H·CUP

HEALTHCARE COST AND UTILIZATION PROJECT

HCUP Methods Series

Calculating Kids' Inpatient Database (KID) Variances

Report # 2005-07

AHRQ
Agency for Healthcare Research and Quality

U.S. Department of Health and Human Services
Agency for Healthcare Research and Quality

Contact Information:
Healthcare Cost and Utilization Project (HCUP)
Agency for Healthcare Research and Quality
540 Gaither Road
Rockville, MD 20850
http://www.hcup-us.ahrq.gov

For Technical Assistance with HCUP Products:

Email: hcup@ahrq.gov

or

Phone: 1-866-290-HCUP

Recommended Citation: Houchens R, Elixhauser A. *Calculating Kids' Inpatient Database (KID) Variances*. 2005. HCUP Methods Series Report # 2005-07 ONLINE. December 16, 2005. U.S. Agency for Healthcare Research and Quality. Available: http://www.hcup-us.ahrq.gov/reports/methods/methods.jsp.

TABLE OF CONTENTS

INDEX OF TABLES

EXECUTIVE SUMMARY

The Healthcare Cost and Utilization Project (HCUP) is a Federal-State-Industry partnership to build a standardized, multi-State health data system. The Kids' Inpatient Database (KID) is a nationwide sample of pediatric discharges from HCUP State Inpatient Databases (SID) community, non-rehabilitation hospitals weighted to all pediatric discharges in the target universe. The target universe includes all pediatric discharges from community hospitals in the United States that were open during any part of the calendar year. Beginning with the 2000 KID, short-term rehabilitation hospitals were excluded from the universe because the type of care provided and the characteristics of the discharges from these facilities were markedly different from other short-term hospitals.

This document describes how to calculate simple statistics, including variances, from the KID, taking into account the sampling design and sample discharge weights. Data from the 2003 KID are used in all examples in this report. The report contains the program code required to calculate sample totals, means, rates, and their variances with three commonly used statistical programming languages that run on personal computers: SAS, Stata, and SUDAAN. This report also provides results of example calculations from all three statistical packages using the 2003 KID, and it demonstrates that the results are virtually identical for all three statistical packages. Two approaches to calculating variances for subpopulations are suggested. The first, described in the body of the report, uses the entire KID sample. The second, described in Appendix B, can be used when computing constraints prevent use of the entire KID.

Finally, we discuss alternative concepts of populations and other methods that could be applied to calculate variances. The KID hospital file contains the population count of hospitals in each stratum. These counts can be used to obtain finite-population estimates of variances, although the finite population correction (fpc) factor is not appropriate for most research applications, because results are usually generalized beyond the specific population of discharges and hospitals that existed in 2003.

INTRODUCTION

The Kids' Inpatient Database (KID) is one of a family of databases and software tools developed as part of the Healthcare Cost and Utilization Project (HCUP), a Federal-State-Industry partnership sponsored by the Agency for Healthcare Research and Quality (AHRQ).

The KID is a unique and powerful database that contains hospital inpatient stays for children. The KID development team designed the database to permit researchers to study a broad range of conditions and procedures related to hospitalizations of children. Researchers and policy makers can use the KID to identify, track, and analyze national trends in hospital utilization, access, charges, quality, and outcomes for children, and it is especially suited for studying relatively rare conditions and procedures.

This document describes how to calculate statistics, including variances, from the KID, taking into account the sampling design and sample discharge weights. Data from the 2003 KID are used in all examples in this report. The report also provides the program code required to calculate sample totals, means, rates, and their variances with three commonly used statistical programming languages that run on personal computers:

1) SAS Version 9.1,
2) Stata SE Version 9.1, and
3) SUDAAN Release 9.01 (SAS-callable standalone version).

.
All three languages have procedures for calculating sample statistics and appropriate variances based on data from complex sampling designs. This is important, because unweighted statistics and weighted analyses that fail to account for the KID sample design could yield biased estimates. Although this report does not cover multivariate statistical procedures such as regression analysis, some concepts introduced in this report carry over to those areas of analysis, as well.

Several statistical programming packages allow weighted analyses. If the user prefers to use a statistical package other than these three, it is likely that the options and statements for that package will be similar to those for one of the three packages covered by this report. For an excellent review of such programs, visit the following web site (users are encouraged to check this site for updated information): http://www.fas.harvard.edu/~stats/survey-soft/survey-soft.html. Appendix A contains a summary of survey analysis capabilities for SAS, Stata, and SUDAAN copied from the web site.

This report also gives the results of example calculations from all three statistical packages. Therefore, the user can run the program code in this document and check the results obtained against the results reported here.

The introduction continues below with a brief overview of the KID sample design and the discharge weights that accompany the KID database.

KID Sample Design

The KID sampling frame included all pediatric discharges from community, non-rehabilitation hospitals in the HCUP State Inpatient Databases (SID) that could be matched to the corresponding AHA survey data (subject to state-specific restrictions). Beginning with the 2000 KID, pediatric discharges were defined as having an age at admission of 20 years or less. This range of ages is a modification compared to the 1997 KID, which included only discharges with an admission age of 18 years or younger. In addition, discharges with missing, invalid, or inconsistent ages were excluded.

The KID includes a sample of pediatric discharges from all hospitals in the sampling frame. For the sample, the pediatric discharges were stratified by uncomplicated in-hospital birth, complicated in-hospital birth, and pediatric non-birth. To further ensure an accurate representation of each hospital's pediatric case-mix, we also sorted the discharges by state, hospital, DRG, and a random number within each DRG. We then used systematic random sampling to select 10 percent of uncomplicated in-hospital births and 80 percent of other pediatric cases from each frame hospital. For more design details including definitions of all stratifiers and hospital characteristics, see the report, *Design of the HCUP Kids' Inpatient Database, 2003.* This report is available on the KID Documentation CD-ROM and on the HCUP User Support Website at www.hcup-us.ahrq.gov.

KID Sample Weights

For the discharge sample weights, we post-stratified hospitals on six characteristics contained in the 2003 American Hospital Association (AHA) Annual Survey Database files. Hospitals were stratified by region (Northeast, Midwest, South, and West), location/teaching status (rural, urban teaching, and urban non-teaching), bed size category (small, medium, and large), ownership (government nonfederal [public], private not-for-profit [voluntary], and private investor-owned [proprietary]), and whether the hospital was a freestanding children's hospital.

Location is defined by the AHA's designation of urban or rural. The teaching status of hospitals identified as children's hospitals by the National Association of Children's Hospitals and Related Institutions (NACHRI) is based on an indicator provided by NACHRI. Other hospitals are considered to be teaching hospitals if they have an American Medical Association (AMA)-approved residency program, are members of the Council of Teaching Hospitals (COTH), or have a ratio of full-time equivalent interns and residents to beds of .25 or higher. Bed size categories are small, medium, and large, with separate size cut points defined for each combination of hospital region, teaching status, and urban/rural designation. Ownership breakdowns are based on the degree of observed ownership variation within each region across bed size categories. Some of the NIS strata definitions were revised for 1998 and subsequent data years, and both the 2000 and 2003 KID files used these revised strata. The discharge sample weights were calculated within each sampling stratum as the ratio of discharges in the universe to discharges in the sample. The number of discharges in the universe was estimated from the 2003 AHA Annual Survey Database.

Missing Values

The procedures presented in this report omit cases with missing values from all calculations. Values that are missing for any reason can compromise the quality of estimates. If the outcome for discharges with missing values is different from the outcome for discharges with valid values, then sample estimates for that outcome will be biased and will not accurately represent the discharge population. There are several techniques available to help overcome this bias. One strategy is to use imputation to replace missing values with acceptable values. Another strategy is to use sample weight adjustments to compensate for missing values[1]. Such data preparation and adjustments are outside the scope of this report; however, if necessary, they should be carried out before using the statistical procedures presented here.

On the other hand, if the cases with and without missing values are assumed to be similar with respect to their outcomes, then no adjustment may be necessary for estimates of means and rates because the means and rates based on nonmissing cases would be representative of the means and rates of missing cases.

It is possible that some adjustment may still be necessary for the estimates of totals. Totals (of non-negative variables) would tend to be underestimated in the presence of missing values of the variable for which the total is estimated because the cases with missing values would be omitted from the calculations. For example, in calculating aggregate charges (the sum of charges), the analyst could impute missing values of hospital charges by using the average charge for the DRG or by using the estimated value from a regression that predicts charges.

The next section establishes some sampling concepts in a short discussion of a formula that could be used to calculate the variance of a total from the KID sample. The following sections contain the program code required to estimate some sample statistics and their variances using each of the three statistical packages. We demonstrate that the results are identical or very similar for all of the programs. Finally, we discuss the finite population correction, alternative concepts of population, and other methods that could be applied to calculate variances.

RATIONALE AND FORMULAS FOR KID VARIANCE CALCULATIONS

For a simple random sample of discharges, the usual variance calculations are appropriate. For example, the unbiased estimate for the variance of hospital length of stay (LOS) based on a sample of n discharges would be calculated as:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

where x_i is the LOS for discharge i, and $\bar{x}$ is the mean LOS over the sample of n discharges. Consequently, the estimated standard error of the mean would be calculated as: $\hat{\sigma}/\sqrt{n}$.

However, the sample of KID discharges is not a simple random sample. Only discharges from hospitals in states that participate in HCUP and agree to contribute to the KID are available for inclusion in the KID. For the purpose of calculating variances, we consider hospitals inside the frame to be similar to hospitals outside the frame within each stratum. Consequently, although *we do not actually draw a sample of hospitals for the KID*, for estimation purposes we nevertheless treat the KID hospitals as though they were randomly selected at the first stage of sampling from the entire universe of hospitals within each stratum.

Standard formulas for stratified, two-stage cluster sample without replacement may be used to calculate statistics and their variances in most KID applications. The proportion of KID hospitals within each stratum varies according to the HCUP participation rate. We select a random

[1] See, for example, Foreman, E.K., *Survey Sampling Principles*, Dekker, New York, 1991, Chapter 10.

sample of hospital discharges from each of the KID hospitals. Consequently, the KID sample *resembles* a stratified two-stage cluster sample with hospitals as the clusters sampled in the first stage and with discharges as the elementary units sampled at a rate of 10 percent for uncomplicated newborns and at a rate of 80 percent for other pediatric discharges (age 20 years and under).

A complication, which we ignore in calculating variances below, is that the hospital sampling frame does not contain the entire universe of U.S. hospitals. The frame contains only hospitals in the 36 states for which all-payer discharge data were made available to the KID. To the extent that states in the frame differ from other states on outcomes within each stratum, this could lead to biased estimates. Consequently, KID users are encouraged to compare estimates from the KID to other benchmarks whenever they are available. For the 1997 KID, we compared a broad range of estimates from the KID to estimates from the National Hospital Discharge Survey (NHDS)[2]. Most estimates were consistent between the two data sources. The updated report for the current KID is expected to be available on the HCUP User Support Website (www.hcup-us.ahrq.gov) in late January, 2006.

The appropriate variance formula, for a stratified two-stage cluster sample, employs weights and components for the two stages of sampling. This is necessary to account for the possibility that sample discharges within hospitals may be more homogeneous in their outcomes than sample discharges between hospitals. If the analyst wants finite population estimates, then factors are also needed to correct for the proportion of the universe included in the sample at each level (finite population correction factors).

The following example is meant to illustrate the "behind the scenes" calculations that statistical programs make for variances based on sample designs. The reader may move on to the next section of this report without understanding the technical details of this example.

For this example, consider the estimated total of a variable Y, calculated as the weighted sum:

$$T = \sum_s \sum_h \sum_d w_{shd} \, y_{shd}$$

where:

y_{shd} = the observed value of variable Y for sample discharge *d* within sample hospital *h* within stratum *s*.

w_{shd} = a set of discharge weights or any other constants over the set of sample discharges, hospitals, and strata. An estimate of the variance of T from the sample is:

$$\hat{\sigma}_T^2 = \sum_s (1 - f_s) n_s V_s + \sum_s f_s \sum_h (1 - f_{sh}) n_{sh} V_{sh} \qquad (1)$$

where:

f_s = the proportion of the universe hospitals sampled in stratum *s*. If we wish to generalize results to a broader set of hospitals and patients outside that year's hospital population, then we would set f_s = 0. This might be desirable, for example, if the analyst wishes to draw inferences about a future year or wishes to use the results to set policy going forward.

[2] 2000 KID Comparison Report.

n_s = the number of hospitals within stratum s.

f_{sh} = the proportion of the discharges in the sample from sample hospital *h* within stratum *s*. For the KID, f_{sh} = .10 for uncomplicated newborns and f_{sh} = .80 for other discharges. However, an analyst may wish to consider this a sample from an infinite population (of possible patients) rather than a finite population, in which case f_{sh} = 0.

n_{sh} = the number of discharges in hospital h within stratum s.

V_s = the component of variance due to the first stage of sampling (variation among hospitals within stratum s):

$$V_s = \frac{\sum_h \left(\sum_d w_{shd}\, y_{shd} - \frac{\sum_h \sum_d w_{shd}\, y_{shd}}{n_s} \right)^2}{(n_s - 1)}$$

Notice that the numerator is the sum of squared deviations of the individual hospital totals from the mean hospital total, and the sum is over all hospitals in stratum s, similar to the familiar calculation for the variance of any sample statistic. Also notice in equation 1 that this term is multiplied by zero if f_s = 1. In that case, all hospitals within stratum s are sampled, and the estimated total for that stratum has no sampling error associated with it.

V_{sh} = the component of variance due to the second stage of sampling (variation among discharges within hospital h in stratum s):

$$V_{sh} = \frac{\sum_d \left(w_{shd}\, y_{shd} - \frac{\sum_d w_{shd}\, y_{shd}}{n_{sh}} \right)^2}{n_{sh} - 1}$$

Again, this calculation of a variance is familiar. The numerator is the sum of squared deviations of the individual weighted discharge totals from the mean weighted discharge total for each hospital h in stratum s. If the sampling rate f_{sh} = 1 for hospital h in stratum s, then this term is multiplied by zero in equation 1 because the hospital total is estimated without error.

Many statistical packages use variance formulas similar to (1) to estimate variances for simple statistics, such as means and totals.

It is important to recognize that these variance calculations assume that the analyst is interested in making inferences to the finite population of hospital discharges in the year of the data. As the sampling fraction *f* approaches 1, the sampling variance approaches zero. If the analyst is interested in making inferences to another population, not the specific discharges represented in the discharge population for the year of the data, then the sampling fraction *f* should be set to zero. *Our examples will not use the finite population correction (fpc) and, in general, most analyses will not use the fpc*. However, we will indicate the effect of the fpc and how the fpc could be incorporated.

EXAMPLES OF KID VARIANCE CALCULATIONS

The example analysis is for a subpopulation of the KID defined by a Clinical Classifications Software (CCS) diagnosis category code equal to 128*: Asthma.* CCS is a tool for clustering patient diagnoses and procedures into a manageable number of clinically meaningful categories developed at the Agency for Healthcare Research and Quality (*Clinical Classifications Software*. Fact Sheet. Agency for Healthcare Research and Quality, Rockville, MD. http://www.ahrq.gov/data/hcup/ccsfact.htm).

To obtain estimates, we created an indicator variable using SAS to identify the subset of discharges with asthma for the analysis (DXCCS1 = 128). For Stata, we also generated a data file in ASCII format. The SAS-callable version of SUDAAN can use the SAS file. However, like Stata, the standalone version of SUDAAN would also require the ASCII format.

The SAS, Stata, and SUDAAN program code for the analysis of the asthma subpopulation are shown below, along with examples of the output produced by each program.

In these examples, the following conventions apply:

- UPPERCASE WORDS denote KID variable names.
- Lowercase words denote keywords and options that are part of the programming language.
- *Italicized words* denote information to be supplied by the researcher.
- **Bold words** denote comments.

Note that the example programs shown here use the entire KID, which is the approach generally recommended for calculating standard errors and which will yield correct standard errors. If computing constraints force the use of a subset of the KID (such as a specific subpopulation defined by a condition or by patient characteristics), then refer to Appendix B for alternative methods.

SAS Programming Statements

```
/* Create analysis file */

libname IN “location of kid file” ;

data ASTHMA ;
  set IN.KID_2003_CORE;
  if DXCCS1 = 128 then ASTHMA = 1; else ASTHMA = 0;
  DISCHGS = 1 ;

/* Obtain estimates:  The following SAS code produces estimates of
the sums, the means, and the standard errors for the number of
discharges, the length of stay, and the total hospital charges */

proc surveymeans data=ASTHMA sum std mean stderr ;
  weight DISCWT ;
  class DIED ;
  cluster HOSPID ;
  strata KID_STRATUM ;
  var DISCHGS LOS DIED TOTCHG ;
  domain ASTHMA ;

/* Note: If finite population estimates of standard errors are   */
/* wanted, then the PROC SURVEYMEANS statement could include the */
/* “total= ” option, indicating a file containing the number     */
/* of population hospitals in each stratum, such as could be     */
/* constructed from the KID hospital file.                       */
```

- The PROC SURVEYMEANS statement invokes the SAS procedure.

- The DATA= option requests that the analysis be performed on the file specified. If this statement is omitted, SAS uses the most recently created dataset.

- The SUM option requests the sum for variables listed in the VAR statement. For example, the variable DISCHGS is set to equal 1 for every record, so its sum estimates the total number of discharges.

- The STD option requests the standard error of the sum.

- The MEAN and STDERR options request that the mean and its standard error be printed. The default statistics are the mean, its standard error, and 95% confidence limits.

- The WEIGHT statement weights each record by the value of the variable DISCWT.

- The CLASS statement identifies DIED as a categorical variable for which a ratio analysis is performed (ratio of sum of DIED to sum of DISCWT).

- The STRATUM statement specifies KID_STRATUM as the stratum identifier.

- The CLUSTER statement specifies HOSPID as the cluster identifier.

- The VAR statement requests the statistics for the variables DISCHGS, LOS, TOTCHG and DIED. If the VAR statement is omitted, statistics will be calculated for all of the variables in the dataset except for those listed in the WEIGHT, STRATUM or CLUSTER statement.

- The DOMAIN statement requests separate statistics for the subpopulations of asthmatics and non-asthmatics.

These commands produced the following output:

SAS Output (for Asthma = 1)

Data Summary	
Number of Strata	61
Number of Clusters	3438
Number of Observations	2984129
Sum of Weights	7409161.76

Class Level Information			
Class Variable	Label	Levels	Values
DIED	Died during hospitalization	2	0 1

Variable	Label	Mean	Std Error of Mean	Sum	Std Dev
dischgs		1.000000	0	173392	5585.186299
LOS	Length of stay (cleaned)	2.328225	0.023386	403695	14645
TOTCHG	Total charges (cleaned)	7650.602681	216.086445	1308540386	63393715
DIED	Died during hospitalization	0.000136	0.000039234	23.654124	6.825513

Stata Programming Statements

```
/* Using SAS, create an ASCII file for use by STATA */

libname IN “location of kid file” ;

data _null_ ;
  set  IN.KID_2003_CORE ;
  if DXCCS1 = 128 then ASTHMA = 1 ; else ASTHMA = 0 ;
  DISCHGS = 1 ;
  file “file to write” ;
  if LOS < 0 then LOS = . ;
  if DIED < 0 then DIED = . ;
  if TOTCHG < 0 then TOTCHG = . ;
  put KID_STRATUM 1-4 HOSPID 6-10 DIED 12 LOS 14-17
      DISCHGS 19 TOTCHG 21-27 ASTHMA 29 +1 DISCWT ;

/* Obtain STATA Version 9.1 estimates */
/* Note: Stata commands should be entered in lowercase text */

set memory 150000
infile KID_STRATUM HOSPID DIED LOS DISCHGS TOTCHG
       ASTHMA DISCWT using “DATASET NAME”
svyset HOSPID [pweight = DISCWT], strata (KID_STRATUM)
svy: total DISCHGS, subpop(ASTHMA)
svy: mean LOS TOTCHG, subpop(ASTHMA)
svy: ratio DIED DISCHGS, subpop(ASTHMA)
```

- The SET MEMORY command increases the memory allocated to Stata to 150 megabytes. The default memory allocation of 1 megabyte is too small for the KID and will need to be changed prior to any analyses. (This command works only for the Windows and Unix versions of Stata. Users of other versions should see the manual specific to their operating system).

- The INFILE command lists the variables to read in from the dataset created for this analysis.

- The SVYSET command identifies the primary sampling unit, weight variable, and the stratification variable. For Stata versions lower than 9.0, the command is: SVYSET [PWEIGHT = discwt], STRATA (kid_stratum) PSU (hospid).

- The SVY: TOTAL command requests the estimate of the total and standard error for the variable listed. For Stata versions lower than 9.0, the command is: svytotal DISCHGS, subpop(ASTHMA).

- The SVY: MEAN command requests the estimate of the mean and its standard error for the variables listed. Note that Stata uses listwise deletion for missing values. Consequently, Stata ignores observations missing *either* LOS *or* TOTCHG. To avoid this, use separate

SVY: Mean commands for LOS and TOTCHG. For Stata versions lower than 9.0, the command is: svymean LOS TOTCHG, subpop(ASTHMA).

- The SVY: RATIO command requests the ratio of the two variables listed, in this case, of those who died to total discharges. For Stata versions lower than 9.0, the command is: svyratio DIED DISCHGS, subpop(ASTHMA).

- The SUBPOP option requests statistics for the subpopulation of asthmatics.

The results are shown on the next page.

These commands produce the following output:

Stata Output (for asthma = 1)

```
Survey: Total estimation

Number of strata =      61          Number of obs    = 2984129
Number of PSUs   =    3438          Population size  = 7.4e+06
                                    Subpop. no. obs  =  102101
                                    Subpop. size     =  173392
                                    Design df        =    3377

--------------------------------------------------------------
             |             Linearized
             |      Total   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
     dischgs |   173391.6   5585.186      162440.9    184342.3

Survey: Mean estimation

Number of strata =      61          Number of obs    = 2916018
Number of PSUs   =    3408          Population size  = 7.2e+06
                                    Subpop. no. obs  =  100483
                                    Subpop. size     =  171038
                                    Design df        =    3347

--------------------------------------------------------------
             |             Linearized
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
         los |   2.335077   .0235326      2.288937    2.381217
      totchg |   7650.603   215.8839      7227.325     8073.88
--------------------------------------------------------------

Survey: Ratio estimation

Number of strata =      61          Number of obs    = 2982507
Number of PSUs   =    3436          Population size  = 7.4e+06
                                    Subpop. no. obs  =  102095
                                    Subpop. size     =  173382
                                    Design df        =    3375

     _ratio_1: died/dischgs

--------------------------------------------------------------
             |             Linearized
             |      Ratio   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
     ratio 1 |   .0001364   .0000392      .0000595    .0002134
```

SUDAAN Programming Statements

```
/* The following code produces the estimate and standard error for
total hospital discharges, mean length of stay, and mean total
charges using the SAS-callable version of SUDAAN */

/* Create analysis file */

libname IN “location of kid file” ;

data ASTHMA ;
  set IN.KID_2003_CORE;
  if DXCCS1 = 50 then ASTHMA = 1 ; else ASTHMA = 2 ;
  DISCHGS = 1 ;

proc sort data=ASTHMA ; by KID_STRATUM HOSPID ;

proc descript data=ASTHMA filetype=sas design=wr ;
  weight DISCWT ;
  nest KID_STRATUM HOSPID ;
  setenv colwidth = 24 decwidth = 5 ;
  var LOS DISCHGS TOTCHG ;
  subgroup ASTHMA ;
  levels 2 ;
  print total setotal mean semean ;

/* SUDAAN does not allow continuous and categorical variables  */
/* to be analyzed in a single step. The following procedure    */
/* calculates statistics for the categorical variable “died”. */

proc descript data=ASTHMA filetype=sas design=wr ;
  weight DISCWT ;
  setenv colwidth = 24 decwidth = 5 ;
  nest KID_STRATUM HOSPID ;
  var DIED ;
  catlevel 1 ; /* Specifies analysis for the category died = 1 */
  subgroup ASTHMA ;
  levels 2 ;
```

- The PROC DESCRIPT statement invokes the procedure.

- The DATA= option specifies the dataset name.

- The FILETYPE option specifies that this is a SAS file.

- The DESIGN= option identifies this as a sample With Replacement (WR). KID hospitals were not sampled with replacement. However, this specification is appropriate when the hospital “population” is considered very large or conceptually infinite.

- The WEIGHT statement identifies DISCWT as the weight variable.

- The SETENV statement increases the column width to allow the printing of numbers larger than the default width and increases the number of digits past the decimal to 5.

- The NEST statement identifies KID_STRATUM as the stratum variable and HOSPID as the primary sampling unit.

- The VAR statement lists the variables to be included in the analysis.

- The CATLEVEL statement specifies that the analysis is for the category of DIED = 1.

- The SUBGROUP statement requests statistics for the subpopulation of asthmatics (and non-asthmatics).

- The LEVELS statement tells SUDAAN that ASTHMA has two levels (1 and 2).

These statements produced the following output:

SUDAAN Output (for Asthma = 1, slightly reformatted for readability)

Variable		ASTHMA 1
Length of stay (cleaned)	Total	403694.62433
	SE Total	14644.98761
	Mean	2.32822
	SE Mean	0.02339
DISCHGS	Total	173391.61547
	SE Total	5585.18630
	Mean	1.00000
	SE Mean	0.00000
Total charges (cleaned)	Total	1308540385.62639
	SE Total	63393715.17573
	Mean	7650.60268
	SE Mean	216.08644

Variable		ASTMHA 1
Died during hospitalization: 1	Sample Size	102095.00000
	Weighted Size	173381.79587
	Total	23.65412
	Lower 95% Limit	10.27157
	Upper 95% Limit	37.03668
	Percent	0.01364
	SE Percent	0.00392
	Lower 95% Limit	0.00776
	Upper 95% Limit	0.02398

Comparison of Estimates

Table 1 displays the estimates from each of the three statistical programming packages using the program code described earlier.

Table 1. Comparison of SAS, Stata, and SUDAAN Results, Asthma Kids' Inpatient Database (KID), 2003

VARIABLE (Standard Error)	SAS	Stata	SUDAAN
Total Discharges	173,392 (5,585)	173,392 (5,585)	173,392 (5,585)
In-hospital Mortality %	0.0136 (0.0039)	0.0136 (0.0039)	0.0136 (0.0039)
Length of Stay	2.328 (.023)	2.335 (.024)	2.328 (.023)
Total Charges	$7,651 (216)	$7,651 (216)	$7.651 (216)

The estimates from SAS and SUDAAN are identical. However, estimates from Stata differ slightly for LOS. This difference stems from the way Stata treats missing values in the calculations. Stata uses "listwise" deletion, which drops an observation from the calculations if *any* of the analysis variables are missing. In contrast, SAS and SUDAAN employ "variable-wise" deletion, which drops an observation from the calculations only for the variable that is missing. Consequently, in Stata, statistics should be calculated for one variable at a time to avoid losing valid values for some of the variables. In the asthma example, the estimated means and standard errors calculated for LOS by Stata are identical to those calculated by SAS and SUDAAN, when statistics for LOS and total charges are calculated separately:

```
svy: mean LOS, subpop(ASTHMA)
svy: mean TOTCHG, subpop(ASTHMA).
```

Finite Population Corrections

The KID sample contains a subset of all hospitals nationwide. Therefore, analysts may want to "correct" variance estimates to account for the fact that sampling error is attributable only to the the hospital population excluded from the KID. The KID excludes only 1,398 of 4,836 hospitals nationally (about 29 percent). Hence, the finite population correction factor (fpc), the multiple of the infinite population variance, is *on average* equal to about 29 percent. This means that the standard errors reported in the above table would be multiplied by about 54 percent (the square-root of 29 percent). While this substantially decreases the estimated standard errors, the fpc should be applied only when inferences are being made to the specific population of patients actually hospitalized during the year of the data, in this case, 2003. Usually analysts prefer not to use the fpc because they are interested in the long-run results for hospitals. For example, interest centers on the true, long-run mortality rate for a hospital rather than on the mortality rate actually observed for the particular group of patients treated in 2003.

Nevertheless, for analysts who do want statistics for finite populations, the 2003 KID hospital file contains the count of population hospitals for each stratum (the variable N_HOSP_U). These counts can be employed by each of the statistical packages to produce finite sample variances. See the program's documentation for details.

Analyzing Subpopulations

For the KID, interest is sometimes limited to a subpopulation (domain, subset, or subgroup) of the sampled population. For example, interest might center on patients with a given medical condition like asthma or heart disease or on patients with certain characteristics like males under the age of 13. Eliminating individuals outside the subpopulation from the KID before variance estimation will yield correct means and totals, but it can yield incorrect standard errors.

In particular, incorrect standard errors could be produced if a hospital is eliminated from the KID in the process of excluding patients outside the subpopulation of interest. For example, the standard errors could be incorrect if the KID was subset to the subpopulation of patients treated for cystic fibrosis and some hospitals had no cystic fibrosis patients in the sample.

The standard errors from a subset will be correct if every sample hospital has at least one observation in the subset. For example, if every hospital treated at least one cystic fibrosis patient, then the standard errors produced by subsetting the data would be correct.

Standard errors will always be calculated appropriately if all of the KID observations are retained in the analysis and the subpopulations are defined by variables in the DOMAIN statement in SAS, by the SUBPOP option in Stata, or by the SUBGROUP statement in SUDAAN. For example, an indicator variable could be created equal to 1 for cystic fibrosis patients and equal to 0 for all other patients, which could then be used with the DOMAIN statement, with the SUBPOP option, or with the SUBGROUP statement. This was the method used to illustrate the asthma analysis in this report.

One potential difficulty analysts will face with this approach is the requirement to perform analyses on the entire sample. The KID contains about three million observations. Therefore, this approach will require more disk space and more CPU time for analyses of subpopulations compared to the subsetting approach. To address this difficulty, we offer an alternative approach. The strategy is to subset the KID to the subpopulation of interest and then augment that subset with an extra "dummy" observation for each hospital. For the 2003 KID, this adds 3,438 observations, one for each hospital in the KID. These additional observations induce the programs to use the correct formula for calculating standard errors. The program code for SAS, Stata, and SUDAAN is contained in Appendix B.

DISCUSSION

Alternative Population Concepts (deciding whether to apply the fpc factor)

Finite sample or finite population model. Occasionally, analysts require variance calculations based on finite-sample theory. According to finite-sample theory, the intent of the estimation process is to obtain estimates that are precise representations of the nationwide population of patients treated during a specific year. In the context of the KID, any estimates that attempt to accurately describe characteristics (such as expenditure and utilization patterns or hospital market factors) and interrelationships among characteristics of hospitals and discharges during a specific year should be governed by finite-sample theory.

Superpopulation, or stochastic model. Alternatively, in the study of hypothetical population outcomes not limited to a specific point in time, analysts may be less interested in specific characteristics from the finite population (and time period) from which the sample was drawn, than they are in hypothetical characteristics of a conceptual "superpopulation" from which any particular finite population in a given year might have been drawn. According to this superpopulation model, the nationwide population in a given year is only a snapshot in time of the possible interrelationships among hospital, market, and discharge characteristics. In a given year, all possible interactions between such characteristics may not have been observed, but analysts may wish to predict or simulate interrelationships that may occur in the future.

Under the *finite-population model*, the variances of estimates approach zero as the sampling fraction approaches one, since the population is defined at that point in time, and because the estimate is for a characteristic as it existed at the time of sampling. This is in contrast to the superpopulation model, which adopts a stochastic viewpoint rather than a deterministic viewpoint. That is, the nationwide population in a particular year is viewed as a random sample of some underlying superpopulation over time.

Estimation Techniques

Different methods are used for calculating variances under the two sample theories. Under the superpopulation (stochastic) model, procedures have been developed to draw inferences using weights from complex samples[3]. In this context, the survey weights are not used to weight the sampled cases to the universe, because the universe is conceptually infinite in size. Instead, these weights are used to produce unbiased estimates of parameters that govern the superpopulation.

In addition to the methods shown in this report, standard errors can be estimated by validation and cross-validation techniques. Given that a very large number of observations will be available for most analyses, it may be feasible to set aside a part of the data for validation purposes. Standard errors and confidence intervals can then be calculated from the validation data.

If the analytical file is too small to set aside a large validation sample, cross-validation techniques may be used. For example, tenfold cross-validation would split the data into ten equal-sized subsets. The estimation would take place in ten iterations. At each iteration, the outcome of interest is predicted for one-tenth of the observations by an estimate based on a model fit to the other nine-tenths of the observations. Unbiased estimates of error variance are then obtained by comparing the actual values to the predicted values obtained in this manner.

CONCLUSIONS

Similar estimates were obtained from three alternative software packages that were used to calculate variances using the KID. Sample program code was provided for each package. Variance estimates will be calculated correctly whenever the entire KID sample is used. However, when computing constraints require the use of a subset of the KID, methods described in Appendix B should be applied to ensure correct variance calculations. The KID

[3] Potthoff, R.F., M.A. Woodbury, and K.G. Manton, *"Equivalent Sample Size" and "Equivalent Degrees of Freedom" Refinements for Inference Using Survey Weights Under Superpopulation Models. Journal of the American Statistical Association*, Vol. 87, (1992), pp. 383-396.

hospital file contains the population count of hospitals in each stratum. These counts can be used to obtain finite-population estimates of variances, although the finite population correction (fpc) factor is not appropriate for most research applications.

APPENDIX A:
SUMMARY OF SURVEY ANALYSIS CAPABILITIES FOR SAS, STATA, AND SUDAAN[4]

Summary of Survey Software: SAS/STAT

Vendor.

SAS Institute Inc.

Types of Designs That Can Be Accommodated.

For the sample selection procedure, the sample design can be a complex multistage sample design that includes stratification, clustering, replication, and unequal probabilities of selection.

For survey data analysis procedures, the sample design can be a complex survey sample design with stratification, clustering, unequal weighting, and with or without replacement.

Types of Estimands and Statistical Analyses That Can Be Accommodated.

SAS/STAT Software provides the SURVEYSELECT procedure for sample selection and the SURVEYMEANS and SURVEYREG procedures for producing descriptive statistics and regression estimates, respectively. These three procedures are available in SAS versions 8 and higher. Beginning with SAS 9, SAS/STAT also includes the SURVEYFREQ procedure for computing crosstabulations and tests of association, and the SURVEYLOGISTIC procedure for performing logistic regression. The analysis procedures can accommodate complex survey designs that include stratification, clustering, and unequal weighting.

- **The SURVEYSELECT procedure** provides a variety of methods for selecting probability-based random samples. The procedure can select a simple random sample, or samples with design features such as stratification, clustering or multistage sampling, or unequal probabilities of selection. It can accommodate very large sampling frames. It can draw a replicated sampling, i.e., a sample composed of a set of replicates, each selected in the same way.
  PROC SURVEYSELECT accepts the sampling frame as a SAS data set. Control language specifies the selection methods, the desired sample size or sampling rate, and other parameters. The output data set contains the selected units, with selection probabilities and sampling weights.
- **The SURVEYMEANS procedure** estimates population totals, means, and ratios (SAS 8.2 and later), with estimates of their variances, confidence limits, and other descriptive statistics, under sample designs that may include stratification, clustering, and unequal weighting.
- **The SURVEYREG procedure** estimates regression coefficients by generalized least squares, using elementwise regression, assuming that the regression coefficients are the same across strata and PSUs.
- **The SURVEYLOGISTIC procedure** fits logistic regression models for discrete response survey data by maximum likelihood, incorporating the sample design into the analysis.
- **The SURVEYFREQ procedure** produces one-way to n-way frequency and crosstabulation tables from sample survey data. These tables include estimates of population totals, population proportions, and corresponding standard errors. Confidence limits, coefficients of variation, and design effects are also available, as are tests of independence (Wald test, Rao-Scott likelihood ratio test, Rao-Scott chi-square test).

[4] This information was copied from the following website http://www.fas.harvard.edu/~stats/survey-soft/survey-soft.html maintained by Harvard University and the Survey Research Methods Section of the American Statistical Association.

Restrictions on Number of Variables or Observations.

None.

Primary Methods Used for Variance Estimation.

Taylor expansion.

General Description of the "Feel" of the Software.

The interface is similar to other SAS procedures. Programs may be entered from command files or through a windowing system. The Explorer window is used to view and manage SAS files. The Program Editor is used to enter, edit, and submit SAS programs, and messages appear in the Log window. Output from SAS programs is viewed in the Output window and navigate and managed in the Results window.

Platforms on which the Software Can Be Run.

Version 9.1 of the SAS System is available on the following platforms:

- Aix (64-bit), Release 5.1+
- HP/UX (64-bit), Release 11i+
- HP/UX Itanium (64-bit), Release 11i+
- Linux for Intel (32-bit): Red Hat 8.0, Advanced Server 2.1, SuSE SLES 8
- Linux for Itanium (64-bit): Red Hat RHEL 3.0
- Open VMS Alpha (64-bit), Release 7.2+ (excluding Release 7.3)
- OS/390, Version 2, Release 10
- Solaris (64-bit), Version 8 or 9+
- Tru64 UNIX (64-bit), Version 5.1A or 5.1B
- Windows (32-bit on Intel) Windows NT 4 Server, Windows 2000 Server, Windows Server 2003, Windows XP Professional
- Windows (64-bit on Itanium) Windows Server 2003
- z/OS, Version 1

Availability, Pricing and Terms.

SAS Software is licensed on an annual basis. Please contact the SAS Institute directly for more information.

Contact Information.

SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513-2414
USA
Telephone: (919) 677-8000
Fax: (919) 677-4444
SAS Home Page: http://www.sas.com/
Statistics and Operations Research: http://www.sas.com/rnd/app/

Additional Information.

Recent papers and documentation on the survey selection and analysis procedures are available from SAS Institute's Statistics and Operations Research website at http://support.sas.com/rnd/app/da/new/dasurvey.html; see links at bottom of that page for papers and documentation.

Summary of Survey Software: Stata

Vendor.

StataCorp

Types of Designs That Can Be Accommodated.

- stratified designs;
- cluster sampling;
- variance estimation for multistage sample data can be carried out through the customary between-PSU-squared-differences calculation;
- finite-population corrections can be calculated for simple random sampling without replacement of sampling units within strata.

Types of Estimands and Statistical Analyses That Can Be Accommodated.

There are currently about 20 Stata commands for various analyses of survey data, including the following analyses and others:

- Estimation of means, totals, ratios, and proportions.
- Linear regression, logistic regression, and probit; also, tobit, interval, censored, instrumental variables, multinomial logit, ordered logit and probit, and Poisson. Point estimates, associated standard errors, confidence intervals, and design effects for the full population or subpopulations are displayed. Auxiliary commands will display all this information for linear combinations (e.g., differences) of estimators, and conduct hypothesis tests.
- Contingency tables with Rao-Scott corrections of chi-squared tests; new survey-corrected regression commands including tobit, interval, censored, instrumental variables, multinomial logit, ordered logit and probit, and Poisson.

Restrictions on Number of Variables or Observations.

Maximum number of observations limited only by computer RAM (virtual memory can be used, but commands run slower). Maximum number of variables is 32,767 with Stata/SE and 2,047 with Intercooled Stata.

Primary Methods Used for Variance Estimation.

Taylor-series linearization is used in the survey analysis commands. There are also commands for jackknife and bootstrap variance estimation; although these are not specifically oriented to survey data, they will accommodate survey features like clustering and stratification.

General Description of the "Feel" of the Software.

Stata is a complete statistical software package with full statistical, data management, and graphical capabilities. It can be run interactively or in batch mode, and is fully programmable. The survey commands are part of the standard software package. Initially, data can be read in from ASCII files and a Stata-format data file created; or data in other file formats can be translated to Stata format using a standalone software package (Stat/Transfer or DBMS/Copy).

Platforms on which the Software Can Be Run.

- Windows (all current versions)
- Power Macintosh (OS X 10.1 or later)
- Alpha AXP running Digital Unix
- HP-9000 with HP-UX
- Intel Pentium with Linux (32-bit)
- RS/6000 running AIX
- SGI running Irix 6.5
- SPARC running Solaris (64-bit and 32-bit)
- Linux 64

Software distributed as precompiled object program.

Availability, Pricing and Terms.

One-time purchase. Upgrade purchases are optional. Generous academic discount. Volume discounts and student discounts.

Contact Information.

StataCorp
4905 Lakeway Drive
College Station, TX 77845
800-782-8272 (U.S.)
800-248-8272 (Canada)
979-696-4600 (Worldwide)
979-696-4601 (Fax)
E-mail: stata@stata.com
Web site: http://www.stata.com

Additional Information.

This software is discussed in the review article from *The Survey Statistician*.

Summary of Survey Software: SUDAAN

Vendor.

Research Triangle Institute

Types of Designs That Can Be Accommodated.

Multiple design options allow users to analyze data from stratified, cluster sample, or multistage sample designs. Sample members may have been selected with unequal probabilities, and either with or without replacement. Any number of strata and stages can be specified. In addition, different design options may be combined in one study if different sampling methods were used for parts of the population.

Types of Estimands and Statistical Analyses That Can Be Accommodated.

SUDAAN includes the following statistical procedures:

- MULTILOG: Fits multinomial logistic regression models to ordinal and nominal categorical data and computes hypothesis tests for model parameters. Estimates odds ratios and their 95% confidence intervals for each model parameter. Has GEE (Generalized Estimating Equation) modeling capabilities for efficient parameter estimation.
- REGRESS: Fits linear regression models to continuous outcomes and performs hypothesis tests concerning the model parameters.
- LOGISTIC: Fits logistic regression models to binary data and computes hypothesis tests for model parameters. Estimates odds ratios and their 95% confidence intervals for each model parameter.
- SURVIVAL: Fits proportional hazards (Cox regression) models to failure time data. Estimates hazard ratios and their 95% confidence intervals for each model parameter.
- CROSSTAB: Computes frequencies, percentage distributions, odds ratios, relative risks, and their standard errors (or confidence intervals) for user-specified cross-tabulations, as well as chi-square tests of independence and the Cochran-Mantel-Haenszel chi-square test for stratified two-way tables.
- DESCRIPT: Computes estimates of means, totals, proportions, percentages, geometric means, quantiles, and their standard errors. Also computes standardized estimates and tests of single degree-of-freedom contrasts among levels of a categorical variable.
- RATIO: Computes estimates and standard errors of generalized ratios of the form (Summation y) / (Summation x), where x and y are observed variables. Also computes standardized estimates and tests single-degree-of-freedom contrasts among levels of a categorical variable.
- The EFFECT statement allows users to specify contrasts of regression coefficients and hypothesis tests using simple effect names.

Restrictions on Number of Variables or Observations.

None.

Primary Methods Used for Variance Estimation.

The Taylor series linearization method (GEE for regression models) is used combined with variance estimation formulas specific to the sample design. The user does not need to develop special replicate weights since the sample design can be specified directly to the program.

Jackknife and Balanced Repeated Replication (BRR) variance estimation is also supported.

General Description of the "Feel" of the Software.

SUDAAN uses a SAS-like language. There are two versions of Sudaan with different data interfaces:

- "SAS-callable" Sudaan: SUDAAN is called directly as a SAS procedure.
- "Standalone Sudaan": Independent program that reads external file formats, including SAS files or SPSS files.

In either case, the same programming language is used.

Platforms on which the Software Can Be Run.

- PCs under Windows 95 or later versions. This is now the primary platform for Sudaan.
- Sun SPARC computers under Solaris 2.6 and up.

SUDAAN is distributed as a precompiled program.

Availability, Pricing and Terms.

See pricing details at Sudaan Web site.

Contact Information.

SUDAAN Product Coordinator
Research Triangle Institute
3040 Cornwallis Road
Research Triangle Park, NC 27709-2194
Telephone: 919-541-6602
FAX: 919-541-7431
Email: SUDAAN@rti.org
URL: http://www.rti.org/sudaan/

Additional Information.

SUDAAN offers public 2-day or 3-day training classes several times each year. Classes can also be taught at user sites.

The following papers about Sudaan are available on-line:

- Full help manual may be viewed on-line.
- Bieler and Williams (1996), "Application of the SUDAAN Software Package to Clustered Data Problems: Pharmaceutical Research."
- "Analyzing Repeated Measures and Cluster-Correlated Data Using SUDAAN Release 7.5" (1997).
- "Analyzing Survey Data Using SUDAAN Release 7.5" (1997, compares Taylor series, jackknife and BRR variance estimates).

An extensive on-line help library is included for interactive use.

This software is discussed in the review article from *The Survey Statistician*.
See also Shah and Barnwell (1993), "Recent developments and future plans for SUDAAN" in *Proceedings of the Survey Research Methods Section, ASA*, 657-661.

APPENDIX B:
CODE FOR ANALYZING SUBPOPULATIONS USING SAS, STATA, AND SUDAAN

The program code in this appendix yields correct estimates of standard errors based on subsets of the KID and can be used when computing constraints prevent use of the entire KID. The generally recommended approach for calculating standard errors – using the entire KID – was illustrated for the asthma subpopulation analyses shown in the main body of this report. In contrast, the example in this appendix first subsets the KID to patients with asthma, then augments this subset with a “dummy” observation for each KID hospital to ensure that the proper formula is used to calculate standard errors. This approach “tricks” the software into believing that all KID hospitals are in the analysis, even though not all hospitals may contribute discharges to the analysis. The results from this approach are identical to the results shown in the body of this report.

SAS Programming Statements

```
/* Create analysis file */
libname IN “location of kid file” ;
data ASTHMA ;
  set IN.KID_2003_CORE;
  if DXCCS1 = 128;
  DISCHGS = 1 ;
Run;

/*  Augment the asthma subset with hospital-level observations  */

data COMBINED ;
  set ASTHMA
      IN.KID_2003_HOSPITAL(IN=INHOSP KEEP=HOSPID KID_STRATUM);
  INSUBSET = 1;
  if INHOSP then do;
      INSUBSET = 2;       * assign a value outside the subset ;
      DISCWT = 1;         * assign a valid weight ;
      DIED = 0;           * set analysis variables to zero ;
      DISCHGS = 0;
      LOS = 0;
      TOTCHG = 0;
  end;
run;

/* Obtain subpopulation estimates */

proc surveymeans data=COMBINED sum std mean stderr;
  weight DISCWT ;
  class DIED ;
  cluster HOSPID ;
  strata KID_STRATUM ;
  var DISCHGS LOS DIED TOTCHG ;
  domain INSUBSET ;
run;

/* Note: If finite population estimates of standard errors are   */
/* wanted, then the PROC SURVEYMEANS statement could include the */
/* “total= ” option, indicating a file containing the number     */
/* of population hospitals in each stratum, such as could be     */
/* constructed from the KID hospital file.                       */
```

The desired statistics correspond to the domain with insubset = 1. Subgroups within the asthma subpopulation (e.g., males and females) can also be analyzed using the DOMAIN statement. For example, if estimates are desired separately for male and female discharges with asthma, female = 0 can be assigned to the hospital-level observations, and the following statement can be added to PROC SURVEYMEANS:

```
domain INSUBSET * FEMALE ;
```

Again, the desired statistics correspond to the domains with insubset = 1.

Stata Programming Statements

In Stata, the hospital-level observations can be assigned zero weights to generate appropriate standard errors. It is not necessary to create a special domain variable, like insubset.

```
/* Using SAS, create an ASCII file for use by STATA */

libname IN “location of KID file” ;

data ASTHMA ;
  set IN.KID_2003_CORE;
  if DXCCS1 = 128 ;
  DISCHGS = 1 ;
run;

/*  Augment the asthma subset with hospital-level observations  */

data COMBINED ;
  set ASTHMA
      IN.KID_2003_HOSPITAL(IN=INHOSP KEEP=HOSPID KID_STRATUM);
  if INHOSP then do;
      DISCWT = 0;          * Assign zero weights ;
      DIED = 0;            * Set analysis variables to zero ;
      DISCHGS = 0;
      LOS = 0;
      TOTCHG = 0;
  end;
run;

data _null_ ;
  set  IN.KID_2003_CORE ;
  if DXCCS1 = 128 then ASTHMA = 1 ; else ASTHMA = 0 ;
  file "file to write" ;
  DISCHGS = 1 ;
  if LOS < 0 then LOS = . ;
  if DIED < 0 then DIED = . ;
  if TOTCHG < 0 then TOTCHG = . ;
  put KID_STRATUM 1-4 HOSPID 6-10 DIED 12 LOS 14-17
      DISCHGS 19 TOTCHG 21-27 ASTHMA 29 +1 DISCWT ;
RUN;

/* Obtain STATA version 9.1 estimates */
/* Note: Stata commands should be entered in lowercase text */

set memory 15000
infile KID_STRATUM HOSPID DIED LOS DISCHGS TOTCHG
       DISCWT using “DATASET NAME””
svyset HOSPID [pweight = DISCWT], strata (KID_STRATUM)
svytotal DISCHGS
svymean LOS TOTCHG
svyratio DIED DISCHGS
```

Subpopulations of the asthmatic subset can also be analyzed using the SUBPOP option, similar to the Stata code shown for the asthma example in the main body of this report. For example, if estimates are desired separately for female discharges with asthma, the following statements can be substituted:

```
svy: total DISCHGS, subpop(FEMALE)
svy: mean LOS TOTCHG, subpop(FEMALE)
svy: ratio DIED DISCHGS, subpop(FEMALE)
```

This will produce estimates for female discharges with asthma (observations with female = 1). If estimates are desired for male discharges with asthma, then an indicator variable for males should be used.

SUDAAN Programming Statements

```
/* Create analysis file */
libname IN “location of KID file” ;

data ASTHMA ;
  set IN.KID_2003_CORE;
  if DXCCS1 = 128 ;
  DISCHGS = 1 ;
run;

/*  Augment the asthma subset with hospital-level observations  */

data COMBINED ;
  set ASTHMA
      IN.KID_2003_HOSPITAL(IN=INHOSP KEEP=HOSPID KID_STRATUM);
  INSUBSET = 1;
  if INHOSP then do;
      INSUBSET = 2;       * assign a value outside the subset ;
      DISCWT = 1;         * assign a valid weight ;
      DIED = 0;           * set analysis variables to zero ;
      DISCHGS = 0;
      LOS = 0;
      TOTCHG = 0;
  end;
run;

proc sort data=COMBINED; by KID_STRATUM HOSPID;
run;

/*  Use the SUBPOPN statement to analyze the asthma subset    */

proc descript data=COMBINED filetype=sas design=wr ;
  subpopn INSUBSET = 1 / name = "Asthma Subpopulation" ;
  weight DISCWT ;
  nest KID_STRATUM HOSPID ;
  setenv colwidth = 24 decwidth = 5 ;
  var LOS DISCHGS TOTCHG ;
  print total setotal mean semean ;

/* Calculate statistics for the categorical variable “died” */

proc descript data=COMBINED filetype=sas design=wr ;
  weight DISCWT ;
  setenv colwidth = 24 decwidth = 5 ;
  subpopn INSUBSET = 1 / NAME = "Asthma Subpopulation" ;
  nest KID_STRATUM HOSPID ;
  var DIED ;
  catlevel 1 ;
run;
```

Subsets of the asthma subpopulation can be analyzed by using the SUBGROUP statement. For example, if estimates are desired separately for male and female discharges with asthma, the following statements can be added to PROC DESCRIPT:

```
recode FEMALE = (0,1) ;
subgroup FEMALE ;
levels 2 ;
```

Variables in the SUBGROUP statement are assumed to have values that are consecutive positive integers. The RECODE statement converts the values of female from (0,1) to (1,2).