Department of Health & Human Services
OIG - Office of Audit Services

RAT-STATS

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

Companion Manual

September 2001

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 that he/she finds helpful. While the RAT-STATS Users Guide gives descriptions of program input and output, the 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 Dr. Janet Fowler (jfowler@oig.hhs.gov). We are trying to make this manual error free, so any corrections will be appreciated.

Dr. Al Kvanli
University of North Texas
Denton, TX

Dr. Janet Fowler
Dept. of HHS, OIG
Washington, D.C.

ATTRIBUTE APPRAISALS

An attribute appraisal is carried out to estimate a particular universe proportion (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 total number of items in the universe (Np) that meet the criteria.

In an attribute sample, each sample item is either a yes response (did meet the criteria) or no response (did not meet the criteria). This version of RAT-STATS contains two 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

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 (see RAT-STATS User Guide, page 3-2). 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

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

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. 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</u> <u>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 1987 article. John P. Buonaccorsi (1987), "A Note on Confidence Intervals for Proportions in Finite Populations," *The American Statistician*, Vol. 43, No. 3, 215 - 218.

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 must be known. 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 (see RAT-STATS User Guide, page 3-11). A universe of 2500 Medicare claims is stratified into in-patient (Stratum 1) and out-patient (Stratum 2) claims. The universe sizes are N_1 = 1000 in-patient claims and N_2 = 1500 out-patient claims. Of interest, is the proportion, p, of claims in error (containing improper charges).

A random sample of n_1 = 100 in-patient claims revealed x_1 = 2 errors and a random sample of n_2 = 100 out-patient 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 1000 were obtained for stratum 1 and 100 random numbers between 1 and 1500 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: 12/7/2000   STRATIFIED ATTRIBUTE APPRAISAL        Time: 10:55
                AUDIT/REVIEW: Attribute - Stratified
```

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

STRATUM	PRECISION AT 80% CL	PRECISION AT 90% CL	PRECISION 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 in-patient claims in error is (.02)(1000) = 20 and the projected number for the out-patient stratum is (.06)(1500) = 90. Consequently, the projected value for the universe is 20 + 90 =110 (highlighted) with a corresponding error rate of (110/2500) x 100% = 4.4% (highlighted).

A look at the in-patient 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 (i.e., 2%, here) in deriving a confidence interval. Consequently, the 90% confidence interval for the proportion of in-patient claims in error is 2% ± 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 in-patient claims in error is 2% ± 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 out-patient stratum: The estimated error rate is 6%. Continuing the discussion from the in-patient stratum, the 90% confidence interval for the proportion of out-patient claims in error is 6% ± 3.793% (highlighted); that is, 2.207% to 9.793%. The corresponding 95% confidence interval is 6% ± 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% ± 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% CL for the ith stratum, replace 1.644853626951 with 1.959963984540. 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 is 1.644853626951 · SE($\hat{p}$) .

The PRECISION AT 95% CL 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.

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.

There are a variety of procedures that can be used to obtain and appraise a variable sample. These sampling strategies are the same as those contained in the ATTRIBUTE APPRAISALS section. They are listed below and described in the sections to follow.

- Unrestricted
- Stratified

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/audit amounts, examined/difference amounts, or audit/difference amounts).

Example 1 (see RAT-STATS User Guide, page 4-5). An unrestricted sample of 50 items resulted in the 50 examined / audit values contained in data set DATASRS.TXT. For this sample, all of the resulting differences (examined value - audit value) were nonzero since all the examined (book) values were unequal to the corresponding audit (actual) values.

Data file DATASRS.TXT

1 300 267 μ
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, audit value separated by one or more spaces, a tab delimiter, or a comma. For samples containing two values (e.g., examined value and audit value) for each sample observation, the data file <u>must</u> be configured as shown here.

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 following output was obtained from the unrestricted variable appraisal program.

```
                    DEPARTMENT OF HEALTH & HUMAN SERVICES
                       OIG - OFFICE OF AUDIT SERVICES
Date: 4/5/2001          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
STANDARD ERROR                      41.88
SKEWNESS                              .32
KURTOSIS                             1.81
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
STANDARD ERROR                      36.14
SKEWNESS                              .30
KURTOSIS                             1.78
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
STANDARD ERROR                           6.81
SKEWNESS                                  .64
KURTOSIS                                 2.98
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 audit 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 is $48.2519\sqrt{\frac{10000-50}{(50)(10000)}}$ = 6.80677. 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 ± (10,000)(2.009575237129)(6.80677) = 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 PERCENT[1].

FORMULAS

$$\text{STANDARD DEVIATION} = s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

$$\text{STANDARD ERROR} = s\sqrt{\frac{N-n}{nN}}$$

where n = sample size, N = universe size

[1]When the POINT ESTIMATE is negative, the PRECISION PERCENT is set equal to zero.

$$\text{SKEWNESS} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x-\overline{x})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(x-\overline{x})^2\right]^{3/2}}$$

$$\text{KURTOSIS} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x-\overline{x})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(x-\overline{x})^2\right]^{2}}$$

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) $\hat{T} = \overline{x} \cdot N$

(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 must be known. The program will develop estimates for each stratum as well as for the entire universe.

Using a Stratified Sample

Discussion.

Purpose: To divide (partition) the universe into separate strata so that variation within individual strata is less than variation within the entire universe.

Simple Illustration:

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 3 strata:

{5 7 8 10}	{55 60 66 70}	{120 133 145 150}
#1	#2	#3

The strata variances are:

Stratum	Variance	
1	3.25	
2	32.69	µ 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.

<u>Reasons for Using Stratified Sampling</u>:

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 not introduce any bias into the final estimate since strata are defined prior 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 provided 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 highly correlated with the principal variable, such as using the number of hospital beds to measure the "size" of a hospital.

Basic rule: Select strata so that their means are as different as possible and their standard deviations are as small as possible.

Guidelines:

-- 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

-- courser divisions of several stratifying variables are preferable to finer divisions of one variable

-- it is better to use unrelated stratifying variables

Example 2 (see RAT-STATS User Guide, page 4-23). Random samples of size 25 were obtained from two strata:

Stratum 1: Examined amounts under $200 ($N_1$ = 5,200)

Stratum 2: Examined amounts $ $200 ($N_2$ = 3,500)

NOTE: These sample sizes are too small to meet OAS standards and are used for illustrative purposes only.

The sample difference amounts for the two strata are stored in data file DATASTRAT.TXT.

NOTE: The lines 9999 3E33 are used to indicate the end of the data for each stratum. The line number 9999 is arbitrary.

Data set DATASTRAT.TXT

1 80
2 43
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
9999 3E33
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
9999 3E33

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: 4/5/2001          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
           STANDARD ERROR                            5.25
           SKEWNESS                                  -.07
           KURTOSIS                                  2.24
           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	
	STANDARD ERROR	7.90	
	SKEWNESS	-.06	
	KURTOSIS	1.85	
	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

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).

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

$$5200^2\left(\frac{5200-25}{5200}\right)\frac{26.3317^2}{25} + 3500^2\left(\frac{3500-25}{25}\right)\frac{39.6432^2}{25}$$

$$= 1,510,906,287$$

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 ± 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

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)

$N = N_1 + N_2 +$ @@@ $+ N_L$

$\bar{y}_i$ = average of sample items in the i-th stratum

2. Estimate of universe total (T)

$$\hat{T} = N \cdot \bar{y}_{st} = N_1 \cdot \bar{y}_1 + N_2 \cdot \bar{y}_2 + \cdots + N_L \cdot \bar{y}_L$$

3. Estimated variance of $\bar{y}_{st}$

$$v(\bar{y}_{st}) = \frac{1}{N^2} \sum_{i=1}^{L} N_i^2 \left(\frac{N_i - n_i}{N_i} \right) \frac{s_i^2}{n_i}$$

where n_i = number of sampled items in i-th stratum

s_i^2 = sample variance for i-th stratum

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

$$v(\hat{T}) = N^2 v(\bar{y}_{st})$$

5. Approximate 95% confidence interval for universe mean (μ):

$\bar{y}_{st} \pm Z_{.025}\sqrt{v(\bar{y}_{st})}$ 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.

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
  - Stratified (Total Sample Size Known)
  - Stratified (Total Sample Size Unknown)
- ■ Attribute

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 non-homogeneous). Any combination of the confidence levels 80%, 90%, 95%, and 99% can be selected.

Variable Sample Size Determination - Unrestricted

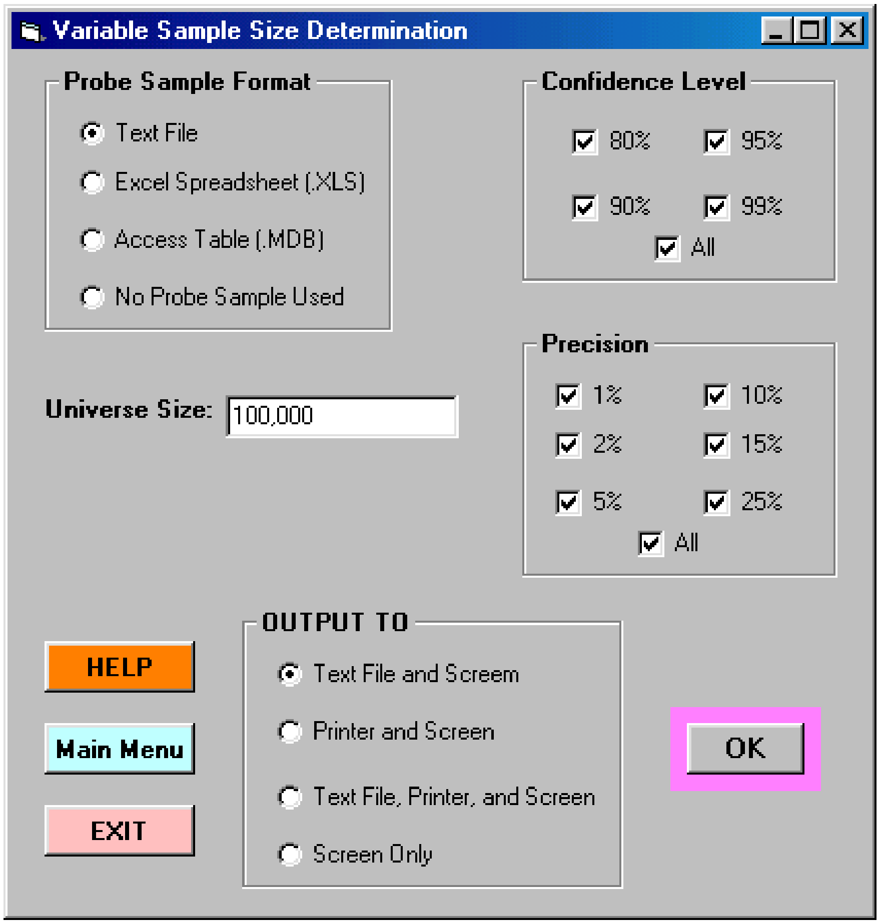
This program allows the user to estimate sample sizes for specified precision percentages and specified confidence levels. The user has the option of having the program read a probe sample file to obtain an estimate of the universe mean and standard deviation or input these two estimates directly without reading a probe sample file. The probe sample may be stored in a text file, an Excel spreadsheet, or an Access table.

Example 1 (See RAT-STATS User Guide, page 5-3). 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 $400.00 and the sample standard deviation is $50.00.

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: 5/11/2001 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 zero, 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 zero and one.
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.

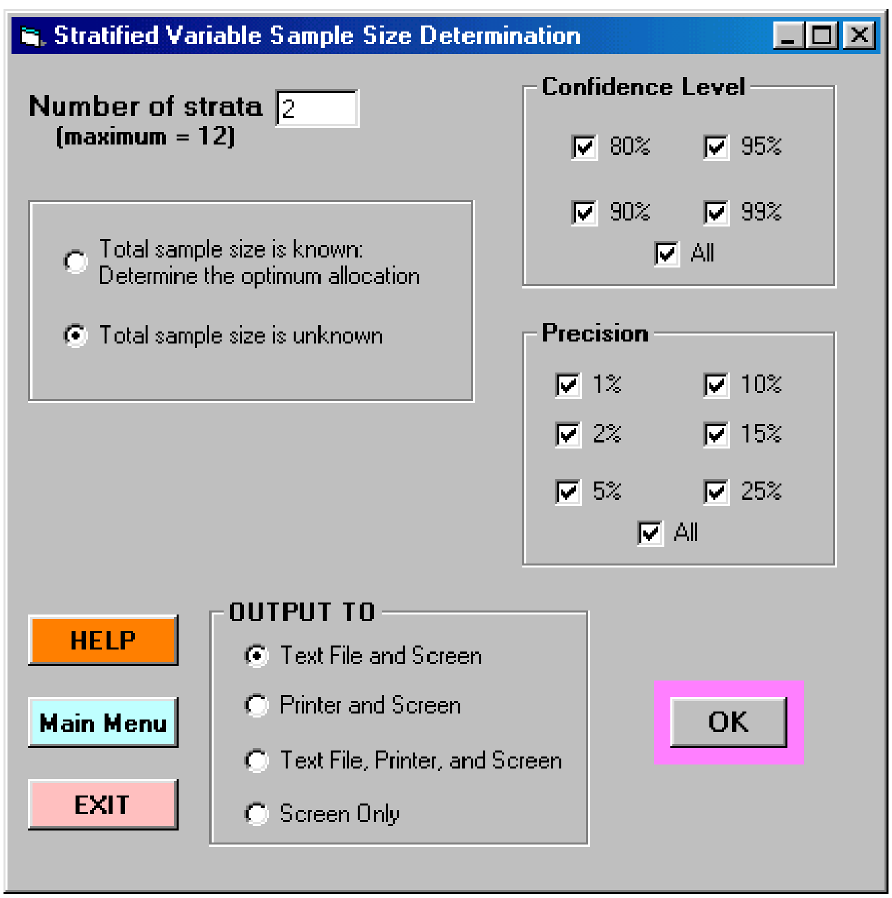
Variable Sample Size Determination - Stratified

Stratified Sample Sizes - Total Sample Size is Unknown

Example 2 (See RAT-STATS User Guide, page 5-16). 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: 12/19/2000             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 see 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: 12/23/2001         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 µ Approx. 5,000
           STANDARD ERROR                         790.42
           SKEWNESS                                 -.22
           KURTOSIS                                 2.30
           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 µ Approx. 4,000
           STANDARD ERROR                         319.17
           SKEWNESS                                  .87
           KURTOSIS                                 3.30
           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% µ	
	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.

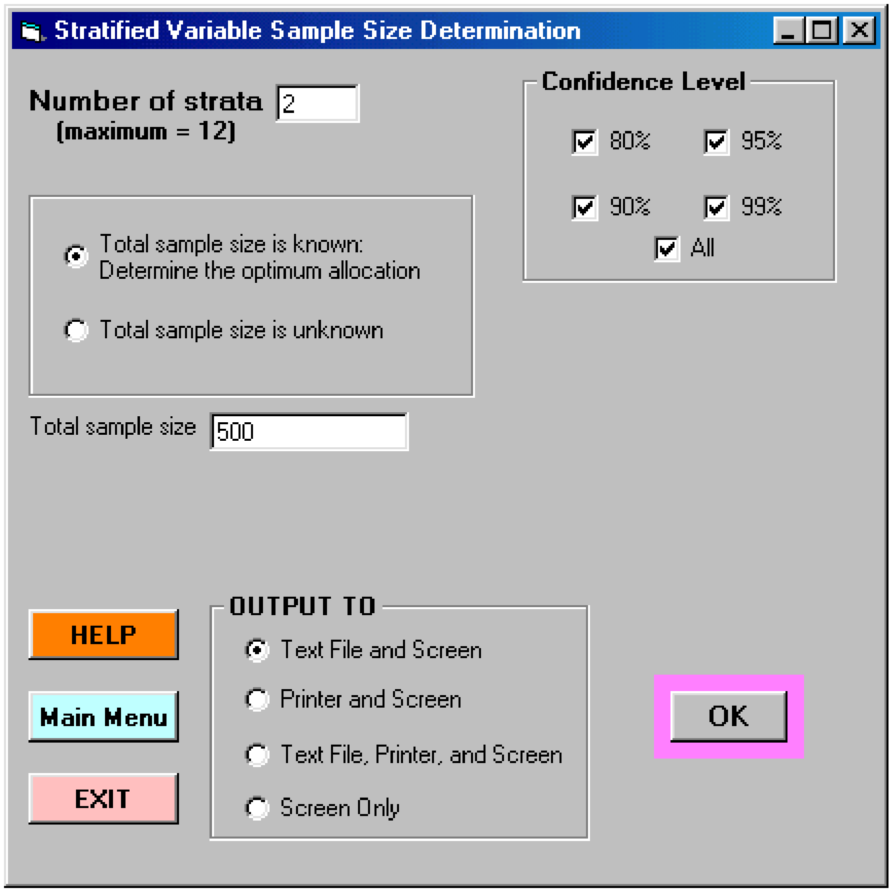
Consequently, the best you can hope for is that the resulting precision percentage will be approximately 10%.

(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 where 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 3 (See RAT-STATS User Guide, page 5-20). This is an illustration of situation 2. The situation is the same as that described in Example 2, 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 2.

```
                     DEPARTMENT OF HEALTH & HUMAN SERVICES
                         OIG - OFFICE OF AUDIT SERVICES
Date: 12/23/2000             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:

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                                     See the Discussion section.
    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 2.

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

$\hat{T}$ = Σ(stratum mean)(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

$$\text{SUM} = \sum_{i=1}^{L} N_i \cdot (StdDev)_i$$

$(Ratio)_i$ = $[N_i \cdot (StdDev)_i]$ / 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 = $\sum_{i=1}^{L} N_i$

$(Mean)_i$ = estimated universe mean for the i-th stratum

UnivTotal = estimated universe total = $\sum_{i=1}^{L} N_i \cdot (Mean)_i$

$(StdDev)_i$ = 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) · (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

$$n_i = n \cdot (\text{Ratio})_i$$

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 ith 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 $\hat{T} = \sum N_i \hat{\mu}_i$.

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.

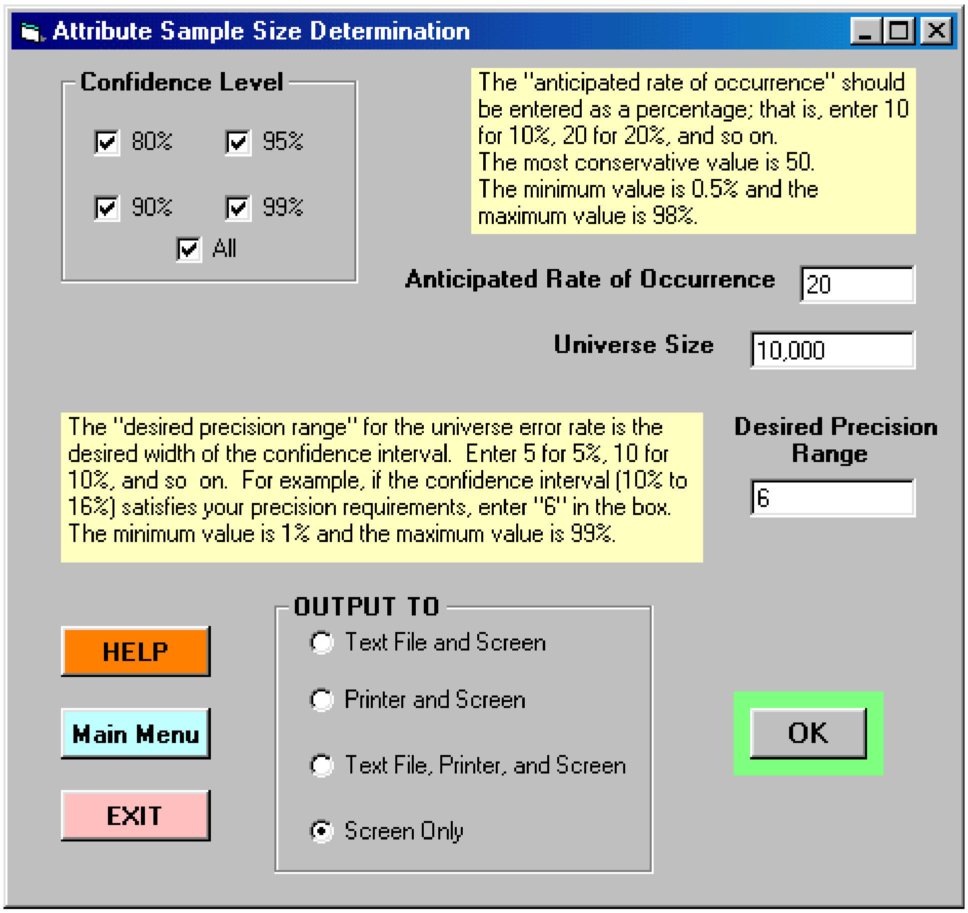
An approximate confidence interval for a universe proportion (discussed in many introductory statistics textbooks) is based on the normal approximation. This particular interval follows the "usual" procedure where the confidence interval is equal to (point estimate) ± (estimated precision); that is, 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 4 (See RAT-STATS User Guide, page 5-26). 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, use $\hat{p}$ = .5. 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: 12/19/2000          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: 2/13/2001         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                                                      1.497%

                                             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 5. Repeat Example 4 where no information is available regarding the proportion of documents not containing the proper signature.

Solution. Here, the user should use 50% ($\hat{p}$ = .5) as the value to enter 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: 2/13/2001            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 5 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 where 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: 2/13/2001        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                                                        1.306%

                                           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 total is say, k_2, where k_2 is the largest 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 smallest 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 the following 1987 article.

John P. Buonaccorsi (1987), "A Note on Confidence Intervals for Proportions in Finite Populations," *The American Statistician*, Vol. 43, No. 3, 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, then 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 is satisfactory, since k_2 - k_1 = 600 and the difference of the two proportions is .06 (i.e., 6%).