RAT-STATS 2007 Companion Manual

Version 2

Department of Health & Human Services
Office of Inspector General
Office of Audit Services

RAT-STATS Companion Manual

TABLE OF CONTENTS

PREFACE

The purpose of this manual is to provide:

- an overview of each program in the Windows version of RAT-STATS,

- examples illustrating the application of the software,

- snapshots of data sets used by the programs,

- some discussion regarding the program output, and

- formulas used within the software.

The intent is for the auditor/specialist to use as much of this discussion as he/she finds helpful.

While the RAT-STATS Users Guide gives descriptions of program input and output, this

Companion Manual should provide insight as to how to better use the software and exactly how

the program derives the results. The formulas are provided so that OAS has a single source for

all formulas in the event that a question is raised as to exactly how a particular result was

obtained.

We hope you find that the manual makes the OAS software easier to understand and easier to

apply. Please pass on any suggestions or corrections to Office of Inspector General, Office of

Public Affairs at paffairs@oig.hhs.gov.

RANDOM NUMBERS

Whatever statistical sampling design you end up using (including stratified and/or multistage), at some point in the data collection you will need one or more random samples. The next section, dealing with Unrestricted Random Sampling, will examine the mechanics and estimation procedures using such a sample in detail, but first it is necessary to discuss procedures for generating a random sample. A number of programs exist for such purposes; namely:

- Single-Stage Random Numbers

- Sets of Two Numbers

- Sets of Three Numbers

- Sets of Four Numbers

- Frames - Single Stage

- Frames - Sets of Two

- RHC Sample Selection

Single-Stage Random Numbers

This program generates an unduplicated quantity of random numbers. Random values (the output from this program) can be output in sequential order, random order, or a mixture of both. For values in sequential order, you will see the values printed in sequential order, beginning with the smallest selected item number and proceeding to the largest item number. For values in random order, the items are printed in the order in which they were selected by the program.

Example 1. A universe contains 1,000 payments and a simple random sample of 10 payments (with four spares) is needed. What items should be selected?

Solution: Using this program and a seed value of 12345, the sampled payments are those numbered as follows:

> 9, 236, 337, 340, 346, 497, 556, 641, 658, and 884

The four spares are payments 404, 624, 927, and 947.

Sets of Two Numbers

This program will generate unduplicated <u>pairs</u> of random numbers. This is useful when sample items are selected through a two-step process (e.g., page number and line number).

Random values (the output from this program) can be output in sequential order, random order, or a mixture of both. Values in sequential order will be printed in sequential order, beginning with the smallest selected item number and proceeding to the largest item number. For values in random order, the items are printed in the order in which they were selected by the program.

Example 2. Items are selected from a computer printout that had pages numbered 1 through 658 and had 66 lines on each page. A simple random sample of 10 items (with four spares) is needed. Which items should be selected?

Solution: Using this program and a seed value of 12345, the sampled payments are:

PAGE:	224	258	266	327	366	400	422	433	561	610
ITEM:	23	37	42	1	16	7	23	59	40	63

The spares are:

PAGE:	579	109	188	330
ITEM:	61	54	50	49

Sets of Three Numbers

This program will generate unduplicated sets of <u>three</u> random numbers. This should be used when sample items are selected through a three-step process (e.g., month, page, and line number).

Random values (the output from this program) can be output in sequential order, random order, or a mixture of both. Values in sequential order will be printed in sequential order, beginning with the smallest selected item number and proceeding to the largest item number. For values in random order, the items are printed in the order in which they were selected by the program.

Example 3. Same as Example 2, where the pages are numbered 1 through 658 for each month. Here the universe consists of 1 year's worth (12 months) of computer printouts. We need four sample items and two spares.

Solution: Using this program and a seed value of 12345, the sampled items are:

MONTH:	3	5	6	8
PAGE:	224	266	6	582
ITEM:	23	42	37	43

The two spares are:

MONTH:	12	8
PAGE:	623	258
ITEM:	57	37

Sets of Four Numbers

This program will generate unduplicated sets of <u>four</u> random numbers. This should be used when sample items are selected through a four-step process (e.g., year, month, page, and line number).

Random values (the output from this program) can be output in sequential order, random order, or a mixture of both. For values in sequential order, you will see the values printed in sequential order, beginning with the smallest selected item number and proceeding to the largest item number. For values in random order, the items are printed in the order in which they were selected by the program.

Example 4. Same as Example 3, where the pages are numbered 1 through 658 for each month and year (total of 5 years). We need three sample items and two spares.

Solution: Using this program and a seed value of 12345, the sampled items are:

YEAR:	2	3	4
MONTH:	5	1	5
PAGE:	433	366	266
ITEM:	59	16	42

The two spares are:

YEAR:	5	2
MONTH:	12	7
PAGE:	561	400
ITEM:	40	7

Frames - Single Stage

This program will generate an unduplicated set of random numbers which is useful when the universe of sampling items either (1) contains gaps of numbers or (2) the numbering system repeats within the universe. For instance, the universe of items consists of two frames, numbered 1 through 1,050 and 8,405 through 9,565.

Random values (the output from this program) can be output in sequential order, random order, or a mixture of both. For values in sequential order, you will see the values printed in sequential order, beginning with the smallest selected item number in the first frame (if any) and proceeding to the largest item number in the last frame (if any). For values in random order, the items are printed in the order in which they were selected by the program.

Example 5. A universe of items that refer to payment of a particular medical procedure are numbered as follows:

      1 - 1,050 (frame 1)

  and 8,405 - 9,565 (frame 2)

A sample of five items is needed. Three of these items should be in sequential order and the remaining two in random order.

Solution: Using this program and a seed number of 12345, the three sample items in sequential order are:

FRAME	ITEM NUMBER
1	20
2	8,452
2	8,584

The two items in random order are

    <u>FRAME</u>   <u>ITEM</u> <u>NUMBER</u>
      1   520
      1   752

Explanation: For this example there are 2,211 items in the frame since there are 1,050 items in the first frame and (9,565 - 8,405 + 1), i.e., 1,161 items, in the second frame. For the sequential items, three values between 1 and 2,211 are generated. These values are 20, 1,098, and 1,230. Since 1,098 is outside the first frame, it is in the second frame; in particular, its location in the second frame would be

$$[8,405 + (1,098 - 1,050) - 1]$$

that is, item number 8,452. Similarly, the value of 1,230 points to item number 8,584.

Similarly, for the two items in random order, the program generated two values between 1 and 2,211. These values were 520 and 752. Since both values are less than 1,050, these locations are items 520 and 752 in the first frame.

Frames - Sets of Two

This program is a combination of two programs, Frames - Single Stage and Sets of Two Numbers. The program will generate an unduplicated set of random numbers which should be used when (1) pairs of numbers are used to locate sample items (as in Sets of Two Numbers) and (2) the universe has gaps or the numbering system repeats (as in Frames - Single Stage).

Random values (the output from this program) can be output in sequential order, random order, or a mixture of both. For values in sequential order, you will see the values printed in sequential order, beginning with the smallest selected item number in the first frame (if any) and proceeding to the largest item number in the last frame (if any). For values in random order, the items are printed in the order in which they were selected by the program.

Example 6. A universe of transactions consists of three sets of computer pages, numbered as follows:

FRAME	RANGE (Page Numbers)
1	1 - 100
2	1 - 456
3	45 - 832

In addition, within each frame there are an equal number of line items per page. The range within each frame is:

FRAME	RANGE (Number of Lines)
1	1 - 66
2	1 - 66
3	1 - 66

A sample of three items in sequential order and two items in random order is needed.

Solution: Using this program and a seed value of 12345, the three sample items in sequential

order are:

FRAME	PAGE NO.	LINE
1	12	45
3	156	21
3	236	14

The two items in random order are:

FRAME	PAGE NO.	LINE
2	216	64
2	357	36

Explanation: For this example, the frame consists of

$$(100)(66) + (456)(66) + (832 - 45 + 1)(66)$$

$$= 6,600 + 30,096 + 52,008 = 88,704 \text{ items.}$$

Random numbers between 1 and 88,704 are generated. Values between 1 and 6,600 will be in

the first frame; values between 6,601 and 6,600 + 30,096 = 36,696 are in the second frame; and

values between 36,697 and 88,704 will come from the third frame.

The three random values (not the spares) generated are 771, 44,043, and 49,316.

To find the value corresponding to 771:

1. Find (771/66) + 1 = 12.682. The integer part of this is 12. So, this value is on page number

(subframe) 12 of frame 1.

2. The decimal part of this number is .682. Multiply this by 66 (the number of lines per page in

the first frame for this example) and round to the nearest integer. This is 45. This item is on line

45 of page 12.

To find the sample value corresponding to 44,043: This value is larger than 36,696 so it is in the third frame.

1. Find $[(44{,}043 - 36{,}696)/66] + 1 = 156.318$. Here, 45 is the low number (input) for frame 3. The integer part of this value is 156. So, this value is on page number (subframe) 156 of frame 2.

2. The decimal part of this number is .318. Multiply this by 66 (the number of lines per page in the second frame for this example) and round to the nearest integer. This is 21. This item is on line 21 of page 156.

Finally, consider the third randomly generated value of 49,316. This is 36,696, so it is in the third frame.

1. Find $[(49{,}316 - 36{,}696)/66] + 45 = 236.212$. The integer part of 236.212 is 236. So, this value is on page number 236 of frame 3.

2. The decimal part of this number is .212. Multiply this by 66 (the number of lines per page) and round to the nearest integer. This is 14. This item is on line 14 of page 236.

RHC Sample Selection

The RHC selection/appraisal procedure is named after three statisticians -- J.N.K. Rao, H.O. Hartley, and William Cochran -- and was originally proposed in 1962. This procedure is essentially the same as single-stage SRS sampling except that the <u>size</u> of each primary unit (cluster) is used to select the sample. It provides a method of sample selection that allows sampling without replacement (the usual procedure) while "maintaining the flavor" of using probability proportional to size. It can be used to select primary units (P.U.s) in a two-stage design or primary and secondary units (S.U.s) in a three-stage design.

<u>Comment</u>: Strictly speaking, you cannot use pure probability proportional to size (pps) sampling when sampling without replacement. To understand why, consider a situation in which a population contains 10 primary units with seven "large" P.U.s and three "small" ones. If the sample size is eight, then one of the small P.U.s must be selected, regardless of its small size. The RHC procedure is not pure pps sampling, but comes very close while allowing the auditor to sample without replacement.

Procedure for Two-Stage RHC Sampling

Suppose that you have N P.U.s and you want a sample of n P.U.s. The procedure is to:

1. <u>Randomly</u> put (partition) the N P.U.s into n groups (no attention to size here).
2. Within each of the n groups, select one P.U. using pps.

Example 7. N = 15, n = 3

1. Generate 3 groups, each containing 5 P.U.s

One possibility: Generate 15 random numbers between 0 and 1. Suppose the smallest value

is in location 8, next largest in location 5, next largest in location 13, next largest in location

12, and the next largest in location 2. The first group consists of P.U.s 2, 5, 8, 12, and 13.

Continue, to get the remaining two groups.

2. Suppose P.U. #7 is put into group 3.

Size of group 3 is 1,000 beds and size of P.U. #7 is 100 beds.

P.U. #7 will be selected from group 3 with probability $100/1000 = .1$.

Example 8. In a particular region of the U.S. there are N = 90 universities with government

research grants. Because these universities are so widespread, it was decided to use a sample of

n = 10 universities. Rather than audit all grants at a selected university, it was decided (based on

available resources) to audit roughly 20% of the grants at each selected university. We know

that there are a total of M = 4,500 grants in all 90 universities.

Size: As a measure of the size for each university, use the total grant dollars.

Each row of the data file will contain:

University ID, number of grants, total grant dollars

i.e.,

ID of P.U., number of S.U.s (universe) in this P.U., size of P.U.

There are 90 rows of data (one for each P.U.) contained in this data set, named UNIVRHC.TXT.

Output: The 10 universities to use in the sample (see last page of computer output) are:

UNIV78, UNIV42, UNIV49, UNIV5, UNIV19, UNIV38, UNIV62, UNIV28,
UNIV60, and UNIV75

This program will create an output file specified by the user (OutRHCsummary.txt for this

illustration) that is used as one of the input files by the RHC appraisal program. Dataset

UNIVRHC.TXT, the program output, and output file OutRHCsummary.txt are contained in the

following pages.

Dataset UNIVRHC.TXT   **< - - - continued - - - >**   **< - - - continued - - - >**
                                  (1)   (2)   (3)

UNIV1	42	8	UNIV31	52	11	UNIV61	66	13
UNIV2	21	4	UNIV32	66	14	UNIV62	77	18
UNIV3	63	13	UNIV33	25	5	UNIV63	31	7
UNIV4	74	16	UNIV34	60	12	UNIV64	46	9
UNIV5	51	11	UNIV35	19	4	UNIV65	32	7
UNIV6	43	9	UNIV36	24	5	UNIV66	68	14
UNIV7	57	11	UNIV37	44	9	UNIV67	41	9
UNIV8	49	10	UNIV38	76	17	UNIV68	28	6
UNIV9	63	13	UNIV39	41	9	UNIV69	66	14
UNIV10	18	4	UNIV40	77	18	UNIV70	31	7
UNIV11	64	13	UNIV41	37	8	UNIV71	27	6
UNIV12	56	11	UNIV42	63	12	UNIV72	33	7
UNIV13	19	4	UNIV43	52	11	UNIV73	23	4
UNIV14	44	9	UNIV44	76	17	UNIV74	71	15
UNIV15	20	4	UNIV45	51	10	UNIV75	75	16
UNIV16	34	7	UNIV46	23	4	UNIV76	47	10
UNIV17	25	6	UNIV47	24	5	UNIV77	50	10
UNIV18	38	9	UNIV48	68	15	UNIV78	37	7
UNIV19	72	16	UNIV49	34	7	UNIV79	77	18
UNIV20	46	10	UNIV50	49	10	UNIV80	49	10
UNIV21	44	9	UNIV51	55	11	UNIV81	76	17
UNIV22	64	13	UNIV52	38	9	UNIV82	66	14
UNIV23	45	9	UNIV53	72	16	UNIV83	28	6
UNIV24	55	11	UNIV54	51	10	UNIV84	77	17
UNIV25	29	7	UNIV55	71	15	UNIV85	27	6
UNIV26	36	7	UNIV56	59	12	UNIV86	75	17
UNIV27	40	9	UNIV57	23	4	UNIV87	71	15
UNIV28	78	18	UNIV58	57	11	UNIV88	59	12
UNIV29	49	10	UNIV59	53	11	UNIV89	71	15
UNIV30	60	12	UNIV60	64	13	UNIV90	72	16

Columns: (1) primary unit ID
         (2) number of grants
         (3) grant dollar amount (x $100,000) ◄ This is the size of the university.

```
                    DEPARTMENT OF HEALTH & HUMAN SERVICES
                       OIG – OFFICE OF AUDIT SERVICES
Date: 10/15/2004      GENERATION OF PRIMARY UNIT SAMPLE        Time: 12:52
                    NAME OF INPUT FILE: C:\TEMP\UNIVRHC.TXT

                          GROUPS OF PRIMARY UNITS
```

```
                     ********* GROUP 1 *********
                                    PRIMARY UNIT    SECONDARY
     PRIMARY UNIT IDENTIFICATION        SIZE         UNIVERSE
     ==============================  =============  =============
     UNIV51                                    11             55
     UNIV44                                    17             76
     UNIV32                                    14             66
     UNIV78 <-- selected                        7             37
     UNIV79                                    18             77
     UNIV2                                      4             21
     UNIV52                                     9             38
     UNIV33                                     5             25
     UNIV47                                     5             24

        GROUP TOTALS:   9                      90            419


                     ********* GROUP 2 *********
                                    PRIMARY UNIT    SECONDARY
     PRIMARY UNIT IDENTIFICATION        SIZE         UNIVERSE
     ==============================  =============  =============
     UNIV6                                      9             43
     UNIV42 <-- selected                       12             63
     UNIV65                                     7             32
     UNIV40                                    18             77
     UNIV45                                    10             51
     UNIV1                                      8             42
     UNIV80                                    10             49
     UNIV36                                     5             24
     UNIV70                                     7             31

        GROUP TOTALS:   9                      86            412


                     ********* GROUP 3 *********
                                    PRIMARY UNIT    SECONDARY
     PRIMARY UNIT IDENTIFICATION        SIZE         UNIVERSE
     ==============================  =============  =============
     UNIV46                                     4             23
     UNIV7                                     11             57
     UNIV90                                    16             72
     UNIV49 <-- selected                        7             34
     UNIV21                                     9             44
     UNIV4                                     16             74
     UNIV54                                    10             51
     UNIV61                                    13             66
     UNIV77                                    10             50

        GROUP TOTALS:   9                      96            471


                     ********* GROUP 4 *********
                                    PRIMARY UNIT    SECONDARY
     PRIMARY UNIT IDENTIFICATION        SIZE         UNIVERSE
     ==============================  =============  =============
     UNIV73                                     4             23
     UNIV50                                    10             49
     UNIV58                                    11             57
     UNIV57                                     4             23
     UNIV82                                    14             66
     UNIV23                                     9             45
     UNIV5  <-- selected                       11             51
     UNIV64                                     9             46
```

```
UNIV34                                              12              60

    GROUP TOTALS:  9                                84             420
```

```
              ********* GROUP 5 *********
                                   PRIMARY UNIT    SECONDARY
PRIMARY UNIT IDENTIFICATION            SIZE        UNIVERSE
============================    =============   =============
UNIV84                                   17              77
UNIV35                                    4              19
UNIV16                                    7              34
UNIV81                                   17              76
UNIV27                                    9              40
UNIV85                                    6              27
UNIV19 <-- selected                      16              72
UNIV68                                    6              28
UNIV26                                    7              36

    GROUP TOTALS:  9                     89             409
```

```
              ********* GROUP 6 *********
                                   PRIMARY UNIT    SECONDARY
PRIMARY UNIT IDENTIFICATION            SIZE        UNIVERSE
============================    =============   =============
UNIV37                                    9              44
UNIV83                                    6              28
UNIV63                                    7              31
UNIV14                                    9              44
UNIV43                                   11              52
UNIV31                                   11              52
UNIV15                                    4              20
UNIV48                                   15              68
UNIV38 <-- selected                      17              76

    GROUP TOTALS:  9                     89             415
```

```
              ********* GROUP 7 *********
                                   PRIMARY UNIT    SECONDARY
PRIMARY UNIT IDENTIFICATION            SIZE        UNIVERSE
============================    =============   =============
UNIV69                                   14              66
UNIV18                                    9              38
UNIV25                                    7              29
UNIV59                                   11              53
UNIV30                                   12              60
UNIV10                                    4              18
UNIV24                                   11              55
UNIV62 <-- selected                      18              77
UNIV17                                    6              25

    GROUP TOTALS:  9                     92             421
```

```
                    ********* GROUP 8 *********
                                   PRIMARY UNIT    SECONDARY
          PRIMARY UNIT IDENTIFICATION      SIZE        UNIVERSE
          =============================  =============  =============
          UNIV28 <-- selected                    18             78
          UNIV41                                  8             37
          UNIV89                                 15             71
          UNIV66                                 14             68
          UNIV11                                 13             64
          UNIV86                                 17             75
          UNIV56                                 12             59
          UNIV12                                 11             56
          UNIV72                                  7             33

              GROUP TOTALS:   9                 115            541


                    ********* GROUP 9 *********
                                   PRIMARY UNIT    SECONDARY
          PRIMARY UNIT IDENTIFICATION      SIZE        UNIVERSE
          =============================  =============  =============
          UNIV71                                  6             27
          UNIV8                                  10             49
          UNIV67                                  9             41
          UNIV3                                  13             63
          UNIV60 <-- selected                    13             64
          UNIV76                                 10             47
          UNIV74                                 15             71
          UNIV9                                  13             63
          UNIV20                                 10             46

              GROUP TOTALS:   9                  99            471


                    ********* GROUP 10 *********
                                   PRIMARY UNIT    SECONDARY
          PRIMARY UNIT IDENTIFICATION      SIZE        UNIVERSE
          =============================  =============  =============
          UNIV22                                 13             64
          UNIV39                                  9             41
          UNIV88                                 12             59
          UNIV55                                 15             71
          UNIV29                                 10             49
          UNIV75 <-- selected                    16             75
          UNIV87                                 15             71
          UNIV13                                  4             19
          UNIV53                                 16             72

              GROUP TOTALS:   9                 110            521
```

```
                      DEPARTMENT OF HEALTH & HUMAN SERVICES
                          OIG - OFFICE OF AUDIT SERVICES
Date: 10/15/2004        GENERATION OF PRIMARY UNIT SAMPLE          Time: 12:52
                   NAME OF OUTPUT FILE: C:\TEMP\OutRHCsummary.txt

  FIRST SEED NUMBER:      100.00     SECOND SEED NUMBER:      200.00

     NUMBER OF PRIMARY UNITS IN THE POPULATION:     90
     NUMBER OF PRIMARY UNITS SAMPLED:               10

                                                                        UNITS
                         SECONDARY       PRIMARY                          IN
PRIMARY UNIT ID           UNIVERSE      UNIT SIZE       GROUP SIZE      GROUP
==========================  ==============  ==============  ==============  =====
UNIV78                            37             7               90         9
UNIV42                            63            12               86         9
UNIV49                            34             7               96         9
UNIV5                             51            11               84         9
UNIV19                            72            16               89         9
UNIV38                            76            17               89         9
UNIV62                            77            18               92         9
UNIV28                            78            18              115         9
UNIV60                            64            13               99         9
UNIV75                            75            16              110         9
```

NOTE: In practice, it is recommended that you not set the two seed values unless you are trying to duplicate prior results.

Output file OutRHCsummary.txt

```
  (1)                    (2)            (3)            (4)            (5)
UNIV78                   37             7              90             9
UNIV42                   63             12             86             9
UNIV49                   34             7              96             9
UNIV5                    51             11             84             9
UNIV19                   72             16             89             9
UNIV38                   76             17             89             9
UNIV62                   77             18             92             9
UNIV28                   78             18             115            9
UNIV60                   64             13             99             9
UNIV75                   75             16             110            9
```

Columns: (1) selected primary unit
            (2) number of grants (secondary units)
            (3) grant dollar amount (x $100,000) ← This is the size of the primary unit.
            (4) size of the group containing this primary unit
            (5) number of universities (primary units) in this group

Procedure for Three-Stage RHC Sampling

1. A sample of primary units (clusters) is obtained as in the two-stage procedure. The size of the primary units is considered for this sample, where pps sampling is used for each group of primary units.

2. A sample of secondary units is obtained within each chosen primary unit by partitioning the primary unit into random groups of secondary units. The numbers of S.U.s in each group are chosen to be as nearly equal as possible. Using pps sampling, and the size of each secondary unit, one secondary unit is chosen from each of the secondary groups.

3. A random sample of third-stage units is obtained for each of the chosen secondary units. No attention is paid to "size" here. This is a random sample.

Example 9. The previous example was expanded to include geographical regions.

> Primary units: 12 regions (select four)
>
> Secondary units: Universities (select 10 from each region)
>
> Third stage units: Grants (audit 20% from each university)

<u>Selection of Primary Units</u>

A file must be constructed containing (for each region) (1) the number of secondary units (universities) in this region and (2) the size of this region (total grant dollars). This file is GRANTSPU.TXT. The selected regions are 4, 6, 8, and 10 using seed values of 100 and 200.

File GRANTSPU.TXT

(1)	(2)	(3)
REGION1	117	1250
REGION2	63	610
REGION3	91	720
REGION4	123	1320
REGION5	107	1160
REGION6	116	1240
REGION7	102	960
REGION8	118	1300
REGION9	122	1320
REGION10	85	640
REGION11	94	930
REGION12	62	550

Columns: (1) region ID
          (2) number of universities (secondary units)
          (3) size (total grant amount x $100,000)

Selection of Secondary Units

The three-stage RHC sample selection procedure requires the user to only obtain information for each **selected** primary unit (i.e., regions 4, 6, 8, and 10 here). The information in each of these four files consists of the size of each secondary unit (university, here) and the number of third-stage units in the universe for each secondary unit. Each of these files should resemble file UNIVRHC.TXT contained in the previous two-stage RHC discussion. Consequently, for each sampled P.U., each line of the corresponding file should contain:

        university ID, number of grants at this university, total grant dollars

i.e.,

        secondary unit ID, no. of third-stage units, size of S.U.

After running the **RHC Sample Selection** program on each of these four regions, the following universities were selected:

REGION	UNIVERSITIES
4	85, 46, 7, 82, 30, 34, 27, 66, 65, 80
6	113, 43, 78, 104, 89, 112, 30, 65, 3, 99
8	112, 6, 77, 93, 75, 111, 62, 115, 70, 99
10	78, 43, 7, 73, 55, 33, 10, 59, 64, 39

<u>Selection of Third-Stage Units</u>

Suppose that approximately 20% of the grants at each selected university are to be audited. Each

of these 40 samples (4 regions x 10 universities) is obtained randomly using the **Single-Stage**

Random Numbers program.

NOTES: (1) The previous five program runs (one at the primary level and four at the
secondary level) created five output files. Using a word processor, these files can
be joined to form one of the input files (the one containing primary/secondary unit
information) for the three-stage RHC appraisal program which calculates the
confidence interval.

(2) This example is examined in more detail in the three-stage RHC appraisal section.

Summary of Input for RHC Sample Selection

<u>RHC Two-Stage</u>

The user must know:

1. The number of P.U.s in the universe and the sample.

2. The size of all P.U.s in the universe.

Procedure: 1. The user can set the number of S.U.s in each universe P.U. equal to one if
these are difficult to determine. This is the middle column in file
UNIVRHC.TXT used in the previous illustration.

2. Next, run the **RHC Two-Stage Sample Selection** program. Store the output
in a text file.

       3.   Using a word processor or spreadsheet, change the number of S.U.s for each sampled P.U. from one to the correct value.

RHC Three-Stage

The user must know:

       1. The number of P.U.s in the universe and the sample.

       2. The size of all P.U.s in the universe.

       3. The number of S.U.s to be sampled within each P.U.

       4. For each sampled P.U.,

          a. The number of S.U.s in the universe within this P.U.

          b. The size of all S.U.s within this P.U.

Procedure: 1. The user can set the number of S.U.s in each universe P.U. equal to 1 if these are difficult to determine.

          2. Run the **RHC Two-Stage Sample Selection** program. Store the output in a text file.

          3. Using a word processor or spreadsheet, change the number of S.U.s for each sampled P.U. from 1 to the correct value.

          4. For each sampled P.U., build a data file where each row consists of
             a. S.U. ID
             b. Number of third-stage units for this S.U. (OK to use a value of 1 here and correct later).
             c. Size of this S.U.

          5. For each sampled P.U., use the data set in step 4 as input to the **RHC Two-Stage Sample Selection** program. Store the output in a text file.

          6. Using a word processor or spreadsheet, change the number of third-stage units for each sampled S.U. from 1 to the correct value.

          7. Merge the results from step 2 and each sampled P.U. into one file. See PUSURHC3.TXT (below) for an example. The values in the second column (123, 54, 44, . . .) can be set to one and later changed to the correct values. This is one of the input files to the **RHC Three-Stage Appraisal** program.

> **NOTE**: Although this procedure allows for substituting 1s for the number of
> second- and third-stage units in the original pass, the required <u>size</u>
> information must be known.

8. Build the data file (file PUSURHC3.TXT for this illustration) containing the
   sampled third-stage units. This is the other input file required by the RHC
   three-stage appraisal program. Using a word processor or spreadsheet, the
   column of sample sizes (highlighted) was added to the files created by the five
   RHC Sample Selection programs.

File PUSURHC3.TXT

```
REGION4  123 10   1320   3410   3
UNIV85   54  11   11   125   12
UNIV46   44   9    9   131   12
UNIV7    77  15   17   119   12
UNIV82   52  10   11   129   12
UNIV30   54  11   11   141   12
UNIV34   50  10   10   140   12
UNIV27   76  15   16   138   12
UNIV66   76  15   16   128   13
UNIV65   62  12   14   125   13
UNIV80   70  14   15   155   13
REGION6  116 10   1240   3100   3
UNIV113  33   7    8   108   11
UNIV43   39   8    7   105   11
UNIV78   63  13   12   104   11
UNIV104  25   5    9    96   11
UNIV89   35   7    7   124   12
UNIV112  27   5   10   108   12
UNIV30   58  12   11    95   12
UNIV65   57  11   10   109   12
UNIV3    56  11   11   115   12
UNIV99   80  16   14   113   12
REGION8  118 10   1300   3170   3
UNIV112  75  15   16   125   11
UNIV6    34   7    8   127   11
UNIV7    51  10   11   120   12
UNIV93   54  11   11   136   12
UNIV75   52  10   11   126   12
UNIV111  84  17   17   134   12
UNIV62   64  13   14   123   12
UNIV115  59  12   13   137   12
UNIV70   65  13   14   143   12
UNIV99   60  12   13   129   12
REGION10 85  10   640   2320   3
UNIV78   39   8    7    62    8
UNIV43   42   8    7    68    8
```

File PUSURHC3.TXT - continued

```
UNIV7    56   11    9    54    8
UNIV73   27    5    5    63    8
UNIV55   78   16   12    70    8
UNIV33   65   13   10    77    9
UNIV10   60   12    9    76    9
UNIV59   52   10    8    71    9
UNIV64   50   10    8    73    9
UNIV39   38    8    6    68    9
```

Generating Spares for RHC Sampling

One question that arises here is what to do if one or more of the selected primary units is unattainable or unusable for some reason. There is a method of generating spares without having to start the sample selection process all over again, once the nonusable primary units have been identified.

A similar approach can be used in a three-stage plan if one or more secondary units are nonusable within a selected primary unit. The following example will illustrate how to recover when one or more primary units are nonusable with a two-stage RHC sampling plan.

Example 10. Population: N = 90 P.U.s (universities)   Sample: n = 5 P.U.s

The final section of the output using the RHC sample selection program is shown below; it turns out that universities 51 (in group 5) and 69 (in group 4) could not be used.

```
                                                                     UNITS
                               SECONDARY         PRIMARY                IN
PRIMARY UNIT ID                UNIVERSE         UNIT SIZE    GROUP SIZE  GROUP
=========================    =============    =============  =============  =====
UNIV47                              24               5             191      18
UNIV64                              46               9             175      18
UNIV52                              38               9             185      18
UNIV69┐                             66              14             203      18
UNIV51┘── can't use                 55              11             196      18
```

The corresponding output file created from the first pass is shown below.

```
   UNIV47                24            5          191           18
   UNIV64                46            9          175           18
   UNIV52                38            9          185           18
   UNIV69                66           14          203           18
   UNIV51                55           11          196           18
```

The section of the output containing the contents of group 4 follows.

```
                 ********* GROUP   4 *********
                                    PRIMARY UNIT    SECONDARY
         PRIMARY UNIT IDENTIFICATION      SIZE        UNIVERSE
         ==============================  ==============  ==============
         UNIV15                                4               20
         UNIV81                               17               76
         UNIV38                               17               76
         UNIV59                               11               53
         UNIV13                                4               19
         UNIV76                               10               47
         UNIV20                               10               46
         UNIV66                               14               68
         UNIV14                                9               44
         UNIV29                               10               49
         UNIV55                               15               71
         UNIV26                                7               36
         UNIV17                                6               25
         UNIV87                               15               71
         UNIV22                               13               64
         UNIV50                               10               49
         UNIV84                               17               77
         UNIV69 <-- Selected, can't use       14               66

              GROUP TOTALS:  18            203              957
```

Remove UNIV69 from the population and this group. Construct a data file (same format as

UNIVRHC.TXT) using this group only. This file (TEMP1.TXT) is shown below.

NOTE: When constructing this file, notice that columns 2 and 3 above (i.e., PRIMARY
              UNIT SIZE and SECONDARY UNIVERSE) need to be switched. This was done
              correctly in TEMP1.TXT.

```
              UNIV15     20      4
              UNIV81     76     17
              UNIV38     76     17
              UNIV59     53     11
              UNIV13     19      4
              UNIV76     47     10
              UNIV20     46     10
              UNIV66     68     14        File TEMP1.TXT
              UNIV14     44      9
              UNIV29     49     10
              UNIV55     71     15
              UNIV26     36      7
              UNIV17     25      6
              UNIV87     71     15
              UNIV22     64     13
              UNIV50     49     10
              UNIV84     77     17
```

Next, run the **RHC Sample Selection** program. Your input file is TEMP1.TXT and your sample

size is 1. This generates another P.U. (university) from this group. The output from this

program is shown below. UNIV22 was selected.

```
                     DEPARTMENT OF HEALTH & HUMAN SERVICES
                        OIG - OFFICE OF AUDIT SERVICES
Date: 10/15/2004        GENERATION OF PRIMARY UNIT SAMPLE          Time: 13:17
                   NAME OF INPUT FILE: C:\TEMP\TEMP1.txt


                          GROUPS OF PRIMARY UNITS

                    ********* GROUP 1 *********
                                      PRIMARY UNIT    SECONDARY
             PRIMARY UNIT IDENTIFICATION        SIZE         UNIVERSE
             ============================= =============  =============
             UNIV81                                17              76
             UNIV76                                10              47
             UNIV15                                 4              20
             UNIV20                                10              46
             UNIV59                                11              53
             UNIV13                                 4              19
             UNIV50                                10              49
             UNIV87                                15              71
             UNIV22  <--- Selected                13              64
             UNIV29                                10              49
             UNIV84                                17              77
             UNIV55                                15              71
             UNIV26                                 7              36
             UNIV66                                14              68
             UNIV38                                17              76
             UNIV14                                 9              44
             UNIV17                                 6              25

                  GROUP TOTALS: 17               189             891


              NAME OF OUTPUT FILE: C:\TEMP\OUTTEMP1.TXT

       FIRST SEED NUMBER:     100.00    SECOND SEED NUMBER:      200.00

       NUMBER OF PRIMARY UNITS IN THE POPULATION:    17
       NUMBER OF PRIMARY UNITS SAMPLED:               1

                                                                 UNITS
                        SECONDARY       PRIMARY                     IN
PRIMARY UNIT ID         UNIVERSE       UNIT SIZE      GROUP SIZE   GROUP
======================= =============  =============  ============= =====
UNIV22                        64             13            189      17
```

Next, repeat this for group 5. Remove UNIV51 from the population and this group. Construct

data file TEMP2.TXT.

NOTE: As before, be sure to switch columns 2 and 3 when building this file.

```
            UNIV63      31      7
            UNIV18      38      9
            UNIV58      57     11
            UNIV31      52     11
            UNIV56      59     12
            UNIV90      72     16
            UNIV65      32      7      File TEMP2.TXT
            UNIV12      56     11
            UNIV16      34      7
            UNIV2       21      4
            UNIV79      77     18
            UNIV74      71     15
            UNIV8       49     10
            UNIV86      75     17
            UNIV53      72     16
            UNIV23      45      9
            UNIV33      25      5
```

Again, run the **RHC Sample Selection** program. The input file is TEMP2.TXT and the sample

size is 1. This generates another P.U. from this group. The output from this program is shown

below. UNIV86 was selected.

```
                    DEPARTMENT OF HEALTH & HUMAN SERVICES
                       OIG - OFFICE OF AUDIT SERVICES
Date: 10/15/2004       GENERATION OF PRIMARY UNIT SAMPLE            Time: 13:24
                    NAME OF INPUT FILE: C:\TEMP\Temp2.txt

                         GROUPS OF PRIMARY UNITS

                    ********* GROUP 1 *********
                                            PRIMARY UNIT     SECONDARY
        PRIMARY UNIT IDENTIFICATION              SIZE        UNIVERSE
        ==============================      =============   =============
         UNIV18                                      9              38
         UNIV90                                     16              72
         UNIV63                                      7              31
         UNIV65                                      7              32
         UNIV31                                     11              52
         UNIV56                                     12              59
         UNIV23                                      9              45
         UNIV86   <--- Selected                     17              75
         UNIV53                                     16              72
         UNIV2                                       4              21
         UNIV33                                      5              25
```

< OUTPUT -- continued >

```
        UNIV79                                        18              77
        UNIV74                                        15              71
        UNIV12                                        11              56
        UNIV58                                        11              57
        UNIV16                                         7              34
        UNIV8                                         10              49

           GROUP TOTALS: 17                          185             866

              NAME OF OUTPUT FILE: C:\TEMP\OUTTEMP2.TXT

      FIRST SEED NUMBER:      100.00     SECOND SEED NUMBER:      200.00

      NUMBER OF PRIMARY UNITS IN THE POPULATION:     17
      NUMBER OF PRIMARY UNITS SAMPLED:                1
```

PRIMARY UNIT ID	SECONDARY UNIVERSE	PRIMARY UNIT SIZE	GROUP SIZE	UNITS IN GROUP
=========================	=============	=============	=============	=====
UNIV86	75	17	185	17

Finally, be sure to update the original output file shown earlier to reflect the two new selected universities. This is shown below. This file is one of the input files to the **RHC Two-Stage Appraisal** program.

Final output file (input file to RHC appraisal program)

```
        UNIV47                24          5            191              18
        UNIV64                46          9            175              18
        UNIV52                38          9            185              18
        UNIV22                64         13            189              17
        UNIV86                75         17            185              17
```

Discussion: RHC Three-Stage sampling

A similar procedure can be used to generate "spare" secondary units. For example, if one of the secondary units within a selected primary unit is nonusable, another secondary unit can be selected from this group using the procedure outlined above.

Discussion: Final-stage units

At the second stage for RHC Two-Stage sampling and the third stage for RHC Three-Stage

sampling, a random sample of units is obtained. Spares for this stage can be obtained in the

usual manner using the single-stage random number generator software (Single Stage Random

Numbers).

Comparison of RHC and Multistage SRS

In general, you can expect greater precision with the RHC procedure, <u>provided</u> there is a

significant correlation between the second and third columns (Number of Units and Size of Unit)

of each file using the RHC sample selection procedure. To illustrate, consider the file containing

the primary unit information used in the three-stage RHC illustration.

```
       (1)       (2)      (3)
   REGION1    117    1250
   REGION2     63     610      Columns:   (1) unit ID
   REGION3     91     720                 (2) number of units
     .                                    (3) size of unit
     .
     .
   REGION10    85     640
   REGION11    94     930
   REGION12    62     550
```

For this example, the correlation between Size of Unit and Number of Units is .958, and we

would expect a two-stage RHC procedure to work quite well. For a three-stage procedure, this

correlation rule must also apply within each of the sampled primary units, at the secondary unit

level.

The benefits of RHC sampling include:

- increased precision if the above correlation rule is satisfied;

- maintaining the flavor of pps sampling, since pps sampling is used to select a unit from each random group;

- relatively simple and straightforward computations;

- unbiased and stable point estimate of the universe total $(\hat{T})$. This implies that when sampling indefinitely, $\hat{T}$, on the average, is equal to the actual universe total, T, and $\hat{T}$ will exhibit relatively small variation.

- a staple point estimate of the variance of $\hat{T}$, producing more reliable confidence intervals. This implies that when sampling indefinitely, the lower confidence limits will exhibit relatively small variation.

ATTRIBUTE APPRAISALS

An attribute appraisal is carried out to estimate a particular universe <u>proportion</u> (p) and its

corresponding sampling error. This proportion is typically an error rate (proportion of the

universe in error) but, more generally, it is the proportion of the universe items that meet (or do

not meet) a specified set of criteria. Also of interest may be the <u>total</u> <u>number</u> of items in the

universe (Np) that meet the criteria.

In an attribute sample, each sample item is either a yes response (met the criteria) or no response

(did not meet the criteria). This version of RAT-STATS contains eight modules that can be used

to appraise an attribute sample. These sampling strategies are listed below and described in the

sections to follow.

- Unrestricted

- Stratified

- Two-Stage Unrestricted

- Three-Stage Unrestricted

- RHC Two Stage

- RHC Three Stage

- Stratified Cluster

- Stratified Multistage

Unrestricted Attribute Appraisal

An **unrestricted** sample is the same as a **simple random sample**. Consequently, every sample of size n has the same chance of being selected. For an unrestricted sample, a sample of size n is randomly obtained and the number of sample elements meeting the criteria (say, x) is recorded.

Example 1. An unrestricted sample of 400 documents was obtained and examined to determine if they had the proper approval signature. In the sample, 82 of the items did not contain the proper signature (were in error). The sample error rate is then 82/400 = .205 (i.e., 20.5%). This is the estimate of p, the error rate for the entire universe. If the universe size is N = 10,000, then the estimated number of universe items in error is (10,000)(.205) = 2,050 items.

Using the RAT-STATS software, the corresponding 90% confidence interval for the total number of universe items in error is from 1,729 to 2,403. The 90% confidence interval for the universe error rate (p) is from 17.29% to 24.03%. Notice that the (point) estimate of 20.5% is between 17.29% and 24.03% but it is not in the center of this interval. The center of the 90% confidence interval is (17.29 + 24.03)/2 = 20.66%. The reason for this result is that this estimation procedure is based on the exact hypergeometric distribution, rather than the normal approximation. The resulting 95% confidence interval for p is 16.73% to 24.70% and for Np (the total number of errors in the universe) is from 1,673 to 2,470.

Discussion. Consider the 90% confidence interval. Define

$$\text{TAIL} = (1 - .90) / 2 = .05$$

The 90% confidence interval for Np is, say, k_1 to k_2. There were x = 82 sample items in error, so

(referring to the Formulas section below) k_1 is the <u>smallest</u> value of k for which the probability

of observing 82 or fewer errors is > .05, where .05 is the value of TAIL. This value of k is

$k_1 = 1,729$. The corresponding error rate is $1,729/10,000 = .1729$ (i.e., 17.29%). To find the

upper limit of the 90% confidence interval, the program determines the <u>largest</u> value of k (say,

k_2) for which the probability of observing x = 82 or more errors is > TAIL = .05. This is $k_2 =$

2,403 with a corresponding error rate of $2,403/10,000 = .2403$ (i.e., 24.03%). A similar

argument applies to the 95% confidence interval, where now the value of TAIL is .025.

NOTES: 1. Using these definitions of k_1 and k_2 for a 90% confidence interval, the user can be assured that the actual confidence level is <u>at least</u> 90%. This also applies to 80% and 95% confidence intervals.

2. In the event that no items having the characteristic(s) of interest are found in the sample, the user has the option of having the program determine both confidence limits or only the upper confidence limits.

3. In the event that the number of items having the characteristic(s) of interest in the sample is the same as the sample size, the user has the option of having the program determine both confidence limits or only the lower confidence limits.

4. The universe size (N) is declared to be a long integer in the RAT-STATS program. Consequently, the largest allowable universe size is $N = 2^{31} - 1 =$ 2,147,483,647.

FORMULAS

To determine a 90% confidence interval for the total number of universe items in error, define

$TAIL = (1 - .90)/2 = .05$.

Upper Limit: Let k_2 = largest value of k for which

$$\sum_{i=0}^{x} \frac{\binom{k}{i}\binom{N-k}{n-i}}{\binom{N}{n}} > .05$$

where N = universe size

n = sample size

k = total number of universe items in error

x = number of sample items in error

Lower Limit: Let k_1 = smallest value of k for which

$$\sum_{i=x}^{n} \frac{\binom{k}{i}\binom{N-k}{n-i}}{\binom{N}{n}} > .05$$

The resulting 90% confidence interval for the total number of universe items in error is from k_1

to k_2 and the corresponding 90% confidence interval for the error rate (p) is k_1/N to k_2/N.

For a 95% confidence interval, use the same two equations, where .05 is replaced with

$TAIL = .025$. For an 80% confidence interval, the value of TAIL is .10.

The procedure used to derive this confidence interval can be found in the following article.

John P. Buonaccorsi (1987), "A Note on Confidence Intervals for Proportions in Finite

Populations," *The American Statistician*, Vol. 41, No. 3, 215-218.

Standard Errors

For universe proportion: Standard Error = $\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n-1}\left(1-\dfrac{n}{N}\right)}$ where $\hat{p}$ = x/n.

For universe total: Standard Error = $N \bullet \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n-1}\left(1-\dfrac{n}{N}\right)}$.

NOTE: RAT-STATS does not use the preceding standard errors when deriving a confidence interval for the universe proportion and universe total. Other software packages use this standard error to derive an <u>approximate</u> confidence interval based on the normal distribution. RAT-STATS derives an <u>exact</u> confidence interval based on the hypergeometric distribution.

Stratified Attribute Appraisal

In a stratified attribute sampling plan, the universe is divided into two or more nonoverlapping categories (strata). As with an unrestricted sample, the intent is to make a statistical estimate for a universe proportion (p) or a universe total (Np) that meets a specified set of criteria. This plan involves obtaining a random sample from each of the strata. The program will request the number of universe items in each stratum and these values <u>must</u> <u>be</u> <u>known</u>. The program will develop estimates for each stratum as well as for the entire universe.

NOTE: In the discussion to follow, we will refer to the proportion, p, as the "error rate."

Example 2. A universe of 2,500 Medicare claims is stratified into inpatient (Stratum 1) and outpatient (Stratum 2) claims. The universe sizes are $N_1 = 1,000$ inpatient claims and $N_2 = 1,500$ outpatient claims. Of interest is the proportion, p, of claims in error (containing improper charges).

A random sample of $n_1 = 100$ inpatient claims revealed $x_1 = 2$ errors and a random sample of $n_2 = 100$ outpatient claims uncovered $x_2 = 6$ errors.

NOTE: Both random samples were obtained using the **Single-Stage Random Numbers** program whereby 100 random numbers between 1 and 1,000 were obtained for stratum 1 and 100 random numbers between 1 and 1,500 were obtained for stratum 2.

The following output was obtained from the stratified attribute appraisal program.

```
                DEPARTMENT OF HEALTH & HUMAN SERVICES
                    OIG - OFFICE OF AUDIT SERVICES
Date: 2/7/2004   STRATIFIED ATTRIBUTE APPRAISAL        Time: 10:55
                AUDIT/REVIEW: Attribute - Stratified


                                                    PROJ. ITEMS
STRATUM      SAMPLE   *ITEMS**   **RATIO*  *UNIVERSE*  IN UNIVERSE
=======      ======   ========   ========  ==========  ===========

      1        100        2      2.000%      1,000           20
      2        100        6      6.000%      1,500           90

COMBINED       200        8      4.400%      2,500          110

STANDARD ERROR:                  1.483%         37


                                 PRECISION   PRECISION   PRECISION
STRATUM                          AT 80% CL   AT 90% CL   AT 95% CL
=======                          =========   =========   =========
      1                           1.711%      2.196%      2.616%
      2                           2.955%      3.793%      4.519%

COMBINED                          1.901%      2.439%      2.907%

LOWER LIMIT - QUANTITY               62          49          37
              PERCENT            2.499%      1.961%      1.493%
UPPER LIMIT - QUANTITY              158         171         183
              PERCENT            6.301%      6.839%      7.307%
```

Discussion. The strata sample error rates are 2% and 6%. The projected number of inpatient claims in error is $(.02)(1,000) = 20$ and the projected number for the outpatient stratum is $(.06)(1,500) = 90$. Consequently, the projected value for the universe is $20 + 90 = 110$ (highlighted) with a corresponding error rate of $(110/2,500) \times 100\% = 4.4\%$ (highlighted).

A look at the inpatient stratum: The estimated error rate is 2%. The corresponding precision at the 90% confidence level is 2.196% (highlighted). The term "precision" refers to the amount that is added and subtracted to the point estimate (2%, here) in deriving a confidence interval.

Consequently, the 90% confidence interval for the proportion of inpatient claims in error is 2% $\pm$ 2.196%; that is, -0.196% to 4.196%. Since the lower limit is negative, it may be set equal to zero. Similarly, the 95% confidence interval for the proportion of inpatient claims in error is 2% $\pm$ 2.616% (highlighted); that is, 0% to 4.616%, once again setting the lower limit equal to zero.

NOTE: These confidence intervals are not actually contained in the program output.

A look at the outpatient stratum: The estimated error rate is 6%. Continuing the discussion from the inpatient stratum, the 90% confidence interval for the proportion of outpatient claims in error is 6% $\pm$ 3.793% (highlighted); that is, 2.207% to 9.793%. The corresponding 95% confidence interval is 6% $\pm$ 4.519% (highlighted); that is, 1.481% to 10.519%. As before, these confidence intervals are not actually contained in the program output.

A look at the overall precision: The precision at the 90% level is 2.439% (highlighted) and so the resulting 90% confidence interval for the universe proportion of claims in error is 4.4% $\pm$ 2.439%; that is, 1.961% to 6.839%. Multiplying these two values by 2,500 (and dividing by 100), the corresponding 90% confidence interval for the total number of universe claims in error is 49 to 171. Notice that these values are rounded to the nearest integer. Using the precision at the 95% confidence level (i.e., 2.907%), the 95% confidence interval in the previous output can be obtained.

FORMULAS

The estimated proportion for stratum i is $\hat{p}_i$ where $\hat{p}_i = x_i / n_i$ and where x_i is the number of sample elements in stratum i in error and n_i is the number of sample items from stratum i. The value of Ratio is $\hat{p}_i$ x 100%. The Projected Items in Universe for stratum i is $(\hat{p}_i)(N_i)$ where N_i is the number of universe items in stratum i. The PRECISION AT 90% CL (CL stands for confidence level) is 1.644853626951 times the standard error of $\hat{p}_i$; that is

$$1.644853626951\sqrt{\frac{N_i - n_i}{N_i} \cdot \frac{\hat{p}_i(1 - \hat{p}_i)}{n_i - 1}}$$

To obtain the Precision at 95% (80%) CL value for the i-th stratum, replace 1.644853626951 with 1.959963984540 (1.281551565545). The estimated standard error of $\hat{p}_i$ is

$$SE(\hat{p}_i) = \sqrt{\frac{N_i - n_i}{N_i} \cdot \frac{\hat{p}_i(1 - \hat{p}_i)}{n_i - 1}}$$

Overall estimates: The estimate of the universe proportion (error rate) is $\hat{p}$ (under the Ratio heading), where

$$\hat{p} = \sum_{i=1}^{L} \left(\frac{N_i}{N}\right)\hat{p}_i$$

the summation is over all of the L strata and $N = \Sigma N_i$ is the total universe size. The estimated standard error of $\hat{p}$ is

$$SE(\hat{p}) = \sqrt{\sum_{i=1}^{L} \left(\frac{N_i}{N}\right)^2 [SE(\hat{p}_i)]^2}$$

The Precision at 90% CL value is $1.644853626951 \cdot SE(\hat{p})$.

The Precision at 95% CL value is obtained by replacing 1.644853626951 by 1.959963984540 in

the above formula and for an 80% confidence interval, 1.644853626951 is replaced by

1.281551565545.

The resulting confidence intervals for the universe proportion (error rate) are

$$\hat{p} \quad \pm (\text{PRECISION})$$

To obtain the confidence intervals for the universe total, multiply both ends of the confidence

interval for the error rate by the universe size, N, and round to the nearest integer.

Two-Stage Unrestricted

This is a special case of **multistage sampling**. Multistage sampling is a very cost-effective

sampling procedure when (1) obtaining a frame that lists all elements in the universe is

very costly or impossible, or (2) the cost of obtaining observations increases as the distance

separating the elements increases. Put another way, multistage sampling is cost effective

when it is more costly to get to the sampling unit than it is to audit the sampling unit. The goal

of multistage sampling is to get the most precise results per unit of examination cost.

General Comments

1. This is a <u>very</u> convenient sampling procedure for many situations because you don't have to
    visit all the locations.

2. For a two-stage procedure, the universe can be broken down into "subgroups."

    Example: 1st Stage: Carriers (P.U.s)

              2nd Stage: Hospitals (S.U.s)

    So, the procedure is to first obtain a random sample of P.U.s. These are called <u>clusters</u>.

    Then, obtain a random sample of S.U.s within each selected P.U.. Notice that at the first

    stage, <u>clusters</u> are the sampling unit (sampling units are not always <u>individual</u> people,

    records, etc.).

3. You can estimate cost overpayments for the entire universe with multistage sampling; it is
    very useful for large, widespread universes.

Example 3. In a particular region of the U.S. there are N = 90 universities with government

research grants. Because these universities are so widespread, it was decided to use a two-stage

sample using 10 universities. Rather than audit all grants at a selected university, it was decided

(based on available resources) to audit roughly 20% of the grants at each selected university to

estimate the proportion of grants containing charges after the scheduled completion of the grant.

The following data were obtained, where a_i (p_i) is the number (proportion) of grants in the

sample from the i-th university containing such charges, m_i is the number of audited (sampled)

grants at the i-th university, and M_i is the total number of grants in the audit universe at the i-th

university.

Univ.	M_i	m_i	a_i	p_i	$M_i p_i$
1	50	10	4	.400	20.00
2	65	13	5	.385	25.00
3	45	9	2	.222	10.00
4	48	10	3	.300	14.40
5	52	10	5	.500	26.00
6	58	12	3	.250	14.50
7	42	8	3	.375	15.75
8	66	13	4	.308	20.31
9	40	8	2	.250	10.00
10	56	11	4	.364	20.36
	522	104			176.32

Define M to be the total number of secondary units (grants) in the universe. In practice, M may

be known or unknown. If M is unknown, the **Two-Stage Unrestricted** program estimates the

universe proportion using a ratio estimator. No estimate of the universe total (total number of

grants containing improper charges) is available. This is illustrated in the computer output to

follow where M is unknown.

```
                    DEPARTMENT OF HEALTH & HUMAN SERVICES
                        OIG - OFFICE OF AUDIT SERVICES
Date: 1/31/2004    TWO STAGE UNRESTRICTED ATTRIBUTE APPRAISAL    Time: 14:13
                           AUDIT/REVIEW: Example
                      DATA FILE: C:\Temp\DATA2STG.TXT
```

PRIMARY UNIT	UNIVERSE	SAMPLE SIZE	SAMPLE ITEMS WITH CHARACTERISTIC(S)	RATIO
=======	============	===========	==================	=====
1	50	10	4	40.00%
2	45	9	2	22.22%
3	52	10	5	50.00%
4	42	8	3	37.50%
5	40	8	2	25.00%
6	65	13	5	38.46%
7	48	10	3	30.00%
8	58	12	3	25.00%
9	66	13	4	30.77%
10	56	11	4	36.36%
TOTALS	522	104	35	

```
TOTAL PRIMARY UNITS IN THE UNIVERSE                        90

OVERALL RATIO                                             33.78%

STANDARD ERROR                                            2.85%
```

CONFIDENCE LEVEL	80 PERCENT	90 PERCENT	95 PERCENT
LOWER LIMIT FOR PROPORTION	30.13%	29.09%	28.19%
UPPER LIMIT FOR PROPORTION	37.43%	38.47%	39.36%

Discussion. The estimate of the universe proportion, p, is

$\hat{p}_r$ = (estimated number of grants containing improper charges in the sampled universities) / (number of grants in the sampled universities)

$= [(50)(.4000) + (45)(.2222) + \cdots + (56)(.3636)] / (50 + 45 + \cdots + 56)$

$= 176.32 / 522 = .3378$ (that is, 33.78%)

The estimated variance of $\hat{p}_r$ is

$v(\hat{p}_r) = $ (Standard Error)$^2 = (.0285)^2 = .000812.$

NOTE: There is a formula for $v(\hat{p}_r)$ in the formula section.

The corresponding approximate 95% confidence interval for the universe proportion is

.3378 ± (Precision at 95% Confidence Level)

The Precision at 95% Confidence Level value is the same as 1.959963984540

times the (Standard Error). So, the resulting 95% confidence interval can also be written

.3378 ± (1.959963984540)(Standard Error)

.3378 ± (1.959963984540)(.0285)

.3378 ± .0559

.2819 to .3937 (28.19% to 39.36%).

NOTE 1: When the value of M is unknown in the formula for $v(\hat{p}_r)$ (the case here), it is acceptable to replace this value with the average value of M for the sample (as was done in this illustration). This value is 522/10 = 52.2. This is an advantage of using this estimator, since it does not require knowledge of M. If **M is known**, the user has two choices: (1) use the above ratio estimator, where now M is known or (2) use an <u>unbiased</u> estimator of p, illustrated in Example 4.

NOTE 2: If the value of M is known, the RAT-STATS software uses the unbiased estimator, illustrated in Example 4.

Example 4. Suppose that in Example 3, it is known that there is a total number of M = 4,500

grants (secondary units) in all 90 universities. As a result, $\overline{M}$ is known and is equal to M/N =

4,500/90 = 50. The following output is obtained. Notice that estimated (projected) totals for

each sampled university (primary unit) and for the entire universe are provided.

```
                    DEPARTMENT OF HEALTH & HUMAN SERVICES
                      OIG - OFFICE OF AUDIT SERVICES
Date: 1/31/2004   TWO STAGE UNRESTRICTED ATTRIBUTE APPRAISAL    Time: 13:52
                          AUDIT/REVIEW: Example
                     DATA FILE: C:\Temp\DATA2STG.TXT

PRIMARY                                SAMPLE ITEMS WITH
 UNIT       UNIVERSE     SAMPLE SIZE   CHARACTERISTIC(S)   RATIO      PROJECTED
=======   ============   ===========   =================   =====      =========

   1              50          10               4          40.00%          20
   2              45           9               2          22.22%          10
   3              52          10               5          50.00%          26
   4              42           8               3          37.50%          16
   5              40           8               2          25.00%          10
   6              65          13               5          38.46%          25
   7              48          10               3          30.00%          14
   8              58          12               3          25.00%          15
   9              66          13               4          30.77%          20
  10              56          11               4          36.36%          20

TOTALS           522         104              35

OVERALL TOTALS
  90           4,500                                      35.26%       1,587

STANDARD ERROR                                            3.67%          165

CONFIDENCE LEVEL                80 PERCENT      90 PERCENT      95 PERCENT
LOWER LIMIT FOR PROPORTION        30.56%          29.22%          28.06%
UPPER LIMIT FOR PROPORTION        39.97%          41.31%          42.47%

LOWER LIMIT FOR TOTAL             1,375           1,315           1,263
UPPER LIMIT FOR TOTAL             1,799           1,859           1,911
```

Discussion. An unbiased estimator of the universe proportion, p, is $\left(\dfrac{N}{M}\right)\left(\dfrac{A}{B}\right)$

where A = estimated number of grants containing improper charges in the sampled universities

      B = number of sampled universities

and A/B = the projected average number of grants containing improper charges for the sampled universities

So, A/B = [(50)(.4000) + (65)(.3846) + ⋯ + (56)(.3636)] / 10

      = 176.32 / 10 = 17.632 grants

The projected <u>number</u> of grants containing improper charges for the universe is $N \cdot (A/B)$; that is,

$(90)(17.632) = 1586.88$ grants. Since there are 4500 grants in the universe, then the estimated

<u>proportion</u> of grants containing improper charges is $1586.88/4,500 = .3526$ (35.26%).

The corresponding approximate 95% confidence interval for the universe proportion is

$\qquad$ .3526 $\pm$ (Precision at 95% Confidence Level),

which is the same as .3526 $\pm$ 1.959963984540(Standard Error). So, the resulting 95%

confidence interval is

$\qquad$ .3526 $\pm$ (1.959963984540)(Standard Error)

$\qquad$ .3526 $\pm$ (1.959963984540)(.0367)

$\qquad$ .3526 $\pm$ .0720

$\qquad$ .2806 to .4247 (28.06% to 42.47%).

The corresponding 95% confidence interval for the <u>total</u> <u>number</u> of grants in the universe

containing improper charges is

$\qquad$ 1587 $\pm$ 324; that is, 1263 to 1911 grants.

NOTE: Formulas for $\hat{p}_u$ and the corresponding confidence interval are contained in the
formula section.

FORMULAS

<u>Case 1</u>: When the total number of secondary units in the universe (M) is unkown, the ratio

estimator for the universe proportion is used. This estimator will be called $\hat{p}_r$. Define:

$\quad M_i$ $=$ number of secondary units in the universe for the i-th sampled primary unit, m_i of
which are sampled

$\hat{p}_i$ = proportion of secondary units having the attribute of interest in the i-th sampled primary unit

n = number of sampled primary units

N = number of primary units in the universe (must be known)

M = number of secondary units in the universe (may be known or unknown)

$\overline{M}$ = average number of secondary units per primary units in the universe. This is equal to M/N if M is known. It can be estimated using $\overline{m}$ if M is unknown, where $\overline{m}$ is the average number of secondary units in the sampled primary units.

- -

The estimate of the universe proportion having the attribute of interest is

$$\hat{p}_r = \frac{\sum_{i=1}^{n} M_i \hat{p}_i}{\sum_{i=1}^{n} M_i}$$

The estimated variance of $\hat{p}_r$ is

$$v(\hat{p}_r) = \left(\frac{N-n}{N}\right)\left(\frac{1}{n\overline{M}^2}\right)\left(\frac{\sum_{i=1}^{n} M_i^2 (\hat{p}_i - \hat{p}_r)^2}{n-1}\right) + \frac{1}{nN\overline{M}^2}\sum_{i=1}^{n} M_i^2 \left(\frac{M_i - m_i}{M_i}\right)\frac{\hat{p}_i(1-\hat{p}_i)}{m_i - 1}$$

NOTE: The standard error of $\hat{p}_r$ is the square root of $v(\hat{p}_r)$.

<u>Case 2</u>: When the total number of secondary units in the universe (M) is known, an unbiased

estimator for the universe proportion is used. This estimator will be called $\hat{p}_u$.

$$\hat{p}_u = \frac{N}{M} \frac{\sum\limits_{i=1}^{n} M_i \hat{p}_i}{n}$$

The estimated variance of $\hat{p}_u$ is

$$v(\hat{p}_u) = \left(\frac{N-n}{N}\right)\left(\frac{1}{n\overline{M}^2}\right)\left(\frac{\sum\limits_{i=1}^{n}(M_i\hat{p}_i - \overline{M}\hat{p}_u)^2}{n-1}\right) + \frac{1}{nN\overline{M}^2}\sum\limits_{i=1}^{n} M_i^2\left(\frac{M_i - m_i}{M_i}\right)\frac{\hat{p}_i(1-\hat{p}_i)}{m_i - 1}$$

NOTE 1: The standard error of $\hat{p}_u$ is the square root of $v(\hat{p}_u)$.

NOTE 2: When estimating the total number of secondary units in the universe having the
attribute of interest, both $\hat{p}_u$ and the standard error of $\hat{p}_u$ are multiplied by M.
The Precision at the 95% Confidence Level value for the universe total is
(1.959963984540)(M)(standard error of $\hat{p}_u$). For the Precision at the 90%
Confidence Level value, replace 1.959963984540 with 1.644853626951 and for the
Precision at the 80% Confidence Level value, replace 1.959963984540 with
1.281551565545.

Three-Stage Unrestricted

Example 5. The situation discussed in Example 4 was extended the following year to a three

stage procedure by defining:

    Stage 1: REGION (select 4 out of 12 regions)

    Stage 2: UNIVERSITY (select 10 from each selected region)

    Stage 3: GRANT (select approximately 20% of all grants at each university)

Using the random number module (**Single-Stage Random Numbers**), regions 5, 7, 8, and 10

were selected as the sampled primary units. Next, 10 universities (secondary units) were

randomly selected (again using program **Single- Stage Random Numbers**) from the available

universities in each of the four selected regions. The following data were obtained, where M_i is

the number of grants in the universe for each university, m_i is the number of audited grants at

each university (chosen to be roughly 20% of M_i), and a_i is the number of grants in the sample

from the i-th university containing charges after the scheduled completion of the grant (in error).

<u>REGION 5</u> (contains 90 secondary units, 10 to be sampled)

Univ.	M_i	m_i	a_i
1	47	9	3
2	51	10	2
3	45	9	4
4	46	9	1
5	46	9	3
6	50	10	1
7	50	10	4
8	57	11	3
9	54	11	4
10	64	13	2

REGION 7 (contains 110 secondary units, 10 to be sampled)

Univ.	M_i	m_i	a_i
1	53	11	2
2	59	12	5
3	52	10	1
4	67	13	3
5	59	12	1
6	73	15	6
7	51	10	3
8	75	15	2
9	66	13	1
10	58	12	4

REGION 8 (contains 85 secondary units, 10 to be sampled)

Univ.	M_i	m_i	a_i
1	45	9	3
2	39	8	2
3	43	9	4
4	34	7	1
5	54	11	2
6	54	11	3
7	34	7	1
8	59	12	1
9	49	10	4
10	43	9	2

REGION 10 (contains 120 secondary units, 10 to be sampled)

Univ.	M_i	m_i	a_i
1	59	12	2
2	68	14	6
3	57	11	3
4	72	14	6
5	70	14	1
6	73	15	2
7	83	17	5
8	89	18	4
9	73	15	3
10	77	15	2

The resulting data set in called DATA3ST.TXT and is shown on the next page. The

corresponding computer output using the **Three-Stage Unrestricted** program immediately

follows. For this illustration, the total number of third-stage units in the universe (S) is

unknown.

--- Data set DATA3ST.TXT ---

```
REGION 5    90    10 ───────
UNIV1    47    9    3
UNIV2    51   10    2
UNIV3    45    9    4
UNIV4    46    9    1
UNIV5    46    9    3
UNIV6    50   10    1
UNIV7    50   10    4
UNIV8    57   11    3
UNIV9    54   11    4
UNIV10   64   13    2
REGION 7   110    10
UNIV1    53   11    2 ───────
UNIV2    59   12    5
UNIV3    52   10    1
UNIV4    67   13    3
UNIV5    59   12    1
UNIV6    73   15    6
UNIV7    51   10    3
UNIV8    75   15    2
UNIV9    66   13    1
UNIV10   58   12    4
REGION 8    85    10
UNIV1    45    9    3
UNIV2    39    8    2
UNIV3    43    9    4
UNIV4    34    7    1
UNIV5    54   11    2
UNIV6    54   11    3
UNIV7    34    7    1
UNIV8    59   12    1
UNIV9    49   10    4
UNIV10   43    9    2
REGION 10  120    10
UNIV1    59   12    2
UNIV2    68   14    6
UNIV3    57   11    3
UNIV4    72   14    6
UNIV5    70   14    1
UNIV6    73   15    2
UNIV7    83   17    5
UNIV8    89   18    4
UNIV9    73   15    3
UNIV10   77   15    2
```

There are 90 secondary units (universities) in this primary unit (region); 10 were audited.

In UNIV1, there were 53 grants (third-stage units). Eleven of these grants were sampled and two contained improper charges.

```
                    DEPARTMENT OF HEALTH & HUMAN SERVICES
                       OIG - OFFICE OF AUDIT SERVICES
Date: 1/31/2004        THREE STAGE ATTRIBUTE APPRAISAL              Time: 15:18
                            AUDIT/REVIEW: Example
                 NAME OF INPUT FILE: C:\Temp\DATA3ST.TXT
```

FIRST STAGE SECOND STAGE	NEXT STAGE UNIVERSE	SAMPLE SIZE	MEETING CRITERIA	RATIO
REGION 5	90	10		
UNIV1	47	9	3	33.33%
UNIV3	51	10	2	20.00%
UNIV3	45	9	4	44.44%
UNIV4	46	9	1	11.11%
UNIV5	46	9	3	33.33%
UNIV6	50	10	1	10.00%
UNIV7	50	10	4	40.00%
UNIV8	57	11	3	27.27%
UNIV9	54	11	4	36.36%
UNIV10	64	13	2	15.38%
TOTALS	510	101	27	
REGION 7	110	10		
UNIV1	53	11	2	18.18%
UNIV2	59	12	5	41.67%
UNIV3	52	10	1	10.00%
UNIV4	67	13	3	23.08%
UNIV5	59	12	1	8.33%
UNIV6	73	15	6	40.00%
UNIV7	51	10	3	30.00%
UNIV8	75	15	2	13.33%
UNIV9	66	13	1	7.69%
UNIV10	58	12	4	33.33%
TOTALS	613	123	28	
REGION 8	85	10		
UNIV1	45	9	3	33.33%
UNIV2	39	8	2	25.00%
UNIV3	43	9	4	44.44%
UNIV4	34	7	1	14.29%
UNIV5	54	11	2	18.18%
UNIV6	54	11	3	27.27%
UNIV7	34	7	1	14.29%
UNIV8	59	12	1	8.33%
UNIV9	49	10	4	40.00%
UNIV10	43	9	2	22.22%
TOTALS	454	93	23	
REGION 10	120	10		
UNIV1	59	12	2	16.67%
UNIV2	68	14	6	42.86%
UNIV3	57	11	3	27.27%
UNIV4	72	14	6	42.86%
UNIV5	70	14	1	7.14%
UNIV6	73	15	2	13.33%
UNIV7	83	17	5	29.41%
UNIV8	89	18	4	22.22%
UNIV9	73	15	3	20.00%

```
   UNIV10                              77        15        2    13.33%

      TOTALS                         721       145       34


OVERALL TOTALS                                  UNIVERSE    SAMPLED
======= ======                                  ========    =======
   FIRST STAGE                                      12          4
  SECOND STAGE                                     405{}        40
   THIRD STAGE                                    2,298{}      462
   SAMPLED ITEMS MEETING CRITERIA                             112
        {} UNIVERSE SIZES FOR THE SECOND AND THIRD STAGES REPRESENT
           THE UNIVERSES FOR THE SAMPLED PRIOR STAGE.

OVERALL POINT ESTIMATE OF THE PROPORTION               24.06%
OVERALL STANDARD ERROR (PROPORTION)                     1.35%

OVERALL POINT ESTIMATE OF UNIVERSE TOTAL               17,210
OVERALL STANDARD ERROR (TOTAL)                          2,415

CONFIDENCE LEVEL              80 PERCENT     90 PERCENT      95 PERCENT
LOWER LIMIT FOR PROPORTION      22.33%         21.84%          21.42%
UPPER LIMIT FOR PROPORTION      25.78%         26.27%          26.70%

LOWER LIMIT FOR TOTAL           14,115         13,238          12,477
UPPER LIMIT FOR TOTAL           20,304         21,182          21,942
```

Highlighted values:

    (1) 33.33% is 3/9

    (2) 405 = 90 (in Region 5) + 110 (in Region 7) + 85 (in Region 8) + 120 (in Region 10)

    (3) 2,298 = (47 + ⋯ + 64) in Region 5 + (53 + ⋯ + 58) in Region 7 + (45 + ⋯ + 43) in Region 8 + (59 + ⋯ + 77) in Region 10.

Discussion. Based on the preceding output, the following results were obtained:

(1) Estimate of the proportion of grants in the universe containing improper charges is .2406 (24.06%). This uses Equation 1 in the Formulas and Definitions section.

(2) The 95% confidence interval for the universe proportion is from .2142 to .2670.

(3) Estimate of the total number of grants in the universe containing improper charges is 17,210. This uses Equation 5 in the Formulas and Definitions section.

(4) The 95% confidence interval for the total number of grants containing improper charges is from 12,477 to 21,942.

The standard error of the proportion estimate is .0135 (1.35%) and the 95% confidence interval

is .2406 ± (1.959963984540)(.0135); that is, .2142 to .2670. The value of .0135 is found by

taking the square root of the value obtained using Equation 2. For the universe total, the

standard error is 2,415 (using the square root of Equation 6) and the 95% confidence interval is

17,210 ± (1.959963984540)(2,415); that is, 12,477 to 21,942.

Is the total number of third-stage units in the universe known or unknown?

Let S = the total number of third-stage units in the universe. Two cases will be considered:

Case 1: S is unknown.

Case 2: S is known.

For Case 1:

To estimate the proportion, p, use the ratio (biased) estimator $(\hat{p}_r)$. To estimate the

number in the population (T) having the attribute of interest, use the unbiased estimator

$(\hat{T}_u)$.

For Case 2:

To estimate the proportion, p, use the unbiased estimator ($\hat{p}_u$). To estimate the number

in the population (T) having the attribute of interest, use the unbiased estimator ($\hat{T}_u$).

NOTE: In the preceding example, the unbiased estimator $\hat{T}_u$ was used where $\hat{T}_u$ = 17,210.
Here S was unknown and so the proportion estimator ($\hat{p}_r$) is from case 1 (Equation 1
in the Formulas and Definitions section). The standard error of $\hat{p}_r$ (.0135) is the
square root of the value obtained from Equation 2.

FORMULAS

<u>Definitions</u>

$S =$ total number of third-stage units in the universe

$N =$ number of primary units in the universe

$n =$ number of primary units in the sample

$M_i =$ number of secondary units (universe) in i-th primary unit

$m_i =$ number of secondary units (sample) in i-th primary unit

$B_{ij} =$ number of third-stage units (universe) in j-th secondary unit within i-th primary unit

$b_{ij} =$ number of third-stage units (sample) in j-th secondary unit within i-th primary unit

$\hat{p}_{ij} =$ proportion of b_{ij} sampled third-stage units in error

Formulas for $\hat{p}_r$ (Case 1) and $\hat{p}_u$ (Case 2)

The ratio estimator $\hat{p}_r$ (Equation 1)

$$\hat{p}_r = \frac{\displaystyle\sum_{i=1}^{n} \frac{M_i}{m_i} \sum_{j=1}^{m_i} B_{ij}\hat{p}_{ij}}{\displaystyle\sum_{i=1}^{n} \frac{M_i}{m_i} \sum_{j=1}^{m_i} B_{ij}}$$

Estimated variance of $\hat{p}_r$ (Equation 2)

$$v(\hat{p}_r) = \frac{N-n}{n(n-1)N\bar{S}^2} \sum_{i=1}^{n} \left[\left(\hat{T}_i - \frac{\hat{T}_u}{N} \right) - \hat{R}\left(\hat{B}_i - \frac{\hat{B}_u}{N} \right) \right]^2$$

$$+ \frac{1}{nN\bar{S}^2} \sum_{i=1}^{n} \frac{M_i(M_i - m_i)}{m_i(m_i - 1)} \sum_{j=1}^{m_i} \left[\left(\hat{T}_{ij} - \frac{\hat{T}_i}{M_i} \right) - \hat{R}\left(B_{ij} - \frac{\hat{B}_i}{M_i} \right) \right]^2$$

$$+ \frac{1}{nN\bar{S}^2} \sum_{i=1}^{n} \frac{M_i}{m_i} \sum_{j=1}^{m_i} \frac{B_{ij}(B_{ij} - b_{ij})}{b_{ij} - 1} \hat{p}_{ij}(1 - \hat{p}_{ij})$$

where $\hat{T}_u = \dfrac{N}{n} \displaystyle\sum_{i=1}^{n} \hat{T}_i$

$\hat{T}_i = \dfrac{M_i}{m_i} \displaystyle\sum_{j=1}^{m_i} B_{ij} \hat{p}_{ij}$

$\hat{T}_{ij} = B_{ij} \hat{p}_{ij}$

$\hat{B}_u = \dfrac{N}{n} \displaystyle\sum_{i=1}^{n} \hat{B}_i$

$\hat{B}_i = \dfrac{M_i}{m_i} \displaystyle\sum_{j=1}^{m_i} B_{ij}$

$\bar{S} = \dfrac{S}{N}$

$\hat{R} = \dfrac{\hat{T}_u}{\hat{B}_u}$

Notes: (1) In Equation 2, $\hat{R} = \hat{p}_r$

(2) To estimate $\bar{S}$, use the sample estimate $\bar{s}$ where

$$\bar{s} = \dfrac{\hat{B}_u}{N} = \dfrac{1}{n} \sum_{i=1}^{n} \dfrac{M_i}{m_i} \sum_{j=1}^{m_i} B_{ij}$$

The unbiased estimator $\hat{p}_u$ (Equation 3)

$$\hat{p}_u = \dfrac{N}{nS} \sum_{i=1}^{n} \dfrac{M_i}{m_i} \sum_{j=1}^{m_i} B_{ij} \hat{p}_{ij}$$

Estimated variance of $\hat{p}_u$ (Equation 4)

This variance can most easily be determined by setting $\hat{R} = 0$ in equation 2. Consequently,

$$
v(\hat{p}_u) = \frac{N-n}{n(n-1)N\overline{S}^2} \sum_{i=1}^{n} \left(\hat{T}_i - \frac{\hat{T}_u}{N} \right)^2
$$

$$
+ \frac{1}{nN\overline{S}^2} \sum_{i=1}^{n} \frac{M_i(M_i - m_i)}{m_i(m_i - 1)} \sum_{j=1}^{m_i} \left(\hat{T}_{ij} - \frac{\hat{T}_i}{M_i} \right)^2
$$

$$
+ \frac{1}{nN\overline{S}^2} \sum_{i=1}^{n} \frac{M_i}{m_i} \sum_{j=1}^{m_i} \frac{B_{ij}(B_{ij} - b_{ij})}{b_{ij} - 1} \hat{p}_{ij}(1 - \hat{p}_{ij})
$$

The unbiased estimator $\hat{T}_u$ (Equation 5)

Since $\hat{T}_u = S \cdot \hat{p}_u$ then

$$
\hat{T}_u = \frac{N}{n} \sum_{i=1}^{n} \frac{M_i}{m_i} \sum_{j=1}^{m_i} B_{ij} \hat{p}_{ij}
$$

NOTE: The value of S is not needed here.

Estimated variance of $\hat{T}_u$ (Equation 6)

Since $v(\hat{T}_u) = S^2 \cdot v(\hat{p}_u)$, then

$$
v(\hat{T}_u) = \frac{N(N-n)}{n(n-1)} \sum_{i=1}^{n} \left(\hat{T}_i - \frac{\hat{T}_u}{N} \right)^2
$$

$$
+ \frac{N}{n} \sum_{i=1}^{n} \frac{M_i(M_i - m_i)}{m_i(m_i - 1)} \sum_{j=1}^{m_i} \left(\hat{T}_{ij} - \frac{\hat{T}_i}{M_i} \right)^2
$$

$$
+ \frac{N}{n} \sum_{i=1}^{n} \frac{M_i}{m_i} \sum_{j=1}^{m_i} \frac{B_{ij}(B_{ij} - b_{ij})}{b_{ij} - 1} \hat{p}_{ij}(1 - \hat{p}_{ij})
$$

NOTE: The value of S is not needed here.

RHC Two Stage

For a discussion on the motivation behind the RHC sampling procedure, refer to the RHC SAMPLE SELECTION section on page 1-11, contained in the RANDOM NUMBERS section of this manual. It provides a method of sample selection that allows sampling without replacement while "maintaining the flavor" of sampling using probability proportional to size. When the P.U.s are selected, the size of each P.U. is considered rather than obtaining a simple random sample of P.U.s.

The size of each P.U. is rather arbitrary and can be the number of people, dollars, beds (for hospitals), and so forth. In general, you can expect improved precision using the RHC procedure if there is a high correlation between the size of each P.U. and the number of S.U.s within each P.U.. In other words, P.U.s having a larger size should contain a larger number of S.U.s.

The P.U.s are selected using the RHC SAMPLE SELECTION program. A random sample is then obtained for each selected P.U. and the number of S.U.s having the attribute of interest (e.g., in error) is recorded.

Example 6. An audit was carried out for state-supported university grants in a particular region. The universe consisted of all charge vouchers recorded for these grants. Because these universities are so widespread, it was decided to employ a two-stage sample using three of the 27 state-supported universities. Rather than audit all the vouchers at a selected university, it was decided (based on available resources) to audit 250 vouchers at each selected university to estimate the proportion of vouchers containing improper charges. The universities (P.U.s) were

to be selected using the RHC procedure where the "size" of each university was the total grant

dollars awarded to that university.

The following file (RHC2STAGE.TXT) was constructed:


```
   (1)      (2)     (3)
  UNIV1    14928     8
  UNIV2    12454     4
  UNIV3    17404    13
  UNIV4    18700    16
  UNIV5    15989    11
  UNIV6    15046     9
  UNIV7    16696    11
  UNIV8    15754    10
  UNIV9    17404    13
  UNIV10   12100     4
  UNIV11   17522    13
  UNIV12   16578    11
  UNIV13   12218     4
  UNIV14   15164     9         Data file RHC2STAGE.TXT
  UNIV15   12336     4
  UNIV16   13986     7
  UNIV17   12925     6
  UNIV18   14457     9
  UNIV19   18464    16
  UNIV20   15400    10
  UNIV21   15164     9
  UNIV22   17522    13
  UNIV23   15282     9
  UNIV24   16461    11
  UNIV25   13396     7
  UNIV26   14222     7
  UNIV27   14693     9
```

Columns: (1) unit ID
               (2) number of vouchers
               (3) size of university (dollar amount of grants x $10,000)

Using the RHC SAMPLE SELECTION program, the following output is produced:

```
                    DEPARTMENT OF HEALTH & HUMAN SERVICES
                        OIG – OFFICE OF AUDIT SERVICES
Date: 10/22/2004       GENERATION OF PRIMARY UNIT SAMPLE         Time: 14:29
                 NAME OF INPUT FILE: C:\TEMP\RHC2STAGE.txt

                          GROUPS OF PRIMARY UNITS

                    ********* GROUP 1 *********
                                       PRIMARY UNIT    SECONDARY
     PRIMARY UNIT IDENTIFICATION          SIZE         UNIVERSE
     ============================== =============   =============
     UNIV27                                     9          14,693
     UNIV2                                      4          12,454
     UNIV6                                      9          15,046
     UNIV1                                      8          14,928
     UNIV7 <-- Selected                        11          16,696
     UNIV21                                     9          15,164
     UNIV4                                     16          18,700
     UNIV5                                     11          15,989
     UNIV16                                     7          13,986

         GROUP TOTALS:  9                      84         137,656

                    ********* GROUP 2 *********
                                       PRIMARY UNIT    SECONDARY
     PRIMARY UNIT IDENTIFICATION          SIZE         UNIVERSE
     ============================== =============   =============
     UNIV19                                    16          18,464
     UNIV20 <-- Selected                       10          15,400
     UNIV14                                     9          15,164
     UNIV26                                     7          14,222
     UNIV15                                     4          12,336
     UNIV18                                     9          14,457
     UNIV24                                    11          16,461
     UNIV10                                     4          12,100
     UNIV23                                     9          15,282

         GROUP TOTALS:  9                      79         133,886

                    ********* GROUP 3 *********
                                       PRIMARY UNIT    SECONDARY
     PRIMARY UNIT IDENTIFICATION          SIZE         UNIVERSE
     ============================== =============   =============
     UNIV17                                     6          12,925
     UNIV11                                    13          17,522
     UNIV12                                    11          16,578
     UNIV8 <-- Selected                        10          15,754
     UNIV3                                     13          17,404
     UNIV9                                     13          17,404
     UNIV22                                    13          17,522
     UNIV25                                     7          13,396
     UNIV13                                     4          12,218

         GROUP TOTALS:  9                      90         140,723
```

```
                         DEPARTMENT OF HEALTH & HUMAN SERVICES
                            OIG - OFFICE OF AUDIT SERVICES
Date: 10/22/2004       GENERATION OF PRIMARY UNIT SAMPLE          Time: 14:29
                NAME OF OUTPUT FILE: C:\TEMP\OutRHCsummary.txt

 FIRST SEED NUMBER:      100.00     SECOND SEED NUMBER:      200.00

     NUMBER OF PRIMARY UNITS IN THE POPULATION:    27
     NUMBER OF PRIMARY UNITS SAMPLED:               3

                                                                     UNITS
                         SECONDARY      PRIMARY                         IN
PRIMARY UNIT ID           UNIVERSE     UNIT SIZE     GROUP SIZE      GROUP
=========================  =============  =============  =============  =====
UNIV7                         16,696         11              84          9
UNIV20                        15,400         10              79          9
UNIV8                         15,754         10              90          9
```

The selected universities are UNIV7, UNIV8, and UNIV20. A sample of 250 vouchers is

obtained at each university with the following results:

	Number of	Number of
University	sampled vouchers	vouchers in error
UNIV7	250	8
UNIV20	250	12
UNIV8	250	5

This information is recorded in the data file (RHC2DATA.TXT) required by the appraisal

program (TWO-STAGE RHC) and is shown below:

```
     1   250    8
     2   250   12     <-- Data file RHC2DATA.TXT
     3   250    5
```

The final portion of the preceding output was stored by the sample selection program in file

RHC2PU.TXT. This file is shown below:

<u>Primary unit file RHC2PU.TXT</u>

```
UNIV7              16696          11          84            9
UNIV20             15400          10          79            9
UNIV8              15754          10          90            9
```

Using these two files, the following output is generated by the TWO-STAGE RHC program:

```
                    DEPARTMENT OF HEALTH & HUMAN SERVICES
                       OIG - OFFICE OF AUDIT SERVICES
Date: 10/22/2004          TWO STAGE RHC ATTRIBUTE APPRAISAL      Time: 14:40
                            AUDIT/REVIEW: Example

               NAME OF DATA FILE: C:\TEMP\RHC2DATA.TXT
             NAME OF PRIMARY UNIT FILE: C:\TEMP\RHC2PU.TXT
                OUTPUT FILE: C:\TEMP\OutRHC2attr.txt


            PRIMARY   SAMPLE   == ATTRIBUTE ==
             UNIT      SIZE     SAMPLE TOTAL
            =======   ======   =============
               1       250            8
               2       250           12
               3       250            5

            TOTALS     750           25
```

```
                                                                     UNITS
 P.U.                           SECONDARY     PRIMARY                  IN
 NBR   PRIMARY UNIT ID          UNIVERSE    UNIT SIZE    GROUP SIZE  GROUP
 ====  ========================= ============= ============= ============= =====
  1    UNIV7                       16,696          11            84       9
  2    UNIV20                      15,400          10            79       9
  3    UNIV8                       15,754          10            90       9

 TOTALS:                          47,850          31           253      27
```

```
     P.U.  SAMPLE                   SECONDARY    SIZES       POINT
     NBR    SIZE    SAMPLE MEAN     UNIVERSE     RATIO     ESTIMATE
     ====  ======  =============   ============= ========  =============
       1    250            .03       16,696       7.636      4,079.90
       2    250            .05       15,400       7.900      5,839.68
       3    250            .02       15,754       9.000      2,835.72

  TOTALS:   750                      47,850                 12,755.30
```

```
                    --- VARIANCE COMPONENTS ---

     P.U.           WITHIN              BETWEEN              TOTAL
     NBR           VARIANCE            VARIANCE            VARIANCE
     ====          ========            ========            ========
       1         260,846.68           32,187.19          293,033.87
       2         338,251.15        4,907,290.25        5,245,541.41
       3         173,034.95        3,618,097.64        3,791,132.59

  TOTALS:        772,132.78        8,557,575.09        9,329,707.87


 PRIMARY UNITS SAMPLED:                                             3
 PRIMARY UNITS NOT SAMPLED:                                        24
 PRIMARY UNITS IN POPULATION:                                      27

 PROJECTED QUANTITY IN UNIVERSE:                               12,755
 STANDARD ERROR:                                                3,054
```

```
CONFIDENCE LEVEL           80 PERCENT        90 PERCENT            95 PERCENT
LOWER LIMIT                     8,841             7,731                 6,769
UPPER LIMIT                    16,670            17,779                18,742
PRECISION AMOUNT                3,914             5,024                 5,987
PRECISION PERCENT              30.69%            39.39%                46.93%
Z-VALUE USED          1.281551565545    1.644853626951        1.959963984540
```

Final results: The point estimate of the total number of vouchers in error is 12,755 with a

corresponding standard error of 3,054. The resulting 90% confidence interval is from 7,731 to

17,779 vouchers. Notice that this a very wide confidence interval. The PRECISION PERCENT

is 39.39%, obtained by multiplying the standard error by 1.644853626951 and dividing by the

point estimate (expressed as a percentage); that is (100)(1.644853626951)(3054)/12755. In

general, this can be reduced by sampling a larger number of P.U.s.

Discussion

For this example, $\hat{p}_1 = 8/250 = .032$, $\hat{p}_2 = 12/250 = .048$, and $\hat{p}_3 = 5/250 = .020$. Referring to

the Formula section on the next page, the estimate for the total number of vouchers containing

improper charges is:

$\hat{T}$ = (84/11)(16696)(.032) + (79/10)(15400)(.048) + (90/10)(15754)(.02)

$= 4,079.90 + 5,839.68 + 2,835.72 = 12,755$ (rounded)

To determine the variance of $\hat{T}$, the first component of this variance is the "within variance"

equal to $V_2 = 772,132.78$. This accounts for the variation within the primary units (universities).

The larger variance component is the variation between the primary units measured by $V_1 =$

8,557,575.09. The total variance is $V_1 + V_2 = 9,329,707.87$ and the estimated standard error of

$\hat{T}$ is $\sqrt{9,329,707.87} = 3,054$.

The 95% confidence interval for T is $12,755 \pm (1.959963984540)(3,054)$; that is 6,769 to

18,742.

FORMULAS

Definitions

   1.   P.U. stands for primary unit and S.U. is secondary unit

   2. A_i = size of i-th P.U.

   3. S_i = (size of i-th P.U.)/(size of entire population) = A_i/(size of entire population)

   4. B_i = total size for i-th group

   5. π_i = (total size for i-th group)/(size of entire population) = B_i/(size of entire population)

   6.   N = number of P.U.s in the population

   7. N_i = number of P.U.s in the i-th group

   8.   n = number of P.U.s in the sample

   9. M_i = number of S.U.s in the i-th sampled P.U. (population)

 10. m_i = number of S.U.s in the i-th sampled P.U. (sample)

Estimator of population total (T)

$$\hat{T} = \sum_{i=1}^{n} \left(\frac{B_i}{A_i} \right) M_i \hat{p}_i$$

where $\hat{p}_i$ = proportion of m_i sampled S.U.s having the attribute of interest.

Estimated variance of $\hat{T}$

$$v(\hat{T}) = V_1 + V_2 \quad \text{where}$$

$$V_1 = \frac{\sum\limits_{i=1}^{n} N_i^2 - N}{N^2 - \sum\limits_{i=1}^{n} N_i^2} \sum\limits_{i=1}^{n} \pi_i \left(\frac{M_i \hat{p}_i}{S_i} - \hat{T} \right)^2$$

and

$$V_2 = \sum\limits_{i=1}^{n} \pi_i \frac{M_i}{S_i} \left(M_i - m_i \right) \frac{\hat{p}_i (1 - \hat{p}_i)}{m_i - 1}$$

NOTE: The estimated standard error of $\hat{T}$ is $\sqrt{v(\hat{T})}$.

Approximate 95% confidence interval for the population total (T)

$$\hat{T} \pm 1.959963984540 \sqrt{v(\hat{T})}$$

NOTE: For a 90% confidence interval, replace 1.959963984540 with 1.644853626951 and for an 80% confidence interval, replace 1.959963984540 with 1.281551565545.

RHC Three Stage

The RHC sampling procedure can used for a three-stage design.

The steps for such a procedure are the following:

1. A sample of primary units (clusters) is obtained as in the one- and two-stage procedures. The size of the primary units is considered for this sample, where pps sampling is used for each group of primary units.

2. A sample of secondary units is obtained within each chosen primary unit by partitioning the primary unit into random groups. The group sizes are chosen to be as nearly equal as possible. Using pps sampling, and the size of each secondary unit, one secondary unit is chosen from each of the secondary groups.

3. A random sample of third-stage units is obtained for each of the chosen secondary units. No attention is paid to "size" here. This is a random sample.

Prior to running the appraisal program, the user must run the RAT-STATS RHC SAMPLE

SELECTION program.

Example 7. The situation discussed in Example 5 in the THREE-STAGE UNRESTRICTED

section will be appraised using the RHC methodology. For this example, the stages are:

Stage 1: REGION (select 4 out of 12 regions)

Stage 2: UNIVERSITY (select 10 from each selected region)

Stage 3: GRANT (select approximately 20% of all grants at each university)

Selection of Primary Units

A file must be constructed containing (for each region) (1) the number of secondary units

(universities) in this region and (2) the size of this region (total dollars of grants). This file is

GRANTSPU.TXT. The selected regions are 4, 6, 8, and 10 and the output file created by the

RHC SAMPLE SELECTION program is GRANTSPUOUT.TXT.

Data set GRANTSPU.TXT

```
   (1)      (2)    (3)
REGION1    117    1250
REGION2     63     610
REGION3     91     720
REGION4    123    1320
REGION5    107    1160
REGION6    116    1240
REGION7    102     960
REGION8    118    1300
REGION9    122    1320
REGION10    85     640
REGION11    94     930
REGION12    62     550
```

NOTE: It is okay to set the number of S.U.s [column (2)] equal to one in this file. The actual number of S.U.s <u>must</u> be known for the selected P.U.s. The correct number of S.U.s must then be inserted into file GRANTSPUOUT.TXT (the highlighted values).

Columns: (1) unit ID

(2) number of universities (S.U.s)

(3) size (total grant dollar amount x $100,000)

--- Data set GRANTSPUOUT.TXT ---

```
REGION6          116      1240      3100      3
REGION4          123      1320      3410      3
REGION8          118      1300      3170      3
REGION10          85       640      2320      3
```

```
                      DEPARTMENT OF HEALTH & HUMAN SERVICES
                          OIG - OFFICE OF AUDIT SERVICES
Date: 10/22/2004      GENERATION OF PRIMARY UNIT SAMPLE          Time: 15:09
                      NAME OF INPUT FILE: C:\TEMP\GRANTSPU.TXT

                             GROUPS OF PRIMARY UNITS

                      ********* GROUP 1 *********
                                         PRIMARY UNIT    SECONDARY
          PRIMARY UNIT IDENTIFICATION        SIZE        UNIVERSE
          ============================= =============  =============
          REGION2                                 610             63
          REGION6 <-- Selected                  1,240            116
          REGION1                               1,250            117

             GROUP TOTALS:  3                   3,100            296

                      ********* GROUP 2 *********
                                         PRIMARY UNIT    SECONDARY
          PRIMARY UNIT IDENTIFICATION        SIZE        UNIVERSE
          ============================= =============  =============
          REGION4 <-- Selected                  1,320            123
          REGION5                               1,160            107
          REGION11                                930             94

             GROUP TOTALS:  3                   3,410            324

                      ********* GROUP 3 *********
                                         PRIMARY UNIT    SECONDARY
          PRIMARY UNIT IDENTIFICATION        SIZE        UNIVERSE
          ============================= =============  =============
          REGION12                                550             62
          REGION8 <-- Selected                  1,300            118
          REGION9                               1,320            122

             GROUP TOTALS:  3                   3,170            302

                      ********* GROUP 4 *********
                                         PRIMARY UNIT    SECONDARY
          PRIMARY UNIT IDENTIFICATION        SIZE        UNIVERSE
          ============================= =============  =============
          REGION3                                 720             91
          REGION7                                 960            102
          REGION10 <-- Selected                   640             85

             GROUP TOTALS:  3                   2,320            278

 FIRST SEED NUMBER:      100.00     SECOND SEED NUMBER:      200.00
```

In practice, do not set these seed values.

```
      NUMBER OF PRIMARY UNITS IN THE POPULATION:     12
      NUMBER OF PRIMARY UNITS SAMPLED:                4
```

< Program output - continued >

PRIMARY UNIT ID	SECONDARY UNIVERSE	PRIMARY UNIT SIZE	GROUP SIZE	UNITS IN GROUP
=========================	=============	=============	=============	=====
REGION6	116	1,240	3,100	3
REGION4	123	1,320	3,410	3
REGION8	118	1,300	3,170	3
REGION10	85	640	2,320	3

NOTE: This is file GRANTSPUOUT.TXT

Selection of Secondary Units

The input for three-stage RHC program can be greatly simplified if you only obtain information

for each **selected** primary unit (i.e., regions 4, 6, 8, and 10 here). The information consists of the

size of each secondary unit (here, university) and the number of third-stage units in the universe

for each secondary unit (it is acceptable to set these equal to one and change later). This input is

shown in files REGION4.TXT, REGION6.TXT, REGION8.TXT, and REGION10.TXT. Each

line in the files contains the number of third-stage units (grants) in the universe and the size of

that secondary unit (total grant $ x $100,000), in that order.

After each of these four files is the computer output using the RHC SAMPLE SELECTION

program. A sample of 10 universities is selected for each region. The results are:

REGION	UNIVERSITIES
4	85, 46, 7, 82, 30, 34, 27, 66, 65, 80
6	113, 43, 78, 104, 89, 112, 30, 65, 3, 99
8	112, 6, 77, 93, 75, 111, 62, 115, 70, 99
10	78, 43, 7, 73, 55, 33, 10, 59, 64, 39

The previous five program runs (one at the primary level and four at the secondary level) created

five output files. Using a word processor or spreadsheet, these files can be joined to form one of

the input files (the one containing primary/secondary unit information) for the three-stage RHC

program which calculates the confidence interval. The file for this example is

PUSURHC3.TXT.

Data set REGION4.TXT   < - - - continued - - - >   < - - - continued - - - >

(1)	(2)	(3)
UNIV1	52	11
UNIV2	37	9
UNIV3	38	9
UNIV4	20	5
UNIV5	69	15
UNIV6	69	15
UNIV7	77	17
UNIV8	32	7
UNIV9	49	10
UNIV10	73	15
UNIV11	21	5
UNIV12	62	13
UNIV13	55	11
UNIV14	59	12
UNIV15	55	11
UNIV16	36	8
UNIV17	51	11
UNIV18	26	7
UNIV19	25	6
UNIV20	73	15
UNIV21	71	15
UNIV22	47	10
UNIV23	34	8
UNIV24	25	6
UNIV25	39	9
UNIV26	49	10
UNIV27	76	16
UNIV28	21	5
UNIV29	33	8
UNIV30	54	11
UNIV31	45	10
UNIV32	74	16
UNIV33	69	14
UNIV34	50	10
UNIV35	29	7
UNIV36	56	12
UNIV37	64	14
UNIV38	66	14
UNIV39	63	14
UNIV40	57	12
UNIV41	71	15
UNIV42	45	10
UNIV43	21	5
UNIV44	46	10
UNIV45	48	10
UNIV46	44	9
UNIV47	71	15
UNIV48	67	14
UNIV49	23	6
UNIV50	54	11

UNIV51	62	13
UNIV52	52	11
UNIV53	56	11
UNIV54	70	15
UNIV55	41	9
UNIV56	65	14
UNIV57	76	16
UNIV58	30	7
UNIV59	75	16
UNIV60	27	7
UNIV61	36	8
UNIV62	61	13
UNIV63	58	12
UNIV64	61	13
UNIV65	62	14
UNIV66	76	16
UNIV67	71	15
UNIV68	34	8
UNIV69	62	13
UNIV70	23	6
UNIV71	28	7
UNIV72	46	10
UNIV73	62	14
UNIV74	67	14
UNIV75	25	6
UNIV76	24	6
UNIV77	57	12
UNIV78	44	10
UNIV79	73	16
UNIV80	70	15
UNIV81	45	10
UNIV82	52	11
UNIV83	34	8
UNIV84	59	12
UNIV85	54	11
UNIV86	31	7
UNIV87	69	14
UNIV88	22	6
UNIV89	47	10
UNIV90	57	12
UNIV91	31	7
UNIV92	73	15
UNIV93	52	11
UNIV94	22	6
UNIV95	22	6
UNIV96	29	7
UNIV97	56	12
UNIV98	74	16
UNIV99	43	9
UNIV100	57	12

UNIV101	34	8
UNIV102	28	7
UNIV103	73	15
UNIV104	65	14
UNIV105	68	14
UNIV106	28	7
UNIV107	55	11
UNIV108	37	9
UNIV109	54	11
UNIV110	47	10
UNIV111	44	9
UNIV112	24	6
UNIV113	50	10
UNIV114	52	11
UNIV115	66	14
UNIV116	50	10
UNIV117	66	14
UNIV118	34	8
UNIV119	73	16
UNIV120	37	8
UNIV121	42	9
UNIV122	59	12
UNIV123	45	11

NOTE: This file has 123 lines.

Columns: (1) unit ID   (2) number of grants   (3) size of university (grant amount x $100,000)

```
                        DEPARTMENT OF HEALTH & HUMAN SERVICES
                            OIG - OFFICE OF AUDIT SERVICES
Date: 10/25/2004      GENERATION OF SECONDARY UNIT SAMPLE       Time: 14:21
                  NAME OF INPUT FILE: C:\TEMP\REGION4.TXT

                        GROUPS OF SECONDARY UNITS

                    ********* GROUP 1 *********
                                        SECONDARY UNIT   3RD STAGE
            SECONDARY UNIT IDENTIFICATION       SIZE         UNIVERSE
            ============================== =============  =============
            UNIV57                                    16             76
            UNIV48                                    14             67
            UNIV35                                     7             29
            UNIV107                                   11             55
            UNIV85 <-- Selected                       11             54
            UNIV103                                   15             73
            UNIV86                                     7             31
            UNIV2                                      9             37
            UNIV81                                    10             45
            UNIV58                                     7             30
            UNIV36                                    12             56
            UNIV49                                     6             23

                GROUP TOTALS: 12                     125            576


                    ********* GROUP 2 *********
                                        SECONDARY UNIT   3RD STAGE
            SECONDARY UNIT IDENTIFICATION       SIZE         UNIVERSE
            ============================== =============  =============
            UNIV52                                    11             52
            UNIV6                                     15             69
            UNIV46 <-- Selected                        9             44
            UNIV69                                    13             62
            UNIV108                                    9             37
            UNIV44                                    10             46
            UNIV50                                    11             54
            UNIV121                                    9             42
            UNIV1                                     11             52
            UNIV43                                     5             21
            UNIV87                                    14             69
            UNIV39                                    14             63

                GROUP TOTALS: 12                     131            611
```

< GROUPS 3 THROUGH 9 ARE OMITTED HERE >

```
                    ********* GROUP 10 *********
                                        SECONDARY UNIT   3RD STAGE
            SECONDARY UNIT IDENTIFICATION       SIZE         UNIVERSE
            ============================== =============  =============
            UNIV53                                    11             56
            UNIV24                                     6             25
            UNIV42                                    10             45
            UNIV120                                    8             37
            UNIV105                                   14             68
            UNIV97                                    12             56
            UNIV119                                   16             73
            UNIV32                                    16             74
            UNIV80 <-- Selected                       15             70
```

```
          UNIV96                                     7              29
          UNIV13                                    11              55
          UNIV62                                    13              61
          UNIV59                                    16              75

                GROUP TOTALS: 13               155             724
```

DEPARTMENT OF HEALTH & HUMAN SERVICES
OIG – OFFICE OF AUDIT SERVICES
Date: 10/25/2004   GENERATION OF SECONDARY UNIT SAMPLE   Time: 14:21
NAME OF OUTPUT FILE: C:\TEMP\OutRegion4.txt

FIRST SEED NUMBER: 100.00   SECOND SEED NUMBER: 200.00

NUMBER OF SECONDARY UNITS IN THE POPULATION: 123
NUMBER OF SECONDARY UNITS SAMPLED: 10

SECONDARY UNIT ID	3RD STAGE UNIVERSE	SECONDARY UNIT SIZE	GROUP SIZE	UNITS IN GROUP
UNIV85	54	11	125	12
UNIV46	44	9	131	12
UNIV7	77	17	119	12
UNIV82	52	11	129	12
UNIV30	54	11	141	12
UNIV34	50	10	140	12
UNIV27	76	16	138	12
UNIV66	76	16	128	13
UNIV65	62	14	125	13
UNIV80	70	15	155	13

UNIV1	56	10
UNIV2	27	5
UNIV3	56	11
UNIV4	23	5
UNIV5	72	13
UNIV6	24	5
UNIV7	61	11
UNIV8	65	12
UNIV9	68	13
UNIV10	40	8
UNIV11	64	12
UNIV12	66	13
UNIV13	80	14
UNIV14	53	9
UNIV15	36	7
UNIV16	53	10
UNIV17	47	9
UNIV18	73	14
UNIV19	41	8
UNIV20	58	11
UNIV21	45	9
UNIV22	43	8
UNIV23	56	10
UNIV24	35	7
UNIV25	34	7
UNIV26	65	13
UNIV27	78	14
UNIV28	35	7
UNIV29	31	6
UNIV30	58	11
UNIV31	29	6
UNIV32	76	14
UNIV33	57	10
UNIV34	42	8
UNIV35	69	13
UNIV36	58	11
UNIV37	31	6
UNIV38	33	6
UNIV39	40	8
UNIV40	51	9
UNIV41	60	11
UNIV42	78	14
UNIV43	39	7
UNIV44	46	9
UNIV45	58	11
UNIV46	59	11
UNIV47	53	10
UNIV48	57	10
UNIV49	28	6
UNIV50	63	12
UNIV51	31	6
UNIV52	60	11
UNIV53	30	6
UNIV54	30	6
UNIV55	40	8
UNIV56	26	5
UNIV57	24	5
UNIV58	44	8

<-- continued -->

UNIV59	67	13
UNIV60	56	10
UNIV61	33	7
UNIV62	40	8
UNIV63	68	13
UNIV64	70	13
UNIV65	57	10
UNIV66	40	7
UNIV67	54	10
UNIV68	65	12
UNIV69	62	12
UNIV70	28	5
UNIV71	56	10
UNIV72	41	8
UNIV73	31	6
UNIV74	31	6
UNIV75	46	9
UNIV76	38	7
UNIV77	62	12
UNIV78	63	12
UNIV79	50	9
UNIV80	53	9
UNIV81	39	7
UNIV82	39	7
UNIV83	39	7
UNIV84	25	5
UNIV85	67	13
UNIV86	47	9
UNIV87	54	10
UNIV88	50	9
UNIV89	35	7
UNIV90	66	13
UNIV91	65	12
UNIV92	71	13
UNIV93	29	6
UNIV94	74	14
UNIV95	66	13
UNIV96	71	13
UNIV97	43	8
UNIV98	62	11
UNIV99	80	14
UNIV100	57	11
UNIV101	22	5
UNIV102	33	6
UNIV103	78	5
UNIV104	25	9
UNIV105	76	8
UNIV106	39	5
UNIV107	48	5
UNIV108	54	7
UNIV109	63	12
UNIV110	28	8
UNIV111	69	8
UNIV112	27	10
UNIV113	33	8
UNIV114	52	7
UNIV115	33	15
UNIV116	23	10

Data set REGION6.TXT

NOTE: This file has 116 lines.

```
                    DEPARTMENT OF HEALTH & HUMAN SERVICES
                        OIG – OFFICE OF AUDIT SERVICES
Date: 10/25/2004      GENERATION OF SECONDARY UNIT SAMPLE      Time: 13:57
                  NAME OF INPUT FILE: C:\TEMP\REGION6.txt

                       GROUPS OF SECONDARY UNITS

                  ********* GROUP 1 *********
                                    SECONDARY UNIT   3RD STAGE
     SECONDARY UNIT IDENTIFICATION        SIZE        UNIVERSE
     ============================== ============= =============
     UNIV52                                   11            60
     UNIV45                                   11            58
     UNIV32                                   14            76
     UNIV86                                    9            47
     UNIV113 <-- Selected                      8            33
     UNIV85                                   13            67
     UNIV109                                  12            63
     UNIV87                                   10            54
     UNIV2                                     5            27
     UNIV80                                    9            53
     UNIV53                                    6            30

          GROUP TOTALS: 11                   108           568

                  ********* GROUP 2 *********
                                    SECONDARY UNIT   3RD STAGE
     SECONDARY UNIT IDENTIFICATION        SIZE        UNIVERSE
     ============================== ============= =============
     UNIV33                                   10            57
     UNIV48                                   10            57
     UNIV6                                     5            24
     UNIV43 <-- Selected                       7            39
     UNIV68                                   12            65
     UNIV41                                   11            60
     UNIV46                                   11            59
     UNIV1                                    10            56
     UNIV40                                    9            51
     UNIV88                                    9            50
     UNIV36                                   11            58

          GROUP TOTALS: 11                   105           576

          < GROUPS 3 THROUGH 9 ARE OMITTED HERE >

                  ********* GROUP 10 *********
                                    SECONDARY UNIT   3RD STAGE
     SECONDARY UNIT IDENTIFICATION        SIZE        UNIVERSE
     ============================== ============= =============
     UNIV20                                   11            58
     UNIV22                                    8            43
     UNIV39                                    8            40
     UNIV111                                   8            69
     UNIV100                                  11            57
     UNIV29                                    6            31
     UNIV105                                   8            76
     UNIV79                                    9            50
     UNIV99 <-- Selected                      14            80
     UNIV13                                   14            80
     UNIV60                                   10            56
```

```
       UNIV54                                    6            30

          GROUP TOTALS: 12                     113           670
```

<div align="center">
DEPARTMENT OF HEALTH & HUMAN SERVICES
OIG – OFFICE OF AUDIT SERVICES
</div>

```
Date: 10/25/2004     GENERATION OF SECONDARY UNIT SAMPLE       Time: 13:57
             NAME OF OUTPUT FILE: C:\TEMP\OutRegion6.txt

 FIRST SEED NUMBER:     100.00    SECOND SEED NUMBER:      200.00

    NUMBER OF SECONDARY UNITS IN THE POPULATION:    116
    NUMBER OF SECONDARY UNITS SAMPLED:               10
```

SECONDARY UNIT ID	3RD STAGE UNIVERSE	SECONDARY UNIT SIZE	GROUP SIZE	UNITS IN GROUP
UNIV113	33	8	108	11
UNIV43	39	7	105	11
UNIV78	63	12	104	11
UNIV104	25	9	96	11
UNIV89	35	7	124	12
UNIV112	27	10	108	12
UNIV30	58	11	95	12
UNIV65	57	10	109	12
UNIV3	56	11	115	12
UNIV99	80	14	113	12

Data set REGION8.TXT   < - - - continued - - - >   < - - - continued - - - >

UNIV1	72	15		UNIV51	77	16		UNIV101	24	5
UNIV2	44	10		UNIV52	36	9		UNIV102	26	6
UNIV3	43	10		UNIV53	75	16		UNIV103	40	10
UNIV4	55	12		UNIV54	68	15		UNIV104	77	16
UNIV5	27	7		UNIV55	34	8		UNIV105	27	6
UNIV6	34	8		UNIV56	55	12		UNIV106	65	15
UNIV7	51	11		UNIV57	42	10		UNIV107	61	13
UNIV8	42	10		UNIV58	36	9		UNIV108	36	9
UNIV9	54	12		UNIV59	36	9		UNIV109	26	6
UNIV10	25	6		UNIV60	66	15		UNIV110	38	9
UNIV11	82	17		UNIV61	61	13		UNIV111	84	17
UNIV12	65	14		UNIV62	64	14		UNIV112	75	16
UNIV13	33	8		UNIV63	72	15		UNIV113	26	6
UNIV14	48	10		UNIV64	65	14		UNIV114	45	10
UNIV15	32	8		UNIV65	58	13		UNIV115	59	13
UNIV16	82	17		UNIV66	49	11		UNIV116	59	13
UNIV17	35	8		UNIV67	30	7		UNIV117	57	12
UNIV18	54	12		UNIV68	75	16		UNIV118	58	12
UNIV19	34	8		UNIV69	33	8				
UNIV20	62	14		UNIV70	65	14		**NOTE: This file has**		
UNIV21	26	6		UNIV71	55	12		**118 lines.**		
UNIV22	31	7		UNIV72	38	9				
UNIV23	58	13		UNIV73	36	9				
UNIV24	61	13		UNIV74	60	13				
UNIV25	61	14		UNIV75	52	11				
UNIV26	54	12		UNIV76	65	14				
UNIV27	53	11		UNIV77	49	10				
UNIV28	56	12		UNIV78	27	7				
UNIV29	57	12		UNIV79	48	10				
UNIV30	26	6		UNIV80	36	9				
UNIV31	25	5		UNIV81	66	15				
UNIV32	37	9		UNIV82	62	14				
UNIV33	79	16		UNIV83	70	15				
UNIV34	60	13		UNIV84	68	15				
UNIV35	57	12		UNIV85	53	11				
UNIV36	27	7		UNIV86	38	9				
UNIV37	31	7		UNIV87	35	8				
UNIV38	75	15		UNIV88	36	9				
UNIV39	26	6		UNIV89	26	6				
UNIV40	36	9		UNIV90	26	6				
UNIV41	36	9		UNIV91	51	11				
UNIV42	49	10		UNIV92	25	5				
UNIV43	83	17		UNIV93	54	11				
UNIV44	71	15		UNIV94	56	12				
UNIV45	31	7		UNIV95	81	17				
UNIV46	42	10		UNIV96	73	15				
UNIV47	62	14		UNIV97	44	10				
UNIV48	54	11		UNIV98	50	11				
UNIV49	31	7		UNIV99	60	13				
UNIV50	80	16		UNIV100	31	7				

```
                   DEPARTMENT OF HEALTH & HUMAN SERVICES
                        OIG – OFFICE OF AUDIT SERVICES
Date: 10/25/2004      GENERATION OF SECONDARY UNIT SAMPLE      Time: 14:03
                 NAME OF INPUT FILE: C:\TEMP\REGION8.TXT

                       GROUPS OF SECONDARY UNITS

                  ********* GROUP 1 *********
                                        SECONDARY UNIT    3RD STAGE
        SECONDARY UNIT IDENTIFICATION        SIZE         UNIVERSE
        ============================== =============  =============
        UNIV54                                    15             68
        UNIV46                                    10             42
        UNIV33                                    16             79
        UNIV86                                     9             38
        UNIV112 <-- Selected                      16             75
        UNIV85                                    11             53
        UNIV108                                    9             36
        UNIV87                                     8             35
        UNIV2                                     10             44
        UNIV55                                     8             34
        UNIV34                                    13             60

            GROUP TOTALS: 11                     125            564

                  ********* GROUP 2 *********
                                        SECONDARY UNIT    3RD STAGE
        SECONDARY UNIT IDENTIFICATION        SIZE         UNIVERSE
        ============================== =============  =============
        UNIV47                                    14             62
        UNIV50                                    16             80
        UNIV6 <-- Selected                         8             34
        UNIV44                                    15             71
        UNIV68                                    16             75
        UNIV42                                    10             49
        UNIV48                                    11             54
        UNIV1                                     15             72
        UNIV41                                     9             36
        UNIV89                                     6             26
        UNIV37                                     7             31

            GROUP TOTALS: 11                     127            590
```

< GROUPS 3 THROUGH 9 ARE OMITTED HERE >

```
                  ********* GROUP 10 *********
                                        SECONDARY UNIT    3RD STAGE
        SECONDARY UNIT IDENTIFICATION        SIZE         UNIVERSE
        ============================== =============  =============
        UNIV21                                     6             26
        UNIV23                                    13             58
        UNIV40                                     9             36
        UNIV110                                    9             38
        UNIV100                                    7             31
        UNIV30                                     6             26
        UNIV104                                   16             77
        UNIV81                                    15             66
        UNIV99 <-- Selected                       13             60
        UNIV13                                     8             33
        UNIV60                                    15             66
```

```
     UNIV56                                         12              55

         GROUP TOTALS: 12                          129             572
```

                    DEPARTMENT OF HEALTH & HUMAN SERVICES
                        OIG – OFFICE OF AUDIT SERVICES
Date: 10/25/2004   GENERATION OF SECONDARY UNIT SAMPLE   Time: 14:03
                NAME OF OUTPUT FILE: C:\TEMP\OutRegion8.txt

 FIRST SEED NUMBER: 100.00   SECOND SEED NUMBER: 200.00

     NUMBER OF SECONDARY UNITS IN THE POPULATION: 118
     NUMBER OF SECONDARY UNITS SAMPLED: 10

SECONDARY UNIT ID	3RD STAGE UNIVERSE	SECONDARY UNIT SIZE	GROUP SIZE	UNITS IN GROUP
UNIV112	75	16	125	11
UNIV6	34	8	127	11
UNIV7	51	11	120	12
UNIV93	54	11	136	12
UNIV75	52	11	126	12
UNIV111	84	17	134	12
UNIV62	64	14	123	12
UNIV115	59	13	137	12
UNIV70	65	14	143	12
UNIV99	60	13	129	12

Data set
REGION10.TXT

<--continued -->

Note: This file has 85 lines.

UNIV1	34	6
UNIV2	32	5
UNIV3	69	10
UNIV4	23	4
UNIV5	60	9
UNIV6	72	11
UNIV7	56	9
UNIV8	28	5
UNIV9	38	6
UNIV10	60	9
UNIV11	58	9
UNIV12	37	6
UNIV13	70	10
UNIV14	37	6
UNIV15	81	12
UNIV16	53	9
UNIV17	63	10
UNIV18	32	5
UNIV19	33	5
UNIV20	37	6
UNIV21	77	11
UNIV22	52	8
UNIV23	63	10
UNIV24	41	7
UNIV25	45	8
UNIV26	34	6
UNIV27	61	10
UNIV28	70	10
UNIV29	34	5
UNIV30	22	4
UNIV31	66	10
UNIV32	69	10
UNIV33	65	10
UNIV34	26	4
UNIV35	43	7
UNIV36	65	10
UNIV37	80	12
UNIV38	74	11
UNIV39	38	6
UNIV40	43	7
UNIV41	47	8
UNIV42	59	9
UNIV43	42	7
UNIV44	54	9
UNIV45	73	11

UNIV46	78	12
UNIV47	72	11
UNIV48	30	5
UNIV49	47	8
UNIV50	52	8
UNIV51	24	4
UNIV52	26	4
UNIV53	22	4
UNIV54	57	9
UNIV55	78	12
UNIV56	62	10
UNIV57	57	9
UNIV58	68	10
UNIV59	52	8
UNIV60	54	9
UNIV61	41	7
UNIV62	61	10
UNIV63	79	12
UNIV64	50	8
UNIV65	54	9
UNIV66	53	9
UNIV67	40	7
UNIV68	44	7
UNIV69	39	7
UNIV70	72	11
UNIV71	76	11
UNIV72	34	5
UNIV73	27	5
UNIV74	40	7
UNIV75	41	7
UNIV76	25	4
UNIV77	41	7
UNIV78	39	7
UNIV79	58	9
UNIV80	71	11
UNIV81	37	6
UNIV82	30	5
UNIV83	78	12
UNIV84	59	9
UNIV85	29	5

```
                    DEPARTMENT OF HEALTH & HUMAN SERVICES
                       OIG - OFFICE OF AUDIT SERVICES
Date: 10/25/2004     GENERATION OF SECONDARY UNIT SAMPLE      Time: 13:49
                   NAME OF INPUT FILE: C:\TEMP\REGION10.TXT

                        GROUPS OF SECONDARY UNITS

                   ********* GROUP 1 *********
                                      SECONDARY UNIT   3RD STAGE
     SECONDARY UNIT IDENTIFICATION         SIZE         UNIVERSE
     ============================== =============  =============
     UNIV44                                     9             54
     UNIV32                                    10             69
     UNIV77                                     7             41
     UNIV78 <-- Selected                        7             39
     UNIV2                                      5             32
     UNIV50                                     8             52
     UNIV34                                     4             26
     UNIV46                                    12             78

          GROUP TOTALS:  8                     62            391

                   ********* GROUP 2 *********
                                      SECONDARY UNIT   3RD STAGE
     SECONDARY UNIT IDENTIFICATION         SIZE         UNIVERSE
     ============================== =============  =============
     UNIV6                                     11             72
     UNIV43 <-- Selected                        7             42
     UNIV62                                    10             61
     UNIV41                                     8             47
     UNIV1                                      6             34
     UNIV40                                     7             43
     UNIV79                                     9             58
     UNIV36                                    10             65

          GROUP TOTALS:  8                     68            422

             < GROUPS 3 THROUGH 9 ARE OMITTED HERE >

                   ********* GROUP 10 *********
                                      SECONDARY UNIT   3RD STAGE
     SECONDARY UNIT IDENTIFICATION         SIZE         UNIVERSE
     ============================== =============  =============
     UNIV71                                    11             76
     UNIV9                                      6             38
     UNIV21                                    11             77
     UNIV23                                    10             63
     UNIV39 <-- Selected                        6             38
     UNIV29                                     5             34
     UNIV72                                     5             34
     UNIV13                                    10             70
     UNIV51                                     4             24

          GROUP TOTALS:  9                     68            454
```

```
                     DEPARTMENT OF HEALTH & HUMAN SERVICES
                        OIG - OFFICE OF AUDIT SERVICES
Date: 10/25/2004      GENERATION OF SECONDARY UNIT SAMPLE      Time: 13:49
                 NAME OF OUTPUT FILE: C:\TEMP\OutRegion10.txt

 FIRST SEED NUMBER:      100.00     SECOND SEED NUMBER:       200.00

     NUMBER OF SECONDARY UNITS IN THE POPULATION:    85
     NUMBER OF SECONDARY UNITS SAMPLED:              10
```

SECONDARY UNIT ID	3RD STAGE UNIVERSE	SECONDARY UNIT SIZE	GROUP SIZE	UNITS IN GROUP
UNIV78	39	7	62	8
UNIV43	42	7	68	8
UNIV7	56	9	54	8
UNIV73	27	5	63	8
UNIV55	78	12	70	8
UNIV33	65	10	77	9
UNIV10	60	9	76	9
UNIV59	52	8	71	9
UNIV64	50	8	73	9
UNIV39	38	6	68	9

--- Data set PUSURHC3.TXT ---

```
REGION4                123      1320      3410         3      10
UNIV85                  54        11       125        12
UNIV46                  44         9       131        12
UNIV7                   77        17       119        12
UNIV82                  52        11       129        12
UNIV30                  54        11       141        12
UNIV34                  50        10       140        12
UNIV27                  76        16       138        12
UNIV66                  76        16       128        13
UNIV65                  62        14       125        13
UNIV80                  70        15       155        13
REGION6                116      1240      3100         3      10
UNIV113                 33         8       108        11
UNIV43                  39         7       105        11
UNIV78                  63        12       104        11
UNIV104                 25         9        96        11
UNIV89                  35         7       124        12
UNIV112                 27        10       108        12
UNIV30                  58        11        95        12
UNIV65                  57        10       109        12
UNIV3                   56        11       115        12
UNIV99                  80        14       113        12
REGION8                118      1300      3170         3      10
UNIV112                 75        16       125        11
UNIV6                   34         8       127        11
UNIV7                   51        11       120        12
UNIV93                  54        11       136        12
UNIV75                  52        11       126        12
UNIV111                 84        17       134        12
UNIV62                  64        14       123        12
UNIV115                 59        13       137        12
UNIV70                  65        14       143        12
UNIV99                  60        13       129        12
REGION10                85       640      2320         3      10
UNIV78                  39         7        62         8
UNIV43                  42         7        68         8
UNIV7                   56         9        54         8
UNIV73                  27         5        63         8
UNIV55                  78        12        70         8
UNIV33                  65        10        77         9
UNIV10                  60         9        76         9
UNIV59                  52         8        71         9
UNIV64                  50         8        73         9
UNIV39                  38         6        68         9
```

NOTE: This is the data file constructed using the RHC SAMPLE SELECTION program to select the primary units (regions) and, within each selected primary unit, the 10 secondary units (universities). The four lines beginning with REGIONx are from the output file created during the primary unit selection (GRANTSPUOUT.TXT). A value of 10 (the number of selected universities for that region) is added to the end of each of these lines. The 10 lines after each REGIONx line consist of the output file created when selecting the universities from each region (OUTREGION4.TXT, . . ., OUTREGION10.TXT).

Selection of Third-Stage Units

Since approximately 20% of the grants at each university in the sample are to be audited, the

following sample sizes are determined:

```
Region 4:   University   Grants in universe   Number to be audited
            UNIV85             54                      11
            UNIV46             44                       9
            UNIV7              77                      15
            UNIV82             52                      10
            UNIV30             54                      11
            UNIV34             50                      10
            UNIV27             76                      15
            UNIV66             76                      15
            UNIV65             62                      12
            UNIV80             70                      14
                                                      122

Region 6:   University   Grants in universe   Number to be audited
            UNIV113            33                       7
            UNIV43             39                       8
            UNIV78             63                      13
            UNIV104            25                       5
            UNIV89             35                       7
            UNIV112            27                       5
            UNIV30             58                      12
            UNIV65             57                      11
            UNIV3              56                      11
            UNIV99             80                      16
                                                       95

Region 8:   University   Grants in universe   Number to be audited
            UNIV112            75                      15
            UNIV6              34                       7
            UNIV7              51                      10
            UNIV93             54                      11
            UNIV75             52                      10
            UNIV111            84                      17
            UNIV62             64                      13
            UNIV115            59                      12
            UNIV70             65                      13
            UNIV99             60                      12
                                                      120

Region 10:  University   Grants in universe   Number to be audited
            UNIV78             39                       8
            UNIV43             42                       8
            UNIV7              56                      11
            UNIV73             27                       5
            UNIV55             78                      16
            UNIV33             65                      13
            UNIV10             60                      12
            UNIV59             52                      10
            UNIV64             50                      10
            UNIV39             38                       8
                                                      101
```

The data file containing the errors for these 438 audited grants is RHC3DATA.TXT. Each line

contains (1) a counter, (2) the number of sampled (audited) secondary units (grants), and (3) the

number of grants containing improper charges (in error).

File RHC3DATA.TXT   **< - - - continued - - - >**

```
1.1    11   2              3.1    15   1
1.2     9   4              3.2     7   3
1.3    15   3              3.3    10   2
1.4    10   2              3.4    11   4
1.5    11   5              3.5    10   3
1.6    10   2              3.6    17   6
1.7    15   2              3.7    13   0        Note: This file
1.8    15   4              3.8    12   1        has 40 lines.
1.9    12   1              3.9    13   2
1.10   14   3              3.10   12   3
2.1     7   2              4.1     8   5
2.2     8   4              4.2     8   1
2.3    13   5              4.3    11   5
2.4     5   1              4.4     5   3
2.5     7   3              4.5    16   3    (*)
2.6     5   2              4.6    13   4
2.7    12   3              4.7    12   0
2.8    11   2              4.8    10   2
2.9    11   4              4.9    10   3
2.10   16   3              4.10    8   3
```

(*) To illustrate, the fifth university in the fourth sampled P.U. (Region 10) had 16 grants (third-

stage units) audited and three of them contained improper charges.

Finally, the three-stage RHC program is run to generate a confidence interval for the universe

total using input files PUSURHC3.TXT and RHC3DATA.TXT. The output from this program

is shown at the end of this section.

Summary of results

The estimate for the number of grants in error for the universe (all 12 regions) is the OVERALL

POINT ESTIMATE of 15,861 with a corresponding estimated OVERALL STANDARD

ERROR of 2,039 grants.

NOTE: This estimate does not require knowing the number of grants in the universe. If this
            value is known, you can convert the point estimate into a proportion. For example, if
            the total number of grants in all 12 regions is 59,200, then the point estimate for the
            proportion of grants in error is 15,861/59,200 = .268 (26.8%) with a corresponding
            standard error of 2,039/59,200 = .034 (3.4%).

The program also provides estimates for the number of grants in error for each sampled P.U.

(region) and for each of the groups of S.U.s (universities) within each sampled region. For

example, the estimated number of grants in error for Region 4 is 1,531 and the estimated number

of grants in error for the group of nine universities containing UNIV85 is 112. The SIZES

RATIO refers to the ratio of the size of the group containing this university to the size of this

university. To illustrate, UNIV85 in Region 4 has a size of 11 and is in a group of size 125 (look

at file REGION4.TXT). The SIZES RATIO here is 125/11 = 11.3636.

The 95% confidence interval for the number of grants in error is 11,865 to 19,857. If the total

number of third-stage units in the universe is known (say, 59,200), this interval can be converted

into an interval for the proportion of grants in error by dividing both limits by this value (here,

.200 to .335).

The PRECISION AMOUNT is the amount added and subtracted to the point estimate (15,861)

to obtain the corresponding confidence interval. In the 95% confidence interval, the lower limit

of 11,865 is obtained by subtracting the precision amount of 3,996 from 15,861. The

PRECISION PERCENT is the precision amount divided by the point estimate, expressed as a

percentage.

```
                     DEPARTMENT OF HEALTH & HUMAN SERVICES
                        OIG – OFFICE OF AUDIT SERVICES
Date: 10/26/2004     THREE STAGE RHC ATTRIBUTE APPRAISAL        Time: 10:04
                        AUDIT/REVIEW: RHC 3-Stage

                   DATA FILE USED: C:\TEMP\RHC3DATA.txt
           PRIMARY/SECONDARY UNIVERSE FILE USED: C:\TEMP\PUSURHC3.txt
                   OUTPUT FILE: C:\TEMP\OutRHC3.txt


                                        *** ATTRIBUTE ***
        **** SAMPLED UNITS ****      THIRD STAGE  SAMPLE    NO. WITH
PRIMARY / SECONDARY IDENTIFICATION     UNIVERSE    SIZE    ATTRIBUTE
================================== =========== ====== ============
REGION4
     UNIV85                                 54      11           2
     UNIV46                                 44       9           4
     UNIV7                                  77      15           3
     UNIV82                                 52      10           2
     UNIV30                                 54      11           5
     UNIV34                                 50      10           2
     UNIV27                                 76      15           2
     UNIV66                                 76      15           4
     UNIV65                                 62      12           1
     UNIV80                                 70      14           3
REGION6
     UNIV113                                33       7           2
     UNIV43                                 39       8           4
     UNIV78                                 63      13           5
     UNIV104                                25       5           1
     UNIV89                                 35       7           3
     UNIV112                                27       5           2
     UNIV30                                 58      12           3
     UNIV65                                 57      11           2
     UNIV3                                  56      11           4
     UNIV99                                 80      16           3
REGION8
     UNIV112                                75      15           1
     UNIV6                                  34       7           3
     UNIV7                                  51      10           2
     UNIV93                                 54      11           4
     UNIV75                                 52      10           3
     UNIV111                                84      17           6
     UNIV62                                 64      13           0
     UNIV115                                59      12           1
     UNIV70                                 65      13           2
     UNIV99                                 60      12           3
REGION10
     UNIV78                                 39       8           5
     UNIV43                                 42       8           1
     UNIV7                                  56      11           5
     UNIV73                                 27       5           3
     UNIV55                                 78      16           3
     UNIV33                                 65      13           4
     UNIV10                                 60      12           0
     UNIV59                                 52      10           2
```

```
    UNIV64                                  50        10            3
    UNIV39                                  38         8            3

        TOTALS                          2,270       438          111


                        --- POINT ESTIMATES ---
                          *** ATTRIBUTE ***
        **** SAMPLED UNITS ****                   SIZES        POINT
    PRIMARY / SECONDARY IDENTIFICATION  SAMPLE MEAN   RATIO      ESTIMATE
    ================================   ===========  ===========  =============
REGION4
    UNIV85  NOTE:  112 is the estimate    0.18      11.3636          112
    UNIV46  for the group containing      0.44      14.5556          285
    UNIV7   UNIV85 (not just UNIV85).     0.20       7.0000          108
    UNIV82  This group contained 12       0.20      11.7273          122
    UNIV30  universities shown earlier    0.45      12.8182          315
    UNIV34  in the output using           0.20      14.0000          140
    UNIV27  data set REGION4.TXT.         0.13       8.6250           87
    UNIV66                                0.27       8.0000          162
    UNIV65                                0.08       8.9286           46
    UNIV80                                0.21      10.3333          155

        TOTAL                        Estimate for Region 4 ➜ 1,531

REGION6
    UNIV113                               0.29      13.5000          127
    UNIV43                                0.50      15.0000          293
    UNIV78                                0.38       8.6667          210
    UNIV104                               0.20      10.6667           53
    UNIV89                                0.43      17.7143          266
    UNIV112                               0.40      10.8000          117
    UNIV30                                0.25       8.6364          125
    UNIV65                                0.18      10.9000          113
    UNIV3                                 0.36      10.4545          213
    UNIV99                                0.19       8.0714          121

        TOTAL                                                     1,638

REGION8
    UNIV112                               0.07       7.8125           39
    UNIV6                                 0.43      15.8750          231
    UNIV7                                 0.20      10.9091          111
    UNIV93                                0.36      12.3636          243
    UNIV75                                0.30      11.4545          179
    UNIV111                               0.35       7.8824          234
    UNIV62                                0.00       8.7857            0
    UNIV115                               0.08      10.5385           52
    UNIV70                                0.15      10.2143          102
    UNIV99                                0.25       9.9231          149

        TOTAL                                                     1,340

REGION10
    UNIV78                                0.63       8.8571          216
    UNIV43                                0.13       9.7143           51
    UNIV7                                 0.45       6.0000          153
    UNIV73                                0.60      12.6000          204
    UNIV55                                0.19       5.8333           85
    UNIV33                                0.31       7.7000          154
```

```
    UNIV10                                0.00      8.4444            0
    UNIV59                                0.20      8.8750           92
    UNIV64                                0.30      9.1250          137
    UNIV39                                0.38     11.3333          162

        TOTAL                                                    1,254
```

--- VARIANCE COMPONENTS FOR PRIMARY UNITS ---

```
 **** SAMPLED UNITS ****           WITHIN        BETWEEN        TOTAL
PRIMARY UNIT IDENTIFICATION       VARIANCE       VARIANCE       VARIANCE
============================== ============== ============== ==============
REGION4                             4,722        59,965         64,687
REGION6                             4,734        59,651         64,385
REGION8                             4,102        68,169         72,272
REGION10                            3,351        47,934         51,285
```
(Values of V_4)   (Values of V_3)

--- COMBINED VARIANCE COMPONENTS ---

```
       STAGE 1              STAGES 2 AND 3         TOTAL VARIANCE
====================== ====================== ======================

      3,466,892               690,212             4,157,104
```
(Value of V_1)   (Value of V_2)

*** ATTRIBUTE ***
--- SUMMARY OF APPRAISAL RESULTS ---
```
PRIMARY UNITS SAMPLED                           4
PRIMARY UNITS NOT SAMPLED                       8
TOTAL PRIMARY UNITS                            12

PROJECTED QUANTITY IN UNIVERSE             15,861
STANDARD ERROR                              2,039

CONFIDENCE LEVEL          80 PERCENT       90 PERCENT       95 PERCENT
LOWER LIMIT                   13,248           12,508           11,865
UPPER LIMIT                   18,474           19,215           19,857
PRECISION AMOUNT               2,613            3,354            3,996
PRECISION PERCENT             16.47%           21.14%           25.19%
Z-VALUE USED          1.281551565545   1.644853626951   1.959963984540
```

FORMULAS

<u>Definitions</u>

1. S_i = (size of i-th P.U.)/(size of entire population)

2. $\pi_i = \Sigma S_i$ over the i-th group of P.U.s

3. S_{ij} = (size of j-th S.U. in the i-th sampled P.U.)/(size of i-th sampled P.U.)

     (**Note**: denominator of S_{ij} = numerator of S_i)

4. $\pi_{ij} = \Sigma S_{ij}$ over the j-th group in i-th sampled P.U.

5.   N = number of P.U.s (population)

6.   n = number of P.U.s (sample)

7. M_i = number of S.U.s in i-th sampled P.U. (population)

8. m_i = number of S.U.s in i-th sampled P.U. (sample)

9. K_{ij} = number of records for j-th sampled S.U. in i-th sampled P.U. (population)

10. k_{ij} = number of records for j-th sampled S.U. in i-th sampled P.U. (sample)

<u>Estimator of population total (T)</u>

$$\hat{T} = \sum_{i=1}^{n} \pi_i \left(\frac{\hat{T}_i}{S_i} \right)$$

where $\hat{T}_i$ = estimator of total for i-th sampled P.U.

$$\sum_{j=1}^{m_i} \pi_{ij} \left(\frac{\hat{T}_{ij}}{S_{ij}} \right) \qquad\qquad \text{(equation 1)}$$

and $\hat{T}_{ij}$ = estimator of population total for j-th sampled S.U. in i-th sampled P.U.

$$= K_{ij} \hat{p}_{ij}$$

with $\hat{p}_{ij}$ = proportion of k_{ij} records having the attribute of interest

NOTE: It can be shown that $\hat{T}$ is an unbiased estimator of T.

<u>Estimated variance of $\hat{T}$</u>

$v(\hat{T}) = V_1 + V_2$ where

$$V_1 = \frac{\sum_{i=1}^{n} N_i^2 - N}{N^2 - \sum_{i=1}^{n} N_i^2} \sum_{i=1}^{n} \pi_i \left(\frac{\hat{T}_i}{S_i} - \hat{T} \right)^2 \qquad \text{(equation 2)}$$

and

$$V_2 = \sum_{i=1}^{n} \left(\frac{\pi_i}{S_i} \right) v(\hat{T}_i) \qquad \text{(equation 3)}$$

and N_i = number of P.U.s in the i-th group after the random split into n groups.

$v(\hat{T}_i)$ is obtained by applying the <u>two</u> stage RHC procedure within the i-th sampled P.U.; i.e., the i-th P.U. is viewed as the entire population. Consequently, $v(\hat{T}_i) = V_{3,i} + V_{4,i}$,where

$$V_{3,i} = \frac{\sum_{j=1}^{m_i} M_{ij}^2 - M_i}{M_i^2 - \sum_{j=1}^{m_i} M_{ij}^2} \sum_{j=1}^{m_i} \pi_{ij} \left(\frac{K_{ij} \hat{p}_{ij}}{S_{ij}} - \hat{T}_i \right)^2$$

and

$$V_{4,i} = \sum_{j=1}^{m_i} \pi_{ij} \frac{K_{ij}}{S_{ij}} (K_{ij} - k_{ij}) \frac{\hat{p}_{ij}(1 - \hat{p}_{ij})}{k_{ij} - 1}$$

and where (1) M_{ij} =   the number of S.U.s in the j-th random group, i-th sampled P.U.

(2) $\hat{p}_{ij}$ =   proportion of the k_{ij} items having the attribute of interest for the j-th sampled S.U. within the i-th sampled P.U.

Comments 1. V_1 is essentially the same expression obtained for the single-stage RHC

procedure and will be referred to as the "between unit" variation.

2. V_2 is the contribution of the second- and third-stage variation and is obtained by

treating each sampled P.U. as the population to be sampled using two

(additional) stages.

3. The estimated standard error of $\hat{T}$ is $\sqrt{v(\hat{T})}$.

Approximate 95% confidence interval for the population total (T)

$$\hat{T} \pm 1.959963984540\sqrt{v(\hat{T})}$$

NOTE: For a 90% confidence interval, replace 1.959963984540 with 1.644853626951
             and for an 80% confidence interval, replace 1.959963984540 with 1.281551565545.

Stratified Cluster

With this procedure, you first stratify, then obtain a cluster (single-stage) sample within each stratum. This is motivated by the discussion in the RAT-STATS User's Guide.

Example 8. In a large section of the U.S., an audit was conducted for 583 universities with health-related research grants. Two strata were defined:

     Stratum 1: state-supported universities and

     Stratum 2: private universities.

It was decided to estimate the proportion of contracts containing charges after the scheduled completion of the contract using the same two strata. The strata sizes were $N_1 = 415$ and $N_2 = 168$. Within each stratum, a **single-stage** cluster sample was obtained with $n_1 = 25$ universities selected from Stratum 1 (state supported universities) and $n_2 = 10$ universities from Stratum 2 (private universities). The total number of grants in the universe for Stratum 1 is 2,500 and for Stratum 2, the total number is 1,000 grants. Consequently, there are 3,500 grants in the entire universe.

NOTE 1: These sample sizes are not adequate according to OAS policy and are used here for illustration purposes only.

NOTE 2: This procedure <u>does</u> require knowledge of the total number of elements in the universe for each stratum.

The following data were obtained, where a_j is the number of grants containing charges after the scheduled completion of the grant for the j-th university, M_j is the number of grants (universe) for this university (all of which are audited), and p_j is a_j/M_j.

Summary Using Computer Output and Corresponding Formulas:

Stratum 1: $\Sigma a_j = 38$, $\Sigma M_j = 151$, $\hat{p}_1 = 38/151 = .2517$

The projected number in the universe for Stratum 1 is

$$\hat{T}_1 = (2500)(.2517) = 629$$

Stratum 2: $\Sigma a_j = 19$, $\Sigma M_j = 49$, $\hat{p}_2 = 19/49 = .3878$

The projected number in the universe for Stratum 2 is

$$\hat{T}_2 = (1000)(.3878) = 388$$

The estimate of the total number of grants in the universe with charges after the

scheduled grant completion is

$$\hat{T} = 629 + 388 = 1017$$

The estimated proportion of grants with such charges is

$$\hat{p} = 1017 / 3500 = .2905$$

The estimated standard error for $\hat{T}$ (using the square root of Equation 5) is 46.

The estimated standard error for $\hat{p}$ (using the square root of Equation 6) is .0132.

Stratum 1 -- State Universities

Univ.	M_j	a_j	p_j	Univ.	M_j	a_j	p_j
1	8	2	.25	14	10	3	.30
2	12	3	.25	15	9	1	.11
3	4	2	.50	16	3	1	.33
4	5	1	.20	17	6	2	.33
5	6	1	.17	18	5	1	.20
6	6	2	.33	19	5	1	.20
7	7	2	.29	20	4	1	.25
8	5	2	.40	21	6	1	.17
9	8	2	.25	22	8	1	.12
10	3	1	.33	23	7	2	.29
11	2	0	.00	24	3	1	.33
12	6	2	.33	25	8	2	.25
13	5	1	.20				

Stratum 2 -- Private Universities

Univ.	M_j	a_j	p_j	Univ.	M_j	a_j	p_j
1	2	1	.50	6	8	3	.37
2	5	2	.40	7	6	2	.33
3	7	2	.29	8	10	4	.40
4	4	2	.50	9	3	1	.33
5	3	1	.33	10	1	1	1.00

These results are combined into data set DATACLUS.TXT (39 lines)

Data set

```
STATE UNIVERSITIES  415  25
2500
UNIV1      8    2
UNIV2     12    3
UNIV3      4    2
UNIV4      5    1
UNIV5      6    1
UNIV6      6    2
UNIV7      7    2
UNIV8      5    2
UNIV9      8    2
UNIV10     3    1
UNIV11     2    0
UNIV12     6    2
UNIV13     5    1
UNIV14    10    3
UNIV15     9    1
UNIV16     3    1
UNIV17     6    2
```

```
UNIV18     5    1
UNIV19     5    1
UNIV20     4    1
UNIV21     6    1
UNIV22     8    1
UNIV23     7    2
UNIV24     3    1
UNIV25     8    2
PRIVATE UNIVERSITIES   168
10   1000
UNIV1      2    1
UNIV2      5    2
UNIV3      7    2
UNIV4      4    2
UNIV5      3    1
UNIV6      8    3
UNIV7      6    2
UNIV8     10    4
UNIV9      3    1
UNIV10     1    1
```

The following computer printout is produced:

```
              DEPARTMENT OF HEALTH & HUMAN SERVICES
                 OIG - OFFICE OF AUDIT SERVICES
Date: 2/16/2004     STRATIFIED CLUSTER ATTRIBUTE APPRAISAL        Time: 15:36
              AUDIT/REVIEW: Attribute - Stratified Cluster
               NAME OF INPUT FILE: C:\Temp\DATACLUS.TXT
```

STRATUM IDENTIFICATION CLUSTER IDENTIFICATION	SAMPLE UNIVERSE	SAMPLE SIZE	MEETING CRITERIA	PERCENT	PROJECTED QUANTITY
STATE UNIVERSITIES	415	25			
UNIV1	8	8	2		
UNIV2	12	12	3		
UNIV3	4	4	2		
UNIV4	5	5	1		
UNIV5	6	6	1		
UNIV6	6	6	2		
UNIV7	7	7	2		
UNIV8	5	5	2		
UNIV9	8	8	2		
UNIV10	3	3	1		
UNIV11	2	2	0		
UNIV12	6	6	2		
UNIV13	5	5	1		
UNIV14	10	10	3		
UNIV15	9	9	1		
UNIV16	3	3	1		
UNIV17	6	6	2		
UNIV18	5	5	1		
UNIV19	5	5	1		
UNIV20	4	4	1		
UNIV21	6	6	1		
UNIV22	8	8	1		

UNIV23	7	7	2	
UNIV24	3	3	1	
UNIV25	8	8	2	
STRATUM TOTALS	2,500	151	38	25.17% 629
PRIVATE UNIVERSITIES	168	10		
UNIV1	2	2	1	
UNIV2	5	5	2	
UNIV3	7	7	2	
UNIV4	4	4	2	
UNIV5	3	3	1	
UNIV6	8	8	3	
UNIV7	6	6	2	
UNIV8	10	10	4	
UNIV9	3	3	1	
UNIV10	1	1	1	
STRATUM TOTALS	1,000	49	19	38.78% 388

STRATA TOTALS	583	35		
CLUSTER UNIT TOTALS	3,500	200	57	
OVERALL TOTALS				29.05% 1,017
OVERALL STANDARD ERROR				1.32% 46

CONFIDENCE LEVEL	80 PERCENT	90 PERCENT	95 PERCENT
LOWER LIMIT FOR PROPORTION	27.36%	26.88%	26.47%
UPPER LIMIT FOR PROPORTION	30.74%	31.22%	31.64%
LOWER LIMIT FOR TOTAL	958	941	926
UPPER LIMIT FOR TOTAL	1,076	1,093	1,107

FORMULAS

1. Estimated proportion in stratum h that possess the attribute of interest

$$\hat{p}_h = \frac{\sum_{j=1}^{n_h} a_{j,h}}{\sum_{j=1}^{n_h} M_{j,h}}$$

where $a_{j,h}$ is the number of elements in the j-th secondary unit in stratum h possessing the attribute of interest, $M_{j,h}$ is the number of secondary units in the j-th primary unit in stratum h, and n_h is the number of sample items in stratum h.

2. Estimated total number of elements in stratum h that possess the attribute of interest

$$\hat{T}_h = M_h \hat{p}_h$$

where M_h = number of secondary units in the universe for stratum h (**must be known**)

3. Estimated universe total having the attribute of interest

$$\hat{T} = \sum_{h=1}^{L} \hat{T}_h \quad \text{summed over the L strata}$$

4. Estimated universe proportion having the attribute of interest is $\hat{p} = \hat{T} / M$ where M is the

    total number of secondary units in the universe and $M = \Sigma M_h$ (summed over the L strata).

5. Estimated variance of $\hat{T}$

$$v(\hat{T}) = \sum_{h=1}^{L} \frac{N_h(N_h - n_h)}{n_h(n_h - 1)} \sum_{j=1}^{n_h} (a_{j,h} - \hat{p}_h M_{j,h})^2$$

where N_h is the number of universe items in stratum h.

6. Estimated variance of $\hat{p}$

$$v(\hat{p}) = v(\hat{T}) / M^2$$

7. Approximate 95% confidence interval for T
$$\hat{T} \pm 1.959963984540\sqrt{v(\hat{T})}$$

8. Approximate 95% confidence interval for p

$$\hat{p} \pm 1.959963984540\sqrt{v(\hat{p})}$$

NOTE: For the Precision at the 90% Confidence Level, replace 1.959963984540 with
            1.644853626951 and for the Precision at the 80% Confidence Level,
            replace1.959963984540 with 1.281551565545.

Stratified Multistage

As with the stratified cluster procedure, you must first stratify the universe. Rather than take a cluster (single-stage) sample within each stratum, you will obtain a multistage (two-stage or three-stage) sample within each stratum. These multistage samples may be random (using the **Two-Stage Unrestricted** or **Three-Stage Unrestricted** programs) or may be obtained using the RHC procedure and the **RHC Two-Stage** or **RHC Three-Stage** programs.

Unlike the **Stratified Cluster** program, this program requires that you first run the appropriate multistage program <u>on</u> <u>each</u> <u>stratum</u> and record the results. The output results are then used as input to the **Stratified Multistage** program. You may store the results from each stratum (point estimate, standard error, universe size) in a file or simply input these values interactively.

NOTE: The "universe size" refers to the number of units at the most detailed level of the multistage sample. For example, if you are obtaining a three-stage sample within each stratum, then the "universe size" refers to the total number of third-stage units within this stratum.

Example 9. This example is similar to Example 8 in the Stratified Cluster section. The universe consisting of university grants is stratified by defining

Stratum 1: state-supported universities (5,600 grants) and

and Stratum 2: private universities (3,500 grants)

Because these universities are so widespread, it was decided to employ a two-stage sample using 20 state-supported universities and 10 private universities. Rather than audit all grants at a selected university, it was decided (based on available resources) to audit roughly 20% of the grants at each selected university to estimate the proportion of grants containing charges after the

scheduled completion of the grant. The following data were obtained, where a_i is the number of

grants in the sample from the i-th university containing such charges, m_i is the number of audited

(sampled) grants at the i-th university, and M_i is the total number of grants in the audit universe

at the i-th university.

State-supported universities				Private universities			
Univ.	M_i	m_i	a_i	Univ.	M_i	m_i	a_i
1	41	8	2	1	11	2	1
2	62	12	3	2	25	5	2
3	21	4	2	3	34	7	2
4	23	5	1	4	18	4	2
5	31	6	1	5	16	3	1
6	32	6	2	6	40	8	3
7	33	7	2	7	31	6	2
8	27	5	2	8	50	10	4
9	41	8	2	9	14	3	1
10	16	3	1	10	12	2	1

These values are stored in data set MULSTAT2.TXT.

Univ.	M_i	m_i	a_i
11	9	2	0
12	31	6	2
13	27	5	1
14	49	10	3
15	46	9	1
16	15	3	1
17	30	6	2
18	24	5	1
19	23	5	1
20	21	4	1

These values are stored in data set MULSTAT1.TXT.

The following two computer outputs are obtained using the **Two-Stage Unrestricted** program.

```
                    DEPARTMENT OF HEALTH & HUMAN SERVICES
                       OIG - OFFICE OF AUDIT SERVICES
Date: 2/11/2004   TWO STAGE UNRESTRICTED ATTRIBUTE APPRAISAL    Time: 15:02
                           AUDIT/REVIEW: Stratum1
                    DATA FILE: C:\Temp\MULSTAT1.TXT

PRIMARY                            SAMPLE ITEMS WITH
 UNIT       UNIVERSE   SAMPLE SIZE  CHARACTERISTIC(S)   RATIO      PROJECTED
=======   ============  ===========  =================   =====     =========

    1           41           8              2          25.00%          10
    2           62          12              3          25.00%          16
    3           21           4              2          50.00%          11
    4           23           5              1          20.00%           5
    5           31           6              1          16.67%           5
    6           32           6              2          33.33%          11
    7           33           7              2          28.57%           9
    8           27           5              2          40.00%          11
    9           41           8              2          25.00%          10
   10           16           3              1          33.33%           5
   11            9           2              0           0.00%           0
   12           31           6              2          33.33%          10
   13           27           5              1          20.00%           5
   14           49          10              3          30.00%          15
   15           46           9              1          11.11%           5
   16           15           3              1          33.33%           5
   17           30           6              2          33.33%          10
   18           24           5              1          20.00%           5
   19           23           5              1          20.00%           5
   20           21           4              1          25.00%           5

TOTALS         602         119             31


OVERALL TOTALS
  120         5,600                                    16.90%         946

STANDARD ERROR                                          1.99%         111

CONFIDENCE LEVEL          80 PERCENT      90 PERCENT      95 PERCENT
LOWER LIMIT FOR PROPORTION    14.35%          13.63%          13.00%
UPPER LIMIT FOR PROPORTION    19.44%          20.17%          20.79%

LOWER LIMIT FOR TOTAL            803             763             728
UPPER LIMIT FOR TOTAL          1,089           1,129           1,164
```

```
                    DEPARTMENT OF HEALTH & HUMAN SERVICES
                       OIG – OFFICE OF AUDIT SERVICES
Date: 2/11/2004    TWO STAGE UNRESTRICTED ATTRIBUTE APPRAISAL    Time: 15:13
                         AUDIT/REVIEW: Stratum 2
                   DATA FILE: C:\Temp\MULSTAT2.TXT

PRIMARY                                  SAMPLE ITEMS WITH
 UNIT        UNIVERSE    SAMPLE SIZE   CHARACTERISTIC(S)    RATIO      PROJECTED
=======    ============  ===========  ==================   =====      =========

   1             11           2              1           50.00%             6
   2             25           5              2           40.00%            10
   3             34           7              2           28.57%            10
   4             18           4              2           50.00%             9
   5             16           3              1           33.33%             5
   6             40           8              3           37.50%            15
   7             31           6              2           33.33%            10
   8             50          10              4           40.00%            20
   9             14           3              1           33.33%             5
  10             12           2              1           50.00%             6

TOTALS          251          50             19

OVERALL TOTALS
  80          3,500                                       21.84%           764

STANDARD ERROR                                            3.57%            125

CONFIDENCE LEVEL              80 PERCENT       90 PERCENT       95 PERCENT
LOWER LIMIT FOR PROPORTION      17.27%           15.97%           14.85%
UPPER LIMIT FOR PROPORTION      26.41%           27.70%           28.83%

LOWER LIMIT FOR TOTAL            604              559              520
UPPER LIMIT FOR TOTAL            924              970            1,009
```
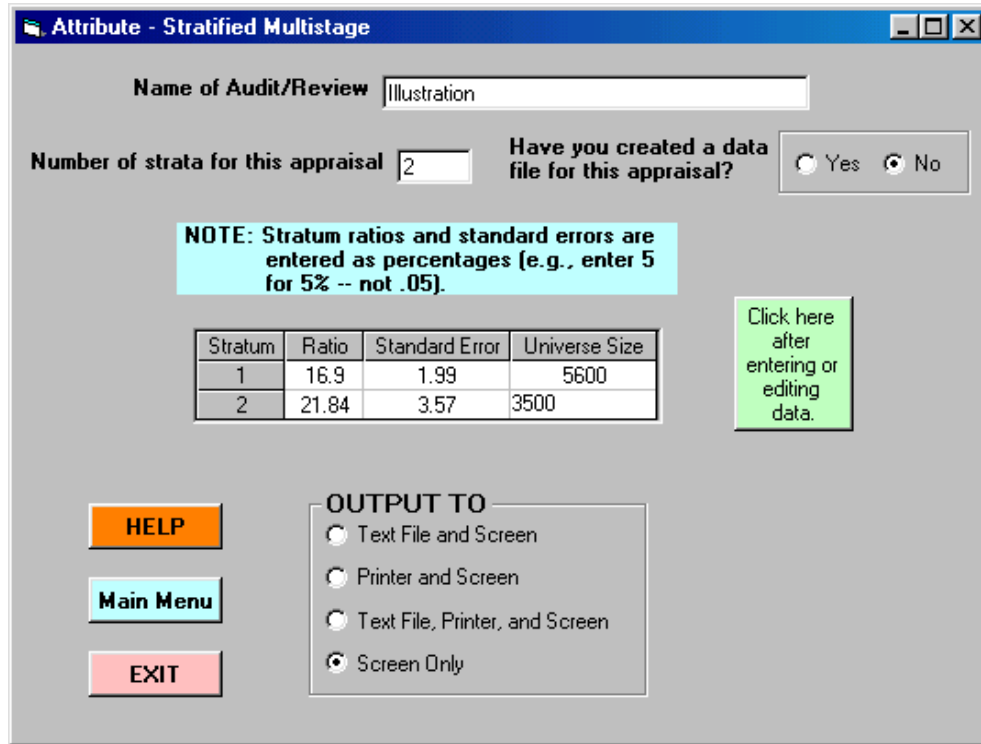
The values used as input to the **Stratified Multistage** program are highlighted in the preceding computer output. The following computer screen illustrates how to enter these values:

The following output is obtained from the **Stratified Multistage** program.

```
                    DEPARTMENT OF HEALTH & HUMAN SERVICES
                       OIG – OFFICE OF AUDIT SERVICES
Date: 2/11/2004      STRATIFIED MULTISTAGE ATTRIBUTE APPRAISAL      Time: 16:08
                        AUDIT/REVIEW: Illustration


             THE ESTIMATORS ARE BASED ON THE FOLLOWING ENTRIES:
   STRATUM          ATTRIBUTE     STANDARD ERROR          UNIVERSE SIZE
      1              16.90%           1.99%                  5,600
      2              21.84%           3.57%                  3,500



= = = = = = = = = = = = = = = = = RESULTS = = = = = = = = = = = = = = = = = = =

ESTIMATED PERCENTAGE:                18.80%
ESTIMATED TOTAL:                      1,711

STANDARD ERROR (PERCENTAGE):          1.84%
STANDARD ERROR (TOTAL):                167

CONFIDENCE LEVEL               80 PERCENT        90 PERCENT        95 PERCENT
LOWER LIMIT FOR PROPORTION        16.44%            15.77%            15.19%
UPPER LIMIT FOR PROPORTION        21.16%            21.83%            22.41%

LOWER LIMIT FOR TOTAL              1,496             1,435             1,383
UPPER LIMIT FOR TOTAL             1,925             1,986             2,039
```

Final results: The point estimate for the percentage of grants containing improper charges is

18.8% (standard error of 1.84%) and the 90% confidence interval for this proportion is from

15.77% to 21.83%. The 90% confidence interval for the universe total is from 1,435 to 1,986.

Discussion. Using Equation 1 in the Formulas section,

$$\hat{p} = (5600/9100)(.169) + (3500/9100)(.2184) = .188 \ (18.8\%)$$

and

$$v(\hat{p}) = (5600/9100)^2(.0199)^2 + (3500/9100)^2(.0357)^2 = .000339$$

The estimated standard error of $\hat{p}$ is $\sqrt{v(\hat{p})} = .0184 \ (1.84\%)$.

The corresponding 90% confidence interval is

$$18.8 \pm 1.644853626951(1.84); \text{ that is } 15.77\% \text{ to } 21.83\%$$

The estimate of the universe total and corresponding confidence interval are obtained by

multiplying the previous results by the total universe size = 5,600 + 3,500 = 9,100.

FORMULAS

1. Estimated universe proportion having the attribute of interest

$$\hat{p} = \sum_{i=1}^{L} \left(\frac{M_i}{M} \right) \hat{p}_i$$

where L = number of strata

M_i = universe size for the most detailed level of the multistage sample

M = total universe size = Σm_i

$\hat{p}_i$ = estimated proportion for the i-th stratum

2. Estimated variance of $\hat{p}$

$$v(\hat{p}) = \sum_{i=1}^{L} \left(\frac{M_i}{M} \right)^2 \text{(standard error of } \hat{p}_i \text{)}^2$$

3. 90% confidence interval for p

$$\hat{p} \pm 1.644853626951\sqrt{v(\hat{p})}$$

4. Estimated universe total having the attribute of interest

$$\hat{T} = M\hat{p}$$

5. Estimated variance of $\hat{T}$

$$v(\hat{T}) = M^2 v(\hat{p})$$

6. 90% confidence interval for T

$$\hat{T} \pm 1.644853626951\sqrt{v(\hat{T})}$$

NOTE: For the Precision at the 95% Confidence Level, replace 1.644853626951 with 1.959963984540 and for the Precision at the 80% Confidence Level, replace 1.644853626951 with 1.281551565545.

VARIABLE APPRAISALS

A variable appraisal is carried out to estimate a particular universe total (T) and its corresponding sampling error. For example, the audit intent may be to determine the dollar value of an inventory or the amount of duplicate payments made by an organization.

A variety of procedures can be used to obtain and appraise a variable sample. There are ten sampling strategies utilized in the Variables Appraisals modules. They are listed below and described in the sections to follow.

- Unrestricted

- Stratified

- Two-Stage Unrestricted

- Three-Stage Unrestricted

- RHC Two Stage

- RHC Three Stage

- Stratified Cluster

- Stratified Multistage

- Post Stratification

- Unknown Universe Size

Unrestricted Variable Appraisal

An **unrestricted** sample is the same as a **simple random sample**. Consequently, every sample of size n has the same chance of being selected. For an unrestricted sample, a sample of size n is randomly obtained and the variable of interest is recorded for each sample item. Actually, the user may input a set of single values (examined amounts, audit amounts, or difference amounts) or a set of two values (examined and audit amounts, examined and difference amounts, or audit and difference amounts).

Example 1. An unrestricted sample of 50 items resulted in the 50 examined/audited values contained in data set DATASRS.TXT. For this sample, all of the resulting differences (examined value - audited value) were nonzero since all the examined (book) values were unequal to the corresponding audit (actual) values.

Data file DATASRS.TXT

1	300	267
2	900	774
3	300	255
4	200	174
5	900	810
6	700	560
7	1000	820
8	100	80
9	900	765
10	700	630
11	700	630
12	400	332
13	300	255
14	100	84
15	200	168
16	100	88
17	600	528
18	400	340

← Each line contains a line counter, examined value, and audited value separated by one or more spaces, a tab delimiter, or a comma.

```
19    900   747
20   1000   800
21   1000   862
22    600   504
23    800   648
24    200   176
25    200   172
26   1000   890
27    900   792
28    600   540
29    500   525
30    200   172
31    200   178
32    500   425
33    200   164
34    500   420
35    500   400
36    400   324
37    200   160
38    600   540
39    500   425
40    300   264
41    900   765
42    100    84
43    100    85
44    900   810
45    300   240
46    500   415
47    500   425
48    300   237
49    500   435
50    100    86
```

The output on the next page was obtained from the **Unrestricted Variable Appraisal** program.

```
                    DEPARTMENT OF HEALTH & HUMAN SERVICES
                       OIG - OFFICE OF AUDIT SERVICES
Date: 4/5/2004         VARIABLE UNRESTRICTED APPRAISAL          Time: 11:23
                         AUDIT/REVIEW: Variable SRS
                     DATA FILE USED: C:\Temp\DATASRS.TXT

      SAMPLE         EXAMINED    NONZERO       TOTAL OF          TOTAL OF
       SIZE            VALUE      DIFFS       DIFF VALUES       AUD VALUES
        50          24,800.00      50          3,530.00        21,270.00


        --------------------- E X A M I N E D -----------------------
        MEAN / UNIVERSE                   496.00              10,000
        STANDARD DEVIATION                296.90
        SKEWNESS                             .32
        KURTOSIS                            1.81
        STANDARD ERROR (MEAN)              41.88
        STANDARD ERROR (TOTAL)           418,823
        POINT ESTIMATE                  4,960,000

                              CONFIDENCE LIMITS
                             80% CONFIDENCE LEVEL
        LOWER LIMIT                     4,415,921
        UPPER LIMIT                     5,504,079
        PRECISION AMOUNT                  544,079
        PRECISION PERCENT                  10.97%
        T-VALUE USED                1.299068784748

                             90% CONFIDENCE LEVEL
        LOWER LIMIT                     4,257,823
        UPPER LIMIT                     5,662,177
        PRECISION AMOUNT                  702,177
        PRECISION PERCENT                  14.16%
        T-VALUE USED                1.676550892617

                             95% CONFIDENCE LEVEL
        LOWER LIMIT                     4,118,344
        UPPER LIMIT                     5,801,656
        PRECISION AMOUNT                  841,656
        PRECISION PERCENT                  16.97%
        T-VALUE USED                2.009575237129

        --------------------- A U D I T E D -------------------------
        MEAN / UNIVERSE                   425.40              10,000
        STANDARD DEVIATION                256.20
        SKEWNESS                             .30
        KURTOSIS                            1.78
        STANDARD ERROR (MEAN)              36.14
        STANDARD ERROR (TOTAL)           361,412
        POINT ESTIMATE                  4,254,000

                              CONFIDENCE LIMITS
                             80% CONFIDENCE LEVEL
        LOWER LIMIT                     3,784,500
        UPPER LIMIT                     4,723,500
        PRECISION AMOUNT                  469,500
        PRECISION PERCENT                  11.04%
        T-VALUE USED                1.299068784748
```

```
                                90% CONFIDENCE LEVEL
        LOWER LIMIT                        3,648,074
        UPPER LIMIT                        4,859,926
        PRECISION AMOUNT                     605,926
        PRECISION PERCENT                     14.24%
        T-VALUE USED                  1.676550892617


                                95% CONFIDENCE LEVEL
        LOWER LIMIT                        3,527,715
        UPPER LIMIT                        4,980,285
        PRECISION AMOUNT                     726,285
        PRECISION PERCENT                     17.07%
        T-VALUE USED                  2.009575237129

        -------------------- D I F F E R E N C E ----------------------
        MEAN / UNIVERSE                       70.60                 10,000
        STANDARD DEVIATION                    48.25
        SKEWNESS                                .64
        KURTOSIS                              2.98
        STANDARD ERROR (MEAN)                  6.81
        STANDARD ERROR (TOTAL)              68,068
        POINT ESTIMATE                      706,000

                                 CONFIDENCE LIMITS
                                80% CONFIDENCE LEVEL
        LOWER LIMIT                          617,575
        UPPER LIMIT                          794,425
        PRECISION AMOUNT                      88,425
        PRECISION PERCENT                     12.52%
        T-VALUE USED                  1.299068784748


                                90% CONFIDENCE LEVEL
        LOWER LIMIT                          591,881
        UPPER LIMIT                          820,119
        PRECISION AMOUNT                     114,119
        PRECISION PERCENT                     16.16%
        T-VALUE USED                  1.676550892617


                                95% CONFIDENCE LEVEL
        LOWER LIMIT                          569,213
        UPPER LIMIT                          842,787
        PRECISION AMOUNT                     136,787
        PRECISION PERCENT                     19.37%
        T-VALUE USED                  2.009575237129
```

Explanation.

NOTE: The following discussion can be applied to the examined values, the audited values, or the difference values. The difference values will be used when discussing the computer output.

The estimated mean of the difference amounts in the universe is the sample mean, $\bar{x} = \$70.60$.

The estimated total difference for the universe (T) is the sample mean times the universe size;

that is, $\hat{T} = (70.60)(10,000) = \$706,000$. This is referred to as the POINT ESTIMATE in the

computer output.

The sample standard deviation is s = 48.2519 and the corresponding (estimated) standard error

for the mean is $48.2519\sqrt{\dfrac{10000 - 50}{(50)(10000)}} = 6.80677$. The (estimated) standard error for the

total is 10,000 x 6.80677 = 68,067.7 (68,068 rounded). The sample skewness is a measure of

the symmetry of the sample data. This value is SKEWNESS = 0.64, indicating a very slight

positive (right-tail) skew. The sample kurtosis is a measure of the sample "peakedness" and is

equal to KURTOSIS = 2.98. Essentially, this value is small whenever the frequency of

observations close to the mean is high and the frequency of observations far from the mean is

low.

The 95% confidence interval for the universe <u>total</u> of the difference amounts is 706,000 ±

(2.009575237129)(68,067.7) = 706,000 ± 136,787; that is, 569,213 to 842,787. The

PRECISION AMOUNT is the amount added and subtracted to the POINT ESTIMATE; that is,

$136,787. This value is 19.37% of the point estimate and is referred to as the PRECISION

PRECENT.[1]

---

[1] When the POINT ESTIMATE is negative, the PRECISION PERCENT is set equal to zero.

FORMULAS

$$\text{STANDARD DEVIATION} = s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

$$\text{STANDARD ERROR (MEAN)} = s\sqrt{\frac{N-n}{nN}} \text{ and STANDARD ERROR (TOTAL)} = Ns\sqrt{\frac{N-n}{nN}}$$

where n = sample size, N = universe size.

$$\text{SKEWNESS} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x - \bar{x})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(x - \bar{x})^2\right]^{3/2}}$$

$$\text{KURTOSIS} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x - \bar{x})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(x - \bar{x})^2\right]^2}$$

95% confidence interval for the universe total (T)

$$\hat{T} \pm t_{.025,n-1} \cdot s \cdot \sqrt{\frac{N(N-n)}{n}}$$

where (1) $\hat{T} = \bar{x} \cdot N$

(2) $t_{.025,n-1}$ is the t-value with n - 1 df having a right-tail area = .025 (RAT-STATS provides t-values accurate to 12 decimal places).

NOTE: For a 90% confidence interval, replace $t_{.025,n-1}$ with $t_{.05,n-1}$ and for an 80% confidence interval, replace $t_{.025,n-1}$ with $t_{.10,n-1}$.

Stratified Variable Appraisal

In a stratified variable sampling plan, the universe is divided into two or more nonoverlapping categories (strata). As with an unrestricted sample, the intent is to make a statistical estimate for a universe total (T) for a particular variable of interest. This plan involves obtaining a random sample from each of the strata. The program will request the number of universe items in each stratum and these values <u>must</u> <u>be</u> <u>known</u>. The program will develop estimates for each stratum as well as for the entire universe.

Using a Stratified Sample

<u>**Purpose**</u>:   To divide (partition) the universe into separate strata so that variation within individual strata is less than variation within the entire universe.

<u>Simple Illustration</u>:

Universe consists of {5 7 8 10 55 60 66 70 120 133 145 150}

Mean of universe is $\mu = 69.08$

Variance of universe is $\sigma^2 = 2871.9$

Partition the universe into three strata:

{5 7 8 10}   {55 60 66 70}   {120 133 145 150}
   #1   #2   #3

The strata variances are: <u>Stratum</u>   <u>Variance</u>
                              1   3.25
                              2   32.69   ← Compare these to $\sigma^2 = 2871.9$
                              3   134.50

Consequently, the individual strata are much more homogeneous.

So, when obtaining a stratified sample, the user can take a larger sample (perhaps 100%) from the stratum containing the large dollar items.

Reasons for Using Stratified Sampling:

A. Improved Sampling Precision

Stratification tends to make the sampling more efficient; that is, the user will obtain narrower confidence intervals for the same sample size. When a sample is skewed or has a high degree of variability, the sample size required to provide a reasonable degree of precision using simple random sampling may be quite large. Precision is improved because each stratum should have a relatively small variance and the weighted sum of the strata variances is less than the variance for the entire universe.

B. Separate Information About Strata and the Universe

Strata may be formed because separate estimates are desired for subuniverses. For example, a nationwide audit of nursing homes can be planned in advance such that separate estimates are published for each state (stratum). When an auditor selects a simple random sample from the entire universe, he/she cannot control the sample size within each stratum. Stratified sampling permits the auditor to also impose different precision requirements on different strata, such as requiring more precise estimates for large accounts.

C. Accommodation of Different Techniques

It may be desirable to employ different sampling methods or audit techniques in various portions of the universe. For example, in a sample of health service employees, the headquarters employees (Stratum 1) may be sampled as individuals and the employees scattered throughout the state (Stratum 2) may be sampled as clusters to save travel time and cost.

Comments:

(1) Defining effective strata is no accident! The user can incorporate all sorts of prior knowledge in defining the strata. Such a technique does <u>not</u> introduce any bias into the final estimate since strata are defined <u>prior</u> to obtaining the sample and each sampling item has a known (although not the same) chance of being selected. As a result, a well-designed stratified plan can provide audit protection and/or improved precision.

(2) Strata can be defined after sample data are obtained <u>provided</u> the proportions of the universe in each stratum are known (with negligible error) and samples of at least 20 are obtained from each stratum.

(3) Generally, it is not a good idea to stratify for convenience (unlike cluster or multistage sampling) since the resulting estimator may be less efficient than the estimator which uses a single simple random sample.

(4) Even though random selection is performed within strata, this does not mean that the user cannot take a close look at the individual findings to determine nature, source, cause, trend and impact.

(5) A careful balance must be maintained between the gains expected in sample precision and the additional time and resources involved in introducing a stratified scheme into the sample design.

Strata Formation

Strata are typically defined using the dollar value of the items being sampled. An alternative is
to stratify using some other variable which is <u>highly</u> <u>correlated</u> with the principal variable, such
as using the number of hospital beds to measure the "size" of a hospital.

<u>**Basic Rule**</u>: Select strata so that their means are as <u>different</u> as possible and their standard
deviations are as small as possible.

<u>**Guidelines**</u>:

- A few strata yield most of the gains (say, 2 to 6).

- Experience, intuition, and the judgment of the auditor are extremely useful in improving
the sampling precision through effective stratification.

- Quantitative rather than qualitative (sex, race, etc.) variables are preferable for defining
strata.

- Coarser divisions of <u>several</u> stratifying variables are preferable to finer divisions of one
variable.

- It is better to use unrelated stratifying variables.

Example 2. Random samples of size 25 were obtained from two strata:

Stratum 1: Examined amounts under $200 ($N_1 = 5,200$) and

Stratum 2: Examined amounts $\geq$ $200 ($N_2 = 3,500$)

NOTE: These sample sizes are too small to meet OAS standards and are used for illustrative
purposes only.

The sample <u>difference</u> amounts for the two strata are stored in data file DATASTRAT.TXT and
the universe/sample sizes are stored in file UNIVSTRAT.TXT.

Data set DATASTRAT.TXT   **Universe File UNIVSTRAT.TXT**

1	80		1	5200	25
2	43		2	3500	25
3	133				
4	125				
5	116				
6	84				
7	111				
8	148				
9	104				
10	114				
11	83				
12	132				
13	96				
14	86				
15	66				
16	89				
17	72				
18	114				
19	135				
20	71				
21	127				
22	105				
23	102				
24	69				
25	76				
26	354				
27	328				
28	313				
29	250				
30	261				
31	294				
32	380				
33	296				
34	248				
35	277				
36	331				
37	305				
38	360				
39	348				
40	318				
41	290				
42	249				
43	362				
44	348				
45	355				
46	295				
47	277				
48	355				
49	314				
50	277				

The sample results are:

Stratum 1: $n_1 = 25$, mean = 99.24, std. dev. = 26.3317

Stratum 2: $n_2 = 25$, mean = 311.40, std. dev. = 39.6432

The following computer output was obtained from the **Stratified Variables Appraisal** program:

```
                    DEPARTMENT OF HEALTH & HUMAN SERVICES
                       OIG - OFFICE OF AUDIT SERVICES
Date: 4/5/2004          STRATIFIED VARIABLE APPRAISAL              Time: 12:17
                     AUDIT/REVIEW: Variable - Stratified

                    DATA FILE USED: C:\Temp\DATASTRAT.TXT

          STRATUM      SAMPLE
           NUMBER       SIZE      VALUE OF SAMPLE   NONZERO ITEMS
             1           25           2,481.00          25
             2           25           7,785.00          25

          TOTALS        50          10,266.00          50

          -------------------- D I F F E R E N C E ----------------------
Stratum 1  MEAN / UNIVERSE                        99.24              5,200
           STANDARD DEVIATION                     26.33
           SKEWNESS                               -.07
           KURTOSIS                              2.24
           STANDARD ERROR (MEAN)                  5.25
           STANDARD ERROR (TOTAL)             27,319
           POINT ESTIMATE                      516,048

                                    CONFIDENCE LIMITS
                                    80% CONFIDENCE LEVEL
           LOWER LIMIT                           480,046
           UPPER LIMIT                           552,050
           PRECISION AMOUNT                       36,002
           PRECISION PERCENT                       6.98%
           T-VALUE USED              1.317835933673

                                    90% CONFIDENCE LEVEL
           LOWER LIMIT                           469,308
           UPPER LIMIT                           562,788
           PRECISION AMOUNT                       46,740
           PRECISION PERCENT                       9.06%
           T-VALUE USED              1.710882079909

                                    95% CONFIDENCE LEVEL
           LOWER LIMIT                           459,664
           UPPER LIMIT                           572,432
           PRECISION AMOUNT                       56,384
           PRECISION PERCENT                      10.93%
           T-VALUE USED              2.063898561628
```

```
Stratum 2   MEAN / UNIVERSE                    311.40                  3,500
            STANDARD DEVIATION                  39.64
            SKEWNESS                            -.06
            KURTOSIS                            1.85
            STANDARD ERROR (MEAN)               7.90
            STANDARD ERROR (TOTAL)             27,651
            POINT ESTIMATE                  1,089,900

                                     CONFIDENCE LIMITS
                                   80% CONFIDENCE LEVEL
            LOWER LIMIT                    1,053,461
            UPPER LIMIT                    1,126,339
            PRECISION AMOUNT                  36,439
            PRECISION PERCENT                  3.34%
            T-VALUE USED              1.317835933673

                                   90% CONFIDENCE LEVEL
            LOWER LIMIT                    1,042,592
            UPPER LIMIT                    1,137,208
            PRECISION AMOUNT                  47,308
            PRECISION PERCENT                  4.34%
            T-VALUE USED              1.710882079909

                                   95% CONFIDENCE LEVEL
            LOWER LIMIT                    1,032,831
            UPPER LIMIT                    1,146,969
            PRECISION AMOUNT                  57,069
            PRECISION PERCENT                  5.24%
            T-VALUE USED              2.063898561628

OVERALL     POINT ESTIMATE / UNIVERSE     1,605,948                  8,700
            STANDARD ERROR                   38,870

                                     CONFIDENCE LIMITS
                                   80% CONFIDENCE LEVEL
            LOWER LIMIT                    1,556,134
            UPPER LIMIT                    1,655,762
            PRECISION AMOUNT                  49,814
            PRECISION PERCENT                  3.10%
            Z-VALUE USED              1.281551565545

                                   90% CONFIDENCE LEVEL
            LOWER LIMIT                    1,542,012
            UPPER LIMIT                    1,669,884
            PRECISION AMOUNT                  63,936
            PRECISION PERCENT                  3.98%
            Z-VALUE USED              1.644853626951

                                   95% CONFIDENCE LEVEL
            LOWER LIMIT                    1,529,764
            UPPER LIMIT                    1,682,132
            PRECISION AMOUNT                  76,184
            PRECISION PERCENT                  4.74%
            Z-VALUE USED              1.959963984540
```

Discussion. The point estimates for the universe total difference amounts are $516,048

(Stratum 1) and $1,089,900 (Stratum 2). Referring to the formula section and the OVERALL

section in the computer output, the estimate of the universe total difference is

$$\hat{T} = (5200)(99.24) + (3500)(311.40) = \$1,605,948$$

The estimated variance of $\hat{T}$ is

$$5200^2\left(\frac{5200-25}{5200}\right)\frac{26.3317^2}{25} + 3500^2\left(\frac{3500-25}{3500}\right)\frac{39.6432^2}{25} \quad = 1,510,906,287$$

The (estimated) standard error of $\hat{T}$ is $\sqrt{1,510,906,287} = 38,870$.

The 95% confidence interval for universe total (T) is

$$1,605,948 \pm (1.959963984540)(38,870)$$

that is, $1,605,948 \pm 76,184$ ($1,529,764 to $1,682,132).

The PRECISION AMOUNT here is $76,184 and is 4.74% of the point estimate, $\hat{T}$.

NOTE: When the POINT ESTIMATE is negative, the PRECISION PERCENT is set equal to zero.

FORMULAS

NOTE: For definitions and formulas of the statistics within each stratum (standard deviation, standard error, skewness, and kurtosis), refer to the previous section (**Unrestricted Variable Appraisal**).

1. Estimate of universe mean (μ):

$$\bar{y}_{st} = (N_1 / N)\bar{y}_1 + (N_2 / N)\bar{y}_2 + \cdots + (N_L / N)\bar{y}_L$$

   where $L =$ number of strata

         $N_i =$ number of items in i-th stratum (universe)

$$N = \ N_1 + N_2 + \cdots + N_L$$

$$\overline{y}_i \ = \ \text{average of sample items in the i-th stratum}$$

2. Estimate of universe total (T):

$$\hat{T} = \ N \cdot \overline{y}_{st} \ = \ N_1 \cdot \overline{y}_1 + N_2 \cdot \overline{y}_2 + \cdots + N_L \cdot \overline{y}_L$$

3. Estimated variance of $\overline{y}_{st}$:

$$v(\overline{y}_{st}) = \ \frac{1}{N^2} \sum_{i=1}^{L} N_i^2 \left(\frac{N_i - n_i}{N_i} \right) \frac{s_i^2}{n_i}$$

   where $n_i = $ number of sampled items in i-th stratum

   $s_i^2 = $ sample variance for i-th stratum

4. Estimated variance of $\hat{T}$:

$$v(\hat{T}) = \ N^2 v(\overline{y}_{st})$$

5. Approximate 95% confidence interval for universe mean (μ):

$$\overline{y}_{st} \pm Z_{.025} \sqrt{v(\overline{y}_{st})} \ \ \text{where } Z_{.025} = 1.959963984540.$$

6. Approximate 95% confidence interval for universe total (T):

$$\hat{T} \pm Z_{.025} \sqrt{v(\hat{T})}$$

NOTES: 1. For a 90% confidence interval, replace $Z_{.025}$ with $Z_{.05} = 1.644853626951$ and for an 80% confidence interval, replace $Z_{.025}$ with $Z_{.10} = 1.281551565545$.

   2. The confidence intervals for each stratum total use t-values that are accurate to 12 decimal places.

Two-Stage Unrestricted Variable Appraisal

This is a special case of **multistage sampling**. Multistage sampling is a very cost-effective

sampling procedure when (1) obtaining a frame that lists all elements in the universe is very

costly or impossible or (2) the cost of obtaining observations increases as the distance separating

the elements increases. Put another way, multistage sampling is cost effective when it is more

costly to get <u>to</u> the sampling unit than it is to audit the sampling unit. The goal of multistage

sampling is to get the most precise results per unit of examination cost.

<u>General Comments</u>

1. This is a <u>very</u> convenient sampling procedure for many situations. You don't have to visit all

    locations.

2. For a two-stage procedure, the universe can be broken down into "subgroups."

    > Example: 1st Stage: Carriers (primary units, P.U.s)

    > 2nd Stage: Hospitals (secondary units, S.U.s)

    So, the procedure is to first obtain a random sample of P.U.s. These are called <u>clusters</u>.

    Then, obtain a random sample of S.U.s within each selected P.U. Notice that at the first

    stage, <u>clusters</u> are the sampling unit (sampling units are not always <u>individual</u> people,

    records, etc.). The program will accept a maximum of 20 clusters.

3. You can estimate cost recoveries for the entire universe with multistage sampling and it is

    very useful for large, widespread universes.

Example 3. In a particular region of the U.S. there are N = 90 universities with government

research grants. Because these universities are so widespread, it was decided to use a cluster

sample of n = 10 universities. The 10 universities to be sampled may be obtained using the

Single-Stage Random Numbers module discussed in the **Random Numbers** section.

Enter the values shown in the following input screen:

The resulting output is shown on the next page. Note that the selected universities are in

sequential order:

Universities: 2, 5, 7, 23, 28, 46, 56, 67, 70, 76

```
                    Department of Health and Human Services
                          OIG - Office of Audit Services
Date: 10/26/2004              Random Number Generator              Time: 11:03
                              AUDIT: select

SEED NUMBER: 1357.00                          FRAME SIZE:              90


     FILE OF RANDOM NUMBERS: C:\TEMP\SELECT.TXT

     TOTAL RANDOM NUMBERS GENERATED: 10

THE NUMBERS ARE IN THE FOLLOWING FORMAT IN YOUR FILE:
   POSITIONS 1 THROUGH  6 - ORDER OF SELECTION
   POSITIONS 7 THROUGH 17 - RANDOM NUMBER
EACH COLUMN OF NUMBERS IS RIGHT JUSTIFIED.


Selection
  Order      Value
    4            2
    7            5
    9            7
   10           23
    5           28
    2           46
    6           56
    1           67
    8           70
    3           76

SUMMATION OF RANDOM NUMBERS = 380
```

Example 3 -- continued. Rather than audit all grants at a selected university, it was decided

(based on available resources) to audit roughly 20% of the grants at each selected university.

There are a total of M = 4500 grants in all 90 universities (see NOTE 5 under Formula 2). The

following data were obtained where y_i is the dollars (in thousands) of improper charges for the

ith

sampled university, M_i is the total number of grants at the i-th university, and m_i is the number of

audited grants at the i-th university. Also, $\overline{y}_i$ and s_i^2 are the mean and variance of the sample

values from the i-th university.

For ease of illustration, university 1 refers to university 2, university 2 refers to university 5, and

so on.

Univ.	M_i	m_i	Dollars (y_i, in thousands)	$\bar{y}_i$	s_i^2
1	50	10	5, 7, 9, 0, 11, 2, 8, 4, 3, 5	5.40	11.38
2	65	13	4, 3, 7, 2, 11, 0, 1, 9, 4, 3, 2, 1, 5	4.00	10.67
3	45	9	5, 6, 4, 11, 12, 0, 1, 8, 4	5.67	16.75
4	48	10	6, 4, 0, 1, 0, 9, 8, 4, 6, 10	4.80	13.29
5	52	10	11, 4, 3, 1, 0, 2, 8, 6, 5, 3	4.30	11.12
6	58	12	12, 11, 3, 4, 2, 0, 0, 1, 4, 3, 2, 4	3.83	14.88
7	42	8	3, 7, 6, 7, 8, 4, 3, 2	5.00	5.14
8	66	13	3, 6, 4, 3, 2, 2, 8, 4, 0, 4, 5, 6, 3	3.85	4.31
9	40	8	6, 4, 7, 3, 9, 1, 4, 5	4.88	6.13
10	56	11	6, 7, 5, 10, 11, 2, 1, 4, 0, 5, 4	5.00	11.80

NOTE: This example violates OAS minimum sample sizes of at least 30 grants at each university. It is used for illustration only. These data are in data set DATA2STG.TXT. The universe sizes (the M_i's) are in data set UNIV2STG.TXT. The program output immediately follows.

Dataset DATA2STG.TXT   **Dataset UNIV2STG.TXT**

1	5		1	50	10
2	7		2	65	13
3	9		3	45	9
4	0		4	48	10
5	11		5	52	10
6	2		6	58	12
7	8		7	42	8
8	4		8	66	13
9	3		9	40	8
10	5		10	56	11
11	4				
12	3				
13	7				
14	2				
15	11				
16	0				
17	1				
18	9				
19	4				
20	3				
21	2				
22	1				

```
23          5
  .
  .
  .
94          6
95          7
96          5
97         10
98         11
99          2
100         1
101         4
102         0
103         5
104         4
```

DEPARTMENT OF HEALTH & HUMAN SERVICES
OIG - OFFICE OF AUDIT SERVICES
Date: 10/29/2004   TWO-STAGE UNRESTRICTED VARIABLE APPRAISAL   Time: 10:07
AUDIT/REVIEW: Variable 2-Stage

DATA FILE USED: C:\Temp\DATA2STG.TXT

--------------------- D I F F E R E N C E ---------------------

UNIT NBR	SAMPLE SIZE/ NONZERO ITEMS	SAMPLE MEAN	VARIANCE	UNIVERSE SIZE	POINT ESTIMATE
1	10/9	5.40	11.38	50	270
2	13/12	4.00	10.67	65	260
3	9/8	5.67	16.75	45	255
4	10/8	4.80	13.29	48	230
5	10/9	4.30	11.12	52	224
6	12/10	3.83	14.88	58	222
7	8/8	5.00	5.14	42	210
8	13/12	3.85	4.31	66	254
9	8/8	4.88	6.13	40	195
10	11/10	5.00	11.80	56	280
	104/94	4.80		522	2,400

```
NOT SAMPLED
   80                                       3,978
OVERALL TOTALS
   90                                       4,500        21,602

STANDARD ERROR                                              867
                         CONFIDENCE LIMITS
                       80% CONFIDENCE LEVEL
          LOWER LIMIT                   20,491
          UPPER LIMIT                   22,712
          PRECISION AMOUNT               1,111
          PRECISION PERCENT              5.14%
          Z-VALUE USED          1.281551565545
```

```
                                    90% CONFIDENCE LEVEL
              LOWER LIMIT                         20,176
              UPPER LIMIT                         23,027
              PRECISION AMOUNT                     1,425
              PRECISION PERCENT                     6.60%
              Z-VALUE USED                 1.644853626951

                                    95% CONFIDENCE LEVEL
              LOWER LIMIT                         19,903
              UPPER LIMIT                         23,300
              PRECISION AMOUNT                     1,699
              PRECISION PERCENT                     7.86%
              Z-VALUE USED                 1.959963984540
```

Discussion.   The point estimate (highlighted) for the universe total difference amount is 21,602

(that is, \$21,602,000).   The 95% confidence interval for this amount is from 19,903 to 23,300

(\$19,903,000 to \$23,300,000).   The PRECISION AMOUNT at the 95% confidence level is

(Z-value)(standard error of $\hat{T}$) = (1.959963984540)(867) = 1,699 (that is, \$1,699,000). This value

is 7.86% of the point estimate.

Notice that the output also contains the estimated totals for each primary unit (university).   For

example, the estimated total difference for university 1 is 270 (\$270,000).   The sample average of

these estimates is 2,400/10 = 240 (\$240,000).   Since there are 90 universities in the universe, the

point estimate for the universe total is (90)(240) = 21,600; that is, \$21,600,000 (more precisely,

21,602,000).

FORMULAS

1. The point estimate for the universe total (T) is

$$\hat{T} = N \frac{\sum\limits_{i=1}^{n} M_i \bar{y}_i}{n}$$

2. The estimated variance of $\hat{T}$ is

$$v(\hat{T}) = \left(\frac{N(N-n)}{n} \right) \left[\frac{\sum_{i=1}^{n} \left(M_i \bar{y}_i - \frac{\hat{T}}{N} \right)^2}{n-1} \right] + \frac{N}{n} \sum_{i=1}^{n} M_i^2 \left(\frac{M_i - m_i}{M_i} \right) \frac{s_i^2}{m_i}$$

NOTES: 1. n = number of primary units in the sample and N = number of primary units in the universe.

2. The STANDARD ERROR of $\hat{T}$ is the square root of v($\hat{T}$).

3. The PRECISION AMOUNT at the 95% confidence level for the universe total is (1.959963984540)(standard error of $\hat{T}$).

4. For the PRECISION AMOUNT at the 90% confidence level, replace 1.959963984540 with 1.644853626951. For the PRECISION AMOUNT at the 80% confidence level, replace 1.959963984540 with 1.281551565545.

5. The total number of secondary units in the universe (M) may be known or unknown and is not used in any of the calculations.

3. The approximate 95% confidence interval for T is

$$\hat{T} \pm 1.959963984540\sqrt{v(\hat{T})}$$

NOTE: For a 90% confidence interval, replace 1.959963984540 with 1.644853626951 and for an 80% confidence interval replace 1.959963984540 with 1.281551565545.

Three-Stage Unrestricted Variable Appraisal

Example 4.   The situation discussed in Example 3 was extended the following year to a three-

stage procedure by defining:

   Stage 1: REGION (select 4 out of 12 regions)

   Stage 2: UNIVERSITY (select 10 from each selected region)

   Stage 3: GRANT (select approximately 20% of all grants at each university)

NOTE: This example violates OAS minimum sample sizes and is used for illustration only.

Using the **Single-Stage Random Numbers** module, regions 5, 7, 8, and 10 were selected as the

sampled primary units. Next, 10 universities (secondary units) were randomly selected from the

available universities in each of the four selected regions. The data on the following four pages

were obtained, where M_i is the number of grants in the universe for each university, m_i is the

number of sampled grants at each university (chosen to be roughly 20% of M_i),

$\bar{y}_{ij}$ is the sample average of the items from the j-th university within the i-th region, and s_{ij} is

the sample standard deviation of the items from the j-th university within the i-th region. The
resulting data are stored in file DATA3STG.TXT.

REGION 5 -- Universe contains 90 universities.

Univ.	M_i	m_i
1	47	9
2	51	10
3	45	9
4	46	9
5	46	9
6	50	10
7	50	10
8	57	11
9	54	11
10	64	13

Data (total thousands of dollars of improper charges)

University

Obs.	1	2	3	4	5	6	7	8	9	10
1	8	13	10	11	14	5	0	0	2	12
2	0	13	0	9	14	11	14	3	9	13
3	6	4	12	12	5	0	4	0	15	14
4	6	6	14	1	1	4	10	6	4	11
5	0	0	13	10	2	13	13	1	13	6
6	13	15	13	11	11	8	10	0	12	10
7	1	12	13	15	14	15	8	3	14	11
8	7	9	2	7	15	2	0	13	6	0
9	2	0	9	8	11	0	7	7	11	7
10		13				14	3	5	0	12
11								4	1	9
12										11
13										7
$\bar{y}_{1j}$	4.78	8.50	9.56	9.33	9.67	7.20	6.90	3.82	7.91	9.46
s_{1j}	4.38	5.64	5.13	3.91	5.52	5.79	5.02	3.92	5.52	3.73

REGION 7 -- Universe contains 110 universities.

Univ.	M_i	m_i
1	53	11
2	59	12
3	52	10
4	67	13
5	59	12
6	73	15
7	51	10
8	75	15
9	66	13
10	58	12

Data (total thousands of dollars of improper charges)

University

Obs.	1	2	3	4	5	6	7	8	9	10
1	0	12	4	5	0	2	4	19	17	16
2	4	0	19	10	4	18	16	13	8	17
3	0	15	16	2	6	0	0	0	15	6
4	10	11	12	10	9	0	8	13	12	10
5	11	0	4	12	13	12	3	4	0	2
6	18	18	2	7	19	17	8	0	20	6
7	18	0	1	3	9	0	13	0	17	13
8	16	17	5	0	18	0	0	0	6	0
9	2	8	1	20	16	16	0	0	9	12

	1	2	3	4	5	6	7	8	9	10
10	18	12	8	10	15	3	9	13	16	20
11	18	13		0	0	0		4	7	19
12		5		0	13	4		16	2	0
13				15		3		13	7	
14						6		0		
15						5		4		
$\bar{y}_{2j}$	10.45	9.25	7.20	7.23	10.17	5.73	6.10	6.60	10.46	10.08
s_{2j}	7.69	6.59	6.41	6.30	6.59	6.65	5.65	7.01	6.21	7.23

REGION 8 -- Universe contains 85 universities.

Univ.	M_i	m_i
1	45	9
2	39	8
3	43	9
4	34	7
5	54	11
6	54	11
7	34	7
8	59	12
9	49	10
10	43	9

Data (total thousands of dollars of improper charges)

					University					
Obs.	1	2	3	4	5	6	7	8	9	10
1	6	0	10	4	13	0	13	0	6	14
2	5	8	15	1	13	0	10	13	8	0
3	1	0	11	2	15	14	10	9	3	0
4	3	1	6	5	10	11	6	3	12	10
5	12	10	12	2	10	14	9	14	4	12
6	7	15	2	9	7	5	8	14	8	0
7	0	1	14	4	0	1	1	12	1	7
8	3	14	7		0	13		11	2	6
9	12		0		3	0		11	4	0
10					13	5		6	3	
11					1	10		15		
12								11		
$\bar{y}_{3j}$	5.44	6.12	8.56	3.86	7.73	6.64	8.14	9.92	5.10	5.44
s_{3j}	4.33	6.40	5.20	2.67	5.78	5.90	3.80	4.66	3.38	5.68

REGION 10 -- Universe contains 120 universities.

Univ.	M_i	m_i
1	59	12
2	68	14
3	57	11
4	72	14
5	70	14
6	73	15
7	83	17
8	89	18
9	73	15
10	77	15

Data (total thousands of dollars of improper charges)

University

Obs.	1	2	3	4	5	6	7	8	9	10
1	0	3	8	9	7	15	18	16	20	18
2	10	13	10	10	6	15	10	2	14	0
3	14	0	1	15	3	7	17	14	5	8
4	0	12	8	16	15	12	18	0	0	4
5	18	12	19	6	12	15	0	0	19	5
6	0	7	18	17	10	4	0	0	0	17
7	8	1	15	5	17	0	8	6	12	0
8	20	13	0	2	0	14	0	19	5	0
9	19	2	0	8	0	18	3	17	15	13
10	0	0	17	9	0	17	18	13	15	8
11	0	16	18	16	0	13	12	12	2	10
12	3	14		7	0	6	0	13	15	7
13		17		0	7	0	11	12	2	10
14		5		0	16	3	19	11	2	0
15						4	0	14	1	0
16							9	13		
17							6	0		
18								1		
$\bar{y}_{4j}$	7.67	8.21	10.36	8.57	6.64	9.53	8.76	9.06	8.47	6.67
s_{4j}	8.25	6.27	7.55	5.80	6.44	6.32	7.38	6.77	7.39	6.17

Using the file construction suggested in the User's Guide for this module, the primary unit file

and secondary unit file could be constructed as shown below. In the secondary unit file, each

line begins with a counter (1, 2, 3, ...), a value identifying the primary unit number in the second

column, and a value identifying the secondary unit number within each primary unit in the third

column.

Data set PRIMARY3STG.TXT
```
1   REGION 5       90    10
2   REGION 7      110    10
3   REGION 8       85    10
4   REGION 10     120    10
```

Data set SECONDARY3STG.TXT
```
1    1   1    UNIV1    47    9
2    1   2    UNIV2    51   10
3    1   3    UNIV3    45    9
4    1   4    UNIV4    46    9
5    1   5    UNIV5    46    9
6    1   6    UNIV6    50   10
7    1   7    UNIV7    50   10
8    1   8    UNIV8    57   11
9    1   9    UNIV9    54   11
10   1   10   UNIV10   64   13
11   2   1    UNIV1    53   11
12   2   2    UNIV2    59   12
13   2   3    UNIV3    52   10
14   2   4    UNIV4    67   13
15   2   5    UNIV5    59   12
16   2   6    UNIV6    73   15
17   2   7    UNIV7    51   10
18   2   8    UNIV8    75   15
19   2   9    UNIV9    66   13
20   2   10   UNIV10   58   12
21   3   1    UNIV1    45    9
22   3   2    UNIV2    39    8
23   3   3    UNIV3    43    9
24   3   4    UNIV4    34    7
25   3   5    UNIV5    54   11
26   3   6    UNIV6    54   11
27   3   7    UNIV7    34    7
28   3   8    UNIV8    59   12
29   3   9    UNIV9    49   10
30   3   10   UNIV10   43    9
31   4   1    UNIV1    59   12
32   4   2    UNIV2    68   14
33   4   3    UNIV3    57   11
34   4   4    UNIV4    72   14
35   4   5    UNIV5    70   14
36   4   6    UNIV6    73   15
37   4   7    UNIV7    83   17
38   4   8    UNIV8    89   18
39   4   9    UNIV9    73   15
40   4   10   UNIV10   77   15
```

Once again using the file construction suggested in the User's Guide for this module, the data

file could be constructed as shown below. The data lines for the first two universities in Region

5 and the last two universities in Region 10 are shown. Each line begins with a counter (1, 2, 3,

...), a value identifying the primary unit number in the second column, a value identifying the

secondary unit number within each primary unit in the third column, and a value identifying the

third-stage unit number within each sampled primary/secondary unit in the fourth column. The

sample value appears in the fifth column.

Data set DATA3STG.TXT
```
1        1        1        1        8
2        1        1        2        0
3        1        1        3        6
4        1        1        4        6
5        1        1        5        0
6        1        1        6       13
7        1        1        7        1
8        1        1        8        7
9        1        1        9        2
10       1        2        1       13
11       1        2        2       13
12       1        2        3        4
13       1        2        4        6
14       1        2        5        0
15       1        2        6       15
16       1        2        7       12
17       1        2        8        9
18       1        2        9        0
19       1        2       10       13
     .
     .
     .
433      4        9        1       20
434      4        9        2       14
435      4        9        3        5
436      4        9        4        0
437      4        9        5       19
438      4        9        6        0
439      4        9        7       12
440      4        9        8        5
441      4        9        9       15
```

```
442     4     9     10     15
443     4     9     11      2
444     4     9     12     15
445     4     9     13      2
446     4     9     14      2
447     4     9     15      1
448     4    10      1     18
449     4    10      2      0
450     4    10      3      8
451     4    10      4      4
452     4    10      5      5
453     4    10      6     17
454     4    10      7      0
455     4    10      8      0
456     4    10      9     13
457     4    10     10      8
458     4    10     11     10
459     4    10     12      7
460     4    10     13     10
461     4    10     14      0
462     4    10     15      0
```

The program output using these three files is shown on the following pages.


```
                     DEPARTMENT OF HEALTH & HUMAN SERVICES
                          OIG - OFFICE OF AUDIT SERVICES
Date: 10/4/2004   THREE-STAGE UNRESTRICTED VARIABLE APPRAISAL      Time: 17:34
                         AUDIT/REVIEW: Variable 3-Stage

                   DATA FILE USED: C:\Temp\DATA3STG.TXT
```

			------------------- D I F F E R E N C E ----------------------			
FIRST STAGE	SAMPLE	NON-				
SECOND STAGE	SIZE	ZEROES	SAMPLE MEAN	VARIANCE	UNIVERSE	POINT ESTIMATE
================	======	======	===========	========	==========	==============
REGION 5						
UNIV1	9	7	4.78	19.19	47	225
UNIV2	10	8	8.50	31.83	51	434
UNIV3	9	8	9.56	26.28	45	430
UNIV4	9	9	9.33	15.25	46	429
UNIV5	9	9	9.67	30.50	46	445
UNIV6	10	8	7.20	33.51	50	360
UNIV7	10	8	6.90	25.21	50	345
UNIV8	11	8	3.82	15.36	57	218
UNIV9	11	10	7.91	30.49	54	427
UNIV10	13	12	9.46	13.94	64	606
COMBINED	101		392		90	35,256

```
REGION 7
  UNIV1          11       9      10.45    59.07        53             554
  UNIV2          12       9       9.25    43.48        59             546
  UNIV3          10      10       7.20    41.07        52             374
  UNIV4          13      10       7.23    39.69        67             484
  UNIV5          12      10      10.17    43.42        59             600
  UNIV6          15      10       5.73    44.21        73             419
  UNIV7          10       7       6.10    31.88        51             311
  UNIV8          15       9       6.60    49.11        75             495
  UNIV9          13      12      10.46    38.60        66             690
  UNIV10         12      10      10.08    52.27        58             585

    COMBINED    123             506                   110          55,643

REGION 8
  UNIV1           9       8       5.44    18.78        45             245
  UNIV2           8       6       6.12    40.98        39             239
  UNIV3           9       8       8.56    27.03        43             368
  UNIV4           7       7       3.86     7.14        34             131
  UNIV5          11       9       7.73    33.42        54             417
  UNIV6          11       8       6.64    34.85        54             358
  UNIV7           7       7       8.14    14.48        34             277
  UNIV8          12      11       9.92    21.72        59             585
  UNIV9          10      10       5.10    11.43        49             250
  UNIV10          9       5       5.44    32.28        43             234

    COMBINED     93             310                    85          26,388


          -------------------- D I F F E R E N C E ---------------------
FIRST STAGE      SAMPLE  NON-
  SECOND STAGE    SIZE   ZEROES SAMPLE MEAN VARIANCE   UNIVERSE   POINT ESTIMATE
================ ====== ====== =========== ========   ==========  ===============
REGION 10
  UNIV1          12       7       7.67    68.06        59             452
  UNIV2          14      12       8.21    39.26        68             559
  UNIV3          11       9      10.36    57.05        57             591
  UNIV4          14      12       8.57    33.65        72             617
  UNIV5          14       9       6.64    41.48        70             465
  UNIV6          15      13       9.53    39.98        73             696
  UNIV7          17      12       8.76    54.44        83             727
  UNIV8          18      14       9.06    45.82        89             806
  UNIV9          15      13       8.47    54.55        73             618
  UNIV10         15      10       6.67    38.10        77             513

    COMBINED    145             604                   120          72,534


                STAGES              UNIVERSE        SAMPLED
                 FIRST                 12              4
                 SECOND               405             40
                 THIRD              2,298            462

OVERALL POINT ESTIMATE                                          569,464
OVERALL STANDARD ERROR                                         102,337

                            CONFIDENCE LIMITS
                          80% CONFIDENCE LEVEL
              LOWER LIMIT                      438,314
              UPPER LIMIT                      700,615
              PRECISION AMOUNT                 131,151
              PRECISION PERCENT                 23.03%
              Z-VALUE USED             1.281551565545
```

```
                              90% CONFIDENCE LEVEL
        LOWER LIMIT                        401,134
        UPPER LIMIT                        737,795
        PRECISION AMOUNT                   168,330
        PRECISION PERCENT                   29.56%
        Z-VALUE USED              1.644853626951

                              95% CONFIDENCE LEVEL
        LOWER LIMIT                        368,887
        UPPER LIMIT                        770,042
        PRECISION AMOUNT                   200,578
        PRECISION PERCENT                   35.22%
        Z-VALUE USED              1.959963984540
```

Some highlighted values: (1) 405 is 90 + 110 + 85 + 120, (2) 2,298 is the total number of third-stage units (universe) for the four sampled primary units, (3) 462 is 101 + 123 + 93 + 145.

The point estimate and confidence intervals: The point estimate (highlighted) for the universe total difference amount is 569,464 (that is, $569,464,000). The 95% confidence interval for this amount is from 368,887 to 770,042 ($368,887,000 to $770,042,000). The PRECISION AMOUNT at the 95% confidence level is (Z-value)(standard error of $\hat{T}$) =

(1.959963984540)(102,337) = 200,578 (that is, $200,578,000). This value is 35.22% of the point estimate.

Notice that the output also contains the estimated totals for each primary unit (region) and each secondary unit (university). For example, the estimated total difference for UNIV1 within region 5 is 225 ($225,000). The sample average of the 10 university estimates in region 5 is 392 ($392,000). Since there are 90 universities in this region, the point estimate for the region 5 total is (90)(392) = 35,280 (more precisely, 35,256 or $35,256,000).

The average of the four regional estimates is (35,256 + 55,643 + 26,388 + 72,534)/4 = 47,455.25 ($47,455,250). Since there are 12 regions in the universe, the (unbiased) point estimate for the

universe total is (12)(47,455,250) = approximately \$569,463,000. The actual amount

(highlighted) is \$569,464,000.

FORMULAS

1. Point estimate of the universe total (T):

$$\hat{T} = \frac{N}{n} \sum_{i=1}^{n} \hat{T}_i$$

where $\hat{T}_i = \dfrac{M_i}{m_i} \sum_{j=1}^{m_i} \hat{T}_{ij}$ is the estimate of the total for the i-th sampled P.U. and where

$$\hat{T}_{ij} = \frac{B_{ij}}{b_{ij}} \sum_{k=1}^{b_{ij}} y_{ijk}$$ is the estimate of the total for the j-th S.U. within the i-th sampled P.U.

Notation: n =   number of primary units (P.U.s) in the sample

  N =   number of P.U.s in the universe

  m_i =   number of sampled secondary units (S.U.s) in the i-th P.U. (i = 1, . . ., n)

  M_i =   number of S.U.s in the universe in the i-th P.U. (i = 1, . . ., n)

  b_{ij} =   number of 3rd-stage items in the sample for the i-th P.U. and j-th S.U.
          (i = 1, . . ., n and j = 1, . . ., m_i)

  B_{ij} =   number of 3rd-stage items in the universe for the i-th P.U. and j-th S.U.
          (i = 1, . . ., n and j = 1, . . ., m_i)

  y_{ijk} =   sample value of the k-th item from the i-th P.U. and j-th S.U.
          (i = 1, . . ., n and j = 1, . . ., m_i, k = 1, . . ., b_{ij})

NOTE: The value of n, N, along with m_i, M_i, b_{ij}, B_{ij} for each **sampled** primary and secondary
          unit must be known.

2. Estimated variance of $\hat{T}$:

$$v(\hat{T}) = \frac{N(N-n)}{n}s^2 + \frac{N}{n}\sum_{i=1}^{n}\frac{M_i(M_i - m_i)}{m_i}s_i{}^2 + \frac{N}{n}\sum_{i=1}^{n}\frac{M_i}{m_i}\sum_{j=1}^{m_i}\frac{B_{ij}(B_{ij} - b_{ij})}{b_{ij}}s_{ij}{}^2$$

where $\quad s^2 = \left[\sum_{i=1}^{n}\hat{T}_i{}^2 - \left(\sum_{i=1}^{n}\hat{T}_i\right)^2 / n\right] / (n-1)$

$s_i{}^2 = \left[\sum_{j=1}^{m_i}\hat{T}_{ij}{}^2 - \left(\sum_{j=1}^{m_i}\hat{T}_{ij}\right)^2 / m_i\right] / (m_i - 1)$

$s_{ij}{}^2 = \left[\sum_{k=1}^{b_{ij}}y_{ijk}{}^2 - \left(\sum_{k=1}^{b_{ij}}y_{ijk}\right)^2 / b_{ij}\right] / (b_{ij} - 1)$

3. The approximate 95% confidence interval for T:

$$\hat{T} \pm 1.959963984540\sqrt{v(\hat{T})}$$

NOTE: For a 90% confidence interval, replace 1.959963984540 with 1.644853626951 and for an 80% confidence interval, replace 1.959963984540 with 1.281551565545.

RHC Two-Stage Variable Sampling

For a discussion on the motivation behind the RHC (developed by Rao, Hartley, and Cochran) sampling procedure, refer to the **RHC Sample Selection** section, contained in the **Random Numbers** section of this manual. It provides a method of sample selection that allows sampling without replacement while maintaining the flavor of sampling using probability proportional to size. When the primary units (P.U.s) are selected, the size of each P.U. is considered rather than obtaining a simple random sample of P.U.s.

The size of each P.U. is rather arbitrary and can be the number of people, dollars, beds (for hospitals), and so forth. In general, you can expect improved precision using the RHC procedure if there is a high correlation between the size of each P.U. and the number of secondary units (S.U.s) within each P.U.. In other words, larger P.U.s should contain a larger number of S.U.s.

The P.U.s are selected using the **RHC Sample Selection** program. A random sample is then obtained for each selected P.U. and the variable(s) of interest (e.g., dollars in error) is/are recorded.

Example 5. (**Note**: This is the same example used in Example 8 in the **RHC Sample Selection** discussion contained in the **Random Numbers** section). In a particular region of the United Sates there are N = 90 universities (primary units) with Government research grants. Because these universities are so widespread, it was decided to use a cluster sample of n = 10 universities. Rather than audit all grants at a selected university, it was decided (based on available resources) to audit roughly 20% of the grants at each selected university. As a measure of the size for each university, the total grant dollars were used.

DATA: University ID, number of grants, total grant dollars (90 rows)

The data are contained in UNIVRHC.TXT.

OUTPUT: The 10 universities to use in the sample (see last page of computer output):

UNIV78, UNIV42, UNIV49, UNIV5, UNIV19,

UNIV38, UNIV62, UNIV28, UNIV60, UNIV75

Here there are 10 groups with 9 universities per group. The output file created by this program

is OUTRHC.TXT. This program is required by the RHC appraisal program.

Data set UNIVRHC.TXT and the program output are contained in the pages to follow.

Dataset UNIVRHC.TXT   < - - - continued - - - >   < - - - continued - - - >
                                    (1)   (2)
                                         (3)

UNIV1	42	8	UNIV31	52	11	UNIV61	66	13
UNIV2	21	4	UNIV32	66	14	UNIV62	77	18
UNIV3	63	13	UNIV33	25	5	UNIV63	31	7
UNIV4	74	16	UNIV34	60	12	UNIV64	46	9
UNIV5	51	11	UNIV35	19	4	UNIV65	32	7
UNIV6	43	9	UNIV36	24	5	UNIV66	68	14
UNIV7	57	11	UNIV37	44	9	UNIV67	41	9
UNIV8	49	10	UNIV38	76	17	UNIV68	28	6
UNIV9	63	13	UNIV39	41	9	UNIV69	66	14
UNIV10	18	4	UNIV40	77	18	UNIV70	31	7
UNIV11	64	13	UNIV41	37	8	UNIV71	27	6
UNIV12	56	11	UNIV42	63	12	UNIV72	33	7
UNIV13	19	4	UNIV43	52	11	UNIV73	23	4
UNIV14	44	9	UNIV44	76	17	UNIV74	71	15
UNIV15	20	4	UNIV45	51	10	UNIV75	75	16
UNIV16	34	7	UNIV46	23	4	UNIV76	47	10
UNIV17	25	6	UNIV47	24	5	UNIV77	50	10
UNIV18	38	9	UNIV48	68	15	UNIV78	37	7
UNIV19	72	16	UNIV49	34	7	UNIV79	77	18
UNIV20	46	10	UNIV50	49	10	UNIV80	49	10
UNIV21	44	9	UNIV51	55	11	UNIV81	76	17
UNIV22	64	13	UNIV52	38	9	UNIV82	66	14
UNIV23	45	9	UNIV53	72	16	UNIV83	28	6
UNIV24	55	11	UNIV54	51	10	UNIV84	77	17
UNIV25	29	7	UNIV55	71	15	UNIV85	27	6
UNIV26	36	7	UNIV56	59	12	UNIV86	75	17
UNIV27	40	9	UNIV57	23	4	UNIV87	71	15
UNIV28	78	18	UNIV58	57	11	UNIV88	59	12
UNIV29	49	10	UNIV59	53	11	UNIV89	71	15
UNIV30	60	12	UNIV60	64	13	UNIV90	72	16

Columns: (1) unit ID

(2) number of grants

(3) grant dollar amount (x $100,000) ◄ This is the size of the university.

```
                    DEPARTMENT OF HEALTH & HUMAN SERVICES
                         OIG – OFFICE OF AUDIT SERVICES
Date: 10/15/2004      GENERATION OF PRIMARY UNIT SAMPLE       Time: 12:52
                  NAME OF INPUT FILE: C:\TEMP\UNIVRHC.TXT

                          GROUPS OF PRIMARY UNITS

                     ********* GROUP 1 *********
                                    PRIMARY UNIT    SECONDARY
    PRIMARY UNIT IDENTIFICATION         SIZE         UNIVERSE
    ============================== =============  =============
    UNIV51                                  11              55
    UNIV44                                  17              76
    UNIV32                                  14              66
    UNIV78 <-- selected                      7              37
    UNIV79                                  18              77
    UNIV2                                    4              21
    UNIV52                                   9              38
    UNIV33                                   5              25
    UNIV47                                   5              24

        GROUP TOTALS:  9                    90             419

                     ********* GROUP 2 *********
                                    PRIMARY UNIT    SECONDARY
    PRIMARY UNIT IDENTIFICATION         SIZE         UNIVERSE
    ============================== =============  =============
    UNIV6                                    9              43
    UNIV42 <-- selected                     12              63
    UNIV65                                   7              32
    UNIV40                                  18              77
    UNIV45                                  10              51
    UNIV1                                    8              42
    UNIV80                                  10              49
    UNIV36                                   5              24
    UNIV70                                   7              31

        GROUP TOTALS:  9                    86             412
```

< Groups 3 Through 9 Are Omitted Here >

```
                     ********* GROUP 10 *********
                                    PRIMARY UNIT    SECONDARY
    PRIMARY UNIT IDENTIFICATION         SIZE         UNIVERSE
    ============================== =============  =============
    UNIV22                                  13              64
    UNIV39                                   9              41
    UNIV88                                  12              59
    UNIV55                                  15              71
    UNIV29                                  10              49
    UNIV75 <-- selected                     16              75
    UNIV87                                  15              71
    UNIV13                                   4              19
    UNIV53                                  16              72

        GROUP TOTALS:  9                   110             521
```

```
                    DEPARTMENT OF HEALTH & HUMAN SERVICES
                       OIG – OFFICE OF AUDIT SERVICES
Date: 10/15/2004       GENERATION OF PRIMARY UNIT SAMPLE          Time: 12:52
             NAME OF OUTPUT FILE: C:\TEMP\OutRHCsummary.txt

  FIRST SEED NUMBER:      100.00     SECOND SEED NUMBER:       200.00

      NUMBER OF PRIMARY UNITS IN THE POPULATION:     90
      NUMBER OF PRIMARY UNITS SAMPLED:               10

                                                                      UNITS
                         SECONDARY       PRIMARY                        IN
PRIMARY UNIT ID          UNIVERSE       UNIT SIZE      GROUP SIZE      GROUP
========================= ============= ============= ============= =====
UNIV78                           37             7             90         9
UNIV42                           63            12             86         9
UNIV49                           34             7             96         9
UNIV5                            51            11             84         9
UNIV19                           72            16             89         9
UNIV38                           76            17             89         9
UNIV62                           77            18             92         9
UNIV28                           78            18            115         9
UNIV60                           64            13             99         9
UNIV75                           75            16            110         9
```

Example--continued. The sample size for each selected university (P.U.) was chosen to be

approximately 20% of the universe size. This leads to the following table where M_i is the total

number of grants at the i-th university, and m_i is the number of audited grants at the i-th

university.

University	M_i	m_i
UNIV78	37	7
UNIV42	63	13
UNIV49	34	7
UNIV5	51	10
UNIV19	72	14
UNIV38	76	15
UNIV62	77	15
UNIV28	78	16
UNIV60	64	13
UNIV75	75	15
		125

Data from these 125 secondary units (grants) were obtained by recording the total amount that

was charged to each grant after the scheduled completion of this grant (dollars in error). The

error amounts (in thousands of dollars) for each grant are contained in data set

DATARHC2.TXT.

The output file created using the sample selection program is also used as input to the appraisal

program in this two-step procedure. This file (PRIMRHC2.TXT), along with

DATARHC2.TXT, are used as the input files for the **RHC Two-Stage** appraisal program. Both

files are listed on the next page and the computer output from the appraisal program immediately

follows. The illustrated data file (DATARHC2.TXT) contains the data for the first two

universities (primary units) and the last two universities.

Dataset DATARHC2.TXT

1   9
2   2
3   9
4   6
5   0
6   5
7   7
8   2
9   7
10 6
11 0
12 6
13 0
14 3
15 4
16 1
17 13
18 8
19 0
20 6
21 11
22 8
23 8
24 0
.
.
.
100   7
101   10
102   2
103   6
104   0
105   8
106   4
107   0
108   10
109   3
110   2
111   5
112   10
113   0
114   0
115   0
116   0

```
117    0
118    8
119    9
120    0
121    2
122    8
123    4
124    6
125    2
```

Output/Input file PRIMRHC2.TXT

```
UNIV78              37      7      7       90      9
UNIV42              63     13     12       86      9
UNIV49              34      7      7       96      9
UNIV5               51     10     11       84      9
UNIV19              72     14     16       89      9
UNIV38              76     15     17       89      9
UNIV62              77     15     18       92      9
UNIV28              78     16     18      115      9
UNIV60              64     13     13       99      9
UNIV75              75     15     15      110      9
```

NOTE: File PRIMRHC2.TXT was created by adding the third column containing sample sizes to the output file created by the **RHC Sample Selection** program.

```
                    DEPARTMENT OF HEALTH & HUMAN SERVICES
                        OIG - OFFICE OF AUDIT SERVICES
Date: 10/26/2004        RHC TWO-STAGE VARIABLE APPRAISAL         Time: 13:26
                         AUDIT/REVIEW: RHC 2-Stage

              DATA FILE USED: C:\TEMP\DATARHC2.TXT


              PRIMARY   SAMPLE   ==DIFFERENCE===    NUMBER OF
               UNIT      SIZE     SAMPLE TOTAL    NONZERO ITEMS

              =======   ======   =============   =============
                 1        7          38.00             6
                 2       13          56.00            10
                 3        7          33.00             4
                 4       10          55.00             8
                 5       14          67.00            13
                 6       15          76.00            14
                 7       15          79.00            14
                 8       16          84.00            13
                 9       13          61.00            11
                10       15          54.00             9

              TOTALS    125         603.00           102


          PRIMARY UNIT FILE USED: C:\TEMP\PRIMRHC2.txt
```

```
                                                                       UNITS
P.U.                          SECONDARY      PRIMARY                     IN
NBR   PRIMARY UNIT ID         UNIVERSE      UNIT SIZE    GROUP SIZE    GROUP
====  =========================  =============  =============  =============  =====
 1    UNIV78                        37             7            90          9
 2    UNIV42                        63            12            86          9
 3    UNIV49                        34             7            96          9
 4    UNIV5                         51            11            84          9
 5    UNIV19                        72            16            89          9
 6    UNIV38                        76            17            89          9
 7    UNIV62                        77            18            92          9
 8    UNIV28                        78            18           115          9
 9    UNIV60                        64            13            99          9
10    UNIV75                        75            15           110          9

      TOTALS                       627           134           950         90
```

* Note: 2,582 is the product of 5.43, 37, and 12.857. This is the point estimate for P.U. 1.

```
                        --- POINT ESTIMATES —

                              ==DIFFERENCE===
     P.U.  SAMPLE                SECONDARY      SIZES        POINT
     NBR    SIZE   SAMPLE MEAN    UNIVERSE      RATIO      ESTIMATE
     ====  ======  =============  =============  ========  =============
       1      7        5.43            37        12.857      2,582    *
       2     13        4.31            63         7.167      1,945
       3      7        4.71            34        13.714      2,198
       4     10        5.50            51         7.636      2,142
       5     14        4.79            72         5.563      1,917
       6     15        5.07            76         5.235      2,016
       7     15        5.27            77         5.111      2,073
```

8	16	5.25	78	6.389	2,616
9	13	4.69	64	7.615	2,287
10	15	3.60	75	7.333	1,980
TOTALS:	125		627		21,756

Note: 21,756 is the point estimate of the universe total ($\hat{T}$).

--- VARIANCE COMPONENTS ---

P.U. NBR	WITHIN VARIANCE	BETWEEN VARIANCE	TOTAL VARIANCE
1	23,689	283,356	307,044
2	25,870	659	26,529
3	38,797	0	38,797
4	22,443	53,230	75,674
5	19,701	15,570	35,271
6	13,977	523	14,500
7	33,107	1,192	34,299
8	29,738	247	29,985
9	23,704	369	24,074
10	31,366	247,927	279,293
TOTALS:	262,392	603,072	865,464

Note: 865,464 is equal to $v(\hat{T})$.

PRIMARY UNITS SAMPLED:	10
PRIMARY UNITS NOT SAMPLED:	80
PRIMARY UNITS IN POPULATION:	90

POINT ESTIMATE OF POPULATION TOTAL: 21,756

STANDARD ERROR   930

CONFIDENCE LIMITS
80% CONFIDENCE LEVEL

LOWER LIMIT	20,564
UPPER LIMIT	22,948
PRECISION AMOUNT	1,192
PRECISION PERCENT	5.48%
Z-VALUE USED	1.281551565545

90% CONFIDENCE LEVEL

LOWER LIMIT	20,226
UPPER LIMIT	23,286
PRECISION AMOUNT	1,530
PRECISION PERCENT	7.03%
Z-VALUE USED	1.644853626951

95% CONFIDENCE LEVEL

LOWER LIMIT	19,933
UPPER LIMIT	23,579
PRECISION AMOUNT	1,823
PRECISION PERCENT	8.38%
Z-VALUE USED	1.959963984540

Discussion: The (highlighted) estimate of the universe total on the previous page obtained

using formula 1 is:

$$\hat{T} = \text{(Estimate of group 1 total)} + \cdots + \text{(Estimate of group 10 total)}$$

$$= (90/7)(5.43)(37) + (86/12)(4.31)(63) + \cdots + (110/15)(3.60)(75)$$

$$= 2{,}582 + 1{,}945 + \cdots + 1{,}980$$

$$= 21{,}756 \ (\$21{,}756{,}000)$$

Using formula 2, the estimated variance of $\hat{T}$ is:

$$v(\hat{T}) = 865{,}464$$

and the estimated standard error of $\hat{T}$ is the square root of 865,464; that is, 930 (highlighted on

the previous page).

The approximate 95% confidence interval is:

$$21{,}756 \pm 1.959963984540(930)$$

$$21{,}756 \pm 1{,}823$$

that is, 19,933 to 23,579 ($19,933,000 to $23,579,000).

FORMULAS

Definitions

1.   P.U. stands for "primary unit" and S.U. is "secondary unit"

2. A_i = size of i-th P.U.

3. S_i = (size of i-th P.U.)/(size of entire population) = A_i/(size of entire population)

4. B_i = total size for i-th group

5. π_i = (total size for i-th group)/(size of entire population) = B_i/(size of entire population)

6.   N = number of P.U.s in the population

7. N_i = number of P.U.s in the i-th group

8.   n = number of P.U.s in the sample

9. M_i = number of S.U.s in the i-th sampled P.U. (population)

10. m_i = number of S.U.s in the i-th sampled P.U. (sample)

Estimate of population total (T)

$$\hat{T} = \sum_{i=1}^{n} \left(\frac{B_i}{A_i} \right) M_i \bar{y}_i$$

where $\bar{y}_i$ = average of m_i sampled S.U.s and B_i/A_i is labeled SIZES RATIO in the computer

output.

Estimated variance of $\hat{T}$

$$v(\hat{T}) = V_1 + V_2 \quad \text{where}$$

$$V_1 = \frac{\sum\limits_{i=1}^{n} N_i^2 - N}{N^2 - \sum\limits_{i=1}^{n} N_i^2} \sum\limits_{i=1}^{n} \pi_i \left(\frac{M_i \bar{y}_i}{S_i} - \hat{T} \right)^2$$

and

$$V_2 = \sum\limits_{i=1}^{n} \pi_i \frac{M_i}{S_i} \left(\frac{M_i - m_i}{m_i} \right) s_i^2$$

where s_i^2 = variance of the m_i sampled S.U.s.

NOTE: The estimated standard error of $\hat{T}$ is $\sqrt{v(\hat{T})}$

Approximate 95% confidence interval for the population total (T)

$$\hat{T} \pm 1.959963984540 \sqrt{v(\hat{T})}$$

NOTE: For a 90% confidence interval, replace 1.959963984540 with 1.644853626951 and for an 80% confidence interval replace 1.959963984540 with 1.281551565545.

RHC Three-Stage Variable Sampling

The RHC sampling procedure can used for a three-stage design.

The steps for such a procedure are the following:

1. A sample of primary units (clusters) is obtained as in the one- and two-stage procedures. The size of the primary units is considered for this sample, where pps sampling is used for each group of primary units.

2. A sample of secondary units is obtained within each chosen primary unit by partitioning the primary unit into random groups. The group sizes are chosen to be as nearly equal as possible. Using pps sampling and the size of each secondary unit, one secondary unit is chosen from each of the secondary groups.

3. A random sample of third-stage units is obtained for each of the chosen secondary units. No attention is paid to "size" here. This is a random sample.

Prior to running the appraisal program, the user must run the **RHC Sample Selection** program in the OAS software.

Example 6. The situation discussed in Example 4 in the **Three-Stage Unrestricted** section will be appraised using the RHC methodology. For this example, the stages are:

> Stage 1: REGION (select 4 out of 12 regions)
>
> Stage 2: UNIVERSITY (select 10 from each selected region)
>
> Stage 3: GRANT (select approximately 20% of all grants at each university)

Selection of Primary Units

A file must be constructed containing (for each region) (1) the number of secondary units (universities) in this region and (2) the size of this region (total dollars of grants). This file is GRANTSPU.TXT. The selected regions are 4, 6, 8, and 10.

NOTE: Seed values of 100 and 200 were used to select the primary units. In practice, it is
recommended that the user not set these seed values.

--- Data set GRANTSPU.TXT ---

	(A)	(B)
REGION1	117	1250
REGION2	63	610
REGION3	91	720
REGION4	123	1320
REGION5	107	1160
REGION6	116	1240
REGION7	102	960
REGION8	118	1300
REGION9	122	1320
REGION10	85	640
REGION11	94	930
REGION12	62	550

NOTE: It is okay to set the number of S.U.s [column (A)]
equal to one in this file. The actual number of
S.U.s <u>must</u> be known for the selected P.U.s. The
correct number of S.U.s must then be inserted into
file GRANTSPUOUT.TXT (the highlighted
values).

Columns: (A) number of universities (S.U.s)
(B) size of each P.U. (total grant amount x $100,000)

--- Data set GRANTSPUOUT.TXT ---

REGION6	116	1240	3100	3
REGION4	123	1320	3410	3
REGION8	118	1300	3170	3
REGION10	85	640	2320	3

```
                         DEPARTMENT OF HEALTH & HUMAN SERVICES
                             OIG – OFFICE OF AUDIT SERVICES
Date: 10/22/2004        GENERATION OF PRIMARY UNIT SAMPLE          Time: 15:09
                     NAME OF INPUT FILE: C:\TEMP\GRANTSPU.TXT

                            GROUPS OF PRIMARY UNITS

                      ********* GROUP 1 *********
                                         PRIMARY UNIT    SECONDARY
           PRIMARY UNIT IDENTIFICATION       SIZE         UNIVERSE
           ============================  =============  =============
           REGION2                               610             63
           REGION6 <-- Selected                1,240            116
           REGION1                             1,250            117

               GROUP TOTALS:  3                3,100            296

                      ********* GROUP 2 *********
                                         PRIMARY UNIT    SECONDARY
           PRIMARY UNIT IDENTIFICATION       SIZE         UNIVERSE
           ============================  =============  =============
           REGION4 <-- Selected                1,320            123
           REGION5                             1,160            107
           REGION11                              930             94

               GROUP TOTALS:  3                3,410            324

                      ********* GROUP 3 *********
                                         PRIMARY UNIT    SECONDARY
           PRIMARY UNIT IDENTIFICATION       SIZE         UNIVERSE
           ============================  =============  =============
           REGION12                              550             62
           REGION8 <-- Selected                1,300            118
           REGION9                             1,320            122

               GROUP TOTALS:  3                3,170            302

                      ********* GROUP 4 *********
                                         PRIMARY UNIT    SECONDARY
           PRIMARY UNIT IDENTIFICATION       SIZE         UNIVERSE
           ============================  =============  =============
           REGION3                               720             91
           REGION7                               960            102
           REGION10 <-- Selected                 640             85

               GROUP TOTALS:  3                2,320            278

    FIRST SEED NUMBER:     100.00     SECOND SEED NUMBER:     200.00
```

In practice, do not set these seed values.

```
       NUMBER OF PRIMARY UNITS IN THE POPULATION:    12
       NUMBER OF PRIMARY UNITS SAMPLED:               4
```

< Program output - continued >

PRIMARY UNIT ID	SECONDARY UNIVERSE	PRIMARY UNIT SIZE	GROUP SIZE	UNITS IN GROUP
=========================	=============	=============	=============	=====
REGION6	116	1,240	3,100	3
REGION4	123	1,320	3,410	3
REGION8	118	1,300	3,170	3
REGION10	85	640	2,320	3

NOTE: The above four lines make up file GRANTSPUOUT.TXT

Selection of Secondary Units

The input for three-stage RHC can be greatly simplified if you only obtain information for each

selected primary unit (that is, Regions 4, 6, 8, and 10 here). The information consists of the size

of each secondary unit (university, here) and the number of third-stage units in the universe for

each secondary unit. This input is shown in files REGION4.TXT, REGION6.TXT,

REGION8.TXT, and REGION10.TXT. Each line in the files contains the number of third-stage

units (grants) in the universe and the size of that secondary unit (total grant dollars x 100,000), in

that order. After each of these four files is the computer output using the **RHC Sample**

Selection program. A sample of 10 universities is selected for each region. The results are:

REGION	UNIVERSITIES
4	85, 46, 7, 82, 30, 34, 27, 66, 65, 80
6	113, 43, 78, 104, 89, 112, 30, 65, 3, 99
8	112, 6, 7, 93, 75, 111, 62, 115, 70, 99
10	78, 43, 7, 73, 55, 33, 10, 59, 64, 39

The previous five program runs (one at the primary level and four at the secondary level) created

five output files. Using a word processor or spreadsheet, these files can be joined to form one of

the input files (the one containing primary/secondary unit information) for the three-stage RHC

program which calculates the confidence interval. The file for this example is

PUSURHC3.TXT.

Data set REGION4.TXT   < - - - continued - - - >   < - - - continued - - - >

(1)	(2)	(3)
UNIV1	52	11
UNIV2	37	9
UNIV3	38	9
UNIV4	20	5
UNIV5	69	15
UNIV6	69	15
UNIV7	77	17
UNIV8	32	7
UNIV9	49	10
UNIV10	73	15
UNIV11	21	5
UNIV12	62	13
UNIV13	55	11
UNIV14	59	12
UNIV15	55	11
UNIV16	36	8
UNIV17	51	11
UNIV18	26	7
UNIV19	25	6
UNIV20	73	15
UNIV21	71	15
UNIV22	47	10
UNIV23	34	8
UNIV24	25	6
UNIV25	39	9
UNIV26	49	10
UNIV27	76	16
UNIV28	21	5
UNIV29	33	8
UNIV30	54	11
UNIV31	45	10
UNIV32	74	16
UNIV33	69	14
UNIV34	50	10
UNIV35	29	7
UNIV36	56	12
UNIV37	64	14
UNIV38	66	14
UNIV39	63	14
UNIV40	57	12
UNIV41	71	15
UNIV42	45	10
UNIV43	21	5
UNIV44	46	10
UNIV45	48	10
UNIV46	44	9
UNIV47	71	15
UNIV48	67	14
UNIV49	23	6
UNIV50	54	11

UNIV51	62	13
UNIV52	52	11
UNIV53	56	11
UNIV54	70	15
UNIV55	41	9
UNIV56	65	14
UNIV57	76	16
UNIV58	30	7
UNIV59	75	16
UNIV60	27	7
UNIV61	36	8
UNIV62	61	13
UNIV63	58	12
UNIV64	61	13
UNIV65	62	14
UNIV66	76	16
UNIV67	71	15
UNIV68	34	8
UNIV69	62	13
UNIV70	23	6
UNIV71	28	7
UNIV72	46	10
UNIV73	62	14
UNIV74	67	14
UNIV75	25	6
UNIV76	24	6
UNIV77	57	12
UNIV78	44	10
UNIV79	73	16
UNIV80	70	15
UNIV81	45	10
UNIV82	52	11
UNIV83	34	8
UNIV84	59	12
UNIV85	54	11
UNIV86	31	7
UNIV87	69	14
UNIV88	22	6
UNIV89	47	10
UNIV90	57	12
UNIV91	31	7
UNIV92	73	15
UNIV93	52	11
UNIV94	22	6
UNIV95	22	6
UNIV96	29	7
UNIV97	56	12
UNIV98	74	16
UNIV99	43	9
UNIV100	57	12

UNIV101	34	8
UNIV102	28	7
UNIV103	73	15
UNIV104	65	14
UNIV105	68	14
UNIV106	28	7
UNIV107	55	11
UNIV108	37	9
UNIV109	54	11
UNIV110	47	10
UNIV111	44	9
UNIV112	24	6
UNIV113	50	10
UNIV114	52	11
UNIV115	66	14
UNIV116	50	10
UNIV117	66	14
UNIV118	34	8
UNIV119	73	16
UNIV120	37	8
UNIV121	42	9
UNIV122	59	12
UNIV123	45	11

NOTE: This file has 123 lines.

Columns: (1) unit ID
         (2) number of grants

(3) size of university (grant amount x $100,000)

```
                    DEPARTMENT OF HEALTH & HUMAN SERVICES
                         OIG – OFFICE OF AUDIT SERVICES
Date: 10/25/2004      GENERATION OF SECONDARY UNIT SAMPLE        Time: 14:21
                    NAME OF INPUT FILE: C:\TEMP\REGION4.TXT


                         GROUPS OF SECONDARY UNITS


                    ********* GROUP 1 *********
                                       SECONDARY UNIT    3RD STAGE
        SECONDARY UNIT IDENTIFICATION       SIZE         UNIVERSE
        ============================== =============   =============
        UNIV57                                   16               76
        UNIV48                                   14               67
        UNIV35                                    7               29
        UNIV107                                  11               55
        UNIV85  <-- Selected                     11               54
        UNIV103                                  15               73
        UNIV86                                    7               31
        UNIV2                                     9               37
        UNIV81                                   10               45
        UNIV58                                    7               30
        UNIV36                                   12               56
        UNIV49                                    6               23

           GROUP TOTALS: 12                     125              576


                    ********* GROUP 2 *********
                                       SECONDARY UNIT    3RD STAGE
        SECONDARY UNIT IDENTIFICATION       SIZE         UNIVERSE
        ============================== =============   =============
        UNIV52                                   11               52
        UNIV6                                    15               69
        UNIV46  <-- Selected                      9               44
        UNIV69                                   13               62
        UNIV108                                   9               37
        UNIV44                                   10               46
        UNIV50                                   11               54
        UNIV121                                   9               42
        UNIV1                                    11               52
        UNIV43                                    5               21
        UNIV87                                   14               69
        UNIV39                                   14               63

           GROUP TOTALS: 12                     131              611
```

< GROUPS 3 THROUGH 9 ARE OMITTED HERE >

```
                    ********* GROUP 10 *********
                                       SECONDARY UNIT    3RD STAGE
        SECONDARY UNIT IDENTIFICATION       SIZE         UNIVERSE
        ============================== =============   =============
        UNIV53                                   11               56
        UNIV24                                    6               25
        UNIV42                                   10               45
        UNIV120                                   8               37
        UNIV105                                  14               68
        UNIV97                                   12               56
```

```
              UNIV119                                    16            73
              UNIV32                                     16            74
              UNIV80 <-- Selected                        15            70
              UNIV96                                      7            29
              UNIV13                                     11            55
              UNIV62                                     13            61
              UNIV59                                     16            75

              GROUP TOTALS: 13                          155           724
```

DEPARTMENT OF HEALTH & HUMAN SERVICES
OIG – OFFICE OF AUDIT SERVICES

Date: 10/25/2004   GENERATION OF SECONDARY UNIT SAMPLE   Time: 14:21
NAME OF OUTPUT FILE: C:\TEMP\OutRegion4.txt

FIRST SEED NUMBER: 100.00   SECOND SEED NUMBER: 200.00

NUMBER OF SECONDARY UNITS IN THE POPULATION: 123
NUMBER OF SECONDARY UNITS SAMPLED: 10

SECONDARY UNIT ID	3RD STAGE UNIVERSE	SECONDARY UNIT SIZE	GROUP SIZE	UNITS IN GROUP
UNIV85	54	11	125	12
UNIV46	44	9	131	12
UNIV7	77	17	119	12
UNIV82	52	11	129	12
UNIV30	54	11	141	12
UNIV34	50	10	140	12
UNIV27	76	16	138	12
UNIV66	76	16	128	13
UNIV65	62	14	125	13
UNIV80	70	15	155	13

UNIV1	56	10
UNIV2	27	5
UNIV3	56	11
UNIV4	23	5
UNIV5	72	13
UNIV6	24	5
UNIV7	61	11
UNIV8	65	12
UNIV9	68	13
UNIV10	40	8
UNIV11	64	12
UNIV12	66	13
UNIV13	80	14
UNIV14	53	9
UNIV15	36	7
UNIV16	53	10
UNIV17	47	9
UNIV18	73	14
UNIV19	41	8
UNIV20	58	11
UNIV21	45	9
UNIV22	43	8
UNIV23	56	10
UNIV24	35	7
UNIV25	34	7
UNIV26	65	13
UNIV27	78	14
UNIV28	35	7
UNIV29	31	6
UNIV30	58	11
UNIV31	29	6
UNIV32	76	14
UNIV33	57	10
UNIV34	42	8
UNIV35	69	13
UNIV36	58	11
UNIV37	31	6
UNIV38	33	6
UNIV39	40	8
UNIV40	51	9
UNIV41	60	11
UNIV42	78	14
UNIV43	39	7
UNIV44	46	9
UNIV45	58	11
UNIV46	59	11
UNIV47	53	10
UNIV48	57	10
UNIV49	28	6
UNIV50	63	12
UNIV51	31	6
UNIV52	60	11
UNIV53	30	6
UNIV54	30	6
UNIV55	40	8
UNIV56	26	5
UNIV57	24	5
UNIV58	44	8

<-- continued -->

UNIV59	67	13
UNIV60	56	10
UNIV61	33	7
UNIV62	40	8
UNIV63	68	13
UNIV64	70	13
UNIV65	57	10
UNIV66	40	7
UNIV67	54	10
UNIV68	65	12
UNIV69	62	12
UNIV70	28	5
UNIV71	56	10
UNIV72	41	8
UNIV73	31	6
UNIV74	31	6
UNIV75	46	9
UNIV76	38	7
UNIV77	62	12
UNIV78	63	12
UNIV79	50	9
UNIV80	53	9
UNIV81	39	7
UNIV82	39	7
UNIV83	39	7
UNIV84	25	5
UNIV85	67	13
UNIV86	47	9
UNIV87	54	10
UNIV88	50	9
UNIV89	35	7
UNIV90	66	13
UNIV91	65	12
UNIV92	71	13
UNIV93	29	6
UNIV94	74	14
UNIV95	66	13
UNIV96	71	13
UNIV97	43	8
UNIV98	62	11
UNIV99	80	14
UNIV100	57	11
UNIV101	22	5
UNIV102	33	6
UNIV103	78	5
UNIV104	25	9
UNIV105	76	8
UNIV106	39	5
UNIV107	48	5
UNIV108	54	7
UNIV109	63	12
UNIV110	28	8
UNIV111	69	8
UNIV112	27	10
UNIV113	33	8
UNIV114	52	7
UNIV115	33	15
UNIV116	23	10

Data set REGION6.TXT

NOTE: This file has 116 lines.

```
                    DEPARTMENT OF HEALTH & HUMAN SERVICES
                        OIG – OFFICE OF AUDIT SERVICES
Date: 10/25/2004      GENERATION OF SECONDARY UNIT SAMPLE        Time: 13:57
                   NAME OF INPUT FILE: C:\TEMP\REGION6.txt

                          GROUPS OF SECONDARY UNITS

                     ********* GROUP 1 *********
                                        SECONDARY UNIT    3RD STAGE
          SECONDARY UNIT IDENTIFICATION        SIZE       UNIVERSE
          ============================== =============  =============
          UNIV52                                   11           60
          UNIV45                                   11           58
          UNIV32                                   14           76
          UNIV86                                    9           47
          UNIV113 <-- Selected                      8           33
          UNIV85                                   13           67
          UNIV109                                  12           63
          UNIV87                                   10           54
          UNIV2                                     5           27
          UNIV80                                    9           53
          UNIV53                                    6           30

              GROUP TOTALS: 11                    108          568

                     ********* GROUP 2 *********
                                        SECONDARY UNIT    3RD STAGE
          SECONDARY UNIT IDENTIFICATION        SIZE       UNIVERSE
          ============================== =============  =============
          UNIV33                                   10           57
          UNIV48                                   10           57
          UNIV6                                     5           24
          UNIV43 <-- Selected                       7           39
          UNIV68                                   12           65
          UNIV41                                   11           60
          UNIV46                                   11           59
          UNIV1                                    10           56
          UNIV40                                    9           51
          UNIV88                                    9           50
          UNIV36                                   11           58

              GROUP TOTALS: 11                    105          576
```

< GROUPS 3 THROUGH 9 ARE OMITTED HERE >

```
                     ********* GROUP 10 *********
                                        SECONDARY UNIT    3RD STAGE
          SECONDARY UNIT IDENTIFICATION        SIZE       UNIVERSE
          ============================== =============  =============
          UNIV20                                   11           58
          UNIV22                                    8           43
          UNIV39                                    8           40
          UNIV111                                   8           69
          UNIV100                                  11           57
          UNIV29                                    6           31
          UNIV105                                   8           76
          UNIV79                                    9           50
          UNIV99 <-- Selected                      14           80
```

```
        UNIV13                                   14            80
        UNIV60                                   10            56
        UNIV54                                    6            30


            GROUP TOTALS: 12                     113           670
```

```
                 DEPARTMENT OF HEALTH & HUMAN SERVICES
                    OIG - OFFICE OF AUDIT SERVICES
   Date: 10/25/2004      GENERATION OF SECONDARY UNIT SAMPLE      Time: 13:57
                 NAME OF OUTPUT FILE: C:\TEMP\OutRegion6.txt


    FIRST SEED NUMBER:     100.00    SECOND SEED NUMBER:      200.00


       NUMBER OF SECONDARY UNITS IN THE POPULATION:    116
       NUMBER OF SECONDARY UNITS SAMPLED:               10
```

| | | | | UNITS |
| | 3RD STAGE | SECONDARY | | IN |
SECONDARY UNIT ID	UNIVERSE	UNIT SIZE	GROUP SIZE	GROUP
UNIV113	33	8	108	11
UNIV43	39	7	105	11
UNIV78	63	12	104	11
UNIV104	25	9	96	11
UNIV89	35	7	124	12
UNIV112	27	10	108	12
UNIV30	58	11	95	12
UNIV65	57	10	109	12
UNIV3	56	11	115	12
UNIV99	80	14	113	12

Data set REGION8.TXT

UNIV1	72	15
UNIV2	44	10
UNIV3	43	10
UNIV4	55	12
UNIV5	27	7
UNIV6	34	8
UNIV7	51	11
UNIV8	42	10
UNIV9	54	12
UNIV10	25	6
UNIV11	82	17
UNIV12	65	14
UNIV13	33	8
UNIV14	48	10
UNIV15	32	8
UNIV16	82	17
UNIV17	35	8
UNIV18	54	12
UNIV19	34	8
UNIV20	62	14
UNIV21	26	6
UNIV22	31	7
UNIV23	58	13
UNIV24	61	13
UNIV25	61	14
UNIV26	54	12
UNIV27	53	11
UNIV28	56	12
UNIV29	57	12
UNIV30	26	6
UNIV31	25	5
UNIV32	37	9
UNIV33	79	16
UNIV34	60	13
UNIV35	57	12
UNIV36	27	7
UNIV37	31	7
UNIV38	75	15
UNIV39	26	6
UNIV40	36	9
UNIV41	36	9
UNIV42	49	10
UNIV43	83	17
UNIV44	71	15
UNIV45	31	7
UNIV46	42	10
UNIV47	62	14
UNIV48	54	11
UNIV49	31	7
UNIV50	80	16

< - - - continued - - - >

UNIV51	77	16
UNIV52	36	9
UNIV53	75	16
UNIV54	68	15
UNIV55	34	8
UNIV56	55	12
UNIV57	42	10
UNIV58	36	9
UNIV59	36	9
UNIV60	66	15
UNIV61	61	13
UNIV62	64	14
UNIV63	72	15
UNIV64	65	14
UNIV65	58	13
UNIV66	49	11
UNIV67	30	7
UNIV68	75	16
UNIV69	33	8
UNIV70	65	14
UNIV71	55	12
UNIV72	38	9
UNIV73	36	9
UNIV74	60	13
UNIV75	52	11
UNIV76	65	14
UNIV77	49	10
UNIV78	27	7
UNIV79	48	10
UNIV80	36	9
UNIV81	66	15
UNIV82	62	14
UNIV83	70	15
UNIV84	68	15
UNIV85	53	11
UNIV86	38	9
UNIV87	35	8
UNIV88	36	9
UNIV89	26	6
UNIV90	26	6
UNIV91	51	11
UNIV92	25	5
UNIV93	54	11
UNIV94	56	12
UNIV95	81	17
UNIV96	73	15
UNIV97	44	10
UNIV98	50	11
UNIV99	60	13
UNIV100	31	7

< - - - continued - - - >

UNIV101	24	5
UNIV102	26	6
UNIV103	40	10
UNIV104	77	16
UNIV105	27	6
UNIV106	65	15
UNIV107	61	13
UNIV108	36	9
UNIV109	26	6
UNIV110	38	9
UNIV111	84	17
UNIV112	75	16
UNIV113	26	6
UNIV114	45	10
UNIV115	59	13
UNIV116	59	13
UNIV117	57	12
UNIV118	58	12

NOTE: This file has 118 lines.

```
                  DEPARTMENT OF HEALTH & HUMAN SERVICES
                     OIG – OFFICE OF AUDIT SERVICES
Date: 10/25/2004      GENERATION OF SECONDARY UNIT SAMPLE        Time: 14:03
                 NAME OF INPUT FILE: C:\TEMP\REGION8.TXT


                      GROUPS OF SECONDARY UNITS


                 ********* GROUP 1 *********
                                      SECONDARY UNIT    3RD STAGE
      SECONDARY UNIT IDENTIFICATION       SIZE          UNIVERSE
      ============================== ============== ==============
      UNIV54                                     15             68
      UNIV46                                     10             42
      UNIV33                                     16             79
      UNIV86                                      9             38
      UNIV112 <-- Selected                       16             75
      UNIV85                                     11             53
      UNIV108                                     9             36
      UNIV87                                      8             35
      UNIV2                                      10             44
      UNIV55                                      8             34
      UNIV34                                     13             60

          GROUP TOTALS: 11                     125            564

                 ********* GROUP 2 *********
                                      SECONDARY UNIT    3RD STAGE
      SECONDARY UNIT IDENTIFICATION       SIZE          UNIVERSE
      ============================== ============== ==============
      UNIV47                                     14             62
      UNIV50                                     16             80
      UNIV6 <-- Selected                          8             34
      UNIV44                                     15             71
      UNIV68                                     16             75
      UNIV42                                     10             49
      UNIV48                                     11             54
      UNIV1                                      15             72
      UNIV41                                      9             36
      UNIV89                                      6             26
      UNIV37                                      7             31

          GROUP TOTALS: 11                     127            590


          < GROUPS 3 THROUGH 9 ARE OMITTED HERE >


                 ********* GROUP 10 *********
                                      SECONDARY UNIT    3RD STAGE
      SECONDARY UNIT IDENTIFICATION       SIZE          UNIVERSE
      ============================== ============== ==============
      UNIV21                                      6             26
      UNIV23                                     13             58
      UNIV40                                      9             36
      UNIV110                                     9             38
      UNIV100                                     7             31
      UNIV30                                      6             26
      UNIV104                                    16             77
      UNIV81                                     15             66
      UNIV99 <-- Selected                        13             60
      UNIV13                                      8             33
```

```
        UNIV60                                    15            66
        UNIV56                                    12            55


             GROUP TOTALS: 12                    129           572
```

DEPARTMENT OF HEALTH & HUMAN SERVICES
OIG – OFFICE OF AUDIT SERVICES
Date: 10/25/2004   GENERATION OF SECONDARY UNIT SAMPLE   Time: 14:03
NAME OF OUTPUT FILE: C:\TEMP\OutRegion8.txt

FIRST SEED NUMBER: 100.00   SECOND SEED NUMBER: 200.00

     NUMBER OF SECONDARY UNITS IN THE POPULATION: 118
     NUMBER OF SECONDARY UNITS SAMPLED: 10

SECONDARY UNIT ID	3RD STAGE UNIVERSE	SECONDARY UNIT SIZE	GROUP SIZE	UNITS IN GROUP
UNIV112	75	16	125	11
UNIV6	34	8	127	11
UNIV7	51	11	120	12
UNIV93	54	11	136	12
UNIV75	52	11	126	12
UNIV111	84	17	134	12
UNIV62	64	14	123	12
UNIV115	59	13	137	12
UNIV70	65	14	143	12
UNIV99	60	13	129	12

Data set
REGION10.TXT

<--continued -->

Note: This file has 85 lines.

UNIV1	34	6		UNIV46	78	12
UNIV2	32	5		UNIV47	72	11
UNIV3	69	10		UNIV48	30	5
UNIV4	23	4		UNIV49	47	8
UNIV5	60	9		UNIV50	52	8
UNIV6	72	11		UNIV51	24	4
UNIV7	56	9		UNIV52	26	4
UNIV8	28	5		UNIV53	22	4
UNIV9	38	6		UNIV54	57	9
UNIV10	60	9		UNIV55	78	12
UNIV11	58	9		UNIV56	62	10
UNIV12	37	6		UNIV57	57	9
UNIV13	70	10		UNIV58	68	10
UNIV14	37	6		UNIV59	52	8
UNIV15	81	12		UNIV60	54	9
UNIV16	53	9		UNIV61	41	7
UNIV17	63	10		UNIV62	61	10
UNIV18	32	5		UNIV63	79	12
UNIV19	33	5		UNIV64	50	8
UNIV20	37	6		UNIV65	54	9
UNIV21	77	11		UNIV66	53	9
UNIV22	52	8		UNIV67	40	7
UNIV23	63	10		UNIV68	44	7
UNIV24	41	7		UNIV69	39	7
UNIV25	45	8		UNIV70	72	11
UNIV26	34	6		UNIV71	76	11
UNIV27	61	10		UNIV72	34	5
UNIV28	70	10		UNIV73	27	5
UNIV29	34	5		UNIV74	40	7
UNIV30	22	4		UNIV75	41	7
UNIV31	66	10		UNIV76	25	4
UNIV32	69	10		UNIV77	41	7
UNIV33	65	10		UNIV78	39	7
UNIV34	26	4		UNIV79	58	9
UNIV35	43	7		UNIV80	71	11
UNIV36	65	10		UNIV81	37	6
UNIV37	80	12		UNIV82	30	5
UNIV38	74	11		UNIV83	78	12
UNIV39	38	6		UNIV84	59	9
UNIV40	43	7		UNIV85	29	5
UNIV41	47	8				
UNIV42	59	9				
UNIV43	42	7				
UNIV44	54	9				
UNIV45	73	11				

```
                    DEPARTMENT OF HEALTH & HUMAN SERVICES
                       OIG - OFFICE OF AUDIT SERVICES
Date: 10/25/2004      GENERATION OF SECONDARY UNIT SAMPLE      Time: 13:49
                   NAME OF INPUT FILE: C:\TEMP\REGION10.TXT


                        GROUPS OF SECONDARY UNITS


                    ********* GROUP 1 *********
                                     SECONDARY UNIT    3RD STAGE
      SECONDARY UNIT IDENTIFICATION        SIZE        UNIVERSE
      ============================== =============  =============
      UNIV44                                    9             54
      UNIV32                                   10             69
      UNIV77                                    7             41
      UNIV78 <-- Selected                       7             39
      UNIV2                                     5             32
      UNIV50                                    8             52
      UNIV34                                    4             26
      UNIV46                                   12             78

         GROUP TOTALS:  8                      62            391


                    ********* GROUP 2 *********
                                     SECONDARY UNIT    3RD STAGE
      SECONDARY UNIT IDENTIFICATION        SIZE        UNIVERSE
      ============================== =============  =============
      UNIV6                                     11            72
      UNIV43 <-- Selected                        7            42
      UNIV62                                    10            61
      UNIV41                                     8            47
      UNIV1                                      6            34
      UNIV40                                     7            43
      UNIV79                                     9            58
      UNIV36                                    10            65

         GROUP TOTALS:  8                       68           422
```

< GROUPS 3 THROUGH 9 ARE OMITTED HERE >

```
                    ********* GROUP 10 *********
                                     SECONDARY UNIT    3RD STAGE
      SECONDARY UNIT IDENTIFICATION        SIZE        UNIVERSE
      ============================== =============  =============
      UNIV71                                   11             76
      UNIV9                                     6             38
      UNIV21                                   11             77
      UNIV23                                   10             63
      UNIV39 <-- Selected                       6             38
      UNIV29                                    5             34
      UNIV72                                    5             34
      UNIV13                                   10             70
      UNIV51                                    4             24

         GROUP TOTALS:  9                      68            454
```

```
                     DEPARTMENT OF HEALTH & HUMAN SERVICES
                        OIG - OFFICE OF AUDIT SERVICES
Date: 10/25/2004      GENERATION OF SECONDARY UNIT SAMPLE        Time: 13:49
               NAME OF OUTPUT FILE: C:\TEMP\OutRegion10.txt

 FIRST SEED NUMBER:      100.00     SECOND SEED NUMBER:       200.00

     NUMBER OF SECONDARY UNITS IN THE POPULATION:    85
     NUMBER OF SECONDARY UNITS SAMPLED:              10

                                                                    UNITS
                        3RD STAGE      SECONDARY                      IN
SECONDARY UNIT ID        UNIVERSE      UNIT SIZE      GROUP SIZE     GROUP
=========================  =============  =============  =============  =====
UNIV78                          39              7             62         8
UNIV43                          42              7             68         8
UNIV7                           56              9             54         8
UNIV73                          27              5             63         8
UNIV55                          78             12             70         8
UNIV33                          65             10             77         9
UNIV10                          60              9             76         9
UNIV59                          52              8             71         9
UNIV64                          50              8             73         9

UNIV39                          38              6             68         9
```

Constructing the data file

The data file for this example (PUSURHC3.TXT) is shown on the next page. This file was

constructed using the **RHC Sample Selection** program to select the primary units (regions) and

within each selected primary unit, the 10 secondary units (universities). The four lines beginning

with REGIONx are from the output file created during the primary unit selection

(GRANTSPUOUT.TXT). The 10 lines after each REGIONx line consist of the output file

created when selecting the universities from each region (OUTREGION4.TXT, . . .,

OUTREGION10.TXT). Using a word processor or spreadsheet, a column containing the sample

sizes (highlighted) must be added to the files created by the five RHC Sample Selection

programs.

--- Data set PUSURHC3.TXT ---

```
REGION4    123    10     1320    3410     3
UNIV85      54    11     11      125     12
UNIV46      44     9      9      131     12
UNIV7       77    15     17      119     12
UNIV82      52    10     11      129     12
UNIV30      54    11     11      141     12
UNIV34      50    10     10      140     12
UNIV27      76    15     16      138     12
UNIV66      76    15     16      128     13
UNIV65      62    12     14      125     13
UNIV80      70    14     15      155     13
REGION6    116    10     1240    3100     3
UNIV113     33     7      8      108     11
UNIV43      39     8      7      105     11
UNIV78      63    13     12      104     11
UNIV104     25     5      9       96     11
UNIV89      35     7      7      124     12
UNIV112     27     5     10      108     12
UNIV30      58    12     11       95     12
UNIV65      57    11     10      109     12
UNIV3       56    11     11      115     12
UNIV99      80    16     14      113     12
REGION8    118    10     1300    3170     3
UNIV112     75    15     16      125     11
UNIV6       34     7      8      127     11
UNIV7       51    10     11      120     12
UNIV93      54    11     11      136     12
UNIV75      52    10     11      126     12
UNIV111     84    17     17      134     12
UNIV62      64    13     14      123     12
UNIV115     59    12     13      137     12
UNIV70      65    13     14      143     12
UNIV99      60    12     13      129     12
REGION10    85    10     640     2320     3
UNIV78      39     8      7       62      8
UNIV43      42     8      7       68      8
UNIV7       56    11      9       54      8
UNIV73      27     5      5       63      8
UNIV55      78    16     12       70      8
UNIV33      65    13     10       77      9
UNIV10      60    12      9       76      9
UNIV59      52    10      8       71      9
UNIV64      50    10      8       73      9
UNIV39      38     8      6       68      9
```

Selection of Third-Stage Units

Since approximately 20% of the grants at each selected university are to be audited, the

following sample sizes are determined:

```
Region 4:  University   Grants in universe      Number to be audited
           UNIV85            54                          11
           UNIV46            44                           9
           UNIV7             77                          15
           UNIV82            52                          10
           UNIV30            54                          11
           UNIV34            50                          10
           UNIV27            76                          15
           UNIV66            76                          15
           UNIV65            62                          12
           UNIV80            70                          14
                                                        122

Region 6:  University   Grants in universe      Number to be audited
           UNIV113           33                           7
           UNIV43            39                           8
           UNIV78            63                          13
           UNIV104           25                           5
           UNIV89            35                           7
           UNIV112           27                           5
           UNIV30            58                          12
           UNIV65            57                          11
           UNIV3             56                          11
           UNIV99            80                          16
                                                         95

Region 8:  University   Grants in universe      Number to be audited
           UNIV112           75                          15
           UNIV6             34                           7
           UNIV7             51                          10
           UNIV93            54                          11
           UNIV75            52                          10
           UNIV111           84                          17
           UNIV62            64                          13
           UNIV115           59                          12
           UNIV70            65                          13
           UNIV99            60                          12
                                                        120

Region 10: University   Grants in universe      Number to be audited
           UNIV78            39                           8
           UNIV43            42                           8
           UNIV7             56                          11
           UNIV73            27                           5
           UNIV55            78                          16
           UNIV33            65                          13
           UNIV10            60                          12
           UNIV59            52                          10
           UNIV64            50                          10
           UNIV39            38                           8
                                                        101
```

The data file containing the errors for these 438 audited grants is DATARHC3.TXT, shown on

the next page. The values for the first two universities in Region 4 and the last two universities

in Region 10 are illustrated. Each sample value is equal to the total charges after the scheduled

completion of the grant (in thousands of dollars). Notice that each line begins with a counter.

Finally, the **RHC Three-Stage** program is run, which generates a confidence interval for the

universe total using input files PUSURHC3.TXT and DATARHC3.TXT. The output from this

program is shown at the end of this section.

--- Data set DATARHC3.TXT ---

```
1     8
2     0
3     6
4     6
5     0
6    13
7     1
8     7
9     2
10   13
11   13
12    4         These are the sample values for UNIV85 and UNIV46 in Region 4.
13    6
14    0
15   15
16   12
17    9
18    0
19   13
20   10
.
.
.
421   0
422   6
423  19
424  17
425  13
426  12
427  13         These are the sample values for UNIV64 and UNIV39 in Region 10.
428  12
429  11
430  14
431  13
432   0
433   1
434   5
435  16
436   0
437   0
438   8
```

Summary of results. Referring to the last page in the computer output, the estimate of the universe total (all 12 regions) is the OVERALL POINT ESTIMATE of 463,526 ($463,526,000) with a corresponding estimated OVERALL STANDARD ERROR of 53,521 ($53,521,000).

NOTE: This estimate does not require knowing the number of grants in the universe.

The program also provides estimates for the total error amount for each sampled P.U. (region) and for each of the groups of S.U.s (universities) within each sampled region. For example, the estimated error amount for Region 4 is 50,529 ($50,529,000) and the estimated error amount for the <u>group</u> of 12 universities containing UNIV85 is 3,849 ($3,849,000). The SIZES RATIO refers to the ratio of the size of the group containing this university to the size of this university. To illustrate, UNIV85 in Region 4 has a size of 11 and is in a group of size 125 (look at file REGION4.TXT). The SIZES RATIO here is 125/11 = 11.3636.

The 80% confidence interval for the total error amount is 394,936 to 532,116 ($394,936,000 to $532,116,000). The PRECISION AMOUNT is the amount added and subtracted to the point estimate (463,526) to obtain the corresponding confidence interval. In the 80% confidence interval, the lower limit of 394,936 is obtained by subtracting the precision amount of 68,590 from 463,526. The PRECISION PERCENT is the precision amount divided by the point estimate, expressed as a percentage.

```
                      DEPARTMENT OF HEALTH & HUMAN SERVICES
                         OIG - OFFICE OF AUDIT SERVICES
Date: 10/26/2004        RHC THREE-STAGE VARIABLE APPRAISAL          Time: 16:27
                          AUDIT/REVIEW: RHC 3-Stage

                    DATA FILE USED: C:\TEMP\DATARHC3.TXT
            PRIMARY/SECONDARY UNIVERSE FILE USED: C:\TEMP\PUSURHC3.txt


                                          *****D I F F E R E N C E*****
     **** SAMPLED UNITS ****        THIRD STAGE   SAMPLE     SAMPLE      NONZERO
PRIMARY / SECONDARY IDENTIFICATION    UNIVERSE      SIZE      VALUE       COUNT
================================== =========== ====== ============ ========
REGION4
     UNIV85                                 54     11        69.00            9
     UNIV46                                 44      9        69.00            7
     UNIV7                                  77     15       145.00           14
     UNIV82                                 52     10        91.00           10
     UNIV30                                 54     11        83.00            9
     UNIV34                                 50     10        69.00            8
     UNIV27                                 76     15        72.00           12
     UNIV66                                 76     15       134.00           13
     UNIV65                                 62     12        97.00           10
     UNIV80                                 70     14       153.00           11
  Total                                    615    122                       103
REGION6
     UNIV113                                33      7        81.00            7
     UNIV43                                 39      8        36.00            8
     UNIV78                                 63     13        83.00            9
     UNIV104                                25      5        56.00            5
     UNIV89                                 35      7        82.00            6
     UNIV112                                27      5        29.00            2
     UNIV30                                 58     12        65.00            9
     UNIV65                                 57     11        82.00            8
     UNIV3                                  56     11        67.00            6
     UNIV99                                 80     16       164.00           15
  Total                                    473     95                        75
REGION8
     UNIV112                                75     15        78.00           12
     UNIV6                                  34      7        60.00            7
     UNIV7                                  51     10        48.00            9
     UNIV93                                 54     11        85.00            9
     UNIV75                                 52     10        63.00            7
     UNIV111                                84     17       154.00           16
     UNIV62                                 64     13        83.00           13
     UNIV115                                59     12        61.00            8
     UNIV70                                 65     13        87.00            9
     UNIV99                                 60     12        90.00            9
  Total                                    598    120                        99
REGION10
     UNIV78                                 39      8        75.00            7
     UNIV43                                 42      8        71.00            6
     UNIV7                                  56     11       123.00           11
     UNIV73                                 27      5        32.00            3
     UNIV55                                 78     16       123.00           11
     UNIV33                                 65     13       113.00           11
     UNIV10                                 60     12       104.00            8
     UNIV59                                 52     10        77.00            7
     UNIV64                                 50     10       117.00            9
     UNIV39                                 38      8        43.00            5
  Total                                    507    101                        78
```

```
                                    *****D I F F E R E N C E*****
        **** SAMPLED UNITS ****    THIRD STAGE  SAMPLE    SAMPLE     NONZERO
PRIMARY / SECONDARY IDENTIFICATION   UNIVERSE    SIZE     VALUE       COUNT
================================== =========== ====== ============ ========

            TOTALS                     2,193      438    3,414.00        355

                        --- POINT ESTIMATES---
                    *****D I F F E R E N C E*****
        **** SAMPLED UNITS ****                    SIZES        POINT
PRIMARY / SECONDARY IDENTIFICATION  SAMPLE MEAN    RATIO       ESTIMATE
================================== =========== =========== =============
REGION4
   UNIV85   Note: 3,849 is the estimate  6.27     11.3636        3,849
   UNIV46   for the group containing     7.67     14.5556        4,910
   UNIV7    UNIV85 (not just UNIV85).    9.67      7.0000        5,210
   UNIV82   This group contained 12      9.10     11.7273        5,549
   UNIV30   universities show earlier    7.55     12.8182        5,223
   UNIV34   in the output using data     6.90     14.0000        4,830
   UNIV27   set REGION4.TXT.             4.80      8.6250        3,146
   UNIV66                                8.93      8.0000        5,431
   UNIV65                                8.08      8.9286        4,475
   UNIV80                               10.93     10.3333        7,905

           TOTAL                 Estimate for Region 4 ➔  50,529

REGION6
   UNIV113                               11.57     13.5000        5,155
   UNIV43                                 4.50     15.0000        2,633
   UNIV78                                 6.38      8.6667        3,486
   UNIV104                               11.20     10.6667        2,987
   UNIV89                                11.71     17.7143        7,263
   UNIV112                                5.80     10.8000        1,691
   UNIV30                                 5.42      8.6364        2,713
   UNIV65                                 7.45     10.9000        4,632
   UNIV3                                  6.09     10.4545        3,566
   UNIV99                                10.25      8.0714        6,619

           TOTAL                                                40,744

REGION8
   UNIV112                                5.20      7.8125        3,047
   UNIV6                                  8.57     15.8750        4,626
   UNIV7                                  4.80     10.9091        2,671
   UNIV93                                 7.73     12.3636        5,159
   UNIV75                                 6.30     11.4545        3,753
   UNIV111                                9.06      7.8824        5,998
   UNIV62                                 6.38      8.7857        3,590
   UNIV115                                5.08     10.5385        3,161
   UNIV70                                 6.69     10.2143        4,443
   UNIV99                                 7.50      9.9231        4,465

           TOTAL                                                40,913

                        --- POINT ESTIMATES---
                    *****D I F F E R E N C E*****
        **** SAMPLED UNITS ****                    SIZES        POINT
PRIMARY / SECONDARY IDENTIFICATION  SAMPLE MEAN    RATIO       ESTIMATE
================================== =========== =========== =============
REGION10
   UNIV78                                 9.38      8.8571        3,238
   UNIV43                                 8.88      9.7143        3,621
   UNIV7                                 11.18      6.0000        3,757
   UNIV73                                 6.40     12.6000        2,177
   UNIV55                                 7.69      5.8333        3,498
```

```
        UNIV33                                  8.69      7.7000        4,351
        UNIV10                                  8.67      8.4444        4,391
        UNIV59                                  7.70      8.8750        3,554
        UNIV64                                 11.70      9.1250        5,338
        UNIV39                                  5.38     11.3333        2,315

                TOTAL                                                  36,240
```

--- VARIANCE COMPONENTS FOR PRIMARY UNITS ---

**** SAMPLED UNITS **** PRIMARY UNIT IDENTIFICATION	WITHIN VARIANCE	BETWEEN VARIANCE	TOTAL VARIANCE
REGION4	757,373	10,383,517	11,140,890
REGION6	798,923	29,209,394	30,008,317
REGION8	710,631	8,065,650	8,776,280
REGION10	781,064	6,460,707	7,241,771

(Values of V_4) (Values of V_3)

--- OVERALL VARIANCE COMPONENTS ---

STAGE 1	STAGES 2 AND 3	TOTAL VARIANCE
2,713,022,822	151,453,466	2,864,476,288

(Value of V_1)   **(Value of V_2)**


```
                    *****D I F F E R E N C E*****
                 --- SUMMARY OF APPRAISAL RESULTS ---
PRIMARY UNITS SAMPLED                               4
PRIMARY UNITS NOT SAMPLED                           8
TOTAL PRIMARY UNITS                                12
OVERALL POINT ESTIMATE                        463,526
OVERALL STANDARD ERROR                         53,521

                          CONFIDENCE LIMITS
                        80% CONFIDENCE LEVEL
              LOWER LIMIT               394,936
              UPPER LIMIT               532,116
              PRECISION AMOUNT           68,590
              PRECISION PERCENT          14.80%
              Z-VALUE USED         1.281551565545

                        90% CONFIDENCE LEVEL
              LOWER LIMIT               375,492
              UPPER LIMIT               551,560
              PRECISION AMOUNT           88,034
              PRECISION PERCENT          18.99%
              Z-VALUE USED         1.644853626951

                        95% CONFIDENCE LEVEL
              LOWER LIMIT               358,627
              UPPER LIMIT               568,425
              PRECISION AMOUNT          104,899
              PRECISION PERCENT          22.63%
              Z-VALUE USED         1.959963984540
```

Discussion. In general, you can expect greater precision with the RHC procedure, <u>provided</u>

there is a significant correlation between NUMBER OF UNITS and SIZE OF UNIT. To

illustrate, consider the file containing the primary unit information used in the three-stage RHC

illustration:

```
              (A)     (B)
   REGION1    117    1250
   REGION2     63     610
   REGION3     91     720
   REGION4    123    1320
   REGION5    107    1160
   REGION6    116    1240
   REGION7    102     960
   REGION8    118    1300
   REGION9    122    1320
   REGION10    85     640
   REGION11    94     930
   REGION12    62     550
```

Columns: (A) number of universities (S.U.s)
                (B) size of each P.U. (grant amount x $100,000)

For this example, the correlation between columns (A) and (B) is .958, and we would expect a

single- and two-stage RHC procedure to work quite well. For a three-stage procedure, this

correlation rule must also apply within the sampled primary units, at the secondary unit level.

As mentioned earlier, the benefits of RHC sampling include the following:

- Precision is increased if the above correlation rule is satisfied.

- The flavor of pps sampling is maintained, since pps sampling is used to select a unit from each random group.

- Computations are relatively simple and straightforward.

- The point estimate $(\hat{T})$ is stable. This implies that when sampling indefinitely, the point estimate will exhibit relatively small variation.

- The point estimate of the variance of $\hat{T}$ is stable, producing more reliable confidence intervals.

This implies that when sampling indefinitely, the lower confidence limits will exhibit relatively small variation.

FORMULAS

Definitions

1. S_i = (size of i-th P.U.)/(size of entire population)

2. $\pi_i = \Sigma S_i$ over the i-th group of P.U.s

3. S_{ij} = (size of j-th S.U. in the i-th sampled P.U.)/(size of i-th sampled P.U.)

   (**Note**: denominator of S_{ij} = numerator of S_i)

4. $\pi_{ij} = \Sigma S_{ij}$ over the j-th group in i-th sampled P.U.

5. N = number of P.U.s (population)

6. n = number of P.U.s (sample)

7. M_i = number of S.U.s in i-th sampled P.U. (population)

8. m_i = number of S.U.s in i-th sampled P.U. (sample)

9. K_{ij} = number of third-stage units for j-th sampled S.U. in i-th sampled P.U. (population)

10. k_{ij} = number of third-stage units for j-th sampled S.U. in i-th sampled P.U. (sample)

Estimator of population total (T)

$$\hat{T} = \sum_{i=1}^{n} \pi_i \left(\frac{\hat{T_i}}{S_i} \right)$$

where $\hat{T_i}$ = estimator of total for i-th sampled P.U.

$$= \sum_{j=1}^{m_i} \pi_{ij} \left(\frac{\hat{T_{ij}}}{S_{ij}} \right) \qquad \text{(equation 1)}$$

and $\hat{T_{ij}}$ = estimator of population total for j-th sampled S.U. in i-th sampled P.U.

$$= K_{ij} \bar{y}_{ij} \text{ where } \bar{y}_{ij} = \text{average of } k_{ij} \text{ units at the third stage}$$

NOTE: It can be shown that $\hat{T}$ is an unbiased estimator of T.

<u>Estimated variance of</u> $\hat{T}$

$$v(\hat{T}) = V_1 + V_2 \quad \text{where}$$

$$V_1 = \frac{\sum\limits_{i=1}^{n} N_i^2 - N}{N^2 - \sum\limits_{i=1}^{n} N_i^2} \sum\limits_{i=1}^{n} \pi_i \left(\frac{\hat{T}_i}{S_i} - \hat{T} \right)^2 \qquad \text{(equation 2)}$$

and

$$V_2 = \sum\limits_{i=1}^{n} \left(\frac{\pi_i}{S_i} \right) v(\hat{T}_i) \qquad \text{(equation 3)}$$

and N_i = number of P.U.s in the i-th group after the random split into n groups.

$v(\hat{T}_i)$ is obtained by applying the <u>two</u>-stage RHC procedure within the i-th sampled P.U.;

i.e., the i-th P.U. is viewed as the entire population. Consequently,

$$v(\hat{T}_i) = V_{3,i} + V_{4,i}$$

where

$$V_{3,i} = \frac{\sum\limits_{j=1}^{m_i} M_{ij}^2 - M_i}{M_i^2 - \sum\limits_{j=1}^{m_i} M_{ij}^2} \sum\limits_{j=1}^{m_i} \pi_{ij} \left(\frac{K_{ij} \bar{y}_{ij}}{S_{ij}} - \hat{T}_i \right)^2$$

and

$$V_{4,i} = \sum_{j=1}^{m_i} \pi_{ij} \frac{K_{ij}}{S_{ij}} \left(\frac{K_{ij} - k_{ij}}{k_{ij}} \right) s_{ij}^2$$

and where

(1) M_{ij} = the number of S.U.s in the j-th random group within the i-th sampled P.U.

(2) $\bar{y}_{ij}$ = average of the k_{ij} items for the j-th sampled S.U. within the i-th sampled P.U.

<u>**Comments**</u>   1. V_1 is essentially the same expression obtained for the single-stage RHC

procedure and will be referred to as the "between unit" variation.

2. V_2 is the contribution of the 2nd- and 3rd-stage variation and is obtained by

treating each sampled P.U. as the population to be sampled using two

(additional) stages.

3.   The estimated standard error of $\hat{T}$ is $\sqrt{v(\hat{T})}$.

<u>Approximate 95% confidence interval for the population total (T)</u>

$$\hat{T} \pm 1.959963984540\sqrt{v(\hat{T})}$$

NOTE: For a 90% confidence interval, replace 1.959963984540 with 1.644853626951 and
for an 80% confidence interval replace 1.959963984540 with 1.281551565545.

Stratified Cluster Variable Appraisal

With this procedure, you first stratify, then obtain a cluster (single-stage) sample within each stratum. This is motivated by the discussion in the RAT-STATS User's Guide.

The estimate of a universe total is the sum of the estimates for each stratum. The estimated variance of this estimator is the sum of the estimated variances for each stratum.

Example 7. In a large section of the U.S., an audit was conducted for 583 universities with health related research grants. It was decided to define two strata:

　　Stratum 1: state-supported universities ($N_1 = 415$)

　　Stratum 2: private universities ($N_2 = 168$)

Within each stratum, a **single-stage cluster** sample was obtained with $n_1 = 25$ universities selected from Stratum 1 and $n_2 = 10$ universities from Stratum 2. For each of the sampled universities, all health-related grants would be audited (since there weren't that many at each university) to determine the amount of charges improperly charged to these grants. The following data were obtained, where y_j is the total of the improper charges (in thousands of dollars) for the j-th university (cluster) and M_j is the number of grants (universe) for this university, all of which are audited.

NOTE:　The number of grants audited at each university (the M_i values) are <u>not</u> used in the program calculations. They are supplied for informational purposes only. For example, if all the M_i values are set equal to 1, the resulting confidence intervals will be unchanged.

Stratum 1

Univ.	M_j	y_j	Univ.	M_j	y_j		
1	8	96	14	10	49		
2	12	121	15	9	53		
3	4	42	16	3	50		
4	5	65	17	6	32		
5	6	52	18	5	22		
6	6	40	19	5	45	$\Sigma M_j = 151$	
7	7	75	20	4	37		
8	5	65	21	6	51	$\Sigma y_j = 1,329$	
9	8	45	22	8	30		
10	3	50	23	7	39		
11	2	85	24	3	47		
12	6	43	25	8	41		
13	5	54					

Stratum 2

Univ.	M_j	y_j	Univ.	M_j	y_j		
1	2	18	6	8	96		
2	5	52	7	6	64	$\Sigma M_j = 49$	
3	7	68	8	10	115		
4	4	36	9	3	41	$\Sigma y_j = 547$	
5	3	45	10	1	12		

These values were stored in data file DATASTRCLUS.TXT. Immediately following the listing

of this data file is the resulting computer output using the VARIABLE STRATIFIED CLUSTER

program.

Data file DATASTRCLUS.TXT
```
STATE UNIVERSITIES  415   25
UNIV1      8      96
UNIV2     12     121
UNIV3      4      42
UNIV4      5      65
UNIV5      6      52
UNIV6      6      40
UNIV7      7      75
UNIV8      5      65
UNIV9      8      45
UNIV10     3      50
UNIV11     2      85
UNIV12     6      43
UNIV13     5      54
UNIV14    10      49
```

```
UNIV15    9    53
UNIV16    3    50
UNIV17    6    32
UNIV18    5    22
UNIV19    5    45
UNIV20    4    37
UNIV21    6    51
UNIV22    8    30
UNIV23    7    39
UNIV24    3    47
UNIV25    8    41
PRIVATE UNIVERSITIES  168  10
UNIV1     2    18
UNIV2     5    52
UNIV3     7    68
UNIV4     4    36
UNIV5     3    45
UNIV6     8    96
UNIV7     6    64
UNIV8    10   115
UNIV9     3    41
UNIV10    1    12
```

```
                    DEPARTMENT OF HEALTH & HUMAN SERVICES
                       OIG - OFFICE OF AUDIT SERVICES
Date: 10/23/2004      STRATIFIED CLUSTER VARIABLE APPRAISAL      Time: 14:36
                   AUDIT/REVIEW: Variable - Stratified Cluster

                   DATA FILE USED: C:\Temp\DATASTRCLUS.TXT
```

STRATUM IDENTIFICATION CLUSTER IDENTIFICATION	SAMPLE UNIVERSE	SAMPLE SIZE	SAMPLED VALUE	POINT ESTIMATE
STATE UNIVERSITIES	415	25		
UNIV1	8	8	96.00	
UNIV2	12	12	121.00	
UNIV3	4	4	42.00	
UNIV4	5	5	65.00	
UNIV5	6	6	52.00	
UNIV6	6	6	40.00	
UNIV7	7	7	75.00	
UNIV8	5	5	65.00	
UNIV9	8	8	45.00	
UNIV10	3	3	50.00	
UNIV11	2	2	85.00	
UNIV12	6	6	43.00	
UNIV13	5	5	54.00	
UNIV14	10	10	49.00	
UNIV15	9	9	53.00	
UNIV16	3	3	50.00	
UNIV17	6	6	32.00	
UNIV18	5	5	22.00	
UNIV19	5	5	45.00	
UNIV20	4	4	37.00	
UNIV21	6	6	51.00	
UNIV22	8	8	30.00	

```
   UNIV23                              7        7         39.00
   UNIV24                              3        3         47.00
   UNIV25                              8        8         41.00
      STRATUM TOTALS                 151      151      1,329.00              22,061

PRIVATE UNIVERSITIES                 168       10
   UNIV1                               2        2         18.00
   UNIV2                               5        5         52.00
   UNIV3                               7        7         68.00
   UNIV4                               4        4         36.00
   UNIV5                               3        3         45.00
   UNIV6                               8        8         96.00
   UNIV7                               6        6         64.00
   UNIV8                              10       10        115.00
   UNIV9                               3        3         41.00
   UNIV10                              1        1         12.00
      STRATUM TOTALS                  49       49        547.00               9,190
```

STRATUM IDENTIFICATION CLUSTER IDENTIFICATION	SAMPLE UNIVERSE	SAMPLE SIZE	SAMPLED VALUE	POINT ESTIMATE
STRATA TOTALS	583	35		
CLUSTER UNIT TOTALS	200	200	1,876.00	
OVERALL POINT ESTIMATE				31,251
OVERALL STANDARD ERROR				2,418

```
CONFIDENCE LEVEL      ---80 PERCENT--      ---90 PERCENT--      ---95 PERCENT--
LOWER LIMIT                    28,152               27,273               26,511
UPPER LIMIT                    34,350               35,229               35,991
PRECISION AMOUNT                3,099                3,978                4,740
PRECISION PERCENT               9.92%               12.73%               15.17%
Z-VALUE USED           1.281551565545       1.644853626951       1.959963984540
```

Discussion. For stratum 1, the unbiased estimate of the universe total is

$$\hat{T}_1 = (415/25)(1,329) = 22,061 \ (\$22,061,000)$$

The unbiased estimate of the universe total for stratum 2 is

$$\hat{T}_2 = (168/10)(547) = 9,190 \ (\$9,190,000)$$

Consequently, an unbiased estimate of the universe total (highlighted) is

$$\hat{T} = \hat{T}_1 + \hat{T}_2 = 31,251 \ (\$31,251,000)$$

Using formula 2, the estimated variance of $\hat{T}$ is $v(\hat{T}) = v(\hat{T}_1) + v(\hat{T}_2) = 5,848,565$ and the

corresponding standard error (highlighted) is 2,418.

The approximate 95% confidence interval for the universe total is

$$31,251 \pm 1.959963984540(2418)$$

that is, 26,511 to 35,991 ($26,511,000 to $35,991,000).

FORMULAS

1. Estimated total in the universe (T)

$$\hat{T} = \sum_{h=1}^{L} \frac{N_h}{n_h} \left(\sum_{j=1}^{n_h} y_{j,h} \right) = \sum_{h=1}^{L} N_h \bar{y}_h$$

where L = number of strata

N_h = number of clusters (universe) for stratum h

n_h = number of clusters (sample) for stratum h

$y_{j,h}$ = total of the variable of interest (e.g., errors) for the j-th P.U. within stratum h

$\bar{y}_h$ = sample average for stratum h

NOTE: Let $\hat{T}_h$ = estimated total for stratum h. Then $\hat{T}_h = N_h \bar{y}_h$ and $\hat{T} = \sum \hat{T}_h$

2. Estimated variance of $\hat{T}$

$$v(\hat{T}) = \sum_{h=1}^{L} \frac{N_h(N_h - n_h)}{n_h(n_h - 1)} \sum_{j=1}^{n_h} (y_{j,h} - \bar{y}_h)^2 = \sum_{h=1}^{L} v(\hat{T}_h)$$

3. Approximate 95% confidence interval for T

$$\hat{T} \pm 1.959963984540 \sqrt{v(\hat{T})}$$

NOTE: For a 90% confidence interval, replace 1.959963984540 with 1.644853626951 and for an 80% confidence interval replace 1.959963984540 with 1.281551565545.

Stratified Multistage Variable Appraisal

As with the stratified cluster procedure, you must first stratify the universe. Rather than take a

cluster (single-stage) sample within each stratum, you will obtain a multistage (two-stage or

three-stage) sample within each stratum. These multistage samples may be random (using the

Two-Stage Unrestricted or **Three-Stage Unrestricted** programs) or may be obtained using the

RHC procedure and the **RHC Two-Stage** or **RHC Three-Stage** programs.

Unlike the **Stratified Cluster** program, this program requires that you first run the appropriate

multistage program <u>on</u> <u>each</u> <u>stratum</u> and record the results. The output results are then used as

input to the **Stratified Multistage** program. You may store the results from each stratum (point

estimate, standard error) in a file or simply input these values interactively.

NOTE: The "universe size" refers to the number of units at the most detailed level of the
multistage sample. For example, if you are obtaining a three-stage sample within
each stratum, then the "universe size" refers to the total number of third-stage units
within this stratum.

Example 8. This example is similar to Example 7 in the Stratified Cluster section. In a

particular region, the universe consisting of university grants is stratified by defining

> Stratum 1: state-supported universities ($N_1 = 120$ univ.)

> and Stratum 2: private universities ($N_2 = 85$ univ.)

Because these universities are so widespread, it was decided to employ a two-stage sample using

15 state supported universities and 10 private universities. Rather than audit all grants at a

selected university, it was decided (based on available resources) to audit roughly 20% of the

grants at each selected university to estimate the amount of charges improperly charged to these

grants. We know that there are a total of 5,800 grants within the universe of the 120 state

supported universities and 4,500 grants within the 85 private universities.

The following data were obtained where $y_{i,j}$ is the dollars (in thousands) of improper charges for

the j-th grant within the i-th sampled university, M_i is the total number of grants at the i-th

university, and m_i is the number of audited grants at the i-th university. Also, $\bar{y}_i$ and s_i^2 are the

mean and variance of the sample values from the i-th university.

NOTE: The 15 state-supported universities and 10 private universities were obtained using
              the **Single-Stage Random Numbers** program. For ease of illustration, they will be
              referred to as University 1, 2, 3, . . . within each stratum.

The corresponding data files are the input files for the **Two-Stage Unrestricted** program. These

are files STRMULT1.TXT and STRMULT2.TXT. The files containing the universe sizes are

UNIV1.TXT and UNIV2.TXT.

State-supported universities

Univ.	M_i	m_i	Dollars ($y_{i,j}$, in thousands)	$\bar{y}_i$	s_i^2
1	60	12	4 0 0 6 7 11 0 5 4 0 8 2	3.92	13.36
2	50	10	4 7 0 0 6 10 3 3 2 0	3.50	11.17
3	45	9	3 1 0 5 5 8 10 0 4	4.00	12.00
4	40	8	2 7 10 0 6 5 4 0	4.25	12.21
5	55	11	7 1 6 0 0 0 5 12 8 2 4	4.09	15.49
6	58	12	0 0 6 5 12 8 2 0 7 7 2 5	4.50	14.27
7	62	12	4 1 0 3 10 7 6 6 0 5 4 4	4.17	8.70
8	52	10	3 8 0 6 2 10 0 0 5 2	3.60	12.49
9	50	10	7 1 4 0 2 6 0 0 0 5	2.50	7.61
10	45	9	2 1 0 0 0 5 8 6 10	3.56	14.53
11	40	8	4 7 3 8 0 1 6 10	4.88	12.13
12	48	10	0 0 6 5 2 3 8 0 12 0	3.60	16.93
13	57	11	3 0 1 10 5 0 6 6 7 3 8	4.45	11.07
14	60	12	4 1 0 0 0 6 8 10 4 3 2 5	3.58	10.63
15	54	11	3 0 1 1 4 8 6 9 5 0 3	3.64	9.65

Private universities

Univ.	M_i	m_i	Dollars ($y_{i,j}$, in thousands)	$\bar{y}_i$	s_i^2
1	66	13	4 4 0 0 0 3 6 7 5 2 0 0 4	2.69	6.40
2	52	10	10 1 6 4 0 0 5 8 12 7	5.30	17.12
3	47	9	2 8 1 0 6 6 11 5 0	4.33	14.75
4	55	11	3 8 0 0 6 5 2 1 12 8 5	4.55	14.47
5	48	10	0 5 8 1 1 6 10 7 9 3	5.00	12.89
6	60	12	7 3 0 5 6 6 0 0 8 2 3 5	3.75	8.02
7	57	11	2 4 0 6 10 7 5 0 0 3 8	4.09	11.89
8	50	10	3 5 1 1 0 0 3 8 10 7	3.80	12.62
9	62	12	8 4 0 2 2 6 5 1 12 5 0 0	3.75	13.66
10	56	11	5 0 1 2 8 7 10 6 0 4 2	4.09	11.49

The data files are shown on the next page. The corresponding computer outputs immediately

follow.

Data file STRMULT1.TXT   **Data file STRMULT2.TXT**

1	4		1	4
2	0		2	4
3	0		3	0
4	6		4	0
5	7		5	0
6	11		6	3
7	0		7	6
8	5		8	7
9	4		9	5
10	0		10	2
11	8		11	0
12	2		12	0
13	4		13	4
14	7		14	10
15	0		15	1
16	0		16	6
17	6		17	4
18	10		18	0
19	3		19	0
20	3		20	5
21	2		21	8
22	0		22	12
			23	7
.			.	
.			.	
.			.	
133	4		.	
134	1		87	8
135	0		88	4
136	0		89	0
137	0		90	2
138	6		91	2
139	8		92	6
140	10		93	5
141	4		94	1
142	3		95	12
143	2		96	5
144	5		97	0
145	3		98	0
146	0		99	5
147	1		100	0
148	1		101	1
149	4		102	2
150	8		103	8
151	6		104	7
152	9		105	10
153	5		106	6
154	0		107	0
155	3		108	4
			109	2

Data file UNIV1.TXT

```
 1   60   12
 2   50   10
 3   45    9
 4   40    8
 5   55   11
 6   58   12
 7   62   12
 8   52   10
 9   50   10
10   45    9
11   40    8
12   48   10
13   57   11
14   60   12
15   54   11
```

Data file UNIV2.TXT

```
 1   66   13
 2   52   10
 3   47    9
 4   55   11
 5   48   10
 6   60   12
 7   57   11
 8   50   10
 9   62   12
10   56   11
```

Output using the state-supported universities

```
                   DEPARTMENT OF HEALTH & HUMAN SERVICES
                       OIG - OFFICE OF AUDIT SERVICES
Date: 10/22/2004    TWO-STAGE UNRESTRICTED VARIABLE APPRAISAL       Time: 14:14
                          AUDIT/REVIEW: Stratum 1

             DATA FILE USED: C:\TEMP\STRMULT1.TXT

        --------------------- D I F F E R E N C E ----------------------
UNIT    SAMPLE SIZE/
 NBR    NONZERO ITEMS    SAMPLE MEAN      VARIANCE    UNIVERSE SIZE   POINT ESTIMATE

  1        12/8            3.92          13.36             60             235
  2        10/7            3.50          11.17             50             175
  3         9/7            4.00          12.00             45             180
  4         8/6            4.25          12.21             40             170
  5        11/8            4.09          15.49             55             225
  6        12/9            4.50          14.27             58             261
  7        12/10           4.17           8.70             62             258
  8        10/7            3.60          12.49             52             187
  9        10/6            2.50           7.61             50             125
 10         9/6            3.56          14.53             45             160
 11         8/7            4.88          12.12             40             195
 12        10/6            3.60          16.93             48             173
 13        11/9            4.45          11.07             57             254
 14        12/9            3.58          10.63             60             215
 15        11/9            3.64           9.65             54             196

          155/114          4.15                          776           3,010

NOT SAMPLED
   105                                          5,024
OVERALL TOTALS
   120                                          5,800          24,077

STANDARD ERROR                                                  1,277
```

```
                            CONFIDENCE LIMITS
                         80% CONFIDENCE LEVEL
            LOWER LIMIT                          22,440
            UPPER LIMIT                          25,714
            PRECISION AMOUNT                       1,637
            PRECISION PERCENT                      6.80%
            Z-VALUE USED                 1.281551565545

                         90% CONFIDENCE LEVEL
            LOWER LIMIT                          21,976
            UPPER LIMIT                          26,178
            PRECISION AMOUNT                       2,101
            PRECISION PERCENT                      8.73%
            Z-VALUE USED                 1.644853626951

                         95% CONFIDENCE LEVEL
            LOWER LIMIT                          21,573
            UPPER LIMIT                          26,580
            PRECISION AMOUNT                       2,503
            PRECISION PERCENT                     10.40%
            Z-VALUE USED                 1.959963984540
```

Output using the private universities

```
                    DEPARTMENT OF HEALTH & HUMAN SERVICES
                       OIG - OFFICE OF AUDIT SERVICES
        Date: 10/22/2004    TWO-STAGE UNRESTRICTED VARIABLE APPRAISAL      Time: 14:13
                          AUDIT/REVIEW: Stratum 2

                     DATA FILE USED: C:\TEMP\STRMULT2.TXT

                 -------------------- D I F F E R E N C E ---------------------
       UNIT     SAMPLE SIZE/
        NBR    NONZERO ITEMS    SAMPLE MEAN      VARIANCE    UNIVERSE SIZE   POINT ESTIMATE

         1        13/8            2.69            6.40            66            178
         2        10/8            5.30           17.12            52            276
         3         9/7            4.33           14.75            47            204
         4        11/9            4.55           14.47            55            250
         5        10/9            5.00           12.89            48            240
         6        12/9            3.75            8.02            60            225
         7        11/8            4.09           11.89            57            233
         8        10/8            3.80           12.62            50            190
         9        12/9            3.75           13.66            62            232
        10        11/9            4.09           11.49            56            229

                 109/84           4.26                           553          2,257

       NOT SAMPLED
           75                                                   3,947
       OVERALL TOTALS
           85                                                   4,500        19,182

       STANDARD ERROR                                                          873
```

```
                            CONFIDENCE LIMITS
                         80% CONFIDENCE LEVEL
        LOWER LIMIT                            18,064
        UPPER LIMIT                            20,300
        PRECISION AMOUNT                        1,118
        PRECISION PERCENT                        5.83%
        Z-VALUE USED                   1.281551565545

                         90% CONFIDENCE LEVEL
        LOWER LIMIT                            17,747
        UPPER LIMIT                            20,617
        PRECISION AMOUNT                        1,435
        PRECISION PERCENT                        7.48%
        Z-VALUE USED                   1.644853626951

                         95% CONFIDENCE LEVEL
        LOWER LIMIT                            17,472
        UPPER LIMIT                            20,892
        PRECISION AMOUNT                        1,710
        PRECISION PERCENT                        8.92%
        Z-VALUE USED                   1.959963984540
```

When running the STRATIFIED MULTISTAGE program, you should see the following input window containing values for the first stratum. The program output immediately follows.

```
                    DEPARTMENT OF HEALTH & HUMAN SERVICES
                      OIG - OFFICE OF AUDIT SERVICES
Date: 10/22/2004    STRATIFIED MULTISTAGE VARIABLE APPRAISAL       Time: 14:41
                      AUDIT/REVIEW: Combining the Strata

              THE ESTIMATORS ARE BASED ON THE FOLLOWING ENTRIES:
   STRATUM          POINT ESTIMATE      STANDARD ERROR
      1                 24,077              1,277
      2                 19,182                873


= = = = = = = = = = = = = = = = = RESULTS = = = = = = = = = = = = = = = = = = =

                 POINT ESTIMATE                STANDARD ERROR

                    43,259                         1,547



CONFIDENCE LEVEL      ---80 PERCENT--      ---90 PERCENT--      ---95 PERCENT--
LOWER LIMIT                   41,277               40,715               40,227
UPPER LIMIT                   45,241               45,803               46,291
PRECISION AMOUNT               1,982                2,544                3,032
PRECISION PERCENT              4.58%                5.88%                7.01%
Z-VALUE USED           1.281551565545       1.644853626951       1.959963984540
```

Discussion. The point estimate for the universe total is the sum of the point estimates for the

two strata; that is, $\hat{T} = 24{,}077 + 19{,}182 = 43{,}259$ ($43,259,000). The estimated variance of $\hat{T}$

is $v(\hat{T}) = (1277)^2 + (873)^2 = 2{,}392{,}858$ and the corresponding standard error is $\sqrt{2{,}392{,}858}$

$= 1{,}547$. The approximate 95% confidence interval for the universe total is

$$43{,}259 \pm 1.959963984540(1{,}547)$$

that is, 40,227 to 46,291 ($40,227,000 to $46,291,000).

FORMULAS

1. Estimated total in the universe (T)

$$\hat{T} = \sum_{h=1}^{L} \hat{T}_h$$

   where $\hat{T}_h$ is the point estimate for the universe total in stratum h and L is the number of strata.

2. Estimated variance of $\hat{T}$

$$v(\hat{T}) = \sum_{h=1}^{L} v(\hat{T}_h)$$

   where $v(\hat{T}_h)$ is the estimated variance of $\hat{T}_h$ and is equal to the square of the standard error of $\hat{T}_h$.

3. Approximate 95% confidence interval for T

$$\hat{T} \pm 1.959963984540\sqrt{v(\hat{T})}$$

NOTE: For a 90% confidence interval, replace 1.959963984540 with 1.644853626951 and for an 80% confidence interval replace 1.959963984540 with 1.281551565545.

Poststratification

Oftentimes sampling problems arise in which the user would like to stratify on a key variable but cannot place the sampling units into their correct strata until after the sample is selected. Another situation arises when an auditor does not recognize a need to stratify prior to obtaining a simple random sample and the sample items are evaluated. The poststratification program is designed for such situations and provides reliable results if the overall sample size is large and the poststratified sample sizes are large (say, at least 20). It is however, less efficient than using a prestratified sample; that is, produces a slightly wider confidence interval for the same sample size.

A key thing to keep in mind here is that **the universe strata sizes must be known**. Consequently, before you define a set of strata, make sure that the you know the number of universe items in each of the strata. The program does not allow you to estimate these universe sizes. Poststratification is often appropriate when a simple random sample is not properly balanced according to major groupings of the population.

Example 9. In a recent hospital audit, the amount of unallowable bad debts was determined for a particular year. For the universe, it was known that there were $N_1 = 373$ inpatient bad debts (Stratum 1) and $N_2 = 1,146$ outpatient bad debts (Stratum 2). The total universe size is $N = N_1 + N_2 = 1,519$. So there are roughly 20% inpatient and 80% outpatient bad debts in the universe. Suppose that a simple random sample of 100 bad debts revealed:

<u>Inpatient bad debts</u>   <u>Outpatient bad debts</u>

$n_1 = 45$ $n_2 = 55$

$\bar{y}_1 = \$240.00$ $\bar{y}_2 = \$30.00$

$s_1^2 = 22.04$ $s_2^2 = 198.56$

The data files (POSTDATA.TXT and UNIVPOST.TXT) are shown on the next page and the

resulting computer output immediately follows.

Data file POSTDATA.TXT   < - continued - >   ## Universe file UNIVPOST.TXT

1	242.27		51	25.99		1	373	45
2	240.43		52	31.75		2	1146	55
3	243.43		53	42.37			
4	241.00		54	28.82			
5	235.71		55	14.80			
6	239.74		56	22.44			
7	249.53		57	12.93			
8	232.61		58	37.22			
9	243.51		59	12.29			
10	243.70		60	21.02			
11	241.36		61	24.76			
12	238.86		62	49.69			
13	241.20		63	50.43			
14	246.14		64	9.03			
15	236.91		65	32.47			
16	244.91		66	28.43			
17	231.91		67	18.54			
18	249.58		68	31.48			
19	227.84		69	4.45			
20	239.07		70	54.76			
21	232.89		71	9.45			
22	242.15		72	19.73			
23	238.93		73	33.71			
24	243.41		74	59.84			
25	241.61		75	22.73			
26	240.04		76	15.54			
27	229.43		77	22.14			
28	238.12		78	21.81			
29	240.82		79	53.10			
30	241.84		80	38.72			
31	236.08		81	36.58			
32	248.11		82	39.26			
33	239.39		83	53.71			
34	236.00		84	40.95			
35	238.67		85	24.82			
36	242.88		86	45.05			
37	241.90		87	25.30			
38	238.96		88	58.31			
39	234.03		89	25.30			
40	237.86		90	9.48			
41	239.01		91	15.93			
42	239.21		92	25.00			
43	245.45		93	29.27			
44	242.08		94	26.29			
45	241.39		95	38.26			
46	46.32		96	31.30			
47	37.59		97	37.97			
48	31.23		98	38.52			
49	22.92		99	11.53			
50	44.96		100	3.64			

```
                    DEPARTMENT OF HEALTH & HUMAN SERVICES
                       OIG – OFFICE OF AUDIT SERVICES
Date: 10/24/2004       POSTSTRATIFIED VARIABLE APPRAISAL        Time: 9:51
                     AUDIT/REVIEW: Poststratification


                   DATA FILE USED: C:\Temp\POSTDATA.TXT


            ---------------------D I F F E R E N C E---------------------
Stratum 1   SAMPLE SIZE / UNIVERSE SIZE              45                 373
            MEAN                                  240.00
            STANDARD DEVIATION                      4.69
            STANDARD ERROR (TOTAL)                347.13
            POINT ESTIMATE                        89,520

                                  CONFIDENCE LIMITS
                                80% CONFIDENCE LEVEL
            LOWER LIMIT                          89,075
            UPPER LIMIT                          89,965
            PRECISION AMOUNT                        445
            PRECISION PERCENT                       .50%
            Z-VALUE USED                  1.281551565545

                                90% CONFIDENCE LEVEL
            LOWER LIMIT                          88,949
            UPPER LIMIT                          90,091
            PRECISION AMOUNT                        571
            PRECISION PERCENT                       .64%
            Z-VALUE USED                  1.644853626951

                                95% CONFIDENCE LEVEL
            LOWER LIMIT                          88,839
            UPPER LIMIT                          90,200
            PRECISION AMOUNT                        680
            PRECISION PERCENT                       .76%
            Z-VALUE USED                  1.959963984540

Stratum 2   SAMPLE SIZE / UNIVERSE SIZE              55               1,146
            MEAN                                   30.00
            STANDARD DEVIATION                     14.09
            STANDARD ERROR (TOTAL)              1,800.03
            POINT ESTIMATE                        34,379

                                  CONFIDENCE LIMITS
                                80% CONFIDENCE LEVEL
            LOWER LIMIT                          32,072
            UPPER LIMIT                          36,685
            PRECISION AMOUNT                       2,307
            PRECISION PERCENT                      6.71%
            Z-VALUE USED                  1.281551565545

                                90% CONFIDENCE LEVEL
            LOWER LIMIT                          31,418
            UPPER LIMIT                          37,339
            PRECISION AMOUNT                       2,961
            PRECISION PERCENT                      8.61%
            Z-VALUE USED                  1.644853626951
```

```
                               95% CONFIDENCE LEVEL
          LOWER LIMIT                           30,851
          UPPER LIMIT                           37,907
          PRECISION AMOUNT                        3,528
          PRECISION PERCENT                       10.26%
          Z-VALUE USED                    1.959963984540

OVERALL   SAMPLE SIZE / UNIVERSE SIZE             100            1,519
          POINT ESTIMATE                       123,898
          STANDARD ERROR                         1,833

                                CONFIDENCE LIMITS
                               80% CONFIDENCE LEVEL
          LOWER LIMIT                          121,549
          UPPER LIMIT                          126,248
          PRECISION AMOUNT                        2,349
          PRECISION PERCENT                        1.90%
          Z-VALUE USED                    1.281551565545

                               90% CONFIDENCE LEVEL
          LOWER LIMIT                          120,883
          UPPER LIMIT                          126,914
          PRECISION AMOUNT                        3,015
          PRECISION PERCENT                        2.43%
          Z-VALUE USED                    1.644853626951

                               95% CONFIDENCE LEVEL
          LOWER LIMIT                          120,305
          UPPER LIMIT                          127,491
          PRECISION AMOUNT                        3,593
          PRECISION PERCENT                        2.90%
          Z-VALUE USED                    1.959963984540
```

Discussion. Using the usual estimator for a simple random sample, the estimate of the

universe mean is

$$\bar{y} = [(45)(240) + (55)(30)]/100 = \$124.50$$

and the estimate of the universe total is

$$\hat{T} = (1519)(124.50) = \$189,115.50$$

Since there is an unusually high number of inpatient bad debts (and low outpatient), a better

procedure would be to use the poststratified estimate of the universe total; namely

$$\hat{T}_{pst} = (373)(240.00) + (1146)(30.00) = \$123,900$$

The (more precise) computer-generated point estimate is $123,898 (highlighted).

Also, the point estimate for the inpatient stratum is $\hat{T}_1 = (373)(240.00) = \$89,520$ (highlighted);

for the outpatient stratum, this estimate is $\hat{T}_2 = (1146)(30.00) = \$34,380$ (more precisely, the

highlighted value of $34,379). Notice that $\hat{T}_{pst} = 89,520 + 34,379 = \$123,899$ (more precisely,

$123,898).

The estimated variance of $\hat{T}_{pst}$ is $v(\hat{T}_{pst}) = v(\hat{T}_1) + v(\hat{T}_2)$

$$= \left[\frac{1419}{100}(373)(4.69484)^2 + \frac{1519}{100^2}(1146)(4.69484)^2\right] + \left[\frac{1419}{100}(1146)(14.09098)^2 + \frac{1519}{100^2}(373)(14.09098)^2\right]$$

$= 120,499.88 + 3,240,111.44 = 3,360,611$ and the estimated standard error is $\sqrt{3,360,611}$ =

$1,833 (highlighted)

NOTE: The estimated standard error for stratum 1 is $\sqrt{120,499.88} = 347.13$ and the
estimated standard error for stratum 2 is $\sqrt{3,240,111.44} = 1,800.03$.

The corresponding 95% confidence interval for the universe total is

$$123,898 \pm 1.959963984540(1,833)$$

that is, $120,305 to $127,491.

Comments

1. Poststratification allows you to obtain a single simple random sample (easier than obtaining a
    simple random sample from each stratum) and then stratify provided the strata sizes in the
    universe (N_i) are known. The value of N_i is multiplied by the sample mean $\bar{y}_i$ to estimate

the total for the i-th stratum. These estimates are then summed over all the strata to estimate the universe total.

2. A minimum of 20 sampling units per stratum is required as well as 6 nonzero items per stratum (OA Policy and Procedures). The total sample size should be at least 100.

3. With poststratification, the sample sizes (n_i) are unknown in advance (random variables). With stratified sampling, the sample sizes are fixed (nonrandom).

FORMULAS

1. Estimate of the universe total for the i-th stratum (T_i)

$$\hat{T}_i = N_i \bar{y}_i$$

where N_i = number of items (universe) in stratum i

$\bar{y}_i$ = average of sample items in i-th stratum

L = number of strata

2. Estimate of universe total (T)

$$\hat{T}_{pst} = \sum_{i=1}^{L} N_i \bar{y}_i = \sum_{i=1}^{L} \hat{T}_i$$

3. Estimated variance of $\hat{T}_{pst} = v(\hat{T}_{pst}) = \sum_{i=1}^{L} v(\hat{T}_i)$ where

$$v(\hat{T}_i) = \left(\frac{N-n}{n} \right) N_i s_i^2 + \frac{N}{n^2} (N - N_i) s_i^2$$

where N = universe total = ΣN_i and n is the total sample size.

4. Estimated standard error of $\hat{T}_i = \sqrt{v(\hat{T}_i)}$

5. Approximate 95% confidence interval for stratum total (T_i):

$$\hat{T}_i \pm 1.959963984540\sqrt{v(\hat{T}_i)}$$

6. Estimated standard error of $\hat{T}_{pst}$ is $\sqrt{v(\hat{T}_{pst})}$

7. Approximate 95% confidence interval for universe total (T):

$$\hat{T}_{pst} \pm 1.959963984540\sqrt{v(\hat{T}_{pst})}$$

NOTE: For a 90% confidence interval in equations 5 and 7, replace 1.959963984540 with 1.644853626951 and for an 80% confidence interval replace 1.959963984540 with 1.281551565545.

Unknown Universe Size

This program calculates a confidence interval for a universe total when using variable sampling, as does the Unrestricted Variable Appraisal program. When using the Unrestricted Variable Appraisal program, one of the user queries is for the universe size (N) and this value must be known. For situations where N is unknown, the Unknown Universe Size program can be used.

Use of this program requires that two random samples be used -- one to estimate the universe size and the other to estimate one or more variable characteristics. Both samples must be appraised using the Unrestricted Variable Appraisal program prior to running this module since this program will ask for the mean and standard deviation of each sample.

The population of interest is a subset of some other universe. For example, the larger sampling frame might consist of 575 file drawers containing a mixture of dental claims and the population of interest consists of all claims related to a particular dental procedure (say, procedure ABC). Suppose we sample 70 of the file drawers and count the number of claims related to procedure ABC. The first step is to estimate the number of claims related to procedure ABC in all 575 drawers. The results on the next page were obtained and are stored in data file DATAUNIV.TXT.

	Number of Claims Related
Sampled Drawer	to Procedure ABC
1	9
2	12
3	9
.	.
.	.
.	.
70	10
	Total 723

Sample summary

$$\bar{x} = 10.33$$
$$s = 2.75$$

Data file DATAUNIV.TXT   <continued>

1	9	36	9
2	12	37	12
3	9	38	8
4	6	39	6
5	12	40	14
6	13	41	9
7	10	42	14
8	9	43	14
9	10	44	6
10	6	45	14
11	13	46	12
12	7	47	9
13	12	48	10
14	7	49	9
15	12	50	9
16	12	51	8
17	10	52	14
18	9	53	9
19	10	54	14
20	10	55	8
21	13	56	12
22	10	57	14
23	10	58	10
24	10	59	12
25	6	60	9
26	8	61	12
27	14	62	12
28	22	63	14
29	10	64	10
30	8	65	10
31	8	66	6
32	8	67	9
33	12	68	8
34	10	69	12
35	8	70	10

NOTE: This file contains 70 lines.

The following computer output was obtained using the **Unrestricted Variable Appraisal**

program. In the **Universe Size** box, the size of the larger universe (575 file drawers here) was

used. With this procedure, you are able to see the estimated size of the universe of interest

($\hat{N}$ = 5,939) in this output.

```
                      DEPARTMENT OF HEALTH & HUMAN SERVICES
                         OIG - OFFICE OF AUDIT SERVICES
    Date: 10/24/2004        VARIABLE UNRESTRICTED APPRAISAL        Time: 12:12
                           AUDIT/REVIEW: First Sample

                      DATA FILE USED: C:\Temp\DATAUNIV.TXT


          SAMPLE SIZE        VALUE OF SAMPLE   NONZERO ITEMS
               70                 723.00              70

         ---------------------- E X A M I N E D ----------------------
         MEAN / UNIVERSE                       10.33————————————————575
         STANDARD DEVIATION                     2.75——┐
         STANDARD ERROR                          .31 NOTE: Input to Unknown
         SKEWNESS                               1.02 Universe Size program.
         KURTOSIS                               5.95
         POINT ESTIMATE                        5,939 <--estimated universe
                                                                  size

                               CONFIDENCE LIMITS
                            80% CONFIDENCE LEVEL
         LOWER LIMIT                          5,710
         UPPER LIMIT                          6,168
         PRECISION AMOUNT                       229
         PRECISION PERCENT                     3.86%
         T-VALUE USED                1.293941609194

                            90% CONFIDENCE LEVEL
         LOWER LIMIT                          5,644
         UPPER LIMIT                          6,234
         PRECISION AMOUNT                       295
         PRECISION PERCENT                     4.97%
         T-VALUE USED                1.667238548669

                            95% CONFIDENCE LEVEL
         LOWER LIMIT                          5,586
         UPPER LIMIT                          6,292
         PRECISION AMOUNT                       353
         PRECISION PERCENT                     5.95%
         T-VALUE USED                1.994945415107
```

The next step is to <u>independently</u> obtain a random sample from the population of interest to

appraise the variable(s) of interest. It was decided to sample 55 claims related to procedure ABC

and record the amount in error for each sampled claim. The results are stored in data file

DATAVAR.TXT.

Data file DATAVAR.TXT   **< - continued - >**

1	15.05			
2	14.05	29	16.07	
3	18.55	30	20.10	
4	13.62	31	17.29	
5	18.78	32	21.38	
6	6.97	33	16.60	
7	17.17	34	6.98	
8	23.75	35	16.85	
9	19.32	36	19.25	**NOTE: This file contains 55**
10	21.32	37	20.30	**lines.**
11	25.17	38	29.51	
12	14.12	39	6.95	
13	15.85	40	19.00	
14	13.50	41	13.93	
15	14.62	42	12.45	
16	6.97	43	15.23	
17	22.62	44	15.09	
18	6.97	45	6.98	
19	15.61	46	20.48	
20	12.69	47	6.98	
21	21.45	48	16.47	
22	21.62	49	19.75	
23	17.95	50	6.98	
24	14.65	51	17.65	
25	6.97	52	15.26	
26	18.89	53	6.98	
27	15.34	54	14.15	
28	6.98	55	22.05	

This file is used as input to the **Unrestricted Variable Appraisal** program. The user can enter

any value in the **Universe Size** box since this value has no effect on the results produced by the

Unknown Universe Size program. The following computer output was produced.


```
                    DEPARTMENT OF HEALTH & HUMAN SERVICES
                       OIG - OFFICE OF AUDIT SERVICES
Date: 10/24/2004        VARIABLE UNRESTRICTED APPRAISAL        Time: 13:12
                       AUDIT/REVIEW: Second Sample

              DATA FILE USED: C:\Temp\DATAVAR.TXT


         SAMPLE SIZE        VALUE OF SAMPLE    NONZERO ITEMS
             55                   860.76              55

                                             Any value can be
                                             used for universe

        -------------------- D I F F E R E N C E ---------------------
        MEAN / UNIVERSE                      15.65              1,000
        STANDARD DEVIATION                    5.45
        STANDARD ERROR                         .71  NOTE: Input to Unknown
        SKEWNESS                              -.14  Universe Size program.
        KURTOSIS                             2.63
        POINT ESTIMATE                     15,650

                              CONFIDENCE LIMITS
                            80% CONFIDENCE LEVEL
        LOWER LIMIT                        14,723
        UPPER LIMIT                        16,577
        PRECISION AMOUNT                      927
        PRECISION PERCENT                    5.92%
        T-VALUE USED                 1.297426488209

                            90% CONFIDENCE LEVEL
        LOWER LIMIT                        14,455
        UPPER LIMIT                        16,846
        PRECISION AMOUNT                    1,196
        PRECISION PERCENT                    7.64%
        T-VALUE USED                 1.673564906352

                            95% CONFIDENCE LEVEL
        LOWER LIMIT                        14,218
        UPPER LIMIT                        17,082
        PRECISION AMOUNT                    1,432
        PRECISION PERCENT                    9.15%
        T-VALUE USED                 2.004879288188
```

Comment: This is the same example used in the RAT-STATS User's Guide illustration. The

following computer output is produced by the Unknown Universe Size program:

```
                    DEPARTMENT OF HEALTH & HUMAN SERVICES
                       OIG - OFFICE OF AUDIT SERVICES
Date: 10/24/2004   VARIABLE APPRAISAL WITH UNKNOWN UNIVERSE SIZE   Time: 13:48
                   AUDIT/REVIEW: Variable Unknown Universe Size

   = = = = = = = = = = = = = = = = = I N P U T = = = = = = = = = = = = = = = = = =
                                      SAMPLE TO                SAMPLE FOR
                                  ESTIMATE POPULATION    VARIABLE ATTRIBUTE
     UNIVERSE                              575
     SAMPLE                                 70                      55
     MEAN                                 10.33                   15.65
     STANDARD DEVIATION                    2.75                    5.45

   = = = = = = = = = = = = = =  E S T I M A T I O N  = = = = = = = = = = = = = =
                      80% CONFIDENCE        90% CONFIDENCE        95% CONFIDENCE

     POINT ESTIMATE     T̂ --> 92,957             92,957                92,957
     STANDARD ERROR          5,152               5,152                 5,152
     LOWER LIMIT            86,355              84,483                82,859
     UPPER LIMIT            99,560             101,431               103,055
     PRECISION AMOUNT        6,603               8,474                10,098
     PRECISION PERCENT       7.10%               9.12%                10.86%
     Z-VALUE USED      1.281551565545     1.644853626951      1.959963984540
```

Discussion. The estimated total dollars in the universe is equal to

> (estimated universe size)(mean of the variable sample)

> $= [(10.33)(575)](15.65) = \$92,957.$ This is $\hat{T}$.

Using the formula section, the value of se_1 is

$$\frac{(2.75)(575)}{\sqrt{70}}\sqrt{1 - \frac{70}{575}} = 177.118$$

and the value of se_2 is

$$\frac{5.45}{\sqrt{55}}\sqrt{1 - \frac{55}{5939}} = .7315$$

The estimated variance of $\hat{T}$ is

$$[(15.65)(177.118)]^2 + [(5939)(.7315)]^2 - [(177.118)(.7315)]^2 = 26,540,250.$$

The estimated standard error of $\hat{T}$ is $\sqrt{26,540,250} = \$5,152$.
The PRECISION AMOUNT at the 90% confidence level is $(5,152)(1.644853626951) = \$8,474$

(highlighted). This amount is 9.12% of the point estimate, since $8,474/92,957 = .0912$.

The corresponding confidence interval is $92,957 \pm 8,474$; that is, \$84,483 to \$101,431.

FORMULAS

Given: Larger universe size: N_1. Sample to estimate universe size: sample size is n_1, mean is

$\overline{x}_1$, and standard deviation is s_1.

Sample to estimate variable of interest: sample size is n_2, mean is $\overline{x}_2$, and standard

deviation is s_2.

1. Estimate of universe (of interest) size

$$\hat{N} = N_1 \bullet \overline{x}_1$$

2. Overall estimate for variable total (e.g., total error amount)

$$\hat{T} = \hat{N} \bullet \overline{x}_2$$

3. Estimated variance of $\hat{T}$

$$v(\hat{T}) = [\overline{x}_2 \bullet se_1]^2 + [\hat{N} \bullet se_2]^2 - [se_1 \bullet se_2]^2$$

where

$$se_1 = \frac{s_1 N_1}{\sqrt{n_1}} \sqrt{1 - \frac{n_1}{N_1}} \quad \text{and} \quad se_2 = \frac{s_2}{\sqrt{n_2}} \sqrt{1 - \frac{n_2}{\hat{N}}}$$

4. Approximate 95% confidence interval for universe total (T):

$$\hat{T} \pm 1.959963984540\sqrt{v(\hat{T})}$$

NOTE: For a 90% confidence interval, replace 1.959963984540 with 1.644853626951 and for an 80% confidence interval replace 1.959963984540 with 1.281551565545.

Reference:
Estimated variance: Kendall's Advanced Theory of Statistics, Volume 1, 5th ed., Alan Stuart and J. Keith Ord, New York: Oxford University Press, 1987, page 343, exercise 10.23.

SAMPLE SIZE DETERMINATION

A commonly encountered question in auditing is "How large a sample is necessary?" When

using an unrestricted (simple random) sample, this depends on the desired precision of the point

estimate. The programs in this section are listed below and are concerned with determining

sample sizes for various data types and sample strategies.

- Variable
  - Unrestricted Using a Probe Sample
  - Unrestricted Using Estimated Error Rate
  - Stratified (Total Sample Size Known)
  - Stratified (Total Sample Size Unknown)

Variable Sample Size Determination

This RAT-STATS module can be used for two situations.

Situation 1: The program will help select the necessary sample size for an unrestricted or stratified variable appraisal. The program output includes sample sizes for each stratum that will provide precision percentages of 1%, 2%, 5%, 10%, 20% and "Other." When selecting "Other," the user will be prompted to enter the desired precision percentage. The user may also select any combination of the following confidence levels: 80%, 90%, 95%, and 99%.

Situation 2: The program also allows the user to determine the optimum distribution of a sample among strata when the overall sample size has already been determined. It will allocate the larger samples to those strata that are larger in size and/or contain a larger amount of variation (are nonhomogeneous). Any combination of the confidence levels 80%, 90%, 95%, and 99% can be selected.

Variable Sample Size Determination - Unrestricted Using a Probe Sample

This program allows the user to estimate sample sizes for specified precision percentages and specified confidence levels. The user has the option of having the program read a probe sample file to obtain an estimate of the universe mean and standard deviation or input these two estimates directly without reading a probe sample file. The probe sample may be stored in a text file, an Excel spreadsheet, or an Access table.

Example 1. This example illustrates Situation 1. A probe sample of 25 examined values was obtained. The audit objective was to determine the necessary sample sizes when estimating the total examined amount for the universe of 100,000 items. The probe sample (SAMPDATA.TXT) is shown below. The sample mean is $400 and the sample standard deviation is $50.

```
321
382
453
459
343
388
313
420
407
395
441
448
447
333
357
395
477
391
356
368
376
350
461
472
447
```

The input screen and resulting text file output are shown on the following page.

The following text file output is obtained using the previous screen. A sample size under 30 will

be flagged using "(*)" and the note immediately following the sample sizes will appear.

```
                  DEPARTMENT OF HEALTH & HUMAN SERVICES
                      OIG - OFFICE OF AUDIT SERVICES
Date: 5/11/2004       Sample Size Determination       Time: 21:52

                              Confidence Level
                  80%         90%           95%           99%
           1%     256         421           597           1026
           2%     64          106           150           259
Precision  5%     10 (*)      17 (*)        24 (*)        41
Level      10%    3 (*)       4 (*)         6 (*)         10 (*)
           15%    1 (*)       2 (*)         3 (*)         5 (*)
           25%    ---         1 (*)         1 (*)         2 (*)

Estimated Mean: 400.00
Estimated Std. Deviation: 50.00
Universe Size: 100,000
```

**NOTE (*): One or more sample sizes were under 30. The
         generated sample sizes were the result of
         mathematical formulas and did not
         incorporate management decisions concerning
         the purpose of the sample or current
         organizational sampling policies. You may
         need to increase the sample sizes in order
         to be in compliance with organizational
         objectives.**

Explanation of Output

The output for each cell in the output table will consist of (1) the necessary sample size or (2) the

text "- - -". The necessary sample size is the number of sample items necessary to obtain the

specified sample precision at the specified confidence level. For example, in this illustration, a

sample size of 106 is necessary to obtain a point estimate having a precision percentage of plus

or minus 2% using a 90% confidence level. If the calculated sample size is 0, a text value of "- -

-" will appear in this cell. This occurred in the lower left cell for the sample illustration.

The output also contains the estimated mean and standard deviation, along with the specified

universe size.

FORMULAS

Let PREC = the precision percentage (e.g., 1 for 1%, 10 for 10%)

ZVAL = the value from the standard normal (Z) distribution having a right-tail area equal to (100 - Confidence Level)/2, where the right-tail area is expressed as a proportion between 0 and 1.

ZVAL is 1.281551565545(80%), 1.644853626951 (90%), 1.959963984540 (95%), and 2.575829303549 (99%).

N = the universe size

Mean = estimated universe mean obtained from the probe sample or specified by the user

StdDev = estimated universe standard deviation obtained from the probe sample or specified by the user

E = maximum error = (PREC/100) · Mean · N

For each selected value of PREC and ZVAL, the sample size is

$$ n = \frac{(StdDev \cdot N)^2}{(E / ZVAL)^2 + N \cdot (StdDev)^2} $$

The value of n is rounded up or down to the nearest integer.

Variable Sample Size Determination - Unrestricted Using Expected Error Rate

This procedure estimates the mean and standard deviation of the difference (error) amounts by assuming (1) any item found to be in error is 100% in error and (2) the mean and standard deviation of the *nonzero* error amounts is the same as the mean and standard deviation of the reported (examined) amounts. The mean and standard deviation of the error amounts are estimated by assuming the percentage of nonzero errors in the error population is equal to the expected error rate (one of the input values) and the nonzero errors resemble the reported amounts; that is, the mean and standard deviation of the nonzero errors are equal to the mean and standard deviation of the reported amounts.

Comment.   Even though these assumptions may not be entirely true, this procedure will often give more reliable sample size estimates than those obtained using the Variable Unrestricted (Using Reported Amounts) module since the expected number of zero values in the error population is factored into the sample size calculation.

Example 2. This example illustrates another method of dealing with Situation 1. The estimated error rate is 15% for a universe of 10,000 transactions. The total reported amount is $3,000,000 and the standard deviation of the reported amounts is $125. Consequently, the mean reported amount is $300. Of interest is the required sample size necessary in order to obtain plus or minus 15% using a 90% confidence level. The corresponding input screen follows where 25% was specified for the "Other" precision level.

The text file output shown on the next page is obtained using this input screen. A sample size under 30 will be flagged using "(*)" in the program output and a note informing the user of this fact will also appear in the program output immediately following the sample sizes.

```
                         DEPARTMENT OF HEALTH & HUMAN SERVICES
                           OIG - OFFICE OF AUDIT SERVICES
           Date: 12/22/2004           Sample Size Determination         Time: 10:14


                                    Confidence Level
                       80%          90%          95%          99%
                1%     9181         9486         9633         9784
                2%     7370         8219         8676         9188
   Precision    5%     3095         4248         5119         6443
   Level       10%     1008         1559         2077         3117
               15%      474          758         1044         1675
               25%      176          287          403          675


   Universe Size: 10,000

   Anticipated Error Rate: 15%

   Reported Amounts - - Total Amount: 3,000,000.00

                       Standard Deviation: 125.00


   Difference Values - - Estimated Mean: 45.00

                        Estimated Standard Deviation: 117.55
```

Explanation of Output

The output for each cell in the output table will consist of (1) the necessary sample size or (2) the

text "- - -". The necessary sample size is the number of sample items necessary to obtain the

specified sample precision at the specified confidence level. For example, in this illustration, a

sample size of 287 is necessary to obtain a point estimate having a precision percentage of plus

or minus 15% using a 90% confidence level. If the calculated sample size is 0, a text value of

"- - -" will appear in this cell.

The output also contains the estimated mean and standard deviation of the difference (error)

values. For this illustration, the estimated mean and standard deviation are $45.00 and $117.55,

respectively.

FORMULAS

Let PREC = the precision percentage (e.g., 1 for 1%, 10 for 10%)

ZVAL =   the value from the standard normal (Z) distribution having a right-tail area equal to (100 - Confidence Level)/2, where the right-tail area is expressed as a proportion between 0 and 1.

ZVAL is 1.281551565545(80%), 1.644853626951 (90%), 1.959963984540 (95%), and 2.575829303549 (99%).

N =   the universe size (input)

T_R = the total reported amount (input)

μ_R = mean reported amount = T_R / N

σ_R = standard deviation of reported amounts (input)

$\hat{p}$ = the estimated error rate (input)

$\hat{\mu}_D$ = the estimated mean of the difference (error) values = $\hat{p}\mu_R$

$\hat{\sigma}_D$ = estimated standard deviation of the difference (error) values

$$= \sqrt{\hat{p}[\sigma_R^2 + (1-\hat{p})\mu_R^2]}$$

E = maximum error = (PREC/100) $\cdot$ $\hat{\mu}_D$ $\cdot$ N

For each selected value of PREC and ZVAL, the sample size is

$$n = \frac{(\hat{\sigma}_D \cdot N)^2}{(E / ZVAL)^2 + N \cdot (\hat{\sigma}_D)^2}$$
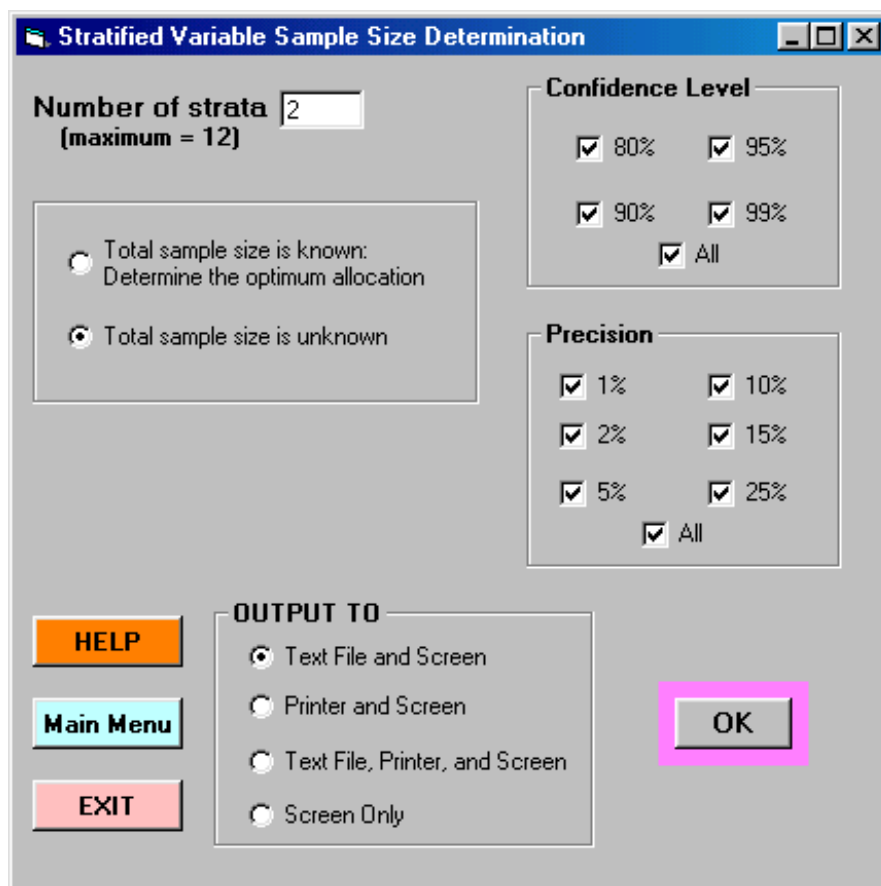
The value of n is rounded up or down to the nearest integer.

Variable Sample Size Determination - Stratified

Stratified Sample Sizes - Total Sample Size is Unknown

Example 3. This example illustrates Situation 1. Two strata have been defined: The high-income stratum (N_1 = 100,000 items) and the low-income stratum (N_2 = 500,000 items). Of interest is the total audit (claimed) amount for the universe. For the high-income stratum, the estimated mean of the audited amounts is $10,000 and the estimated standard deviation is $5,000. These values for the low-income stratum are $5,000 (mean) and $4,000 (standard deviation). At a confidence level of 95%, what sample size is required to obtain a precision percentage of ± 10%?

Solution. The following input screen was used for this example.

The following output is obtained using the previous screen. If one or more of the sample sizes

are under 30, the note immediately following the total sample sizes will appear.

```
                    DEPARTMENT OF HEALTH & HUMAN SERVICES
                        OIG - OFFICE OF AUDIT SERVICES
Date: 10/19/2004            Sample Size Determination              Time: 12:02


            THE ESTIMATES ARE BASED ON THE FOLLOWING ENTRIES:

NBR    DESCRIPTION       -- MEAN --   -- STD.DEV. --   -- UNIVERSE --   -- RATIO --
  1    High Income       10,000.00        5,000.00         100,000          20.00%
  2    Low Income         5,000.00        4,000.00         500,000          80.00%

- TOTALS -                5,833.33        4,579.54         600,000


= = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = =

Sample Sizes for Stratum 1: High Income

                          Confidence Level
                80%         90%          95%           99%
          1%    1653        2699         3795          6406
          2%    418         687          972           1669
Precision 5%    67          111          157           271
Level     10%   17 (*)      28 (*)       40            68
          15%   8 (*)       13 (*)       18 (*)        31
          25%   3 (*)       5 (*)        7 (*)         11 (*)

Sample Sizes for Stratum 2: Low Income

                          Confidence Level
                80%         90%          95%           99%
          1%    6611        10793        15180         25624
          2%    1671        2745         3888          6676
Precision 5%    268         442          627           1081
Level     10%   68          111          157           271
          15%   30          50           70            121
          25%   11 (*)      18 (*)       26 (*)        44

Total Sample Sizes

                          Confidence Level
                80%         90%          95%           99%
          1%    8264        13492        18975         32030
          2%    2089        3432         4860          8345
Precision 5%    335         553          784           1352
Level     10%   85          139          197           339
          15%   38          63           88            152
          25%   14 (*)      23 (*)       33            55
```

```
NOTE (*):   One or more sample sizes were under 30.  The generated
            sample sizes were the result of mathematical formulas
            and did not incorporate management decisions
            concerning the purpose of the sample or current
            organizational sampling policies.  You may need to
            increase the sample sizes in order to be in compliance
            with organizational objectives.
```

If any of the calculated samples sizes exceeds the corresponding universe size, the program will

conclude with the following reminder:

```
NOTE (#):   The formulas calculated a sample size greater than the
            universe size.  The program reduced the calculated
            sample size to the universe size.  The additional
            sampling units were then distributed among the
            remaining strata based on optimal allocation formulas.
```

Discussion. For 10% precision and 95% confidence, the total sample size required is n = 197

with $n_1 = 40$ items to be obtained from the high-income stratum and $n_2 = 157$ from the low-

income stratum. Consequently, a 95% confidence interval based on these sample sizes should

result in a precision percentage of ±10%. This assumes that the resulting sample means and

standard deviations are the same as the values used as input to this program.

To demonstrate this, a data set was constructed that contained 40 items from stratum 1 with a

sample mean and standard deviation of $10,000 and $5,000, respectively, and 157 items from

stratum 2 with a sample mean and standard deviation of $5,000 and $4,000, respectively. When

this data set (named STRATA.TXT) was used as input to the STRATIFIED VARIABLE

APPRAISAL module, the computer output on the next page was obtained. In the final portion of

the output, notice that the resulting point estimate for the universe total is 3,500,000,000. At the

95% confidence level, the precision amount is 349,043,863 and is in fact (approximately) 10% of

the point estimate.

```
                    DEPARTMENT OF HEALTH & HUMAN SERVICES
                       OIG - OFFICE OF AUDIT SERVICES
Date: 10/23/2004          STRATIFIED VARIABLE APPRAISAL          Time: 10:36
                       AUDIT/REVIEW: Two Strata Example

                    DATA FILE USED: C:\Temp\STRATA.TXT


        STRATUM      SAMPLE
         NUMBER       SIZE      VALUE OF SAMPLE    NONZERO ITEMS
            1          40         400,000.00                 40
            2         157         785,000.00                157

        TOTALS        197       1,185,000.00                197


               -------------------- D I F F E R E N C E --------------------
Stratum 1  MEAN / UNIVERSE                    10,000.00              100,000
           STANDARD DEVIATION                  5,000.04  ◄ Approx. 5,000
           STANDARD ERROR                        790.42
           SKEWNESS                               -.22
           KURTOSIS                               2.30
           STANDARD ERROR (MEAN)                 790.42
           STANDARD ERROR (TOTAL)            79,041,756
           POINT ESTIMATE                1,000,000,000

                               CONFIDENCE LIMITS
                             80% CONFIDENCE LEVEL
               LOWER LIMIT                  896,958,117
               UPPER LIMIT                1,103,041,883
               PRECISION AMOUNT             103,041,883
               PRECISION PERCENT                 10.30%
               T-VALUE USED              1.303638588621

                             90% CONFIDENCE LEVEL
               LOWER LIMIT                  866,824,512
               UPPER LIMIT                1,133,175,488
               PRECISION AMOUNT             133,175,488
               PRECISION PERCENT                 13.32%
               T-VALUE USED              1.684875121711

                             95% CONFIDENCE LEVEL
               LOWER LIMIT                  840,122,958
               UPPER LIMIT                1,159,877,042
               PRECISION AMOUNT             159,877,042
               PRECISION PERCENT                 15.99%
               T-VALUE USED              2.022690920037

Stratum 2  MEAN / UNIVERSE                     5,000.00              500,000
           STANDARD DEVIATION                  3,999.81  ◄ Approx. 4,000
           SKEWNESS                                .87
           KURTOSIS                               3.30
           STANDARD ERROR (MEAN)                 319.17
           STANDARD ERROR (TOTAL)           159,584,887
           POINT ESTIMATE                2,500,000,000
```

```
                                    CONFIDENCE LIMITS
                                  80% CONFIDENCE LEVEL
            LOWER LIMIT                   2,294,613,944
            UPPER LIMIT                   2,705,386,056
            PRECISION AMOUNT                205,386,056
            PRECISION PERCENT                     8.22%
            T-VALUE USED                  1.287001917850

                                  90% CONFIDENCE LEVEL
            LOWER LIMIT                   2,235,938,079
            UPPER LIMIT                   2,764,061,921
            PRECISION AMOUNT                264,061,921
            PRECISION PERCENT                    10.56%
            T-VALUE USED                  1.654679995672

                                  95% CONFIDENCE LEVEL
            LOWER LIMIT                   2,184,773,966
            UPPER LIMIT                   2,815,226,034
            PRECISION AMOUNT                315,226,034
            PRECISION PERCENT                    12.61%
            T-VALUE USED                  1.975287507703

OVERALL     POINT ESTIMATE / UNIVERSE    3,500,000,000          600,000
            STANDARD ERROR                 178,086,876

                                    CONFIDENCE LIMITS
                                  80% CONFIDENCE LEVEL
            LOWER LIMIT                   3,271,772,485
            UPPER LIMIT                   3,728,227,515
            PRECISION AMOUNT                228,227,515
            PRECISION PERCENT                     6.52%
            Z-VALUE USED                  1.281551565545

                                  90% CONFIDENCE LEVEL
            LOWER LIMIT                   3,207,073,156
            UPPER LIMIT                   3,792,926,844
            PRECISION AMOUNT                292,926,844
            PRECISION PERCENT                     8.37%
            Z-VALUE USED                  1.644853626951

                                  95% CONFIDENCE LEVEL
            LOWER LIMIT                   3,150,956,137
            UPPER LIMIT                   3,849,043,863
            PRECISION AMOUNT                349,043,863
            PRECISION PERCENT                     9.97%   ◄
            Z-VALUE USED                  1.959963984540
```

Comments

(1)   When the sample of size n = 197 is obtained, the values of the sample mean and standard

deviation will likely not be exactly those specified in the input to this program.

Consequently, the best the user can hope for is that the resulting precision percentage will be <u>approximately</u> 10%.

(2) For the preceding example, the specified precision was 10% of the point estimate. The point estimate for the universe total was 3,500,000,000. In the formula section, E is the desired precision amount expressed as a percentage of the point estimate for the universe total. Here this would be E = 350,000,000.

(3) For situations in which you do not have an estimate of the universe standard deviation (σ) from previous audit results, a rough approximation for σ can be obtained for each stratum by estimating (1) the largest value (L) that you expect to see in the sample for this stratum and (2) the smallest value (S) that you expect to see in this stratum. Then, the approximate value of σ for this stratum is $\hat{\sigma} = \dfrac{L - S}{4}$. In the previous example, if the largest audit amount that you expect to see in the LOW INCOME stratum is L = $15,000 and the smallest value is S = $1,000, then the estimated standard deviation is $\hat{\sigma}$ = (15,000 - 1,000)/4 = $3,500.

Stratified Sample Sizes - Total Sample Size is Known

Example 4. This is an illustration of situation 2. The situation is the same as that described in Example 3, which used two strata -- the high-income stratum and the low-income stratum. The total sample size is set at 500. The input screen on the following page was used for this example. Notice that the user is unable to set the precision percentages for this situation.

The following estimates were used as input to the program:

Stratum	Estimated Mean	Estimated Standard Deviation	Estimated Universe Size
High Income	10,000	5,000	100,000
Low Income	5,000	4,000	500,000

The program output on the next page is obtained. Notice that the resulting strata ratios (i.e., 20% and 80%) are identical to those obtained in Example 3.

```
                    DEPARTMENT OF HEALTH & HUMAN SERVICES
                        OIG - OFFICE OF AUDIT SERVICES
Date: 10/23/2004            Sample Size Determination              Time: 13:07

          THE ESTIMATES ARE BASED ON THE FOLLOWING ENTRIES:

NBR    DESCRIPTION         -- MEAN --    -- STD.DEV. --    -- UNIVERSE --
  1    High Income         10,000.00          5,000.00          100,000
  2    Low Income           5,000.00          4,000.00          500,000

- TOTALS -                  5,833.33          4,579.54          600,000
= = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = =

Precision Values:                          NOTE:  See the Discussion section.


Confidence Level  80%          90%            95%               99%
                 4.09%        5.25%          6.26%             8.22%

The following sample sizes are based on a total sample size of 500.

Stratum 1: High Income
  Sample
    Size       Ratio
     100      20.00%

Stratum 2: Low Income
  Sample
    Size       Ratio

     400      80.00%
```

Discussion. The two sample sizes are $n_1 = 100$ and $n_2 = 400$, which total n = 500. For this

example, $\sum N_i \hat{\sigma}_i$ is (100,000)(5,000) + (500,000)(4,000) = 2,500,000,000. Call this SUM.

The ratio value for stratum 1 is (100,000)(5,000) divided by SUM; that is .2. So, 20% of the

sample size is allocated to stratum 1; that is, n_1 is (500)(.2) = 100. Similarly, the ratio for stratum

2 is .8 and n_2 is (500)(.8) = 400. **NOTE**: This same discussion applies to Example 3.

What is the precision amount for this sampling design? This will be the value obtained by the

Stratified Variable Appraisal program using these sample sizes and estimated standard

deviations. This formula (borrowed from the **Stratified Variable Appraisal** formula section) is

contained in the formula section to follow. For this example, the precision amount will be

$$1.95996\sqrt{100{,}000^2\left(\frac{100{,}000-100}{100{,}000}\right)\frac{5{,}000^2}{100}+500{,}000^2\left(\frac{500{,}000-400}{500{,}000}\right)\frac{4{,}000^2}{400}}$$

$$=219{,}038{,}136.$$

The estimated universe total is

$$\hat{T} = \Sigma(\text{stratum mean})(\text{stratum size}) = (10{,}000)(100{,}000) + (5{,}000)(500{,}000) = 3{,}500{,}000{,}000.$$

The resulting precision percentage is

$$100 \cdot (219{,}038{,}136 / 3{,}500{,}000{,}000) = 6.26\%.$$

This value is called PERC in the formula section to follow and matches with the highlighted

value in the computer output.

FORMULAS

Total Sample Size (n) is Known

Notation

$L =$ Number of strata

$N_i =$ the universe size for the i-th stratum

$(StdDev)_i =$ estimated universe standard deviation for the i-th stratum

$$SUM = \sum_{i=1}^{L} N_i \cdot (StdDev)_i$$

$(Ratio)_i =$ $[N_i \cdot (StdDev)_i] / SUM$

The resulting sample size allocated to the i-th stratum is $n_i = n \cdot (Ratio)_i$.

Total Sample Size (n) is Unknown

Notation

L = Number of strata

N_i = the universe size for the i-th stratum

N = the total universe size = $\displaystyle\sum_{i=1}^{L} N_i$

$(Mean)_i$ = estimated universe mean for the i-th stratum

$UnivTotal$ = estimated universe total = $\displaystyle\sum_{i=1}^{L} N_i \cdot (Mean)_i$

$(StdDev)_i$ = estimated universe standard deviation for the i-th stratum

$$SUM1 = \sum_{i=1}^{L} N_i \cdot (StdDev)_i$$

$$SUM2 = \sum_{i=1}^{L} N_i \cdot (StdDev)_i^{2}$$

$(Ratio)_i$ = $[N_i \cdot (StdDev)_i] / SUM1$

$PREC$ = the precision percentage (e.g., 1 for 1%, 10 for 10%)

$ZVAL$ = the value from the standard normal (Z) distribution having a right-tail area equal to (100 - Confidence Level)/2, where the right-tail area is expressed as a proportion between zero and one.

ZVAL is 1.281551565545 (80%), 1.644853626951 (90%), 1.959963984540 (95%), and 2.575829303549 (99%).

E = the precision amount = $(PREC/100) \cdot (UnivTotal)$

For each selected value of PREC and ZVAL,

(1) the total sample size is

$$n = \frac{(SUM1)^2}{(E / ZVAL)^2 + SUM2}$$

(2) the sample size allocated to the i-th stratum is

$$n_i = n \cdot (Ratio)_i$$

Comments

1. In the preceding calculation, the value of n is treated as a floating point number (e.g., n = 487.263) and the strata sample sizes (n_i) are calculated using this value. The n_i values are then rounded up to the nearest integer. After all strata sample sizes have been determined, n is reset to the sum of the n_i.

2. If the computed sample size for stratum i (n_i) is larger than the universe size N_i, then n_i is set equal to N_i. The remaining sample sizes are then obtained by applying the above formula and (1) omitting the i-th stratum in the denominator and (2) replacing n with n - N_i (the total sample size for the remaining L-1 strata).

The precision percentage at the 95% confidence level is ± PERC, where

$$PERC = \frac{1.959963984540}{\hat{T}} \sqrt{\sum_{i=1}^{L} N_i^2 \left(\frac{N_i - n_i}{N_i} \right) \frac{(StdDev)_i^2}{n_i}}$$

and where $\hat{T}$ is the estimated total for the universe. The value of $\hat{T}$ is obtained by multiplying

N_i by the estimated mean for stratum i and summing over the L strata; that is, $\hat{T} = \sum_{i=1}^{L} N_i \hat{\mu}_i$

NOTE: Replace 1.959963984540 with 1.281551565545 for an 80% interval, 1.644853626951 for a 90% interval, and 2.575829303549 for a 99% interval.

Attribute Sample Size Determination

This program determines the sample size for an attribute simple random sample. The sample size is determined for specified degrees of precision (using the desired width of the confidence intervals) and for various levels of confidence. The resulting sample size is the smallest sample size capable of meeting the specified precision requirement at each of the specified confidence levels. The user may select any combination of the following confidence levels: 80%, 90%, 95%, and 99%.

Confidence intervals for attribute sampling are exact and are based on the hypergeometric distribution. As a result, such confidence intervals are usually not symmetric about the point estimate. For example, the point estimate might be 3% and the corresponding 95% confidence interval might be 2% to 6%. For this illustration, the width of the confidence interval is 4% and the confidence level is 95%. Consequently, attribute confidence intervals differ from the usual interval obtained by deriving the point estimate plus or minus the estimated precision, where the estimated precision is half the width of the resulting confidence interval. Because of this, the "desired precision" for the attribute sampling procedure must be specified as the desired width (rather than the half-width) of the confidence interval.

An approximate confidence interval for a universe proportion (discussed in many introductory statistics textbooks) is based on the normal approximation. This particular interval follows the "usual" procedure where the confidence interval is equal to (point estimate) ± (estimated precision); that is, this interval is symmetric about the point estimate. However, this confidence interval is approximate and is unreliable whenever the estimated proportion is very small or very

large, unless the sample size is extremely large. The confidence interval using the RAT-STATS

attribute sample size module discussed here is always <u>exact</u>.

The input screen includes (1) the size of the universe and (2) the anticipated rate of occurrence in

the universe. This rate of occurrence is generally estimated from past experience, either from

similar systems or a past review of this universe. If no information concerning the rate of

occurrence is available, the most conservative procedure is to specify 50% for this value. If the

actual rate of occurrence differs from the user-specified rate of occurrence, this in no way affects

the sample's validity but the resulting precision (confidence interval width) may not meet the

desired precision requirement.

Example 5. An audit is to be carried out using a universe of N = 10,000 documents to determine

what proportion (p) of the documents do not have the proper approval signature. A confidence

level of 95% will be used. It is estimated that 20% of the documents will not have the proper

signature. Consequently, the estimate of p is $\hat{p} = .20$.

NOTE: This may be a rough guess if little information regarding this estimate is available
             from previous audit experience. If the user has no idea as to the value of p, $\hat{p} = .5$
             should be used. This will produce the largest possible sample size (for fixed values
             of N and precision range) but the user will be guaranteed that the resulting confidence
             interval will meet the desired precision range.

Suppose that the desired precision range is 6%. This is equal to the desired value of (upper

confidence limit - lower confidence limit). If the confidence limits were symmetric about the

point estimate, the user would have specified the precision as ± 3% for this situation, where 3%

is half the width of the resulting confidence interval. Since the exact procedure used in this

program usually does not produce an interval symmetric about the point estimate, the user must

specify the desired total width of the confidence interval. The following input screen is used for

this example:

The resulting computer output (saved to a text file) is shown on the next page.

```
                    DEPARTMENT OF HEALTH & HUMAN SERVICES
                       OIG - OFFICE OF AUDIT SERVICES
Date: 10/19/2004        Sample Size Determination                Time: 8:46

                            Confidence Level
                80%         90%          95%              99%
Sample Size     314         488          666            1,079
```

Anticipated Rate of Occurrence: 20%

Desired Precision Range: 6%

Universe Size: 10,000

Explanation of Output

The output for each cell in the output table will consist of (1) the necessary sample size or (2) the text "- - -". The necessary sample size is the number of sample items necessary to obtain the specified sample precision at each confidence level. For example, in this illustration a sample size of 488 is necessary to obtain a confidence interval having a width of 6% using a 90% confidence level. If the calculated sample size is zero, a text value of "- - -" will appear in this cell.

Discussion. The necessary sample size (highlighted) is n = 666. As a result, after the sample of 666 is obtained, if the resulting point estimate is close to $\hat{p} = .20$, then the resulting 95% confidence interval for p should have a width approximately equal to .06 (such as .1710 to .2310, with a width of .2310 - .1710 = .06). If the resulting sample produced 133 documents not containing the proper signature, then the rate of occurrence in this sample would be 133/666; that is, 20%.

The resulting confidence interval will have a width equal to .06 (i.e., 6%). This can be seen in the computer output below, obtained using the **Unrestricted Attribute Appraisal** module. The

width of this 95% confidence interval is 23.10% -17.10% = 6% (the desired precision range).

Notice that $\hat{p} = .20$ (i.e., 20%) is inside this interval (it always is), but it is not in the center.

```
                    Department of Health and Human Services
                         OIG - Office of Audit Services
Date: 10/19/2004      Single Stage Attribute Appraisal        Time: 12:24
                          AUDIT/REVIEW: Example


        UNIVERSE SIZE                                          10,000
        SAMPLE SIZE                                            666
        CHARACTERISTIC(S) OF INTEREST
          QUANTITY IDENTIFIED IN SAMPLE                        133
          PROJECTED QUANTITY IN UNIVERSE                       1,997
          PERCENT                                              19.970%
        STANDARD ERROR
          PROJECTED QUANTITY                                   150
          PERCENT                                              1.498%

                               CONFIDENCE LIMITS
                              80% CONFIDENCE LEVEL
        LOWER LIMIT - QUANTITY              1,805
                      PERCENT               18.050%
        UPPER LIMIT - QUANTITY              2,202
                      PERCENT               22.020%

                              90% CONFIDENCE LEVEL
        LOWER LIMIT - QUANTITY              1,754
                      PERCENT               17.540%
        UPPER LIMIT - QUANTITY              2,259
                      PERCENT               22.590%

                              95% CONFIDENCE LEVEL
        LOWER LIMIT - QUANTITY              1,710
                      PERCENT               17.100%
        UPPER LIMIT - QUANTITY              2,310
                      PERCENT               23.100%
```

Example 6. Repeat Example 5 where no information is available regarding the proportion of documents not containing the proper signature.

Solution. Here, the user should enter 50% ($\hat{p} = .5$) in the Anticipated Error Rate box. The resulting computer output is shown below. The necessary sample size (highlighted) is now n = 991, approximately 50% larger than the previous sample size of 666.

```
                    DEPARTMENT OF HEALTH & HUMAN SERVICES
                       OIG - OFFICE OF AUDIT SERVICES
Date: 10/19/2004            Sample Size Determination            Time: 12:33


                          Confidence Level
              80%             90%             95%             99%

Sample Size    466            725             991            1,580

Anticipated Rate of Occurrence: 50%

Desired Precision Range: 6%

Universe Size: 10,000
```

Discussion. Example 6 illustrates how using $\hat{p} = .5$ produces a very large value of n. **The user should be encouraged to use even a rough guess for the value of $\hat{p}$**. Using $\hat{p} = .5$ is a very conservative procedure because with a sample of size n = 991, quite likely the resulting confidence interval will have a width considerably less than the desired precision range of 6%. To illustrate, the computer output below was obtained when the sample of 991 documents produced 248 not containing the proper signature. Here, $\hat{p} = 248/991 = .250$ and the confidence interval width (using the highlighted values in the following computer output) is 5.21%. This value is less than 6%, but the user did have the guarantee that this value would be no more than 6%.

```
                      Department of Health and Human Services
                           OIG - Office of Audit Services
Date: 10/19/2004         Single Stage Attribute Appraisal          Time: 12:43
                              AUDIT/REVIEW: Example


        UNIVERSE SIZE                                            10,000
        SAMPLE SIZE                                              991
        CHARACTERISTIC(S) OF INTEREST
          QUANTITY IDENTIFIED IN SAMPLE                          248
          PROJECTED QUANTITY IN UNIVERSE                         2,503
          PERCENT                                                25.025%
        STANDARD ERROR
          PROJECTED QUANTITY                                     131
          PERCENT                                                1.307%

                                CONFIDENCE LIMITS
                              80% CONFIDENCE LEVEL
        LOWER LIMIT - QUANTITY                  2,334
                      PERCENT                   23.340%
        UPPER LIMIT - QUANTITY                  2,678
                      PERCENT                   26.780%


                              90% CONFIDENCE LEVEL
        LOWER LIMIT - QUANTITY                  2,288
                      PERCENT                   22.880%
        UPPER LIMIT - QUANTITY                  2,727
                      PERCENT                   27.270%


                              95% CONFIDENCE LEVEL
        LOWER LIMIT - QUANTITY                  2,249
                      PERCENT                   22.490%
        UPPER LIMIT - QUANTITY                  2,770
                      PERCENT                   27.700%
```

FORMULAS

In the discussion to follow, a sample item having the attribute of interest will be referred to as an item "in error." Consequently, the universe proportion, p, will be the "error rate."

Consider the case where the specified confidence level is 95%. The upper limit of the 95% confidence interval for the universe <u>total</u> is, say, k_2, where k_2 is the <u>largest</u> value of k for which

$$\sum_{i=0}^{x} \frac{\binom{k}{i}\binom{N-k}{n-i}}{\binom{N}{n}} > .025$$

where N = universe size

   n = sample size

   k = total number of universe items in error

   x = number of sample items in error

  .025 = [1 - (confidence level)]/2

NOTE: Here, the "confidence level" is expressed as .95.

The lower limit of the 95% confidence interval is, say, k_1, where k_1 is the <u>smallest</u> value of k for which

$$\sum_{i=x}^{n} \frac{\binom{k}{i}\binom{N-k}{n-i}}{\binom{N}{n}} > .025$$

The resulting 95% confidence interval for the total number of universe items in error is k_1 to k_2.

The procedure used to derive this confidence interval can be found in John P. Buonaccorsi (1987), "A Note on Confidence Intervals for Proportions in Finite Populations," *The American Statistician*, Vol. 41, No. 3, pp. 215-218.

Suppose that the universe size is N = 10,000, the anticipated rate of occurrence (i.e., error rate) is 20%, and the desired precision range is 6%. Since (10,000)(.06) is 600, we know that $k_2 = k_1 + 600$; that is, the upper confidence limit must be 600 more than the lower limit. The anticipated rate of occurrence is used to specify the number of sample items that contain the characteristic of interest. Here, it would be 20% of n, where n is the sample size determined by this program.

For example, suppose that n = 300 and (300)(.20) = 60 (call this x). If the values, N = 10,000, n = 300, and x = 60 are used as input to the **Unrestricted Attribute Appraisal** program, the resulting 95% confidence interval for the universe proportion (p) has a lower limit of .1569 [i.e., $k_1 = (10,000)(.1569) = 1,569$] and an upper limit of .2490 [i.e., $k_2 = (10,000)(.2490) = 2,490$]. But this is not a satisfactory value of n since $k_2 - k_1 = 2,490 - 1,569 = 921$, which must equal 600 according to the previous discussion.

Summary of program procedure. For a specified confidence level of 95%, this program searches for the value of n that produces a confidence interval (k_1 to k_2) such that k_1 and k_2 satisfy the preceding two inequalities and $k_2 - k_1 = 600$, where, in general, 600 is equal to N · (desired precision range). For the preceding example, if n = 666, then (666)(.20) ≈ 133. If the values N = 10,000, n = 666, and x = 133 are used as input to the **Unrestricted Attribute Appraisal** module, the resulting 95% confidence interval for the universe proportion (p) has a lower limit of .1710 [i.e., $k_1 = (10,000)(.1710) = 1,710$] and an upper limit of .2310 [i.e.,

$k_2 = (10{,}000)(.2310) = 2{,}310$]. This <u>is</u> satisfactory, since $k_2 - k_1 = 600$ and the difference of the

two proportions is .06 (i.e., 6%).