in turn producing "too many" statistically significant findings. For stratified random or systematic sampling, like PRAMS, use of standard statistical software likely will overestimate the standard errors and may lead to missing statistically significant associations.

SUDAAN probably will not appear as friendly as the statistical software packages with which you have experience. We hope that friendliness will increase as new developments are added to the package. However, we recommend that SUDAAN, or some other survey software package, be used to analyze sample survey data.

The formulas in SUDAAN for estimating variances and standard errors are approximations based on the Taylor Series linearization technique. Methods other than the Taylor Series linearization are available for obtaining these approximate formulas, e.g., balanced repeated replication and jackknifing, but they use much more computer time. Almost all survey software packages use the Taylor Series linearization technique to get approximate standard errors and variances.

SUDAAN and SAS

SUDAAN is a stand-alone software package, whereas its mainframe predecessor operated within SAS as a SAS procedure. SUDAAN will take a SAS data set as the data input file, and this is definitely the easiest way to use SUDAAN. The only other current method of data input is via an ASCII file, but

the required statements to specify variable names and location of variables in the file seem cumbersome. Other methods of data input are under development.

Probably because of its history as a SAS procedure, the SUDAAN programming language is similar to the SAS language. Also, some of the analyses in SUDAAN parallel available

procedures in SAS, e.g., PROC FREQ and PROC MEANS. Those who are familiar with SAS may learn SUDAAN more easily.

SUDAAN Procedures and Output

SUDAAN currently contains eight PROCS or procedures. These guidelines will emphasize two of the procedures: CROSSTAB and DESCRIPT.

PROC CROSSTAB produces cross tabulations of categorical variables and does chi-square tests. It is similar to PROC FREQ in SAS.

PROC DESCRIPT gives means and percentiles for metric variables and will compare strata or domains on a given metric variable. PROC DESCRIPT can also summarize categorical variables and be used to compute pairwise comparisons of categorical variables. Options in PROC DESCRIPT are similar to PROC MEANS, PROC UNIVARIATE, and PROC TTEST in SAS.

PROC RATIO is used for ratio estimators. PROC DESGCHK will check your description of the survey design for inconsistencies. PROC REGRESS fits general linear models to survey data, including linear regression, analysis of variance and analysis of covariance. PROC LOGISTIC fits a logistic regression model to survey data. In PROC REGRESS, the dependent variable is continuous (or assumed to be so), whereas in PROC LOGISTIC the dependent variable is dichotomous or ordinal. PROC SURVIVAL fits the discrete proportional hazards model or Cox's proportional hazards model to survey data. PROC CATAN provides linear and loglinear modeling capabilities for contingency table analyses of survey data.

Standard output from SUDAAN includes point estimates, their corresponding standard errors and *design effects.* The design effect for a given point estimate is defined as the variance (i.e., the square of the standard error) of the point estimate obtained from SUDAAN divided by the variance of the same point estimate which would have been obtained under simple random sampling (i.e., a different survey design with the same sample size). The design effect compares the efficiency of the PRAMS survey design with the hypothetical efficiency that would have resulted if simple random sampling had been used instead with the same sample size. (Statisticians often consider simple random sampling as the "gold standard" among survey designs.) In sample surveys that involve clustering, the design effect generally is larger than 1.0, indicating that cluster sampling is although almost always less expensive than simple random sampling is not as efficient. In the PRAMS stratified random sample the stratum-specific design effects will be around 1.0, or somewhat larger, whereas the statewide design effects are likely to be larger than the stratum-specific design effects because of the wide variation in analysis weights.

Introduction to Writing a SUDAAN Program

To perform an analysis using the SUDAAN software, you must write a program in the SUDAAN language. Each program has two major components. One component describes the PRAMS survey to SUDAAN so the software knows which formulas to use when calculating point estimates, standard errors, and statistical tests of significance. The second component requests a particular statistical analysis, e.g., a crosstab of two categorical variables (with PROC CROSSTAB) or a comparison of the mean of a continuous variable across several strata (with PROC DESCRIPT). The program you write then is submitted to SUDAAN, as you would submit your SAS program to the SAS software package. The following statement typed at the keyboard directs SUDAAN to look for the program in the file work1.pgm and to send the output from the analysis to the file workl.out.

```
SUDPROC WORK1.PGM  WORK1.OUT
```

The output also can be directed to the terminal.

Part 6CWriting a SUDAAN Program

Describing PRAMS to SUDAAN

This section describes statements that are needed in all SUDAAN programs. These statements give SUDAAN a basic description of the PRAMS survey design. The following discussion is based on this example:

```
1  PROC xxxxxxx DESIGN=STRWOR FILETYPE=SAS DATA=WVFOUR;
2  NEST STRATUM;
3  TOTCNT TOTCNT;
4  SAMCNT SAMCNT;
5  WEIGHT WTANAL;
```

On the first line (the PROC statement), DESIGN=STRWOR indicates stratified random sampling without replacement. (States that used stratified systematic sampling may want to note that, for these computational purposes, stratified systematic sampling is equivalent to stratified random sampling). The PROC statement also specifies that the data input for SUDAAN is a SAS data set (FILETYPE=SAS) and gives the name of the SAS data set, which in this example is WVFOUR (DATA=WVFOUR). Only the first part of the two-part data set name is required on the PROC statement; the second part of the name is SSD. The first part of the statement, PROC xxxxxx, depends on the analysis that you want to do. For analysis of a categorical variable, you would enter PROC CROSSTAB (or PROC DESCRIPT). For analysis of a metric variable, you would enter PROC DESCRIPT.

On lines 2 through 5, the variables STRATUM, TOTCNT, SAMCNT, and WTANAL are used. These variables are created in SAS and are included in the input SAS data set. Every mother in the data set must have a value for each of these variables.

The NEST statement (line 2) tells SUDAAN the name of the variable that indicates the stratum in which each respondent mother is. In this example, this variable is named STRATUM. The input SAS data set to SUDAAN must be sorted by all variables on the NEST statement, in this instance just the one variable STRATUM. If you fail to sort your data set, SUDAAN may run anyway and yield incorrect results.

On lines 3 and 4, the TOTCNT and SAMCNT statements tell SUDAAN to take into account the sampling fractions in each stratum when it calculates estimated standard errors. (Recall that SUDAAN already takes the sampling fraction into account in calculating point estimates because the sampling fraction is a component of the analysis weight.) It is advantageous to take into account the sampling fractions for standard error calculation whenever some stratum-specific sampling fractions are fairly high, e.g.,above 5% or 10%. Since this condition is satisfied in the PRAMS surveysCin particular, for

mothers of low birthweight infants who are sampled at very high ratesCwe recommend that SUDAAN be used with the TOTCNT and SAMCNT statements.
The TOTCNT statement gives the name of the variable that indicates the number of women on the frame for a given stratum. The SAMCNT statement indicates the name of the variable that gives the number of respondent women in a given stratum. For PRAMS analyses, we use the variable names TOTCNT and SAMCNT, although we did not need to use the same names as the statement name.

In our example, for every mother in stratum 1, the value of the TOTCNT variable is 678 and the value of the SAMCNT variable is 344. For every mother in stratum 4 the value for the TOTCNT and SAMCNT variables are 11342 and 522, respectively. SUDAAN uses the ratio of the two variables SAMCNT to TOTCNT for each stratum in its calculations of standard errors. Since the ratio of SAMCNT to TOTCNT is about 1/3 to 1/2 in the low birthweight strata (strata one and two) in our example data set, you generally will find smaller standard errors and smaller design effects in these two strata compared to the normal birthweight strata (strata three and four).

On line 5, the WEIGHT statement tells SUDAAN the variable name in the data set for the analysis weight. In the example data set this statement is WEIGHT WTANAL;.

When you want to specify a particular analysis, the PROC statement is modified and additional statements are added to the SUDAAN program.

Introduction to PROC CROSSTAB

PROC CROSSTAB is used whenever all variables in the desired analysis are categorical (or analyzed as categorical). PROC CROSSTAB is similar to PROC FREQ in SAS except, of course, that SUDAAN calculates standard errors according to the description of the survey design. PROC CROSSTAB will produce contingency tables, or cross tabulations, with row percentages, column percentages, and total percentages. (You can suppress some of this printout if you want to). PROC CROSSTAB also will do a chi square test to see whether the categorical variables in the table are independent of each other. Again, the chi square test is calculated by taking into account the survey design.

EXAMPLE1.SUD: Crosstabulation of Stratum and Prepregnancy Smoking

Consider the variable "Did you smoke during the three months prior to pregnancy?", i.e., CIG3BEFG with 1=no and 2=yes and don't know and no answer coded as dot ".". Suppose we want to generate the distribution of this variable for each of the four strata and get standard errors for the point estimates. The following SUDAAN program, EXAMPLE1.SUD, will accomplish this objective.

```
1    PROC CROSSTAB FILETYPE=SAS DESIGN=STRWOR
                                                    DATA=WVFC
```

```
2   NEST STRATUM;
3   TOTCNT TOTCNT;
4   SAMCNT SAMCNT;
5   WEIGHT WTANAL;
6   SUBGROUP STRATUM CIG3BEFG;
7   LEVELS 4 2;
8   TABLES STRATUM*CIG3BEFG;
9   TEST CHISQ;
10  PRINT NSUM WSUM SEWGT ROWPER SEROW DEFFROW ATLEV1
    CHISQ CHISQP CHISQDF;
```

On line 1, the PROC statement specifies the procedure CROSSTAB. DEFT (or DEFF) on the PROC statement requests SUDAAN to calculate the design effect for all point estimates. ATLEVEL1=1 on the PROC statement requests SUDAAN to count the number of strata that are summarized for any point estimate; you can use this information to check your output.

Lines 2 through 5 are the same as described in the previous section.

On line 6, the SUBGROUP statement identifies the categorical variables to be used in the analyses and the LEVELS statement (line 7) indicates the number of levels of each categorical variable. In this example, STRATUM is a categorical variable at four levels and CIG3BEFG is a categorical variable at two levels. (Note that CIG3BEFG has a third level, don't know/no answer (DK) coded as "." SUDAAN does not include "." or missing values in analyses. If you are interested in analyzing the missing values of a categorical variable, which is generally recommended, you must create a category for missing in SAS. Then on line 7, the variable CIG3B3FG would have 3 levels.)

On line 8, the TABLES statement requests a contingency table or crosstabulation of STRATUM by CIG3BEFG. Because STRATUM appears first in the statement, STRATUM will comprise the rows of the crosstabulation and CIG3BEFG will comprise the columns.

On line 9, the TEST statement requests a chi square test of the null hypothesis of no association between stratum and smoking during the three months prior to pregnancy. On line 10, the PRINT statement specifies the information to be included in the output. Because we are interested in an analysis by stratum, only row percents (and not column percents and not total percents) are requested. If the PRINT statement is missing, the default option for output prints a lot of material, much of which is not useful. The PRINT statement in this example requests the number of respondent mothers in each cell of the contingency table (NSUM); the sum of the analysis weights of the women in each cell (WSUM); the estimated standard error of WSUM; the row percentage (ROWPER), e.g.,

the percentage of mothers in a given stratum who did or did not smoke in the three months prior to pregnancy; the estimated standard error of the row percentage (SEROW); the design effect for the row percentage (DEFFROW); the number of strata added over to do the given calculation (ATLEV1); the calculated chi square value for the test of no association between the two categorical variables (CHISQ); the p-value of the calculated chi square value (CHISQP); and the degrees of freedom for the chi square statistic (CHISQDF).

The SUDAAN program in this example puts out a 5 X 3 contingency table, with row 1 being the entire sample (i.e., adding over all strata) and rows 2 through 5 being strata 1 through 4. Column 1 is the total sample and columns 2 and 3 are no and yes, respectively, for smoking during the three months before pregnancy. You can delete the row and column marginals, i.e., the row and column totals, from the calculations and the output by using the option NOMARG on the PROC statement. It is included here, though, because it gives statewide estimates, which are of interest.

An Overview of the Output From EXAMPLE1.SUD

As a first step when reviewing your output, it is always useful to check the sample size on the printout in case you have made some programming errors or your data set has unexpectedly changed. The SUDAAN output includes the number of observations read in the data set; you should know what this number is and hence can check to see if SUDAAN read the entire data set. The sample size in the contingency table is the number of subjects included in the analysis. This number should equal the total number of respondents minus those who did not give an informative answer to the question being analyzed.

A condensation of the SUDAAN output EXAMPLE1.OUT is below.

TABLE A

Percentage of Mothers Who Smoked During 3 Months Before Pregnancy, by Stratum (Birthweight and Adequacy of Prenatal Care)

STRATUM	PERCENT WHO SMOKED	STANDARD ERROR	DESIGN EFFECT
1-TOTAL	41.4%	1.5%	1.50
2-INADQ/LBW	58.5%	2.1%	0.10
3-ADQ/LBW	50.8%	2.0%	0.07

4-INADQ/NBW	**48.8%**	**2.4%**	**1.39**
5-ADQ/NBW	**35.0%**	**2.1%**	**2.09**

Line 2 of Table A states that, among state residents with inadequate prenatal care and who delivered a low birthweight baby during 1988 in the state, 58.5% smoked during the 3 months before pregnancy. The estimated standard error of this point estimate 58.5% is 2.1%. A 95% confidence interval on the percentage who smoked during the three months prior to pregnancy for the population of inference (i.e., stratum 1) is 58.5% +/- 1.96 (2.1%) or (54.4%, 62.6%). The multiplier 1.96 is from the standard normal distribution and is used for a two-sided 95% confidence interval.

There seems to be an association between stratum and smoking during three months before pregnancy, since the four percentages for "yes" in the table seem different from each other. The chi square statistic to test this association was calculated as 64.35 with 3 degrees of freedom and a p-value of .00,
i.e., < .01. Hence, there is a statistically significant association between stratum and smoking during the three months prior to pregnancy. The chi square test with 3 degrees of freedom tells you that the four population percentages, for which the table contains the point estimates, are not all equal to each other. However, this chi square test does not indicate which strata are actually significantly different from other strata. This can be done via pairwise comparisons between strata using PROC DESCRIPT, to be illustrated later.

In the meantime, we can "eyeball" the percentages to see which strata seem to differ. Women in stratum 1 (inadequate prenatal care and low birthweight) seem most likely to have smoked, women in strata 2 and 3 are less likely to have smoked and probably not different from each other, and women in stratum 4 (adequate prenatal care and normal birthweight) are least likely to have smoked.

Because the four strata are composed of two categorical variables, each at two levels, it is possible to draw some "eyeball" conclusions about adequacy of prenatal care and child's birthweight. Comparing the two low birthweight strata (1 and 2) indicates that mothers with inadequate prenatal care were more likely to have smoked; the same conclusion is reached by comparing the two normal birthweight strata (3 and 4). Hence, inadequate prenatal care seems to be associated with a higher prevalence of prepregnancy smoking. Comparing strata 1 and 3 (both inadequate prenatal care) indicates that mothers in the low birthweight stratum were more likely to have smoked; the same conclusion is drawn from comparing strata 2 and 4. Thus, low birthweight babies seem to be associated with a higher prevalence of prepregnancy smoking. Again, these "eyeball" analyses can be made more rigorous and statistical with other techniques to be illustrated later.

Using PROC DESCRIPT for Analysis of Categorical Variables

PROC DESCRIPT also can be used to produce a percentage distribution for categorical variables. You need to specify which level(s) of the categorical variable you want to summarize. In contrast to PROC DESCRIPT, PROC CROSSTAB summarizes ALL levels of each categorical variable (up to the highest level on the LEVELS statement) by default. Hence, PROC DESCRIPT can be used to limit the output if you only wish to estimate the percent who said yes to a yes/no question, since you know that the percent who said no will be 100% minus the percent who said yes.

PROC DESCRIPT has two useful options, PAIRWISE and CONTRAST, which allow the comparison of strata or domains to each other on a categorical variable. This may give more specific information than the overall chi square test in PROC CROSSTAB.

Using PROC DESCRIPT to Analyze Metric Variables

PROC DESCRIPT will produce means and estimated standard errors of the means for metric variables. Point estimates of quantiles (e.g., median, 40th percentile) and corresponding estimated standard errors are also available. As discussed above, the PAIRWISE and the CONTRAST statements permit the comparison of strata or domains with each other on either a metric or categorical variable. Many other features are available in DESCRIPT, such as rate standardization, orthogonal polynomials and poststratification.

Part 7CHow to Handle Missing Data (Item Nonresponse)

Two Types of Missing Data: Unit Nonresponse and Item Nonresponse

Unit nonresponse **occurs when a mother who was included in the sample does not participate (i.e., does not complete a questionnaire). Because she did not participate, no questionnaire data are available for her. Adjustment for unit nonresponse forms one component of the analysis weight (see Part 1 and Appendix 1). In contrast to unit nonresponse, *item nonresponse* occurs when a mother who returned a questionnaire did not answer one or more of the items (i.e., questions). For example, on the PRAMS questionnaire, a mother may have left an item blank, either by error or on purpose. For analysis purposes, this results in the absence of data for that item, which we refer to as item nonresponse.**

In analyzing PRAMS questionnaire data, you must distinguish failure to answer an item from either a legitimate skip of that item or a "don't know" response or refusal. A legitimate skip (and hence a blank answer) occurs when the item was not applicable. For example, Q15 (satisfaction with prenatal care) is not applicable for women who did not get prenatal care. In contrast, a "don't know" response often occurs because the respondent did not know the requested information.

The issue of how to treat legitimate skips, "don't know" responses, and blanks that represent incorrect skips will need to be decided on an item-by-item basis, taking into consideration the aim of the particular analysis at hand. For example, for the PRAMS questions relating to content of prenatal care (e.g., did a doctor/nurse/health worker talk with you about how smoking/drinking could affect your baby...), we can infer that the answer for respondents who said that they did not have prenatal care was "no." For some items, "don't know" was offered as a response option. Depending on the goal of a particular analysis, you may decide to retain "don't know" as a response category, treat it as a "missing" answer, or consider it equivalent to "yes" or "no." Items that were not answered when they should have been answered are truly missing. The two options for handling missing data are described next.

Options for Dealing With Item Nonresponse

The first option is to omit the person with the missing data from the analysis; the second option is to impute a value for the missing answer. Neither approach is entirely satisfactory. However, for PRAMS, for the analysis of a particular item, we favor the approach of omitting persons with missing data.

1. Omitting Persons With Item Nonresponse

In practice, this means that women who did not answer a particular item will not be included in the analysis of that item. This approach has the advantage of avoiding the imputation process (described below). It rests on the assumption that, for a particular item, the average of the answers of the mothers who responded to it is similar to the hypothetical average of the answers of the mothers who did *not* respond to it. While this assumption seems reasonable when the item nonresponse rate is low (i.e., when most mothers answered the item), it seems less justified when fewer than 75% of the mothers who should have answered the item actually did so. For this reason, when interpreting the results for a particular item, we recommend always considering the magnitude of the item nonresponse. As a rule of thumb, if item nonresponse is greater than 25%, the results may not accurately represent the population. Another consideration is that as the number of women for whom an answer is available decreases, the effective sample size decreases, leading to an undesirable increase in the variance of the statistics.

a. Calculating the Item Nonresponse Rate

Suppose that the answers to a question have the following frequencies:

Answer	Frequency
yes	**a**
no	**b**
DK	**c**
Not Applicable	**d**
Missing	**e**

If DK (don't know) is counted as an acceptable answer, the item nonresponse rate is:

$$\frac{e}{a + b + c + e}$$

If DK is counted as missing, the item nonresponse rate is:

$$\frac{c + e}{a + b + c + e}$$

As discussed above, the way that you decide to count "don't know" answers will depend on the goals of the analysis you are conducting.

b. Multivariate Analysis

An analysis that uses several variables (e.g., a "multivariate" analysis) can cause special problems if the patterns of item nonresponse differ among the variables. For example, consider an analysis examining the relationship between the number of nights the newborn spent in the hospital (Q35) and the following variables: maternal participation in WIC (Q22), maternal prepregnancy cigarette smoking (Q28), and wantedness/timing of the pregnancy (Q7). This analysis includes four items: newborn days; WIC participation; maternal cigarette smoking; and wantedness of pregnancy. To determine the impact of nonresponse in this analysis, you need to compute the percentage of mothers for whom answers are missing (blank) for one *or more* of these four items. This is because the multivariate analysis will exclude any mother for whom the answer to one or more items is missing. If the same 10% of mothers did not respond to each of the four items, the item nonresponse for the multivariate analysis would be 10%Can acceptable level. In contrast, consider the worst case, where for each of the four items, a different 10% of mothers had a missing answer. The result would be 40% item nonresponse (4 x 10%) for the multivariate analysisCa level that would call the validity of the results into question as well as substantially diminish the effective sample size.

We know that the rate of item nonresponse varies among questions. A practical result of omitting subjects who did not respond to a particular item is that the number of subjects available for analysis will vary among items. Analysis of items with very low rates of item nonresponse (Ql, for example) will be based on a larger number of subjects than analyses of items with higher rates of item nonresponse.

2. Imputing a Value for Missing Answers

Another approach for handling item nonresponse is to estimate, or "impute," an answer for the item. Although this can be done in a number of ways, each method involves identifying another respondent (or set of respondents) who is (are) "similar" to the mother with no answer to the item and assigning the missing mother the same answer as the other respondent(s). This approach assumes that these similar mothers would respond in the same way. The task of imputing responses for each item for each mother who did not answer it can require a large effort.

Imputation has the advantage of retaining all of the respondents to the survey in the analysis. However, many researchers prefer not to use it. This is partly because of concerns about the validity of the assumption on which imputation is based, but also because the methods for identifying similar respondents can be statistically complex and time-consuming.

Part 8CSmall Area Analyses or Subpopulation Analyses

The PRAMS surveys are designed to focus on statewide estimates, with additional interest in the subpopulations defined by stratification variables such as birth weight of infant and/or demographic characteristics of the mother. Some examples of subpopulations follow:

1. **One or more specific strata (e.g., mothers with inadequate prenatal care and a low birth weight infant).**

2. **A domain in the population such as white mothers who smoked during the three months prior to pregnancy.**

3. **A domain in the population such as women who deliver at a particular hospital or a certain group of hospitals.**

4. **A geographic area smaller than the state (e.g., mothers residing in the western counties of a state or in one particular county).**

Note that in cases 2, 3, and 4, members of the subpopulation may be found in all or most or only some of the strata, whereas in case 1 above members of the subpopulation are not found in all the strata.

There are two important issues to consider when analyses are desired for subpopulations. First, the sample size needs to be adequate for analysis of subpopulations. Guidelines on sample size requirements are given below. Second, estimates for subpopulations may be biased because of the methods used to adjust for survey nonresponse. Circumstances that may lead to biased estimates are discussed below.

Sample Size Considerations for Subpopulations

The issue of sample size is whether the sample size is large enough to use legitimately the computational techniques in SUDAAN when interest is in a subpopulation rather than the statewide population. Under the assumption of stratified random sampling, which is the design for all PRAMS surveys, the answer to this question is not difficult. The minimum number of mothers needed in any subpopulation analysis is 30 respondents plus the number of strata in the survey. Thus, if there are four strata in the survey design, then a minimum of 34 respondents in a subpopulation is needed in order to make statistical inference to that subpopulation.

This minimum required sample size will almost certainly be satisfied for a stratum-specific analysis. Hence, there should be no problem with respect to sample size in conducting a

stratum-specific analysis. Of course, standard errors for point estimates within a specific stratum will be much larger than the standard errors for these same point estimates on a statewide basis, just because the sample size in a specific stratum is smaller than the statewide sample size. Thus, confidence intervals on population parameters for a particular stratum will be much wider than confidence intervals on those same parameters for the statewide population.

The minimum sample size requirement is also likely to be satisfied for particular domains of interest, e.g., mothers who smoked three months prior to pregnancy, black mothers who smoked during pregnancy or white mothers with a low family income. Recall that the word domain refers to a subpopulation that is not defined only by stratification variables. Hence, the members of the domain will be found in all or some of the strata.

The minimum sample size requirement is also likely to be satisfied for moderately sized geographic areas of interest, e.g., the southern counties of a state or the rural or metropolitan area of a state. The minimum sample size may also be available to analyze data on a county specific basis in counties that have a moderately large percentage of the state's population. Even smaller counties may have the required minimum sample size if the data cover a longer time period such as one or two years.

The minimum sample size requirement may not be present if one wishes to define the subpopulation as mothers from geographic areas with a smaller population, e.g., a very small county or a particular catchment area of a public health clinic. Another case where the minimum sample size may not be present is when the subpopulation is a particular minority group which represents a small percentage of the statewide population.

Given the structure of the PRAMS surveys and their sample sizes, the minimum required sample sizes should be attained for most subpopulations of interest. Recall, however, that even if the minimum required sample size is present, the standard errors for these analyses will be much larger than for statewide point estimates. Once the variability in a point estimate for a subpopulation is considered, e.g., with a 95% confidence interval on the population parameter, the confidence interval may be so wide as to be not useful for health planning.

If it is desired to make inference to a subpopulation where the sample size does not meet the required minimum size, then other procedures may be used. These techniques have been developed recently in the research area of "small area analyses." They are not discussed in these guidelines and are not implemented directly in SUDAAN.

Potentially Biased Estimators

Because nonresponse adjustments are made for the analysis weights on a statewide basis, although separately for each stratum, there are some particular circumstances under which subpopulation estimates may be biased. These circumstances probably are unlikely to occur, but they are mentioned here so that precautions can be taken.

The analysis weights are adjusted in an attempt to compensate for nonresponse to the questionnaire. Within each stratum, nonresponse adjustment cells are defined based on variables related to response rate, e.g., education, age and marital status of mother. Within a stratum and within an adjustment cell formed by (for example) the crossclassification of education, age and marital status, it is assumed that, on the average, the nonrespondents are similar to the respondents. Note that the adjustment cell within a stratum (e.g., mothers with a low birth weight infant) can contain mothers from throughout the state. In most cases the nonresponse adjustment is a statewide adjustment, which is appropriate when the intent of the analyses is statewide inference.

When it is desired, however, to analyze data on a subpopulation, e.g., mothers who are resident in only part of the state, the statewide nonresponse adjustments to the weights may not describe the specific subpopulation appropriately. This would be true if the pattern of nonresponse in the subpopulation differs substantially from the statewide pattern of nonresponse. In this situation, the estimates for the subpopulation may be biased.

A method for preventing biased estimates when the subpopulations are geographic areas of the state is to make nonresponse adjustments within each geographic area of interest. This would require additional resources devoted to the development of analysis weights for the survey, utilizing information from the state birth certificates. One state recalculated nonresponse adjustments based on geographic areas of the state, resulting in new analysis weights for the statewide survey. Analysis of the data for geographic subpopulations using both sets of analysis weights showed that the two sets of point estimates and

estimated standard errors were similar. Hence, the state decided that reweighting the data set was not necessary for the analysis of geographic subpopulations.

In most instances it will be appropriate to analyze subpopulations without concern for potentially biased estimators. Concern should be raised when the pattern of nonresponse in the subpopulations differs substantially from the pattern of nonresponse in the statewide population.

Part 9CModeling Procedures in SUDAAN

These guidelines discuss PRAMS analyses which utilize descriptive statistics and which compare subpopulations on variables of interest, resulting in an emphasis on PROC DESCRIPT and PROC CROSSTAB. PROC RATIO is also a descriptive procedure but most likely would be rarely used in analysis of PRAMS surveys.

SUDAAN also contains several modeling procedures, including PROCS REGRESS, LOGISTIC, SURVIVAL and CATAN. These procedures can be used to analyze sample survey data using the following well known statistical procedures:

1. **Linear regression, analysis of variance and analysis of covariance (i.e., linear models) for a continuous dependent variable and independent variables which can be categorical or metric (REGRESS).**

2. **Logistic regression for a dependent variable which is dichotomous or ordinal and independent variables which can be categorical or metric (LOGISTIC).**

3. **Discrete proportional hazards model or Cox's proportional hazards model for survival analysis (SURVIVAL).**

4. **Linear or loglinear models for contingency table analyses on categorical variables (CATAN).**

When implementing these modeling procedures with sample survey data, there are different opinions about whether the survey design and/or the analysis weight needs to be incorporated into the analyses. Sample survey experts who are "design based" generally would include the survey design and the analysis weight, whereas sample survey experts who are "model based" generally would not include the survey design and the analysis weight. There are additional options which involve a combination of the two approaches. This issue cannot be discussed thoroughly in these guidelines.

SUDAAN has the capability of performing the four modeling procedures mentioned above with the "design based" approach, i.e., taking into account the survey design and the analysis weight. SUDAAN also has the capability of performing these four modeling procedures using the "model based" approach by using the option SRS (simple random sampling) as the method of selecting

the sample. Of course, standard software packages such as SAS, BMDP, SPSS-X, SYSTAT and EPI-INFO also can be used for the model-based approach to analysis of sample survey data.
The purpose of this section of the guidelines is not to inform you how to choose between the

"design based" and "model based" approaches and how to implement them. Rather, the purpose is to alert you to the different approaches which can be taken and that there are differences of opinion about which approach is "best."